

Internet, pluralisme et redondance de l'information.

Recherche exploratoire soutenue en 2008
par la Maison des Sciences de l'Homme Paris-Nord.

- Rapport scientifique (avril 2009) -

Responsable scientifique : **Franck Rebillard**
Maître de conférences à l'Université Lyon 2
Laboratoire Elico (EA 4147 – Université de Lyon)
Bénéficiaire d'un Congé Recherche (71 ème section CNU) lors du 1er semestre 2008-
2009.
Franck.Rebillard@univ-lyon2.fr

I Présentation générale	2
1. Descriptif du projet	2
2. Organisation du travail	3
3. Participants à la recherche	5
II Travaux réalisés	6
1. Axe « analyse de contenu des sites web »	6
1.1 Constitution des données	6
1.1.1 Délimitation du corpus de sites web	6
1.1.2 Extraction des contenus	9
1.2 Analyse quantitative du pluralisme offert	11
1.2.1 Méthode semi-automatisée d'analyse de contenu	11
1.2.2 Calculs de variété et d'équilibre	13
1.2.3 Approfondissements autour de l'hypothèse de « longue traîne »	17
1.3 Prolongements et compléments de la recherche	24
1.3.1 Analyse du traitement médiatique de l'élection de Barack Obama.	24
1.3.2 Représentation graphique des proximités éditoriales	29
1.3.3 Suivi diachronique des informations	34
2. Axe « socio-économie des modalités de production »	42
2.1 Choix des structures observées	42
2.2 Modalités de production observées	44
2.2.1 Libération.fr : atouts et contraintes d'une grande entreprise de presse.	44
2.2.2 Bakchich.info : un webzine dans la filiation des radios libres ?	47
2.2.3 Rue89 : le « participatif encadré »	51
2.2.4 PaperBlog : tirer partie de la richesse du web, à peu de frais.	54
2.3 Analyse comparative	57
2.3.1 Des similitudes au niveau des personnels de rédaction.	57
2.3.2 Des modalités de production alternatives pour des informations originales ?	58
3. Conclusion provisoire : des conditions de production à la nature des contenus, une interaction continue.	61
3.1 Analyse des segments répétés	61
3.2 Répartition des segments répétés et des hapax dans les différentes sources	63
III Bilan et perspectives	66
1. Rappel des objectifs initiaux	66
2. Résultats obtenus	66
3. Continuation de la recherche	67
4. Communications et publications scientifiques	68
<i>Références bibliographiques</i>	69
<i>Annexes</i>	71

I Présentation générale

1. Descriptif du projet

La question du pluralisme de l'information est essentielle dans les sociétés démocratiques dès lors que l'on considère les médias comme des moyens d'expression et d'échange entre diverses opinions, autant que comme des outils d'appréhension de la vie sociale. Afin d'assurer ce pluralisme, les pouvoirs publics assurent une régulation du secteur de la presse écrite (ex : aides de l'État pour les quotidiens nationaux d'information politique et générale à faibles ressources publicitaires) et de l'audiovisuel (ex : contrôle du temps de parole des différentes formations politiques par le Conseil Supérieur de l'Audiovisuel).

Pour l'internet en revanche, aucune mesure d'envergure n'a été prise à ce jour en France : le pluralisme est supposé découler de la multiplicité des sources d'information disponibles, comme le laissent penser de récents rapports publics consacrés à ces questions (Lancelot, 2005 ; Tessier, 2007). Or, une étude plus approfondie fait apparaître une piste d'investigation inverse : la multiplicité des espaces de publication sur l'internet pourrait favoriser le pluralisme autant qu'une certaine redondance des informations en circulation. En effet, nombre d'articles publiés sur des sites de presse en ligne, fondés dans bien des cas sur des dépêches d'agence, se voient par exemple compilés par des agrégateurs ou commentés sur des blogs (Rebillard, 2006).

La recherche exploratoire nommée *Internet, pluralisme et redondance de l'information* (IPRI) visait à apporter un début de réponse empirique à ces hypothèses. En prenant notamment appui sur des méthodes quantitatives, ce qui était jusqu'ici inédit, elle a notamment amené des chercheurs de disciplines différentes (Informatique et Sciences de l'information et de la communication) à bâtir ensemble un protocole d'observation innovant. A travers une première expérimentation portant sur un corpus de plusieurs dizaines de sites, représentatifs des différentes catégories de sites d'information en ligne, une analyse semi-automatisée a livré quelques premiers résultats intéressants quant à la part de création originale dans les contenus d'actualité en circulation sur l'internet. Ces résultats partiels viennent compléter les connaissances relatives aux modalités de production de l'information en ligne, que le présent projet de recherche a contribué à étendre en parallèle. Au final, cette recherche exploratoire IPRI jette les bases d'un programme plus large qui viserait à rendre compte plus précisément - toujours sur le double plan de l'analyse de contenu et des enquêtes socio-économiques mais cette fois-ci à plus large échelle - de la dialectique entre pluralisme et redondance de l'information sur l'internet.

2. Organisation du travail

Le travail s'est organisé principalement en deux phases, et autour de deux approches du pluralisme de l'information sur l'internet : analyse de contenu des sites web, et enquêtes socio-économiques auprès des acteurs de la filière de l'information en ligne.

La première phase a eu une vocation préparatoire. Elle a consisté principalement en des séminaires de présentation mutuelle des travaux des chercheurs et d'élaboration d'une méthodologie commune.

- Un premier séminaire plénier s'est tenu à la MSH Paris-Nord (avec visioconférence à l'UQAM) les 5 et 6 juin 2008. Les chercheurs participant au projet IPRI, provenant d'horizons disciplinaires différents (Informatique et Sciences de l'information et de la communication) voire de traditions différentes à l'intérieur même de l'interdiscipline des Sic (entre analyse du discours et socio-économie par exemple), ont pu présenter leurs méthodes et travaux respectifs.

A l'issue de ce séminaire introductif, il est apparu que les efforts de mise en commun des connaissances pour bâtir une méthodologie commune devaient avant tout porter sur l'axe « analyse de contenu des sites web », car l'axe « socio-économie de l'information en ligne » était déjà bien rôdé suite à un programme de recherche précédent (Programme *Diversité culturelle et mutations des industries de la culture, de l'information, et de la communication*, Ministère de la culture et de la communication / MSH Paris-Nord, 2007).

- Deux réunions du sous-groupe « analyse de contenu des sites web » ont eu lieu à l'Insa de Lyon, le 17 juin et le 26 juin 2008. Avec pour objectif de passer d'une analyse qualitative, voire spéculative, à une analyse quantitative et formalisable du pluralisme de l'information sur l'internet. Les échanges lors de cette réunion ont permis d'identifier les différents niveaux possibles (lexicaux, sémantiques) d'une analyse automatisée des similarités de contenus journalistiques ; de préciser les critères caractérisant le pluralisme de l'information, à l'issue d'une synthèse critique d'études antérieures portant sur la diversité culturelle ; de procéder à une première délimitation du corpus visé ; de dessiner un premier protocole d'expérimentation.

- Les propositions du sous-groupe « analyse de contenu des sites web » ont été soumises à l'ensemble des chercheurs lors du second séminaire plénier IPRI, le 3 juillet 2008, à la MSH Paris-Nord (avec visioconférence à l'Insa de Lyon). Sur cette base ont alors été définis respectivement les modalités de coordination entre l'axe « analyse de contenu des sites web » et l'axe « socio-économie de l'information en ligne », un corpus commun, ainsi que le planning opérationnel de la phase suivante d'expérimentation.

La seconde phase a donc été dévolue à une première expérimentation.

- Les mois de septembre et octobre ont servi à formaliser et tester le protocole d'expérimentation pour l'axe « analyse de contenu des sites web » : développement du logiciel d'extraction des fils RSS des sites du corpus ; constitution de la base de données et de ses fonctionnalités ; mise au point d'une méthode semi-automatisée combinant analyse lexicométrique (via le logiciel Lexico) et identification « manuelle » des sujets des articles ; premier essai de mesure du pluralisme offert sur l'échantillon du 23-24 septembre 2008.

Cette période a été ponctuée par deux réunions du sous-groupe « analyse de contenu des sites web », le 17 septembre et le 15 octobre 2008, à l'Insa de Lyon.

- Les mois de novembre et décembre ont été ceux de l'expérimentation en tant que telle.

Pour l'axe « analyse de contenu des sites web » : enregistrement en continu des flux RSS durant la période allant du 4 au 20 novembre 2008 ; analyse semi-automatisée des échantillons du 6 et du 10 novembre 2008 ; analyse quantitative du degré de pluralisme offert ; prolongements de l'analyse (analyse qualitative pour un sujet particulièrement redondant, suivi diachronique des contenus, représentation graphique des points de vue).

Une réunion de mise en commun et discussion des résultats de l'axe « analyse de contenu » a eu lieu le 11 décembre 2008 à l'Insa de Lyon (avec visioconférence à Toulouse 3).

Pour l'axe « socio-économie de l'information en ligne » : des enquêtes ont été menées auprès de quatre structures de production d'information en ligne, choisies notamment en raison de leur appartenance à des catégories de sites du corpus qui n'avaient pas été étudiées lors du programme de recherche précédent (webzine, site participatif, agrégateur de blogs). Les chercheurs de l'axe « socio-économie de l'information en ligne », moins nombreux que dans l'autre axe mais ayant davantage l'habitude de travailler ensemble, ont réalisé un guide d'entretien commun en amont, et mis en commun leurs résultats en aval, par le biais d'échange à distance (courrier électronique et téléphone).

- Une réunion bilan s'est tenue le 15 janvier 2009 à l'Insa de Lyon (avec visioconférence à Toulouse 3). Les résultats des deux axes ont été confrontés pour essayer d'en voir les articulations. Plus largement, des conclusions ont été tirées de cette année de recherche IPRI, et les opportunités pour la valoriser et la prolonger ont été décidées : propositions de publication, inscription dans un programme de recherche plus large (ANR).

3. Participants à la recherche

Le travail a reposé principalement sur la collaboration de trois laboratoires – Elico, Lerass, Liris - ainsi que de chercheurs associés de façon plus individuelle.

Initialement, une autre structure de recherche (Web Atlas) devait être partie prenante du projet, mais cela n'a pas été possible pour des raisons conjoncturelles d'incompatibilité de calendrier. Au total, ce sont quinze chercheurs qui ont participé à la recherche IPRI en 2008, de façon régulière ou plus ponctuelle.

Participation régulière

Caplat, Guy, Maître de conférences à l'Insa de Lyon, Informatique, LCI
Calabretto, Sylvie, Maître de conférences HDR à l'Insa de Lyon, Informatique, Liris
Damian-Gaillard, Béatrice, Maître de conférences à l'IUT de Lannion, Sciences de l'information et de la communication, Crape
Egyed-Zsigmond Elöd, Maître de conférences à l'Insa de Lyon, Informatique, Liris
George, Eric, Professeur adjoint à l'UQAM (Montréal), Sciences de l'information et de la communication, Gricis
Gesche, Loïc, Attaché temporaire d'enseignement et de recherche à l'Insa de Lyon, Informatique, Liris
Lallich-Boidin, Geneviève, Professeur à l'Université Lyon 1, Sciences de l'information et de la communication, Elico
Maisonasse, Loïc, Attaché temporaire d'enseignement et de recherche à l'Insa de Lyon, Informatique, Liris
Marty Emmanuel, Doctorant à l'Université Toulouse 3, Sciences de l'information et de la communication, Lerass
Rebillard, Franck, Maître de conférences à l'Université Lyon 2, Sciences de l'information et de la communication, Elico
Smyrnaio, Nikos, Maître de conférences à l'Université Toulouse 3, Sciences de l'information et de la communication, Lerass
Touboul, Annelise, Maître de conférences à l'Université Lyon 2, Sciences de l'information et de la communication, Elico

Participation ponctuelle

Lafouge, Thierry, Professeur à l'Université Lyon 1, Sciences de l'information et de la communication, Elico.
Ouakrat, Alan, Doctorant en Sciences de l'information et de la communication à l'Université Paris 2, IFP, et Cifre à Orange Labs, Sense.
Pouchot, Stéphanie, Maître de conférences à l'Université Lyon 1, Sciences de l'information et de la communication, Elico

II Travaux réalisés

Nous respectons ici, dans le compte-rendu de nos travaux, la partition entre les deux axes.

Le premier axe, relatif à l'analyse de contenu, a été bien plus développé car il constituait l'originalité même de cette recherche exploratoire. Le second axe tend davantage à actualiser et compléter les connaissances acquises au sujet des modalités socio-économiques de production de l'information en ligne. Celles-ci peuvent alors, *in fine*, entrer en résonance avec les résultats issus de l'analyse de contenu des sites web.

1. Axe « analyse de contenu des sites web »

Les travaux réalisés au sein de l'axe « analyse de contenu des sites web » sont allés bien au-delà de ce qui était imaginé au départ. Le développement rapide d'un prototype logiciel, apte à récupérer en temps réel les contenus diffusés sur les fils RSS d'actualité des sites, a offert un ensemble de données (présentation dans la partie 1.1) que les chercheurs ont pu exploiter dans des directions variées. D'une part, l'objectif central d'une quantification du pluralisme de l'information offert a été atteint (partie 1.2) et a même été affiné grâce à l'apport en compétences de spécialistes d'infométrie. D'autre part, des investigations complémentaires ont été développées (partie 1.3) : présentant un caractère moins abouti, plus prospectif, elles démontrent néanmoins l'effervescence de cette année de travail et les pistes qui s'ouvrent pour une étude de plus large ampleur.

1.1 Constitution des données

Avant de passer à la présentation des résultats de l'analyse des contenus des sites web, un préalable nécessaire est d'explicitier la méthode de constitution des données sur lesquelles se fonde cette analyse. Dans un premier temps, nous présenterons les choix qui ont été faits pour restreindre la population des sites web proposant des informations d'actualité à un corpus d'une centaine d'unités. Dans un second temps, nous préciserons quels contenus ont été extraits de ce corpus et comment ils sont venus alimenter la base de données dans laquelle les chercheurs ont puisé par la suite.

1.1.1 Délimitation du corpus de sites web

Chercheurs impliqués¹ : Franck Rebillard, Nikos Smyrniaios, Eric George, Annelise Touboul.

La population de sites visée par cette recherche est celle des sites proposant de l'information d'actualité et/ou journalistique. Il n'existe pas de définition unanime et consensuelle de l'information d'actualité et/ou journalistique, en particulier parce que le qualificatif journalistique est l'objet d'une captation symbolique et statutaire par le corps professionnel des journalistes qui ne « reconnaît » pas certains des supports de ces informations d'actualité (Ruellan, 2007). Dans le cadre de cette recherche, et *a fortiori* parce que l'internet constitue un support relativement récent

¹ Au début de chaque section détaillant une partie du travail, nous faisons apparaître les noms des chercheurs plus particulièrement impliqués dans sa réalisation. Ce qui n'enlève rien, bien entendu, à la dimension collective de l'ensemble du projet.

et susceptible d'accueillir des contenus ne rentrant pas forcément dans les canons traditionnels de l'information journalistique, nous avons privilégié une conception extensive de l'information d'actualité. Sa définition (très) synthétique serait : description-représentation de faits relativement récents, caractérisée par sa périodicité régulière (ex : différence entre un journal et un livre) et la distinction entre l'instance éditoriale et les instances médiatisées (ex : différence avec un journal municipal ou un journal d'entreprise).

Une telle définition rejoint celle employée par la Direction du Développement des Médias (DDM) pour délimiter le périmètre de la « *presse éditeur* », par opposition aux « *autres types de presse (presse de groupement, presse administrative, journaux d'entreprise, lettres confidentielles)* » (<http://ddm.gouv.fr>).

Une fois cette définition posée, reste à repérer l'ensemble des sites web qui proposent de l'information d'actualité, des sites francophones et non uniquement français afin de conserver le caractère international de l'internet dans notre analyse. Pour ce faire, faute de ne pouvoir hélas compter sur l'aide de Web Atlas en matière de cartographie du web, nous avons établi un recensement à partir de répertoires spécialisés suivants : bases de Google Actualités, de Wikio, de Rezo.net et du site professionnel IPLJ (« Internet pour les journalistes »).

Ces répertoires présentent l'avantage d'être variés (certains renvoient à des sites d'information « grand public », d'autres sont davantage axés sur les blogs ou les webzines d'opinion) et d'offrir une profondeur intéressante (jusqu'à 550 sites répertoriés).

Dans chacun de ces répertoires, les chercheurs du groupe IPRI les plus experts en matière d'information d'actualité sur le web (en France et à l'étranger, Québec mais aussi pays francophones d'Afrique ou d'Europe) ont repéré les sites correspondant à la définition de l'information d'actualité retenue pour cette recherche, en aboutissant toujours à une population de plusieurs centaines d'unités.

Ce volume était trop considérable pour une analyse de contenu, quand bien même fût-elle en partie automatisée. Il a été donc décidé de la restreindre à un corpus plus viable d'une centaine de sites d'information d'actualité, selon les critères suivants correspondant aux visées de la recherche IPRI :

- Tout d'abord, nous avons limité le corpus aux sites proposant de l'information politique et générale, type d'information le plus central aux regard des problématiques de pluralisme de l'information.

Les publications d'information générale et politique sont désignées par la DDM comme « *apportant de façon permanente informations et commentaires sur l'actualité politique* », par opposition aux publications d'information spécialisée grand public « *traitant d'un thème principal plus ou moins spécialisé (le cyclisme ou le sport en général, la mode...) dont le contenu est accessible à tous publics* », et aux publications d'information technique et professionnelle « *dont la teneur des articles est trop technique pour susciter l'intérêt au-delà des professions correspondant aux sujets traités.* » (<http://ddm.gouv.fr>).

Ainsi, ont été écartés de notre corpus les nombreux sites thématiques, spécialisés notamment dans l'actualité des nouvelles technologies ou l'actualité sportive.

- A l'issue de ce premier « écrémage », nous avons fait en sorte de nous assurer de la représentativité de ces sites web au regard de l'ensemble des modalités de publication sur le web. Car le processus de publication sur le web, ainsi que l'a montré un programme de recherche précédent dédié de façon plus large aux contenus informationnels et culturels (Chartron, Rebillard, 2007), ne se limite pas au modèle

classique de diffusion mass-médiatique : il comprend aussi le registre de l'autopublication (ex : blogs), de la publication distribuée (ex : *peer-to-peer*), et le niveau méta-éditorial (ex : agrégateurs). Pour les contenus journalistiques, cela revient à considérer que la logique de l'*editorial content* voisine avec celle de la *public connectivity* comme l'a exprimé Mark Deuze dans un article de référence sur le sujet (2003) où il distinguait les catégories de *mainstream news sites, index and category sites, meta and comment sites, share and discussion sites*.

Cette typologie est reprise dans de nombreux travaux anglophones se penchant sur le journalisme en ligne mais sans jamais vraiment entraîner son analyse dans la globalité. Y compris dans les travaux les plus récents qui, devant pourtant prendre acte de cette existence plurielle du journalisme en ligne, se limitent au mieux à analyser les interactions entre deux catégories d'espaces de publication d'informations d'actualité : par exemple entre les sites de médias traditionnels et leurs pourvoyeurs d'audience transnationaux que sont les agrégateurs (agrégateurs « manuels » comme le *Drudge Report* ou « automatisés » comme *Google News*, in Thurman, 2007) ; entre blogs amateurs et sites professionnels, malgré l'annonce d'une analyse de la « *global news arena* » (Reese *et al.*, 2008).

Parce que s'intéresser au pluralisme de l'information sur l'internet implique d'intégrer les particularités de la publication de contenus sur le web, en comparaison des supports antérieurs comme l'imprimé ou l'audiovisuel, nous nous sommes efforcés de poursuivre le travail de défrichage initié par Deuze, jusqu'à aboutir à une typologie complète et actualisée des différents espaces web de publication d'information d'actualité :

1. versions internet de médias existants (presse en ligne)
2. versions internet d'agences de presse (agences de presse en ligne)
3. publications individuelles exclusivement internet (blogs)
4. publications collectives exclusivement internet (webzines)
5. publications « collaboratives » exclusivement internet (sites de journalisme participatif)
6. composantes informationnelles de plateformes multiservices (rubriques *Actualités* des portails)
7. regroupements automatisés d'informations d'actualité (agrégateurs de nouvelles)

En accord avec la problématique de cette recherche, nous avons ainsi fait en sorte que le corpus retenu contienne des sites représentant tout l'éventail des catégories énumérées ci-dessus. Pour chacune de ces catégories, nous avons retenu au moins 5 sites dans le corpus. Ce qui permet d'aboutir à un premier sous total de (7 catégories x 5 sites) = 35 sites. Les sites restants ont été distribués dans chacune des catégories en fonction de leur volume sur le web (ex : les sites des catégories 1. presse en ligne, 3. blogs, et 4. webzines, sont par exemple plus nombreux dans notre corpus), jusqu'à parvenir à un total de 103 sites.

1.1.2 Extraction des contenus

Chercheurs impliqués : Elöd Egyed-Zsigmond, Loïc Maisonnasse, Franck Rebillard.

De cette première liste d'une centaine de sites, n'ont pu finalement être exploités que 89 d'entre eux, ceux qui présentaient un flux RSS « A la Une » ou « Actualités », afin de répondre à deux exigences :

- La limitation aux sites disposant d'un flux RSS répond à une exigence technique. L'application informatique d'extraction des contenus, développée spécifiquement pour les besoins de la présente recherche, repose en effet sur une « aspiration » (*crawling*), en temps réel, des éléments composant les fils RSS des différents sites du corpus. Ont ainsi pu être collectés, pour chaque article publié sur un fil RSS, son titre, sa description (équivalent du "chapô" dans la presse écrite), le lien vers son texte intégral ainsi que la date et l'horaire de sa mise en ligne. La base de données ainsi constituée comprend l'ensemble de ces champs ainsi que deux entrées supplémentaires liées à l'implantation d'un système de lemmatisation (identification des « mots-racines » employés) : lemmes du titre et lemmes de la description de l'article. Cette lemmatisation s'est avérée utile pour augmenter la dimension automatisée de l'exploitation de la base de données et ainsi pouvoir procéder à un suivi diachronique sur l'ensemble des données collectées (partie 1.3.3). Elle s'avère par ailleurs complémentaire de la méthode d'analyse de contenu semi-automatisée plus approfondie sur certains échantillons de données (partie 1.2.1).

- Le fait de se focaliser sur les seuls fils RSS « A la Une » ou « Actualités », plutôt que d'enregistrer l'ensemble des fils RSS, satisfait elle une exigence de cohérence par rapport à la problématique de recherche. Comme nous l'avons déjà explicité dans la partie précédente, la problématique du pluralisme de l'information se cristallise essentiellement autour de l'information politique et générale, et c'est ce type de contenus que l'on retrouve dans les fils « A la Une » ou « Actualités ». Par ailleurs, ces deux fils présentent l'avantage de reprendre en grande partie les informations générales publiées en page d'accueil du site.

Cette limitation aux fils RSS « Actualités » ou « A la Une » a conduit à exclure du corpus certains des sites ne présentant pas ces dispositions : Ouest-France, Aujourd'hui Le Maroc, Vox Populi, Le Grand Soir, Le Monde Citoyen, Oulala.net, Samizdat.net, Daily Motion (chaîne Actu), et Net Vibes.

Plus spécifiquement, ce procédé aura dans la pratique montré d'autres limites concernant deux catégories de sites n'ayant pas (ou peu) de flux RSS : les agences de presse et les agrégateurs. Pour la première catégorie, cet obstacle a pu être anticipé et contourné en relevant « à la main », pour chaque journée d'observation plus poussée (voir partie 1.2), les titres des dépêches publiées dans les catégories « A la Une » et « Actualités » des sites de AFP, Reuters, Xinhuan, et La Presse Canadienne. Pour l'autre catégorie concernée, celles des agrégateurs, un problème technique d'enregistrement des flux a malheureusement révélé *a posteriori* la mise à l'écart involontaire d'un site majeur comme Google News, alors que son alter ego Wikio a lui vu ses flux correctement captés par notre application informatique.

Malgré ces quelques limites, nous estimons avoir pu travailler sur un ensemble de données satisfaisant par sa taille et sa globalité, puisqu'il rassemble plusieurs dizaines de milliers d'articles, publiés par environ 90 sources d'informations, couvrant toutes les catégories de sites. (voir le détail du corpus final en annexes)

Il s'agit là d'une ampleur rarement atteinte dans les analyses de contenu des médias : au niveau des échantillons jour par jour, c'est ainsi la production quotidienne d'environ soixante sources d'information qui a été rendue disponible pour l'observation, à comparer avec la taille standard des corpus ordinairement mobilisés dans les études sur les seuls grands quotidiens nationaux ou sur les seules chaînes nationales hertziennes, dépassant rarement la dizaine de sources.

1.2 Analyse quantitative du pluralisme offert

Dans cette partie, nous présentons les résultats de l'analyse qui était le plus au cœur de ce projet de recherche. Elle touche en effet à une mesure quantitative de la variété et de la répartition des contenus d'actualité publiés sur le web. Celle-ci a porté de façon approfondie sur deux échantillons de la base de données précédemment constituée, correspondant à deux journées :

- échantillon du 6 novembre 2008 : 2617 articles issus de 61 sources ;
- échantillon du 10 novembre 2008 : 2040 articles issus de 60 sources².

Début novembre 2008 coïncidait avec un événement politique – et médiatique – majeur : l'élection de Barack Obama à la Présidence des Etats-Unis d'Amérique. Le scrutin a eu lieu le 4 novembre et ses résultats n'ont été connus qu'à partir du 5 novembre en France. Le fait de s'attacher à deux échantillons, l'un proche temporellement de cet événement, et l'autre plus éloigné, visait à analyser autant une journée d'actualité « chaude » qu'une journée plus ordinaire sur le plan journalistique.

1.2.1 Méthode semi-automatisée d'analyse de contenu

Chercheurs impliqués : Emmanuel Marty, Nikos Smyrnaio.

Une analyse de contenu semi-automatisée des titres³ a d'abord été déployée. Elle vise à saisir la quantité de sujets abordés au cours d'une journée, et surtout l'importance relative de chacun d'entre eux au sein de l'agenda médiatique numérique.

La classification des articles en sujets est une étape déterminante vis-à-vis des résultats obtenus. Pour répondre à l'objectif de mesure quantitative du pluralisme de l'information sur l'internet, les articles collectés ont en effet été regroupés par sujets, sur la base de leurs titres respectifs. La méthode employée pour parvenir à ces classifications est semi-automatisée et inductive. Fondée au départ sur une analyse lexicométrique (automatisée) des titres qui permet d'opérer un premier défrichage, cette méthode repose conjointement sur une identification par les chercheurs des sujets constituant l'actualité du jour au sein de chaque échantillon.

Par *sujet*, nous désignons une expérience factualisée, telle que définie par Jean-Pierre Esquenazi (2002) : la « factualisation de l'expérience » revient « dans le temps et l'espace, à délimiter, circonscrire le fait en le présentant comme une discontinuité ». En prenant appui, avec Esquenazi, sur la notion de cadre d'Erving Goffman (1991), on considère le sujet d'actualité comme une « expérience cadrée », résultant de l'application du cadre primaire de la perception. La rédaction de l'article de presse, en tant que production discursive médiatique, consiste à appliquer à cette réalité perçue et cadrée un cadre médiatique second, un « recadrage » (Esquenazi, op. cit.). Notre conception du sujet n'est pas relative à ce cadre médiatique second mais au cadre primaire du sujet d'actualité en amont de son traitement médiatique, c'est-à-dire *avant la situation de rédaction*. En effet, comme le soulignent Roselyne Ringoot et Yvon Rochard (2005), le « rapport au terrain va déterminer la pratique

² Le fait que la journée du 10 novembre compte 60 sources et non plus 61 est lié à la méthode d'aspiration utilisée, ne retenant que les sources qui avaient publié de manière effective dans les 24 heures concernées.

³ Pour lever toute ambiguïté terminologique, précisons que l'emploi du mot « titre » fera ici toujours référence à la titraile de premier niveau, qui constitue notre matériau d'analyse, et non aux supports ni aux acteurs de cette information, qui seront toujours désignés respectivement par les mots « sites » et « sources ». Une confusion est en effet possible avec la notion de titre de presse, communément employée, désignant la source du message.

professionnelle du journaliste en articulant la perception du réel en situation de recherche d'informations et le processus d'écriture en situation de rédaction.» Pour résumer notre démarche de regroupement des articles en sujets, nous nous sommes attachés à l'identification du sujet d'actualité avant qu'il ne devienne un sujet journalistique, auquel on a apposé « angles, format, genres, positionnement dans le journal » (Ringoot et Rochard, op.cit.).

C'est sur la base de cette définition, tentant d'objectiver les différents sujets d'actualité et d'aboutir à un niveau de classification homogène pour chacun d'entre eux, que ces derniers ont été progressivement inventoriés, à l'issue de plusieurs étapes.

L'étape automatisée, tout d'abord, a été réalisée à l'aide du logiciel d'analyse des données textuelles Lexico⁴. Une des fonctionnalités de ce logiciel est le comptage des segments répétés. Il s'agit de repérer les groupes de mots très fréquents dans un corpus. A cette étape de l'analyse, cet outil nous permet de repérer de manière chiffrée les titres identiques (ou présentant une forte proximité lexicale) les plus présents au sein chaque échantillon. Ce premier tri automatique ne permet pas simplement de regrouper les articles dont les titres sont similaires ; il permet aussi de faire émerger un certain nombre de sujets particulièrement récurrents.

Dans un second temps, sont pris en considération les titres plus originaux sur le plan lexical et donc non repérés par l'outil de lexicométrie. Ceux-ci sont analysés un à un par les chercheurs : certains sont rattachés aux sujets déjà mis en lumière dans l'étape précédente ; les autres titres conduisent à l'identification de nouveaux sujets, au sein desquels ils sont regroupés.

Dans un dernier temps, après constitution définitive des sujets, des codes sont affectés à chacun d'eux selon une numérotation classique. L'indexation des titres consiste donc en l'attribution à chacun d'eux d'un code sujet allant de 1 à n. Après traitement, chacun des deux échantillons (6 novembre et 10 novembre) se présente donc sous la forme d'un tableau qui comporte autant de lignes que d'items récoltés automatiquement pendant la journée en question ; et trois colonnes : titre de l'article, code sujet, nom de la source. Ces bases de titres indexés sont alors utilisées pour les calculs de variété et d'équilibre des sujets.

Pour obtenir une vision plus synthétique de la production observée, nous avons procédé en parallèle à un regroupement des sujets en thématiques, qui constitue le niveau supérieur de généralité dans l'analyse de contenu ainsi effectuée.

Les thématiques ont été créées *a posteriori* à partir des sujets, qui sont alors regroupés pour leur cohérence. L'objectif est de fournir un niveau d'analyse supplémentaire, notamment avec des volumes plus conséquents. Certaines thématiques peuvent apparaître comme relevant du rubriquage (ex : « Faits divers ») mais la démarche est plutôt de regrouper des champs de connaissance généraux (« Actualités littéraires et culturelles » ; « Diplomatie internationale-conflits armés ») à la manière des classifications mises en place dans les bibliothèques ou l'univers de la documentation.

⁴ Le logiciel effectue tout d'abord un comptage des occurrences de formes ou groupes de formes similaires, puis construit un tableau des formes récurrentes avec leur fréquence d'apparition dans le corpus. Pour aller plus loin, voir Lebart et Salem (1994).

1.2.2 Calculs de variété et d'équilibre

Chercheurs impliqués : Franck Rebillard, Guy Caplat.

Pour réaliser une mesure chiffrée du pluralisme des sujets traités au sein de nos échantillons de données, nous avons appliqué au cas de l'information journalistique des formules déjà expérimentées pour calculer le degré de diversité d'autres types de contenus culturels et médiatiques, comme cela a été fait par exemple dans le secteur du livre (Benhamou, Peltier, 2006). Notre étude reprend donc des axes conceptuels déjà éprouvés, à travers les notions de variété et d'équilibre de l'information. La variété correspond au nombre de sujets au sein de l'échantillon considéré, l'équilibre est pour sa part déterminé par la répartition des sujets entre les différents articles⁵.

- Échantillon du 6 novembre 2008 : Sur un total de 2617 articles, 385 sujets différents ont été identifiés. On observe ainsi une moyenne de 7 articles par sujet, avec d'immenses écarts à la moyenne et une médiane de 1 article par sujet. Dans le détail, alors que les cinq sujets les plus fréquents regroupent chacun plus de 100 articles (voir tableau 1), 301 des 385 sujets identifiés regroupent moins de 5 articles chacun.

Tableau 1 : Sujets les plus fréquemment abordés lors de la journée du 6 novembre 2008

Sujet	Nb articles	% échantillon
L'élection américaine de Barack Obama	435	16,62%
Le congrès du PS et son actualité	187	7,15%
Les institutions internationales face à la crise financière	178	6,80%
L'actualité boursière	121	4,62%
Grève à la SNCF	120	4,59%

- Échantillon du 10 novembre 2008 : Sur un total de 2040 articles, 309 sujets différents ont été identifiés. On observe là encore une moyenne de 7 articles par sujet, et les écarts à la moyenne restent colossaux, bien qu'un peu plus lissés que dans l'échantillon du 6 novembre, en particulier parce qu'il n'y a plus de sujet extrêmement dominant comme pouvait l'être l'élection de Barack Obama. La médiane est ici aussi proche de 1 : en fait 169 sujets sur 309 correspondent à un seul article, inclus dans un total de 236 sujets à comptabiliser moins de 5 articles chacun. Les cinq sujets les plus fréquents totalisent chacun plus de 80 articles. Ce nombre, bien qu'inférieur à celui du 6 novembre, reste néanmoins élevé.

Tableau 2 : Sujets les plus fréquemment abordés lors de la journée du 10 novembre 2008

Sujet	Nb articles	% échantillon
La question de la gouvernance du PS après le congrès	129	6,32%
Obama rencontre Bush à la Maison Blanche	114	5,59%
Nouvel acte de malveillance contre la SNCF	112	5,49%
Le prix Goncourt décerné à l'auteur franco-afghan A. Rahimi	108	5,29%
La situation en République Démocratique du Congo	83	4,07%

⁵ En retenant ces notions de *variété* et d'*équilibre*, nous avons bien conscience de gagner en vision panoramique et relativement exhaustive du pluralisme ce que nous perdons en finesse d'analyse. Voilà pourquoi nous avons attribué à la troisième notion de *disparité* (différence de traitement à l'intérieur de chaque sujet) élaborée dans les travaux susmentionnés, un caractère beaucoup plus qualitatif : cette dimension qualitative est explorée via une analyse de discours des articles les plus redondants, autre étape de cette recherche IPRI (voir partie 1.3.1), complément indispensable selon nous à l'apport quantitatif des calculs de *variété* et d'*équilibre*.

Les deux échantillons présentent des caractéristiques très voisines : une grande variété de sujets abordés, mais très inégalement représentés. Quelques sujets agglomèrent des quantités conséquentes d'articles (plusieurs centaines), tandis que la majorité des sujets ne connaît de traitement que dans un seul article. La production de l'information telle que nous l'avons mesurée sur le web affiche donc une structure relativement constante, partagée entre une concentration extrême des articles sur un nombre réduit de sujets et une dispersion très forte des articles restants sur presque autant de sujets isolés.

Une concentration extrême sur quelques sujets dominants.

Malgré le poids médiatique exceptionnel pris le 6 novembre par le sujet de l'élection de Barack Obama (environ 17% des sujets), en raison de la proximité temporelle avec le scrutin présidentiel aux Etats-Unis, on observe une structure très semblable de la distribution des sujets au sein des deux échantillons.

Les articles se concentrent dans les deux premiers déciles de sujets de l'échantillon du 6 novembre 2008 : 20% des sujets rassemblent 82% des articles (en sachant que les seuls 10% des sujets les plus traités rassemblent 71% des articles).

La concentration est quasi-identique dans l'échantillon du 10 novembre 2008: 20% des sujets rassemblent 81% des articles (en sachant que les seuls 10% des sujets les plus traités rassemblent 69% des articles).

Une forte dispersion en de multiples sujets isolés.

La concentration est ainsi très forte sur les sujets les plus mis en avant sur la scène médiatique, fut-elle numérique. Parallèlement, à l'autre extrémité de la courbe, apparaît un phénomène inverse de très forte dispersion. Mais si les sujets originaux repérés sur la toile sont nombreux, ils sont loin d'entraîner un engouement médiatique, en comparaison des sujets qui s'imposent à la une de l'actualité. Ainsi, pour l'échantillon du 6 novembre, 50% des sujets agrègent 93% des articles tandis que l'autre moitié n'est présente que dans 7% des articles de l'échantillon.

Les chiffres sont quasiment identiques pour l'échantillon du 10 novembre : 50% des sujets agrègent 92% des articles tandis que l'autre moitié n'est présente que dans 8% des articles de l'échantillon.

Analyse au niveau des thématiques

- Échantillon du 6 novembre 2008

Variété : 34 thématiques

Distribution : globalement, la répartition des articles par thématiques est plus équilibrée que par sujets (en raison du caractère moins éclaté de la catégorisation en thématiques), mais les écarts à la moyenne de 77 restent énormes. Seules 10 thématiques sur 34 atteignent ce seuil et lorsque c'est le cas, elles le dépassent largement pour la plupart.

Tableau 3 : Thématiques les plus fréquentes lors de la journée du 6 novembre 2008

Thématique	Nb articles	% échantillon
PIAM-Politique intérieure américaine	448	17,12%
VPF-Vie politique française	252	9,63%
RECF-Répercussions économiques de la crise financière	240	9,17%
AEE-Actualités économiques et des entreprises	229	8,75%
DICA-Diplomatie internationale / Conflits armés	210	8,02%
AS-Actualités sportives	179	6,84%
GPSR-Grèves, plans sociaux, restructurations	140	5,35%
FD-Faits divers	132	5,04%
TAE-Terrorisme : attentats et enlèvements	98	3,74%
PIC-Politique intérieure canadienne	77	2,94%

- Échantillon du 10 novembre 2008

Variété : 34 thématiques.

Variété strictement égale en nombre à celle constatée le 6 novembre. Toutefois, quelques thématiques diffèrent entre le 6 et le 10 novembre (disparition des thématiques AC-Aménagement/collectivités et PISA-Politique intérieure sud-américaine ; apparition des thématiques ESD-Exploration spatiale, découverte et PIA-Politique intérieure des Antilles).

Distribution :

Répartition très déséquilibrée, mais un peu moins que dans l'échantillon du 6 novembre : aucune thématique n'est ultra-dominante (comme pouvait l'être PIAM précédemment), les écarts supérieurs à la moyenne restent importants en concernant une thématique de plus (11 thématiques sur 34 au lieu de 10 sur 34 le 6 novembre).

Tableau 4 : Thématiques les plus fréquentes lors de la journée du 10 novembre 2008

Thématique	Nb articles	Pourcentage
TAE-Terrorisme : attentats et enlèvements	239	11,72%
DICA-Diplomatie internationale / Conflits armés	223	10,93%
AS-Actualités sportives	200	9,80%
VPF-Vie politique française	192	9,41%
ALC-Actualités littéraires et culturelles	157	7,70%
AEE-Actualités économiques et des entreprises	150	7,35%
PIAM-Politique intérieure américaine	137	6,72%
FD-Faits divers	118	5,78%
ARTCM-Actualités radio-télé-ciné-musique	78	3,82%
RECF-Répercussions économiques de la crise financière	70	3,43%
RCI-Rubriques, chroniques et inclassables	68	3,33%

Analyse recentrée au niveau national : une structure identique à celle de l'international.

La très forte dispersion constatée pourrait s'expliquer par la délimitation du corpus présidant à l'échantillonnage. En choisissant de faire figurer des sites étrangers francophones pour conserver la dimension internationale de l'internet dans notre

analyse, nous aurions pu créer un émiettement artificiel des sujets, via la collecte d'articles concernant un sujet très spécifique à un pays (ex : élections provinciales dans les sites québécois).

Une réduction de l'échantillon aux seuls sites français a donc été effectuée pour tester l'existence d'un tel biais. Elle confirme la permanence de la structure déjà observée, celle d'une grande variété de sujets inégalement répartis entre concentration extrême et forte dispersion. En effet, sur chacune des deux journées, les cinq premiers sujets traités sont les mêmes et les ordres de grandeur sont très comparables : la médiane reste à 1 et 20% des sujets continuent à rassembler environ 80% des articles.

1.2.3 Approfondissements autour de l'hypothèse de « longue traîne »

Chercheurs impliqués : Thierry Lafouge, Franck Rebillard, Stéphanie Pouchot

Cette première analyse quantitative a permis d'offrir, à propos d'échantillons de données conséquents (plus de 2000 articles par jour collectés sur une soixantaine de sites) une représentation chiffrée de la production des informations d'actualité sur internet. Celle-ci apparaît donc à la fois comme variée (plus de 300 sujets) et comme très inégalement répartie : aux côtés de quelques sujets omniprésents, quantité de sujets n'émergent que par la grâce d'un article ou d'une poignée d'articles. Si bien qu'un grand pluralisme (forte variété des sujets) voisine sur la toile avec une redondance prononcée de l'information (concentration extrême sur les sujets dominants).

La représentation graphique de la distribution des sujets fait apparaître une structure très inégalitaire, proche de la *distribution de Pareto* suivant la règle classique dite des 20/80 (20% des éléments d'une population s'arrogent 80% des ressources).

La forme prise par la courbe des résultats IPRI évoque par ailleurs une analogie avec les représentations graphiques de la « longue traîne », hypothèse très en vogue pour caractériser la diffusion des biens culturels et médiatiques sur l'internet, que nous proposons de discuter ici.

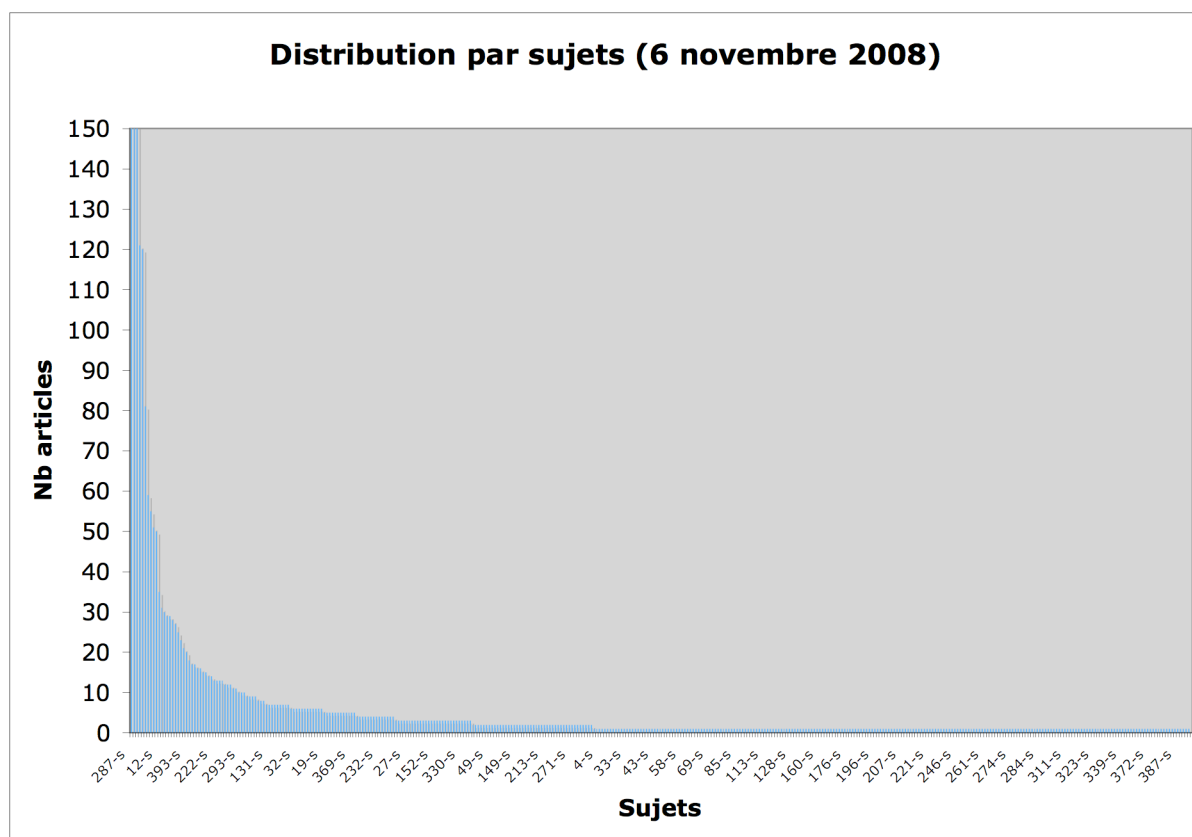


Figure 1

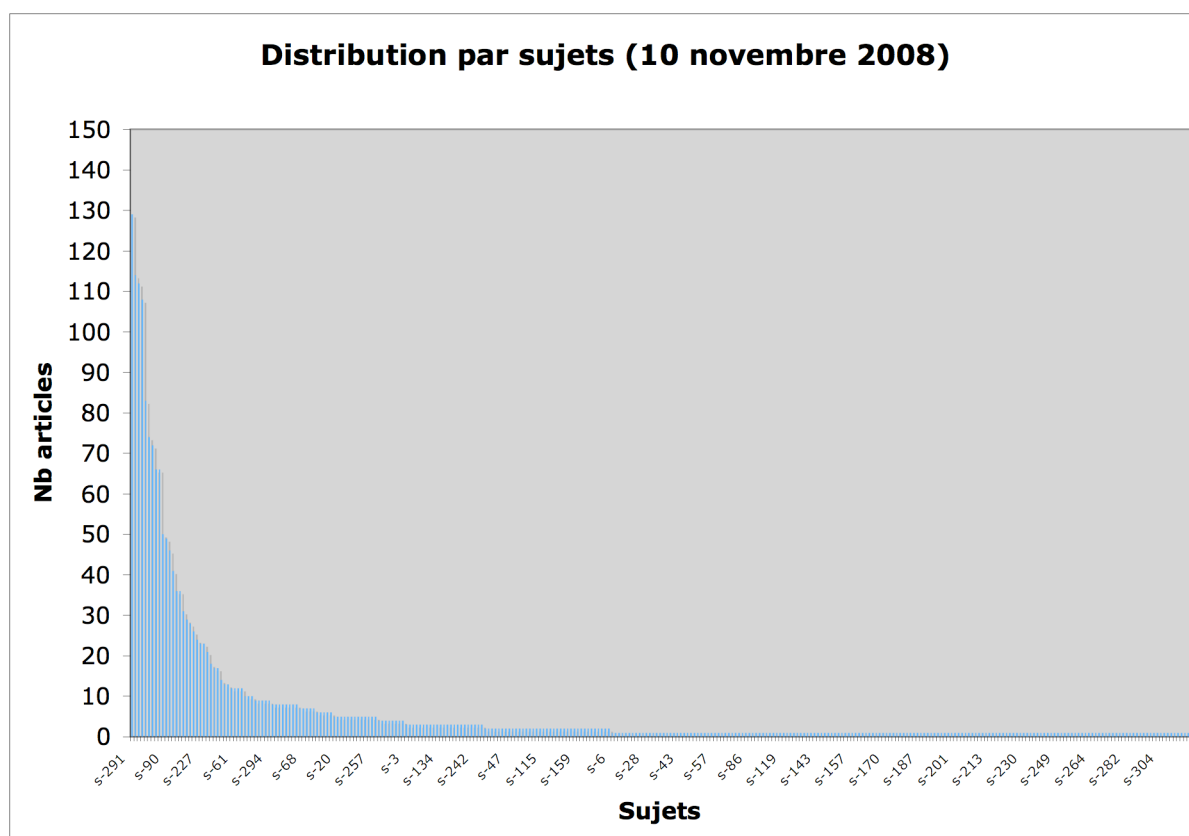


Figure 2

L'hypothèse de « longue traîne » est celle d'une moindre concentration et d'une répartition plus étalée de la consommation des biens culturels sur l'internet, par rapport au commerce traditionnel (limité dans ses possibilités logistiques de stockage). Formulée dans un essai il y a quelques années (Anderson, 2006), cette hypothèse n'a toujours pas trouvé de véritable confirmation empirique (Benghozi et Benhamou, 2008). Elle reste pourtant très présente et continue d'alimenter l'espoir d'un succès plus grand des produits de niche ou extrêmement rares sur l'internet. Or, pour les produits vidéo par exemple, il a été démontré que les films de « fond de catalogue », œuvres anciennes ou artistiquement plus exigeantes que les succès habituels du box-office, ne trouvaient pas forcément plus d'amateurs en proportion sur le web (Elberse, Oberholzer-Gee, 2007) : pour que la théorie soit vérifiée, il faut que la traîne soit non seulement longue, mais également « épaisse ». Mais dans le cas de l'étude susmentionnée, il apparaît que les visionnages de « petits » films, même agrégés dans un même ensemble, sont très loin d'atteindre le nombre de visionnages des blockbusters.

Dans un même esprit, nous avons souhaité confronter les résultats issus de cette recherche IPRI avec les essais de théorisation de la « longue traîne ». Plus encore, sur les conseils de Jean-Michel Salaün qui avait déjà réfléchi à cette question pour la production de documents numériques, et grâce à l'expertise de collègues d'Elico spécialistes d'infométrie, nous avons pris pour cadre théorique la comparaison avec certains modèles et lois de puissance inverse autres que celui de Pareto, et davantage en lien avec la production de documents : loi de Zipf en lexicométrie et loi de Lotka en scientométrie. Plus largement, cette mise en parallèle amène à mieux cerner ce que pourrait être l'originalité de la « longue traîne » pour la production web de documents et voir si les résultats IPRI en fournissent ou non une illustration.

Comparaison IPRI / Loi de Zipf

La loi de Zipf établit que la distribution des occurrences verbales au sein d'un texte est constante. Établie au milieu du XXe siècle, cette loi lexicométrique a pu être vérifiée pour de nombreux textes et a même été appliquée, par extrapolation, dans d'autres domaines comme par exemple les requêtes sur les moteurs de recherche. Plus largement, elle a donné lieu à des représentations graphiques de la relation entre fréquence et rang d'apparition d'un item, sous forme de courbes hyperboliques précisément appelées « zipfiennes ».

Parce que la représentation des résultats IPRI ressemblait à ce type de courbe, il a paru intéressant de la confronter de façon plus précise à celle de Zipf.

La mise en correspondance de la courbe IPRI pour l'échantillon du 6 novembre 2008 et la courbe théorique de Zipf (puissance inverse) est livrée dans le schéma qui suit.

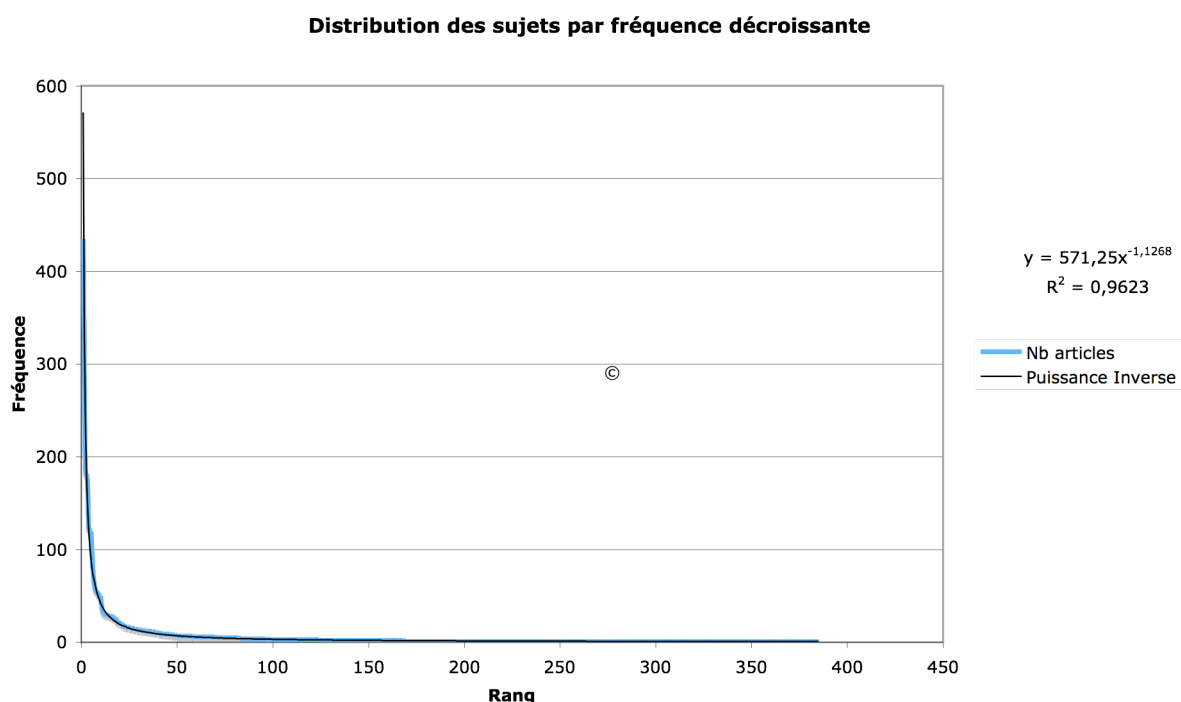


Figure 3

On peut observer une très forte homologie entre les deux courbes.

Cette proximité IPRI/Zipf apparaît encore plus nettement dans le cadre d'une représentation graphique sous échelle logarithmique, selon les différents « paliers » de fréquence des sujets traités (voir schéma suivant).

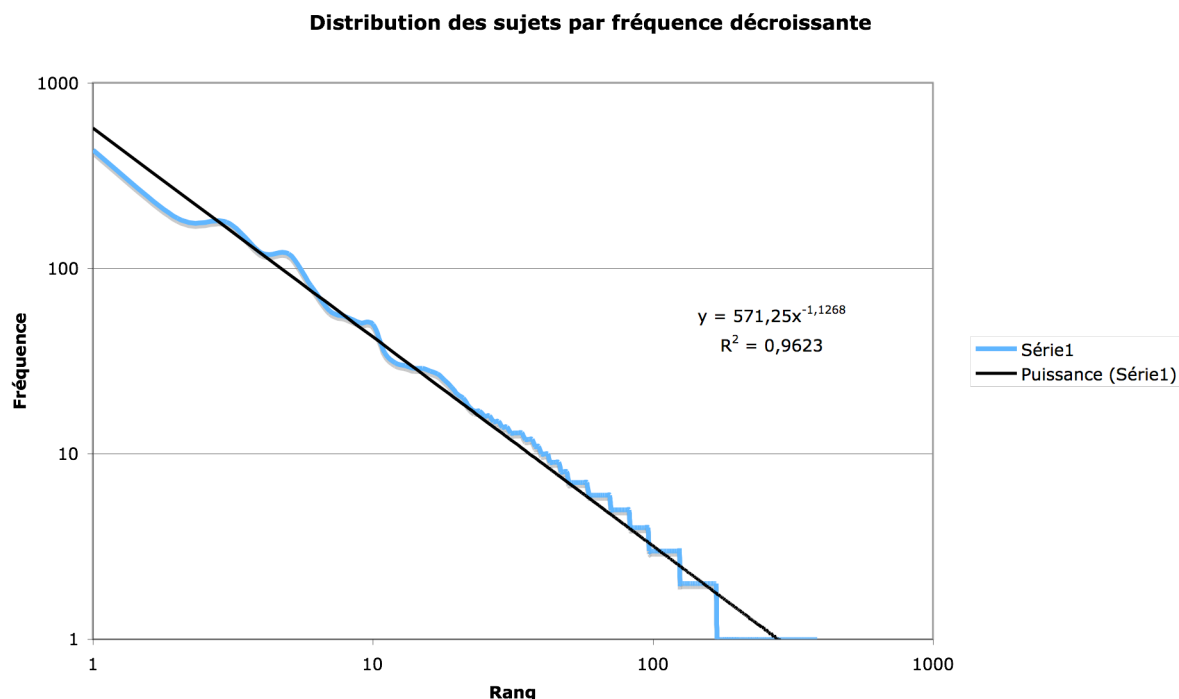


Figure 4

Ici, on voit bien à quel point la représentation de la distribution des sujets IPRI épouse de très près la droite représentant elle la fonction puissance en échelle logarithmique.

Sur le plan mathématique, ceci est confirmé par le fait que le carré de la valeur R est très proche de 1. Dans le cas étudié ici de l'échantillon du 6 novembre, il est de 0,9623. En réduisant l'échantillon aux seuls sites français, on aboutit à 0,9698. Et pour l'échantillon du 10 novembre : 0,9541 pour l'ensemble des sites et 0,9486 pour les seuls sites français.

Comparaison IPRI / Loi de Lotka

La loi de Zipf s'intéresse aux modalités de distribution de différents items au sein d'un corpus textuel. La loi de Lotka fait de même mais en regroupant chaque item dans des classes d'items possédant la même fréquence d'apparition.

Élaborée au début du XXe siècle, la loi de Lotka portait sur la production scientifique des chercheurs, et constatait de forts écarts entre un faible nombre d'auteurs de très nombreux articles et à l'inverse une grande majorité de chercheurs publiant en quantité limitée. Cette inégalité dans la productivité scientifique s'est là encore retrouvée confirmée au fil du temps. Surtout, on peut y distinguer une distribution hyperbolique des contenus non plus considérés individuellement mais collectivement, c'est-à-dire au sein d'effectifs de contenus possédant la même fréquence d'apparition.

Concrètement, cela nous amène pour le cas IPRI à regrouper les sujets selon leur niveau de fréquence d'apparition. Pour l'échantillon du 6 novembre, ont ainsi été regroupés les sujets selon les niveaux suivants, depuis les 217 sujets traités isolément

(dans un seul article) jusqu'aux sujets donnant lieu à plusieurs articles (en s'arrêtant au seuil de 18 articles, à partir duquel les niveaux de fréquence deviennent trop discontinus pour une représentation graphique satisfaisante).

Tableau 5 : base de calcul utilisée pour la représentation graphique « type Lotka »

Fréquence	Effectifs
1	217
2	44
3	28
4	14
5	12
6	12
7	9
8	3
9	4
10	3
11	2
12	3
13	4
14	2
15	2
16	2
17	2
18	1

La courbe hyperbolique tirée de cette classification des sujets selon leur niveau de fréquence d'apparition entraîne la représentation hyperbolique suivante, que nous mettons en parallèle avec la courbe théorique de puissance inverse de Lotka.

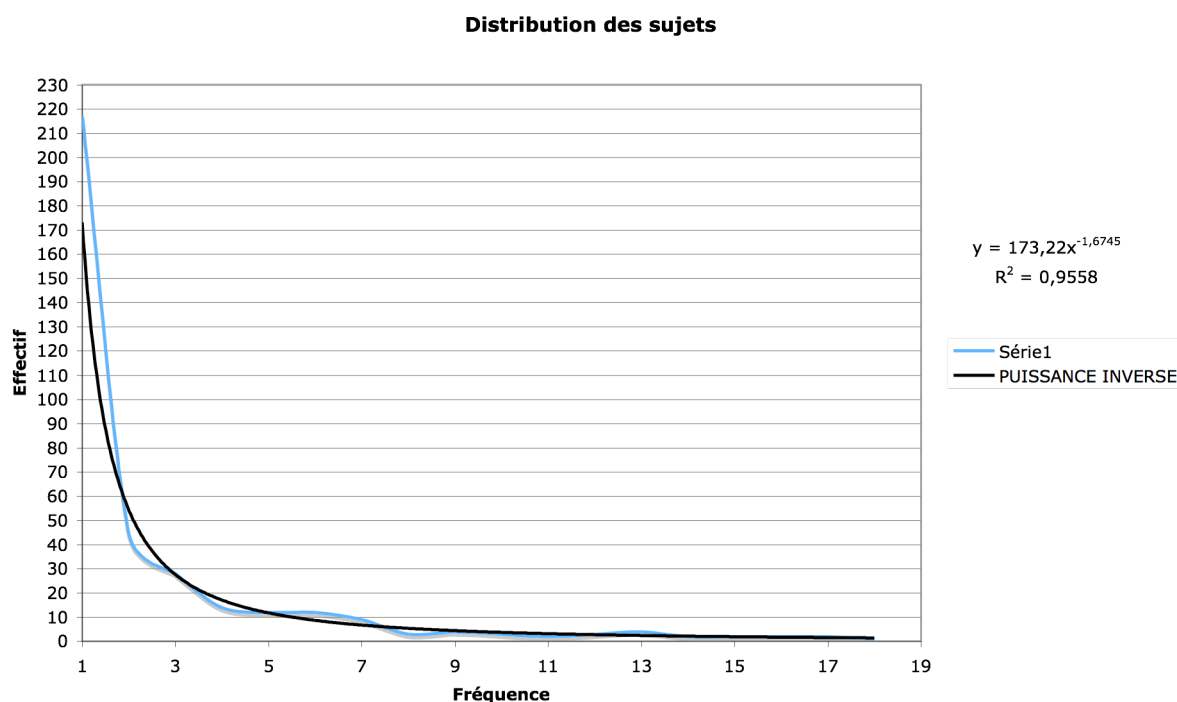


Figure 5

Là encore ressort de cette comparaison graphique une relative homologie entre les deux courbes, IPRI et Lotka. On remarquera toutefois que, pour le niveau de fréquence le plus bas (sujets qui ne sont traités que dans un seul article), la courbe IPRI dépasse assez nettement celle de Lotka. Une légère disjonction des courbes étayée de façon plus transversale par le calcul du carré de la valeur R, bien moins proche de 1 que dans les calculs précédents pour Zipf : 0,9558 le 6 novembre et 0,8013 lorsque limité aux seuls sites français ; 0,9088 le 10 novembre et 0,8719 lorsque limité aux seuls sites français.

Cette dernière observation laisse penser que les sujets « mono-traités » sont particulièrement nombreux sur le web sans pour autant que les sujets faiblement traités dans leur ensemble n'échappent eux à une distribution relativement classique en termes de production de contenu. Ceci nous amène plus largement à revenir sur notre interrogation de départ concernant le caractère heuristique de l'hypothèse de « longue traîne » sur le web.

Discussions des résultats IPRI au regard des lois infométriques existantes et de l'hypothèse de « longue traîne » sur l'internet.

A l'issue de ces approfondissements de l'analyse quantitative du pluralisme offert sur le web, le constat peut être dressé d'une correspondance entre les résultats obtenus dans le cadre de cette recherche IPRI et des lois infométriques plus anciennes. Qu'en déduire ?

- Tout d'abord que la structure duale [extrême concentration + forte dispersion] est courante dans la production de documents, et que les résultats IPRI n'échappent pas à cette règle. Au contraire, la comparaison IPRI/Zipf conduit plutôt à mettre en évidence une superposition des deux courbes ainsi que des valeurs quasi-constantes. Pour résumer, gardons donc cela à l'esprit : les résultats obtenus pour IPRI, et notamment les grandes inégalités constatées dans la focalisation ou au contraire l'abandon de certains sujets d'actualité, n'ont rien d'extraordinaire... sauf à être extrêmement proches de ce que devrait être la distribution ordinaire de contenus au sein de larges corpus.

- Cela signifie par ailleurs que les résultats obtenus pour IPRI ne viennent pas en appui de l'hypothèse de la « longue traîne » sur l'internet, celle-ci se voulant originale et inédite par rapport à ce qu'ont été les formes précédentes de production et de distribution de contenus. En particulier, l'idée d'une traîne aussi consistante qu'étalée, n'est pas vérifiée pour IPRI (pas plus qu'elle ne l'a été dans d'autres recherches d'ailleurs). Si on prend le cas de la représentation graphique inspirée de Lotka, on a par exemple vu que les sujets mono-traités sont particulièrement nombreux mais que dans le même temps les sujets faiblement traités ne sont pas spécialement surreprésentés en comparaison de ce qui a été observé dans d'autres domaines de production de contenus (domaines parfois très anciens comme par exemple la production scientifique imprimée).

- Pour terminer, il convient toutefois de ne pas surinterpréter les enseignements de telles comparaisons. Élaborées à partir de la production d'articles scientifiques ou de la simple comptabilisation lexicale, les lois infométriques mobilisées ici n'ont que peu à voir avec une analyse de contenus d'informations d'actualité entre plusieurs sources sur le web.

Pourquoi les mettre en parallèle alors ? Encore une fois pour ne pas s'illusionner sur l'apparente nouveauté d'une telle distribution, et aussi pour rappeler qu'une telle structure correspond en définitive davantage à la réalité de la production de contenus qu'une répartition équiprobable où chaque sujet serait abordé de façon proportionnelle. Si l'on considère que le pluralisme de l'information doit se matérialiser autant par une variété des sujets traités que par une égalité de leur distribution, alors sans doute faut-il réaffirmer le fait qu'une intervention publique est nécessaire pour le garantir et en quelque sorte le « construire artificiellement ». Car les lois de la production de contenus, tout comme celles des avantages cumulés auxquelles se rapporte notamment celle de Pareto, sont implacables dans l'autre sens : la concentration sur les uns coexiste avec une dispersion des autres dans les lois de puissance inverse, et à cela l'internet ne semble pas changer grand chose quoi qu'en disent les hypothèses techno-déterminées de « longue traîne ».

1.3 Prolongements et compléments de la recherche

Au-delà de son objectif premier d'une expérimentation en matière d'analyse quantitative du pluralisme de l'information sur le web, la recherche exploratoire IPRI a ouvert d'autres voies. Celles-ci complètent les résultats précédents, en puisant dans la même base de données initialement constituée. Elles ne connaissent pas toutes le même niveau d'aboutissement aujourd'hui, mais font néanmoins surgir des chantiers de recherche intéressants à creuser à l'avenir pour aborder plus largement la question du pluralisme de l'information sur l'internet : analyse du traitement d'un sujet particulièrement redondant, selon les différentes catégories de sites ; représentation graphique des proximités éditoriales entre sources ; suivi diachronique de la propagation des informations.

1.3.1 Analyse du traitement médiatique de l'élection de Barack Obama.

Chercheuse impliquée : Annelise Touboul

Préambule : le travail d'analyse qui suit ne peut être qualifié de purement qualitatif. En effet, ce travail a été réalisé à partir de l'enregistrement, puis du traitement des flux RSS de la journée du 6 novembre 2008 et prend donc pour point de départ un matériau quantitativement important. Nous expliquerons ci-après les différentes étapes qui nous ont permis de constituer un sous-ensemble moins volumineux de façon à pouvoir travailler sur la formulation de quelques titres de flux RSS de cette journée du 6 novembre.

Aucune étape présentée ci-après n'a été automatisée, ce qui comporte certains inconvénients (temps long de réalisation notamment) mais ce qui permet aussi d'éprouver et surmonter certaines difficultés liées à la nécessité de sous diviser l'échantillon choisi (nécessité d'un travail de définition et d'appréciation identique à celui de l'étape de traitement lexicologique, mais à une échelle moindre). En avançant souvent à « petits pas », on prend conscience du fait que chacune des étapes pose des questions méthodologiques et fournit des résultats qui, sans pouvoir être qualifiés de strictement qualitatifs, permettent de vérifier des hypothèses posées par le projet de recherche IPRI.

Parce qu'une analyse qualitative suppose un certain approfondissement du travail sur les données du corpus, un préalable consiste à réduire la quantité d'items sur lesquels mener l'étude. Nous avons fait le choix de travailler sur le seul sujet de « L'élection américaine de Barack Obama », de loin le plus traité ce 6 novembre 2008 (435 titres). Nous avons conscience du caractère exceptionnel de cet événement et de son traitement médiatique, c'est pourquoi, initialement, nous avons envisagé d'utiliser la même méthodologie sur des sujets générant un nombre d'articles plus conforme à une moyenne quotidienne. Faute de temps – cette phase plus qualitative d'analyse est nécessairement arrivée en bout de chaîne de la présente recherche - nous repoussons ce complément d'étude à une période ultérieure. Même s'il faut admettre que la redondance de l'information a pu être accrue du fait du caractère exceptionnel d'une élection présidentielle aux Etats-Unis dont le candidat vedette est pour la première fois de l'histoire une personne de couleur, on peut aussi considérer que la passion suscitée par cette élection constitue une occasion idéale d'expression citoyenne variée ou au moins décalée par rapport à l'information des médias dominants. Ce sujet si particulier peut donc, en théorie, conduire à autant de redondance que de diversité dans son mode de traitement.

Le projet général de notre étude a donc consisté à travailler sur les répétitions et les écarts au sein d'un même sujet. Pour lui donner davantage de relief et le relier encore plus précisément aux objectifs de la recherche IPRI, nous avons en outre croisé ces différenciations (à rapprocher de la notion de *disparité* mobilisée dans les études sur la diversité culturelle) dans le mode de traitement du sujet « L'élection américaine de Barack Obama » avec une prise en compte des catégories de sites web. Nous présenterons ici successivement les différentes étapes de cette démarche et leurs résultats respectifs. Pour finir, nous signalerons les limites de ce travail et ses prolongements possibles.

Répartition du sujet « L'élection américaine de Barack Obama » selon les différentes catégories de sites web

A partir du travail d'indexation des articles en sujets, nous avons repris la liste des titres relatifs à « L'élection américaine de Barak Obama » (sujet n°287 comportant 435 items).

Voici le graphique qui permet de visualiser la répartition des titres en fonction des différentes catégories de sites définies pour la constitution de la base de données.

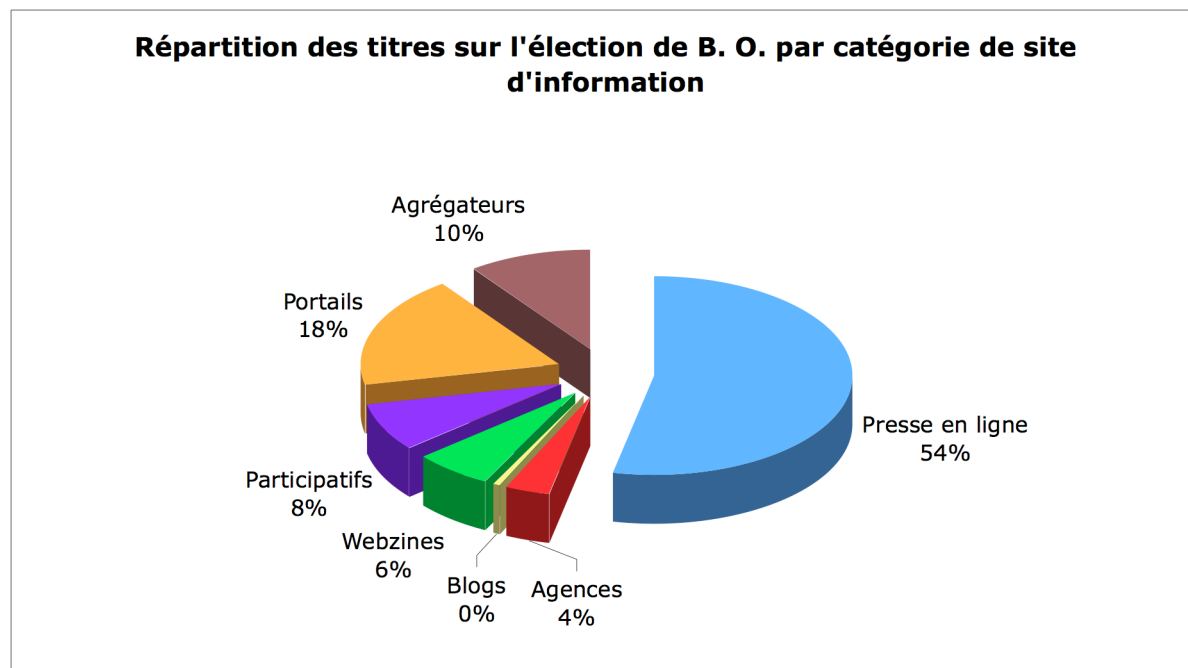


Figure 6

Voici les valeurs précises qui ont permis de réaliser le graphique :

Tableau 6

catégorie de site d'information	nombre de titres sur l'élection de B. Obama le 6/11/2008
Presse en ligne	232
Agences	16
Blogs	2
Webzines	28
Participatifs	34
Portails	80
Agrégateurs	43
Total titres sujet 287	435

A ce stade, deux constats s'imposent :

- D'une part, la médiatisation « en ligne » de ce sujet si redondant lors de la journée du 6 novembre 2008, est en grande partie l'œuvre des sites de presse en ligne. Ces derniers étant toutefois les plus représentés au sein de notre corpus, difficile toutefois d'en tirer des conclusions. En l'absence de pondération des résultats, encore une fois faute de temps, mieux vaut considérer cet élément comme une indication sur la place toujours centrale des médias traditionnels, *via* leurs éditions numériques, dans *l'agenda setting*.

- D'autre part, les blogs se sont très peu saisis de ce sujet, pourtant au centre de l'agenda médiatique ce jour-là, sur l'internet comme sans doute sur les autres supports (presse écrite ou radio et télévision). Là encore, en raison des mêmes réserves que celles exprimées précédemment, conservons à ce résultat son caractère indicatif.

Pour affiner l'analyse, nous avons ensuite repéré et comptabilisé les répétitions identiques de titres complets de la part d'une même source. Ceci nous donne quelques indications sur les stratégies d'occupation du territoire informationnel et de « matraquage » de certains acteurs.

Le regroupement des doublons au sein d'une même source a permis de produire le graphique suivant, rapporté aux différentes catégories de sites web

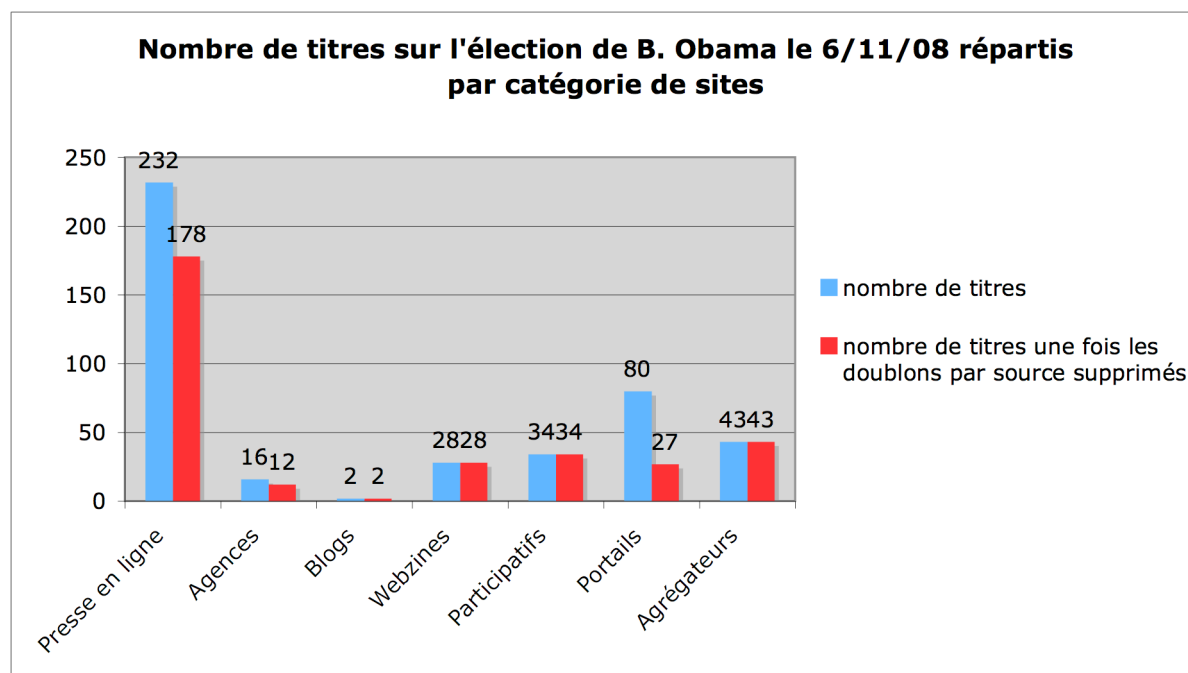


Figure 7

Ce graphique met en évidence les stratégies de répétition intensive des portails et dans une moindre mesure de la presse en ligne. Nous émettons l'hypothèse que, à défaut de créer des contenus originaux, le « matraquage » permet d'occuper le territoire (notamment en terme de référencement) en occultant éventuellement la présence d'acteurs moins « bruyants ».

Différenciations dans le mode de traitement du sujet

Face à ce sujet volumineux en termes d'articles suscités, nous avons ensuite opéré des regroupements selon les « sous-sujets » ou les différents angles de traitement de ce sujet. Par rapport à l'analyse semi-automatisée menée préalablement sur la base du cadre primaire de leur perception (voir partie 1.2.1), cela revient ici à discriminer les titres portant sur un même sujet selon leur cadre médiatique second. A l'issue d'une analyse en ce sens, les titres traitant du sujet « L'élection américaine de Barack Obama » ont ainsi été classés dans 11 sous-catégories (dont une regroupant les « indéfinis ») :

Sous-sujets de « L'élection américaine de Barack Obama » :

1. Les actions et l'agenda politique d'Obama au lendemain de son élection
2. Les défis qui l'attendent
3. Bush, son bilan et ce qu'il va faire
4. Retour sur l'élection, les chiffres, les raisons de la victoire ou de l'échec, les États qui basculent
5. La symbolique de l'élection, l'enthousiasme qu'elle déclenche
6. Les réactions dans le monde
7. Les réactions en France
8. Barack Obama, l'homme, emblème
9. Les opposants, les critiques, les sujets pièges
10. Opinion d'expert, sondage du public, débat
11. Indéfini

Pour pousser encore plus loin l'analyse de la différenciation dans le mode de traitement du sujet, nous avons observé la répartition d'un sous-sujet en particulier (en l'occurrence, le sous-sujet n°1 « Les actions et l'agenda politique d'Obama au lendemain de son élection ») parmi les différentes catégories de sites.

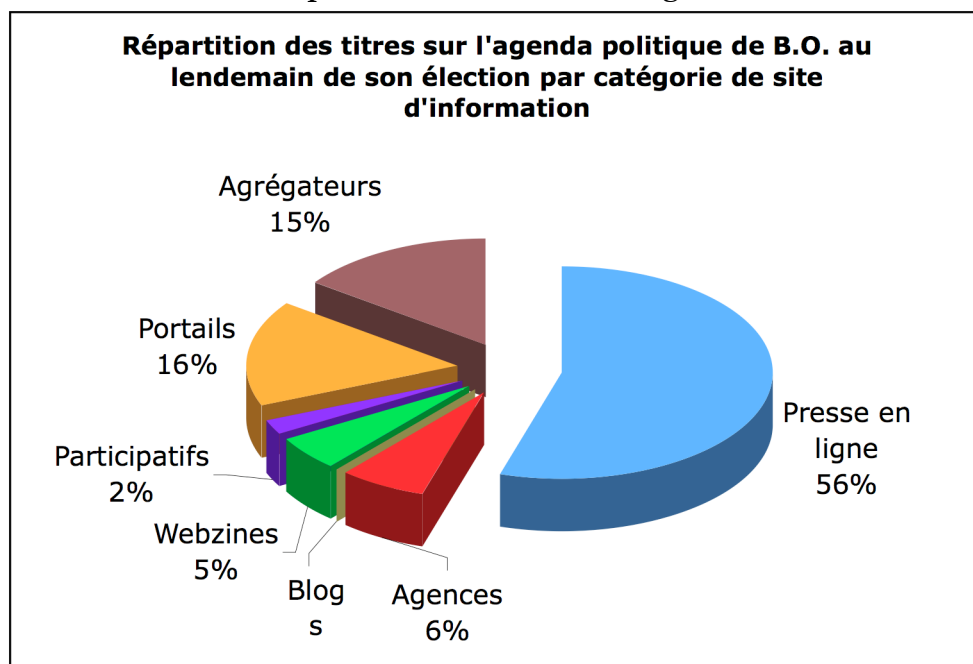


Figure 8

La structure de répartition, entre catégories de sites, du sous-sujet « Les actions et l'agenda politique d'Obama au lendemain de son élection » s'avère comparable à celle observée précédemment pour la répartition du sujet « L'élection américaine de Barack Obama ». Il semblerait ainsi que le mode de traitement d'un sujet ne soit pas un facteur modifiant l'occupation de l'espace médiatique numérique par les différentes catégories de sites. Une hiérarchie semble demeurer, à l'avantage des structures les plus professionnalisées et industrialisées (presse en ligne, portails, agrégateurs) par rapport aux structures amateurs et alternatives (sites participatifs, webzines, blogs).

Une fois de plus, il faudrait prolonger l'analyse pour pouvoir affirmer avec plus de force un tel résultat indicatif. Prendre par exemple un sous-sujet appelant à moins de factuel que « Les actions et l'agenda politique d'Obama au lendemain de son élection » : un sous-sujet comme « La symbolique de l'élection, l'enthousiasme qu'elle déclenche » amènerait peut-être davantage d'information de commentaire et donc une présence plus favorable des blogs, webzines, et sites participatifs.

Cette piste fait partie des nombreux chantiers qu'une analyse qualitative plus consistante, avec davantage de temps, appelle. En rentrant dans une approche plus sémiologique du texte intégral de chaque article, on pourrait alors s'intéresser aux évolutions d'un énoncé repris avec des modifications, soit par un seul émetteur (évolution dans le temps) soit par des émetteurs différents (dire un peu différemment ce que tous disent). Cette étape permettrait de voir s'il est simplement question d'une apparence de variété (c'est à dire le masquage de la redondance sous l'apparence d'un travail éditorial de réécriture) ou bien d'une réelle diversité d'approche au sein d'un sujet circonscrit. Cela suppose de définir le niveau de variation, considérant que le changement de certains termes ne modifie pas en profondeur le sens d'un énoncé. L'évaluation pose bien entendu la question de l'interprétation du sens des titres et suppose d'adosser les décisions à une argumentation détaillée. Enfin, une ultime phase de l'analyse consisterait à noter qu'il existe des titres originaux, produits par une seule source (éventuellement repris par un agrégateur) et de vérifier la nature des émetteurs. Il sera aussi possible d'affirmer combien de propositions différentes ont été faites au sein d'un même sous-sujet, de mesurer la multiplication de ces énoncés de base et de vérifier les circuits d'amplification de certains énoncés finissant par noyer et recouvrir le travail d'interprétation produit par certains sites.

1.3.2 Représentation graphique des proximités éditoriales

Chercheurs impliqués : Samuel Gesche, Guy Caplat , Sylvie Calabretto.

L'application ArcEnCiel développée par Samuel Gesche dans le cadre de son travail de doctorat permet, entre autres, de représenter graphiquement les proximités sémantiques à l'intérieur d'un corpus. A son origine, l'application ArcEnCiel est dédiée à la confrontation de points de vue / opinions. De tels points de vue sont définis comme étant des points de vue d'experts, comprenant autant une dimension de connaissance disciplinaire qu'un positionnement personnel et une influence communautaire. Dans le cadre de la recherche IPRI, une telle définition peut être assez facilement transposée au cas des rédactions des sources-médias que nous avons exploitées : les rédactions sont considérées comme experts du traitement de l'information journalistique, et font montre d'un positionnement personnel (par rapport aux autres rédactions) et d'une influence communautaire (dans la mesure où l'on peut les positionner sur ce que l'on a coutume d'appeler un « échiquier politique », qui en réalité inclut bien plus que les simples préoccupations politiques).

Association des rubriques communes aux différents sites

En amont, ont été extraites des flux RSS des sources médias les catégories (en fait, rubriques) dans lesquelles elles classaient l'information. Ainsi, nous pouvions ranger les articles par rubrique, lier les rubriques de différentes sources média si elles étaient équivalentes et, dans le cas des sources qui ne donnaient pas de telles rubriques, les inférer à partir de la similarité du contenu de leurs articles avec celui d'articles classés. La figure 9 donne un exemple de visualisation de ces rubriques, sans faire apparaître les articles. Chaque pastille de couleur correspond à une source média.

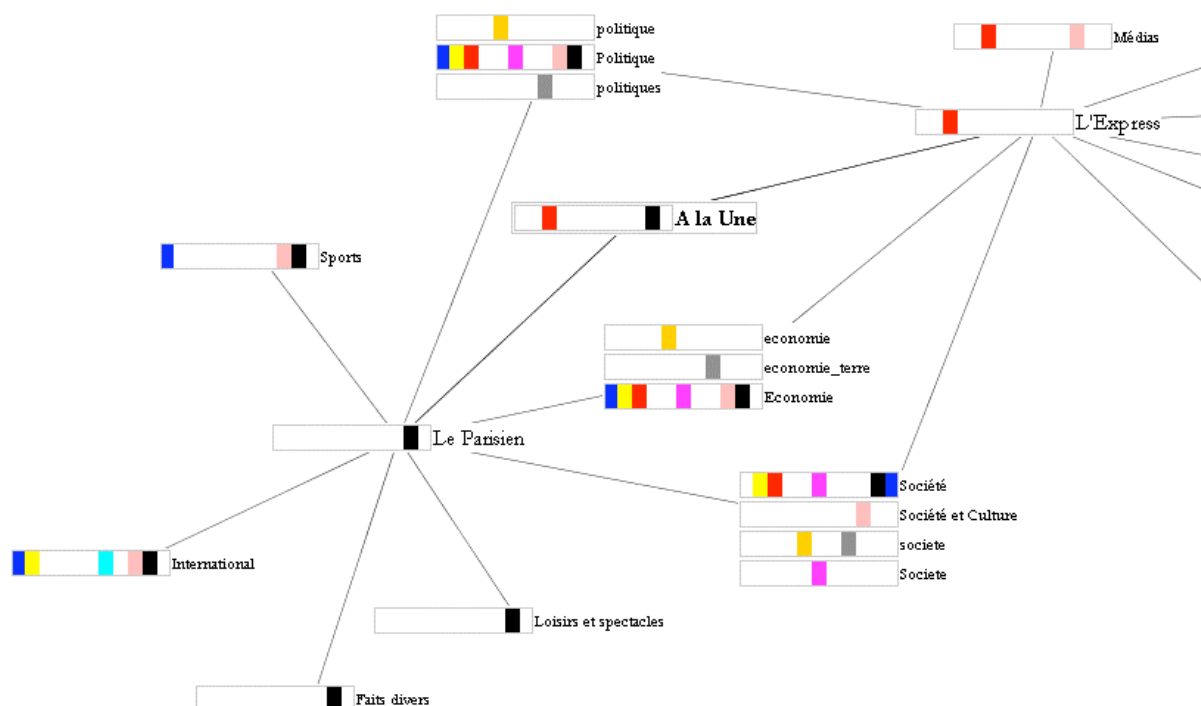


Figure 9 : visualisation et mise en relation des rubriques de différentes sources média

Dans la mesure où de meilleures techniques étaient disponibles par ailleurs, et permettaient d'inférer des catégories plus descriptives à partir du contenu, nous avons pu nous concentrer davantage sur l'aspect de la confrontation.

Confrontation des « points de vue » éditoriaux des sites

La campagne d'échantillonnage des 6 et 10 novembre 2008 nous a fourni entre 4000 et 5000 articles, classés par thématiques et sujets à l'issue de l'analyse semi-automatisée de contenu. Nous avons donc pu visualiser, au sein d'une même thématique, les différents articles qui avaient été publiés. Pour des raisons techniques, nous n'avons eu accès qu'à leurs titres. Cependant, étant donnée l'importance du titre dans le contexte journalistique, les résultats que l'on peut inférer du titre sont tout à fait intéressants.

Nous avons donc, à partir de la hiérarchie de thématiques, sujets et articles, construit pour chaque source un document XML constituant le point de vue de cette source sur l'actualité, selon la DTD suivante :

```
<! ELEMENT SOURCE (NOM, THEME*) >
<! ELEMENT THEME (NOM, SUJET*) >
<! ELEMENT SUJET (NOM, ARTICLE*) >
<! ELEMENT NOM (#PCDATA) >
<! ELEMENT ARTICLE (#PCDATA) >
```

Ces points de vue sont ensuite importés dans l'environnement ArcEnCiel, qui permet de visualiser « la différence autour de la similarité ». Les points de vue étant constitués autour des sources, et puisque la hiérarchie unique de thématiques et de sujets fournissait une similarité transversale, nous pouvons facilement étudier dans quelle mesure les différentes sources abordent les différentes thématiques et sujets.

La visualisation en parallèle des différentes sources média nous a permis de visualiser aisément la redondance de l'information, ainsi que sa diversité. Nous avons pu ainsi repérer quelques phénomènes intéressants :

a) Similitudes et écarts dans les thématiques abordées

L'environnement ArcEnCiel permet la visualisation des points de vue au sein de graphes par l'intermédiaire de pastilles de couleur. Ainsi, on peut facilement, pour une thématique ou un sujet donné, savoir quelles sources l'abordent et lesquelles n'en parlent pas. La figure 10 présente un exemple sur six thématiques.



Figure 10 : visualisation du traitement de l'information

b) Redondance de sources

Certaines sources média, bien que différentes par leur nom, diffusent la même information. C'est notamment le cas de France 2 et France 3. La figure 11 illustre cela : tous les articles qui apparaissent dans l'une des sources apparaissent, avec le même titre, dans l'autre source (les sources sont différenciées par des pastilles couleur). En d'autres termes, le flux RSS est identique. Une visite de leurs sites respectifs permet d'ailleurs de s'apercevoir de la similarité très forte, non seulement entre les flux, mais dans l'ensemble du site. Cela semble indiquer que le centre de production des articles en ligne est lié à France Télévision, propriétaire des deux chaînes, et que chacun des sites pointe simplement sur ces articles (la seule différence dans l'URL des articles étant le '2' ou '3' caractéristique de la chaîne).

	Acheminement, par train puis camions, de déchets radioactifs jusqu'à Dannenberg en Alle
	ACTUALITES LITTERAIRES ET CULTURELLES {Ces termes sont identiques}
	ACTUALITES RADIO-TELE-CINE-MUSIQUE {Ces termes sont identiques}
	Alerte aux vents violents dans 17 départements de France {Ces termes sont identiques}
	Arrestation, en Allemagne et à la demande de la France, de Rose Kabuye ; protestations au
	CATASTROPHES NATURELLES ET PHENOMENES CLIMATIQUES {Ces termes sont identiques}
	Commémoration du 11 novembre 1918 {Ces termes sont identiques}
	Deux morts dans l'incendie d'un appartement à Paris {Ces termes sont identiques}
	DIPLOMATIE INTERNATIONALE-CONFLITS ARMES {Ces termes sont identiques}
	ECOLOGIE/ENVIRONNEMENT {Ces termes sont identiques}
	FAITS DIVERS {Ces termes sont identiques}
	Il y a 90 ans, l'armistice du 11 novembre 1918 {Ces termes sont identiques}
	Il y a 90 ans, l'armistice du 11 novembre 1918 (2) {Ces termes sont identiques}
	Il y a 90 ans, l'armistice du 11 novembre 1918 (3) {Ces termes sont identiques}
	Il y a 90 ans, l'armistice du 11 novembre 1918 (4) {Ces termes sont identiques}
	Il y a 90 ans, l'armistice du 11 novembre 1918 (5) {Ces termes sont identiques}
	Incendie à Paris: deux morts {Ces termes sont identiques}
	Incendie à Paris: deux morts, trois blessés graves {Ces termes sont identiques}
	La chanteuse Miriam Makeba est décédée d'une crise cardiaque {Ces termes sont identiques}
	La chanteuse Miriam Makeba est morte {Ces termes sont identiques}

Figure 11: deux sources média identiques

c) Répétition intra-flux

Certains articles sont publiés plusieurs fois dans la même journée. ArcEnCiel, pour des raisons d'unicité des termes employés, assigne un numéro à des termes identiques au sein d'un même point de vue. Cela se traduit, dans le cas qui nous intéresse, par des articles au titre identique ayant des numéros. L'exemple de la figure

12, découpé dans la figure 11, présente cinq occurrences du même article au sein des mêmes flux dans la même journée. Dans certains cas, c'est non moins de vingt fois que l'on retrouve le même article.

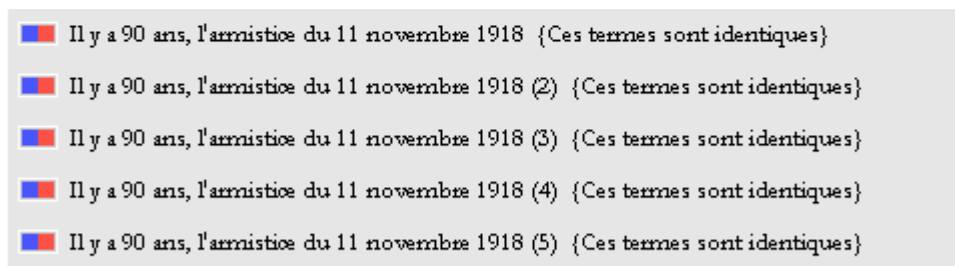


Figure 12 : répétition d'articles

Ce type de redondance a été remarqué et étudié de manière complémentaire dans les parties dédiées à l'analyse qualitative et au suivi diachronique. Il s'agit bien de multiples occurrences du même titre sur le même flux et non, par exemple, d'une erreur liée à la récupération des articles sur les flux.

d) Positionnement politique sur un même sujet

Mais le résultat le plus marquant de cette utilisation d'ArcEnCiel dans le cadre de la recherche IPRI est sans conteste la visualisation du positionnement des sources sur les sujets traités. Ainsi, comme on peut le voir dans la figure 13, nous pouvons visualiser les titres donnés par les différentes sources à leurs articles, et en inférer leur position sur le sujet considéré. Le positionnement politique, en particulier, peut être vu aisément.

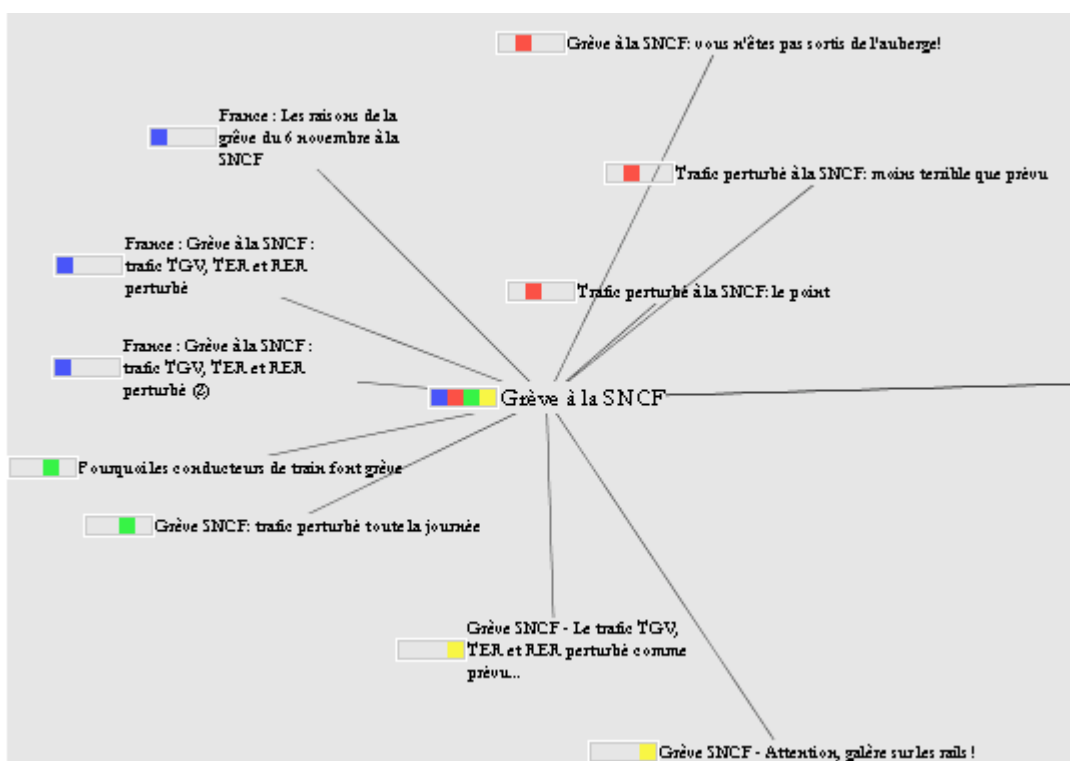


Figure 13 : Positionnement politique sur un sujet
(Bleu : La Croix; Vert : Libération; Rouge : Le Post; Jaune : TF1)

Ici, on rejoint le niveau de la différenciation ou des disparités dans le traitement d'un sujet. Sur un sujet aussi conflictuel qu'une grève, toutes les sources médias affichent au moins un titre factuel, résumé par la formule redondante de « trafic perturbé » ; mais ils expriment une prise de position plus distincte au travers d'autres titres dont l'angle va de la compréhension des raisons de la perturbation pour Libération et La Croix (avec respectivement « Les raisons de la grève du 6 novembre à la SNCF » et « Pourquoi les conducteurs de train font grève ») à l'alerte aux usagers pour Le Post et TF1 (avec respectivement « vous n'êtes pas sortis de l'auberge ! » et « Attention, galère sur les rails ! »).

1.3.3 Suivi diachronique des informations

Chercheurs impliqués : Loïc Maisonnasse, Samuel Gesche, Előd Egyed-Zsigmond, Geneviève Lallich-Boidin, Sylvie Calabretto, Guy Caplat

Dans les parties précédentes, l'analyse s'est le plus souvent limitée à une étude synchronique (ex : sur une journée) du pluralisme et de la redondance. Un tel découpage est nécessaire pour produire des analyses sur de grandes quantités de données : comme nous l'avons souligné, identifier les sujets de plus de 2000 articles représente par exemple une opération lourde, même si l'analyse de contenu est alors semi-automatisée. Une telle limitation temporelle obère toutefois le caractère dynamique de la production d'information et de l'*agenda setting* dans les médias, et doit donc se compléter d'un suivi diachronique de la propagation des informations. En effet, certaines informations peu redondantes le jour J pourraient en fait constituer la terminaison d'un mouvement de redondance entamé à J-1 ou J-10, ou inversement le tout début d'un mouvement de redondance, se poursuivant en J+1, J+2, etc.

De ce point de vue, le système de lemmatisation mis au point en même temps que la base de données s'est avéré tout à fait utile pour retrouver - de façon cette fois totalement automatisée - la trace, antérieure et ultérieure, d'apparition des informations, sur la base de leurs titres. Il a permis d'analyser la production des 89 flux RSS enregistrés en continu pendant 2 semaines, du 4 novembre 2008 au 20 novembre 2008.

Présentation de la démarche

L'analyse s'est effectuée en deux étapes. La première consiste à déterminer des scores de similarité entre les items produits par les différentes sources d'informations (ici des flux RSS). Une deuxième étape consiste, à partir de ces liens, à créer un graphe orienté temporellement reliant les sources d'informations en fonction de la quantité d'information reprise temporellement d'une source à l'autre. Ce qui permet d'évaluer la redondance et la propagation de l'information entre les différentes sources.

Similarité entre items

Le titre est important, à tel point qu'il est souvent suffisant pour déterminer si deux articles ont un contenu semblable⁶. Par conséquent nous ne nous intéresserons dans cette partie qu'aux mots contenus dans le titre des articles. Cependant différentes similarité peuvent être employés pour déterminer si deux titres sont similaires. Les mots employés dans un titre devant être des mots significatifs nous émettons l'hypothèse, certes réductrice, que la similarité thématique entre deux titres peut s'obtenir par comparaison des mots pleins qui constituent ces deux titres. De plus, ces mots pouvant être affectés par des variations lexicales nous établissons que deux mots sont similaires si leur lemme est identique. Pour obtenir les mots *pleins* et les lemmes correspondant, nous avons utilisé l'analyseur morphosyntaxique TreeTagger⁷. Plusieurs métriques sont ensuite disponibles pour comparer deux chaînes de mots. Certaines de ces métriques dérivent de la recherche d'information. D'autres s'intéressent plus à la comparaison de chaînes de mots.

⁶ Afin d'éviter toute confusion, soulignons que nous n'employons pas ici la notion de sujet telle que définie au moment de l'analyse semi-automatisée, c'est-à-dire un sujet identifié en grande partie par les chercheurs à la lecture des titres et des descriptions voire du texte intégral des articles. Dans cette partie, la parenté entre des articles sujet est automatiquement déduite d'un calcul de similarité des lemmes entre titres.

⁷ <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

Nous proposons ici d'utiliser une mesure ensembliste simple donnée par :

$$score(t1, t2) = \frac{|t1 \cap t2|^2}{|t1| \times |t2|}$$

L'ensemble des scores forme alors une matrice symétrique par exemple :

	item 1	item 2	Item 3
item 1	1	0,4	0,2
item 2	0,4	1	0,8
item 3	0,2	0,8	1

TABLEAU 7 : Exemple de matrice.

Nous ne conservons dans cette matrice que les couples (item1,item2) tel que $date(item1) < date(item2)$. En plus des scores, dans cette matrice nous conservons un ensemble d'informations, tel que la date de l'item, son nombre de lemmes et le nombre de lemme en commun. Cette matrice est stockée dans la table suivante :
MatriceFinale(item1,date1,flux1,nb1,item2,date2,flux2,nb2,nbCommun,score)

Cette matrice est obtenue par l'enchaînement de requêtes SQL, pour réduire la taille de la matrice, cette matrice qui ne contient que les items qui ont au moins deux lemmes en commun dans le titre. Part la suite nous référerons à cette matrice par *MF*.

Nous créons également une deuxième version de cette matrice dans laquelle nous filtrons les redondances propres à un même flux. En effet dans de nombreux flux RSS, dans le but de mettre en avant une actualité (volonté éditoriale) ou suite à des problèmes de publications, un item peut être publié plusieurs fois. Pour cela la matrice *MF* est filtrée pour ne prendre en compte qu'une occurrence de chaque couple d'item du même flux identique sur le titre (titre, flux source). Comme nous souhaitons étudier l'ordre d'apparition de la nouveauté sur le corpus pour un couple nous ne conservons que l'item le plus ancien dans le temps. Par la suite nous référerons à cette matrice par *MFF*.

Puisque que la redondance au sein même d'une source peut découler d'une volonté éditoriale ou marketing, l'étude des deux matrices est intéressante. Pour cette raison notre analyse de la redondance portera sur les deux matrices (*MF* et *MFF*).

Synthèse temporelle

Les matrices (*MF* et *MFF*) obtenues par l'analyse de similarité représentent un graphe orienté (par les dates) entre items, mais ce graphe est trop grand, il considère tous les couples d'items ayant deux lemmes commun.

Une deuxième étape consiste donc à le réduire. Cela consiste à donner des contraintes plus fortes pour considérer que deux items sont reliés. Deux types de contraintes (complémentaires) ont été retenus :

- Le filtrage par rapport au score, c'est à dire conserver les titres les plus similaires
- Le filtrage par rapport au nombre de lemmes en commun

Par exemple deux items sont reliés s'ils ont 4 lemmes en commun et si leur score de proximité est supérieur à 0,5. On peut ainsi choisir de varier ces deux contraintes pour définir si deux items sont liés.

A l'aide de ces contraintes, nous obtenons une matrice de liens entre items binaire qui définit une relation ou non entre deux items. Ces liens sont orientés par rapport aux dates d'apparition des items. Cette matrice peut être représentée par un graphe qui permet de suivre l'évolution temporelle des reprises, ou redondances, au sein des items. Mais cette matrice ne permet pas de donner des conclusions quantitatives sur les liens entre les producteurs de flux

Pour pouvoir donner des indications sur la redondance au niveau des sources d'informations, il est possible de synthétiser ces relations au niveau du flux RSS. Un lien entre deux flux est créé en fonction d'un certain nombre de conditions sur les liens entre items. Nous posons les contraintes suivantes sur l'apparition d'une relation entre deux flux à partir des relations entre items :

- D'une part par rapport au nombre de relations regroupé : un flux entretient une relation avec un autre flux si au moins N de leurs items entretiennent une relation.
- D'autre part, sur la moyenne des scores : un flux entretient une relation avec un autre si la moyenne des scores des relations qu'il regroupe est supérieure à Min .

Ces deux contraintes, exécutées à l'aide d'une requête SQL, permettent d'obtenir une matrice de liens entre flux.

Résultats

Les résultats présentés dans cette partie ont été obtenus sur un extrait de la base de flux rss allant du 4 novembre au 20 novembre 2008. Nous présentons dans un premier temps des résultats d'une étude globale du corpus. Dans un deuxième temps nous présentons les résultats de notre méthode d'analyse temporelle.

a) Étude globale de la base

Nous étudions d'abord la base de manière globale, pour extraire des renseignements sur la diversité de production et des thèmes sur le web. Pour cela nous étudions dans le tableau 8 la fréquence des lemmes de mots plein dans les titres et dans les descriptions des items de flux RSS.

Lemme	countTitle	countDesc	Lemme	countDesc	countTitle
obama	2825	3708	avoir	28463	1545
être	2197	20923	être	20923	2197
A	2113	0	président	3982	488
france	1562	2288	obama	3708	2825
avoir	1545	28463	premier	3676	848
Ps	1396	1136	source	3265	23
sncf	1143	662	barack	3027	766
royal	992	1420	mardi	3013	111
crise	947	1987	faire	2977	796
nouveau	929	2783	an	2963	859
An	859	2963	deux	2819	532
premier	848	3676	nouveau	2783	929
mort	804	970	annoncer	2770	369
faire	796	2977	vendredi	2718	177

sabotage	795	342	abonner	2597	0
barack	766	3027	mercredi	2538	154
grève	704	903	américain	2521	702
américain	702	2521	jeudi	2422	150
congrès	637	1245	français	2303	481
20	628	1093	lundi	2295	105
g20	622	608	france	2288	1562
aller	606	1399	devoir	2018	306
marché	603	1045	crise	1987	947
ségolène	587	1188	jour	1893	291
élection	568	1187	pouvoir	1837	337
minute	556	730	dernier	1789	126
sarkozy	550	1270	etats-unis	1782	547
etats-unis	547	1782	pays	1691	283
deux	532	2819	ministre	1641	233
Fréquence des 29 lemmes les plus fréquents sur les titres			Fréquence des 29 lemmes les plus fréquents sur les descriptions :		

TABLEAU 8 : lemmes les plus fréquents sur l'extrait de la base.

Les résultats nous montrent en effet que les lemmes les plus fréquents dans les titres ne sont pas obligatoirement ceux qui sont les plus fréquents dans les descriptions. En effet, dans les titres, les journalistes utilisent des termes plus courts que dans les descriptions, ainsi les journalistes utilisent plus facilement « Obama » seul dans le titre mais « Barack Obama » dans la description. D'après ce résultat on peut supposer que la variabilité des termes sur les titres est moins grande que sur les descriptions.

b) Production des lemmes

Nous étudions ensuite la quantité de production de lemmes au cours du temps avec deux échelle ; l'une montrant la production de lemmes par heures (figure 14) et l'autre la production de lemmes par jours (figure 15)

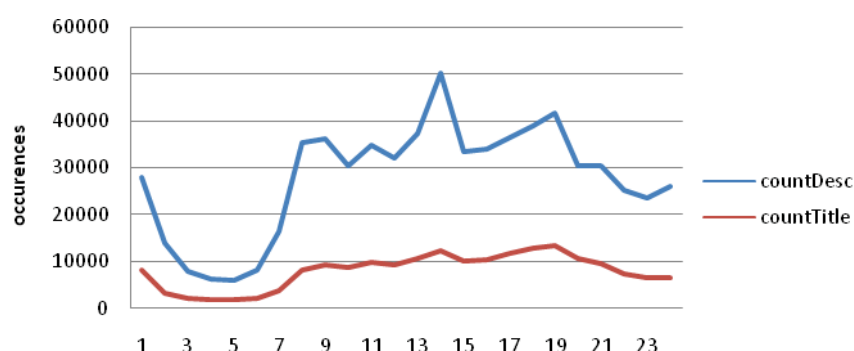


FIGURE 14 : graphe de production de lemmes en fonction des heures de la journée.

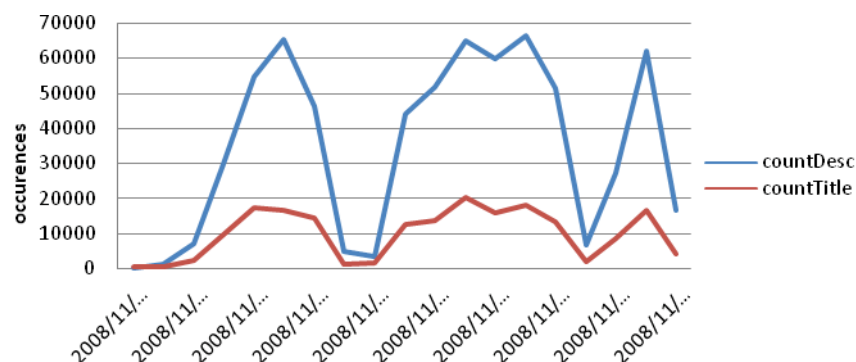


FIGURE 15 – graphe de la production de lemmes en fonction des jours.

Les résultats sur ces deux figures montrent des comportements de production où les heures de la nuit sont des périodes de faible production et de manière identique pour les week-ends et les jours férié.

Ceci est expliqué par les enquêtes socio-économiques au sein des rédactions.

La moindre activité de production de contenus est tout simplement liée aux temps de repos (quotidiens et hebdomadaire) des employés des structures éditrices. Si prosaïque soit-elle, cette précision a le mérite de rappeler que l'internet, pas plus que les autres technologies d'information et de communication, n'échappe aux rythmes de vie sociale des individus. Et que l'idée souvent avancée d'une information en continu sur le web est en grande partie illusoire et à ranger aux oubliettes des prophéties techno-déterminées.

Il est néanmoins plus intéressant et instructif de se pencher sur les trois pics de production de lemmes au cours de la journée : autour de 9h, puis de 13h, et enfin de 18h. Ces trois moments de surproduction éditoriale (tout du moins au niveau des descriptions des articles) correspondent aux horaires de bureau que les éditeurs savent les plus propices à la consultation d'information sur l'internet. C'est donc en toute connaissance de cause que le éditeurs renouvellent leur production dans ces créneaux horaires particulièrement stratégiques, comme le corroborent des témoignages de responsables de Libération.fr ou de Rue89 dans les enquêtes socio-économiques qui suivent.

c) Analyse de cas

La période d'analyse sélectionnée couvrant les élections américaines, nous montrons que notre base permet de suivre les pics d'intérêt pour un thème représenté ici par la quantité de l'utilisation d'un lemme. Cette approche est celle proposée par de nombreux systèmes de suivi de tendances sur l'internet tels que trendpedia⁸

⁸ <http://www.trendpedia.com/>

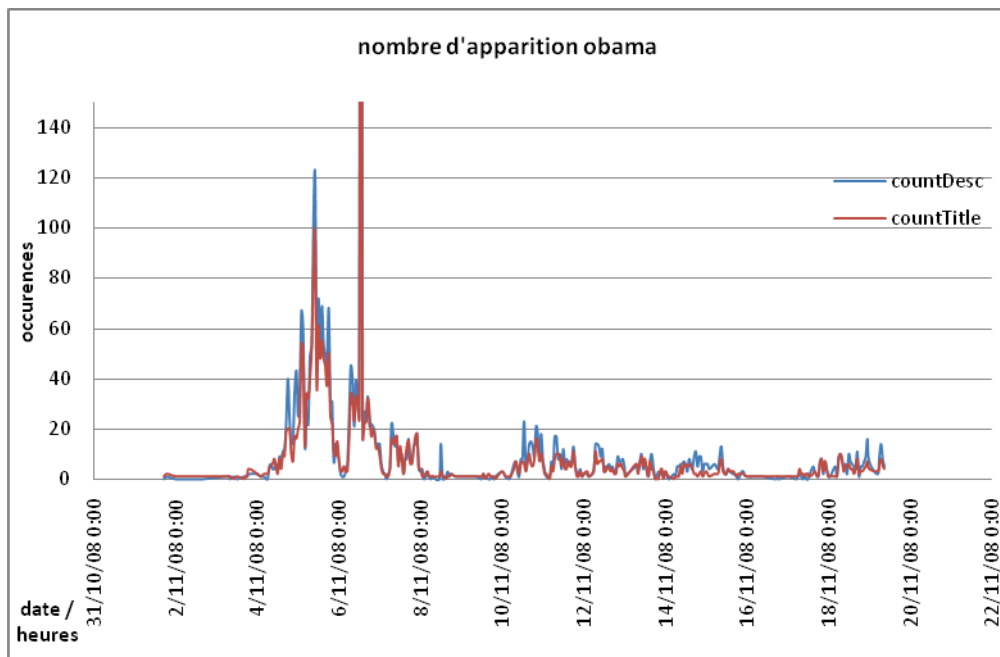


FIGURE 16 : Suivi temporel de l'utilisation du mot Obama au cours de la période d'étude.

d) Analyse temporelle

Nous présentons dans la suite des graphes de liens entre flux obtenus à partir des matrices MF et MFF et en appliquant des contraintes sur les couples et sur le regroupement des couples. Pour afficher le graphe final, nous utilisons ici un ensemble de scripts pour transformer cette matrice en graphe au format xml GraphML(REF) et nous utilisons l'éditeur de graph yEd⁹ pour afficher le graphe.

Dans ces graphes, les nœuds sont des flux RSS, les relations orientées sont les liens montrant des reprises (le numéro associé est le nombre de relations entre items synthétisé par un lien entre flux), les couleurs représentent un type de flux, autrement dit une catégorie de sites (blog, presse, agrégateur)

Avec répétition intra flux (matrice MF)

Nous utilisons ici notre analyse quantitative sur la matrice MF qui contient des répétitions intra flux. Sur l'extrait du 4 novembre au 20 novembre, la matrice MF des liens entre items contient environ 2 600 000 couples dont :

- 52 000 couples avec un score égal à 1.
- 98 000 couples avec un score supérieur à 0.5
- 554 000 couples avec au moins deux lemmes en commun
- 42 000 couples avec au moins six lemmes en commun

A partir de cette matrice nous présentons ici dans la figure 17 un graphe obtenu en appliquant une contrainte sur les items et une sur leur regroupement. Cette figure permet alors de suivre la reprise de l'information de manière quantitative entre sources d'information. On remarque par exemple sur le graphique que les portails d'actualité (ici en rose) sont des maillons favorisant de la redondance de l'information.

⁹ www.yworks.com

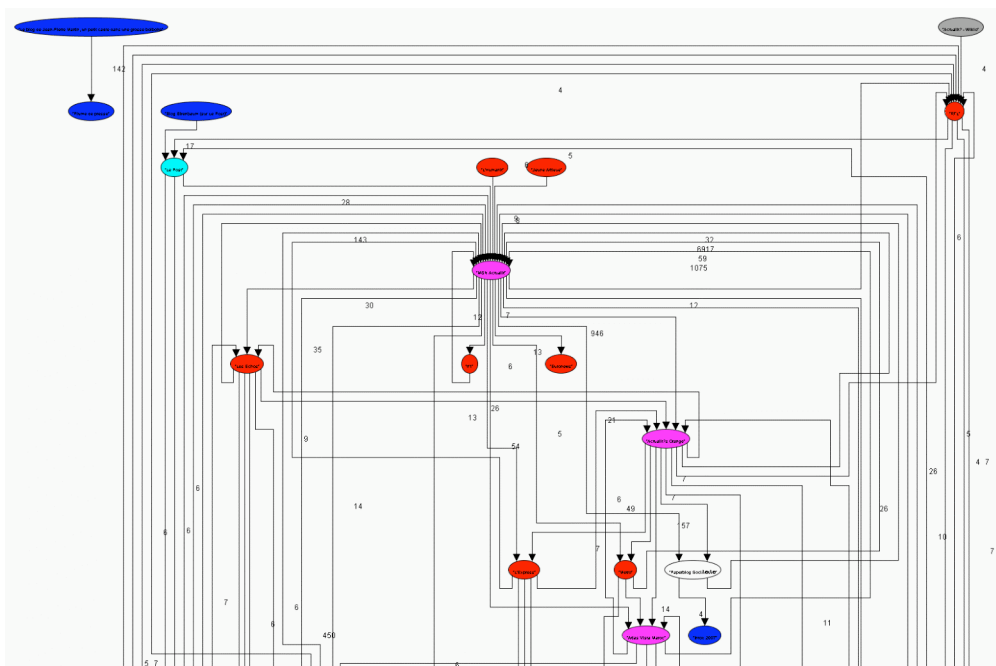


FIGURE 17 : extrait du graphe obtenu avec en ne conservant que les relations liant des items avec les titres identiques (**score=1**) et en ne créant une relation entre flux que si les flux sont reliés par **4 relations** entre items sur la matrice MF.

Le nombre de sources considérées dans notre corpus étant grand, il peut être intéressant d'exclure certains types de source d'information pour mieux appréhender la circulation de reprise. Par exemple, dans la figure 18, nous avons exclu les sites de presse, cela permet par exemple de mieux observer les reprises d'information dans les autres sources de sources d'informations.

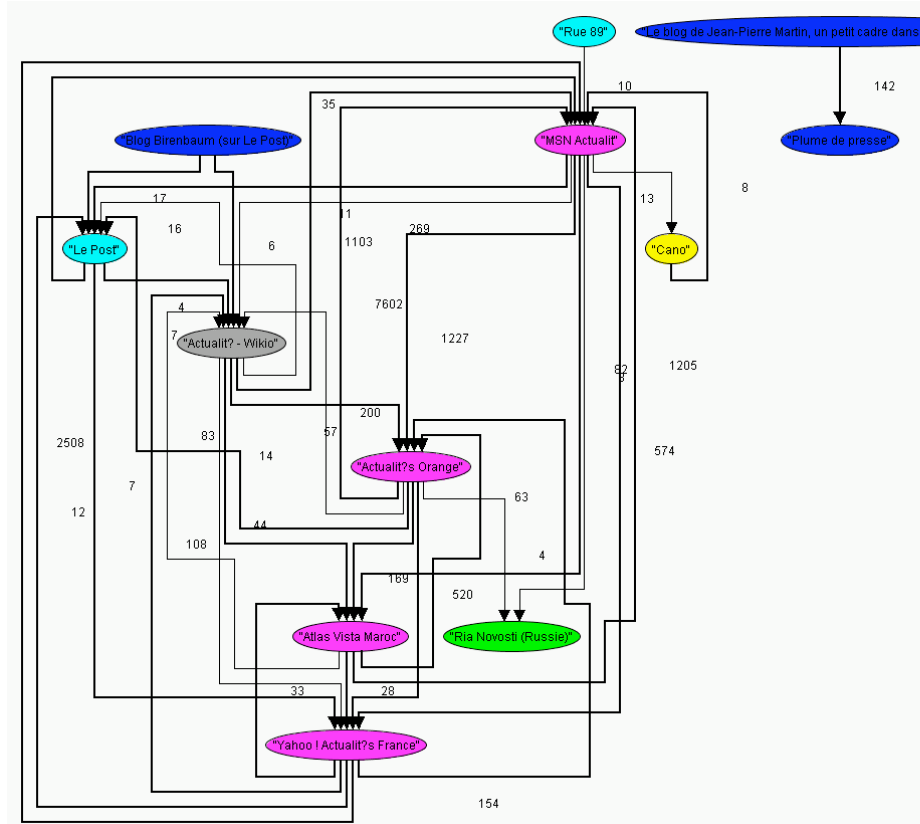


FIGURE 18 : Extrait du graphe obtenu en ne conservant que les relations liant des items avec un score de similarité supérieur à 0.7 et en ne créant une relation entre flux que si les flux sont reliés par **3 relations** entre items et en excluant au préalable les sites de presse sur la matrice MF.

Nous utilisons dans cette deuxième partie notre analyse quantitative sur la matrice *MFF* qui ne contient pas de répétition intra flux. Sur l'extrait du 4 novembre au 20 novembre, cette matrice contient environ 1 150 000 couples dont :

- On remarque que le nombre de liens est divisé par deux et notamment que le nombre de liens avec un score de 1 a fortement diminué. Nous pouvons donc en déduire que la redondance est très forte au sein même de chaque source d'information. En effet certaines source republient des items en ne les modifiant qu'en partie. Par exemple en modifiant le descriptif en lui ajoutant une référence.

[illegible]

Conclusion

D'autre part les graphes obtenus à partir de l'analyse temporelle montrent un intérêt pour déterminer les flux de reprise de l'information entre les sources d'informations.

2. Axe « socio-économie des modalités de production »

L'offre de contenus d'actualité sur le web émane d'une diversité d'acteurs, appartenant au secteur professionnel du journalisme ou non, intégrés à l'équipe rédactionnelle d'un média existant par ailleurs (journal, radio, télévision) ou exclusivement numérique (*Internet pure player*¹⁰). On pourrait supposer que cette diversité des acteurs de l'information web se traduise par une certaine variété de leurs modes de production, et en particulier par l'émergence de pratiques de construction de l'information alternatives à celles établies dans l'univers professionnel du journalisme, reposant par exemple davantage sur la logique de la *public connectivity* plutôt que sur celle de l'*editorial content* (Deuze, 2003). Ces pratiques alternatives se définiraient par le recours à des sources nouvelles et moins officielles, par un appel accru à la participation des lecteurs-internautes, tant pour la création de contenus (*user-generated contents*) que pour la veille (signalement des sujets, *buzz*) ou l'animation du site (commentaires, *forums*), et, enfin, par une volonté éditoriale délibérée de s'extirper de l'agenda médiatique impulsé par les médias dominants.

Plusieurs enquêtes ont déjà été menées à ce sujet, qui montrent notamment un faible recours à la participation des internautes, de la part des responsables de rédaction web (Hermida, Thurman, 2008) ou des journalistes eux-mêmes (Paulussen, Ugille, 2008). Mais ces études portent essentiellement sur les extensions numériques des médias traditionnels, et en particulier imprimés (sites web de quotidiens ou de magazines). Reste donc à savoir si leurs concurrents exclusivement présents sur l'internet, les *pure players*, ont eux adopté ces modalités plus originales de production de l'information d'actualité. Pour le savoir, nous avons privilégié les *pure-players* de l'Internet dans nos enquêtes (un site de journalisme participatif, un webzine d'opinion, un agrégateur de blogs), tout en gardant un œil sur les développements numériques des médias plus traditionnels (un site de quotidien national) à titre de comparaison.

Plusieurs des données empiriques issues de ces enquêtes seront présentées et mises en perspective dans cette synthèse. Nous présenterons en premier lieu les modalités de production en vigueur au sein de chacune des rédactions observées dans le cadre de cette recherche IPRI : Libération.fr, Bakchich, PaperBlog, Rue89. Élargissant la perspective aux connaissances rassemblées lors d'un précédent programme de recherche (Rebillard *et al.*, 2007), nous entamerons alors une démarche comparative pour voir quelles différences existent dans les modalités de production des différents sites, et dans quelle mesure celles-ci peuvent peser sur les projets éditoriaux et plus largement sur le pluralisme des informations ainsi produites.

2.1 Choix des structures observées

Chercheurs impliqués: Béatrice Damian-Gaillard, Franck Rebillard, Nikos Smyrnaio.

Pour donner davantage de cohérence à la recherche IPRI, les corpus respectifs de l'analyse de contenus et des enquêtes socio-économiques auprès des organisations éditrices ont été reliés, dans la mesure du possible.

L'idée imaginée au départ, consistant à être simultanément présents durant une journée dans plusieurs salles de rédaction et à réaliser ensuite une analyse des

¹⁰ Terme anglais qui désigne les entreprises qui développent leur activité sur le net.

contenus publiés durant cette même journée, n'a pas pu être mise en pratique, encore une fois eu égard aux moyens limités de la recherche IPRI. En revanche, il a été possible d'enquêter au cours d'une période identique (automne 2008) auprès d'organisations éditrices dont les sites figurent dans le corpus de l'analyse de contenu.

Ce fut l'un des critères de choix des organisations éditrices sur lesquelles enquêter. L'autre fut, comme pour celui du corpus de l'analyse de contenu, d'avoir une représentation de toutes les catégories de sites web.

Dans le cadre du programme de recherche précédent (programme « Diversité », 2007), nous avons pu enquêter sur les modalités de production au sein de plusieurs sites faisant partie du corpus sélectionné pour l'analyse de contenu IPRI : une agence de presse (AFP), des médias traditionnels (Les Echos ; Journal du Dimanche et Paris-Match au sein du groupe Lagardère ; La Croix au sein du Groupe Bayard-Milan), des agrégateurs (Wikio et Google). Nous avons souhaité compléter cette analyse pour la recherche IPRI en enquêtant auprès de sites faisant partie de ce même corpus et rattachés à des catégories plus « alternatives » : un webzine (Bakchich.info), un site participatif (Rue89), un agrégateur spécialisé dans les blogs (PaperBlog).

Ainsi toutes les catégories de site ou presque étaient représentées. Dans l'idéal, le spectre d'observation aurait pu être complété avec des enquêtes auprès de blogueurs en tant que tels et de sites étrangers (Québec), mais ces enquêtes n'ont pu être intégrées à la recherche IPRI pour des raisons d'emploi du temps des chercheurs concernés.

En revanche, une opportunité d'enquête auprès de la rédaction numérique d'un quotidien national d'information générale et politique (Libération), sous-catégorie de site parmi les plus centrales dans la production de l'information d'actualité sur le web, a pu être saisie.

Dans cette partie, nous livrerons des éléments descriptifs concernant les rédactions observées spécifiquement au cours de cette année 2008 de la recherche IPRI : Libération.fr, Bakchich.info, Rue89, et Paperblog.

Pour chacune de ces structures, les méthodes d'enquête mobilisées ont mêlé des entretiens semi-directifs auprès de responsables éditoriaux et de rédacteurs avec des observations ethnographiques des conditions de production de l'information dans les salles de rédaction. Nous avons ainsi pu obtenir des données empiriques de première main concernant des questions importantes au sujet de l'évolution des modes de production de l'information (Neveu, 2001) : Selon quels modes et quels critères se divise et se coordonne le travail entre les journalistes ? Quelle est la place de la technique dans la définition du projet et des rôles éditoriaux, la distribution des postes de travail ? Quelles sont les conventions, les routines professionnelles à l'œuvre dans les rédactions ?

2.2 Modalités de production observées

Sous la forme de compte-rendus d'observation, sont dressés pour chaque site les conditions de l'enquête, une description détaillée des modalités de production, ainsi qu'une synthèse des principaux constats effectués.

2.2.1 Libération.fr : atouts et contraintes d'une grande entreprise de presse.

Chercheur impliqué: Franck Rebillard.

Date et lieu de l'observation et des entretiens

Vendredi 28 novembre 2008, 9h35-16h45.

Siège de Libération : 11 rue Béranger - 75003 PARIS

Lieux de production

Libération occupe une partie du rez-de-chaussée (accueil), et surtout les étages 4 à 7 de l'immeuble du 11, rue Béranger.

La rédaction de Libération.fr est installée au 6^{ème} étage, dans un *open space* où elle côtoie la Direction de la rédaction (séparée uniquement par une vitre sur un côté) ainsi que les services de l'édition, de l'iconographie, de la fabrication, et du prépresse. L'Internet apparaît ainsi comme une composante rédactionnelle de l'entreprise, à la fois proche de la direction et simultanément située « en bout de chaîne » (Internet comme maillon de la diffusion autant que de la création des informations), en comparaison des autres composantes de la rédaction (Politique, International, Sport, etc.) qui sont placées à d'autres étages.

À noter :

- Le site de Libération.fr affiche 6 onglets (La Une, Labo, Rebond, Ecrans, Next, Voyages) ainsi que des éditions régionales. Seuls les personnels réalisant La Une (en fait, le journal numérique) et Labo sont regroupés physiquement au 4^{ème} étage, au sein de la rédaction de Libération.fr. Les autres parties du site sont réalisées en dehors de ce périmètre spatial (et organisationnel). Par exemple, Ecrans est alimenté en contenus par des journalistes qui sont logés à l'étage des services Médias et Culture du journal papier, tandis que ce sont les correspondants de Libération en région qui alimentent les éditions régionales (en fait, blogs locaux) en contenus, en plus des articles qu'ils peuvent écrire pour l'édition papier.

- Le studio d'enregistrement du LibéLabo (partie multimédia, surtout sonore, du site Libération.fr) se situe au 4^{ème} étage, à côté du laboratoire de développement photo (chambre noire) et du service de Documentation. Le studio est constitué d'une table avec micros (3-4 places pour des débats ou plateaux d'invités), de caméras légères, et d'outils de montage. Les deux permanents du LibéLabo ont un bureau au sein de la rédaction de Libération.fr.

Effectifs de Libération.fr

- Rédacteur en chef : Ludovic Blécher (absent le jour de l'observation)
- Rédacteur en chef adjoint : Olivier Costemalle
- Responsable LibéLabo : Florent Latrive (à son bureau l'après-midi, au studio ou dans d'autres services le matin)
- Chef d'édition (Éditeur Web) : Richard Poirot (absent le jour de l'observation)

- Informaticien : René Saget (absent le jour de l'observation), qui s'occupe en fait des statistiques, des bannières, et des LibéVilles.
- Responsable blogs : Marie-Dominique Arrighi

- Rédacteurs Libération.fr : Cordélia Bonal, Alexis Danjon + 5 autres permanents (au moins un absent le jour de l'observation).

Les 7 rédacteurs se répartissent entre les spécialités suivantes : Société (2 rédacteurs) ; Monde (1 rédacteur) ; Économie (1 rédacteur) ; Politique (1 rédacteur) ; Sport (1 rédacteur) ; Environnement (1 rédacteur).

- Rédacteur LibéLabo / Reporter audio et vidéo : Hervé Marchon

+ Chargé prise de vue et de son : David Ctiborsky

- Chargé diaporama et assistance webmastering : Dominique Poirer

- Chargée gestion des *chats* : Annie Parise

A noter : Ci-dessus, organigramme des personnels travaillant pour réaliser les onglets La Une et Labo. Il faudrait pour être complet y rajouter les personnels réalisant des contenus pour les autres onglets (notamment les journalistes titulaires de Ecrans), l'ensemble formant le service des « Éditions électroniques » disposant d'une enveloppe budgétaire propre selon Hervé Marchon.

D'un point de vue économique et organisationnel, on peut aussi inclure les personnels des services informatiques de Libération, sollicités pour gérer l'infrastructure technique du site web et notamment le basculement automatique des articles de la version papier sur le site.

Processus quotidien de production

- 6h : Alimentation du site à distance, par un permanent.

Un journaliste de la rédaction web, depuis son domicile, sélectionne 4 articles de l'édition papier du jour (les articles de l'édition papier ne sont pas intégralement reproduits sur le web pour éviter la supposée cannibalisation du papier par le web), et confectionne des articles à partir des fils d'agences (AFP en particulier) et des journaux radios. Ceci jusqu'à 10h-10h30.

- 9h / 9h30 : Réunion du service Web

Réunion du service web pour décider du « Homier » (composition de la Home Page ?), en rebondissant éventuellement sur les articles de l'édition papier sélectionnés préalablement, et attribution des tâches à chacun pour la journée.

-10h : Comité de rédaction de Libération

Réunion du Comité de Rédaction, au 7^{ème} étage, entre les directeurs de la rédaction, les responsables de rubriques, et les responsables de l'édition.

De tous les chefs de services, c'est le responsable du service web qui s'exprime en premier. C'est une nouveauté tout à fait notable : il y a encore quelques mois, le responsable du service web était le dernier des chefs de service à prendre la parole. Après discussion sur ce point avec Olivier Costemalle, il apparaît que ce changement traduise une meilleure intégration du web dans le continuum de production de l'entreprise. Des rebonds sur l'actualité entre papier et web mais aussi entre web et papier sont désormais envisagés, dans les deux sens donc.

A noter :

Une réaction lors du Comité de rédaction a montré les réticences toujours vives des professionnels du papier vis-à-vis de l'ouverture offerte par le web. Olivier Costemalle annonce au Comité de rédaction que l'un des blogs de photo a reçu un prix, pour s'en

féliciter. La réaction de ses collègues du comité de rédaction a été de déplorer que le lauréat était « *un extérieur* » (non membre de la rédaction).

-17h : Basculement des articles de l'édition papier

De façon automatisée, tous les articles de l'édition papier sont rendus disponibles pour le web. La rédaction web les insère au sein des différentes rubriques, en fixant notamment leur hiérarchisation vis-à-vis des autres articles (ceux produits par le service web au cours de la journée).

Selon Olivier Costemalle, l'information web connaît trois pics de consultation, calqués sur les horaires de travail au bureau : arrivée le matin, pause déjeuner, sortie du bureau. Et, toujours selon Olivier Costemalle, les gens n'achètent plus les journaux papier (sauf Le Monde) à partir de la fin de l'après-midi, et le risque de cannibalisation du papier par le web s'estompe à partir de cet horaire. Le déroulé chronologique de l'activité du service web paraît ainsi calqué sur cette connaissance, en partie empirique et intuitive, de la consommation médiatique (web et papier).

Constats principaux

On retrouve dans cette observation de Libération.fr plusieurs des traits déjà entrevus lors des études réalisées auprès de rédactions web de médias traditionnels :

- Modalités de travail très différentes entre journalistes papier et journalistes web (lors d'un entretien avec Alexis Danjon, collaborant à la rubrique Sports de l'édition papier, celui-ci explique qu'on peut aller de 10 minutes pour la confection d'un article web reposant sur du « *bâtonnage de dépêches* » jusqu'à un travail de 6 heures pour la réalisation d'un article papier).

- Version web qui trouve progressivement ses complémentarités avec l'édition papier, mais après moult tâtonnements (ex : horaire de basculement des articles de l'édition papier), malgré des réticences encore présentes de la part des rédacteurs papiers (travail d'évangélisation du service web auprès des journalistes papier via les blogs, stigmatisation de l'« *extérieur* » lors du Comité de Rédaction, etc.), et en étant marquée par les expérimentations antérieures (aspect *patchwork* du site de Libération.fr, agrégeant LibéLabo et Ecrans, par exemple).

2.2.2 Bakchich.info : un webzine dans la filiation des radios libres ?

Chercheurs impliqués: Béatrice Damian-Gaillard, Nikos Smyrnaio.

Date et lieux de l'observation

Jeudi 16 octobre 2008, 10h-19h30.
11 bis rue Euryale Dehaynin – 75019 PARIS

Lieu de production

La rédaction est hébergée à titre gracieux dans les locaux d'une société de production audiovisuelle (fiction et films publicitaires) nommée Kino et occupe un rez-de-chaussée disposé en U.

La rédaction se répartit en deux espaces : une salle multimédia et une salle vidéo. S'ajoutent à ces deux espaces un lieu de détente avec des fauteuils, un canapé et un « coin » café, et le long du couloir qui mène à la partie commune les bureaux des responsables éditoriaux : Nicolas Beau (directeur de la rédaction), Xavier Monnier (directeur adjoint de la rédaction) et Laurent Léger (rédacteur en chef).

Dans la salle multimédia, nous trouvons :

- Des affiches sur les règles de l'editing (pour chaque article : titre descriptif, sous-titre, chapô, liens internes et externes, intertitres et mots-clés).
- Le planning de la rubrique *Le guichet des posts*, des chroniques.
- Les codes vidéos (DailyMotion, YouTube).
- Statistiques Bakchich : pages vues, visiteurs uniques.
- L'affichage de dessins humoristiques sur les fondements de l'information indépendante (« se fendre la margoulette », « le mieux est l'ennemi du bien », etc.).
- Une bibliothèque : ouvrages sur la politique, la culture, le monde, la société...
- Deux postes de travail avec un ordinateur fixe chacun (un poste sur lequel s'effectue la mise en page du journal papier, un poste dédié au journaliste en charge des dessins). Les autres journalistes sont équipés d'ordinateurs portables personnels (et non payés par le journal).

Effectifs

Les fonctions de direction semblent être attribuées selon le parcours professionnel de chacun et sa place dans le journal (son rôle, son importance stratégique) :

- Xavier Monnier (directeur adjoint de la rédaction et en charge du sport), fondateur du site et garant de sa ligne éditoriale, reste le seul fondateur présent dans la rédaction. Ils étaient trois à l'origine : un est parti au Monde diplomatique, l'autre a rejoint une société d'audit en lien avec ses études de commerce.
- Nicolas Beau (directeur de la rédaction), ancien journaliste au *Canard enchaîné*,
- Laurent Léger (rédacteur en chef) a passé 12 ans à Paris Match, travaille sur les enquêtes de « long cours ».

Ils sont aidés dans leur tâche d'administrateur par un enseignant de l'université de Cergy-Pontoise, Éric de Saint-Léger, agrégé de gestion, directeur de département. Il est nommé à la fonction de directeur général de Bakchich. Le soutien logistique est assuré par un conseiller technique : Vincent Caron, et un développeur venu du site *De source sûre*.

L'équipe, composée d'une dizaine de personnes, comptabilise 50% de journalistes diplômés d'une école de journalisme reconnue par la profession, et 50% de

rédacteurs non formés dans une école reconnue. Selon Xavier Monnier, l'avantage de cette équipe est qu'elle n'a pas ou peu expérimenté le formatage de la presse écrite, du fait de la jeunesse de ses membres et de leur parcours. Cet état de fait faciliterait la polyvalence des journalistes vis-à-vis des thématiques et des sujets traités. Par ailleurs, si la parité est atteinte dans la rédaction multimédia, et, ce depuis le mois de mars 2008 en termes de nombre de postes occupés proportionnellement par les journalistes hommes et femmes, elle n'est pas la règle en ce qui concerne les fonctions à responsabilité éditoriale. Il est à noter toutefois davantage d'hommes chez les pigistes, du fait notamment du poids en nombre de la catégorie des dessinateurs.

Les journalistes de la rédaction sont :

- Sébastien Fontenelle (responsable vidéo),
- Anne Giudicelli (spécialiste du monde arabe et musulman),
- Catherine Graciet (chef du service international),
- Maïté Labat (santé),
- Eric Laffitte (affaires politico-judiciaires),
- Marion Mourgue, (information politique)
- Simon Piel (médias et responsables du vivier des dessinateurs),
- Bertrand Rothé (économie),
- Anaëlle Verzaux (culture et société).

Processus de production

L'affectation de spécialités journalistiques est fondée sur le rubriquage du site (international, santé, consommation, culture, affaires politico-judiciaires, sport...). Les affectations se décident selon les profils de formation, les contacts, les parcours et les centres d'intérêt des journalistes. Toutefois, l'organisation du travail couple à cette distribution des spécialités journalistiques un dispositif d'organisation de certains postes de travail :

- Chaque journaliste est l'interlocuteur d'un vivier de pigistes en lien avec sa spécialité journalistique (réfèrent thématique). Il ou elle a en charge de récupérer les papiers des pigistes, de les transmettre au responsable *editing*.
- La fonction de responsable *editing* se fonde sur un poste de travail tournant chaque semaine. Ce poste est dédié à la mise en ligne de l'information sous la direction de N. Beau, directeur de publication. Une modération intervient *a priori* avant la mise en ligne des articles et il a la responsabilité des activités de travail de back office telles que la vérification de la mise en page, la création des mots-clés, la mise en place des liens internes et externes. Ce responsable est également chargé du suivi de l'avancée des papiers des pigistes (relances). Il cumule donc les activités de responsable de l'*editing* et de secrétaire de rédaction. Pour la gestion des papiers produits par les pigistes, il ou elle peut s'adresser au réfèrent thématique. Il ou elle se charge également des forums, avec l'aide d'autres journalistes les week-end et en semaine (généralement le stagiaire).
- Le responsable du *guichet des posts* alimente une rubrique écrite tous les deux jours qui consiste en un commentaire construit à partir des commentaires des internautes. Tous les journalistes tournent sur ce poste (y compris stagiaire), .
- Il existe également un pôle vidéos, constitué de 3 à 4 personnes sous la direction d'un rédacteur en chef spécifique. Elles sont en charge de la production de vidéos. La plupart de ces journalistes viennent du site *De source sûre*, récemment racheté par Bakchich, et dont le nom apparaît encore sur le portail Bakchich.TV.

L'organisation du travail des journalistes de Bakchich.info se structure autour de quatre conférences de rédaction, auxquelles tous ne sont pas conviés et dont certaines sont quotidiennes alors que d'autres sont hebdomadaires. Ainsi :

- Le lundi, le responsable *editing* de la semaine (un journaliste permanent) réunit les pigistes (une quinzaine) et note les informations sur qui fait quo, et prend éventuellement des contacts.
- Le matin se tient, aux alentours de 10h30 une conférence de rédaction entre les permanents de la rédaction.
- L'après-midi, les journalistes se retrouvent à nouveau pour faire le point sur les sujets, les ajuster si nécessaire, et les hiérarchiser.
- La responsable de l'information internationale a une réunion le jeudi à 11h avec ses propres pigistes.

La production quotidienne est plus difficile à établir que pour le cas de Libération.fr, dans la mesure où certaines productions se réactualisent tous les jours alors que d'autres s'étalent sur une durée hebdomadaire. Ainsi, sur une semaine type, la production globale se répartit de la façon suivante :

- 5 papiers de Une sont réactualisés par jour dont une vidéo.
- 1 « coup de gueule » est rédigé par jour par un journaliste qui souhaite s'exprimer sur une question en lien avec l'actualité.
- 1 chronique est rédigée par jour, sans journaliste spécifiquement assigné à ce type d'articles.
- 3 à 5 vidéos sont produites par semaine.

Ainsi, nous pouvons comptabiliser en moyenne 7-8 productions inédites par jour pour une équipe d'une dizaine de permanents (à laquelle s'ajoute une quinzaine de pigistes). Dans ces conditions, la Une reste stable en moyenne sur une journée/une demi-journée. Les papiers de la vieille descendent au fur et à mesure ainsi que les vidéos. Il est à noter que les papiers peuvent être mis en ligne au fur et à mesure de leur rédaction, à la condition d'être validés. Toutefois, la mise en ligne peut durer jusqu'à tard dans la nuit, jusqu'à minuit.

Cette organisation flexible, en termes de postes de travail mais aussi de temps de travail, se couple à une absence de formalisation des statuts, des règles structurant cette polyvalence et les obligations professionnelles. Ce mode de fonctionnement crée certes de la souplesse dans la coordination du travail des journalistes, mais, revers de la médaille, ce flou engendre également des tensions, des conflits sur la répartition de la charge de travail supportée par le collectif. Par exemple, lors de notre observation, les journalistes ont demandé une réunion ad hoc pour repenser le *desk* (la répartition des postes et de la charge de travail). Une première réunion informelle a eu lieu entre la rédaction et X. Monnier (directeur en chef adjoint), puis, une seconde lorsque Nicolas Beau (directeur de la rédaction) est arrivé dans les locaux.

Constats principaux

Ce site d'information en ligne se singularise par un projet éditorial centré sur un mode de traitement de l'information : celui de l'enquête, couplé avec un ton qualifié par le directeur adjoint de la rédaction de « *mauvais esprit* ».

Ainsi, ces choix éditoriaux s'incarnent à la fois dans une volonté de traiter de sujets non traités par les autres journaux (en ligne ou papier) et dans la présence à l'écran de nombreux dessins, ou encore la prédominance accordée (notamment en Une) aux

affaires politico-financières. Ces éléments constitueraient les fondamentaux de la marque Bakchich.

Suivant cette logique journalistique, la rédaction n'est pas abonnée au fil des dépêches de l'AFP. De fait, les journalistes entretiennent des rapports suivis avec des informateurs choisis en fonction de leur spécialité journaliste au sein d'instances officielles (champ politique, administratif, économique, médiatique, culturelles, etc.), mais également dans des espaces sociaux moins institutionnalisés.

La nécessité de se démarquer par rapport aux autres acteurs de l'information en ligne ou sur papier est d'autant plus vitale que ce site n'est pas adossé à un groupe de presse ou de communication. En tant que structure indépendante, il doit supporter seul les coûts de production de l'information, dont la seule masse salariale s'élève à 60 000 euros par mois. Il ne peut donc pas mettre en œuvre des stratégies d'économies d'échelle telles que des synergies organisationnelles ou financières. De même, il ne bénéficie pas d'une reconnaissance supportée par l'image de marque d'un groupe de presse traditionnelle ou d'un journal papier. Cette absence de reconnaissance auprès des sources d'information, des investisseurs, des autres médias, des publics, des annonceurs a motivé en partie le recrutement du directeur actuel de la rédaction, ancien journaliste « historique » au *Canard Enchaîné*. La notoriété de cette figure du journalisme satirique, ainsi que sa connaissance du milieu médiatique et ses relations dans les espaces journalistique, économique, politique, devraient servir à accroître celle de Bakchich.info. Dans la répartition des activités attachées aux fonctions de direction, le directeur de publication est en charge du développement des relations publiques (actionnaires et autres médias), alors que le directeur adjoint de la rédaction se concentre sur l'encadrement des rédactions (papier et Web).

De cette capacité à susciter de la reconnaissance par la production de scoops dépend la survie de l'organisation, SA au capital de 37 000 euros, et qui a dépensé depuis sa création 1 million d'euros. En juin 2008, de nouveaux financiers avaient apporté leur contribution à une recapitalisation, siégeant dès lors à un conseil de surveillance. Au moment de l'observation ethnographique de la rédaction en novembre 2008, l'état de la trésorerie ne permettait pas le paiement des salaires d'une dizaine de journalistes permanents. C'est certainement dans une volonté de retirer quelques revenus de la production journalistique que, depuis le 8 décembre 2008, les *Off de Bakchich*, c'est-à-dire les informations en exclusivité, sont devenues payantes, à raison de 35 euros par an ou de 10 euros par trimestre.

Dans ces conditions, dicit le directeur de la publication, « *l'important n'est pas de produire de l'information, mais de se battre pour être repris* ». Il faut non seulement trouver des sujets non traités par ailleurs, mais que ces sujets suscitent des visites uniques, provoquent des reprises à la fois par des internautes sur leurs propres *blogs*, mais aussi des médias *mainstream* en citant la source Bakchich. Ce faisant, le site deviendra plus visible dans les moteurs de recherche et sera plus à même d'attirer des capitaux. En ce sens, une offre publicitaire commune a été mise en place en septembre 2008 entre Bakchich, qui venait de racheter *De source sûre*, *AgoraVox* et *La télé libre*. Par cette alliance, les quatre sites en ligne cumulaient une audience d'au moins un million de visiteurs uniques, seuil d'accès à la reconnaissance des acteurs du marché publicitaire.

2.2.3 Rue89 : le « participatif encadré »

Chercheur impliqué: Franck Rebillard.

Date et lieu de l'observation

Jeudi 27 novembre 2008, 10h15-13h30, puis 14h30-16h.
Siège de Rue 89 : 80, rue des Haies – 75020 PARIS

Lieu de production

- L'entreprise Rue89 est installée au 20 rue des Haies, Paris 20^{ème} (quartier en restructuration) avec d'autres sociétés travaillant dans le domaine du multimédia (5 à 10 sociétés), dont notamment PaperBlog. Espace fermé, dénommé « Pépinière contenus numériques » (financeur principal : Ville de Paris).
- Locaux de Rue 89 relativement exigus pour la douzaine de personnes qui y travaillent. Une seule pièce, rectangulaire, *open space*, sans absolument aucune séparation physique entre les différents bureaux : « *esprit start-up* », dicit Arnaud Aubron.
- Au site principal Rue89 est accolé un site secondaire Eco89 (spécialisé dans l'actualité économique comme son nom l'indique) : les rédacteurs de Rue 89 et ceux de Eco89 partagent les mêmes locaux, sachant qu'ils sont même appelés à intervenir sur l'un ou l'autre site à l'occasion (rédaction d'articles et commentaires).

Effectifs de Rue 89

- Président, Directeur de la publication : Pierre Haski (souvent en déplacement, « *en représentation, à des colloques* », dicit Arnaud Aubron)
 - Directeur Général : Laurent Mauriac
 - Assistante Édition et Direction : Corinne Antoine
- Rôle-clé dans la structure, puisque gérant l'administratif au quotidien ainsi que l'interactivité ordinaire du site (tri des courriels envoyés à la rubrique « Contacts » et qui signalent notamment des commentaires à caractère illicite).
Physiquement, l'assistante de direction se situe au centre de la pièce, à proximité des trois dirigeants fondateurs présents en permanence (Mauriac, Aubron, Riché).
- Directeur Technique : Damien Cirotteau
 - Développeur informatique : Sébastien Foll
-
- Webmaster-Editeur et DRH : Arnaud Aubron
 - Webmasters éditoriaux : 2 permanents (Yann Guégan – incluant partie GRH - et Marie-Sophie Keller) + une stagiaire (formation continue après des années de SR dans la presse écrite)
-
- Rédacteur en chef : Pascal Riché
 - Rédacteurs Rue89 : 5 permanents (Julien Martin, Guillemette Faure, Zineb Dryef, David Servenay, Chloë Leprince)
 - Rédacteurs Eco89 : 2,5 permanents dont Sophie Verney-Caillat et François Krug (+ Augustin Scalbert à mi-temps ?)
- Au total, selon Arnaud Aubron, la rédaction est composée de 5 rédacteurs Rue89 + 2,5 rédacteurs Eco89 + Aubron partiellement + Riché + Haski + 1pige/jour environ + 5

blogueurs « forfaitisés » + non-journalistes autres = production quotidienne moyenne de 15 articles/jour dont 3 articles issus de non-journalistes.

A noter :

- Cinq fondateurs au départ (printemps 2007), anciens de Libération : Haski, Mauriac, Aubron, Riché + Michel Lévy-Provençal qui est parti en février 2008 chez France 24 (Responsable Nouveaux Médias, service « Les Observateurs – Vos yeux aux quatre coins du monde »).

Raisons du départ de Michel Lévy-Provençal données sur son blog : il estime que le projet initial de Rue89 d'une info à trois voix a été trahi, le site se restreignant de plus en plus au journalisme professionnel et n'utilisant le participatif que comme alibi marketing...

- Pas de commerciaux parmi les permanents. L'activité publicitaire est sous-traitée à la régie de Hi-Media, éditeur de sites web (sport, actu *people*, etc.) revendiquant 30 millions de visiteurs uniques au total, avec lequel Rue89 a récemment noué des liens capitalistiques.

Processus de production

Le travail journalistique à Rue89 se déroule largement au bureau (même s'il y a aussi des déplacements, notamment pour des prises de vue ou de son), mais ce n'est pas pour autant un travail de *desk* au sens habituel où on l'entend (appui sur les dépêches et autres communiqués de presse). Pour chaque sujet abordé, les journalistes de Rue89 passent du temps sur leurs papiers, enquêtent et souhaitent avoir des sources propres.

Surtout, les membres de Rue 89 sont invités à passer beaucoup de temps dans la gestion des interactions avec les internautes, puisant auprès de ces derniers des propositions de sujets (que les journalistes se chargent le plus souvent de réaliser eux-mêmes).

Les webmestres éditoriaux supervisent la gestion des commentaires des internautes, en collaboration avec les journalistes :

- 1^{er} niveau : la modération des commentaires, les réponses que les membres de la rédaction peuvent apporter. A ce niveau, la modération (et non la sélection) est assurée par l'ensemble de l'équipe, pas uniquement par l'auteur de l'article commenté.

- 2^{ème} niveau : la sélection et la mise en avant de commentaires, choisis par les journalistes. Ce sont ces commentaires (2 ou 3 en général) qui seront placés immédiatement à la suite de l'article, au sommet de la longue liste de l'ensemble des commentaires.

- 3^{ème} niveau : Certains commentaires sont mis en valeur sous forme d'article « Vos Réactions ». Parfois même, un article est écrit à partir d'une revue de commentaires compilant 10 ou 12 commentaires.

- 4^{ème} niveau : L'auteur d'un commentaire peut être relancé par la rédaction pour la production d'un contenu voire d'un blog.

La nécessité de médiation des journalistes professionnels pour valider l'information est toutefois réaffirmée. Rue89 = « *participatif encadré* », « *utiliser les sources données par les lecteurs, mais tout est contrôlé par la rédaction* », « *Les internautes ne peuvent pas proposer des enquêtes ou des articles validés, car c'est un métier ;*

mais des tribunes, de l'opinion, des expertises, des témoignages, oui. » (Pascal Riché).

Il y a une conférence de rédaction hebdomadaire le vendredi, à 11h. Chacun propose 3 ou 4 sujets, et la façon de les traiter est discutée collectivement.

Parallèlement à cette conférence de rédaction, des sujets sont choisis au jour le jour, avec consultation des autres rédacteurs et principalement d'Arnaud Aubron et de Pascal Riché. Idée de « *bouclage permanent* » (Sophie Verney-Caillat)

Les rédacteurs d'Eco89 peuvent être amenés à également rédiger des articles pour Rue89. L'un des journalistes est d'ailleurs un peu à cheval entre les deux éditions. Plus largement, les journalistes doivent assurer des permanences le week-end (et s'il s'agit d'un journaliste normalement plutôt affecté à Eco89, il devra le week-end travailler tout autant pour Rue89).

« *Il y a des spécialisations, par exemple entre Politique et Société, mais ce n'est pas très tranché* » (Arnaud Aubron). Les rédacteurs, même s'ils sont plus ou moins spécialisés dans un domaine, sont également appelés à couvrir d'autres domaines. Avec en filigrane, la volonté de ne pas tomber dans une routine sclérosante, comme dans certaines rédactions papier.

Constats principaux

Les journalistes présentent des propriétés sociodémographiques relativement classiques (notamment dans le cursus : sciences-po + école de journalisme), mais s'adonnent à un traitement de l'information un peu différent (en l'occurrence, appui sur les réactions des internautes).

Par ailleurs, certaines contraintes économiques semblent constituer simultanément (et involontairement ?) des moyens de différenciation médiatique :

- Non-abonnement à des fils d'agence selon les motifs suivants : pour des raisons économiques (trop coûteux), mais aussi pour « *échapper à l'agenda médiatique* ».
- Gestion de l'interactivité en interne (animation et commentaires) et non sous-traitée à des sociétés spécialisées comme Concileo, là encore pour des raisons économiques mais aussi pour se différencier des concurrents par la prise en compte des réactions du public (participatif = stratégie d' « *image* » explicite).

2.2.4 PaperBlog : tirer partie de la richesse du web, à peu de frais.

Chercheur impliqué: Nikos Smyrnaio.

Date et lieu de l'observation

14 octobre 2008, 15h-17h.

Siège de Paperblog : 80, rue des Haies – 75020 PARIS

Lieu de production

- Emplacement voisin de celui de Rue89 : Paperblog est installé au 20 rue des Haies, Paris 20^{ème} (quartier en restructuration) avec d'autres sociétés travaillant dans le domaine du multimédia (5 à 10 sociétés). Espace fermé, dénommé « Pépinière contenus numériques » (financeur principal : Ville de Paris).
- Locaux de Paperblog relativement exigus. Une seule pièce, rectangulaire, *open space*, sans séparation entre les différents bureaux, à peu près la moitié de la superficie des locaux de Rue89.

Effectifs de Paperblog

- Fondateur et directeur : Nicolas Verdier, 26 ans, diplômé d'École de commerce, ancien blogueur.
- Responsable éditoriale : Johara Boukabous, 26 ans, diplômée en Sciences politiques de l'Université de Marne la Vallée (Master 1 et 2), blogueuse régulière.

L'équipe inclut également une autre journaliste, une responsable de communication et des ingénieurs : en tout entre 5 et 8 personnes selon le moment (4 CDI et les autres en contrats précaires).

Paperblog revendique 5.000 auteurs, 2.000 nouveaux articles par jour, 2 millions de visiteurs uniques par mois selon la mesure *site-centric*, 1,3 selon Nielsen Médiametrie - c'est ce dernier chiffre qui compte auprès des annonceurs -, il occupe la 17^{ème} place parmi tous les sites d'information consultés depuis la France.

Processus de production

Il n'y a pas de production de contenu à proprement parler chez Paperblog, mais plutôt la reprise organisée et automatisée de contenus en provenance de milliers de blogs différents. L'objectif du service tel qu'il est défini par son fondateur est de rendre accessible au grand public la production très riche des blogueurs francophones. Paperblog est donc un infomédiaire entre les lecteurs et les « bons » contenus qui sont identifiés, valorisés et mis à disposition.

Les deux opérations effectuées par le service sont essentiellement la « thématisation » du contenu (ranger les articles au sein de différentes catégories) et sa « qualification » (évaluer le contenu en termes de qualité). Pour ce faire, dans un premier temps, il y a la constitution d'une base de sources d'où sont aspirés les contenus (par différentes méthodes mais essentiellement en utilisant les fils RSS). Ensuite ces contenus sont contrôlés en fonction de plusieurs critères : l'orthographe, la taille (la longueur idéale étant de 300 à 400 mots), la présence d'illustrations, la date de publication et l'historique de l'auteur. Enfin, au sein du site il existe également

des classements en fonction des critères de popularité (les articles les plus lus, commentés, etc.). Le processus de thématisation est automatisé à 60% sur la base d'une analyse lexicale et de l'utilisation des tags (mots-clés attribués à leurs articles par les auteurs). L'ensemble des technologies utilisées, qualifiées de « sémantiques », sont propriétaires. Le tout est complété par l'intervention de l'équipe éditoriale qui choisit les articles de la Une et de la rubrique « Coup de cœur » et des dossiers thématiques.

Les deux journalistes sont des jeunes personnes, pour lesquelles c'est le premier emploi. Leur travail consiste en un *mix* entre le travail d'un rédacteur en chef et d'un secrétaire de rédaction.

Le travail de la responsable éditoriale consiste à sélectionner les meilleurs articles et à les mettre en valeur. En pratique, elle passe sa journée devant l'ordinateur à surveiller le *BackOffice* du site et voir en temps réel les articles en provenance de blogs qui sont repris par Paperblog. Sur les 2 000 articles qui arrivent par jour, elle en lit quelques dizaines, parmi lesquels elle note les plus intéressants. A partir de ce premier tri, elle sélectionne les quatre contenus qui font la Une du site ainsi que d'autres qu'elle fait entrer dans la rubrique Coup de cœur. Pour ce faire, elle a déjà en tête un ensemble de sites qu'elle sait publier régulièrement sur des sujets « *qui plaisent* ». Par ailleurs, elle dispose d'une sorte de planning des thématiques (sport, divertissement, culture) qui font la Une chaque jour de la semaine. Par exemple, la veille du mercredi, elle essaye de mettre en avant un sujet lié au cinéma (critique de film ou autre), en fin de semaine elle choisit des sujets sur les sorties (critique restaurant, concert etc.). Il s'agit aussi de « fidéliser » le lecteur.

Globalement, par la sélection de la Une, la responsable éditoriale essaye de se démarquer de l'actualité dominante. Le raisonnement étant que les événements majeurs sont évoqués par les grands médias et donc sont connus des internautes. Ainsi, elle choisit de mettre en avant des sujets relativement décalés. Par contre, pour approfondir l'actualité chaude (crise financière, mort de Guillaume Depardieu etc.) sont constitués des dossiers thématiques où sont regroupés les articles correspondant. Certaines fois la valeur ajoutée de la Une provient du goût personnel de la responsable éditoriale, par exemple pour la musique africaine, ce qui peut aboutir à ce que des informations relativement pointues ou spécialisées trouvent en Une. Dans l'ensemble, Johara Boukabous détient la prérogative éditoriale mais semble aussi se tourner vers Nicolas Verdier quand il s'agit de prendre des décisions importantes.

Constats principaux

- Malgré le caractère innovant du site, l'aspect participatif est peu développé techniquement et surtout au niveau de l'engagement des utilisateurs. Selon le fondateur du site, ce point faible est dû aux moyens humains très limités pour animer et structurer les communautés.
- Paperblog est très dépendant de la production des blogueurs. Cependant, les relations du service avec eux sont complexes. D'une part plusieurs milliers d'auteurs diffusent leur production via Paperblog, d'autre part beaucoup en sont partis, souvent en mauvais termes. Le problème provient de la particularité de Paperblog qui, contrairement à d'autres sites comme Wikio, Fuzz ou même Google Blogs, reprend non seulement les titres et chapô des articles mais aussi l'intégralité du contenu. Le reproche fait par de nombreux blogueurs à Paperblog est d'exploiter leur contenu pour être mieux référencé que les blogs eux mêmes.

- L'ambiguïté de Paperblog est forte à la lecture des conditions d'utilisation que les blogueurs acceptent pour voir leurs contenus publiés. Il s'agit d'un contrat par lequel les auteurs cèdent en fait le droit patrimonial de leur production à Paperblog et permettent donc à celui-ci de l'exploiter comme il l'entend, dans les conditions définies par la loi (durée de protection, conditions de rémunération etc.).
- L'automaticité des procédés de Paperblog constitue son premier avantage en même temps qu'elle apparaît également comme son principal défaut : le fonctionnement du service est basé sur l'aspiration du contenu en provenance des blogs dans son intégralité et sa republication automatique sur Paperblog sans contrôle des possibles atteintes à la propriété intellectuelle. En pratique, on trouve très souvent sur le site des articles qui comportent des photos, des musiques ou des textes dont les blogueurs ne sont pas propriétaires. Par conséquent, la reproduction systématique de contenus non contrôlés de la part de Paperblog l'expose à la reproduction des infractions au droit d'auteur qu'auraient commis les auteurs (blogueurs) des articles.

2.3 Analyse comparative

Chercheurs impliqués: Béatrice Damian-Gaillard, Franck Rebillard, Nikos Smyrnaiois.

Pour cette analyse comparative, nous avons mêlé des éléments de connaissance issues du programme IPRI mais également du programme « Diversité » de 2007.

Il en ressort que l'opposition entre sites *pure—players* et sites d'information adossés à des structures médiatiques traditionnelles ne peut être complète, du moins du point de vue de l'organisation de la production d'information. Certaines similitudes peuvent être observées, notamment du côté du profil sociodémographique des rédacteurs et de leurs pratiques, qui restent globalement ancrées dans les référents professionnels du journalisme traditionnel (Ruellan, 2007). Parallèlement, des différences significatives existent dans le choix des informations, entre un pôle « productiviste » reposant largement sur le recours à des matériaux produits par des tiers « établis » (dépêches d'agences, communiqués d'entreprises, autres médias) et un pôle « angliste » s'attachant volontairement à des sujets originaux (au moins dans leur mode de traitement) par le recours, notamment, aux ressources offertes par l'internet : productions des blogs et commentaires des internautes.

2.3.1 Des similitudes au niveau des personnels de rédaction.

Si l'on met de côté le cas particulier des agrégateurs dont l'activité est au mieux l'éditorialisation de contenus (contenus qui n'ont pas été produits par eux-mêmes) comme chez Paperblog, les autres sites emploient des rédacteurs qui s'inspirent des modèles professionnels journalistiques pour structurer leurs pratiques. Des modèles proches de ceux mis en œuvre dans les salles de rédaction de journaux papier.

Nous retrouvons les principes fondamentaux enseignés dans les formations en journalisme, censés encadrer le rapport aux sources, les valeurs de vérification de l'information, les formats d'écriture journalistique (les genres journalistiques tels que l'enquête, l'entretien...).

De même, nous retrouvons certains dispositifs d'organisation quant à la division et la coordination du travail des rédacteurs : par exemple, dans trois des quatre sites observés lors de cette année IPRI (*Rue89.com*, *Libération.fr* et *Bakchich.info*), les mêmes modalités de division horizontale du travail apparaissent, qui concernent plus spécifiquement la question de l'attribution de dossiers journalistiques.

Les journalistes sont plus spécifiquement dédiés à une thématique en lien avec des sources d'information spécifiques. Les modalités de leur attribution se fondent sur des principes de légitimité qui s'appuient parfois sur une expertise, constituée à partir d'une formation initiale et/ou d'un parcours professionnel, militant, car certains dossiers sont, en effet, considérés comme plus techniques que d'autres, à l'exemple de la justice ou de la santé. Les principes de légitimité se fondent également sur les "entrées" dans les espaces sociaux concernés (les carnets d'adresse), ou encore sur les centres d'intérêts des journalistes et éventuellement leurs pratiques amateurs (culturelles notamment).

Toutefois, on notera que chez les *pure-players*, une polyvalence plus importante est attendue des journalistes, qui peuvent être incités à traiter de sujets extérieurs à leur domaine de spécialisation, en raison souvent d'équipes de production plus restreintes.

Par ailleurs, il est intéressant de souligner que même si les *pure-players* se défendent de reproduire un traitement de l'information inspiré de la presse écrite, la plupart des

encadrants éditoriaux des trois sites mentionnés sont issus pour la plupart de la presse écrite, et assez symptomatiquement de la presse nationale d'information politique et générale. Le directeur de la rédaction de *Bakchich.info* est un ancien journaliste du *Canard Enchaîné*, son rédacteur en chef a travaillé pendant 12 ans pour *Paris Match*. Les quatre encadrants éditoriaux de Rue89 sont des anciens de Libération. Ces affiliations sont potentiellement des atouts pour la reconnaissance professionnelle des sites auprès des autres médias, des autres journalistes (les pairs), des sources d'information, des investisseurs, des annonceurs. Mais, par les fonctions qu'ils occupent, ces responsables influent aussi sur les définitions des projets éditoriaux proposés à la consommation, sur la structuration des thématiques et des sujets traités, les angles adoptés. De même, ils participent à importer sur le Web les représentations et les pratiques professionnelles défendues dans le champ journalistique et construites principalement en référence à l'histoire de la presse écrite.

Toujours sur ce plan du cursus et des référents professionnels, ajoutons en complément que les journalistes de base, très jeunes, sont pour la plupart issus des écoles reconnues de journalisme. Cet emploi dans une rédaction web constitue souvent pour eux leur première expérience professionnelle significative : leur entrée dans le métier se passe donc à travers ce secteur du web, stimulant par son caractère innovant en comparaison d'autres supports, mais bien décevant aussi par l'écart qu'il procure entre la vision idéale du journalisme d'origine et sa traduction souvent beaucoup plus prosaïque « en ligne » (Estienne, 2007). En effet, l'organisation à la fois spatiale et organisationnelle du travail dans les rédactions web amplifie et accélère des tendances déjà observées dans les médias de presse écrite, à savoir une convergence entre les activités rédactionnelles et celles d'editing (Damian *et al.*, 2002). Le journaliste doit non seulement produire de l'information, mais également la mettre en forme et, enfin, la mettre en ligne. Par ailleurs, la polyvalence des tâches à effectuer, des postes de travail à occuper, d'une part, les lieux de travail agencés en *open space*, d'autre part, renforcent la difficulté pour le journaliste de s'isoler du collectif et de ses interférences pour écrire. En effet, la coordination du travail s'appuie principalement sur des mécanismes d'ajustement mutuel entre les opérateurs et donc, ces mécanismes engendrent une multitude d'interactions croisées portant sur la hiérarchie de l'information, l'arrivée des sujets, les contacts avec les sources...

2.3.2 Des modalités de production alternatives pour des informations originales ?

Les similitudes constatées précédemment entre sites de *pure players* et sites dépendants de médias traditionnels ne doivent en aucun cas masquer les différences entre ces deux types de producteurs d'information. Ces dernières s'expliquent en partie par les spécificités de leurs structures organisationnelles respectives, en lien avec l'histoire et les appartenances de chaque site. Elles tiennent aussi, pour le point qui nous intéresse plus spécifiquement au sein de cette recherche IPRI, à de vraies différences dans les stratégies de couverture de l'actualité.

La gestion du personnel dans des sites tels que *Bakchich.info* ou Rue89 les rapprochent fortement des start-up de la « nouvelle économie » des années 90, avec une direction par objectif ou des procédures de management par projet (Gadrey, 2001). Ces structures se caractérisent par une faible étanchéité entre temps libre et temps de travail, des temps de travail extensifs au nom de l'engagement dans un projet. Par ailleurs, ces organisations ont généralement pour principale

préoccupation leur survie économique, du fait de la prégnance des difficultés économiques. Il est attendu de tous, y compris des encadrants, un engagement total. Il s'objective notamment par une adhésion aux normes et aux valeurs de régulation des relations professionnelles. Cette adhésion est requise comme condition d'appartenance au projet et à l'entreprise. Ainsi, au nom d'un projet, figure de proue du « nouvel esprit du capitalisme » (Boltanski, Chiapello, 1999), dans un contexte de survie -l'entreprise ne sait pas si elle existera dans un mois-, il est attendu, sans discussion, du salarié qu'il « se donne sans compter », qu'il soit toujours joignable et disponible, qu'il ne demande pas une rémunération à la hauteur de son engagement professionnel (en temps, en activités), et qu'il place l'intérêt général de la structure avant celui de son intérêt personnel.

Par comparaison, les rédactions numériques des grandes entreprises de presse et *a fortiori* leur nécessaire coordination avec les rédactions papier (Libération) ou les autres entités du groupe (Lagardère) paraissent souffrir du poids « bureaucratique » intrinsèque à toute organisation de grande taille (Weber, 1971). Le processus de travail est beaucoup plus cadré et moins souple, l'innovation éditoriale se situe dans des niches (exemple : LibéLabo ou studios multimédias dans certaines rédactions de magazines de Lagardère), bref les marges de manœuvre sont plus étroites par rapport au traitement de l'actualité.

Mais il faut dire aussi que les sites de médias traditionnels ont une « obligation » (en tout cas ressentie comme telle par les acteurs concernés) de couverture continue et exhaustive de l'actualité. Ceci pousse à fortement rationaliser la production de l'information, à résoudre l'équation d'une production quantitativement abondante avec des moyens limités. Nous avons décrit cette orientation « productiviste » des sites de médias traditionnels dans notre précédente recherche :

« Pour mettre sur pied des rédactions spécialement dédiées aux éditions numériques des publications, les groupes de presse ont engagé de nouveaux journalistes en sus des équipes existantes. Mais ceci n'équivaut pas à un accroissement proportionnel de la capacité rédactionnelle des publications et donc à une diversité accrue des informations. Car les journalistes spécialisés dans le numérique effectuent un travail qui est rarement un travail de création et beaucoup plus souvent un travail de retraitement ou de réécriture d'informations existantes » (Rebillard et al., 2007, p. 16).

Nous avons retrouvé une situation comparable en 2008 au sein de la rédaction de *Libération.fr*, due à la nécessité pour un site généraliste et relativement *mainstream* de couvrir l'ensemble de l'actualité du jour. Cette recherche d'exhaustivité oblige à produire de façon quantitative, sans avoir forcément le temps de se focaliser sur un angle, une information originale, ou une enquête approfondie.

À l'inverse, les *pure-players* ont délibérément adopté un mode plus discontinu de production de l'information, sélectionnant davantage les sujets à traiter et s'inscrivant dans une temporalité plus étalée.

Paperblog par exemple ne cherche pas à coller à l'actualité chaude, mais s'intéresse plus spécifiquement à des contenus de type magazine (loisir, cuisine, voyage, culture, etc.) qui, selon son dirigeant, comportent un double intérêt. Premièrement, il n'est pas nécessaire, pour ce type de produits éditoriaux, que les journalistes vérifient l'information ; ce qui, du coup, implique un moindre coût de traitement. Deuxièmement, ce sont des contenus dont la valeur économique est diachronique et pas seulement liée à l'actualité immédiate. En clair, il s'agit de ce que nous avons identifié ailleurs comme *information de stock*, sui, comme son nom l'indique, participe à l'enrichissement d'une base de données (Rebillard et al., 2007). Cette information peut générer de l'audience à long terme, essentiellement par le biais de moteurs de recherche. Cette logique de rentabilisation des contenus a des incidences

sur les modalités de production et de hiérarchisation de l'information. Ainsi, considérées comme économiquement peu attractives par Paperblog, les thématiques généralistes telles que *politique* ou *société* sont présentes sur le site, mais rarement mises en avant sur sa page d'accueil.

Dans le même ordre d'idées, le site d'information en ligne *Bakchich.info* se singularise par un projet éditorial centré sur un mode de traitement de l'information fondé sur l'enquête et le « mauvais esprit », que nous avons déjà explicité.

Cette volonté, plus ou moins subie, de ne pas s'appuyer sur les sources habituelles des entreprises de médias au premier rang desquelles les agences de presse et notamment l'AFP en France, s'inscrit plus largement dans une stratégie de positionnement en dehors de l'agenda médiatique. On la retrouve à la fois chez Bakchich, Rue89, mais aussi chez PaperBlog.

Il faut bien être conscient des deux versants d'un tel positionnement. Si les *pure-players* l'ont adopté, c'est parce qu'ils ne pouvaient pas lutter sur ce terrain de la couverture continue et exhaustive de l'actualité face à des entreprises médiatiques beaucoup plus armées sur ce plan. Et c'est aussi parce que leur acculturation à l'internet est beaucoup plus forte. Ici, le recours aux commentaires des internautes ou aux billets publiés sur les blogs tient à la fois de ressource et de contrainte. Ainsi, Rue89 essaie de puiser dans les *mails* et les commentaires de ses visiteurs des idées originales de sujets, c'est-à-dire celles qui n'ont pas encore émergé à la surface de l'actualité, ou qui complètent cette dernière avec un angle ou un élément d'information nouveaux. PaperBlog recherche, dans la multitude des blogs qu'il indexe, des sujets qui ne font pas forcément la Une de l'actualité. Quant à Bakchich, son approche décalée de l'actualité – décalage à la fois temporel et politique – semble constituer son principal trait « distinctif ».

Cette modalité d'un recours « alternatif » à l'internet comme source de la part des *pure-players* - un terrain que, par ailleurs, leurs concurrents issus des médias traditionnels délaissent ou ont délaissé (à l'image de leur sous-traitance généralisée des commentaires et forums, ou de leur hébergement « propriétaire » de blogs) – débouche-t-elle effectivement sur la création d'informations originales ?

Il sera intéressant, en conclusion de cette présentation des travaux de la recherche IPRI, d'examiner l'existence d'une telle corrélation entre les résultats de l'axe « analyse de contenu » et ceux de l'axe « socio-économie de l'information en ligne ».

3. Conclusion provisoire : des conditions de production à la nature des contenus, une interaction continue.

Chercheurs impliqués: Emmanuel Marty, Franck Rebillard, Nikos Smyrnaio.

Pour les résumer à grands traits, disons que les travaux menés au sein de l'axe « analyse de contenu » arrivent à la conclusion générale selon laquelle les informations d'actualité offertes sur le web se caractérisent à la fois par une grande diversité et par une forte redondance. De leur côté, les enquêtes socio-économiques menées spécifiquement auprès des *pure players* de l'internet pointent, au sein de modalités de production qui s'inscrivent dans des référents professionnels relativement homogènes, quelques velléités certaines de développer des alternatives au rythme médiatique moutonnier traditionnellement imposé par les agences de presse et désormais efficacement relayé par les portails et autres agrégateurs.

Une fois posés ces deux mouvements, la tentation est naturellement grande de vouloir établir des liens directs entre eux. Avec cette question qui porte dès lors tous les atours de l'évidence : les informations originales seraient-elles principalement issues des webzines, blogs et sites participatifs du pôle « angliste », tandis que la répétition serait elle à mettre au passif des médias de masse du pôle « productiviste » ?

Nous pouvons trouver, à l'intérieur des différents résultats exposés dans les pages précédentes, des éléments qui vont en ce sens. A commencer par les plus basiques relevés qui mettent en avant la propension des portails surtout, et des sites de médias traditionnels dans une moindre mesure, à multiplier les doublons avec pour seule finalité l'occupation de l'espace médiatique numérique. Mais à ce stade de la réflexion, nous avons souhaité produire une analyse plus fine de cette possible corrélation entre originalité des contenus et nature de la source productrice : si elle confirme en grande partie les intuitions ou impressions éparses laissées précédemment, elle a surtout l'avantage de souligner la complexité de la dialectique entre pluralisme et redondance de l'information sur l'internet.

Pour ce faire, l'analyse lexicométrique des titres a été poussée afin de fournir des indicateurs linguistiques tangibles du positionnement éditorial propre à chaque source, via l'identification d'une production originale de contenu, ou au contraire de reprises de formules stéréotypées.

Après avoir mis en évidence et quantifié la variété et l'équilibre - ou plus exactement le déséquilibre - de l'information, il nous appartient à présent d'en distinguer les acteurs. L'enjeu est d'être en mesure de situer les différentes sources sur un continuum allant des acteurs de la redondance aux acteurs du pluralisme. Pour rentrer dans le détail de la production propre à chaque source, et voir dans quelle mesure chacune pèse sur l'ensemble de l'échantillon analysé, l'outil Lexico a vu ses fonctionnalités mobilisées au-delà du tri initial des différents titres en présence.

3.1 Analyse des segments répétés

Sur chacun des deux échantillons prélevés (6 et 10 novembre 2008), deux analyses ont été effectuées au niveau des segments répétés. La première concerne l'identification des segments extrêmement stéréotypés : le seuil était alors fixé à 10 mots et plus avec une fréquence de 15 occurrences et plus. La seconde est relative à la répartition sur les sources de segments plus courts mais encore relativement

caractéristiques (7 mots et plus), avec une fréquence de 2 occurrences et plus, c'est-à-dire répétés au moins une fois.

Les segments répétés longs à forte fréquence constituent un marqueur tangible de redondance de l'information, correspondant soit à la répétition volontaire d'un titre de la part d'une source, soit à une reprise (au moins partielle) par une deuxième source de la formulation du titre de la première. Ceci apporte un complément d'analyse aux calculs de variété et d'équilibre des sujets, en appréhendant la problématique du pluralisme et de la redondance sous l'angle de la production originale de titres (ou au contraire de la reprise/répétition de sa formulation originelle) et non plus uniquement du point de vue du sujet traité. En d'autres termes, nous ne nous situons plus dans une approche uniquement éditoriale, nous l'enrichissons d'une démarche liée à la manipulation du langage à travers la question des formulations. Il s'agit d'une étape intermédiaire dans notre progression vers l'analyse des particularités des différentes sources.

Dans nos résultats (voir tableaux 9 et 10), il n'est pas étonnant de retrouver dans chaque journée les sujets les plus traités. Mais les seuils choisis, 10 mots d'une part et 15 occurrences d'autre part, nous permettent d'identifier parmi ces sujets ceux qui ont été abordés de la manière la plus redondante sur le plan lexical, et de voir quelle forme textuelle récurrente a été employée.

Tableau 9 : Segments de 10 mots minimum apparaissant au moins 15 fois (6 novembre 2008)

Segment	Longueur	Fréquence
Japon des tissus cérébraux créés à partir de cellules souches	10	28
Obama le premier président noir élu face à des défis colossaux	11	27
perpétuait pour le beau père 30 ans pour la mère	10	23
Les marchés replongent malgré une baisse des taux en Europe	10	21
Enfant mort de coups perpétuait pour le beau père 30 ans	11	19
Obama met en place l'équipe chargée de préparer la transition	11	18
La BCE abaisse ses taux et garde des cartouches pour décembre	11	18
Ligue des champions Lyon en tête de son groupe la Juventus	11	17
Le FMI annonce pour 2009 la 1ère récession dans les pays	11	15
Discrimination la Halde recommande de lutter via les programmes et manuels	11	15

Tableau 10 : Segments de 10 mots minimum apparaissant au moins 15 fois (10 nov. 2008)

Segment	Longueur	Fréquence
attentat à Bagdad fait 28 morts et des dizaines de blessés	11	46
Le Goncourt à l'Afghan Atiq Rahimi le Renaudot au Guinée	11	33
SNCF nouvel acte de malveillance le parquet anti terroriste saisi	10	23
Etats Unis Obama doit rencontrer Bush à la Maison Blanche	10	23
l'UE lance une opération navale historique contre les pirates	10	21
Gilles Simon bat Roger Federer son 2e succès face au Suisse	11	19
les deux camps se font face dans l'est sans affrontement	11	19
Le plan de relance chinois applaudi par des marchés en nette	11	18
Partenariat avec la Russie l'UE prête à reprendre les négociations	11	16
PS Ségolène Royal réunit ses représentants lundi après midi au Sénat	11	15
Le plan de relance chinois et le G20 offrent un bref	11	15

Au regard de ces résultats, on peut affirmer que les sujets occupant le plus d'espace en termes de diffusion et de choix éditoriaux sont également ceux qui sont exprimés de la manière la moins créative au niveau des titres, avec une très forte récurrence de formulations longues similaires (jusqu'à 46 occurrences en un seul jour pour « attentat à Bagdad fait 28 morts et des dizaines de blessés »). On aurait pu penser

que les sujets sur-traités tirent de ce sur-traitement une certaine diversité dans la manière dont ils sont appréhendés et mis en mots : c'est loin d'être le cas. Ces résultats nous permettent de poser l'existence d'une corrélation entre redondance éditoriale et redondance lexicale. Pour le dire autrement, toute information semble d'autant moins riche lexicalement qu'elle est importante éditorialement. L'élection de Barack Obama en constitue un bon exemple : largement sur-traité, ce sujet est principalement abordé à travers la reprise de formules stéréotypées. On peut alors poser l'hypothèse de l'existence d'un lien entre des informations dominantes et des sources dominantes, caractérisées par une intense activité de diffusion d'articles en même temps qu'une forte redondance dans les sujets et leur formulation. Là réside l'enjeu d'une distinction des sources dans les productions textuelles.

3.2 Répartition des segments répétés et des hapax dans les différentes sources

Le calcul des spécificités lexicales pour chaque source repose sur le modèle hypergéométrique de Lafon (1984), qui a abouti à la mise au point de cette fonctionnalité sur le logiciel Lexico. Les spécificités sont exprimées en terme positif, négatif ou nul : ces spécificités correspondent à la fréquence d'apparition d'une ou plusieurs formes lexicales dans une source, au regard à la fois de la taille de cette dernière dans le corpus et de la fréquence moyenne d'apparition de la ou les formes en question dans les différentes sources. C'est pourquoi la fréquence spécifique de formes dans une source s'exprime en spécificité positive ou négative : cela signifie qu'une source peut de manière statistiquement significative sur-employer ou sous-employer cette forme par rapport à un emploi également réparti. Lorsque le nombre de formes spécifiques (segments répétés ou hapax par exemple) employées par une source se situe dans la moyenne, la spécificité est alors nulle car non significative¹¹.

Les résultats sur lesquels nous nous appuyons sont, d'une part, la répartition de la totalité des segments répétés (à partir de deux occurrences, constituant de fait une répétition) de longueur minimum de 7 mots, longueur qui constitue une proposition ou segment de phrase caractéristique (Ex : « Les États baltes ne comprennent plus Moscou ») ; d'autre part la répartition des hapax. Les hapax sont les mots qui n'apparaissent qu'une seule fois dans l'ensemble de l'échantillon. Ils constituent donc un indice de richesse lexicale, plus qu'éditoriale, étant entendu que l'on peut parler des mêmes sujets avec des mots différents. La surreprésentation d'hapax apporte donc un complément d'information à mettre en regard avec la sous-représentation de segments, et inversement entre la sous-représentation d'hapax et la surreprésentation de segments.

Sur l'ensemble des deux journées, les sources apparaissant comme les plus redondantes en termes à la fois de choix de sujets et de reprise de formulations stéréotypées sont les trois portails - MSN Actualités, Yahoo Actualités et Orange Actualités - ainsi que l'AFP. Ce résultat vient confirmer l'importance de l'espace occupé par ces sources et illustre leur politique de flux continu d'information, qui privilégie la réactivité plutôt que la créativité. Viennent ensuite de manière encore

¹¹ En termes statistiques, les spécificités sont considérées comme significatives à 0,05, qui signifie une probabilité de 95% pour que cette répartition ne soit pas due au hasard. Dans nos résultats les niveaux de significativité vont de 2 à 51, traduisant une probabilité comprise entre 10^2 et 10^{51} , soit bien au-delà du seuil de significativité minimum de 0,05.

très significative¹² les sites de télévisions françaises, particulièrement France 2 et France 3, très proches des modèles proposés par l'AFP (TF1 et France 24 apparaissant comme moins redondantes en comparaison) ainsi que la radio RTL.

Du côté de la diversité, les sources les plus remarquables sur les deux journées sont assez logiquement situées à l'étranger : il s'agit d'Afrique en ligne, Canoë (Canada) et Ria Novosti (Russie). Viennent ensuite, parmi les sources situées en France, quatre types de médias qui se distinguent par leur production originale de titres : il s'agit, dans l'ordre, du site participatif AgoraVox, du webzine d'opinion Bakchich, de la version internet du magazine d'opinion Politis, d'un autre site participatif Le Post, et de certains blogs. Viennent seulement ensuite les déclinaisons numériques de la presse plus traditionnelle avec Le Journal du Dimanche, Le Point, Les Échos ; et dans une moindre mesure, par ordre décroissant Libération, L'Humanité et Le Monde, qui restent dans une identité lexicale assez affirmée.

Enfin certaines sources apparaissent au carrefour de la diversité et de ce que l'on appelle la « banalité lexicale », c'est-à-dire « un recours massif aux mots les plus fréquents » (Marchand, 2008). Cette position est flagrante pour le Nouvel Observateur, mais c'est également le cas dans une moindre mesure pour 20Minutes, Métro, RFI, et RMC : ces sources associent globalement une sous-représentation à la fois des segments répétés et des hapax, ce qui témoigne d'une réelle activité de production de titres, mais également de l'emploi d'un langage sinon pauvre, tout au moins très commun et employé également par les autres sources. A la lumière d'études de nature socio-économique préalables qui montrent une tendance à la recherche de productivité dans les rédactions numériques des groupes de presse (Estienne, 2007 ; Rebillard *et al.*, 2007), on peut alors émettre l'hypothèse d'un tropisme de certains sites pour le journalisme « assis » (*desk*), via la réécriture de dépêches ou d'articles produits par des tiers, au détriment d'un travail d'élaboration de contenus originaux, passant lui plutôt par la voie du journalisme « debout ». A cet égard, et pour renforcer cette hypothèse d'un lien entre conditions pratiques de production (liées aux modèles économiques des différents médias) et formules éditoriales et lexicales adoptées, il est intéressant de noter qu'hormis le Nouvel Observateur, les médias concernés par cette caractéristique sont des radios et des titres de presse gratuite.

Des liens peuvent alors se dessiner entre la disparité des conditions de production du contenu, l'enjeu, la nature et enfin les attentes liées à la consommation de ce contenu, qui diffèrent sensiblement selon les types de source. L'hypothèse de l'existence de ces liens entre en résonance avec la notion de « contrat de communication médiatique » avancée par Charaudeau (1997). Réactivité et exhaustivité pour l'AFP, les agrégateurs et les sites de médias audiovisuels ; créativité et re-traitement réflexif ou critique pour les pure-players et les indépendants (professionnels ou non), voilà les deux pôles du continuum redondance-pluralisme. En ce qui concerne les blogs, leurs spécificités langagières et communicationnelles ont déjà été explicitées et interprétées. Elles semblent liées au rôle, joué par ceux-ci, de re-médiation et de re-traitement subjectivé de l'information première (Serfaty, 2006). Les autres types de source se situent entre ces deux pôles dans une position intermédiaire : équilibre plus ou moins précaire entre alignement sur la concurrence et stratégies de distinction pour la

¹² Les sources sont ici présentées par ordre décroissant de significativité, au regard de la répartition des segments répétés et des hapax sur l'ensemble des deux journées. Mais toutes les sources évoquées ici présentent des spécificités statistiquement significatives, y compris les dernières mentionnées (les taux allant de manière décroissante de 10^{51} à 10^{23}).

presse traditionnelle, procédés de retraitement rapide et minimal de l'information pour les gratuits et certaines radios, telles peuvent être les valeurs implicitement partagées par les producteurs et les consommateurs de l'information en ligne¹³, partage cristallisé dans cette disparité de contenus.

On peut envisager à l'avenir des analyses plus poussées et centrées sur des types de médias particuliers, ou sur des périodes plus longues, ou encore focalisées non plus sur les titres mais sur les contenus d'articles, dans lesquelles certaines hypothèses sur les spécificités des différents types de sources pourraient être testées de manière plus solide et approfondie. En effet, si une redondance de titre apparaît comme un indice tangible de redondance éditoriale, l'inverse (la diversité éditoriale) est plus difficilement déductible d'une originalité de titre, principalement du fait de la stratégie de certaines sources que nous venons d'évoquer (notamment les gratuits, mais pas seulement) dont l'activité se concentre précisément sur la reprise des dépêches d'agence (principalement l'AFP) et la reformulation (syntaxique et/ou lexicale) quasi-exclusive des titres.

¹³ Cette correspondance entre une offre et une demande d'information minimale, courte et factuelle, a déjà été démontrée dans le cas de la presse écrite pour les quotidiens gratuits (Augey et al., 2005)

III Bilan et perspectives

1. Rappel des objectifs initiaux

L'objectif général du projet IPRI était d'analyser la part de création originale dans les contenus d'actualité en circulation sur l'internet. A cet égard, le projet IPRI était envisagé comme une recherche exploratoire, étape préparatoire vis-à-vis de cet objectif à long terme, et possédant donc ses propres objectifs de court et moyen terme :

- Identifier les équipes et chercheurs compétents sur les aspects précités et les associer au projet de recherche pour élaborer une méthodologie d'ensemble.
- Transposer le cadre théorique issu d'observations socio-économiques de nature qualitative en questionnements traitables quantitativement : degré d'originalité des informations ; itinéraires de circulation entre les différents types de sites.
- Compléter les savoirs existants à propos du travail journalistique au sein des entreprises médiatiques traditionnelles, par une observation des pratiques alternatives de rédaction web de contenus d'actualité : blogs, webzines, sites participatifs.
- Procéder à de premières expérimentations.

2. Résultats obtenus

Les objectifs fixés au départ ont pour la plupart été atteints. Certains ont même été dépassés, en allant au-delà d'une simple visée exploratoire. Pendant assez longue des initiatives multiples qui ont animé cette année de travaux, la recherche a toutefois pu parfois pêcher par son caractère quelque peu décousu.

Un des premiers succès de la recherche IPRI réside dans son caractère interdisciplinaire. Les chercheurs en Informatique et en Sciences de l'information et de la communication impliqués dans ce projet ont vraiment travaillé en étroite collaboration, comme l'attestent les nombreuses réunions organisées au cours de l'année et surtout l'élaboration d'une méthodologie commune pour ce qui concerne l'axe « analyse de contenu ». Désormais, ces chercheurs se connaissent bien, ils identifient beaucoup plus clairement ce que chacun peut apporter à l'autre et emprunter à l'autre, en termes de savoir-faire et de compétences.

Cette collaboration s'est vraiment nouée dans les choix initiaux de corpus et de constitution des données, ainsi que dans l'analyse quantitative du pluralisme offert. Les travaux empiriques complémentaires réalisés par la suite ont chacun apporté leur part à l'analyse du pluralisme offert sur le web. Mais menés avec leurs propres préoccupations et traditions « disciplinaires » de recherche, ils ont été souvent conduits de façon trop indépendante les uns des autres, perdant en opportunités de collaboration (ex : entre l'analyse du traitement de l'élection de Barack Obama et les possibilités de visualisation offertes par l'application ArcEnCiel) ce qu'ils ont néanmoins fait gagner en termes d'émergence de nouvelles pistes d'investigation.

Une autre réussite de la recherche IPRI, sans doute sa principale réussite au final, tient dans les multiples expérimentations menées au niveau de l'analyse quantitative de contenu. Même si la vaste base de données disponibles n'a pas pu être exploitée dans toute sa profondeur faute de temps, les travaux réalisés au cours de cette année

ont permis de défricher très largement ce territoire scientifique jusqu'ici inexploré : la mesure chiffrée de la production d'informations d'actualité sur le web. Des résultats significatifs ont été obtenus sur ce plan, allant pour certains au-delà de la simple expérimentation, et en particulier à propos de la problématique-cœur de la recherche : le pluralisme et la redondance de l'information.

L'axe « socio-économie des modalités de production » a été développé en conformité avec les objectifs de départ. L'observation de structures alternatives, *pure players* de l'internet (site participatif, webzine, agrégateur de blogs), a apporté plusieurs enseignements utiles. La comparaison avec ce que l'on connaissait des modalités de production web dans les groupes de communication plus traditionnels, a ainsi été rendue possible.

Ceci étant, l'axe « socio-économie des modalités de production » a été moins développé au cours de cette année que l'axe « analyse de contenu », dont nous avons souligné précédemment le caractère foisonnant en termes d'expérimentations. Avec comme conséquence une articulation entre ces deux pans de la recherche qui s'est dans l'ensemble plutôt opérée *a posteriori*, et a fait naître quelques frustrations au final. Si les deux axes avaient évolué au même rythme, nul doute que des corrélations encore plus fines auraient pu être établies - ou relativisées - entre modalités de production et degré d'originalité des contenus.

3. Continuation de la recherche

Stimulés par les résultats obtenus au cours de cette première année de travail en commun, les chercheurs et laboratoires réunis au sein de la recherche exploratoire IPRI ont cherché les moyens de la transformer en un programme plus pérenne et à plus large échelle.

Une demande en ce sens a été faite auprès de l'Agence Nationale de la Recherche, déposée par le responsable scientifique actuel dans le cadre de l'appel à projets Jeunes Chercheuses et Jeunes Chercheurs. Une réponse positive a été obtenue pour financer la recherche IPRI sur la période 2009-2012. Dans cette seconde phase ANR, la recherche IPRI visera en particulier les objectifs suivants :

- Pour la partie analyse de contenu : aller sur le terrain des textes intégraux des articles et donc dépasser les titres et descriptions.
- Pour la partie socio-économique : multiplier les enquêtes auprès des producteurs d'information et surtout les articuler régulièrement avec les résultats de l'analyse de contenu.
- Sur le plan théorique et empirique : doubler l'approche initiale en termes de pluralisme « offert » d'une approche du pluralisme « consommé », via une pondération relative à l'audience des sites.
- Sur le plan institutionnel : pérenniser la recherche avec un Observatoire du pluralisme sur l'internet, à partir d'une modélisation de la méthode éprouvée ainsi que de son application à des thématiques d'information spécialisées (ex : information culturelle) ou des langues étrangères (ex : partenariats de recherche pour étudier le pluralisme dans les pays anglophones).

4. Communications et publications scientifiques

- Deux communications (avec actes) ont été acceptées pour le colloque international *New media and information* :

Smyrnaioi N., Marty E., Rebillard F., « Does the 'Long Tail' exist for online news ? A quantitative analysis of French-speaking websites », *International Conference on New Media and Information*, Panteion University, Athens, May 2009.

Damian B., Rebillard F., Smyrnaioi N., « La production de l'information web : quelles alternatives ? Une comparaison entre médias traditionnels et *pure players* de l'internet. », *International Conference on New Media and Information*, Panteion University, Athens, May 2009

- Une proposition de communication portant sur le suivi diachronique des informations a été soumise à la Conférence CORIA 09 :

Maisonnasse L., Gesche S., Egyed-Zsigmond E., Lallich-Boidin G., Calabretto S., Caplat G., « Internet et pluralisme des opinions »

- Une proposition d'article portant sur l'analyse lexicométrique est en cours d'évaluation par la revue *Mots* :

Marty E., Rebillard F., Smyrnaioi N., Touboul A., « Pluralisme et redondance de l'information sur l'internet. Analyses quantitatives des titres de l'actualité en ligne ».

- Un article incluant des résultats de l'analyse socio-économique est en préparation, à la suite d'une sollicitation de la revue *Réseaux*, en vue de figurer dans un dossier « Médias en ligne » à paraître début 2010.

RÉFÉRENCES BIBLIOGRAPHIQUES

- Anderson C., 2006, *The Long Tail. Why the Future of Business Is Selling Less of More*, New-York, Hyperion.
- Augey D., Lipani-Vaissade M.-C., Ruellan D., Utard J.-M., 2005, « La presse quotidienne gratuite ou le marketing du don », in R. Ringoot et J.-M. Utard (dir.), *Le journalisme en invention – Nouvelles pratiques, nouveaux acteurs*, Rennes, PUR.
- Boltanski, L., Chiapello, E., 2007, *Le Nouvel esprit du capitalisme*. Paris, Gallimard
- Benghozi P.J., Benhamou F., 2008, “The Long Tail: digital leverage for cultural diversity?”, *Note de recherche pour le Département des études, de la prospective et des statistiques*. [<http://www2.culture.gouv.fr/deps/fr/traine.pdf>]
- Benhamou F., Peltier S., 2006, « Une méthode multicritère d'évaluation de la diversité culturelle : application à l'édition de livres en France. », in X. Greffe (dir.), *Création et diversité au miroir des industries culturelles – Actes des journées d'économie de la culture*, Paris, La Documentation Française, p. 313-344.
- Charaudeau P., 1997, *Le discours d'information médiatique. La construction du miroir social*, Paris, Nathan / Institut national de l'audiovisuel (coll. « Médias-Recherches »), 286 p.
- Chartron G., Rebillard F., 2007, « La publication sur le web, entre filiations et innovations éditoriales », in R.-T. Pedauque, *La redocumentarisation du monde*, Toulouse : Editions Cépaduès, p. 201-213.
- Damian B., Ringoot R., Ruellan D., Thierry D., (dir.), 2002, *Inform@tion.local. Le paysage médiatique régional à l'ère d'internet*, Paris: L'Harmattan.
- Deuze M., 2003, « The web and its journalism : considering the consequences of different types of newsmedia online », *New Media and Society*, vol. 5, n°2, p. 203-230.
- Domingo D., Quandt T., Heinonen A., Paulussen S., Singer J.B., Vujnovic M., 2008, “Participatory journalism practices in the media and beyond. An international comparative study of initiatives in online newspapers”, *Journalism Practice*, vol. 2, n°3, p. 326-342.
- Elberse A., Oberholzer-Gee F., 2007, “Superstars and Underdogs : an Examination of the Long Tail Phenomenon in Video Sales”, *Harvard Business School Working Papers Series*, n° 07-015.
- Esquenazi J.P., 2002, *L'écriture de l'actualité; pour une sociologie du discours médiatique*, Grenoble, Presses Universitaires de Grenoble.
- Estienne Y., 2007, *Le journalisme après Internet*, Paris, L'Harmattan, 313 p.
- Gadrey, J., 2001, “Nouvel esprit du capitalisme et idéologie néo-libérale”, *Sociologie du travail*, vol. 43, n°3, p. 389-402.
- Goffman E., 1991, *Les cadres de l'expérience*. Paris, Minuit. 576 p.
- Hermida A., Thurman N., 2008, “A clash of cultures. The integration of user-generated content within professional journalistic frameworks at British newspaper websites”, *Journalism Practice*, vol. 2, n°3, p. 343-356.
- Lafon P., 1984. *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion, 217 p.
- Lancelot A., 2005, *Les problèmes de concentration dans le domaine des médias*, Rapport pour le Premier ministre.
- Lebart L., Salem A., 1994, *Statistique textuelle*, Paris, Dunod, 342 p.

Marchand P., 2008, « La déclaration de politique générale a-t-elle vécu ? », *Parlement(s), Revue d'histoire politique*, n° 10 –2008/2, p. 152-164

Neveu, E., 2001, *Sociologie du journalisme*, Paris, La Découverte.

Paulussen S., Ugille P., 2008, “User Generated Content in the Newsroom: Professional and Organisational Constraints on Participatory Journalism”, *Westminster Papers in Communication and Culture*, vol. 5, n° 2, p. 24–41.

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

Rebillard F., Cabedoche B., Damian B., Smyrnaio N., 2007, *Les mutations de la filière Presse et information – Note sectorielle pour le Ministère de la Culture et de la Communication*, Programme Diversité culturelle et mutations des industries de la culture, de l'information, et de la communication (MSH Paris-Nord), 33 p.

Reese S. D., 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, p. 235–261.

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

Ruellan D., 2007, *Le Journalisme ou le professionnalisme du flou*, Grenoble, PUG.

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, p. 25-35.

Tessier M. (et Baffert M.), 2007, *La presse au défi du numérique*, Rapport pour le Ministre de la culture et de la communication.

Thurman N., 2007, “The globalization of journalism online. A transatlantic study of news websites and their international readers”, *Journalism*, vol. 8, n° 3, p. 285-307.

Weber M., 1971 [1921], *Economie et société*, Paris: Plon.

ANNEXES

Corpus IPRI 2008

Catégorie	Site	URL	RSS
Presse en ligne	Le Monde	http://www.lemonde.fr	http://www.lemonde.fr/rss/une.xml
	Libération	http://www.liberation.fr/	http://www.liberation.fr/interactif/rss/actualites/
	Le Figaro	http://www.lefigaro.fr/	http://rss.lefigaro.fr/lefigaro/laune
		http://www.leparisien.fr/home/actualites-informations-direct-videos-parisien.htm	http://www.leparisien.fr/une/rss.xml
	Le Parisien		http://www.humanite.fr/backend.html
	L'Humanité	http://www.humanite.fr/	http://www.humanite.fr/backend.html
	La Croix	http://www.la-croix.com/	http://www.la-croix.com/rss/une/
	Le Journal du Dimanche	http://www.lejdd.fr/	http://www.lejdd.fr/rss.xml
	Les Echos	http://www.lesechos.fr/	http://syndication.lesechos.fr/rss/rss_une.xml
	La Tribune	http://www.latribune.fr/	http://www.latribune.fr/rss
	20 Minutes	http://www.20minutes.fr/	http://www.20minutes.fr/rss/flux/une.xml?xts=290428&xtor=RSS-1
	Metro	http://www.metrofrance.com/paris.xml	http://www.metrofrance.com/rss.xml?c=section-news
	Fakir	http://www.fakirpresse.info/frontoffice/main.php?rub=accueil.php	http://www.fakirpresse.info/frontoffice/rss.php
			http://www.leplanb.org/index2.php?option=com_rss&feed=RSS1.0&no_html=1
	Le Plan B	http://www.leplanb.org/	
	Le Nouvel Observateur	http://tempsreel.nouvelobs.com/	http://tempsreel.nouvelobs.com/file/rss_perm/rss_permanent.xml
	L'Express	http://www.lexpress.fr/	http://rss.feedsportal.com/c/568/f/9909/index.rss
	Le Point	http://www.lepoint.fr/	http://www.lepoint.fr/content/system/rss/a_la_une/a_la_une_doc.xml
		http://www.valeursactuelles.com/public/valeurs-actuelles/html/fr/accueil/index.php?page_symlink=page-accueil&type_page=1	http://www.valeursactuelles.com/public/valeurs-actuelles/html/fr/actualites/index.php
	Valeurs Actuelles		http://www.parismatch.com/parismatch/rss_export/feed/rss-match-actu
	Paris Match	http://www.parismatch.com/parismatch	http://www.marianne2.fr/xml/atom.xml
	Marianne	http://www.marianne2.fr/	
	Courrier International	http://www.courrierinternational.com/gabarits/html/default_online.asp	http://www.courrierinternational.com/rss/rss_a_la_une.xml
	Le Monde Diplomatique	http://www.monde-diplomatique.fr/	http://www.monde-diplomatique.fr/rss/
	Politis	http://www.politis.fr/	http://www.politis.fr/spip.php?page=backend
			(Pas de flux RSS --> rubrique "à la Une") http://www.ouest-france.fr/actu/actuLst.php?idCla=3631
	Ouest-France	http://www.ouest-france.fr/	
	La Dépêche	http://www.ladepeche.fr/	http://www.ladepeche.fr/rss/39.rss
	TF1/LCI	http://tf1.lci.fr/	http://tf1.lci.fr/xml/rss/0,,10,00.xml
	France 2	http://info.france2.fr/	http://feeds.feedburner.com/France2-ALaUne
	France 3	http://info.france3.fr/	http://feeds.feedburner.com/France3-ALaUne
	France 24	http://www.france24.com/fr/	http://www.france24.com/fr/derniere-nouvelle/rss
	Euronews	http://www.euronews.net/	http://feeds.feedburner.com/euronews/fr/home/
	RTL	http://www.rtl.fr/	http://rss.rtl.fr/une.xml
	Europe 1	http://www.europe1.fr/	http://www.europe1.fr/rss_export/feed/info

Agences	RFI	http://www.rfi.fr/	http://www.rfi.fr/actufr/pages/001/accueil.xml
	RMC	http://www.rmc.fr/	http://www.rmc.fr/rss/livestickers/actus.xml
	Cyberpresse	http://www.cyberpresse.ca/	http://www.cyberpresse.ca/apps/pbcs.dll/section?Category=CPRSS&profile=1019&mime=xml&cacheid=1019
	Radio Canada	http://www.radio-canada.ca/index.shtml	http://rss.radio-canada.ca/fils/nouvelles/nouvelles.xml
	Radio Suisse Romande	http://www.rsr.ch/	http://www.rsr.ch/podcast.aspx?rss=info
	Le Temps	http://www.letemps.ch/	http://www.letemps.ch/rss.asp
	RTL (Belgique)	http://www.rtlinfo.be/index.php	http://feeds.feedburner.com/Rtlinfo-s-ALaUne
	RTBF	http://www.rtbf.be/	http://www.rtbf.be/info/accueil/rss
	Aujourd'hui Le Maroc	http://www.aujourd'hui.ma/	Pas de flux RSS
	El Watan (Algérie)	http://www.elwatan.com/	http://www.elwatan.com/?page=backend
	L'Orient - Le Jour (Liban)	http://www.lorientlejour.com/page.aspx?page=main-page	http://rss.feedsportal.com/feed/jeuneafrique/actualites
	Jeune Afrique	http://www.jeuneafrique.com/	
	AFP	http://www.afp.com/francais/home/ http://today.reuters.fr/news/default.aspx	Pas de flux RSS
	Reuters		Pas de flux RSS
	Ria Novosti (Russie)	http://fr.rian.ru/	http://fr.rian.ru/export/rss2/index.xml
	Xinhuan (Chine)	http://www.french.xinhuanet.com/french/index.htm	Pas de flux RSS
	La Presse Canadienne	http://www.thecanadianpress.com/	Pas de flux RSS
Blogs	Blog Birenbaum (sur Le Post)	http://www.lepost.fr/perso/birenbaum/	http://www.lepost.fr/perso/birenbaum/rss.xml
	Embruns - carnet web de Laurent Gloaguen	http://embruns.net/	http://embruns.net/logbook/atom.xml
	Coulisses de Bruxelles (Jean Quatremer)	http://bruxelles.blogs.liberation.fr/coulisses/	http://bruxelles.blogs.liberation.fr/coulisses/rss.xml
	Le blog politique de Luc Mandret	http://777socrate.blogspot.com/	http://feeds.feedburner.com/Narcisse
	SarkoFrance	http://sarkofrance.blogspot.com/	http://feeds.feedburner.com/sarkofrance
	Intox 2007	http://www.intox2007.info/	http://www.intox2007.info/index.php?feed/rss2
	Partageons mon avis	http://jegpol.blogspot.com/	http://jegpol.blogspot.com/feeds/posts/default
	Marc Vasseur (jeune élu en retraite)	http://marc.vasseur.over-blog.com/	http://marc.vasseur.over-blog.com/rss2.php
	Diner's room	http://dinersroom.free.fr/	http://dinersroom.free.fr/rss.php
	Koztousjours, tu m'intéresses	http://www.koztousjours.fr/	http://feeds.feedburner.com/KoztousjoursTuMintresses

Plume de presse	http://www.plumedepresse.com/	http://www.plumedepresse.com/spip.php?page=backend
Claude Guillon	http://claudeguillon.internetdown.org/	http://claudeguillon.internetdown.org/backend.php3
DavDuf	http://www.davduf.net/	http://www.davduf.net/spip.php?page=backend
Le blog de Colette		
Braeckman (sur Le Soir)	http://blogs.lesoir.be/colette-braeckman/	http://blogs.lesoir.be/colette-braeckman/feed/
Olivier		
Bouba-Olga - Blog d'un économiste qui suit et commente l'actualité	http://obouba.over-blog.com/	http://obouba.over-blog.com/rss2.php
Le Monolecte	http://blog.monolecte.fr/	http://blog.monolecte.fr/feed/atom
Les eaux glacées du calcul égoïste - Le blog de Marc Laimé	http://www.eauxglacees.com/	http://www.eauxglacees.com/backend.php3
Le blog de Jean-Pierre Martin, un petit cadre dans une grosse boîte	http://onsefechier-anatic6.blogspot.com/	http://onsefechier-anatic6.blogspot.com/feeds/posts/default
Jean-Michel Apathie (sur RTL)	http://blogs.rtl.fr/aphatie/	http://blogs.rtl.fr/aphatie/index.php/feed/rss2
Nouvelles d'Orient - Les blogs du Diplo	http://blog.mondediplo.net/-Nouvelles-d-Orient-	http://blog.mondediplo.net/spip.php?page=backend
"Blog québécois 1"	http://blogues.lactualite.com/hebert/	http://blogues.lactualite.com/hebert/?feed=rss2
"Blog québécois 2"	http://blogues.lactualite.com/vastel/	http://blogues.lactualite.com/vastel/?feed=rss2

Webzines	Bakchich Journal Chrétien	http://www.bakchich.info/	http://www.bakchich.info/spip.php?page=backend
	Afrik.com	http://www.afrik.com/	http://journalchretien.net/spip.php?page=backend
	Les infos.tv	http://www.lesinfos.tv/	http://www.afrik.com/affiliation/backend.php
	Politique.net Contre Info.info	http://www.politique.net/	http://www.lesinfos.tv/dotclear/rss.php
	Rebelyon De Source Sûre	http://contreinfo.info/	http://feeds.feedburner.com/politique-net
	Indymedia Paris	http://rebelyon.info/	http://contreinfo.info/backend.php3
	Vox Populi	http://desourcesure.com/une/	http://rebelyon.info/backend.php
	Le Grand Soir	http://paris.indymedia.org/	http://feeds.feedburner.com/desourcesure/rss
	Le Monde Citoyen	http://www.vox-populi.net/	(site inaccessible depuis quelques jours, dernière tentative : 9 sept. 2008)
	Oulala.net	http://www.legrandsoir.info/	Pas de flux RSS
Participatifs	Samizdat.net	http://www.lemondeditoyen.com/	(RSS de la partie "Dazibao", espace offert aux viviteurs pour publier...) http://www.legrandsoir.info/spip.php?page=backend
	Canoë Afrique en Ligne	http://www.oulala.net/Portail/index.php3	Pas de flux RSS
	Rue 89	http://www.samizdat.net/	(renvoi vers les 10 derniers articles publiés, mais le lien ne fonctionne pas)
	MediaPart	http://fr.canoe.ca/	http://www.oulala.net/Portail/index.php3
	AgoraVox	http://www.afriquenligne.fr/	("rendez-vous le 15 septembre"...) http://www.canoe.com/rss/feed/nouvelles/aujourd'hui.xml
	Le Post	http://www.rue89.com/	http://www.afriquenligne.fr/rss.feed
	Daily Motion - Chaîne Actu et Politique	http://www.mediapart.fr/	http://www.rue89.com/homepage/feed
	Cent Papiers	http://www.agoravox.fr/	http://www.mediapart.fr/articles/feed
		http://www.lepost.fr/	http://rss.feedsportal.com/feed/agoravox/main
		http://www.dailymotion.com/cluster/news	http://www.lepost.fr/rss.xml

Portails	Yahoo !		
	Actualités		
	France	http://fr.news.yahoo.com/	http://fr.news.yahoo.com/rss/a-la-une.xml
	Actualités		
	Orange	http://actu.orange.fr/	http://actu.orange.fr/Inc/RSS/rss_une.xml
	MSN		
	Actualités	http://news.fr.msn.com/	http://news.fr.msn.com/RSS/News_RSS_Alaune.aspx
	Les		
	manchettes		
	de Yahoo !		(Fil "Monde")
Agrégateurs	Actualités	http://qc.news.yahoo.com/	http://d.yimg.com/cf.rss.news.yahoo.com/rss/monde
	Atlas Vista		
	Maroc	http://www.avmaroc.com/	http://www.avmaroc.com/rss/actualite/alaune-c0.xml
	Google		
	Actualités		
	France	http://news.google.fr/	http://news.google.fr/?output=rss
	Actualité -	http://www.wikio.fr/dernieres-actualites	
	Wikio		http://rss.wikio.fr/a_la_une.rss
	Netvibes	http://www.netvibes.com/#General	Pas de flux RSS (infos de la colonne de droite)
	Rezo.net	http://rezo.net/	http://rezo.net/backend/portail
	Google		
	Actualités		
	Canada	http://news.google.ca/	http://news.google.ca/?output=rss
	Paper Blog	http://www.paperblog.fr/	http://www.paperblog.fr/rss/tags/politique/