questions de communication

Questions de communication

20 | 2011
Évoquer la mort

L'étude des médias est-elle soluble dans l'informatique et la physique ? À propos du recours aux *digital methods* dans l'analyse de l'information en ligne

Will Media Studies be mixed within computer Science and Physics? The Use of digital Methods in online News Research

Franck Rebillard

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/questionsdecommunication/2148
DOI : 10.4000/questionsdecommunication.2148
ISSN : 2259-8901

Éditeur
Presses universitaires de Lorraine

Édition imprimée
Date de publication : 31 décembre 2011
Pagination : 353-376
ISBN : 978-2-8143-0108-5
ISSN : 1633-5961

Référence électronique
Franck Rebillard, « L'étude des médias est-elle soluble dans l'informatique et la physique ? À propos du recours aux *digital methods* dans l'analyse de l'information en ligne », *Questions de communication* [En ligne], 20 | 2011, mis en ligne le 01 février 2014, consulté le 20 avril 2019. URL : http://journals.openedition.org/questionsdecommunication/2148 ; DOI : 10.4000/questionsdecommunication.2148

questions de communication, 2011, 20, 353-376

> NOTES DE RECHERCHE

FRANCK REBILLARD
Communication, information, médias
Université Sorbonne Nouvelle-Paris 3
franck.rebillard@univ-paris3.fr

L'ÉTUDE DES MÉDIAS EST-ELLE SOLUBLE DANS L'INFORMATIQUE ET LA PHYSIQUE ?
À PROPOS DU RECOURS AUX *DIGITAL METHODS* DANS L'ANALYSE DE L'INFORMATION EN LIGNE

Résumé. — Les recherches portant sur l'information en ligne sont marquées par une certaine emprise des méthodes quantitatives et statistiques. Elles rejoignent pour partie une autoproclamée *nouvelle science des réseaux*, appuyée sur la physique et les mathématiques, ainsi que le plaidoyer pour des méthodes d'observation spécifiques (*digital methods*), tirant profit des capacités de traitement informatique et des matériaux numériques de l'internet. L'examen détaillé de tels travaux permet d'en cerner les apports (exhaustivité du corpus, représentation synthétique des résultats) et les limites (analyses parfois insuffisamment approfondies, catégorisations réductrices pour les besoins d'automatisation). Il fait également apparaître la nécessité de croiser les *digital methods* avec des méthodes plus classiques en sciences humaines et sociales, afin de saisir les multiples dimensions des phénomènes médiatiques.

Mots clés. — Étude des médias, information en ligne, méthodes, *digital methods*, nouvelle science des réseaux, informatique.

L'étude des médias s'intéresse aux formes, aux significations, aux modalités de production et de réception, au contexte historique et aux enjeux politiques des contenus d'actualité diffusés par voie imprimée, audiovisuelle ou numérique[1]. Qu'elle soit regroupée sous l'appellation *media studies* dans le monde académique anglophone, qu'elle se déploie en France dans l'espace interdisciplinaire des sciences de l'information et de la communication (SIC) ou à partir de branches spécialisées de la sociologie comme des sciences du langage, l'étude des médias réunit des approches plurielles mais relevant très majoritairement des sciences humaines et sociales (SHS). Toutefois, depuis quelques années, avec le développement de l'internet, les médias ont été l'objet d'analyses issues de disciplines scientifiques bien différentes. Des concepts et démarches sont importés des sciences physiques et de l'informatique pour mener des observations – à dimension fortement quantitative – de la diffusion d'informations sur l'internet. Ce phénomène est loin d'être marginal. Au contraire, il se caractérise par la grande visibilité – au niveau international et jusqu'auprès des professionnels des médias en ligne – de travaux réalisés par des physiciens et des informaticiens. Dans une telle situation, l'étude des médias n'est plus le domaine réservé ou de prédilection des sciences humaines et sociales. Plus exactement, dans plusieurs travaux sur l'information en ligne, les méthodes de collecte à grande échelle et de traitement statistique d'énormes volumes de données s'imposent, ou sont reprises par des chercheurs en SHS qui les mixent avec leurs méthodes plus habituelles. Afin de lever toute ambigüité, précisons que l'emploi de méthodes quantitatives outillées par l'informatique ne constitue bien évidemment pas une nouveauté en soi pour l'étude des médias. Par exemple, des logiciels de lexicométrie ou des outils de traitement automatique de la langue sont employés depuis plusieurs décennies pour analyser le discours de presse. De plus en plus fréquemment rangées sous l'appellation de *digital methods*, les méthodes auxquelles nous consacrons cet article sont également quantitatives et outillées par l'informatique. Mais avec l'internet, elles ont ceci de nouveau : pouvoir s'appliquer immédiatement à des corpus quasi infinis car nativement numériques.

C'est donc bien dans l'envergure et la rapidité de leur application que réside la nouveauté de l'emploi des *digital methods* en étude des médias. Derrière l'avalanche souvent impressionnante des chiffres et représentations graphiques livrés dans les travaux de ce type, l'objectif est ici de pointer les apports et les limites du recours à de telles méthodes. Celles-ci offrent des possibilités nouvelles à l'étude des médias, notamment au niveau de la taille des corpus rendus disponibles pour l'observation et le rendu synthétique des résultats. Mais, parallèlement, ces méthodes souffrent dans bien des cas d'une certaine superficialité des indicateurs et d'un manque d'approfondissement qualitatif

[1] Plusieurs des réflexions présentées ici sont issues d'une présentation orale, lors du séminaire *Médias et SHS* de l'université Paris 2, à l'invitation de G. Goasdoué. Et une première mouture de ce texte a bénéficié de la relecture attentive de B. Rieder. Que ces collègues en soient remerciés.

de l'analyse des contenus médiatiques dus aux nécessités de systématisation et d'automatisation de l'observation à grande échelle. Parfois aussi, un déficit plus général de problématisation en amont de la recherche, voire un manque d'intégration de la dimension sociologique de la production et de la réception des contenus médiatiques, peut conduire à des interprétations assez sommaires. Pour illustrer ces éléments de réflexion, nous reviendrons sur quelques recherches consacrées à la question de la diversité de l'information en ligne.

Le choix de cette thématique est dû à plusieurs raisons. D'abord, étudier la diversité de l'information en ligne incite à comparer l'offre d'informations dans des espaces médiatiques pluriels, en l'occurrence plusieurs sites *web* ou catégories de sites (sites de journaux, de radios, de télévisions, mais aussi sites participatifs, blogs, etc.). Ceci rend particulièrement opportunes les possibilités d'analyse sur de larges corpus fournies par les *digital methods*. De ce fait, nous examinerons trois recherches exploitant graduellement les possibilités de telles méthodes. Par ailleurs, le choix de cette thématique est intrinsèquement lié à nos propres activités de recherche : en pratique, nous avons expérimenté ce type de méthodes en coordonnant une recherche collective sur la diversité de l'information en ligne, associant des chercheurs en informatique et en sciences de l'information et de la communication. C'est donc un témoignage réflexif qui est également proposé.

Mais avant de présenter ces recherches et ce retour d'expérience, il paraît nécessaire de situer le contexte scientifique dans lequel ils surviennent et de préciser que ces travaux prennent place dans un cadre beaucoup plus large où se rejoignent l'affirmation d'une « nouvelle science des réseaux » – très largement appuyée sur la physique et les mathématiques – ainsi que le plaidoyer pour des méthodes d'observation spécifiques (*digital methods*), tirant profit des capacités de traitement informatique et des matériaux numériques de l'internet. Ce cadrage théorique initial permettra de mieux comprendre l'emprise grandissante des démarches quantitatives et statistiques à propos de l'information en ligne, et les enjeux épistémologiques plus profonds qui en découlent.

Mise en contexte

Dans les recherches qui sont au centre de notre propos, les références aux textes fondateurs de l'étude des médias voisinent avec des renvois à des publications en informatique, mathématiques, physique ou encore biologie. À titre d'exemple, dans la bibliographie de l'une d'entre elles (Leskovec *et al.*, 2009), l'ouvrage *The People's Choice* de Paul Lazarsfeld, Bernard Berelson et Hazel Gaudet (1944), considéré comme un travail inaugural en matière de sociologie empirique des médias de masse, est référencé juste après une modélisation de la maturation folliculaire, issue d'un acte de colloque en biologie mathématique (Lacker, Peskin, 1981). De manière semblable, dans une autre de ces recherches (Asur *et al.*,

2011), un article rétrospectif de Maxwell E. McCombs et Donald L. Shaw (1993) sur leur propre concept d'*agenda setting*, concept largement travaillé dans l'étude des médias, précède une synthèse de Michaël Mitzenmacher (2004) sur les lois de puissance et les distributions logarithmiques publiée dans la revue *Internet Mathematics*. Un tel voisinage de références, en apparence très hétéroclite, surprendra certainement le lecteur non familier de ce type d'écrits. Il s'inscrit pourtant dans une vaste entreprise scientifique engagée maintenant depuis plus d'une dizaine d'années. Depuis que l'internet constitue un objet de recherche en tant que tel, des chercheurs en sciences dites exactes et plus précisément en sciences physiques et en mathématiques ont tenté d'en modéliser la structuration, à travers une autoproclamée « nouvelle science des réseaux ». Dans son versant empirique, cette perspective théorique s'appuie très fortement sur des méthodes de traitement informatique de données numériques. Au point de revendiquer des méthodes nouvelles, des *digital methods* propres à l'analyse de l'internet, appliquées à l'étude de l'information en ligne dans les recherches examinées ici.

La nouvelle science des réseaux

Depuis la fin des années 90 se succèdent des travaux, initiés principalement par des physiciens et des informaticiens, consistant à « mesurer » l'internet. Une première consécration académique est intervenue avec la publication de ces derniers dans les revues *Science* et *Nature*. Dans la livraison 401 de cette dernière, paru en septembre 1999, figurent deux résumés emblématiques de telles démarches. Un premier article s'évertue à estimer le diamètre du *World Wide Web*, indiquant que la distance par hyperliens entre toutes les pages web est en moyenne de 19 clics, et en conclue à une structure hautement connectée de l'internet (Albert *et al.*, 1999). Après avoir observé une répartition typique des lois de puissance entre un faible nombre de sites très fourni en pages web et de nombreux sites formés d'une poignée de pages, le second article propose une modélisation de l'expansion future de l'internet (Huberman, Adamic, 1999). Ces recherches pionnières, et leur grande visibilité conférée par la publication dans des revues d'envergure mondiale, entraînent dans leur sillage une multitude d'autres travaux tout au long de la décennie 2000. Un nouvel espace scientifique va ainsi progressivement se former, revendiqué comme spécifique et dénommé « nouvelle science des réseaux ». L'idée est de faire du réseau un concept pandisciplinaire, de s'appuyer sur les éléments de connaissance issus de différentes disciplines au cours du temps (théorie mathématique des graphes, physique fractale, biologie cellulaire, sociologie des réseaux sociaux), puis de les appliquer à la description et à la compréhension de l'ensemble des phénomènes reposant sur une structure réticulaire, à commencer par l'internet. Un tel projet est notamment explicité dans un ouvrage coordonné par trois des figures de proue de cette « nouvelle science des réseaux » : Mark Newman, Albert-Laszlo Barabasi, et Duncan J. Watts (2006). Consistant essentiellement en une anthologie des recherches passées et contemporaines recourant au concept de réseau,

l'ouvrage débute avec l'annonce d'une nouvelle ère scientifique en la matière. D'après les auteurs, les recherches sur les réseaux sont restées jusqu'ici trop théoriques du côté des sciences dites exactes – la sophistication artificielle des modèles mathématiques de graphes est regrettée –, et trop peu systématiques du côté des SHS. À cet égard, les velléités empiriques de la sociologie des réseaux sociaux sont jugées très opportunes mais jusqu'ici limitées par des méthodes d'observation et de recueil insuffisamment quantitatives. L'arrivée de l'internet, expliquent les auteurs, a permis de dépasser ces deux limites, et d'ouvrir ainsi l'horizon pour (re-)fonder une science des réseaux qui traverse toutes les disciplines scientifiques. D'une part, estiment-ils, l'internet fait partie des « *real-world networks* », des réseaux animés dont l'évolution et la dynamique sont mues par les pratiques des acteurs sociaux. D'autre part, ces pratiques sociales peuvent être analysées à grande échelle, à partir de traitements automatisés, en raison de la nature numérique des immenses jeux de données mis à disposition avec l'internet :

> « *Not only has the Internet focused popular and scientific attention alike on the topic of networks and networked systems, but it has led to data collections methods for social and other networks that avoid many of the difficulties of traditional sociometry* »[2] (Newman *et al.*, 2006 : 6).

C'est donc dans cette double optique qu'émerge la « nouvelle science des réseaux ». Il s'agit non seulement de rassembler des connaissances éparses touchant aux réseaux pour élaborer un paradigme transversal et unifiant. Pour ces chercheurs acculturés aux méthodes quantitatives, il s'agit aussi de saisir l'occasion fournie par la composante numérique de l'internet, informatiquement exploitable, pour tester et élaborer des modélisations mathématiques à grande échelle de ce qu'ils estiment être un réseau « réel », autrement dit en lien avec des activités sociales.

Des *digitized methods* aux *digital methods*

Sans toujours s'inscrire explicitement dans ce courant de la « nouvelle science des réseaux », plusieurs chercheurs vont néanmoins partager une même posture, consistant à tirer profit des ressources numériques de l'internet pour y déployer des méthodes d'analyse quantitative embrassant de larges corpus de données. Expérimentées de façon assez dispersée dans un premier temps, ces méthodes ont maintenant une assise plus stable, jusqu'à gagner une certaine autonomie. En France, le collectif *Web Atlas* initié par Franck Ghitalla a joué un rôle important dans la diffusion de ces méthodes, avec notamment la mise au point d'outils de cartographie de l'internet (*Navicrawler*, *Gephi*), et en essaimant autant vers des

[2] « Non seulement, l'internet a attiré l'attention des scientifiques comme des profanes sur la thématique des réseaux et des systèmes en réseau, il a aussi conduit à des méthodes de collecte de données concernant les réseaux sociaux ou d'autre nature qui dépassent nombre de difficultés rencontrées par la sociométrie traditionnelle ».

applications industrielles (*Linkfluence*) que vers la mise en place de structures d'appui aux recherches en SHS (*Médialab* de Sciences Po).

Au niveau international, c'est sans doute Richard Rogers qui a le plus fortement affirmé cette volonté avec la *Digital Methods Initiative*, coordonnée depuis l'université d'Amsterdam. En effet, Richard Rogers (2010 : 242) plaide pour le développement de méthodes d'observation spécifiques à l'internet (*digital methods*) plutôt que l'adaptation des méthodes classiques des SHS à ce nouvel objet (*digitized methods*) :

> « *A series of volumes and handbooks has now appeared where the researchers continue to develop quite a classic social scientific armature, which includes interviews, surveys, observation, and others. What I would like to point out in particular is these could be categorized or conceptualized as digitized methods. That is, taking methods – existing methods – and trying to move them online* »[3].

Contre cet emploi des bonnes vieilles méthodes pour étudier l'internet, Richard Rogers milite pour des méthodes nouvelles qui soient au plus près du terrain, terrain numérique s'entend :

> « *What I'd like to try to do [...] is introduce a new era in Internet-related research where we no longer need to go off-line, or to digitize method, in order to study the online. Rather, in studying the online, we make and ground findings about society and culture with the Internet. Thus, the Internet is a research site where one can ground findings about reality* »[4] (*ibid.* : 243).

En comparaison des promoteurs de la « nouvelle science des réseaux » évoqués précédemment, Richard Rogers et les chercheurs gravitant autour de la *Digital Methods Initiative* possèdent une proximité beaucoup plus forte avec les SHS. Leur approche de l'internet est moins formalisée sur le plan mathématique, ils s'interrogent de façon approfondie sur les interprétations à apposer aux données collectées, et n'érigent pas le réseau en référent paradigmatique absolu. Néanmoins, malgré ces différences d'ordre théorique et méthodologique, on peut déceler plusieurs points de convergence entre les courants de la « nouvelle science des réseaux » et de la *Digital Methods Initiative* quant à l'appréhension de l'internet. Premier point commun : l'internet est considéré comme un objet de recherche en soi, digne d'intérêt à lui seul. Plus encore, Richard Rogers liste les sous-domaines d'étude de l'internet qu'il convient d'investiguer en tant que

[3] « Une série d'ouvrages et de manuels est maintenant apparue où les chercheurs continuent à développer une armature tout à fait classique en sciences sociales, qui inclue entretiens, questionnaires, observations directes, et autres. Ce que je voudrais particulièrement souligner, c'est que de telles méthodes pourraient être conceptualisées ou catégorisées comme des méthodes numérisées. Ce qui revient à prendre des méthodes – des méthodes existantes – et à essayer de les transposer en ligne ».

[4] « Ce que je voudrais tenter de faire, c'est introduire une nouvelle ère dans la recherche relative à l'internet, où l'on n'aurait plus besoin d'aller hors-ligne, ou de numériser des méthodes existantes, pour étudier ce qui se passe en ligne. À la place, en cherchant directement en ligne, nous produisons et accomplissons grâce à l'internet des découvertes concernant la société et la culture. Ainsi, l'internet est-il un terrain de recherche d'où l'on peut extraire des résultats concernant le monde réel ».

tels : les sites, mais également les liens hypertextuels, les moteurs de recherche, les sphères du *web*, etc. Deuxième point de convergence : les éléments observés directement sur l'internet ne sont pas vus comme de purs artefacts, ils sont aussi censés renseigner sur les pratiques sociales qui les activent. À titre d'illustration, voici le premier exemple donné par l'auteur dans son article-manifeste de 2010 : le projet *Google Flu Trends* (Ginsberg *et al.*, 2009) de cartographie de la propagation internationale de la grippe. Ces cartes sont élaborées à partir de la géolocalisation des requêtes des internautes sur le terme flu dans *Google*, et s'avèrent très proches de celles rendues avec quelques jours de décalage par les organismes institutionnels de surveillance de l'épidémie. Troisième position commune : la matérialité entièrement numérique de l'internet est vue comme un terrain propice à des observations exhaustives. Ce dernier point justifie le recours à des méthodes outillées par l'informatique comme moyen de collecte et de traitement d'immenses jeux de données. Ainsi Richard Rogers et son équipe développent-ils plusieurs outils logiciels destinés à exploiter les ressources numériques de l'internet, avec en outre un soin tout particulier porté au *design* – couleurs différenciées des liens entrants et sortants, nuages de *tags* pour les analyses de contenu, etc. – pour faciliter l'appropriation de ces observations extrêmement volumineuses. L'application de telles prises de position théoriques et méthodologiques à l'étude des médias n'est pas sans bouleverser cette dernière. On voit ici les apports résidant dans les capacités de réaliser des analyses à grande échelle et, en même temps, les limites consistant dans bien des cas à inférer des pratiques sociales à partir de simples agrégats statistiques construits directement à partir du *web*. Nous proposons de les examiner plus en détail à partir du cas d'analyses portant sur la diversité de l'information en ligne.

Quelques travaux récents sur la diversité de l'information en ligne

En matière d'étude des médias, la question du pluralisme de l'information est très ancienne. Elle interroge principalement la propriété des différents médias en présence, ainsi que la diversité des contenus (sujets abordés dans l'agenda médiatique, angles choisis pour les traiter, opinions et points de vue exprimés, etc.) proposés aux lecteurs, auditeurs, téléspectateurs. Cette question a connu un regain d'intérêt avec le développement de l'internet. Outre le fait de s'ajouter à la presse écrite, à la radio, et à la télévision, l'internet a lui-même fait naître une pluralité de nouveaux espaces médiatiques : *webzines* et sites de journalisme participatif, *blogs*, services dits de réseaux sociaux, en plus des déclinaisons numériques des médias existants. Une telle profusion a conduit les chercheurs à se pencher sur la diversité de l'information en ligne. Parmi eux, certains ont tenté d'évaluer la pluralité des informations offertes sur plusieurs des espaces médiatiques de l'internet, expérimentant au passage les opportunités de travailler sur des corpus élargis grâce aux *digital methods*. Les trois recherches examinées

dans la présente section poursuivent ce double objectif. La première (Carpenter, 2010) compare les thématiques abordées dans les sites de quotidiens imprimés avec les thématiques mises en avant par les sites de journalisme participatif. La deuxième (Leskovec *et al.*, 2009) vise à observer la propagation des principales informations à la « Une » de l'agenda médiatique numérique et à déterminer la réactivité des *blogs* par rapport aux sites de médias traditionnels. La troisième (Asur *et al.*, 2011) s'attache aux principaux sujets de discussion sur *Twitter* pour voir s'ils sont lancés par les internautes eux-mêmes ou s'ils trouvent leur origine dans les sites de médias traditionnels. Ces recherches ont été choisies parmi les plus récentes autour de la diversité de l'information en ligne, et surtout parce qu'elles recourent à des méthodes quantitatives outillées par l'informatique. Toutefois, elles le font de manière plus ou moins intensive : de la simple assistance pour la constitution de corpus à l'analyse entièrement automatisée, apparaît une gradation variable dans le recours aux *digital methods*.

Entre *digitized* et *digital methods* : une comparaison entre sites participatifs et sites de quotidiens

Des trois recherches examinées, celle de Serena Carpenter, professeure assistante dans une école de journalisme (Arizona State University) et titulaire d'un doctorat en *Media and Information Studies*, est sans aucun doute la moins engagée dans les *digital methods*. Pour reprendre la dichotomie proposée par Richard Rogers, on pourrait même dire que cette recherche emprunte à la fois aux *digital methods* et aux *digitized methods*. Les *digital methods* sont mobilisées pour constituer le corpus et atteindre une certaine exhaustivité vis-à-vis de l'objectif visé par la recherche. Mais l'analyse des données ainsi récoltées est beaucoup plus usuelle et artisanale, pratiquée « à la main » diraient les chercheurs en informatique. Avant d'entrer dans le détail de cette recherche, rappelons-en l'objectif : il s'agit de comparer la diversité de l'information offerte par les sites de journalisme participatif avec celle des versions *web* de quotidiens « papier ». Ici, la matérialité numérique des informations et leur facilité de collecte à partir de l'internet ont été exploitées pour conduire une étude qui couvre l'ensemble du territoire des États-Unis. De façon exhaustive, les 50 États ont été couverts par l'étude. Dans chacun des États, ont été identifiés les sites participatifs et en particulier leur ville d'implantation. Et si un quotidien imprimé était diffusé dans une de ces villes, alors son site *web* était retenu. On en arrive à une liste de 122 sites *web*, se répartissant entre 72 sites de journalisme participatif et 50 sites de quotidiens « papier », avec tout juste un quotidien par État donc. Les articles apparaissant sur la page d'accueil de ces 122 sites ont été collectés pendant une période de un mois (mars 2007) pour aboutir à un total de 6485 articles.

Au-delà du caractère volumineux du corpus, il convient de souligner la plus-value apportée par l'internet comme moyen d'accès commode à des données concernant des espaces géographiquement très dispersés. S'il avait fallu recueillir

les articles « papier » de tels quotidiens, il aurait fallu soit contacter directement chacun des médias visés, soit passer par des services d'archive ou des centres de documentation sans forcément l'assurance d'un accès unique à toutes les données souhaitées. On voit donc là les facilités procurées au chercheur utilisant des méthodes de collecte adaptées au terrain numérique de l'internet : il peut beaucoup plus librement déterminer son corpus, sans véritable contrainte extérieure, et se l'approprier concrètement de façon plutôt confortable (rapatriement et thésaurisation de données à distance, depuis son propre bureau... d'ordinateur). Dans la présente recherche, la logique des méthodes spécifiques à l'internet n'est toutefois pas poussée jusqu'à son terme. En effet, une fois le corpus rassemblé, et alors que sa matérialité numérique autoriserait des traitements automatiques, le choix est au contraire fait de retourner à des méthodes d'analyse beaucoup plus traditionnelles. À cet égard, une opération est révélatrice : les articles, après avoir été partitionnés en deux échantillons (482 articles pour les sites de journalisme participatif ; 480 pour les sites de quotidiens), sont imprimés sur des feuilles séparées. Chaque article est alors soumis à l'expertise d'étudiants recrutés pour en identifier les thématiques (économie et finance, international, environnement, etc.), en comptabiliser les hyperliens, et en relever les éléments multimédias (extraits sonores et vidéos). Une fois passés au tamis de cette grille d'analyse, le contenu et la structure des articles sont appréhendés sous forme de statistiques pour les besoins de la comparaison. En ce qui concerne les thématiques abordées par exemple, on apprend ainsi que les sites de journalisme participatif tout comme les sites de quotidiens « papier » se concentrent sur les affaires gouvernementales, y consacrant respectivement 27 % et 22 % de leurs articles, mais que les premiers privilégient ensuite le divertissement (16 % de leurs articles), tandis que les seconds accordent une place équivalente à l'économie et à la finance (16 % des articles aussi). Plus généralement, en prenant en compte la répartition de l'ensemble des thématiques abordées par chaque type de site, la recherche en vient à conclure que l'agenda des sites participatifs est légèrement plus équilibré que celui des sites de quotidiens « papier ». En guise de bilan, on retiendra l'intérêt principal résidant dans le spectre d'observation : ont été comparés les sites participatifs et leurs équivalents *mainstream* dans l'ensemble des États-Unis. Avec une rapidité de collecte des données sans équivalent au regard de ce qu'aurait été le travail de documentation et d'archivage papier ou audiovisuel. Mais, à partir d'un corpus aussi intéressant, l'analyse s'avère assez sommaire, se limitant notamment, à une identification des grandes thématiques abordées sans rentrer dans le détail des événements médiatiques et de leur mode de traitement par chaque publication. Bref, l'analyse n'est pas très fouillée, faute sans doute d'une insuffisance de moyens affectée à l'observation de ce très volumineux corpus. Car cela en est bien la contrepartie : si l'internet permet la collecte de gros volumes de données, leur analyse n'en est que plus conséquente. Face à ce défi, des ressources humaines plus consistantes auraient pu constituer une solution. Surtout, des méthodes de traitement automatisé, complément en quelque sorte logique aux procédés informatiques de collecte des données, auraient pu être

mises en œuvre. C'est le cas dans d'autres recherches où la démarche des *digital methods* est plus aboutie.

Les *digital methods* en action : tracer la propagation de l'information entre sites et blogs

La recherche sur laquelle nous allons maintenant nous pencher a été présentée en 2009 par Jure Leskovec, Lars Backstrom et John Kleinberg, qui tous officient dans des départements d'informatique (*Computer science*), le premier à Stanford et les deux autres à Cornell. Bien que spécialistes d'informatique, ces trois chercheurs placent très clairement leur travail dans le domaine de l'étude des médias, se posant comme initiateurs d'une alternative quantitative aux recherches qualitatives menées jusqu'à présent en SHS. Sur le point particulier du cycle des nouvelles, c'est-à-dire du renouvellement des événements majeurs portés à la « Une » de l'actualité et de leur circulation entre les différents médias, ils entendent ainsi prolonger et dépasser les recherches accomplies jusqu'ici :

> « *While the dynamics of the news cycle has been a subject of intense interest to researchers in media and the political process, the focus has been mainly qualitative, with a corresponding lack of techniques for undertaking quantitative analysis of the news cycle as a whole* »[5] (Leskovec *et al.*, 2009 : 1).

Pour atteindre un tel objectif, l'artillerie lourde des *digital methods* est déployée, tant sur le plan de la constitution du corpus, que sur celui de l'analyse automatisée et de la visualisation des résultats obtenus. Le corpus affiché est énorme : durant une période de 3 mois, ceux précédant l'élection présidentielle de 2008 aux États-Unis, environ 20 000 sites *web* (les sites figurant dans le répertoire *Google News*) et 1,6 millions de « *blogs, forums and other media* » ont été explorés à partir de l'API (interface de programmation) Spinn3r. Plus de 90 millions de documents, articles de sites et billets de *blogs*, auront été ainsi collectés. Enfin, dernière étape dans cette constitution de corpus, ont été extraits de cette somme de documents *web* environ 112 millions de citations (discours rapportés à l'intérieur des articles et billets), toutes soumises à un traitement automatisé. Ainsi les citations ont-elles été regroupées en fonction de leurs proximités lexicales. Les phrases identiques ainsi que les segments de phrase identiques ou voisins ont constitué des agrégats (*clusters*) qualifiés de *memes* par les auteurs. Ces *memes* sont censés représenter une idée semblable ou un « concept », celui exprimé dans une citation et ses multiples variations. À partir de là, sont comptabilisés les articles et billets relatifs à chaque *meme*. Un des principaux résultats issus de cette observation réside dans l'identification des 50 principaux *memes* et de leur succession au cours du temps. Ce résultat est représenté sous forme graphique dans le texte scientifique

[5] « Bien que le cycle des nouvelles soit un sujet d'intérêt fort pour les chercheurs spécialistes des médias et des processus politiques, la focale a été principalement qualitative, marquée par un manque de techniques pour entreprendre une analyse quantitative du cycle des nouvelles dans sa totalité ».

soumis ici à examen, il est également proposé à la manipulation « *interactive* » de l'internaute sur le site *Meme-tracker* attenant à la recherche (figure 1). Ici, on remarquera le soin apporté à la visualisation des résultats, fréquent avec les *digital methods*.

Figure 1 : Citations les plus reprises durant la campagne présidentielle de 2008[6].

Ce graphique laisse apparaître quelles ont été les petites phrases de la campagne présidentielle les plus reprises sur le *web* ainsi que la dynamique de leur apparition, montée en puissance, puis disparition. D'autres résultats sont également intéressants au regard de la question de la diversité de l'information, et de son évolution avec l'internet. En effet, une partie de la recherche va s'atteler à découper le corpus entre sites de *mainstream media* et *blogs*. Comparant leur réactivité respective vis-à-vis des 1 000 *memes* les plus importants en volume, les chercheurs calculent un écart moyen de 2,5 heures à l'avantage des sites de *mainstream media* et en concluent à un certain suivisme des blogs. En complément, il apparaît que les blogs ne sont à l'initiative que de 3,5 % des principaux *memes* propagés sur l'internet : leur contribution à un renouvellement de la diversité de l'information, au regard des *mainstream media*, est alors fortement relativisé par les auteurs. Ces différents résultats sont plutôt stimulants en première analyse, et leur rendu chiffré quelque peu séduisant et novateur à l'aune des pratiques habituelles en étude des médias. Toutefois, ils ne sont pas exempts de critiques, loin de là. L'analyse reste assez rudimentaire malgré la taille du corpus et la sophistication des outils employés. Et les conclusions tirées des observations souffrent parfois de coupables raccourcis. Pour rentrer dans le détail de ces critiques, on peut commencer par le corpus : s'il est colossal, il n'est pas complètement contrôlé. Les 20 000 sites catalogués comme sites de *mainstream media* sont issus de *Google News*, dont le répertoire contient

[6] Graphique extrait du site http://memetracker.org/. Consulté le 15/03/2011.

certes des sites généralistes à large audience (à commencer par les émanations *web* des principaux journaux et des grandes chaînes de télévision) mais abrite aussi quantité de sites *pure players*, pour certains participatifs ou très spécialisés (information de niche). Quant aux blogs, c'est un véritable travail de fourmi qu'il faudrait engager pour s'assurer de la conformité de chacun des 1,6 millions de blogs aux objectifs de la recherche, plutôt que de se reposer entièrement sur l'expertise et les critères de sélection de *Spinn3r*. En résumé, la cohérence du corpus est loin d'être assurée, et la revendication d'une comparaison entre sites de *mainstream media* et *blogs* d'information semble quelque peu hasardeuse. Une seconde critique tient aux critères employés pour l'analyse de contenu pratiquée. Entièrement fondée sur des similarités lexicales (reformulations d'une suite d'occurrences verbales), elle est présentée comme atteignant un stade sémantique à travers la notion de *meme*. C'est bien à partir de cette notion que sont distingués ou rapprochés les articles et billets du corpus. Or, le contenu de ceux-ci ne se réduit pas forcément à la reprise plus ou moins variée d'une petite phrase. Celle-ci peut être noyée dans un long article abordant un tout autre sujet, être connotée de façon positive ou négative en fonction de l'orientation politique du site ou du *blog*, etc. Si les chercheurs restent le plus souvent prudents dans leurs propos en parlant d'agrégats de citations, ils franchissent parfois le pas en prétendant à l'analyse de *stories* (récits ou sujets d'actualité) et affichent tout de même l'analyse du cycle des nouvelles comme objectif d'ensemble. Une troisième et dernière critique relèverait d'une autre prétention, celle de rendre compte de pratiques sociales par-delà les données et traces numériques laissées sur le *web*. Dès l'introduction de leur article, les auteurs mentionnent l'objectif d'examiner la circulation de l'information à la fois sur l'internet et entre les individus : « *Which propagation on the web and between people typically occurs ?* »[7] (Leskovec *et al.*, 2009 : 1). Ils affirment par ailleurs que la citation (ou le *meme*) constitue un niveau pertinent d'analyse du rapport des internautes à l'information – « *it is at this intermediate temporal and textual granularity of memes and phrases that people experience news and current events* »[8] (*ibid.*) – sans mentionner aucune enquête sociologique de réception à ce sujet ni la nécessité de devoir en mener. Plus encore, cette conception a-sociologique des phénomènes médiatiques est attestée dans les velléités de modélisation des cycles de propagation de l'information. D'abord, les modèles de variation lexicale des citations sont inspirés d'analogies avec les mutations des signatures génétiques. Ensuite, de façon plus large, les sciences du vivant sont convoquées dans une vision darwinienne de la circulation des nouvelles expliquant pourquoi certaines en viennent à remplacer d'autres :

[7] « Comment se déroule généralement la propagation sur le web et entre les individus ? ».

[8] « C'est dans cette granularité temporelle et textuelle intermédiaire des *memes* et des citations que les gens saisissent les nouvelles et les évènement courants ».

> « *One could imagine the news cycle as a kind of species interaction within an ecosystem [...], where threads play the role of species competing for resources (in this case media attention, which is constant over time), and selectively reproducing (by occupying future articles and posts)* »[9] (*ibid.* : 6).

Ces remarques laissent entrevoir certaines des impasses des *digital methods*, lorsque ces dernières prétendent déduire de macroanalyses de sites web les pratiques sociales afférentes. Toutefois, le développement des services dits de réseaux sociaux – dans lesquelles l'appropriation de l'actualité par les internautes peut s'accomplir – et le développement d'outils permettant d'en enregistrer les traces, ont ouvert la voie à des recherches complémentaires. Nous allons le voir à propos de *Twitter*, celles-ci ambitionnent de rendre compte des comportements des internautes en matière de production et de diffusion de l'information sur l'internet.

Les *digital methods* poussées à l'extrême : *Twitter* comme objet et comme moyen d'observation

Dans ce dernier travail, mené par trois chercheurs en informatique des *hp Labs* (Sitaram Asur, Bernardo A. Huberman et Gabor Szabo) et un chercheur en physique appliquée de Stanford (Wang Chunyan), l'automatisation est encore plus poussée. Comme dans les travaux étudiés précédemment, les *digital methods* sont ici aussi employées pour constituer un très vaste corpus. Mais la part humaine très restreinte au niveau de l'analyse du contenu de l'information en ligne distingue ce dernier travail qui, pourtant, se targue d'observer les comportements des internautes en matière de transmission de l'information. En effet, cette recherche vise à décrire la circulation de l'information en ligne, à travers l'exemple de la plateforme *Twitter*. Il s'agit de voir quels sont les sujets les plus discutés dans les *tweets*, et surtout d'expliquer les processus de focalisation puis de dilution de l'attention sur ces sujets. À ce titre, les auteurs réinscrivent leur visée scientifique dans des problématiques très courantes de l'étude des médias, évoquant très explicitement les théories de l'*agenda-setting* pour mieux cerner le rôle des *social media* comme *Twitter*. Pour les besoins de l'étude, un échantillon de plus de 16 millions de *tweets* a été recueilli, sur une période de 40 jours, entre septembre et octobre 2010. À partir d'une requête effectuée toutes les 20 minutes sur l'API de *Twitter*, ont ainsi été extraits le texte des *tweets* considérés (texte limité à un maximum de 140 caractères), le nom des comptes *Twitter* destinateurs, ainsi que l'heure de publication. Ces *tweets* ont été sélectionnés sur la base de mots-clés, correspondant aux *trending topics* affichés périodiquement par le service *Twitter* lui-même. Il faut donc bien comprendre que la catégorisation en *trending topics*

[9] « On pourrait se représenter le cycle des nouvelles comme une sorte d'interactions entre espèces à l'intérieur d'un écosystème [...], dans lequel les fils [de nouvelles dominantes] jouent le rôle d'espèces luttant pour des ressources (dans le cas présent, l'attention des médias, constante au cours du temps) et se reproduisant de façon sélective (en prenant place dans les futurs articles et billets) ».

effectuée par *Twitter* est reprise telle quelle par les chercheurs. Ces derniers n'interrogent pas véritablement les critères d'élaboration de cette catégorie, se contentant de supposer qu'elle est vraisemblablement fondée sur une analyse de fréquence des occurrences verbales au sein des *tweets*. Ils interrogent encore moins la qualité de la classification opérée par *Twitter*, la jugeant digne de confiance et la reprenant donc à leur propre compte. Ainsi les *trending topics* sont-elles considérées comme une catégorie équivalente à celle des sujets d'actualité dominants dans l'agenda, et la mesure effectuée par *Twitter* comme fiable :

> « *The trending topics, which are shown on the main website, represent those pieces of content that bubble to the surface on Twitter owing to frequent mentions by the community. Thus they can be equated to crowdsourced popularity* »[10] (Asur *et al.*, *op. cit.* : 1).

Cet appui sur les catégorisations et mesures livrées directement par les acteurs de l'internet n'est pas rare dans l'emploi des *digital methods*. Par exemple, plusieurs des applications de la *Digital Methods Initiative* s'appuient sur des services de *Google*, avec toutefois la justification de vouloir précisément interroger le rôle central de cet acteur dans la configuration de l'internet, sa prééminence dans la mise en visibilité et l'organisation de l'information en ligne. Dans les SHS (et dans d'autres disciplines scientifiques), l'exploitation de données de seconde main se pratique également de façon courante, tout en s'accompagnant de précautions méthodologiques quant au statut des données récupérées. Ici, dans cette étude sur *Twitter*, rien de tel : l'intervention des chercheurs se trouve reléguée au traitement de données, sur la base de catégories et de mesures réalisées par des tiers, très peu interrogées et pas contrôlées du tout. Cet écueil méthodologique pose d'autant plus problème que la présente recherche, tout en accentuant la part automatisée et non contrôlée de l'analyse, prétend pouvoir en tirer des enseignements quant aux pratiques des internautes en matière de diffusion de l'information. En effet, elle s'intéresse au rôle joué par les twittonautes, essayant de déceler quels sont les plus influents dans la propagation des *trending topics* sur *Twitter*. Une telle problématique place ce travail dans le domaine de la diffusion interpersonnelle de l'information. Il s'agit d'un domaine dont les origines remontent aux premiers travaux empiriques en étude des médias autour de la notion de « leader d'opinion », un domaine entre-temps beaucoup développé dans les sciences de gestion et le marketing (*word-of-mouth*), et connaissant une nouvelle vigueur avec l'idée de « marketing viral » à laquelle la « nouvelle science des réseaux » tente d'appliquer ses modèles (Mellet, 2009). Dans cette recherche, nous ne comptons donc pas nous en tenir aux seuls facteurs de la nouveauté et de la concurrence entre les *topics* pour expliquer la focalisation sur tel ou tel d'entre eux, même si l'on reprend la métaphore de la compétition sélective héritée des sciences du vivant : « *On Twitter each topic competes with*

[10] « Les *trending topics*, qui apparaissent sur le site web principal, représentent ces morceaux de contenu qui émergent à la surface de *Twitter* en raison de leur reprise récurrente au sein de la communauté. Ils peuvent donc être assimilés à une forme de popularité issue de la foule [des *twittonautes*] ».

the others to survive on the trending page »[11] (Asur *et al.*, *ibid.* : 5). On ajoute à cela un facteur présenté comme autrement plus déterminant : le rôle des utilisateurs. À partir de là, sont recherchés les utilisateurs les plus influents, les propagateurs principaux de l'information, qualifiés de « *trend-setters* » à la manière des *agenda-setters* et autres *gatekeepers* en étude des médias. Après des calculs portant sur les comptes *Twitter* le plus souvent présents dans la propagation des *trending topics*, les chercheurs avancent que les variables de l'activité (nombre de *tweets* postés par jour) ou de la taille du réseau de correspondants (le nombre de *followers*) ne sont pas significatives. En revanche – et ceci est présenté comme la principale découverte de la recherche – figurent parmi les *comptes* les plus fréquemment « re-tweetés » à la fois des comptes difficilement identifiables (Vovo Panico, Keshasuja, LadyGonga, etc.) et aussi plusieurs comptes évoquant des noms d'organisations médiatiques (Cnnbrk, Nytimes, Espn, El Pais, Reuters, etc.).

De ce constat, est tiré l'enseignement suivant : « *This illustrates that social media, far from being an alternate source of news, functions more as a filter and an amplifier for interesting news from traditional media* »[12] (*ibid.* : 7). Fort stimulante, car contre-intuitive vis-à-vis d'hypothèses imaginant une information en ligne plus diversifiée grâce à la multiplication de sources nouvelles sur l'internet, une telle conclusion mériterait d'être formulée avec beaucoup plus de précaution. D'abord pour les raisons déjà énoncées : les *trending topics* constituent une catégorie créée et mesurée par le service *Twitter* lui-même, sans renseignements précis sur les conditions de son élaboration. Ensuite, parce que cette conclusion est fondée sur deux extrapolations ne reposant sur aucune investigation sérieuse. D'une part, l'intitulé du compte *Twitter* est pris comme indicateur de l'identité de son propriétaire, mais les chercheurs ne démontrent pas que CNN est bien derrière le compte Cnnbrk ou que le compte Vovo Panico est celui d'un amateur et non d'une structure professionnelle d'information. D'autre part, même en admettant que certains comptes appartiennent bien à des organisations médiatiques, ces dernières peuvent propager des *tweets* dont elles ne sont elles-mêmes pas forcément à l'origine. Autrement dit, il faudrait une analyse beaucoup plus resserrée des *tweets* et de leurs auteurs pour voir qui est cité puis éventuellement re-cité, qui est véritablement à l'origine de l'information. On réalise ici toute la nécessité de conduire une enquête de type sociologique auprès des acteurs propriétaires des comptes *Twitter* et, de façon plus générale, l'incomplétude d'une logique exclusive de *digital methods*.

[11] « Sur *Twitter*, chaque sujet lutte avec les autres pour survivre sur la page des *trending topics* ».

[12] « Ceci illustre le fait que les médias sociaux, loin d'être une source alternative d'information, agissent plus comme filtres et amplificateurs des nouvelles en provenance des médias traditionnels ».

Retours d'expérience : une recherche collective et interdisciplinaire sur le pluralisme de l'information en ligne

Dans le cadre d'un travail collectif sur le pluralisme de l'information en ligne[13], nous avons expérimenté les *digital methods* pour des raisons analogues à celles mises en avant dans les recherches présentées précédemment : la possibilité d'accès à un corpus quasi-exhaustif – l'ensemble des sites français d'information générale et politique dans le cas présent, soit plus de 200 sites – et les nécessités de systématisation du traitement afférentes à la taille d'un tel corpus. Mais nous avons rencontré assez rapidement les limites pratiques et théoriques de l'emploi de telles méthodes. Ceci a conduit à les hybrider avec des démarches beaucoup plus contrôlées scientifiquement et éprouvées dans le domaine de l'étude des médias : le traitement des données n'a été finalement que semi-automatisé, laissant une grande place à l'expertise humaine des chercheurs en SHS ; l'analyse quantitative s'est complétée d'un indispensable volet qualitatif pour atteindre une connaissance plus fine du pluralisme de l'information ; une analyse socioéconomique des modalités de production et de consommation de l'information a été ajoutée pour mieux saisir les résultats de l'analyse de contenu. Nous proposons de développer ces éléments à travers un témoignage réflexif. Celui-ci ne prétend pas au statut d'auto-analyse rétrospective, puisqu'il porte sur un travail en cours de réalisation et n'a pas été soumis à un strict protocole d'objectivation. Sans pouvoir parvenir à ce stade, nous nous efforcerons toutefois de livrer un retour d'expérience présentant aussi bien les opportunités que les difficultés rencontrées. Avec une volonté double : prolonger la réflexion méthodologique sur les *digital methods*, et signaler aux chercheurs tentés par cette voie – dont tout laisse à penser qu'elle est amenée à se développer – les écueils possibles mais aussi quelques pistes intéressantes à suivre.

Mêler traitement informatique et intervention humaine : utiliser les *digital methods* de façon contrôlée

La visée principale de cette recherche est d'évaluer le niveau de pluralisme de l'information en ligne. Plus précisément, le pluralisme de l'information est considéré au regard de la diversité des sujets d'actualité abordés sur l'internet. Il s'agit donc d'examiner quels sujets sont mis à la « Une » de l'agenda médiatique numérique, de voir si une pluralité d'informations découle de la multiplicité des sites (sites participatifs, *blogs*, *webzines* d'opinion, en sus des extensions numériques de journaux, radios, télévisions) ou si, à l'inverse, une certaine redondance émane

[13] Programme de recherche IPRI – Internet, pluralisme et redondance de l'information (ANR-09-JCJC-0125-01b), soutenu par l'Agence nationale de la recherche et regroupant des laboratoires en information-communication (CIM, université Paris 3 ; ELICO, université de Lyon ; LERASS, université Toulouse 3 ; CRAPE, université Rennes 1 ; GRICIS, UQAM Montréal) et en informatique (LIRIS, INSA Lyon).

du travail de compilation effectué par certains acteurs dominants de l'internet (portails, agrégateurs de nouvelles). Le caractère éclaté de l'internet, en une myriade de sites, est à la base de la problématique et des hypothèses de notre recherche. Par voie de conséquence, celles-ci ne peuvent être validées ou infirmées empiriquement qu'à la condition d'une prise en compte de l'ensemble des espaces de publication et de diffusion de l'information en ligne. Les *digital methods* permettent d'atteindre un tel objectif. Elles ont été employées pour rassembler l'information dispersée sur de nombreux sites, et analyser la masse très volumineuse de données ainsi collectées. Pour autant, à chacune de ces étapes, constitution du corpus comme analyse de contenu, le recours aux *digital methods* s'est avéré nécessaire mais non suffisant : continuellement, l'expertise des chercheurs spécialistes de l'étude des médias a été sollicitée pour statuer sur les différents choix en amont et interpréter les résultats des processus d'automatisation en aval. Les méthodes de constitution du corpus de sites sont tout à fait révélatrices d'une telle imbrication. Recours aux outils informatiques et intervention humaine ont été croisés dans les trois méthodes qui ont permis de délimiter un corpus de 209 sites d'information d'actualité en février 2011, corpus que l'on peut considérer comme quasi exhaustif pour ce qui est de l'information générale et politique au niveau de la France.

La première méthode a consisté à explorer les répertoires de sites d'information d'actualité disponibles en ligne (répertoires de *Google Actualités*, IPLJ, *Rezo*, et *Wikio*), en ne retenant que les sites correspondant au périmètre préalablement défini (sites français d'information générale et politique). La deuxième a concerné plus directement une catégorie de sites, les *blogs*, particulièrement nombreux et difficiles à repérer. À partir d'une liste-racine de blogs déjà identifiés, le logiciel *Navicrawler* a été utilisé pour découvrir les sites voisins liés hypertextuellement. Leur adéquation avec la définition d'information générale et politique a ensuite été vérifiée pour autoriser ou non leur intégration au corpus, comportant 111 *blogs*. Une troisième méthode s'est appuyée sur l'analyse des liens pointés *via Twitter* : les URL pointés par 22 000 comptes, comptes identifiés dans la recherche comme étant particulièrement en prise avec l'actualité générale et politique, ont été mis en correspondance avec les URL des sites déjà présents dans le corpus. Ainsi celui-ci a-t-il pu être augmenté de quelques sites grâce à cette troisième méthode de nature plus inductive. Au niveau de l'analyse de contenu, les *digital methods* se sont révélées un appui à l'intervention humaine plutôt qu'une solution exclusive. Au tout début du projet de recherche, il avait pourtant été envisagé, notamment par des collègues du laboratoire d'informatique, de réaliser une analyse entièrement automatisée du contenu des articles. Ceci était notamment motivé par le volume d'articles produits au sein de notre corpus de 209 sites : environ 3 500 articles d'information générale et politique, en moyenne, paraissent chaque jour. Mais cette perspective d'automatisation complète de l'analyse de contenu s'est heurtée à deux obstacles majeurs. Le premier concerne la collecte des données : les articles sont difficilement isolables au sein d'une page web par un logiciel d'aspiration qui a tendance à

charrier l'ensemble des contenus (articles, mais aussi publicités, liens, etc.). Une alternative plus commode réside alors dans la collecte des flux RSS d'information générale des sites où chaque article est identifiable par son titre, son chapeau, sa date de publication, et son URL. Cette solution a été retenue pour la collecte automatique des données, mais elle ne permet pas d'accéder à la totalité du contenu de l'article : au moment de l'analyse, si le chercheur ne parvient pas à identifier le sujet de l'article à partir de son seul titre ou de son chapeau, alors il doit se rendre sur l'URL d'origine et consulter la totalité du texte. Le second obstacle à cette automatisation intégrale tient à l'objectif de la présente analyse de contenu. En effet, ce qui est recherché est le sujet d'actualité traité dans l'article, correspondant au cadrage médiatique primaire d'une expérience factualisée, au-delà de son cadrage médiatique second – angle, opinion, etc. (pour une définition détaillée, voir Marty *et al.*, 2010). Or, l'identification du sujet d'actualité d'un article est une opération qui n'est pas réalisable par une machine et un logiciel, contrairement, par exemple, au repérage de la duplication d'une citation et des ses variantes tel qu'effectué par le *Meme-Tracker* (Leskovec *et al.*, 2009). Au mieux, les *digital methods* peuvent servir d'aide à la classification, pour permettre un premier défrichage de l'énorme volume d'articles : le logiciel de collecte et traitement des données spécialement développé pour ce programme de recherche (logiciel IPRINA-IPRI *News Analyzer*[14]) dispose ainsi d'un module d'identification des segments répétés qui permet de rassembler les articles dont les titres et les chapeaux présentent des similarités lexicales ; en parallèle, les *clusterings* d'articles effectués par *Google Actualités* sont également récupérés comme indicateurs de possibles sujets d'actualité dominants. Mais, s'il permet de préparer le terrain et de gagner un temps précieux, ce premier débroussaillage ne se substitue pas au travail humain d'identification des sujets d'actualité : tous les articles, y compris ceux pré-indexés automatiquement, sont soumis à l'analyse de chercheurs spécialistes de l'étude des médias. Ainsi les *digital methods*, même si elles sont mobilisées dans ce travail d'analyse de contenu, sont-elles encadrées par l'expertise humaine qui permet de rester en quelque sorte maître de la recherche. Et cela distingue cette recherche de celles examinées précédemment : par exemple, là où les *trending topics* de *Twitter* sont adoptés tels quels comme catégories valides d'analyse (Asur *et al.*, 2011), ici la pertinence des *clusters* de nouvelles de *Google Actualités* est évaluée et au besoin rejetée afin de garder une cohérence vis-à-vis des objectifs et catégories d'analyse propres au programme de recherche (notion de sujet d'actualité). Cependant, il faut le savoir, sur d'aussi vastes volumes de données, le contrôle des *digital methods* et leur croisement avec une analyse « à la main » a un coût, justement humain. À titre d'illustration, l'identification des sujets d'actualité et la classification des 3 500 articles d'une journée requièrent près d'une semaine de travail à temps plein, répartie entre les deux chercheurs qui assurent la procédure d'intercodage.

[14] Logiciel IPRINA mis à disposition sous licence *Creative Commons*. Accès : http://liris.cnrs.fr/ipri/pmwiki/index.php?n=Public.IpriNA.

Mixer le quantitatif et le qualitatif, l'analyse de contenu et l'analyse sociologique : compléter les apports des *digital methods*

L'identification des sujets d'actualité au sein du corpus permet de dessiner l'agenda médiatique de l'internet, sur une période donnée. Nous pouvons même avoir une estimation quantitative de son degré de pluralisme, en reprenant certains critères élaborés pour mesurer la diversité culturelle (Benhamou, Peltier, 2006). Appliqué au cas de l'information en ligne, le critère de variété correspond ainsi au nombre de sujets abordés. Et un second critère de diversité, l'équilibre, correspond lui à la répartition des sujets entre les différents articles. À travers une première observation exploratoire, réalisée sur deux journées de novembre 2008, nous avons pu constater une grande variété des sujets abordés sur l'internet (plus de 300 pour chacune des journées) et en même temps un grand déséquilibre de l'information en ligne : quelques sujets ultra-dominants se retrouvent dans de très nombreux articles tandis que les autres sujets ne sont abordés que de façon isolée (dans un article ou une poignée d'articles). Plus précisément, une telle répartition où 20 % des sujets occupent 80 % de l'agenda médiatique de l'internet semble typique des distributions *parétiennes*, et peu innovante en regard de certaines hypothèses de diversité accrue sur l'internet, énoncées par exemple sous la formule de *Long Tail* (Smyrnaios *et al.*, 2010). Au cours du printemps 2011, une nouvelle observation est menée à plus grande échelle (temporalité d'un mois). Elle reprend les critères quantitatifs de variété et d'équilibre, tout en ajoutant un critère supplémentaire, de nature qualitative cette fois : la disparité. À l'intérieur d'un même sujet d'actualité, il s'agit d'analyser la disparité de traitement journalistique entre tel ou tel article, entre tel ou tel site. Après identification des différentes classes de discours portés sur un sujet d'actualité dominant (explosion dans une centrale nucléaire au Japon ou intervention militaire aérienne en Libye), un échantillon d'une trentaine d'articles, représentatif de ces classes de discours et des différentes catégories de sites, est constitué. Il est alors l'objet d'une analyse sémiopragmatique (modes d'énonciation éditoriale, relations texte-image, représentations du lecteur, registres discursifs) qui vise à cerner les différents cadrages (cadrages seconds), opinions et points de vue exprimés.

Une telle démarche permet d'atteindre un stade d'évaluation du pluralisme (la disparité de traitement journalistique) plus fin que de seules mesures quantitatives (variété et distribution des sujets). Cette analyse qualitative et sémiopragmatique de discours apporte donc des compléments aux analyses quantitatives et souvent purement lexicales autorisées par les *digital methods*. Et celles-ci fournissent en retour une assise scientifique supplémentaire aux analyses de discours plus traditionnelles. Ces analyses sémiopragmatiques de nature qualitative, puisqu'issues d'un échantillon représentatif de l'ensemble des articles produits sur un même sujet (plusieurs centaines pour les sujets ultra-dominants du jour), peuvent prétendre à une portée et à une généralisation

jusqu'ici difficiles à affirmer dans le domaine de l'étude des médias. Une telle complémentarité entre quantitatif et qualitatif, appuyée sur la combinaison entre *digital methods* et méthodes plus traditionnelles, peut ainsi enrichir l'analyse de la diversité de l'information en ligne. Et cette analyse de contenu peut elle-même se doubler d'une approche plus sociologique des modalités de production, diffusion et réception de l'information en ligne, afin d'investiguer les points laissés aveugles par les *digital methods* – et éviter les extrapolations parfois douteuses que nous avons pu pointer dans l'examen précédent de certaines recherches.

Par exemple, lors de la phase exploratoire de 2008, nous avons tenté de coupler notre analyse quantitative de contenu avec des enquêtes (entretiens et observation directe) auprès des rédactions de sites *web* de notre corpus. Ceci a permis de distinguer deux modalités de production (Damian *et al.*, 2009). D'un côté, des rédactions « productivistes » doivent couvrir l'ensemble des domaines de l'actualité du jour avec des moyens limités et ont donc tendance à s'appuyer sur des matières informationnelles déjà produites par des tiers (agences, autres médias). De l'autre côté, des rédactions « anglistes » choisissent délibérément de ne pas suivre l'agenda médiatique général ou bien de traiter certains sujets avec un angle ou un approfondissement particulier. Une corrélation entre cette dualité des rédactions, et la tendance ambivalente à la variété et au déséquilibre de l'information en ligne, a alors pu être envisagée. Les enquêtes au niveau des conditions de production viennent ainsi éclairer les résultats de l'analyse de contenu. Au cours de l'année 2011, l'analyse de contenu sera mise en regard des modalités de diffusion et de réception de l'information. Une analyse des articles recommandés et mis en circulation sur *Twitter* d'une part, et une prise en compte de l'audience des sites d'autre part, donneront des indications quant aux sujets d'actualité qui retiennent l'attention des internautes. En d'autres termes, une certaine mesure du pluralisme « consommé » (Benhamou, Peltier, 2006) pourra ici être comparée avec celle du pluralisme « offert » fournie par l'analyse de contenu. Toutefois, de telles indications statistiques ne formeront qu'une analyse partielle de la réception de l'information en ligne. Pour être pleinement étudiée, cette dernière mériterait de recourir à des techniques d'enquête beaucoup plus qualitatives[15]. Nous touchons donc l'une des limites de la présente recherche collective qui, faute de moyens humains suffisants pour mener une enquête qualitative de réception en propre, devra essentiellement s'appuyer sur le traitement de données secondaires ou sur la littérature existante[16]. Néanmoins, même de manière imparfaite, cette recherche s'efforcera de replacer les résultats de l'analyse de contenu, en partie appuyée sur les *digital methods*, dans le contexte social de production, diffusion et consommation de l'information en ligne.

[15] Voir par exemple l'étude en réception, expérimentée par V. Jeanne-Perrier (2010), de la réappropriation des contenus médiatiques par les utilisateurs de *Twitter*.

[16] À cet égard, le travail ethnographique de P.-J. Boczkowski (2010) décrivant à la fois la production en flux tendu dans certaines rédactions *web*, et la réception majoritairement située sur le lieu de travail de la part des internautes, sera d'une grande aide pour interpréter l'aspect factuel et mimétique de l'information en ligne.

Conclusion

Les exemples rassemblés ici illustrent l'emprise croissante, dans l'étude des médias et en particulier de l'information en ligne, des *digital methods*. Celles-ci ont permis de parvenir à un niveau d'exhaustivité dans les corpus observés, niveau inatteignable sans leur concours. Elles ont concomitamment fourni des solutions de traitement automatique adaptées à la taille des vastes jeux de données collectés sur l'internet. Enfin, sur le plan de la visualisation des résultats, elles offrent également une palette d'outils à l'ergonomie très travaillée, rendant la lecture beaucoup plus assimilable. Pour autant, l'examen détaillé de certaines recherches montre que les *digital methods* ne peuvent se suffire à elles-mêmes, en tout cas lorsqu'elles sont appliquées dans une perspective a-sociologique en phase avec la philosophie première de la « nouvelle science des réseaux ». De fait, la perspective d'une automatisation totale de l'analyse s'avère quelque peu illusoire, sauf à accepter des catégorisations souvent approximatives et un certain impressionnisme quantitativiste :

> « La réalité comme réalité vécue par des êtres humains qui font partie de différents contextes et cultures – la réalité telle qu'elle intéresse les SHS – revendique une épaisseur qui ne se laisse pas si facilement aplatir. Par nécessité méthodologique, la NSR [nouvelle science des réseaux] est donc forcée d'épurer un phénomène pour le faire entrer dans sa grille d'analyse formelle » (Rieder, 2007 : 15).

Dans bien des cas examinés ici, les compléments fournis par une approche qualitative et par une démarche d'interprétation des résultats appuyée sur les sciences humaines et sociales, s'avèrent indispensables. De façon plus générale, les limites de recherches prétendant appréhender les pratiques de communication à partir de seuls artefacts numériques doivent être prises en considération et éventuellement comblées par des observations sociologiques au plus près des acteurs, afin de ne pas confondre traces d'usages et usages sociaux. En effet, au-delà du cas de l'information d'actualité ou citoyenne en ligne, les études portant sur l'internet de façon plus globale (pratiques de sociabilité sur les réseaux socionumériques, travail à distance, jeux en ligne, etc.) présentent une tendance à la « quantophrénie statistique » (Jouët, 2011 : 80) qui mériterait pourtant d'être contrebalancée par une remise en contexte social des phénomènes communicationnels. Cette dernière remarque, en forme de recommandation méthodologique, pourrait s'apparenter à un vœu pieux si l'on en croit les ambitions récentes des chantres de la « nouvelle science des réseaux ». En effet, c'est tout le contraire qui est visé : profiter des derniers développements technologiques, et notamment de la disponibilité des données de géolocalisation issues de l'emploi croissant des terminaux mobiles, pour retracer à partir de là les parcours des individus et leurs relations au sein de la société. Albert-Lazlo Barabasi (2009 : 413) fait de cet objectif une « nouvelle frontière » pour la désormais « science contemporaine des réseaux », dans un écrit à la fois rétrospectif et prospectif paru exactement 10 ans après les articles inauguraux publiés dans *Science* et *Nature* :

> « *Indeed, the sudden emergence of large and reliable network maps drove the development of network theory during the past decade. If data of similar detail capturing the dynamics of processes taking place on networks were to emerge in the coming years, our imagination will be the only limitation to progress. If I dare to make a prediction for the next decade, it is this : Thanks to the proliferation of the many electronic devices that we use on a daily basis, from cell phones to Global Positioning Systems and the Internet, that capture everything from our communications to our whereabouts [...], the complex system that we are most likely to tackle first in a truly quantitative fashion may not be the cell or the Internet but rather society itself* »[17].

Ainsi l'objectif est-il clairement affiché. Après avoir modélisé la topographie statique de l'internet et dévoilé un ordre mathématique derrière l'apparente « *arbitrary nature of the Web* »[18] (Huberman, 2001 : 97), il s'agit désormais de décrire les dynamiques sociales dans leur ensemble, en ligne et hors ligne. Ceci sur la base des seules traces d'usage des terminaux numériques, alors même que la manipulation de ces outils devrait en parallèle appeler des enquêtes ethnographiques restituant l'ancrage sociétal de telles pratiques. Une telle volonté de modéliser l'organisation de la société, d'en prévoir l'évolution à partir de comparaisons avec celles des réseaux cellulaires, comporte un arrière plan organiciste et néo-darwinien qui résulte autant d'une vision politique que d'un projet scientifique (Bautier, 2007). L'exportation de théorisations issues des sciences du vivant et des sciences exactes, dans des domaines d'étude relevant habituellement des sciences humaines et sociales, pourrait conduire à paradoxalement en dénier la composante humaine et sociale. Le risque n'est pas mince. La vigilance devrait même être de mise si l'on considère l'encadrement institutionnel contemporain de la science – pouvoir de preuve et d'objectivité prioritairement attribué aux démonstrations chiffrées et statistiques, exigence de « livrables » sous forme de logiciels ou de bases de données, etc. – car il s'agit du cadre dans lequel émerge progressivement le domaine des *digital humanities* ou « humanités numériques » (Rieder, Röhle, 2010). Celles-ci seront-elles alors contraintes de s'orienter vers un pôle en quelque sorte scientiste et néo-positiviste ? Nous militons au contraire pour la recherche d'un juste équilibre[19]. En effet, nous sommes à la fois convaincus du potentiel d'innovation

[17] « Assurément, l'apparition soudaine de vastes et fiables cartes du réseau a entraîné le développement de la théorie du réseau au cours de la décennie passée. Si des données aussi détaillées, captant les dymaniques des processus se déroulant sur les réseaux, viennent à apparaître dans les prochaines années, notre imagination sera la seule limite au progrès. Si j'ose émettre une prédiction concernant la prochaine décennie, c'est celle-ci : Grâce à la prolifération des nombreux appareils numériques que nous utilisons au quotidien, des téléphones mobiles aux GPS et à l'internet, qui enregistrent tout depuis nos communications jusqu'à notre localisation [...], le système complexe que nous sommes près d'aborder de façon quantitative pour la première fois pourrait être non pas la cellule ou l'internet mais bien la société elle-même ».

[18] « Nature arbitraire du *web* ».

[19] Un équilibre devra être trouvé à l'occasion d'un nouveau programme de recherche, associant des chercheurs en informatique (VIF – INA, DL Web – INA, INRIA, LIA – université d'Avignon et des pays du Vaucluse), en information-communication (CIM, université Paris 3), ainsi que deux entreprises (AFP, Syllabs). En outre, ce programme présente la particularité d'aller au-delà de l'information en ligne pour l'analyser sur tous les supports (presse, radio, télévision, internet) : programme OT-Media – Observatoire TransMedia (ANR 2010 CORD 015 06).

scientifique apporté par les *digital methods* et de leur possible intégration dans des problématiques et démarches propres aux SHS. La combinaison de ces deux approches, si elle est équilibrée est maîtrisée, est susceptible de faire progresser les connaissances concernant l'étude des médias et d'en saisir des dimensions nouvelles.

Références

Albert R., Jeong H., Barabasi A.-L., 1999, « Diameter of the world-wide web », *Nature*, 401, p. 130.

Asur S., Huberman B. A., Szabo G., Wang C., 2011, « Trends in Social Media : Persistance and Decay », *Eprint arXiv,* 1102.1402. Accès : http://www.hpl.hp.com/research/scl/papers/trends/trends_web.pdf.

Barabasi A.-L., 2009, « Scale-Free Networks : A Decade and Beyond », *Science*, 325, pp. 412-413.

Bautier R., 2007, « Les réseaux de l'internet : des artefacts bien trop vivants », *Les Enjeux de l'information et de la communication.* Accès : http://w3.u-grenoble3.fr/les_enjeux/2007-meotic/Bautier/index.html.

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

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

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

Damian B., Rebillard F., Smyrnaios N., 2009, « 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*, Athens, Panteion University. Accès : http://nikos.smyrnaios.free.fr/com_2009_New_Media_Athens_Damian_Rebillard_Smyrnaios.pdf.

Ginsberg, J., Mohebbi M. H., Patel K. S., Brammer L., Smolinski M. S., Brilliant L., 2009, « Detecting influenza epidemics using search engine query data », *Nature,* 457, pp. 1012-1014.

Huberman B. A., 2001, *The Laws of the Web. Patterns in the Ecology of Information*, MIT Press.

Huberman B. A., Adamic L. A., 1999, « Growth dynamics of the World-Wide Web », *Nature*, 401, p. 131.

Jeanne-Perrier V., 2010, « Parler de la télévision sur Twitter : une "réception" oblique à partir d'une "conversation médiatique" ? », *Communication et langages*, 168, pp. 129-150.

Jouët J., 2011, « Des usages de la télématique aux *Internet Studies* », pp. 45-89, *in* : Denouël J., Granjon F., dirs, *Communiquer à l'ère numérique. Regards croisés sur la sociologie des usages*, Paris, Presses des Mines/ParisTech.

Lacker M., Peskin C., 1981, « Control of ovulation number in a model of ovarian follicular maturation », *Lectures on Mathematics in the Life Sciences*, 14, pp. 21-58.

Lazarsfeld P.-F., Berelson B., Gaudet H., 1944, *The People's Choice : How the voter makes up his mind in a presidential campaign*, New York, Duell, Sloan and Pearce.

Leskovec J., Backstrom L., Kleinberg J., 2009, « Meme-tracking and the dynamics of the news cycle », *kdd'09 – International Conference on Knowledge Discovery and Data Mining*, Paris. Accès : http://www.cs.cornell.edu/home/kleinber/kdd09-quotes.pdf.

Marty E., Rebillard F., Smyrnaios N., Touboul A., 2010, « Variété et distribution des sujets d'actualité sur l'internet. Une analyse quantitative de l'information en ligne », *Mots. Les langages du politique*, 93, pp. 107-126.

McCombs M. E., Shaw D. L., 1993, « The Evolution of Agenda-Setting Research : Twenty Five Years in the Marketplace of Ideas », *Journal of Communication*, 43 (2), pp. 68-84.

Mellet K., 2009, « Aux sources du marketing viral », *Réseaux*, 157, pp. 268-292.

Mitzenmacher M., 2004, « A brief history of generative models for power law and lognormal distributions », *Internet Mathematics*, 1, pp. 226-251.

Newman M., Barabasi A.-L., Watts D. J., 2006, *The Structure and Dynamics of Networks*, Princeton, Princeton University Press.

Rieder B., 2007, « Étudier les réseaux comme phénomènes hétérogènes : quelle place pour la "nouvelle science des réseaux" en sciences humaines et sociales ? », Journées d'étude *Dynamiques de réseaux – Information, complexité et non-linéarité*, université de Bordeaux. Accès : http://archivesic.ccsd.cnrs.fr/sic_00379526/.

Rieder B., Röhle T., 2010, « Digital Methods : Five Challenges », *The Computational Turn Conference*, Swansea University. Accès : http://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxkbWJlcnJ5fGd4OjUwMGQzZGYxZDY3ZGUxMWU.

Rogers R., 2010, « Internet Research : The Question of Method », *Journal of Information Technology and Politics*, 7 (2/3), pp. 241-260. Accès : http://www.digitalmethods.net/.

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