A multifaceted study of online news diversity: issues and methods

Emmanuel Marty, Nikos Smyrnaios and Franck Rebillard

In Ramón Salaverría (ed.), *Diversity of Journalisms*, Proceedings of the ECREA Journalism Studies Section and 26th International Conference of Communication (CICOM) at University of Navarra, Pamplona, 4-5 July 2011, p. 228-242

Abstract

Evaluating the pluralism of online journalism is both a major issue for democracy and an academic challenge for researchers. The multiplicity of online news outlets and the complexity of news consumption patterns make it particularly difficult to estimate the degree of pluralism that the web is supposed to embody. In 2010 and 2011 we carried out a research project called IPRI (Internet, Pluralism and Redundancy of Information) that combined quantitative and qualitative methods and aimed at measuring the diversity of online news in France through a transdisciplinary study. The purpose of this paper is to present the theoretical and methodological issues of our study.

Keywords: Agenda-setting, pluralism, diversity, redundancy, media, journalism, methodology

Evaluating the pluralism of online journalism is both a major issue for democracy and an academic challenge for researchers. The multiplicity of online news outlets and the complexity of news consumption patterns make it particularly difficult to estimate the degree of pluralism that the web is supposed to embody. Thus, only a combination of multileveled analyses combining quantitative and qualitative research methods is able to give a satisfying answer. Our research project called IPRI (Internet, Pluralism and Redundancy of Information)[1] aimed at measuring the diversity of online news in France, through a transdisciplinary study of several categories of websites (online media, portals, blogs, pure-players). Our purpose here is to present the theoretical issues and the methods of our study.

Theoretical framework

The degree of pluralism of opinions in the public sphere is a major political and social issue that greatly depends on the diversity of journalism and media: the more diverse are the media, the better will the public be informed on current affairs and social stakes. This ideal type of the public sphere, largely derived from the work of Jürgen Habermas (1991), has inspired journalists, essayists as well as policy makers into considering the internet as a sort of paragon of democracy (Hindman, 2009). The reasoning is simple: since online publication is much easier and cheaper than traditional forms of mass communication, then mediated public expression becomes affordable to every citizen that has access to the internet. This argument has been used for instance in several governmental reports (Lancelot 2005; Tessier 2007) but also in reports produced by supranational organizations such as the UN and the EU on the advent of the so called “ information society ” (Bangemann, 1994; UN, 2005). More generally, the web is seen as a means for marginal cultural and informational products to reach easily a larger public than they do through traditional distribution channels, to the point of overtopping traditional best-sellers. This idea was particularly popularized by Chris Anderson and his Long Tail theory (Anderson, 2006).

This seducing but somewhat simplistic vision has been since challenged by empirical research (Benghozi & Benhamou, 2010; Elberse & Oberholzer-Gee, 2008; Brynjolfsson et al., 2006). Among other things this deterministic approach ignores the fact that, over the last years, the internet has become a field of fierce competition between social groups, political organizations and giant corporations for the distribution of power to control the digital communication outlets (Mansell, 2004). As a result, the contemporary online news sector is the result of a complex set of relations established between professional media, amateur content producing communities and powerful intermediaries such as Google (Rebillard & Smyrnaios, 2010). From this point of view, the electronic public sphere is more likely to be considered as a conflicting arena, rather than a peaceful marketplace of ideas (Peters, 2004), where news and politics embody rival editorial, political and industrial strategies (Mosco, 2009; Fenton, 2009). Thus, online news diversity depends greatly on the outcome of those strategies and the balance of power between players.

Furthermore, such a complex media environment is characterized by two other particular aspects, compared to traditional media: first, the enormous quantity of information daily produced and reproduced online by a large spectrum of entities provokes a situation of oversupply – a tendency that is intrinsic to the cultural industries, but has been exacerbated on the web (Hesmondhalgh, 2007); secondly, this information is systematically computerized, that is stored and/or processed in systems of computers and networks. This digital content supply is then provided to millions of users around the globe through a multiplicity of channels and tools (rss feeds, search engines, social networks, personalized portals, blogs etc.) that allow complex social interaction – from mere transmission to in depth transformation of the information (Im et al., 2011). This multilayered process of production and circulation of online news produces coexisting and contradictory tendencies balancing between redundancy and pluralism (Smyrnaios et al., 2010).

Thus, if one aims in a deep and comprehensive study of such a complex and challenging issue as the question of online news diversity, traditional methods in social science fall short. In order to harness research grounds that produce vast quantities of data one needs to combine an approach moored in humanities and social sciences with automated computerized methods. One set of methods in this field is that of classic quantitative content analysis applied on media messages (Riffe et al., 2005). Another set of innovative methods is provided by the so-called "digital methods" (Rogers, 2009).

Content analysis

Quantitative computerized methods have been used at least since the 90's in order to analyze large corpuses of media content. Such approaches aimed for example in exposing journalistic biases in news coverage that can be explained by corporate ownership of media (Gilens & Hertzman, 2000), culturally and nationally bound journalistic practices (Brossard et al., 2004) or journalist gender differences (Rodgers & Thorson, 2003). The rise of perpetually evolving online content brought about the need for tools that can perform content analysis close to real-time (Krstajic & al., 2010). Furthermore, computerized content analysis has also been applied in order to test the agenda setting effect (Meijer & Kleinnijenhuis, 2006) as defined in media studies (Dearing & Rogers, 1992). Other research focused on methodological problems. For instance Jörg Matthes and Matthias Kohring (2008) provided an alternative procedure that aims at improving reliability and validity of content analysis on media frames, based on the definition advanced by R. M. Entman (1993). For Baumgartner and Mahoney, « methodological advances in computer science now allow much greater use of complex analytic schemes, assisted by computer technologies (not driven by them) to measure the relative use of different frames by different actors in the process. » (2008: 447).) In a comparable approach, Thomas Koenig (2006) introduced a step-by-step program for frame identification and measurement. Frames are identified with analysis techniques borrowed from sociolinguistics, discourse analysis, and content analysis in order to perform international comparative studies. Koening insists on the epistemological necessity of combining quantitative and qualitative methods in content analysis. Indeed, purely quantitative methods tend "to be deductive in that theory-based categorical schemes and coding rules are developed before conducting the analysis of data from subjects or documents" (Waltz et al., 2010: 279). Moreover, such procedures may be deeply positivist assuming that concepts should arise unmediated from the data and falsely "objective" (Koenig, 2005).

Digital methods

One of the most active fields of research that puts in use computerized data is the "new science of networks" that combines a long tradition of network analysis in social science with graph theory in discrete mathematics (Watts, 2004). The basic idea of this approach is that real-world networks are both partly ordered and partly random and that some of their properties can be embodied by mathematical models. In addition to that, the structure of a network is not considered as neutral, but rather as having major implications for the collective dynamics of the system the network represents. This theoretical framework has been applied in numerous empirical studies including metabolic reaction networks, biological neural networks or transportation and information networks. Henceforth, the rise of the internet as a mass media provides researchers in social sciences with a critical and abundant material because it concentrates and records mediated human activity in an unprecedented scale.

According to Richard Rogers, alongside with classic social scientific armature, like interviews, surveys and observations should be applied natively digital methods that ground claims about cultural change and societal conditions in online dynamics (Rogers, 2010). This consists in inverting dominant epistemological approaches and asking what claims about reality may be made on the basis of digital measures. Such an approach allows large-scale analyses of big corpuses, which suppose a certain level of abstraction, while still considering the data as *indices* in the sense of Charles S. Peirce: in this case they take the form of "digital traces" of social activity left on web servers.

Empirical grounds

Recently, number of scientific surveys has applied a mix of digital methods and traditional methods in social sciences in order to address issues related to online news circulation and diversity. For instance, Serena Carpenter has carried out a comparative study on content diversity in order to determine whether online citizen journalism and online newspaper publications were serving this function in the USA (Carpenter, 2010). Even though in this case the process of collecting content online was largely automated, the actual analysis was mostly manual. Undoubtedly, this constraint impacts the quality of the results because of the disproportion that emerges between a very voluminous corpus on the one hand and qualitative analysis on the other, that limits itself in the identification of big news stories. Leskovec et al. (2009) showed how a meme-tracking approach can provide with a representation of what the authors call the "news cycle", meaning the patterns of news circulation through websites, blogs and social networks over time. In this case, both the tracking and the analysis methods were computerized and allowed the authors to process a huge amount of data (90 million news articles).

This kind of primarily quantitative approach of online information was boosted recently by the development of Twitter. Many quantitative studies conducted through automated data extraction from the Twitter API are designed to map the flow of information inside the network and classify user groups and message groups (Cha et al., 2010; Kwak et al., 2010). Others process huge corpuses extracted from Twitter in order to examine the structure of distribution of connections and activity among the network's members (Huberman et al.; 2010, Heil & Piskorski, 2009; Krishnamurthy et al. 2008; Java et al., 2007).

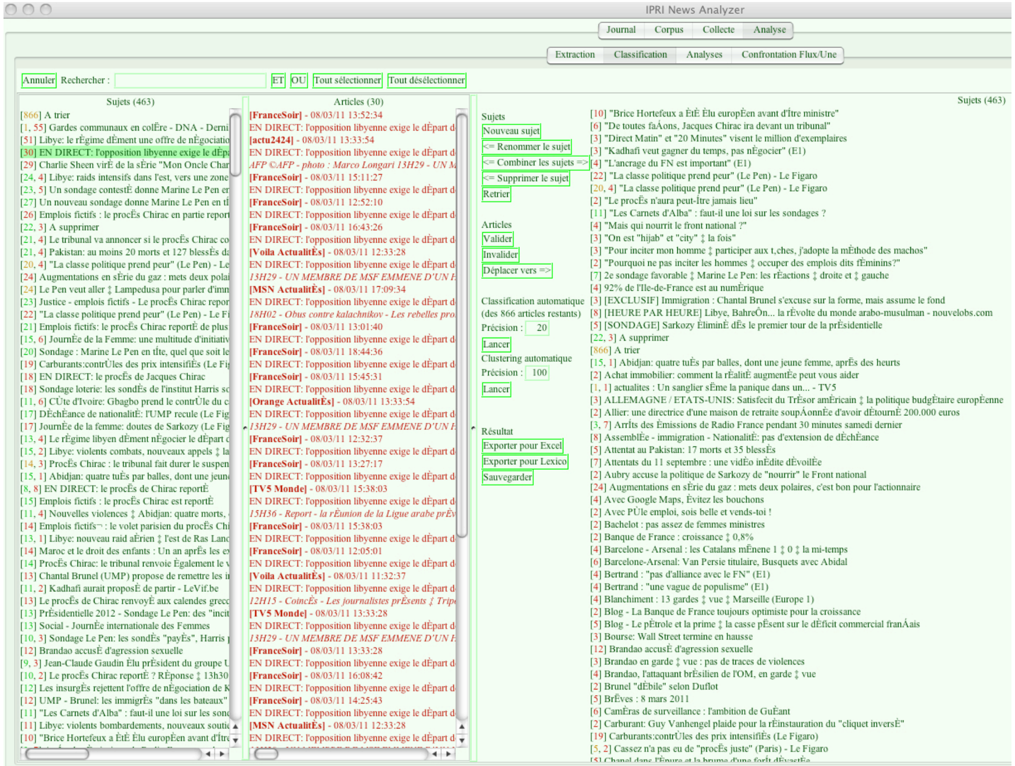
Closer to our subject, a recent study by Asur et al. (2010) explicitly tried to test the agenda-setting hypothesis and its time patterns on data collected through the Twitter API. The findings showed that there are few topics that last for a long time in users' messages, while most of them fade out rapidly. They also revealed that traditional notions of user influence such as the frequency of posting and the number of followers are not the main drivers of trends. What actually triggers a long lasting trend on Twitter over a certain issue is the dominant position of this issue in the media agenda. In this respect, Twitter, and social media in general, behave as a "selective amplifiers" for the content produced by traditional media. Similar findings were also made by Yang and Leskovec (2011) regarding time patterns of dissemination of news on social networks and agenda setting effects of traditional media. Their study revealed that both the adoption of *hashtags* in Twitter and the propagation of quoted phrases on the web exhibit nearly identical temporal patterns. For instance press agency news exhibits a very rapid rise followed by a relatively slow decay, whereas news stories that are discussed by bloggers may experience several rebounds in popularity.

These empirical implementations of the theoretical frame of the new science of networks and of automated content analysis produces interesting results but at the same time raises serious epistemological and methodological questions from the social scientist's point of view. Indeed, frequently there is a tendency of overreliance on concepts, categories and figures provided by the network operators themselves (e.g. Twitter's Trending Topics, Google Trends' Unique Visitors or Facebook's Like Buttons statistics). Yet for a humanity scholar these are artifacts that need to be deconstructed and explained. If they are completely opaque, like the examples mentioned above, the researcher finds himself confronted to a technological black-box effect (Rieder & Röhle, 2010). Similarly, not all digital traces left on the web have the same value as explanatory tools of social practices: writing a long blog post isn't the same thing as poking a Facebook "friend" or performing a search on Google. This means that the race to the "biggest corpus" is useless without suitable reflection prior to the collection of data on what one measures and why. Finally, as Duncan Watts points out, interpreting empirical data of this nature is tricky: "in a symbolic relationship it is frequently unclear how network metrics such as degree, path length, or centrality should be interpreted with respect to their consequences for some particular social, physical, or biological process (…), these relationships involve different kinds of social interactions, but because the interactions themselves are underspecified, the network alone cannot be said to reveal much about actual social processes" (2004: 254). This means that real-world networks metrics should systematically be put in perspective in regard to the socio-economic determinants of social relations.

A case study: the IPRI research project

In 2010 and 2011 we carried out a research project called IPRI (Internet, Pluralism and Redundancy of Information) aimed at measuring the diversity of online news in France through a transdisciplinary study. Its main aspect was a quantitative analysis of a sample of thousands of articles: we created a software called IPRI News Analyzer (IPRI-NA)[2] to collect and process automatically headlines from tenths of news sites through rss feeds. We then developed a semi-manual classification method based on the data collected by IPRI-NA as means to test the agenda-setting effect in online news. This revealed the variety of issues and the types of websites generating diversity, compared to those leading to redundancy. We then doubled the sample of news headlines composed by IPRI-NA with a collection of tweets produced by a sample of French users in order to test potential discrepancies between online news agenda setting and Twitter users' preferences. In order to deepen our overview of French online journalism landscape, several other analyses were led. First, the "offered pluralism", as measured by the initial quantitative study, was confronted to "consumed pluralism" through traffic analysis based on statistics of news sites audiences. Then, a qualitative study of full text news articles allowed us to identify the use of particular media frames, strategic cues and linguistic routines. Finally, we made a comparison between our results concerning the internet and another study related to TV news, to understand how the internet contributes to the diversity of news in a larger media landscape.

Figure 1: a capture of IPRI News Analyzer



Sujets (463)
[866] A trier
[1, 55] Gardes communaux en colËre - DNA - Derni
[51] Libye: le rÈgime dÈment une offre de nÈgociatio
[30] EN DIRECT: l'opposition libyenne exige le dÈpa
[29] Charlie Sheen virÈ de la sÈrie "Mon Oncle Char
[24, 4] Libye: raids intensifs dans l'est, vers une zone
[23, 5] Un sondage contestÈ donne Marine Le Pen en
[27] Un nouveau sondage donne Marine Le Pen en tÍ
[26] Emplois fictifs : le procËs Chirac en partie report
[22, 3] A supprimer
[21, 4] Le tribunal va annoncer si le procËs Chirac co
[21, 4] Pakistan: au moins 20 morts et 127 blessÈs da
[20, 4] "La classe politique prend peur" (Le Pen) - Le
[24] Augmentations en sÈrie du gaz : mets deux polai
[24] Le Pen veut aller ‡ Lampedusa pour parler d'imm
[23] Justice - emplois fictifs - Le procËs Chirac repor
[22] "La classe politique prend peur" (Le Pen) - Le Fi
[21] Emplois fictifs: le procËs Chirac reportÈ de plus
[15, 6] JournÈe de la Femme: une multitude d'initiativ
[20] Sondage : Marine Le Pen en tÍte, quel que soit le
[19] Carburants:contrÙles des prix intensifiÈs (Le Fig
[18] EN DIRECT: le procËs de Jacques Chirac
[18] Sondage loterie: les sondÈs de l'institut Harris so
[11, 6] CÙte d'Ivoire: Gbagbo prend le contrÙle du c
[17] DÈchÈance de nationalitÈ: l'UMP recule (Le Fig
[17] JournÈe de la femme: doutes de Sarkozy (Le Fig
[13, 4] Le rÈgime libyen dÈment nÈgocier le dÈpart d
[15, 2] Libye: violents combats, nouveaux appels ‡ la
[14, 3] ProcËs Chirac : le tribunal fait durer le suspen
[15, 1] Abidjan: quatre tuÈs par balles, dont une jeune
[8, 8] EN DIRECT: le procËs de Chirac reportÈ
[15] Emplois fictifs : le procËs Chirac est reportÈ
[11, 4] Nouvelles violences ‡ Abidjan: quatre morts,
[14] Emplois fictifs¬ : le volet parisien du procËs Chi
[13, 1] Libye: nouveau raid aÈrien ‡ l'est de Ras Lanc
[14] Maroc et le droit des enfants : Un an aprËs les ex
[14] ProcËs Chirac: le tribunal renvoie Ègalement le v
[13] Chantal Brunel (UMP) propose de remettre les i
[11, 2] Kadhafi aurait proposÈ de partir - LeVif.be
[13] Le procËs de Chirac renvoyÈ aux calendes grecq
[13] PrÈsidentielle 2012 - Sondage Le Pen: des "incit
[13] Social - JournÈe internationale des Femmes
[10, 3] Sondage Le Pen: les sondÈs "payÈs", Harris p
[12] Brandao accusÈ d'agression sexuelle
[9, 3] Jean-Claude Gaudin Èlu prÈsident du groupe U
[10, 2] Le procËs Chirac reportÈ ? RÈponse ‡ 13h30
[12] Les insurgÈs rejettent l'offre de nÈgociation de K
[12] UMP - Brunel: les immigrÈs "dans les bateaux"
[11] "Les Carnets d'Alba" : faut-il une loi sur les sond
[11] Libye: violents bombardements, nouveaux soutie
[10] "Brice Hortefeux a ÈtÈ Èlu europÈen avant d'Ítre

Articles (30)
[FranceSoir] - 08/03/11 13:52:34
EN DIRECT: l'opposition libyenne exige le dÈpart d
[actu2424] - 08/03/11 13:33:54
EN DIRECT: l'opposition libyenne exige le dÈpart d
AFP ©AFP - photo : Marco Longari 13H29 - UN M
[FranceSoir] - 08/03/11 15:11:27
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 12:52:10
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 16:43:26
EN DIRECT: l'opposition libyenne exige le dÈpart d
[Voila ActualitÈs] - 08/03/11 12:33:28
EN DIRECT: l'opposition libyenne exige le dÈpart d
13H29 - UN MEMBRE DE MSF EMMENE D'UN H
[MSN ActualitÈs] - 08/03/11 17:09:34
EN DIRECT: l'opposition libyenne exige le dÈpart d
18H02 - Obus contre kalachnikov - Les rebelles pro
[FranceSoir] - 08/03/11 13:01:40
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 18:44:36
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 15:45:31
EN DIRECT: l'opposition libyenne exige le dÈpart d
[Orange ActualitÈs] - 08/03/11 13:33:54
EN DIRECT: l'opposition libyenne exige le dÈpart d
13H29 - UN MEMBRE DE MSF EMMENE D'UN H
[FranceSoir] - 08/03/11 12:32:37
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 13:27:17
EN DIRECT: l'opposition libyenne exige le dÈpart d
[TV5 Monde] - 08/03/11 15:38:03
EN DIRECT: l'opposition libyenne exige le dÈpart d
15H36 - Report - la rÈunion de la Ligue arabe prÈv
[FranceSoir] - 08/03/11 15:38:03
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 12:05:01
EN DIRECT: l'opposition libyenne exige le dÈpart d
[Voila ActualitÈs] - 08/03/11 11:32:37
EN DIRECT: l'opposition libyenne exige le dÈpart d
12H15 - CoincÈs - Les journalistes prÈsents ‡ Tripo
[TV5 Monde] - 08/03/11 13:33:28
EN DIRECT: l'opposition libyenne exige le dÈpart d
13H29 - UN MEMBRE DE MSF EMMENE D'UN H
[FranceSoir] - 08/03/11 13:33:28
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 16:08:42
EN DIRECT: l'opposition libyenne exige le dÈpart d
[FranceSoir] - 08/03/11 14:25:43
EN DIRECT: l'opposition libyenne exige le dÈpart d
[MSN ActualitÈs] - 08/03/11 12:33:28
EN DIRECT: l'opposition libyenne exige le dÈpart d

Classification automatique (des 866 articles restants)
Précision : 20
Clustering automatique
Précision : 100

Sujets (463)
[10] "Brice Hortefeux a ÈtÈ Èlu europÈen avant d'Ítre ministre"
[6] "De toutes faÁons, Jacques Chirac ira devant un tribunal"
[3] "Direct Matin" et "20 Minutes" visent le million d'exemplaires
[3] "Kadhafi veut gagner du temps, pas nÈgocier" (E1)
[4] "L'ancrage du FN est important" (E1)
[22] "La classe politique prend peur" (Le Pen) - Le Figaro
[20, 4] "La classe politique prend peur" (Le Pen) - Le Figaro
[2] "Le procËs n'aura peut-Ítre jamais lieu"
[11] "Les Carnets d'Alba" : faut-il une loi sur les sondages ?
[4] "Mais qui nourrit le front national ?"
[3] "On est "hijab" et "city" ‡ la fois"
[3] "Pour inciter mon homme ‡ participer aux t‚ches, j'adopte la mÈthode des machos"
[2] "Pourquoi ne pas inciter les hommes ‡ occuper des emplois dits fÈminins?"
[7] 2e sondage favorable ‡ Marine Le Pen: les rÈactions ‡ droite et ‡ gauche
[4] 92% de l'Ile-de-France est au numÈrique
[3] [EXCLUSIF] Immigration : Chantal Brunel s'excuse sur la forme, mais assume le fond
[8] [HEURE PAR HEURE] Libye, BahreÔn... la rÈvolte du monde arabo-musulman - nouvelobs.com
[5] [SONDAGE] Sarkozy ÈliminÈ dËs le premier tour de la prÈsidentielle
[22, 3] A supprimer
[866] A trier
[15, 1] Abidjan: quatre tuÈs par balles, dont une jeune femme, aprËs des heurts
[2] Achat immobilier: comment la rÈalitÈ augmentÈe peut vous aider
[1, 1] actualites : Un sanglier sËme la panique dans un... - TV5
[3] ALLEMAGNE / ETATS-UNIS: Satisfecit du TrÈsor amÈricain ‡ la politique budgÈtaire europÈenne
[2] Allier: une directrice d'une maison de retraite soupÁonnÈe d'avoir dÈtournÈ 200.000 euros
[3, 7] ArrÍts des Èmissions de Radio France pendant 30 minutes samedi dernier
[8] AssemblÈe - immigration - NationalitÈ: pas d'extension de dÈchÈance
[5] Attentat au Pakistan: 17 morts et 35 blessÈs
[7] Attentats du 11 septembre : une vidÈo inÈdite dÈvoilÈe
[2] Aubry accuse la politique de Sarkozy de "nourrir" le Front national
[24] Augmentations en sÈrie du gaz : mets deux polaires, c'est bon pour l'actionnaire
[4] Avec Google Maps, Èvitez les bouchons
[2] Avec PÙle emploi, sois belle et vends-toi !
[2] Bachelot : pas assez de femmes ministres
[2] Banque de France : croissance ‡ 0,8%
[4] Barcelone - Arsenal : les Catalans mËnene 1 ‡ 0 ‡ la mi-temps
[6] Barcelone-Arsenal: Van Persie titulaire, Busquets avec Abidal
[4] Bertrand : "pas d'alliance avec le FN" (E1)
[4] Bertrand : "une vague de populisme" (E1)
[4] Blanchiment : 13 gardes ‡ vue ‡ Marseille (Europe 1)
[2] Blog - La Banque de France toujours optimiste pour la croissance
[5] Blog - Le pÈtrole et la prime ‡ la casse pËsent sur le dÈficit commercial franÁais
[3] Bourse: Wall Street termine en hausse
[12] Brandao accusÈ d'agression sexuelle
[3] Brandao en garde ‡ vue : pas de traces de violences
[4] Brandao, l'attaquant brÈsilien de l'OM, en garde ‡ vue
[2] Brunel "dÈbile" selon Duflot
[5] BrËves : 8 mars 2011
[6] CamÈras de surveillance : l'ambition de GuÈant
[2] Carburant: Guy Vanhengel plaide pour la rÈinstauration du "cliquet inversÈ"
[19] Carburants:contrÙles des prix intensifiÈs (Le Figaro)
[5, 2] Cassez n'a pas eu de "procËs juste" (Paris) - Le Figaro
[5] Chanel dans l'Èpure et la brume d'une forÍt dÈvastÈe

Sampling process of sources

The first step was to choose the news websites that would be part of our study. The perimeter that we defined is that of French websites that cover mainly current affairs and politics. Consequently we excluded from our sample sources that focus on particular domains such as sport, finance or technology for example. Through a meticulous census of different directories we established an exhaustive list of 98 general interest French news websites composed as follows:

- 42 online media: digital outlets of traditional media firms such as newspapers, magazines, press agencies, television and radio stations. These structures employ professional journalists and are often part of conglomerates and media groups.

- 14 portals and aggregators: high-traffic websites, belonging to large corporations of the telecom and web services industry, which outsource their news pages to press agencies and other media (e.g. Yahoo, Orange or MSN) or package and deliver deep links to news content on third-party sites (e.g. Wikio and Google News).

- 42 pure-players: news websites without an offline counterpart that employ professional journalists (e.g. Slate.fr) and participatory journalism websites that publish user-generated content (e.g. Agoravox).

We added to this list 111 blogs that we selected out of several hundred that we discovered using the Navicrawler[3] software over blog directories. We then made sure through thorough observation that the selected blogs met the criteria of covering mainly current affairs and politics. The whole sample of sources in its final version included 209 news sites and blogs.

Data collection, topic identification and analysis

The second step was the data collection made via the IPRI-NA software. The software performed real-time crawling of the sources, using their rss feeds, in order to extract different sorts of data: the headlines of the articles, the first lines of the articles included in the rss feeds which we refer to as *descriptions*, the name of the source that published each article and the hour and date of publication. The crawling took place throughout March 2011 on a 24-hour basis. The average number of headlines collected by IPRI-NA each day was 3,500. The gathered data were processed in order to extract the most frequent *lemmatic keywords* (canonical forms of lexemes in the text) which gave us a diachronic and global overlook of the media agenda through the studied period.

The third step was the identification and measurement of the news agenda through a smaller period. We focused on a particular period of eleven days, between March 7 and 18 2011, in order to perform a more detailed statistical analysis of the headlines: we classified all the headlines published by our sample sources in that period according to the topic they were related to. Our definition of a 'topic' is that of an event that occurred in a specific spatiotemporal context. A 'topic' becomes a 'story' or an article after it has been recounted as such by journalists (Esquenazi, 2002; Ringoot & Rochard, 2005). A topic is much broader than a story, in the sense that it can be approached through different angles or frames, but still refers to the same facts. For example the headline « *Libya: violent battles in Ras Lanouf* » was classified in the topic « *Insurrection in Libya against Muammar Gaddafi's regime* ». In order to identify the different topics of the news agenda throughout the eleven days and to link each headline to one of those topics, we used a computer-assisted and inductive method. This classification allowed us to measure the number of articles dedicated to each topic. By proceeding in this way, we were able embrace the spectrum of issues covered by French news sites throughout the period of eleven days.

We carried out the operation of classification using IPRI-NA. First, we applied a clustering method to our corpus of article headlines on the basis of repeated phrase segments. Second, we overlapped the URLs of collected articles in our database with URLs of articles from Google News that were already clustered by that service. These two preliminary processes allowed us to spot the most redundant topics in our data and to attach an important number of headlines to them. Finally, it was necessary to carry out a manual categorization of headlines into less frequent topics that it was impossible to identify through a computerized method. Such a qualitative approach raises the question of arbitrariness. Nevertheless, this solution appeared to be the least biased since the nature of editorial content does not allow a purely automated categorization. This initial treatment led to the creation of a database including several parameters: the name of the web source, the article headlines and descriptions that it published, the hour and date of publication and the topic to which each headline referred to.

In the fourth and final step of this phase of our research, relying on this classification, we calculated the degree of headline distribution among topics, by applying the concepts of *variety* and *balance* that have been previously used in the field of cultural industries (Benhamou & Peltier, 2006; 2007). Variety in this case depends on the number of topics that we isolated in our sample of headlines. The more topics there are in the media agenda during a given period the more pluralistic this agenda can potentially be. Balance on the other hand depends on the number of headlines per topic. If a great number of headlines are concentrated in a few topics then the news agenda is redundant. Finally, we focused on the wording of the headlines. For this analysis we used textual statistics (Lebart et al., 1998), also called *lexicometrics*, to compare the vocabulary used by the different categories of sources, as previously described. The combination of this wide range of methods gives us a multifaceted view of the issue of pluralism in online news production. Our findings suggest that both tendencies, redundancy and diversity, appear to coexist in our sample, each one being enhanced by different categories of websites.

Qualitative analysis

As we saw previously tracking pluralism and redundancy in the online media discourse demands not only strictly quantitative methods, such as statistics about variety and balance, but also more qualitative methods based on discourse analysis and semiotics. An example of pertinent use of basically qualitative content analysis methods in order to record the evolution of the media agenda over time is the News Coverage Index of the Pew Research Center's Project for Excellence in Journalism (State of the Media, 2011). In complement we held such a qualitative analysis upon a full text corpus of all articles released in a single day and dealing with two particular topics « *Insurrection in Libya against Muammar Gaddafi's regime* » and « *Marine Le Pen tops French Presidential poll rating* ». This analysis was carried out through different steps: a manual collecting of the articles, a computer assisted pre-analysis leading to a rational sampling of representative articles, and finally a two-sided content analysis of discourse and image.

First, full text articles were gathered directly from their respective websites on the basis of keyword requests. The gathering consisted basically in a manual copy-paste of the text and a screenshot of the article (thanks to the software ScreenGrab[4]). Then, the textual corpus was pre-analyzed with the lexicometrics software IRAMUTEQ[5], based on hierarchical cluster analysis (for technical details see Reinert, 2007) in order to identify the different frames conveyed by the different articles. Based on this first analysis, having identified several frames through their specific use of particular words, we managed to process a rational sampling of the corpus, leading to a selection of twenty articles, representative of both the different media frames and the different categories of sites using those frames.

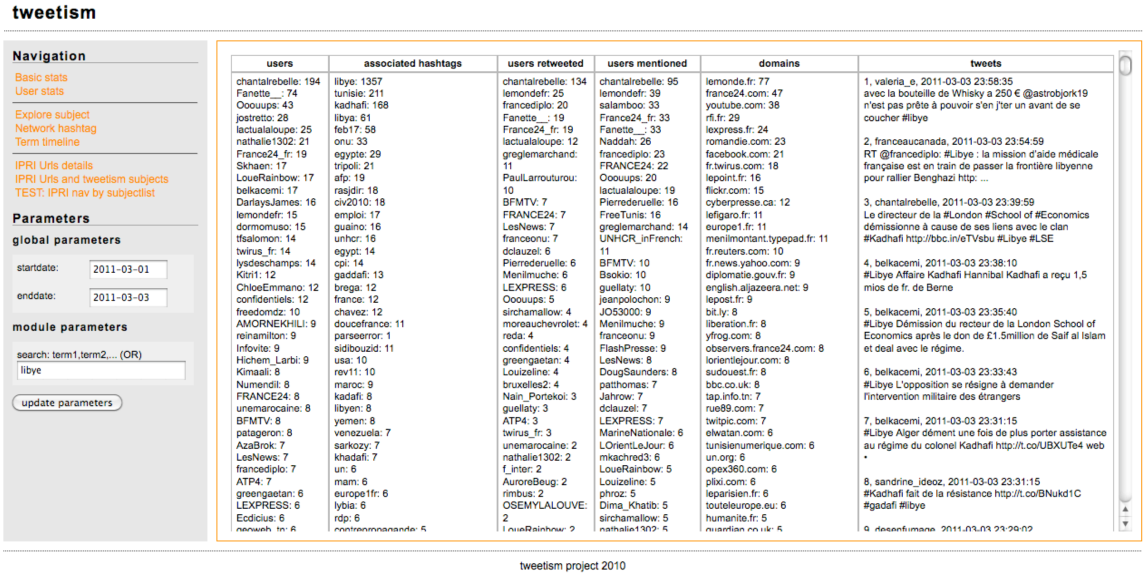
Finally, the qualitative analysis itself was made including textual approach of speech acts and semiotic analysis of images and of graphic settings of the web pages. The discursive analysis was mainly based on the identification of the "voices" taking on the different frames, the use of reported speech by journalists in their stories and media discursive strategies (narrative, descriptive or argumentative), linked to the distinction made by Iyengar (1991) between episodic and thematic framing. Concerning graphic and iconic settings of the sites, the analysis based on the screenshots aimed at identifying the weight of technical constraints on the graphic landscape of the sites as well as their interactive potentialities. It also qualified crucial editorial and journalistic choices, such as the type of picture used, their nature and their symbolic relation with the text. All these elements combined are seen as a means to create a particular relationship to the web users of the site that the study aimed at identifying.

Consumed diversity and social intermediation

A crucial factor concerning the question of diversity and pluralism is their dual nature (Benhamou & Peltier, 2006). On the one hand, there is the *diversity offered*, that is the spectrum of choices offered to the public. On the other hand, there is the *diversity consumed*, meaning the actual choices that the public makes among those that are possible. Even if the existence of diversity offered is a *sine qua non* condition for diversity consumed, it is not a sufficient one. In the case of an extreme concentration of online audiences into a small number of sources, even if the spectrum of online information is very wide, pluralism is not effective. Furthermore, consumers of online news are not limited in simple reception and interpretation of messages. They're also engaged in online activities, such as sharing, rebroadcasting and commenting news, that constitute techno-social intermediations.

In the first part of our study we focused on the problem of news diversity on the supply side. In order to take into account the consumption side, through consumed diversity and social intermediation, we associated the topics we found through our quantitative analysis of the online media agenda with traffic and unique visitors' statistics as well as with data collected on Twitter. Combining traffic data of the AT Internet site-centric solution with user-centric measures from Médiamétrie[6] we were able to determinate which news topics were the most popular ones among those present in our sample of news headlines. In other words, we were able to compare the online media agenda to the preferences of the public as they were measured by two different methods and institutes.

Figure 2: a capture of Tweetism



global parameters

startdate: 2011-03-01

enddate: 2011-03-03

module parameters

search: term1,term2,... (OR)
libye

users

chantalrebelle: 194
Fanette__: 74
Ooouups: 43
jostretto: 28
lactualaloupe: 25
nathalie1302: 21
France24_fr: 19
Skhaen: 17
LoueRainbow: 17
belkacemi: 17
DarlaysJames: 16
lemondefr: 15
dormomuso: 15
tfsalomon: 14
twirus_fr: 14
lysdeschamps: 14
Kitri1: 12
ChloeEmmano: 12
confidentiels: 12
freedomdz: 10
AMORNEKHILI: 9
reinamilton: 9
Infovite: 9
Hichem_Larbi: 9
Kimaali: 8
Numendil: 8
FRANCE24: 8
unemarocaine: 8
BFMTV: 8
patageron: 8
AzaBrok: 7
LesNews: 7
francediplo: 7
ATP4: 7
greengaetan: 6
LEXPRESS: 6
Ecdicius: 6
geoweb_tn: 6

associated hashtags

libye: 1357
tunisie: 211
kadhafi: 168
libya: 61
feb17: 58
onu: 33
egypte: 29
tripoli: 21
afp: 19
rasjdir: 18
civ2010: 18
emploi: 17
guaino: 16
unhcr: 16
egypt: 14
cpi: 14
gaddafi: 13
brega: 12
france: 12
chavez: 12
doucefrance: 11
parseerror: 1
sidibouzid: 11
usa: 10
rev11: 10
maroc: 9
kadafi: 8
libyen: 8
yemen: 8
venezuela: 7
sarkozy: 7
khadafi: 7
un: 6
mam: 6
europe1fr: 6
lybia: 6
rdp: 6
contrepropagande: 5

users retweeted

chantalrebelle: 134
lemondefr: 25
francediplo: 20
Fanette__: 19
France24_fr: 19
lactualaloupe: 12
greglemarchand: 11
PaulLarrouturou: 10
BFMTV: 7
FRANCE24: 7
LesNews: 7
franceonu: 7
dclauzel: 6
Pierrederuelle: 6
Menilmuche: 6
LEXPRESS: 6
Ooouups: 5
sirchamallow: 4
moreauchevrolet: 4
reda: 4
confidentiels: 4
greengaetan: 4
Louizeline: 4
bruxelles2: 4
Nain_Portekoi: 3
guellaty: 3
ATP4: 3
twirus_fr: 3
unemarocaine: 2
nathalie1302: 2
f_inter: 2
AuroreBeug: 2
rimbus: 2
OSEMYLALOUVE: 2
LoueRainbow: 2

users mentioned

chantalrebelle: 95
lemondefr: 39
salamboo: 33
France24_fr: 33
Fanette__: 33
Naddah: 26
francediplo: 23
FRANCE24: 22
Ooouups: 20
lactualaloupe: 19
Pierrederuelle: 16
FreeTunis: 16
greglemarchand: 14
UNHCR_inFrench: 11
BFMTV: 10
Bsokio: 10
guellaty: 10
jeanpolochon: 9
JO53000: 9
Menilmuche: 9
franceonu: 9
FlashPresse: 9
LesNews: 8
DougSaunders: 8
patthomas: 7
Jahrow: 7
dclauzel: 7
LEXPRESS: 7
MarineNationale: 6
LOrientLeJour: 6
mkachred3: 6
LoueRainbow: 5
Louizeline: 5
phroz: 5
Dima_Khatib: 5
sirchamallow: 5
nathalie1302: 5

domains

lemonde.fr: 77
france24.com: 47
youtube.com: 38
rfi.fr: 29
lexpress.fr: 24
romandie.com: 23
facebook.com: 21
fr.twirus.com: 18
lepoint.fr: 16
flickr.com: 15
cyberpresse.ca: 12
lefigaro.fr: 11
europe1.fr: 11
menilmontant.typepad.fr: 11
fr.reuters.com: 10
fr.news.yahoo.com: 9
diplomatie.gouv.fr: 9
english.aljazeera.net: 9
lepost.fr: 9
bit.ly: 8
liberation.fr: 8
yfrog.com: 8
observers.france24.com: 8
lorientlejour.com: 8
sudouest.fr: 8
bbc.co.uk: 8
tap.info.tn: 7
rue89.com: 7
twitpic.com: 7
elwatan.com: 6
tunisienumerique.com: 6
un.org: 6
opex360.com: 6
plixi.com: 6
leparisien.fr: 6
touteleurope.eu: 6
humanite.fr: 5
guardian.co.uk: 5

tweets

1, valeria_e, 2011-03-03 23:58:35
avec la bouteille de Whisky a 250 € @astrobjork19 n'est pas prête à pouvoir s'en j'ter un avant de se coucher #libye

2, franceaucanada, 2011-03-03 23:54:59
RT @francediplo: #Libye : la mission d'aide médicale française est en train de passer la frontière libyenne pour rallier Benghazi http: ...

3, chantalrebelle, 2011-03-03 23:39:59
Le directeur de la #London #School of #Economics démissionne à cause de ses liens avec le clan #Kadhafi http://bbc.in/eTVsbu #Libye #LSE

4, belkacemi, 2011-03-03 23:38:10
#Libye Affaire Kadhafi Hannibal Kadhafi a reçu 1,5 mios de fr. de Berne

5, belkacemi, 2011-03-03 23:35:40
#Libye Démission du recteur de la London School of Economics après le don de £1.5million de Saif al Islam et deal avec le régime.

6, belkacemi, 2011-03-03 23:33:43
#Libye L'opposition se résigne à demander l'intervention militaire des étrangers

7, belkacemi, 2011-03-03 23:31:15
#Libye Alger dément une fois de plus porter assistance au régime du colonel Kadhafi http://t.co/UBXUTe4 web
•

8, sandrine_ideoz, 2011-03-03 23:31:15
#Kadhafi fait de la résistance http://t.co/BNukd1C #gadafi #libye

9, desenfumage, 2011-03-03 23:29:02

tweetism project 2010

At the same time, we operated a similar comparison of the online news agenda to sharing preferences of French Twitter users using a software we created called Tweetism[7]. The sampling technique we opted for was a combination of manual and automated procedures. Our goal was not to obtain the most voluminous sample of tweets possible, but to meet a crucial criterion: collect the messages of users with special interest in French current affairs and politics. After an initial exploratory survey, we hand-picked a base sample of 400 accounts, composed of French journalists, politicians, and internet activists. A quick check showed that these highly visible individuals were, in fact, among the most mentioned and retweeted accounts in the French Twittersphere. Starting with these accounts, we explored the network vicinity by recuperating all friends and followers (n+1). To keep our sample manageable, we reduced this number through several techniques netting the sample to around 22 000 Twitter accounts. We then compared over the same period of one month the URLs of news articles shared by these users in Twitter to those collected initially by IPRI-NA and organized into topics. In this way we managed to shed light on discrepancies that arouse between what online media consider to be "big stories", thus much covered by many different sources, and what users share on Twitter. Each time we explored topic propagation on Twitter we examined in parallel the particular characteristics (nationality, language, profession, age, interests, network of friends) of the users that were actively sharing links to articles related to particular topics in order to identify patterns, profiles and key user groups.

Transmedia comparison and relative diversity

In a first version of this project (Smyrnaios et al., op. cit.) we met justified criticism on a crucial aspect of our interpretation method: measuring online news alone is not sufficient in order to obtain an evaluation of its relative diversity compared to other media. So the last facet of the study focused on a juxtaposition of that kind. Thanks to a collaboration with the Institut National de l'Audiovisuel (INA)[8], we were able to compare the online news agenda as we measured it to the one of French television. Since many years, the INA institute weekly publishes a list with the news topics (variety) treated by all the channels in France, established through a manual method. This list includes the time devoted to each one of the topics by all the TV news shows (balance). Such a comparison allowed us to put our results in perspective and open new horizons for future research.

References

Anderson, C. (2006). *The Long Tail*: *Why the Future of Business is Selling Less of More.* New York: Hyperion.

Asur S., Huberman B. A., Szabo G., & Wang C. (2011). Trends in Social Media : Persistance and Decay. Available at SSRN, February 5 2011 [http://ssrn.com/abstract=1755748].

Bangemann, M. (1994). *Europe and the Global Information Society, Recommendations to the European council*. Brussels: EU.

Baumgartner, F. R. & Mahoney, C. (2008). The Two Faces of Framing: Individual-Level Framing and Collective Issue Definition in the European Union. *European Union Politics*, 9 (3), p. 435-449.

Benghozi J.P., & Benhamou F. (2010). The Long Tail : Myth or reality ?. *Marketing Management, volume 12, number 3*, p. 43-53.

Benhamou, F. & Peltier, S. (2006). Une méthode multicritère d'évaluation de la diversité culturelle: application à l'édition de livres en France. In: Greffe X (ed.) *Création et diversité au miroir des industries culturelles*. Actes des journées d'économie de la culture. Paris: La Documentation Française, p. 313-344.

Benhamou, F. & Peltier, S. (2007). « How should cultural diversity be measured? An application using the French publishing industry. », *Journal of Cultural Economics*, 31(2), p.85–107.

Brossard, D, Shanahan, J., & McComas, K. (2004). Are Issue-Cycles Culturally Constructed? A Comparison of French and American Coverage of Global Climate Change.
Mass Communication and Society, Volume 7, Issue 3, p. 359-377.

Brynjolfsson, E., Hu, Y.J., & Smith, M. D. (2006). From Niches to Riches: Anatomy of the Long Tail. *Sloan Management Review, Vol. 47, No. 4*, p. 67-71.

Carpenter S. (2010). A study of content diversity in online citizen journalism and online newspaper articles. *New Media and Society, 12 (7)*, p. 1064-1084.

Cha, M.., Haddad, H., Benevenuto F. & Gummadi, K. P., (2010). « Measuring User Influence in Twitter: The Million Follower Fallacy », Association for the Advancement of Artificial Intelligence, *4th International Conference on Weblogs and Social Media*, May 23-26, George Washington University, Washington, DC.

Dearing, J.W. & Rogers, E.M. (1992). *Communication Concepts 6: Agenda-Setting.* Thousand Oaks, CA: Sage.

Elberse, A., & Oberholzer-Gee F. (2008). Superstars and Underdogs: An Examination of the Long Tail Phenomenon in Video Sales. *Harvard Business School Working Paper No. 07-015*.

Entman, R. M. (1993). *Framing: Toward Clarification of a Fractured Paradigm*. *Journal of Communication 43 (4)*, p. 51-58.

Esquenazi, J-P. (2002). *L'Écriture de l'actualité. Pour une sociologie du discours médiatique*. Grenoble: Presses Universitaires de Grenoble.

Fenton, N. (2009). *New Media, Old News: Journalism & Democracy in the Digital Age*. London: Sage Publications Ltd.

Gilens M., & Hertzman, G. (2000). Corporate Ownership and News Bias: Newspaper Coverage of the 1996 Telecommunications Act. *The Journal of Politics, 62*, p. 369-386.

Habermas, Jürgen (1991). *The Structural Transformation of the Public Sphere: An Inquiry into a category of Bourgeois Society*. Cambridge, MA: MIT Press.

Heil, B. & Piskorski M. (2009). *New Twitter Research: Men Follow Men and Nobody Tweets*, Working Paper, Harvard Business School.

Hesmondhalgh, D. (2007). *The Cultural Industries*. London : Sage Publications.

Hindman M. (2009). *The Myth of Digital Democracy*. Princeton, NJ and Oxford: Princeton University Press.

Huberman, B. A., Romero, D. M. & Wu F. (2009). « Social networks that matter: Twitter Under the microscope », *First Monday*, *Volume 14, Number 1 – 5*, January.

Im, Y. H., Kim, E. M., Kim, K. & Kim, Y. (2011). « The Emerging Mediascape, Same Old Theories? A Case Study of Online News Diffusion in Korea ». *New Media & Society*, available online ahead of print on March 2011 at [http://nms.sagepub.com/content/early/2010/12/16/1461444810377916.abstract]

Iyengar, S. (1991). *Is anyone responsible ? How television frames political issues*. Chicago: The University of Chicago Press.

Java, A., Finin T., Song X. & Tseng B. (2007). « Why We Twitter: Understanding Microblogging Usage and Communities », *9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis*, August 12, San Jose, California.

Koenig, T. (2005), *Identification and Measurement of Frames*. Available in March 2011 at [http://www.ccsr.ac.uk/methods/publications/frameanalysis/measurement.html]

Koenig, T. (2006). *Compounding mixed-methods problems in frame analysis through comparative research*. *Qualitative Research, vol. 6 no. 1*, p. 61-76.

Krishnamurthy, B., Gill, P. & Arlitt, M. (2008). « A few chirps about twitter », *1st ACM SIGCOMM Workshop on Social Networks*, Seattle, WA, August.

Krstajic, M., Mansmann, F. Stoffel, A., Atkinson, M. & Keim, D. A. (2010). « Processing online news streams for large-scale semantic analysis ». In *Data Engineering Workshops, 22nd International Conference on*, Los Alamitos, CA, USA: IEEE Computer Society, p. 215-220.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). « What is Twitter, a Social Network or a News Media?», *19th International World Wide Web Conference*, April 26-30, Raleigh NC (USA).

Lancelot, A. (2005). *Les problèmes de concentration dans le domaine des médias, Rapport pour le Premier ministre*. Paris: La Documentation française.

Lebart, L., Salem, A. & Berry, L. (1998). *Exploring Textual Data*. Dordrecht: Kluwer.

Leskovec J., Backstrom L., & Kleinberg J. (2009).“Meme-tracking and the dynamics of the news cycle”. *KDD’09 - International Conference on Knowledge Discovery and Data Mining*. Paris.

Lowry, D. & Xie, L. (2007). "Agenda-Setting and Framing by Topic Proximity: A New Technique for the Computerized Content Analysis of Network TV News Presidential Campaign Coverage". *Annual Meeting of the International Communication Association*, TBA, San Francisco, CA, May 23, 2007 .

Mansell, R. (2004). *Political economy, power and new media*. London: LSE Research Online.

Matthes, J. & Kohring, M. (2008). The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication, Volume 58, Issue 2*, p. 258–279.

Meijer, M. & Kleinnijenhuis, J. (206). Issue News and Corporate Reputation: Applying the Theories of Agenda Setting and Issue Ownership in the Field of Business Communication. *Journal of Communication, Volume 56, Issue 3*, p. 543–559.

Mosco, V. (2009). Review Essay: Approaching Digital Democracy. *New Media & Society*, Published online before print November 24.

Pew Research Center's Project for Excellence in Journalism, (2011). *State of the News Media Report*. Available in March 2011 at [http://stateofthemedia.org/]

Rebillard, F. & Smyrnaios, N. (2010). « Les infomédiaires au cœur de la filière de l’information en ligne. Les cas de Google, Wikio et Paperblog », *Réseaux n° 160-161/2010*, p. 163-194.

Reinert, M. (2007). Contenu des discours et approche statistique. In C. Gauzente, & D. Peyrat-Guillard (Eds.). *Analyse statistique de données textuelles en sciences de gestion*, Cormelles-le-Royal: EMS, p. 21-45.

Rieder, B. & Röhle, T. (2010). « Digital Methods : Five Challenges », *The Computational Turn Conference*, Swansea University.

Riffe, D., Lacy, S., & Fico, F.G. (2005). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*, London: Taylor & Francis, 2nd Edition.

Ringoot, R. & Rochard, Y. (2005). « Proximité éditoriale: normes et usages des genres journalistiques ». *Mots. Les langages du politique, 77*, p. 73–90.

Rodgers, S. & Thorson, E. (2003). A Socialization Perspective on Male and Female Reporting. *Journal of Communication, Volume 53, Issue 4*, p. 658–675.

Rogers, R. (2009). *The End of the Virtual: Digital Methods*, Amsterdam: Amsterdam University Press.

Smyrnaios, N., Marty, E. & Rebillard, F. (2010). Does the Long Tail apply to online news? A quantitative study of French-speaking news websites. *New Media and Society* 12 (8), p. 1244-1261.

Tessier, M. (2007). *La presse au défi du numérique,* Rapport pour le ministre de la Culture et de la Communication. Paris.

UN, (2005). *Tunis Agenda for the Information Society*, World Summit on the Information Society (WSIS) Available in March 2011 at [http://www.itu.int/wsis/documents/doc_multi.asp?lang=en&id=2267|0]

Waltz, C. F., Strickland, O. L. & Lenz, E. R. (2010). *Measurement in Nursing and Health Research*, Berlin: Springer.

Watts, D.J., (2004). The "New" Science of Networks. *Annual Review of Sociology, 30*, p. 243–70.

Yang, J. & Leskovec, J. (2011). "Patterns of temporal variation in online media", *WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining*, Hong Kong, February.

[1] The research programme IPRI (Internet, pluralisme et redondance de l'information) was supported by a grant from the Agence Nationale de la Recherche (ANR-09-JCJC-0125–01b). Several research teams specialized in media studies and computing science are involved in the program: CIM (University of Paris 3, France), ELICO (University of Lyon, France), LERASS (University of Toulouse 3, France), CRAPE (University of Rennes 1, France), GRICIS (UQAM – Montreal, Canada), LIRIS (INSA Lyon, France).

[2] The IPRI News Analyzer software was developed by Samuel Gesche, Elöd Egyed-Zsigmond and Cyril Laitang. It is distributed under a Creative Commons Licence http://liris.cnrs.fr/ipri/pmwiki/index.php?n=Public.IpriNA

[3] http://addons.mozilla.org/en-US/firefox/addon/navicrawler/

[4] http://www.screengrab.org/

[5] IRAMUTEQ is a free software developed by Pierre Ratinaud http://sourceforge.net/projects/iramuteq/

[6] http://www.mediametrie.fr/ and http://en.atinternet.com/

[7] Tweetism was developed by Raphaël Velt and Bernhard Rieder.

[8] The Institut National de l'Audiovisuel is a French Institute in charge of the archiving and promotion of mainly radio and TV broadcasting.