Published in Réseaux , 2012, vol. 30, no. 176, p. 29-74 which should be cited to refer to this work

Diversité et concentration de l'information sur le web. Une analyse à grande échelle des sites d'actualité français.

Emmanuel Marty (Univ. Nice), Franck Rebillard (Univ. Paris 3), Stéphanie Pouchot (HEG Genève), Thierry Lafouge (Univ. Lyon 1)

L'internet a pris une place croissante en tant que média d'information, et suscite à ce titre de nombreuses interrogations. Celles-ci sont largement motivées par la volonté de mieux cerner son rôle d'interface entre un espace public en possible reconfiguration (Flichy, 2008 ; Cardon, 2010 ; Miège, 2010) et des industries culturelles et médiatiques en prise avec de sensibles évolutions économiques et technologiques (Bouquillion, Matthews, 2010 ; Charon, 2010). Dans un tel contexte, la nature de l'information en ligne, et en particulier la diversité des contenus médiatiques offerts aux internautes, devient une question centrale : l'arrivée de nouveaux acteurs sur le terrain de l'information (industriels issus d'autres secteurs, amateurs profitant des facilités d'expression numérique) entraîne-t-elle une originalité accrue ou au contraire une certaine redondance des nouvelles ? Autrement dit, en matière d'information en ligne, la quantité est-elle synonyme de qualité ? La question du pluralisme de l'information, et des ses enjeux fondamentaux pour la vie démocratique, se trouve ainsi posée à nouveaux frais avec l'internet.

Elle se pose avec acuité sur le web, lieu majeur pour l'information d'actualité en ligne. Les espaces de publication s'y sont développés tous azimuts, empruntant autant la voie du journalisme participatif que de l'agrégation automatisée de nouvelles, aux côtés des sites de journaux, radios ou télévisions. En résulte-t-il pour autant une plus-value informationnelle, en termes d'originalité et de diversité de l'information ? A l'inverse, cette prolifération des espaces de publication sur le web n'a-t-elle pas engendré un vaste système de « re-traitement » d'une même matière première informationnelle (Rebillard, 2006) ? Le présent article se donne précisément pour objectif de déterminer si le pluralisme de l'information est, dans les faits, favorisé ou non avec la multiplication des sites web. Elle vient enrichir, pour le cas de la France, une littérature scientifique largement étoffée ces dernières années sur le plan international, passant d'un stade plutôt spéculatif à une dimension empirique de plus en plus aboutie.

Le web, opportunité ou menace pour la diversité de l'information ?

Le web constitue, indéniablement, un lieu potentiel de pluralisme pour l'information. Plusieurs chercheurs se sont en particulier intéressés à ce que l'amateurisme pouvait apporter à l'information en ligne (pour une synthèse, cf. Dagiral, Parasie, 2010), à travers l'étude des blogs (Serfaty, 2006), ou en questionnant les relations entre blogueurs et journalistes (Reese et al., 2007). Affirmant que les journalistes ne sont à présent plus les seuls maîtres de l'agenda médiatique en ligne, Bruns (2008) est l'un des auteurs les plus repris à ce sujet. Selon lui, le *gatekeeping* aurait fait place à un *gatewatching* : les internautes contributeurs auraient acquis une capacité de mobilisation collective à même d'influencer les choix opérés par les journalistes dans la sélection de l'information. Dans la même perspective, l'interactivité supposée de l'internet est considérée comme un facteur contribuant à mettre le débat démocratique et l'expression politique au premier plan de l'information médiatique. Ceci permettrait alors au citoyen de se forger une opinion sur le monde social, éventuellement de prendre part à un engagement politique (pour une synthèse, cf. Greffet, Wojcik, 2008).

L'internet cependant, bien loin d'une « *peaceful market-place of ideas* » telle que dénoncée par Peters (2004), constitue une arène où différents acteurs se livrent une compétition pour

l'accès à une tribune médiatique. Les contenus offerts aux internautes sont d'abord le résultat du travail réalisé par les acteurs de l'information en ligne. Et ils sont très souvent liés aux sources que constituent les services de communication des organisations et les agences de presse. Leur pouvoir de fixation de l'agenda est dans bien des cas inversement proportionnel à la capacité matérielle et temporelle des entités médiatiques à produire un contenu propre, ou à exploiter de manière créative le matériau discursif qui leur est fourni (Marty, 2010). Cette logique du système médiatique, aboutissant à une situation assez classique de « *circulation circulaire de l'information* » telle que vulgarisée par Bourdieu (1996), est rendue encore plus complexe sur l'internet : face au succès d'infomédiaires tels que *Google Actualités*, la politique des différents éditeurs est ambigüe, voire ambivalente, faisant se côtoyer mise en cause d'une concurrence considérée comme déloyale et souci presque obsessionnel d'un bon référencement, le tout pesant sur la nature des contenus ainsi produits (Smyrnaios, Rebillard, 2009).

« More is Less » ? La nécessaire analyse des informations offertes aux internautes.

Précisément, des travaux plus empiriques se sont employés à analyser les contenus d'actualité effectivement offerts sur le web. Tous semblent se rejoindre autour d'un constat assez semblable, résumé dans la formule-titre « More is Less » d'une des premières recherches à cet égard (Paterson, 2007) : davantage de sites certes, mais moins d'informations originales au final.

Chris Paterson s'est évertué dès le début des années 2000 à analyser les articles produits par les nouveaux entrants dans le domaine journalistique que constituaient les grands sites portails et agrégateurs (*AOL*, *Yahoo*, *Excite*, *AltaVista*, ...), et à les comparer avec les sites de médias plus établis (*CNN*, *BBC*, *ABC*, *SkyTV*, *New York Times*, ...). Il s'est avéré que ces sites reprenaient en grande partie des dépêches d'agence, et dans des proportions croissantes au fil du temps : de 2001 à 2006, la part de textes provenant de *Reuters* et d'*Associated Press* passe de 34% à 50% dans les articles publiés par les sites de médias, et de 68% à 85% dans les articles publiés par les portails et agrégateurs. Plus récemment, un groupe de chercheurs conduit par Natalie Fenton (2009) s'est penché sur les sites d'information britanniques. En sus des sites de médias imprimés et audiovisuels, et des portails/agrégateurs, ont été pris en compte des sites d'information alternative tels que *IndyMedia* ou *OpenDemocracy*. Ces derniers sont les seuls à ne pas suivre l'agenda médiatique dominant de l'année 2008 au Royaume-Uni, ils sont aussi les moins consultés. Les autres sites, à l'exception notable de celui de la chaîne participative *Current TV* et de la page "Have Your Say" de la *BBC*, présentent des contenus très fréquemment similaires (écrits comme images) et les traitent à partir d'un angle souvent unique. Les chercheurs relient cette similarité dans l'information analysée aux pratiques professionnelles de veille sur la concurrence observées parallèlement dans les salles de rédaction. La même pratique de *monitoring* a été mise en relief par Pablo Boczkowski (2010) lors de son enquête ethnographique au sein du quotidien argentin *Clarín*, et plus particulièrement dans les bureaux du service web *Ultimo Momento* jonchés d'écrans scrutant la concurrence jusqu'aux chaînes d'information en continu. Un phénomène d'imitation dans les contenus en découle : Boczkowski l'a observé en 2005 lors d'une comparaison entre les sites de *Clarín*, *La Nacion* (autre grand quotidien argentin) et *Infobae* (site d'information sans équivalent papier). Il le rattache également à l'anticipation d'une attente d'information factuelle et rapidement renouvelée en réception, en raison des habitudes de consultation de l'internet sur le lieu de travail. Une semblable incitation à la productivité dans l'information en ligne, et son corollaire consistant à s'appuyer essentiellement sur des dépêches d'agence ou des informations de seconde main, avaient été documentés par des enquêtes auprès de cinq sites d'information en Allemagne (Quandt, 2008).

C'est une même démarche, visant à mettre à l'épreuve des faits l'hypothèse d'une contribution du web au pluralisme de l'information, qui a animé notre propre recherche en France. A la différence des travaux évoqués précédemment, nous ne traiterons toutefois dans cet article que de l'analyse des informations offertes sur le web. Les deux autres volets, connexes, de la production et de la réception de l'information, sont en partie abordés dans un autre article de ce dossier (Rieder, Smyrnaios, 2012). Laissant donc momentanément de côté les ressorts sociaux de l'information, nous centrons notre analyse de discours sur les informations produites en les rapportant à leurs sites énonciateurs, en allant plus loin dans cette voie que ne l'ont fait les recherches passées en revue. Afin de pouvoir pleinement statuer sur le degré de diversité des informations disponibles sur le web, nous avons fait en sorte d'inspecter l'ensemble des sites proposant des contenus d'actualité. Cela nécessitait, par rapport aux recherches existantes, d'élargir très sensiblement la focale afin d'intégrer des sites susceptibles de proposer une information plus originale tels que les blogs. Il s'est agi, en somme, de déplacer le centre de gravité de l'analyse vis-à-vis des sites de médias "traditionnels", encore trop souvent pris comme référents de départ.

Les résultats livrés dans cet article présentent ainsi un caractère novateur. Ils sont issus de l'appréhension exhaustive d'un espace national d'informations d'actualité sur le web, et permettent à ce titre de pleinement embrasser la problématique de l'adéquation entre multiplicité des sites et diversité des nouvelles. Ils constituent l'aboutissement d'une recherche au long cours, base du programme Ipri[2], qui a déjà été précédée par des travaux exploratoires (Marty et al., 2010 ; Smyrnaios et al., 2010). Ces premières incursions dans l'analyse de l'information sur le web nous ont permis d'éprouver la validité d'une typologie des sites d'actualité et d'établir des corrélations entre modes de publication de l'information et nature des contenus offerts. Néanmoins, nous pointions dans nos précédents travaux un certain nombre de limites, dépassées dans les dernières phases de notre recherche. D'une part, le nombre de sites web considérés a été substantiellement augmenté, d'un échantillon d'environ 60 à près de 200, jusqu'à représenter la totalité de la population des sites proposant des informations générales et politiques en France. D'autre part, la période d'observation s'est étalée sur une dizaine de jours consécutifs (du 7 au 17 mars 2011), plutôt que sur quelques journées isolées, et a ainsi permis de prendre en compte tous les rythmes de publication des sites, y compris ceux beaucoup plus lents des blogs. Au final, nous avons ainsi disposé d'un terrain à même de fournir une évaluation globale, pour la France, du niveau de pluralisme de l'information sur le web.

[2] Programme de recherche IPRI - Internet, pluralisme et redondance de l'information (ANR-09-JCJC-0125-01b), soutenu par l'Agence nationale de la recherche et regroupant des laboratoires en information-communication (CIM, université Paris 3 ; ELICO, université de Lyon ;LERASS, université Toulouse 3 ; CRAPE, université Rennes 1 ; GRICIS, UQAM Montréal) et en informatique (LIRIS, INSA Lyon). Ce programme ANR (2009-2012) avait été précédé par la constitution d'une équipe-projet soutenue en 2008 par la MSH Paris-Nord.

SAISIR LE PLURALISME DE L'INFORMATION SUR LE WEB : ELEMENTS DE METHODE

Notre visée dans cette recherche a donc été d'épouser le spectre complet des sites d'actualité, dans toute leur diversité. Ceci afin de voir si lui fait écho une diversité dans les informations mises à l'agenda médiatique, dans les façons de rendre compte des différents événements survenant dans le monde social, à même d'éclairer le jugement des citoyens.

Une analyse de l'agenda médiatique, adaptée à la configuration du web

On sait notamment depuis Berger et Luckmann (1966) que l'activité de construction sociale de la réalité, par le biais du langage et des médias, aboutit à livrer un miroir inévitablement déformant et pluriel de notre environnement (Véron, 1981; Charaudeau, 2005). Les différentes étapes de l'élaboration des nouvelles, la sélection des sujets dignes d'être traités, les critères de la « *newsworthiness* » évoquée par Gamson et Modigliani (1989), sont étroitement dépendantes des logiques économiques et éditoriales des différents médias. Cette dimension de l'activité médiatique, située en amont de la diffusion de l'information, est peu visible au public. Elle n'en a pas moins un rôle essentiel : celui de définir la physionomie de l'agenda médiatique.
Depuis les travaux de McCombs et Shaw (1972) et leur théorisation de l'*agenda-setting*, de nombreuses études ont démontré la tendance des principaux médias à traiter des mêmes sujets d'actualité au même moment (Dearing, Rogers, 1992). Il s'agit de ce que Scheufele (2000) a en réalité identifié comme étant l'activité d'*agenda-building*, littéralement de « construction de l'agenda », le terme *setting* désignant plutôt les corrélations entre les sujets privilégiés par les médias et ceux considérés comme importants par le public. Sans préjuger de la manière dont peut se répercuter l'*agenda-setting* auprès des publics, notre propos ici est bien d'identifier l'état de l'*agenda-building* en ligne, en se donnant pour objet de questionner les spécificités des différents sites d'information. L'analyse de l'agenda médiatique sur le web a donc du être rapportée à l'ensemble des types de sites d'actualité. Plus précisément encore, dans une optique d'étude du pluralisme de l'information web en France, il s'est agi de répertorier tous les sites publiant des informations et/ou commentaires sur l'actualité générale et politique, à caractère essentiellement national.

Une appréhension de l'ensemble des catégories de sites

Le paysage des sites web d'information générale et politique avait fait l'objet d'un premier recensement (Marty et al., 2010) distinguant les catégories de *presse en ligne* (versions internet de médias existants), *webzines* (publications collectives exclusivement internet), blogs (publications individuelles exclusivement internet), *sites participatifs* (publications collaboratives exclusivement internet), *portails* (composantes informationnelles de plateformes multiservices), et *agrégateurs* (regroupements automatisés d'informations d'actualité). Cette première catégorisation s'appuyait sur des typologies relativement anciennes (Deuze, 2003 ; Rebillard, 2006) qui méritent d'être actualisées.
Tout d'abord, la frontière entre agrégateurs et portails est devenue plus floue, dans la mesure où ces derniers emploient eux aussi des procédés d'automatisation de la recherche d'actualités. Ces deux types de site officient de plus selon une logique similaire, qui est celle de l'infomédiation telle qu'évoquée plus haut. Dans la présente étude, *portails* et *agrégateurs* sont donc rassemblés et désignés sous l'appellation commune d'*infomédiaires*. Par ailleurs, en ce qui concerne l'appellation *presse en ligne*, on peut légitimement avancer qu'elle ne traduit plus suffisamment la montée en puissance des versions internet de médias audiovisuels. Il est

alors sans doute préférable d'employer la formule plus large de *médias en ligne*, recouvrant le déploiement sur la toile de la presse écrite mais aussi de la radio et de la télévision. Parallèlement, une nouvelle appellation, celle de *pure players*, s'est progressivement imposée jusqu'à subsumer les catégories *webzines* et *participatifs* (Ouakrat, 2011), lesquelles incluent désormais un degré variable de modération des contenus par la communauté des internautes ou par la rédaction professionnelle du site. Pour autant, les supports de diffusion ont également évolué : plusieurs *pure players* développent ou ont un temps développé une version écrite (*Bakchich*, *Rue89*), rendant inadaptée l'appellation de sites *exclusivement internet*. Par conséquent, l'appellation de sites *nés en ligne* (Mercier, Pignard-Cheynel, 2011), ou plus exactement *natifs de l'internet*, leur sera préférée.
Nous proposons donc à présent une catégorisation mise à jour et simplifiée, distinguant les médias en ligne, les sites natifs de l'internet, les blogs et les infomédiaires. A l'intérieur de cette catégorisation, une attention particulière sera toutefois portée au caractère professionnel ou amateur[3] de l'énonciateur, nous amenant parfois à distinguer certains sites à l'intérieur d'une même catégorie. C'est particulièrement vrai pour la catégorie des blogs, lesquels peuvent être des lieux d'expression individuelle des journalistes professionnels, hébergés par le site de leur titre de presse (ex : *Yvan Riouffol* sur le site du *Figaro*), ou relever d'une initiative individuelle et amateure (blogs *Partageons mon avis*, *A perdre la raison*, *CSP*, etc.). C'est le cas également de plusieurs sites natifs de l'internet, voire de certains infomédiaires, tels que *Rezo.net*, qui compile des publications à la fois amateures et professionnelles.

Le recensement des sites et leur exploration

Une fois cette typologie déterminée, la recherche des sites web d'information générale et politique, à caractère essentiellement national, a visé l'exhaustivité. Trois étapes méthodologiques se sont succédé à cette fin.
Une première liste de sites a d'abord pu être établie sur la base des sites identifiés lors des travaux exploratoires précédents de 2008-2010. Cette base initiale a été contrôlée et ainsi mise à jour en janvier 2011. Elle a abouti à l'intégration des 43 médias en ligne et des 14 infomédiaires de notre échantillon final de sites. Des médias natifs de l'internet et des blogs avaient été également repérés lors de nos précédents travaux : ils ont eux aussi été soumis à un nouvel examen en janvier 2011. Mais pour ces deux dernières catégories, blogs et sites natifs de l'internet, des démarches complémentaires ont été mises en œuvre.
La méthode du navi-crawling[4], consistant en une exploration systématique des liens hypertextuels d'un site, a été employée. Plus précisément, elle a été appliquée à un échantillon-racine constitué par les blogs occupant les premières places du classement Wikio, dont l'activité est attestée et dont la place est relativement centrale dans la blogosphère française. Ainsi, plusieurs sites (majoritairement des blogs, mais également certains sites natifs de l'internet) ont été découverts et intégrés à la suite de ce navi-crawling.

[3] Sachant que l'activité et le statut ne se recouvrent pas toujours dans le domaine journalistique (Ruellan, 2007), nous avons classé comme amateurs les non-professionnels du journalisme au sens d'individus n'étant pas rémunérés à titre principal pour cette activité, fût-elle reconnue ou non par les instances de consécration ou de labellisation statutaire (pouvoirs publics, structures paritaires comme la Commission de la carte d'identité professionnelle des journalistes français, etc.). Par conséquent, à l'intérieur de cette catégorie transversale des amateurs, on pourra trouver à la fois des passionnés exerçant bénévolement une activité de production d'information d'actualité et des individus éventuellement rémunérés pour cette activité sans que cela ne constitue pour autant leur activité principale (c'est le cas notamment de travailleurs intellectuels comme les juristes ou les chercheurs).
[4] A l'aide de l'extension Firefox développée par Web Atlas: http://webatlas.fr/wp/navicrawler/

Enfin, d'autres blogs et sites natifs de l'internet ont été ajoutés à partir de liens pointés par les internautes sur Twitter. Ce travail s'est appuyé sur la Twitter REST API et son exploitation depuis la plateforme *Tweetism* développée par Bernhard Rieder et Raphaël Velt.
Ces multiples méthodes ont permis au final d'identifier 110 blogs et 42 sites natifs de l'internet répondant aux critères précédemment évoqués. Ceci a porté le nombre total de sites recensés à 209 en février 2011, un ensemble que l'on peut estimer très proche de la population entière des sites français dédiés à l'information générale et politique au moment de notre période d'observation (7 au 17 mars 2011).

La collecte des articles

Pour procéder à l'analyse des informations publiées par ces sites, un logiciel a été spécialement développé dans le cadre de notre projet par les chercheurs du laboratoire d'informatique LIRIS[5]. Ce logiciel, appelé IPRI - News Analyzer (IPRI-NA), a dans un premier temps permis de collecter les articles publiés sur les flux RSS des sites. Dans l'idéal, il aurait été préférable de collecter les articles publiés directement sur les pages web, mais les procédés techniques existants étaient peu satisfaisants, charriant avec les articles et illustrations des encarts publicitaires ou des liens hypertextuels susceptibles de complètement brouiller l'analyse. Les flux RSS fournissent de ce point de vue un rendu bien plus adéquat, avec titre de l'article, descriptif, lien vers l'URL d'origine, et date de publication. Pour chacun des sites a été capté le flux RSS *A la Une* ou *Actualités* rassemblant les articles correspondant à notre critère d'information générale et politique, à caractère essentiellement national. Il s'est avéré que ces flux RSS n'étaient pas toujours homogènes. Par exemple, l'un d'entre eux (celui de *Free Actualités*) s'est distingué par nombre d'informations à caractère régional. Plus globalement, les flux *A la Une* et *Actualités* intégraient pour la plupart des informations sur les grands évènements sportifs et l'activité boursière des principales places financières, alors que ces dernières informations étaient parfois réservées à des flux thématiques (flux *Sport*, *Economie*) dans certains sites. Ces phénomènes, même minoritaires, peuvent être vus comme une source de biais pour la présente recherche, considérant qu'une dose de diversité est introduite dès l'amont, *via* le mode de collecte des données. Cependant, on peut aussi considérer, par analogie avec l'analyse de la structuration sémiotique de la presse écrite (Mouillaud, Tétu, 1989), que la partition en flux RSS des sites est assimilable à la partition en pages et rubriques des journaux, et traduit en cela un découpage représentationnel de la réalité sociale par chacun des sites. Le fait de placer tel ou tel article dans les flux *A la Une* ou *Actualités* relève donc des choix éditoriaux de sélection et de hiérarchisation de l'information, dont notre analyse de discours entend rendre compte -sans forcément se prononcer sur la pertinence ou la cohérence des choix effectués par les éditeurs- pour précisément voir quelles nouvelles sont mises à l'agenda.
La collecte s'est déroulée sur la quasi-totalité du mois de mars 2011. Comme indiqué précédemment, l'analyse en elle-même a porté sur une période d'une dizaine de jours consécutifs afin d'intégrer les temporalités de publication respectives des différentes catégories de sites. De façon aléatoire, la période du 7 au 17 mars a été délimitée, planifiée plusieurs semaines avant la collecte afin de s'assurer de son bon déroulement pratique (nécessité pour les chercheurs de prévoir une disponibilité pour surveiller le bon fonctionnement du logiciel durant cette période). Cette période s'est révélée comporter le tsunami au Japon en son sein (vendredi 11 mars). Sa physionomie a alors donné la possibilité de travailler d'une part sur des journées « ordinaires » au niveau de l'actualité, en amont du

[5] Le logiciel IPRI News Analyzer (IPRI-NA) a été développé par Samuel Gesche, Elöd Egyed-Zsigmond et Cyril Laitang. Il est distribué sous licence Creative Commons http://liris.cnrs.fr/ipri/pmwiki/index.php?n=Public.IpriNA

tsunami, d'autre part sur les journées plus exceptionnelles qui ont suivi. Après collecte et évacuation de quelques artefacts techniques, le corpus comprend 37 569 articles sur la période du 7 au 17 mars 2011, publiés par 199 sites[6] répartis entre les quatre catégories : médias en ligne, sites natifs de l'internet, blogs, et infomédiaires.

La classification des articles en sujets d'actualité

Les milliers d'articles ainsi collectés ont été soumis à une analyse de discours, afin de déterminer les sujets d'actualité abordés par les différents sites web, et de reconstituer à partir de là l'agenda médiatique. Dans cette recherche, orientée par une problématique de mise à l'agenda, nous considérons en effet l'agenda médiatique comme une sorte de mosaïque composée des différents sujets d'actualité abordés par l'ensemble des sites. Les sujets d'actualité peuvent constituer des évènements médiatiques de plus ou moins grande importance, en fonction du volume d'articles qui leur sont consacrés ou du nombre de sites qui leur prêtent attention. La notion de sujet d'actualité est donc la pierre angulaire de ce travail, à la fois opérateur de la classification sémiotique des articles, et unité de base de l'agenda médiatique pour les analyses ultérieures. Elle nécessite à ce titre d'être définie de façon étayée.

La notion de *sujet d'actualité*

La notion de sujet d'actualité a été construite sur la base d'une distinction entre le cadrage primaire et le cadrage secondaire, au sein du processus de *constitution* médiatique de la réalité (Arquembourg, 2011). Le cadrage primaire se rapporte au concept de « fait », lequel est issu de l'activité de perception d'une expérience par les sens (Goffman, 1991) et va faire l'objet d'une sélection journalistique parmi les occurrences du réel (Neveu, Quéré, 1996), avant d'être soumis à un traitement différencié par chacun des médias. Le cadrage secondaire, en revanche, correspond à ce traitement médiatique d'un même fait (Esquenazi, 2002 ; Marty et al., 2010), auquel les journalistes apposent des angles, des lignes éditoriales, des points de vue (Ringoot & Rochard, 2005).
Un sujet d'actualité est ici entendu comme un fait, une expérience passée au prisme d'un cadrage médiatique primaire, en amont du cadrage secondaire choisi pour le traiter. Les articles s'étant rejoints dans le cadrage médiatique primaire d'un même fait ont ainsi été considérés comme relevant d'un même sujet d'actualité, quel que soit le cadrage médiatique secondaire adopté par chacun d'eux.
Par exemple, tous les articles abordant l'ouverture du procès de Jacques Chirac concernant les emplois fictifs à la Ville de Paris ont été regroupés, considérés comme relevant d'un même sujet d'actualité, et différant par exemple d'articles sur les soupçons d'espionnage au sein de l'entreprise Renault ou plus encore d'articles sur le conflit en Libye. Ceci en sachant que, au sein de ce sujet d'actualité *Procès Chirac - emplois fictifs*, certains articles présenteront un point de vue sévère vis-à-vis de l'ancien Président de la République, tandis que d'autres se montreront bienveillants eu égard à l'état de santé de l'accusé ou à l'antériorité des faits jugés.
Le passage par la notion de sujet d'actualité permet ainsi de reconstituer l'agenda médiatique, d'évaluer l'importance relative de chaque sujet dans l'actualité du web. Il peut également être un préalable à des analyses plus approfondies s'intéressant cette fois aux traitements différenciés d'un même sujet d'actualité par chacun des médias, objectif d'un autre article de ce dossier (Touboul et al., 2012).

Partant de cette définition du sujet d'actualité, la classification des articles a alors été établie sur chacune des journées, selon une méthode semi-automatisée grâce au logiciel IPRI-NA. Par rapport à d'autres recherches portant sur la diversité de l'information en ligne, principalement menées aux Etats-Unis, notre méthode se situe à mi-chemin entre un

[6] Sur les 209 sites initialement identifiés, seuls huit blogs et deux sites natifs de l'internet n'ont pas vu d'article publié et collecté par le logiciel IPRI-NA pendant la période en question. La liste des 199 sites ayant publié est fournie en annexes.

traitement lexicométrique complètement automatisé (cf. Leskovec et al., 2009, à propos de la médiatisation de la campagne présidentielle US en 2008 entre blogs et sites web), et ce qui resterait au stade d'une analyse manuelle par codage thématique (cf. Carpenter, 2010, comparant les articles des sites de médias traditionnels et ceux publiés dans les sites de médias citoyens). Nous avons pour notre part procédé à un premier défrichage informatique des similarités lexicales entre articles, permettant de traiter automatiquement un tiers de notre corpus, avant de vérifier cette première classification et de la compléter par un codage manuel, en sujets d'actualité, des deux-tiers restants du corpus[8].

Une telle méthode, même fondée sur des critères théoriques relativement solides, pose la question de l'arbitraire du codeur au moment de la classification des articles en sujets. Il a donc été décidé de procéder à une mesure statistique de la fiabilité intra-codeur et inter-codeurs. Le test intra-codeur consistait en une nouvelle classification par le même codeur plusieurs mois après (juillet 2011 / novembre 2011), tandis que le test inter-codeurs établissait une comparaison avec un second codeur. Ces tests ont abouti à un taux de cohérence intra-codeur de 96,5% et de plus de 92 % pour la cohérence inter-codeurs, taux très élevés accréditant la fiabilité de la classification[9].

[8] A titre informatif : le codage manuel d'une journée moyenne de la période d'observation (autour de 3 500 articles à classer en sujets d'actualité) demande l'équivalent de cinq jours de travail à temps plein. Ce travail, répété pour les onze journées, a été principalement réalisé par Emmanuel Marty lors de son post-doctorat au sein du programme Ipri.

[9] Les tests ont été réalisés sur un échantillon constitué avec une méthode aléatoire simple, sur un échantillon de 350 articles soit environ 10% de la production éditoriale d'une journée, celle-ci ayant été également été tirée au hasard puisque chaque journée présente globalement les mêmes caractéristiques. Les taux obtenus sont élevés car les probabilités, au vu de taille de l'échantillon, donnent une erreur inférieure à 3%.

Ces tests ont permis d'évaluer la concordance dans l'indexation des articles en catégories (sujets) prédéfinies. Pour avoir un test complet, il aurait également fallu évaluer la pertinence de chacune des catégories, au moment même de leur élaboration (création d'un sujet). Toutefois, une garantie est apportée par l'expertise du chercheur impliqué dans la définition des sujets en 2011 suite à sa participation aux premières phases des travaux en 2008 et 2010 et à leur caractère collectif : la définition des sujets a été faite pour les corpus concernés en collaboration avec un autre chercheur du programme Ipri ainsi qu'avec la responsable du traitement documentaire TV à l'Ina (nos remerciements à Dominique Fackler pour sa contribution à cette étape du projet).

LA PHYSIONOMIE DE L'AGENDA MEDIATIQUE SUR LE WEB

La classification des articles en sujets d'actualité nous permet d'évaluer le pluralisme de l'information en ligne de façon quantitative. En nous inspirant de la typologie élaborée par Benhamou et Peltier pour l'évaluation de la diversité culturelle (2006), nous retenons trois critères similaires pour le pluralisme de l'information. Le critère de variété, premier critère, sera ici constitué par le nombre de sujets d'actualité abordés. Le second critère, celui de l'équilibre, concernera dans notre cas la répartition des articles au sein des différents sujets d'actualité, indiquant le poids relatif de ces derniers entre sujets majeurs occupant la Une de l'actualité sur le web et sujets mineurs relégués dans les tréfonds de l'internet. Un troisième critère, celui de la disparité, vise à identifier les différents modes de traitement pour un même sujet. Nécessitant une analyse de nature plus qualitative, ce dernier critère de disparité de traitement journalistique fait l'objet d'un autre article dans ce dossier (Touboul et al., 2012).
Le présent travail est donc centré sur la variété et l'équilibre des sujets d'actualité. Cette mesure quantitative du pluralisme de l'information aboutit à dessiner l'agenda médiatique tel qu'il s'est présenté sur le web, durant notre période d'observation, entre le 7 et le 17 mars 2011.

La variété des sujets abordés au jour le jour

Rappelons que, au cours de cette période, 199 sites ont publié un total de 37 569 articles. Sur chaque journée, ce sont entre 300 et 700 sujets d'actualité différents qui sont abordés sur le web, témoignant d'un niveau élevé de variété éditoriale. Dans le détail, le décompte s'effectue comme suit (voir tableau 1).

Journée	Nb articles	Nb sujets	Nb sites
Lundi 7 mars	3418	593	143
Mardi 8 mars	3496	625	145
Mercredi 9 mars	3552	670	141
Jeudi 10 mars	3369	678	140
Vendredi 11 mars	**4068**	**573**	145
Samedi 12 mars	2415	342	117
Dimanche 13 mars	2417	306	116
Lundi 14 mars	3527	**534**	**152**
Mardi 15 mars	3757	562	141
Mercredi 16 mars	3723	556	144
Jeudi 17 mars	3827	613	141

Tableau 1 - Variété des sujets abordés entre le 7 et le 17 mars 2011

Le nombre d'articles est relativement stable pour chacune des journées, compris entre 3300 et 3900, hormis lors de trois journées particulières. Deux journées, d'abord, se caractérisent par un nombre nettement plus réduit d'articles (environ 2 400). Il s'agit des deux journées de week-end, lors desquelles les rédactions en ligne sont beaucoup moins fournies en personnel. Une autre journée s'avère ensuite exceptionnelle durant notre période d'observation : le vendredi 11 mars 2011, jour du tsunami au Japon, où 4068 articles ont été produits. Sur le plan des sujets, cette journée constitue parallèlement un moment où la variété éditoriale baisse très nettement, perdant une centaine de sujets d'actualité par rapport à la veille. Cette décrue se poursuit le lundi, où la variété atteint son plus bas niveau -excepté le week-end- avec 534

sujets d'actualité. Ce nombre plancher de sujets correspond, à l'inverse, à un nombre plafond de sites : 152 d'entre eux ont publié des articles ce jour-là. L'évènement exceptionnel survenu au Japon a donc suscité la production chez nombre de sites et focalisé leur attention sur ce sujet précis, ainsi que sur les sujets afférents (débat sur le nucléaire après l'accident des centrales suite au tsunami, voir tableau 2) dans les jours suivants.

Rang	Sujet d'actualité	Nb articles
1	Accidents de centrales au Japon et risque nucléaire	4147
2	Conflit en Libye	4131
3	Violent séisme au Japon et tsunami	3101
4	Le débat sur le nucléaire relancé en France et dans le monde	1031
5	Un sondage Harris donne Marine Le Pen en tête des intentions de vote	962
6	Procès Chirac emplois fictifs	858
7	Affaire Renault	757
8	Conflit en Côte d'Ivoire	683
9	La campagne des élections cantonales	571
10	Manifestations et répression au Bahreïn	475

Tableau 2 - Sujets d'actualité majeurs entre le 7 et le 17 mars 2011 (top 10)

Le vendredi 11 mars (et le week-end qui l'accompagne) a donc constitué une césure. Avant, c'est-à-dire entre le lundi 7 et le jeudi 10 mars, s'étale une phase d'actualité ordinaire, avec une moyenne quotidienne de 3459 articles pour 641 sujets. La phase qui suit, entre le lundi 14 et le jeudi 17 mars, affiche elle des moyennes de 3708 articles pour 566 sujets. Avec un nombre plus élevé d'articles pour moins de sujets, la deuxième phase se caractérise donc par une plus grande concentration de l'information en ligne. Cette modalité particulière de distribution de l'information renvoie au deuxième critère de pluralisme, l'équilibre, sur lequel nous allons nous pencher de façon plus précise.

Un agenda médiatique en déséquilibre quotidien

L'agenda médiatique, sur le web, est constitué d'une multitude de sujets d'actualité comme nous venons de le voir. Reste maintenant à déterminer comment cet agenda médiatique est structuré, en analysant la place accordée à chacun des sujets. Cette hiérarchie de l'information sera évaluée ici en considérant l'importance d'un sujet d'actualité à l'aune du nombre d'articles qu'il rassemble.

De telles mesures font apparaître les résultats suivants : un nombre réduit de sujets concentre la majorité des articles et, inversement, de très nombreux sujets sont abordés de façon isolée dans un seul article ou une poignée d'articles. Cette dualité entre forte concentration d'un côté et extrême dispersion de l'autre, sur laquelle nous nous sommes déjà arrêtés pour notamment montrer son caractère relativement classique au regard de l'hypothèse de « longue traîne » (Marty et al., 2010; Smyrnaios et al., 2010), correspond de façon plus générale à des distributions parétiennes de type 20/80[11]. Concernant le pluralisme de l'information sur le web, cela signifierait que 20% des sujets rassemblent à eux seuls environ 80% des articles. En mars 2011, les journées du 7 au 10 mars présentent des valeurs semblables : 20% des sujets rassemblent entre 80 et 83% des articles. En revanche, lors des

[11] Par exemple, avec un coefficient d'ajustement de 0,91, les mathématiques (Egghe, 2005) permettent de dire : 10% des sujets totalisent 81% des articles, 20% des sujets 86,4% des articles, et enfin 50% des sujets représentent 93,9% des articles.

journées du 11 au 17 mars, 20% des sujets rassemblent entre 85 et 88 % des articles (voir tableau 3).

Date	% d'articles rassemblés par 10% des sujets	% d'articles rassemblés par 20% des sujets	% d'articles rassemblés par 50% des sujets
Lundi 7 mars	75	83	91
Mardi 8 mars	73	81	91
Mercredi 9 mars	71	80	91
Jeudi 10 mars	71	80	90
Vendredi 11 mars	**80**	**86**	**93**
Samedi 12 mars	**81**	**87**	**93**
Dimanche 13 mars	**82**	**88**	**94**
Lundi 14 mars	**79**	**85**	**92**
Mardi 15 mars	**79**	**86**	**93**
Mercredi 16 mars	**80**	**86**	**93**
Jeudi 17 mars	**80**	**85**	**92**

Tableau 3 - Concentration des articles dans les principaux sujets entre le 7 et le 17 mars 2011

Nous retrouvons ici la césure identifiée précédemment : un surcroît de concentration de l'information intervient à partir du vendredi 11 mars, jour du tsunami au Japon. On peut donc considérer que cet évènement est à l'origine d'une telle surconcentration de l'information.
Cette hypothèse est étayée par l'observation du premier décile des sujets lors de chaque journée : 10% des sujets rassemblent entre 71 et 75% des articles du 7 au 10 mars, puis entre 79 et 82% des articles du 11 au 17 mars. Ceci laisse penser que la différence entre les journées se joue bien au niveau du premier décile de sujets d'actualité, comprenant ceux relatifs au tsunami, aux accidents dans les centrales et au débat sur le nucléaire, à compter du 11 mars et par la suite.
Par ailleurs, si l'on s'intéresse cette fois à la médiane des sujets, on remarque que 50% des sujets ont rassemblé entre 90 et 91% des articles du 7 au 10 mars, puis entre 92 et 94% du 11 au 17 mars. Donc à nouveau une césure autour du 11 mars, mais moins prononcée que pour les 10% et 20 % de sujets principaux. En miroir, cela veut aussi dire que l'autre moitié des sujets n'est vraiment abordée qu'avec parcimonie, toujours confinée dans moins d'un dixième des articles. Ceci confirme le déséquilibre quotidien de l'agenda médiatique en ligne, entre surexposition d'une minorité de sujets d'actualité et confidentialité de la majorité d'entre eux.

LA CONTRIBUTION DES DIFFERENTS SITES AU PLURALISME

L'agenda médiatique se caractérise donc, sur le web, par une sorte de grand écart quotidien. Il est constamment étiré entre la focalisation sur une minorité de sujets ultra-médiatisés, et l'ouverture à une myriade d'autres sujets beaucoup plus originaux. Ce tableau général, que nous venons de dessiner, comprend toutefois plusieurs nuances de couleurs. Elles correspondent à la contribution de chacun des sites observés. Ceux-ci pèsent d'un poids respectif fort différent sur le degré de pluralisme général mesuré précédemment. Nous allons à présent nous pencher sur ces différents sites de façon plus individualisée, faisant succéder à l'analyse du pluralisme externe ou *inter media* un regard sur le niveau de pluralisme interne ou *intra medium* (Mc Quail, 1992) propre à chaque site.

Variété et éclectisme éditorial des sites

Notre premier critère d'évaluation quantitative du pluralisme, la variété, tient au nombre de sujets abordés. Appliqué de façon individuelle à chacun des sites, il fournit des résultats intéressants et en même temps perfectibles.

Le décompte brut permet de repérer quels sites ont abordé le plus grand nombre de sujets au cours de la période d'observation (entre le 7 et le 17 mars 2011). Dans les cinq premières places, on retrouve alors quatre infomédiaires (*Free*[12], *Wikio*, *Voila*, *Newspeg*) avec près de 500 sujets, voire plus, abordés au cours de cette période d'une dizaine de jours. Nous remarquons aussi dans le haut de ce classement, aux côtés de versions numériques de plusieurs médias appartenant à des groupes industriels de communication (*Le Journal du Dimanche* et *Europe 1* du groupe Lagardère, *L'Express* et *L'Expansion* du groupe Roularta), plusieurs sites natifs de l'internet présentant une forme participative d'ouverture aux contributions d'amateurs (*François Desouche*, *Bellaciao*, *Agora Vox*)[13].
Inversement, lorsque l'on s'intéresse cette fois aux sites ayant abordé le moins de sujets au cours de la période, c'est-à-dire à la queue d'un tel classement, apparaissent 101 sites à avoir abordé 11 sujets ou moins au cours de la période de 11 jours, c'est-à-dire au mieux un sujet par jour. Ces derniers sont très majoritairement des blogs (précisément 78 blogs sur 101 sites). Une telle différence dans le nombre de sujets abordés, selon les catégories de sites, s'explique en grande partie par leur volume de production d'articles, lui-même très inégal. Les infomédiaires se caractérisent par une forte productivité, continue tout au long de la journée (75 articles diffusés par jour en moyenne), s'alimentant le plus souvent automatiquement auprès d'éditeurs tiers. Les blogs fonctionnent avec un rythme de publication beaucoup plus étalé et moins régulier, que notre période d'observation allongée à onze jours permet tout juste d'embrasser pour certains d'entre eux.

Ce constat amène à considérer la valeur brute du nombre de sujets abordés comme un indicateur nécessaire mais non suffisant de pluralisme interne. Une solution consiste à le pondérer par le nombre d'articles produit par chacun des sites. On arrive ainsi à une valeur relative apte à refléter plus équitablement le niveau de variété d'un site, son niveau d'éclectisme éditorial en quelque sorte.

---

[12] Le nombre de sujets obtenus pour Free est en partie biaisé. Cet infomédiaire agrège dans son flux RSS d'actualités générales et politiques des informations d'intérêt à la fois national et local. Cette dernière particularité, consistant à ajouter des nouvelles locales, explique en partie pourquoi il aborde un nombre de sujets bien supérieur aux autres sites dont les fils d'actualités sont exclusivement nationaux.

[13] Le Post occupait une place ambivalente à cet égard, site natif de l'internet à vocation participative tout en étant intégré à la filiale d'un groupe industriel de communication (Le Monde Interactif).

En procédant à un calcul de ce type, on aboutit à un nouveau classement qui place dans sa première partie tous les sites ayant produit autant d'articles qu'ils n'auront abordé de sujets. Au total, 63 sites auront respecté cette « ligne de conduite éditoriale » consistant à consacrer un seul article à chaque sujet abordé. Il s'agit pour l'essentiel de blogs et, dans une moindre mesure, de sites natifs de l'internet. Parmi eux, le Bondy Blog se distingue en particulier par le fait d'avoir produit un nombre assez consistant d'articles sur la période (24 articles en 11 jours), chaque fois sur un sujet différent (24 sujets différents au total donc). L'éclectisme éditorial de ce site au niveau des contenus, tel que mesuré ici, pourrait être mis en relation avec son mode de réalisation singulier, associant professionnels et amateurs, ancré dans des réalités sociales, celles des « banlieues », bien différentes de celles caractérisant d'autres expériences de journalisme participatif (Sedel, 2011).
A l'inverse, en queue de ce nouveau classement vont se retrouver des sites qui eux s'illustrent par un certain « matraquage » éditorial, n'hésitant pas à multiplier les articles autour d'un même sujet. Parmi eux se trouvent assez logiquement des infomédiaires et plus précisément des agrégateurs de nouvelles, comme *Google Actualités* (14 articles par sujet en moyenne) ou *actu2424* (7 articles par sujet). Les rejoignent des sites de médias professionnels connus pour être très actifs dans la production d'articles afin de, précisément, favoriser leur référencement auprès des agrégateurs et moteurs de recherche (Smyrnaios, Rebillard, 2009). *RTL* (13 articles par sujet) et *Le Nouvel Observateur* (10 articles par sujet) sont par exemple les seuls à passer la barre des 10 articles par sujet, à la suite de Google justement.
Le critère de variété, ainsi pondéré par le nombre d'articles produits, constitue un indicateur de l'éclectisme éditorial caractérisant la production de chaque site. Permettant d'évaluer la propension d'un site à aborder plusieurs sujets d'actualité, il ne renseigne toutefois pas sur la nature de ces derniers. Leur répartition entre sujets utlra-médiatisés et sujets rares constitue un critère complémentaire, critère d'équilibre interne, à l'évaluation du pluralisme d'un site d'actualité.

Originalité éditoriale des sites

Le critère d'équilibre, dans le cadre d'une appréhension *externe* de l'agenda médiatique général, nous amenait à identifier la distribution des articles entre les différents sujets, certains sujets s'avérant dominants tandis que d'autres se montraient minoritaires. Dans le cadre du pluralisme *interne*, il convient désormais de déterminer si chacun des sites s'attache à des sujets très prisés, ou au contraire peu relayés par d'autres. En somme, il s'agit de mesurer l'originalité d'un site en matière de choix éditoriaux, ici au prisme du degré de rareté des sujets qu'il aborde[14].
Le degré de rareté éditoriale d'un sujet sera ainsi évalué à l'aune de sa couverture médiatique, par l'ensemble des sites produisant des informations d'actualité sur le web. Moins un sujet aura été abordé par des sites au cours de la période d'observation, et plus il sera considéré comme rare dans les choix éditoriaux. Nous avons considéré ici qu'un sujet s'avérait rare sur le plan éditorial s'il était abordé par moins de cinq sites au cours d'une journée. A partir de la mise en place de ce seuil, les articles portant sur des sujets rares ont pu être identifiés. Et leur proportion au sein de l'ensemble des articles produits par un site a dès lors pu être calculée. Le pourcentage ainsi obtenu traduit l'originalité éditoriale d'un site, son degré d'attention pour des sujets rares.

[14] Nous remercions notre collègue Bernhard Rieder pour l'aide fournie sur ce point. Les calculs et graphes présentés dans la suite de cet article ont été réalisés par ses soins. Par rapport aux évaluations précédentes sur le niveau de pluralisme global, l'importance d'un sujet n'est désormais plus seulement relative au nombre d'articles qu'il rassemble mais aussi au nombre de sites qui l'abordent.

Une telle évaluation du pluralisme interne des différents sites amène des résultats particuliers pour les blogs. En raison du faible volume d'articles produits par certains d'entre eux (un, deux ou trois articles produits durant la période d'observation, dans bien des cas), les pourcentages calculés n'ont pas forcément grande signification statistique. Par exemple, le blog d'*Abadinte* a produit un article au cours de la période, correspondant à un sujet rare, et se retrouve donc avec un taux poussé au maximum de 100%. Quant au blog *Le Monolecte*, il a lui aussi publié un seul billet, mais sur un sujet abordé par plus de 5 autres sites, et se trouve donc affublé d'un score de 0% de sujets rares, loin de refléter son niveau d'originalité. Afin d'éviter ces écueils statistiques, nous limiterons notre évaluation du pluralisme interne aux sites qui ont produit au moins un article par jour en moyenne, soit 11 articles sur la période. Sur les 199 sites ayant produit des articles au cours de la période, 115 sites atteignent ou dépassent ce seuil minimal. Parmi eux, ils sont une vingtaine à consacrer la moitié de leur production d'articles à des sujets rares (voir tableau 4).

Site	Catégorie de site d'actualité	Nb articles total	Nb articles de sujets rares	% articles de sujets rares
Michel Abhervé	Blog	26	23	88%
France Matin	Site natif de l'internet	43	38	88%
Fluctuat.net	Site natif de l'internet	133	116	87%
Politique.net	Site natif de l'internet	11	9	82%
Bondy blog	Site natif de l'internet	24	19	79%
Minutebuzz	Site natif de l'internet	393	309	79%
Enquête et débat	Site natif de l'internet	90	65	72%
PaperBlog	Infomédiaire	27	19	70%
Le salon beige	Blog	268	185	69%
IndyMedia Paris	Site natif de l'internet	135	92	68%
François Desouche	Site natif de l'internet	555	334	60%
Bivouac-ID	Blog	35	21	60%
Pèlerin Magazine	Média en ligne	15	9	60%
Les mots ont un sens	Blog	81	48	59%
Philippe Méoule	Blog	12	7	58%
Cafebabel	Site natif de l'internet	44	25	57%
Rezo.net	Infomédiaire	84	47	56%
Tian	Blog	80	45	56%
Torapamavoa Nicolas	Blog	13	7	54%
L'Humanité	Média en ligne	148	79	53%
Rue 89	Site natif de	98	51	52%

	l'internet			
Le Grand Soir	Site natif de l'internet	53	27	51%
Contrepoints	Site natif de l'internet	115	58	50%
HNS-Info	Site natif de l'internet	78	38	49%
Agora Vox	Site natif de l'internet	425	205	48%

Tableau 4 - Sites présentant la plus grande proportion de sujets rares parmi leurs articles (seuil de 11 articles produits entre le 7 et le 17 mars 2011)

On retrouve en tête de ce classement des sites déjà repérés lors de notre évaluation de l'éclectisme éditorial et qui cumulent donc cette propriété avec celle d'une originalité dans le choix de leurs sujets (*Bondyblog* et *Michel Abhervé*, blog d'un spécialiste de l'économie solidaire hébergé par le mensuel *Alternatives Économiques*).
De façon plus générale, on retrouve là aussi une majorité de blogs et de sites natifs de l'internet, à la différence près que plusieurs de ces sites relaient des informations à partir de positions idéologiques très radicales voire extrêmes, à droite (*Enquête et Débat*, *Le Salon beige*, *François Desouche*, *Bivouac-ID*) comme à gauche (*Indymedia Paris*, *Le Grand Soir*, *HNS-Infos*). Ainsi, notre indicateur de rareté éditoriale des sujets peut avoir tendance à faire ressortir des sites envisageant l'information comme nécessairement engagée. Pour preuve, les deux seuls médias en ligne présents dans ce haut du classement sont les déclinaisons numériques de journaux d'opinion : *Pèlerin Magazine* et *L'Humanité*. Et sur les deux infomédiaires, l'un est spécialisé dans l'agrégation de billets de blogs (*PaperBlog*) dont on vient de voir le caractère souvent orienté politiquement, et l'autre assure sciemment une veille sur des sources se situant sur la gauche de l'échiquier politique (*Rezo.net*).
Aux côtés de ces sites politiquement très marqués figurent des initiatives dont le projet éditorial revenait explicitement à proposer une alternative aux médias existants. Ces sites présentent le point commun d'avoir assis leur mode de réalisation sur une ouverture aux amateurs, sur un mode participatif : *Bondyblog* donc, mais aussi *Cafebabel* et *Rue 89* ou encore *AgoraVox*. Cette fois-ci, l'indicateur de rareté permet d'identifier des sites réalisant, davantage qu'un filtrage idéologique singulier dans leur appréhension de l'information, une opération de décentrage éditorial par rapport à l'agenda médiatique général.

Catégorie de site d'actualité	Nb articles total	Nb articles de sujets rares	% articles de sujets rares
Sites natifs de l'internet	5 149	2 244	44%
Blogs	1 431	623	44%
Infomédiaires	11 557	2 322	20%
Médias en ligne	19 432	3 265	17%
Total - ensemble des sites	*37 569*	*8 454*	*23%*

Tableau 5 - Proportion de sujets rares parmi les articles, selon les différentes catégories de sites d'actualité (entre le 7 et le 17 mars 2011)

Passant au niveau macro des différentes catégories de sites, les calculs présentés dans le tableau 5 confirment les observations précédentes issues d'un niveau plus micro. Les sites

natifs de l'internet et les blogs, tout en produisant relativement peu d'articles, privilégient des sujets rares.
A l'inverse, les infomédiaires et les médias en ligne multiplient les articles, mais pour aborder des sujets déjà largement traités par d'autres. Certains de ces sites paraissent même complètement immergés dans cet agenda médiatique général. Pour plusieurs d'entre eux, la proportion de sujets rares tombe bien en deçà des 10%, autant dire que la quasi-totalité de leurs articles s'inscrit dans le *mainstream* de l'information en ligne. Sans trop de surprise, on retrouve des sites qui s'étaient déjà illustrés par leur redondance éditoriale, occupant une place de *hub* (Weber, Monge, 2011) dans la circulation de l'information : les infomédiaires *Google Actualités* (seulement 3% des articles portant sur des sujets rares) et *actu2424* (également 3%) rejoints par *MSN Actualités* (5%) et *Orange Actualités* (5%), ainsi que les médias en ligne *Le Nouvel Observateur* (6%) et dans une moindre mesure *RTL* (10%). Parmi les autres médias en ligne, on notera la présence spécifique de déclinaisons numériques de chaînes de télévision : *France 3* et *France 2*[15] (4%), *TV5 Monde* (5%), *i-télé* (8%).
Entre les deux extrémités présentées précédemment, c'est-à-dire entre les sites moteurs du *mainstream* médiatique évoqués à l'instant et les sites éditorialement très dissonants présentés au préalable (tableau 4), se situent l'essentiel des sites d'actualité (80 sites). Leur publication d'articles s'inscrit majoritairement dans l'agenda médiatique général du web tout en se réservant, dans des proportions variables, des incursions vers des sujets moins courants.

[15] Les deux chaînes partagent un même fil d'informations sur l'internet.

UNE REPRESENTATION CARTOGRAPHIQUE DE L'INFORMATION SUR LE WEB

Les résultats livrés jusqu'à présent nous ont fourni deux grands types d'enseignements. D'une part, nous savons désormais que l'information offerte sur le web est très variée et en même temps très concentrée, autour de quelques sujets d'actualité qui relèguent des centaines d'autres dans les confins de l'internet. D'autre part, nous avons également appris que les sites présentent des degrés très différents d'originalité éditoriale : certains privilégient des sujets rares tandis que d'autres se rejoignent dans une même redondance informationnelle, une majorité de sites se situant entre ces deux positionnements éditoriaux.
Dans la dernière partie de cet article, nous proposons d'établir une synthèse qui croise tous ces éléments, relatifs au pluralisme autant externe qu'interne. A cette fin, nous avons réalisé une représentation cartographique qui à la fois rende compte de la physionomie particulière de l'agenda médiatique sur le web et situe le rôle de chaque site, dans leurs relations au sein de cet ensemble complexe. A bien des égards, la perspective ainsi adoptée relève d'une analyse du champ de l'information d'actualité sur le web, si l'on considère que l'application du concept de *champ* à l'espace journalistique permet « *tout à la fois de montrer ce qui fait l'unité et la diversité de cet espace et, surtout, de l'étudier en termes relationnels* » (Marchetti, 2002, p. 23).

Une représentation axée sur les similarités éditoriales

La cartographie élaborée repose sur une base double, à laquelle correspond un même double objectif. Premièrement, elle intègre la production d'information dans sa totalité, autrement dit les plus de 37 000 articles produits durant la période, et se doit de représenter leur répartition -déséquilibrée- entre les différents sujets d'actualité. Deuxièmement, elle s'appuie sur les près de 200 sites ayant publié au cours de la période, et doit indiquer leur propension -variable- à s'attacher à des sujets rares ou au contraire quasi-incontournables. Pour joindre ces deux visées, un algorithme a été employé pour calculer la similarité des sites par le biais de la similarité de leur production éditoriale[16]. Il aboutit à rapprocher les sites qui partagent des sujets d'actualité communs, en procédant à des regroupements différenciés selon que ces sujets communs sont des sujets plutôt dominants ou plutôt rares au sein du corpus total d'articles.
La représentation cartographique qui en résulte, de type graphe de réseau[17] et conçue à l'aide du logiciel libre Gephi, place ainsi en son centre les sites qui partagent les sujets les plus

[16] Plus précisément, l'algorithme employé repose sur une adaptation du modèle *vector space* (Salton et al., 1975), issu des sciences de l'information, consistant à calculer des "distances" entre tous les objets (sites), en fonction de leurs caractéristiques (articles), dans un espace à *n* dimensions (sujets). Un script a ainsi été spécifiquement développé pour calculer la "position" de chaque site, en fonction de sa production d'articles sur tel ou tel sujet d'actualité, prenant en compte la fréquence des sujets (inspiration de la logique *Term Frequency - Inverse Document Frequency*). A partir de la représentation des sites en forme de vecteurs, les distances entre les sites ont été calculées sur la base d'une similarité cosinus : lorsque cette dernière atteignait le seuil significatif de 0,005, des liens entre sites ont été affichés sur le graphe, dont la "spatialisation" a été faite à l'aide du logiciel de visualisation de réseaux Gephi et l'algorithme Force Atlas 2.

[17] Le plus souvent, dans le domaine au sens large des *Internet studies*, le graphe de réseau est utilisé pour représenter des relations sociales (par exemple sur les réseaux socionumériques) ou des liaisons hyertextuelles entre sites web (par exemple dans la blogosphère). Il existe alors une forme d'analogie entre le lien signifié, qu'il soit social ou hypertextuel, et le lien signifiant, c'est-à-dire l'arête reliant les nœuds du réseau dans le graphe. Nous avertissons donc le lecteur, afin qu'il ne soit pas dérouté, que le graphe de réseau est ici utilisé pour représenter non pas des liaisons hypertextuelles mais des proximités thématiques entre les sites d'actualité analysés. En cela, il s'agit plutôt d'un graphe de réseau de nature sémantique, que certains chercheurs travaillant sur l'information en ligne (Weber, Monge, 2011) considèrent comme un complément essentiel aux graphes de réseau de nature hypertextuelle.

traités et en périphérie les sites qui partagent des sujets beaucoup plus isolés. Les nœuds (points) du graphe correspondent à chacun des sites d'actualité, et la taille de chaque nœud est relative au nombre d'articles produits. Les arêtes (liens) entre les nœuds représentent les proximités thématiques, reliant des sites lorsqu'ils abordent des sujets communs. En raison du volume de données traitées (dizaines de milliers d'articles et centaines de sites), la représentation cartographique ainsi générée peut être aisément visualisée sur un écran en couleurs mais s'avère en revanche peu lisible sur un imprimé en noir et blanc. Pour cette raison, nous la présenterons en plusieurs temps. Nous commençons par reproduire la carte dans son intégralité, de façon à montrer sa forme générale et à décrire ses principales zones de façon schématique (figure 1). Nous présenterons ensuite des extraits « zoomés » de cette carte (figures suivantes), afin d'observer plus en détail les territoires informationnels concernés.

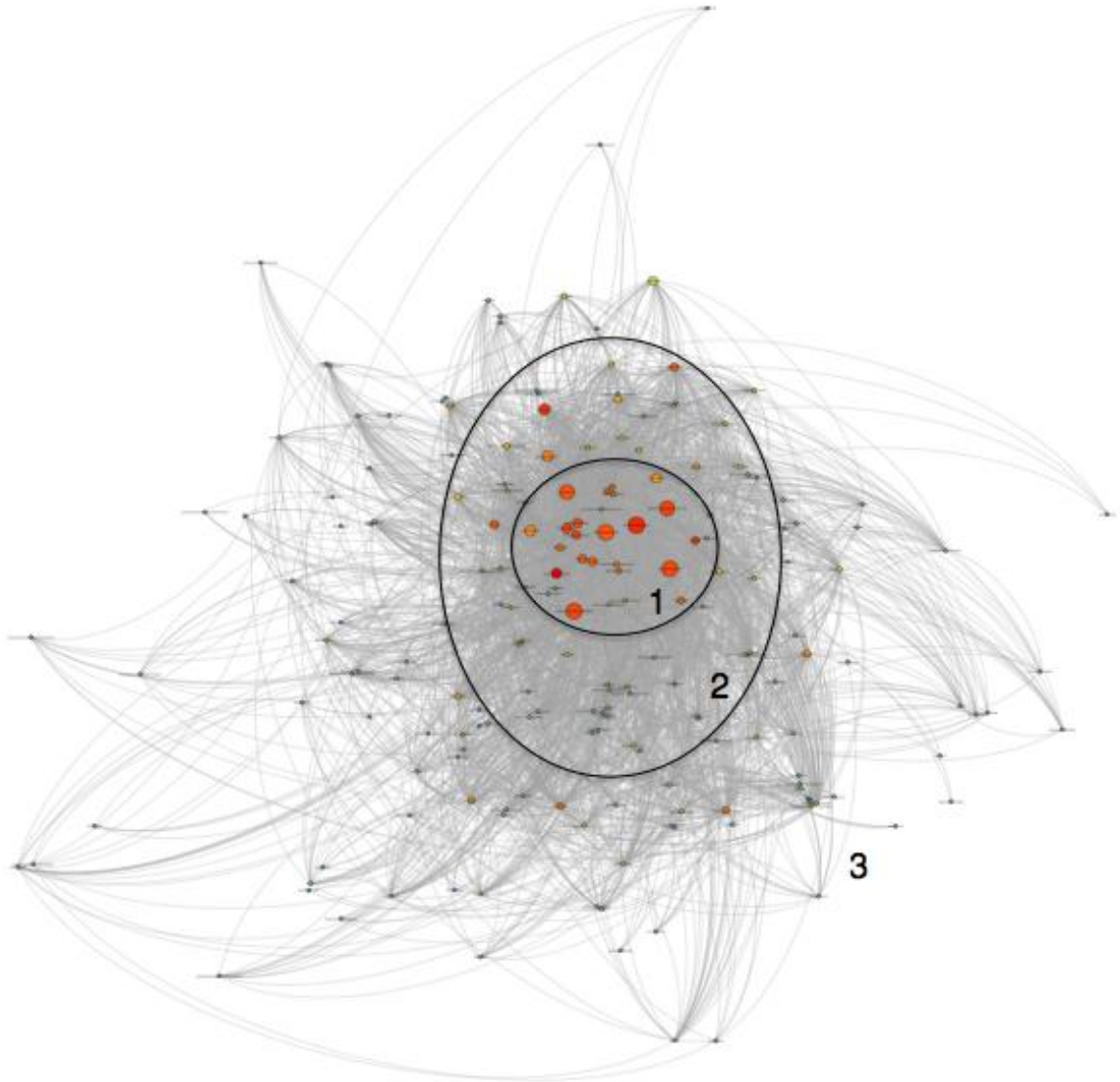

Figure 1 - Représentation cartographique des proximités éditoriales entre sites d'actualités
Nœud = site / Proximité spatiale = proximité éditoriale / Lien = partage de sujets d'actualité
Zone 1 : *mainstream* médiatique / Zone 2 : espace intermédiaire / Zone 3 : information alternative

La représentation cartographique fait apparaître une structure correspondant à la physionomie à la fois concentrée et dispersée de l'information sur le web. Ainsi, selon une logique de cercles concentriques, on s'éloigne progressivement d'un épicentre où les similarités éditoriales entre sites sont très denses autour des sujets les plus traités, pour aller vers des relations éditoriales plus affinitaires autour de sujets moins courants. Plus précisément, nous pouvons distinguer trois grandes zones sur cette carte, qui recouvrent les trois degrés d'originalité éditoriale des sites mis en évidence dans la partie précédente de cet article[18]. La zone 1 est celle du *mainstream* médiatique. Elle rassemble les sites les plus productifs (larges nœuds), qui partagent entre eux un nombre considérable de sujets d'actualité à la « Une » de l'agenda médiatique. La zone 3 est celle de l'information alternative, faite de sujets d'actualité beaucoup moins repris, par des sites eux-mêmes plus restrictifs dans leur production. La zone 2 constitue un espace éditorial intermédiaire, où des sites d'envergure très inégale restent branchés sur l'actualité dominante tout en étant également attentifs à des sujets plus iconoclastes.

Les différents territoires de l'information

Un tel découpage de la carte de l'information sur le web, s'il présente l'avantage d'en dégager la structure générale, n'en demeure pas moins schématique et assez grossier. Il reste maintenant à explorer plus profondément les territoires de l'information composant chacune des zones identifiées. Avec pour valeur ajoutée, au regard de nos résultats antérieurs, de pouvoir mieux apprécier les proximités éditoriales entre sites, que celles-ci soient bilatérales ou multilatérales.

[18] La zone 3 se distingue par le fait de comprendre, pour l'essentiel, des sites présentant plus de 50% d'articles sur des sujets rares, la zone 2 des sites présentant entre 40% et 50% d'articles sur des sujets rares, la zone 1 des sites présentant bien moins de 40% d'articles sur des sujets rares.

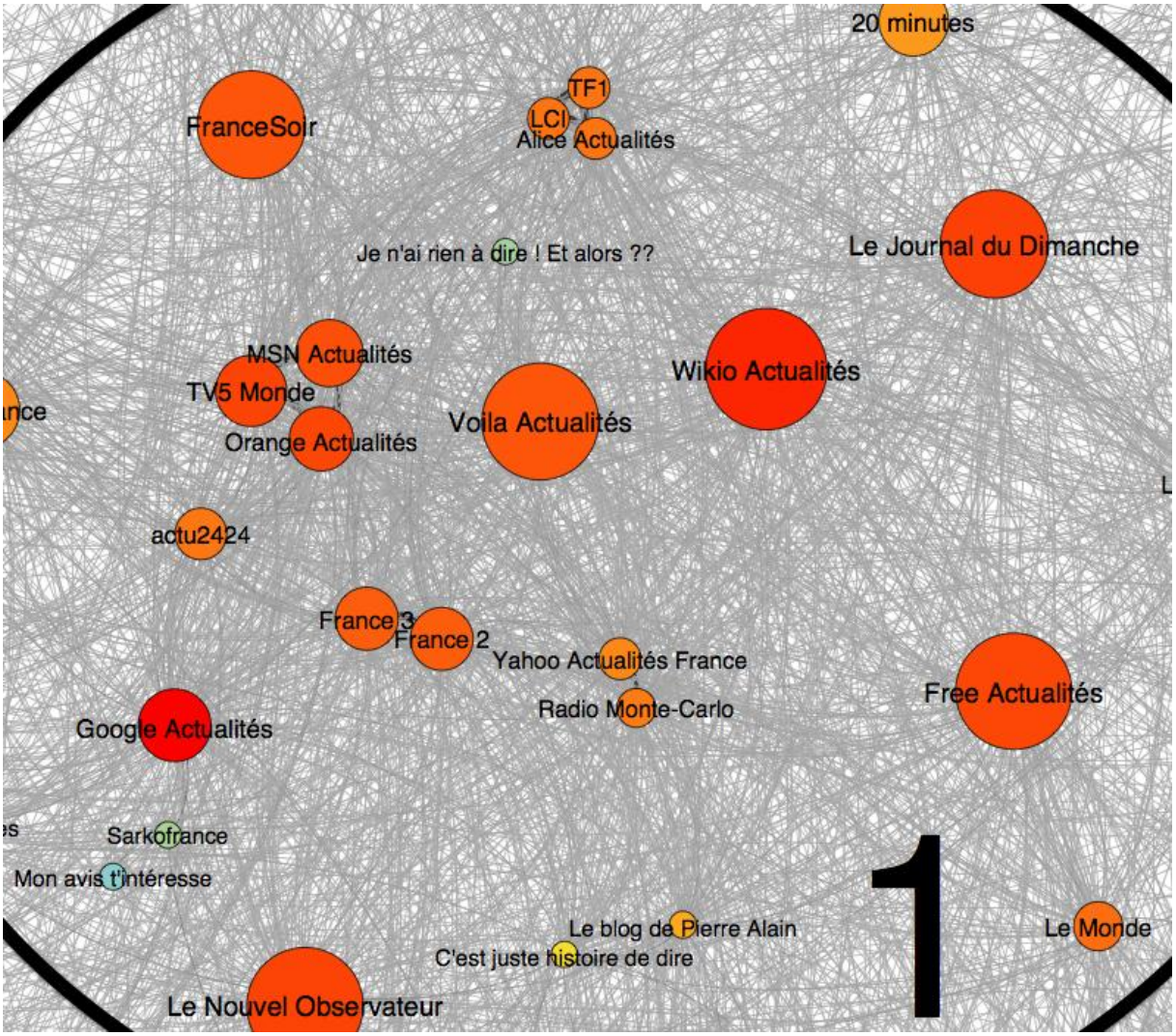

Figure 2 - Extrait de la représentation cartographique - *Mainstream* médiatique (zone 1)

Un premier extrait de ce graphe concerne la zone 1 du *mainstream* médiatique. Il montre la place centrale occupée par les infomédiaires et les sites de médias en ligne (cf. figure 2).
Nous pouvons distinguer aisément les sites les plus productifs en volume d'articles : *Voilà*, *Wikio*, *Free*, du côté des infomédiaires; *France Soir*, *Le Journal du Dimanche*, *Le Nouvel Observateur*, du côté des médias en ligne, ou plus exactement des journaux en ligne. Ces sites constituent ainsi les principaux moteurs du *mainstream* médiatique sur le web en termes d'agenda.
Par ailleurs, la proximité éditoriale entre certains sites est frappante, au point de former de véritables agrégats : ainsi des trios *TF1 - LCI - Alice Actualités* et *TV5 Monde - MSN Actualités - Orange Actualités*, ou des duos *France 2 - France 3* et *Yahoo Actualités - RMC*. Dans certains cas, cette redondance éditoriale s'explique tout simplement par la diffusion d'un même fil d'informations au sein d'un groupe industriel de médias (c'est le cas pour les chaînes de télévision en ligne du groupe TF1 et de France Télévisions). Elle peut également provenir de la fourniture d'informations par une même source, et en particulier par une même agence de presse.
A côté de ces mastodontes de l'information sur le web, on remarque aussi la présence de quelques blogs : *Je n'ai rien à dire ! Et alors ??*, *Sarkofrance*, *Mon avis t'intéresse*, *Le blog de Pierre Alain*, *C'est juste histoire de dire*. Leur présence dans le *mainstream* médiatique peut

s'expliquer par une forte propension à commenter les informations à la Une de l'agenda médiatique. Notons, pour terminer, la position de certains sites en lisière de cette zone 1 : *France Soir*, *20 Minutes*, *Le Nouvel Observateur*, *Le Monde*, *RTL*, *Le Post* (ces deux deniers sites ne sont pas visibles sur l'extrait précédent). Ces sites restent très intégrés au *mainstream* médiatique, mais sans en être au cœur, car ils partagent également quelques sujets d'actualité avec des sites de la zone 2.

La zone 2 constitue, rappelons-le, un espace éditorial intermédiaire, pour une part connecté au *mainstream* médiatique, et pour une autre part assez ouvert à des sujets d'actualité plus rares. Il est composé de médias en ligne, de médias natifs de l'internet, et de blogs, qui toutefois ne se répartissent pas de façon équitable dans toute la zone.
Au sein de cette zone 2, deux grands territoires informationnels peuvent en effet être distingués. La "partie Nord" est surtout dominée par des médias en ligne, qui semblent se rapprocher d'autres sites en fonction de proximités éditoriales et idéologiques. La "partie Sud", elle, fourmille de blogs, dessinant des agrégats éditoriaux typiques de la dimension conversationnelle en vigueur dans la blogosphère.

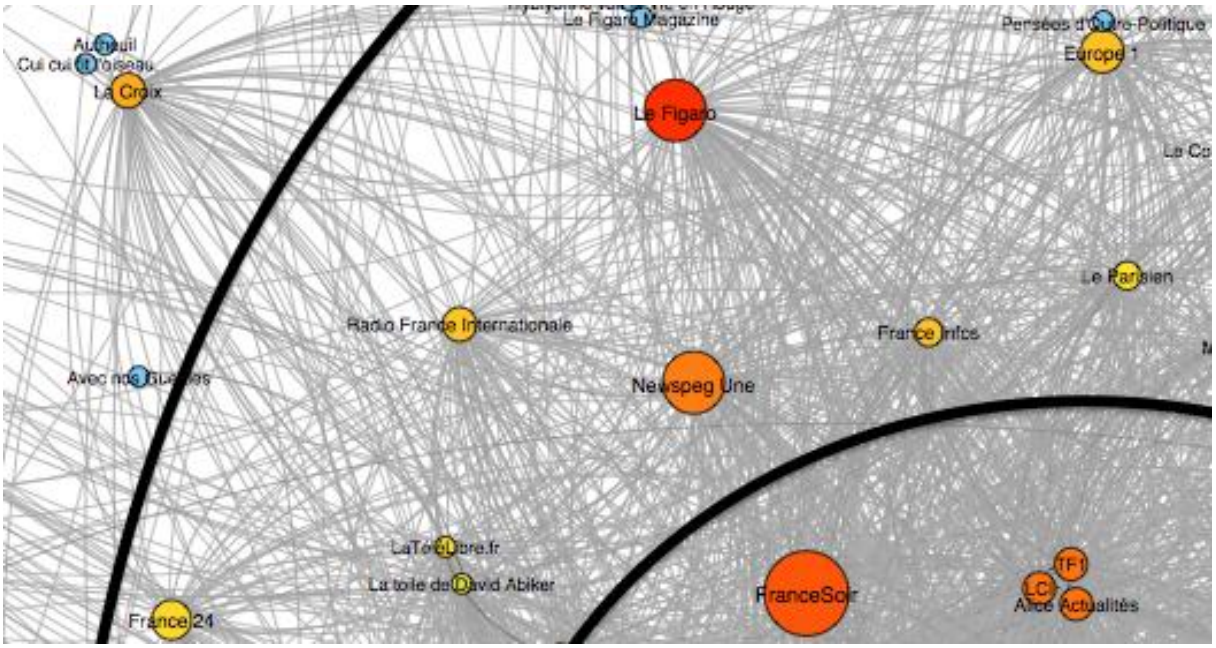

Figure 3 - Extrait de la représentation cartographique - "partie Nord-ouest" de la zone 2

La "partie Nord" de la zone 2 abrite plusieurs médias en ligne à la production d'articles assez conséquente : *Le Point*, *Le Figaro*, *Europe 1*, *Libération*. Ils semblent faire le pont entre le *mainstream* médiatique et l'information alternative. Pour autant, ce rôle de passeur ne s'exerce pas de façon neutre au niveau éditorial : les sujets d'actualité abordés par ces sites, qu'ils soient dominants ou au contraire rares, sont souvent partagés avec des sites qui présentent une proximité éditoriale et idéologique.
Pour illustrer une telle configuration, observons la position occupée par *Le Figaro* (cf. figure 3), souvent classé comme le leader de l'information d'actualité sur le web en termes d'audience. Du côté de la zone 3, il partage des sujets d'actualité avec le site du quotidien catholique *La Croix* et un blog situé dans son orbite immédiate, *Authueil*, dont l'auteur

revendique un ancrage politique dans la démocratie chrétienne[19]. On peut alors émettre l'hypothèse que les sujets rares abordés par *Le Figaro* sont liés à des préoccupations d'ordre religieux. Du côté de la zone 1, *Le Figaro* ne partage pas non plus n'importe quels sujets d'actualité dominants, ou disons plutôt qu'il les partage prioritairement avec certains sites : *France Soir*, et le groupe TF1 dans une moindre mesure. Ici c'est l'hypothèse d'un média "populaire" et plutôt incliné à droite qui ressort. Elle semble être confirmée par le positionnement du *Figaro* au sein même de la zone 2. Outre la proximité avec son satellite *Le Figaro Magazine* considéré comme plus à droite, *Le Figaro* partage également plusieurs sujets d'actualité avec le site d'*Europe 1*, lui-même accolé au blog *Pensées d'outre-politique*, tenu par un « *catholique convaincu* » électeur de centre-droit. Notons aussi le partage de sujets d'actualité avec les radios publiques *France Infos* et *Radio France Internationale* (RFI) qui, là, pourrait mettre en relief la place de l'actualité internationale, traditionnellement forte, dans la couverture médiatique du *Figaro*.
Une configuration assez semblable se retrouve dans la partie cette fois "Nord-Est" de la zone 2, autour du site de *Libération*. Du côté de la zone 3, celui-ci partage des sujets rares avec par exemple *StreetPress.com*, un site participatif « *qui porte les questionnements sociaux, politiques et culturels des jeunes* ». Et du côté de la zone 1, Libération partage des sujets dominants avec *20 Minutes*, quotidien gratuit dont la stratégie de ciblage marketing des "jeunes urbains" est encore plus poussée sur le web[20]. Enfin, au sein de la zone 2, sa proximité la plus immédiate est avec le *Blog d'Eric Mainville*, centré sur l'économie solidaire.

Ainsi, la "partie Nord" de la zone 2 semble polarisée autour d'une poignée de médias en ligne qui s'inscrivent dans le *mainstream* médiatique à partir d'une orientation éditoriale et idéologique particulière, base du partage de sujets d'actualités plus rares. En comparaison, la "partie Sud" de la zone 2 paraît beaucoup plus homogène.
Au sein de cette "partie Sud", aucun site ne se détache véritablement par l'ampleur de sa production. Beaucoup de blogs, ayant produit un ou quelques articles par jour au cours de la période, parsèment ce territoire informationnel. Ils partagent bon nombre de sujets d'actualité entre eux, au point de parfois présenter une grande similarité dans leurs choix éditoriaux.

[19] Il faut noter la présence d'un autre site tout proche d'*Authueil* et *La Croix* : *Cui cui fit l'oiseau*, blog s'ingéniant à tourner en dérision les déclarations les plus conservatrices émaillant l'actualité politique. Ceci rappelle que notre analyse à partir de la notion de sujet d'actualité peut amener à identifier des sites adoptant un cadrage primaire proche alors que leur cadrage médiatique secondaire peut être opposé.
[20] Nicolas Hubé (2008), en combinant analyse des Unes de la presse écrite et observation des rédactions dans la première moitié des années 2000, a bien montré comment la conversion des quotidiens français au « *référentiel de marché* » s'était traduite par des positionnements autour de sujets de société plutôt que d'enjeux politiques. *Libération* était à la pointe de ce mouvement, et son positionnement sur le web pourrait en marquer la continuité.

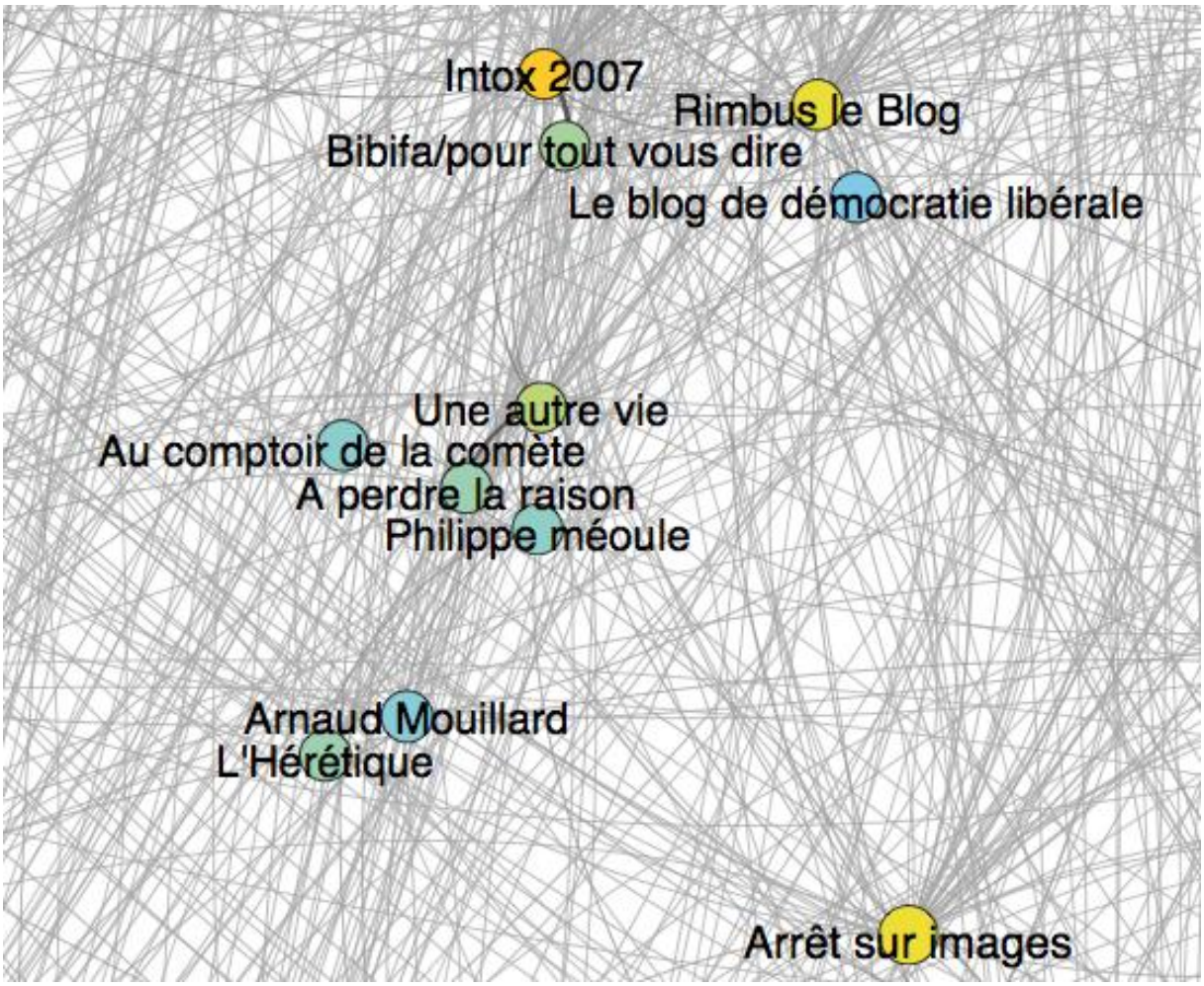

Figure 4 - Extrait de la représentation cartographique - "partie Sud" de la zone 2

Des agrégats de blogs se forment ainsi sur la carte (cf. figure 4), reflétant une dynamique de conversation autant que de publication, typique de la blogosphère (Cardon, Delaunay-Teterel, 2006). Les blogueurs se répondent par le biais de commentaires ou par articles interposés. Ces échanges sont accompagnés par des dispositifs tels que le *blogroll*, liste de blogs suivis de façon privilégiée. Parfois, il s'agit même de blogrolls spécialisés, tels que le *Leftblogs*, regroupant des blogs de gauche donc, et utilisé ici notamment par *Arnaud Mouillard*, *A perdre la raison*, et *Une autre vie*. Pour la plupart d'entre eux, les blogs agrégés dans l'extrait précédent de la représentation cartographique (figure 4), à l'exception du *Blog de démocratie libérale* et de *L'Hérétique* lié aux Modem, sont animés par des militants du Parti socialiste ou du Front de Gauche. En mars 2011, ils se retrouvent autour de positions "antisarkozystes", saisissant les mesures du Président de la République ou du Gouvernement faisant l'actualité pour les commenter de façon critique.
Dans cette "partie Sud" de la zone 2, l'information semble donc constituer un support pour les échanges entre auteurs de blogs. Il s'agit d'une zone de commentaires sur l'actualité, dans laquelle s'intègrent aussi quelques rares sites professionnels comme le site natif de l'internet *Arrêt sur images* (cf. figure 4), les médias en ligne *Marianne* ou *Les Echos*. Ces deux derniers sites se placent dans des parties de la zone 2 où se retrouvent des blogs penchant plus à droite ou revendiquant une forme d'apolitisme.
Les sites de la zone 2, majoritairement des blogs, puisent des sujets d'actualité dans le *mainstream* médiatique, dont ils extraient des morceaux choisis et sur lesquels ils vont

apposer leur point de vue. Les sites de la zone 1 avec lequel ils partagent le plus de sujets d'actualité dominants sont *Le Nouvel Observateur* et *Le Monde*, médias communément situés au centre-gauche ou en tout cas modérés. C'est tout le contraire des sites de la zone 3 avec lesquels quelques sujets d'actualité sont partagés, mais dont les prises de position peuvent s'avérer beaucoup plus radicales.

La zone 3 concerne l'information alternative. Elle est composée de sites qui s'attachent à des sujets d'actualité à l'écart du *mainstream* médiatique. Un tel décalage peut provenir de deux postures : d'un prisme idéologique assumé dans la sélection de l'information comme d'une volonté explicite de se démarquer de l'agenda médiatique traditionnel, en prenant appui sur les contributions d'amateurs.
La première facette de l'information alternative est donc celle d'une information engagée, voire très engagée. La présence dans la zone 3 de *Valeurs Actuelles, Pèlerin Magazine* ou *La Croix*, laisse ainsi penser que les journaux d'opinion ou d'obédience religieuse continuent, sur le web, à apporter leur singularité médiatique, s'attachant à des sujets d'actualité peu ou pas abordés par d'autres. Mais de nouveaux entrants ont poussé encore plus loin ce parti-pris idéologique, portant leur attention de façon quasi-exclusive sur les sujets d'actualité qui rencontrent leurs préoccupations politiques premières. On retrouve ainsi dans cette zone 3 des sites dont nous avions déjà souligné l'originalité éditoriale, en liant cette dernière à un extrémisme politique. La représentation cartographique des proximités éditoriales nous apporte un nouvel d'élément d'intellection, mettant en relief le partage de sujets d'actualité entre sites d'extrême droite et sites d'extrême gauche : *François Desouche* et *Enquête et débat* ont visiblement des préoccupations communes avec l'infomédiaire *Rezo.net* et la galaxie de sites auxquels il renvoie, ou encore avec le site du quotidien communiste *L'Humanité* (cf. figure 5). Sans céder à une interprétation hâtive et usitée selon laquelle « les extrêmes se rejoignent », nous remarquerons néanmoins que ces sites semblent se focaliser sur des sujets d'actualité assez identiques (par exemple la question de l'immigration), sur lesquels ils greffent sans doute des points de vue opposés.

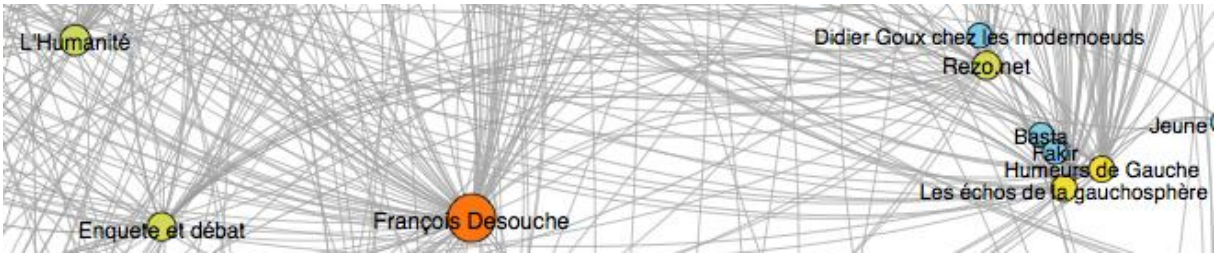

Figure 5 - Extrait de la représentation cartographique - Information alternative très engagée (zone 3)

Aux côtés de ces sites à la ligne éditoriale ouvertement soumise à des partis-pris idéologiques, apparaissent dans cette zone 3 des sites moins unanimement engagés sur le plan politique. Fait notable toutefois, plusieurs s'étaient donnés comme objectif de transformer le modèle médiatique dominant, en ouvrant leur espace de publication aux contributions d'amateurs. Plus exactement, hormis *Le Post* dans la zone 1, tous les sites de notre échantillon possédant une dimension participative se logent dans cette zone 3 : *AgoraVox, Alterinfo, Bellaciao, Cafebabel, Charlie Enchaîné, Citizenside, Enquête et débat, François Desouche, HNS-Info, IndyMedia, Le Gaulois, Le Grand Soir, Mediapart, Minutebuzz, Réseau Voltaire, Rue 89, StreetPress.com, Voie militante, WikiNews.* Certains d'entre eux, tels qu'*Agora Vox, BellaCiao* ou *Mediapart*, se distinguent par leur volume de production d'articles et le fait d'agglomérer autour d'eux des blogs, dans une certaine proximité éditoriale. Ces mêmes sites

voyaient dans le journalisme participatif un moyen de révolutionner l'information. Leur pari semble en partie tenu, dans la mesure où ils sont porteurs de sujets rares, qui sans eux n'auraient peut-être jamais connu d'existence médiatique. Néanmoins, leur éloignement du cœur du *mainstream* médiatique, et le risque de constituer des niches d'information relativement fermées et confidentielles, nous amène à réfléchir de façon plus globale à la structuration de l'information sur le web.

Retours sur d'autres cartographies du web

La structuration de l'information sur le web, telle que nous l'avons identifiée à travers cette analyse des proximités éditoriales, entre en résonance avec d'autres travaux reposant, eux, sur une cartographie des hyperliens entre sites.

Aux Etats-Unis, la représentation de la « *global news arena* » (Reese et al., 2007) avait mis en évidence l'arrimage hypertextuel des blogs aux sites de médias traditionnels. Cette relation est confirmée ici au niveau des contenus, tout en étant affinée : si les blogs s'appuient dans bien des cas sur l'actualité divulguée par des sites professionnels pour la commenter, ils vont surtout s'arrêter sur les sujets livrés par des sites présentant une orientation éditoriale et idéologique compatible avec la leur. Ce type de relations éditoriales apparaît dans notre recherche entre les blogs de la "partie Sud" de la zone 2 et certains médias en ligne de la zone 1, sur lesquels ils viennent prendre appui. Cependant, les relations éditoriales les plus nourries des blogs s'établissent avec les sites natifs de l'internet, en particulier au sein de la zone 3, le plus souvent sur la base de proximités idéologiques là encore.

Il s'agit d'ailleurs là sans doute de l'axe le plus structurant dans la représentation cartographique que nous avons analysée : les blogs et nombre de sites natifs de l'internet semblent relégués à la périphérie de l'information sur le web tandis que, globalement, les médias en ligne s'avèrent davantage en prise avec les infomédiaires pour ce qui est de la constitution de l'agenda médiatique général. A cet égard, notons qu'une possible fermeture de la liaison entre médias en ligne et blogs avait déjà été signalée par une cartographie des liens hypertextuels des principaux sites de médias français (De Maeyer, 2010). Cette cartographie avait même montré une tendance au repli sur soi des médias en ligne, réservant leurs liens aux autres entités de leur groupe industriel de communication : certaines des fortes proximités éditoriales, voire des identités éditoriales uniques au sein de la zone 1 (ex : entre chaînes de télévision du même groupe), relevées via notre analyse de discours, pourraient s'inscrire dans une même logique.

Au-delà de l'information sur le web, des rapprochements peuvent également être établis avec des recherches portant sur la politique en ligne. Depuis le travail inaugural d'Adamic et Glance (2005) sur la blogosphère états-unienne au moment de la campagne présidentielle américaine de 2004, on sait en effet que les sites web ont tendance à privilégier un entre-soi hypertextuel (Flichy, 2008). Les travaux menés en France ont abouti au même constat d'une partition de la blogosphère en fonction des différents partis politiques ou courants de pensée, avec toutefois des passerelles hypertextuelles plus nombreuses au fil des ans (Cardon et al., 2011). Ces travaux ont aussi mis en évidence le rôle des sites de médias et de commentateurs, correspondant aux sites d'actualité, en tant que relais hypertextuels entre les différents camps politiques.

Notre recherche apporte sur ce point une couche d'intelligibilité supplémentaire, en renseignant sur la nature des contenus des sites d'actualité. Les plus engagés politiquement parmi eux sont aussi ceux qui traitent des sujets d'actualité les moins partagés. Ceci n'est pas de nature à conférer une accessibilité médiatique à l'ensemble des questions de société, et à les mettre et en discussion, comme on pourrait l'espérer dans une perspective habermassienne de l'espace public. Cependant, il faut remarquer les quelques relations thématiques de certains

sites engagés politiquement, aux avis moins tranchés, avec d'autres sites d'actualité davantage reliés au *mainstream* médiatique. On peut donc faire l'hypothèse que ces sites intermédiaires entre centre et périphérie du paysage de l'information sur le web pourraient constituer -occasionnellement- les courroies de transmission par lesquelles des idées ou des informations alternatives quelque peu confinées dans les méandres de l'internet viennent alimenter un espace médiatique plus large. C'est plus généralement l'hypothèse d'un espace public réticularisé au sens de la *networked public sphere* de Benkler (2006), qui trouverait là une description plus empirique de ses ramifications sémantiques autant qu'hypertextuelles.

CONCLUSION

Avec cette recherche, l'offre d'informations d'actualité sur le web en France a pu être appréhendée dans sa globalité. En intégrant des sites de médias déjà existants (issus de l'imprimé ou de l'audiovisuel) autant que des nouveaux entrants (blogs, sites participatifs, portails et agrégateurs), tous les types de sites ont été considérés. Leurs temporalités respectives de publication ont été également prises en compte, au cours d'une analyse à grande échelle, sur une dizaine de jours. Ainsi, tout au long d'une période d'observation faisant succéder une phase exceptionnellement chargée sur le plan de l'actualité à une phase plus ordinaire, ce sont plusieurs dizaines de milliers d'articles, produits par près de deux cents sites, qui ont été soumis à analyse.

Les résultats obtenus permettent d'évaluer le niveau de pluralisme de l'information sur le web, en France.
Tout d'abord, l'identification des sujets d'actualité autorise à se prononcer sur le niveau de variété éditoriale. Indéniablement, ce niveau est élevé voire très élevé sur le web français, avec plusieurs centaines de sujets abordés chaque jour. L'agenda médiatique qui en résulte n'est pas pour autant équilibré, bien au contraire. Le présent travail met en relief une dichotomie permanente entre l'ultra-médiatisation de certains sujets, retraités à l'envi, et la dissémination de sujets peinant à trouver une visibilité en dehors de leurs sources d'origine.
Les choix éditoriaux conduisant à une telle situation ont été également approchés, autre apport de cette étude, par une appréciation du pluralisme interne à chacun des sites. Selon qu'ils s'inscrivent pleinement dans le courant médiatique dominant, ou à l'inverse dans une information alternative, ou encore qu'ils soient à l'interface entre *mainstream* et originalité éditoriale, les sites concourent à dessiner des territoires informationnels spécifiques.
La présente recherche a ainsi permis de cartographier l'espace web des sites d'actualité français, et d'analyser la nature des informations qui y sont proposées. De ce point de vue, une ligne de fracture semble se dessiner entre les différentes catégories de sites. Les infomédiaires et les médias en ligne les plus productifs constituent les moteurs de la redondance éditoriale sur le web, alors que les blogs, et les sites natifs de l'internet dans une moindre mesure, créent eux des foyers d'originalité éditoriale. Notre recherche ne conforte ainsi que pour partie la thèse du « *More is Less* » avancée dans de précédents travaux (Paterson, 2006; Fenton, 2009, Boczkowski, 2010) : en intégrant de façon beaucoup plus significative certaines catégories de sites (des dizaines de sites natifs de l'internet et surtout de blogs), elle a mis en évidence qu'une telle multiplication des espaces de pubication sur le web se doublait de relations éditoriales tout aussi complexes, autour de sujets portés à la Une de l'agenda médiatique mais aussi autour du partage de sujets d'actualité plus rares.

Afin de pouvoir traiter un corpus de cette envergure, notre analyse requérait une posture quantitative. Elle gagnera à être prolongée par une perspective plus qualitative. Un même sujet d'actualité peut en effet être traité de façon fort disparate, et l'analyse de discours sémiopragmatique, comme proposée dans un autre article de ce dossier (Touboul et al., 2012), viendra alors opportunément en rendre compte. L'analyse du pluralisme de l'information, même avec cette granularité plus fine, n'en reste pas moins une analyse de la diversité *offerte*, à laquelle échappe la diversité *consommée* (Benhamou, Peltier, 2006). Pour passer, en d'autres termes, du stade de *diversity as sent* au stade de *diversity as received* (Van der Wurff, 2011), il faut cette fois inscrire ces discours médiatiques dans leur contexte social de réception. L'article consacré à la circulation des sujets d'actualité sur Twitter (Rieder et al., 2012) nous fournira des éléments à ce sujet. Mais nul doute que ce travail mériterait d'être amplifié à

l'observation des pratiques des internautes dans leur ensemble pour savoir ce qui, entre la redondance et l'originalité, l'emporte lors de la circulation sociale des informations.

RÉFÉRENCES

ADAMIC L., GLANCE N., 2005, « The Political Blogosphere and the 2004 US Election; Divided They Blog », *Proceedings of the 3rd international workshop on Link discovery - LinkDD '05*, en ligne, [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.9009], consulté le 27 mars 2012.

ARQUEMBOURG, J., 2011, *L'évènement et les médias. Les récits médiatiques des tsunamis et les débats publics (1755-2004)*, Paris, Editions des Archives Contemporaines.

BENHAMOU F., PELTIER S., 2006, « Une méthode multicritère d'évaluation de la diversité culturelle. Application à l'édition de livres en France », in *Création et diversité au miroir des industries culturelles. Actes des journées d'économie de la culture*, X. Greffe (éd.), Paris : La Documentation française, pp. 313-344.

BENKLER Y., 2006, *The Wealth of Networks. How Social Production Transforms Markets and Freedom*, New Heaven / London : Yale University Press.

BERGER P., LUCKMANN T., 1996, *La Construction sociale de la Réalité*, Paris : Masson/Armand Colin [première publication anglophone : 1966].

BOCZKOWSKI P. J., 2010, *News at Work. Imitation in an Age of Information Abundance*, Chicago : University of Chicago Press.

BOUQUILLION P., MATTHEWS J., 2010, *Le Web collaboratif. Mutations des industries de la culture et de la communication*, Grenoble : PUG.

BOURDIEU P., 1996, *Sur la télévision (suivi de L'emprise du journalisme)*, Paris : Liber.

BRUNS A., 2008, « The Active Audience : Transforming Journalism from Gatekeeping to Gatewatching », in *Making Online News. The Ethnography of New Media Production*, C. Paterson and D. Domingo ed., New York : Peter Lang, pp. 171-184.

CARDON D., 2010, *La démocratie Internet. Promesses et limites*, Paris : Seuil.

CARDON D., DELAUNAY-TETEREL H., 2006, « La production de soi comme technique relationnelle. Un essai de typologie des blogs par leurs publics », *Réseaux*, n° 138, pp. 15-71.

CARDON D., FOUETILLOU G., LERONDEAU C., PRIEUR C., 2011, « Esquisse de géographie de la blogosphère politique (2007-2009) », in *Continuer la lutte.com*, F. Greffet (dir.), Paris : Presses de Sciences Po.

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

CHARAUDEAU P., 2005, *Les médias et l'information. L'impossible transparence du discours*, Bruxelles : De Boeck-Ina.

CHARON J.-M., 2010, « Stratégies pluri-médias des groupes de presse », in *Les Cahiers du journalisme*, n° 20, pp. 54-74.

DAGIRAL E., PARASIE S., 2010, « Presse en ligne : où en est la recherche ? », in *Réseaux*, La Découverte, n° 160-161, pp. 13-42.

DEARING J. W., ROGERS E. M., 1992, *Communication Concepts 6. Agenda-setting*, Thousand Oaks: Sage.

DEUZE M., 2003, "The web and its journalisms: Considering the consequences of different types of newsmedia online", in *New Media & Society* 2(5), pp.203–30.

EGGHE L., 2005, *Power Laws in the Information Production Process : Lotkaian Informetrics*, Oxford: Elsevier Academic Press.

ESQUENAZI J.-P., 2002, *L'écriture de l'actualité : pour une sociologie du discours médiatique*, Grenoble : PUG.

FENTON N. (ed.), 2010, *New Media, Old News. Journalism and Democracy in the Digital Age*, London, Thousand Oaks, New Delhi, Singapore : Sage.

FLICHY P., 2008, « Internet et le débat démocratique », in *Réseaux*, La Découverte, n° 150, pp. 159-185.

GAMSON W., MODIGLIANI A., 1989, "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach", in *American Journal of Sociology*, 95 (1), pp. 1-38.

GOFFMAN E., 1991, *Les cadres de l'expérience*, Paris : Les Editions De Minuit.

GREFFET F., WOJCIK S., 2008, « Parler politique en ligne : une revue des travaux français et anglo-saxons », in *Réseaux*, La Découverte, n° 150, pp. 19-50.

HUBE N., 2008, *Décrocher la « Une ». Le choix des titres de première page de la presse quotidienne en France et en Allemagne (1945-2005)*, Strasbourg : P.U.S.

LANCELOT A., 2005, *Les problèmes de concentration dans le domaine des médias*, Rapport pour le Premier ministre, Paris : La Documentation française, en ligne [http://lesrapports.ladocumentationfrancaise.fr/BRP/064000035/0000.pdf], consulté le 27 mars 2012.

LESKOVEC J., BACKSTROM L., KLEINBERG J., 2009, « Meme-tracking and the dynamics of the news cycle », in *kdd'09 - International Conference on Knowledge Discovery and Data Mining*, Paris, en ligne, [http://www.cs.cornell.edu/home/kleinber/kdd09-quotes.pdf], consulté le 27 mars 2012.

MAEYER (de) J., 2010, "Mapping the hyperlinked environment of online news. Issues and challenges for the French news sites", in *IAMCR Conference*, en ligne [http://juliettedm.files.wordpress.com/2010/06/demaeyer_mapping-hyperlink-environment.pdf], consulté le 27 mars 2012.

MARCHETTI D., 2002, « Sociologie de la production de l'information. Retour sur quelques expériences de recherche », *Cahiers de la recherche sur l'éducation et les savoirs*, n° 1, pp. 17-32.

MARTY E., 2010. *Journalismes, discours et publics : une approche comparative de trois types de presse, de la production à la réception de l'information*, Thèse de doctorat en Sciences de l'information et de la communication sous la direction de P. Marchand & A. Burguet, Université de Toulouse.

MARTY E., REBILLARD F., SMYRNAIOS N., TOUBOUL A., 2010, « Variété et distribution des sujets d'actualité sur l'internet. Une analyse quantitative de l'information en ligne », in *Mots. Les langages du politique*, n° 93, pp. 107-126.

NEVEU E., QUERE L., 1996, « Le temps de l'évènement », *Réseaux*, n° 75, pp. 7-21

MC QUAIL D., 1992, *Media Performance: Mass Communication and the Public Interest*, London: Sage.

McCOMBS M., SHAW D., 1972. « The agenda-setting function of mass media », *Public Opinion Quarterly* 36 (2), pp. 176-187.

MERCIER A., PIGNARD-CHEYNEL N., 2011, « L'appropriation des réseaux sociaux par les webjournalistes en France », colloque *Médias'011*, Université d'Aix-en-Provence, en ligne, [http://www.medias011.univ-cezanne.fr/fileadmin/Medias11/Documents/A4/MERCIER_PICHARD.pdf], consulté le 27 mars 2012.

MIÈGE B., 2010, *L'espace public contemporain. Approche info-communicationnelle*, Grenoble : PUG.

MOUILLAUD M., TETU J.-F., 1989, *Le journal quotidien*, Lyon : PUL.

OUAKRAT A., 2011, *Publicité en ligne sur les sites de presse issus de l'imprimé. Construction du marché, logiques de fonctionnement et perspectives d'évolution*, Thèse de doctorat en Sciences de l'information et de la communication sous la direction de N. Sonnac, Université Panthéon-Assas Paris 2.

Paterson C., 2007, « International News on the Internet : Why More is Less », Ethical Space: The International Journal of Communication Ethics, vol. 4, n° 1/2, pp. 57-66.

PETERS J., 2004, « The market-place of ideas: A history of a concept », in *Toward a Political Economy of Culture: Capitalism and Communication in the Twenty-first Century*, A. Calabrese & C. Sparks (eds), Lanham, MD: Rowman and Littlefield, pp. 65–82.

QUANDT T., 2008, « News Tuning and Content Management : An Observation Study of Old and New Routines in German Online Newsrooms », in *Making Online News. The Ethnography of New Media Production*, C. Paterson & D. Domingo (eds.), New York : Peter Lang, pp. 77-97.

REBILLARD F., 2006, « Du traitement de l'information à son retraitement. La publication de l'information journalistique sur l'internet », in *Réseaux*, La Découverte, n° 137, pp. 29-68.

REESE S., RUTIGLIANO L., HYUN K., JEONG J., 2007, « Mapping the blogosphere. Professional and citizen-based media in the global news arena », *Journalism*, vol. 8, n° 3, pp. 235-261.

RIEDER B., SMYRNAIOS N., 2012, « Pluralisme et infomédiation sociale de l'actualité : le cas de Twitter », Réseaux, n° 176.

RINGOOT R., ROCHARD Y., 2005, « Proximité éditoriale. Normes et usages des genres journalistiques », *Mots. Les langages du politique*, n° 77, *Proximité*, pp. 73-90.

RUELLAN D., 2007, *Le journalisme ou le professionnalisme du flou*, Grenoble : PUG.

SALTON G., WONG A., YANG C.S., 1975, « A Vector Space Model for Automatic Indexing », *Communications of the ACM*, vol. 18, n° 11, pp. 613-620.

SCHEUFELE D. A., 2000, « Agenda-Setting, Priming, and Framing Revisited: Another Look at Cognitive Effects of Political Communication », *Mass Communication & Society, 3*(2&3), pp.297-316.

SEDEL J., 2011, « Bondy Blog. Le travail de représentation des « habitants de la banlieue » par un média d'information participative », *Réseaux*, La Découverte, n° 170, pp. 103-133.

SERFATY V., 2006, « Les blogs et leurs usages politiques lors de la campagne présidentielle de 2004 aux États-Unis », *Mots. Les langages du politique*, n° 80, *La politique mise au net*, pp. 25-35.

SMYRNAIOS N., MARTY E., REBILLARD F., 2010, “Does the ‘Long Tail’ apply to online news? A quantitative analysis of French-speaking websites”, *New Media and Society*, Sage Publications, vol. 12, n° 8, pp. 1244-1261.

SMYRNAIOS N., REBILLARD F., 2009, « L’actualité selon Google. L’emprise du principal moteur de recherche sur l’information en ligne », *Communication et langages*, n° 160, pp. 95-109.

TESSIER M., 2007, *La presse au défi du numérique*, Rapport pour le Ministre de la culture et de la communication, en ligne, [http://www.ddm.gouv.fr/IMG/pdf/rapport-tessier-fev2007.pdf], consulté le 27 mars 2012.

TOUBOUL et al., 2012

VAN DER WURFF R., 2011, « Do audiences receive diverse ideas from news media? Exposure to a variety of news media and personal characteristics as determinants of diversity as received », in *European Journal of Communication*, vol. 26, n° 4, pp. 328-342.

VERON E., 1981, *Construire l’événement. Les médias et l’accident de Three Miles Island*, Paris : Éditions de Minuit.

WEBER M.S., MONGE P., 2011, « The Flow of Digital News in a Network of Sources, Authorities, and Hubs », *Journal of Communication*, n° 61, pp. 1062–1081.

ANNEXES
Echantillon final - Sites ayant publié des articles entre le 7 et le 17 mars 2011.

Médias en ligne : 43
Cette catégorie inclut des sites dépendant de journaux, radios, et télévisions.

20 minutes
Alternatives Economiques
Arrêt sur images
Arte Radio
Causeur
Europe 1
Fakir
France 2
France 24
France 3
France Infos
FranceSoir
Itélé
L'Expansion
L'Express
L'Humanité
La Croix
La Tribune
LCI
Le Figaro
Le Figaro Magazine
Le Journal du Dimanche
Le Monde
Le Monde Diplomatique
Le Nouvel Observateur
Le Parisien
Le Point
Les 4 Vérités
Les Echos
Libération
Marianne
Metro
Notre Temps
Paris Match
Pèlerin Magazine
Politis
Radio France Internationale
Radio Monte-Carlo
RTL France
TF1
TV5 Monde
Valeurs Actuelles
VSD

Sites natifs de l'internet : 40
Cette catégorie inclut des sites à dominante participative pour certains d'entre eux.

Acrimed
Actualité Française
Agora Vox
Alterinfonet
Basta
Bellaciao
Bondy blog
Brave Patrie
Cafebabel
Charlie Enchaîné
Citizenside
Contrepoints
Délits d'Opinion
Enquete et débat
Fluctuat.net
France Matin
François Desouche
HNS-Info
Indymédia Paris
LaTéléLibre.fr
Le Gaulois
Le Grand Soir
Le Post
Le volontaire
Les infos
Mediapart
Minutebuzz
Oulala
Owni
Planet.fr
Politique.net
Réseau Voltaire
Respublica
Riposte laïque
Rue 89
Slate.fr
StreetPress.com
Toulouse 7
Voie militante
WikiNews

Blogs : 102

Cette catégorie inclut des blogs indépendants et des blogs rattachés à d'autres sites.

[Unhuman]
A perdre la raison
A toi l'honneur
Abadinte
Alain Godard
Arnaud Mouillard
Article 11
Au comptoir de la comète
Aurélien veron
Autheuil
Avec nos Gueules
Bah !? by CC
Bibifa/pour tout vous dire
Bivouac-ID
Blog de Claude Guillon
Blog de Paul Jorion
Blog gaulliste libre
Bruno Roger petit
Bug brother
C'est juste histoire de dire
Carnet de notes de Yann Savidan
Chez Homer
Christophe Ginisty
Clémentine Autain
Comité de salut public
Coulisses de Bruxelles
Coulisses de Sarkofrance
Cpolitic
Cui cui fit l'oiseau
Des Pas Perdus
Didier Goux
Didier Goux chez les modernoeuds
Diner's Room
Exprimeo.fr
Extrême centre
Fucking disgrace
Hashtable
Humeurs de Gauche
Intox 2007
Ivan Rioufol
Je n'ai rien à dire ! Et alors ??
Jean Gadrey
Jeune garde 87
Koztoujours
L'Hérétique
L'insolent
La chronique du yéti
La maison du faucon
La plume d'Aliocha
La république du peuple
La toile de David Abiker
Lait d'Beu
Le blog d'Eric Mainville
Le blog de démocratie libérale
Le blog de Gabale
Le Blog de Guy Birenbaum
Le blog de Jean-Michel Apathie
Le blog de Pierre Alain
Le blog de Polluxe
Le Blog de Superno
Le blog politique de Dédalus
Le Coucou de Claviers
Le grumeau
Le Monolecte
Le salon beige
Les eaux glacées du calcul égoïste
Les échos de la gauchosphère
Les Entrailles de Mademoiselle
Les Jours et l'Ennui de Seb Musset
Les mots ont un sens
Les Privilégiés parlent aux français
Marc Vasseur
Merle moqueur
Michel Abhervé
Mon avis t'intéresse
Mon Mulhouse
Nouvel Hermes
Objectif Liberté
Olympe et le Plafond de Verre
Partageons mon avis
Pensées d'Outre-Politique
Philippe méoule
Philippe Sage
Piratages
Plume de presse
Rebelles.info
Richard trois
Rimbus le Blog
Ruminances
Sarkofrance
Sarkostique
Sarkozy an 3 / Christophe Barbier
SLOVAR les nouvelles
Thierry Crouzet
Tian
Tizel
Torapamavoa Nicolas

Toreador.fr
Trublyonne voit la Vie en Rouge
Une autre vie
Variae
Yves Thréard

Infomédiaires : 14

Cette catégorie inclut des services d'actualité de portails et des services d'agrégation automatique de nouvelles.

actu2424
Alice Actualités
EditoWeb Magazine
Free Actualités
Google Actualités
Info2424.info
MSN Actualités
Newspeg Une
Orange Actualités
PaperBlog
Rezo.net
Voila Actualités
Wikio Actualités
Yahoo Actualités France