La visibilité politique en ligne : Contribution à la mesure de l'e-reputation politique d'un maire urbain

Vincent Labatut, Guillaume Marrel

Big Data et visibilité en ligne
Un enjeu pluridisciplinaire de l'économie numérique
6-7 et 8 novembre 2017 – Université des Antilles, Fort de France, Martinique

Axe 4 : Influence, mesure d'exposition et méthode d'évaluation de la visibilité en ligne

La visibilité politique en ligne
Contribution à la mesure de l'e-reputation politique d'un maire urbain

Vincent Labatut (LIA-UAPV) et Guillaume Marrel (LBNC-UAPV)

Quelles traces le travail politique quotidien d'un maire urbain laisse-t-il sur la toile, en dehors des campagnes électorales ? Le Web peut-il être appréhendé comme un espace de mise en scène du travail politique ordinaire ? Comment circonscrire l'e-réputation d'une personnalité politique locale ? Peut-on identifier des modèles de diffusion de l'information concernant les événements de l'activité du maire sur le Web, les acteurs qui y prennent part, la rapidité et la structure de la propagation d'une information traitée de manière positive ou négative ? On se propose dans cette contribution de présenter les premiers résultats d'une enquête pluridisciplinaire en cours sur l'extraction et l'exploitation des données du Web concernant l'agenda politique d'un leader local.

L'enquête originale s'intéresse aux enjeux de la visibilité en ligne dans le champ politique et plus précisément à la gestion de l'e-reputation des élus du suffrage universel (Théviot, 2016). Les entrepreneurs politiques et leurs équipes sont désormais très sensibles à la nécessité de connaître et maîtriser leur identité et leur image numérique, en cours de mandat et plus encore en période de campagne (Greffet, 2012). Avec l'essor des médias Web, l'activité quotidienne d'un élu est de plus en plus reportée et commentée en ligne par son propre service de communication, les professionnels de la politique et de l'information qui l'entourent, mais aussi de plus en plus par des « citoyens ordinaires », sur les pages institutionnelles ou personnelles, les blogs, certains forums, les murs Facebook, Twitter et autres réseaux socio-numériques. Le Web peut ainsi être saisi comme un grand miroir pouvant refléter l'activité politique, mais un prisme déformant du fait de certaines réfractions. La maîtrise de cette déformation médiatique devient sans doute un enjeu stratégique de premier ordre avec l'Internet dont la dimension conversationnelle (Cardon, 2010) change le rapport des leaders aux instruments de contrôle de leur image (Roginsky et Perrier, 2014), à tel point que l'e-reputation politique fait aujourd'hui l'objet de conseils spécialisés (Frochot et Molinaro, 2010). L'objectif de la recherche est donc de parvenir à objectiver ces déformations pour ensuite les interpréter et, à terme, dégager les variables et les usages de l'écho Web-médiatique du travail politique et partant une part importante de l'activité de légitimation politique (Lagroye, 1985) au début du XXIe siècle.

1 Le projet : fouiller la légitimation politique 2.0

La contribution explore de manière expérimentale et monographique les premières étapes d'un programme de recherche pluridisciplinaire plus large qui poursuit trois principaux objectifs :

1. Donner une représentation de ce que le Web restitue sous toutes ses formes, à un moment donné, pour une période donnée, de l'activité quotidienne d'une personnalité publique telle qu'un maire urbain en France en 2017. Comment fixer de manière objective le reflet Web-médiatique par nature instable et incomparable (parce qu'individualisé) du travail politique des élus de la démocratie représentative ? Quels types d'événements transparaissent en ligne, avec quelle fréquence et selon quelle modalité ? Comment les extraire, les compiler et les filtrer de manière fiable et automatique ? Comment les représenter de manière lisible ?
2. Décrire les informations contenues dans ces éléments constitutifs de l'écho Web-médiatique du travail politique : leurs sources (identité et position des émetteurs dans les configurations d'acteurs locaux), les supports et formats, leur fréquence (effet de buzz) rapportés aux différents types d'activité publicisés et médiatisés, la polarité des textes relatifs aux faits ou à la personne du maire, etc. Qui publicise, communique, médiatise ou dévoile l'action politique, à quel rythme, sur quels supports et sur quels objets, à quelles fins ?
3. Interpréter et comprendre la nature, les formes et les transformations de cet écho Web-médiatique dans l'espace public local. Qu'est-ce que les « traces numériques » des discours, des réunions publiques, des rendez-vous plus ou moins médiatisés, des déplacements, des réceptions et autres cérémonies auxquelles l'élu participe chaque semaine, et qui sont repérables en ligne, doivent aux différents producteurs de messages en concurrence pour le contrôle de la légitimité démocratique (communication de l'équipe politique, suivi journalistique du travail édilitaire, billets ou posts des interlocuteurs de l'élu, des membres du conseil municipal, ou bien encore des acteurs militants, économiques ou associatifs et des citoyens plus ou moins mobilisés du territoire et présents en ligne, sur leurs pages Facebook, leurs blogs et les réseaux sociaux) ? Y-a-t-il un pilote de l'e-reputation politique ? Que sait-on de la réception de cet écho Web-médiatique ? Comment interpréter les trous noirs de la publicisation du travail politique ?

Le champ de la veille d'e-réputation relève de pratiques expérimentales et d'outils innovants en plein développement, issus des domaines de la fouille de texte et du traitement automatique du langage naturel (TALN). Concrètement, la veille numérique repose cependant sur des pratiques encore peu automatisées et des approches très qualitatives. Notre démarche pluridisciplinaire propose ici une méthode et un outillage exploratoire et monographique pour la reconstitution de l'écho Web-médiatique de l'agenda d'une personnalité publique locale. La contribution présente les résultats d'un dispositif de détection qui repose sur l'exploitation de pages Web textuelles renvoyées par plusieurs moteurs de recherche Web. On détecte dans celles-ci des entités nommées (personnes, lieux, organisations) et des dates, afin d'identifier les événements spatio-temporels qui y sont décrits. La version proposée ici est exploratoire, et sujette à certaines limitations dues à la difficulté des tâches de TALN abordées, notamment la résolution de coréférences (Ng, 2010). On filtre ensuite ces événements, de manière à se concentrer sur la personne ciblée et à procéder à l'analyse contextuelle des résultats.

En croisant sociologie du pouvoir local (Cadiou, 2009), analyse de la communication publique (Aldrin et al., 2014), sociologie des usages des techniques d'information (Proulx, 2015) et développement des outils informatiques de traitement du langage naturel et de détection d'événements (Aggarwal et Zhai, 2012), nous proposons ici de mieux cerner les transformations du travail collectif de légitimation du pouvoir et de l'action publique à l'heure du Web 2.0.

2 Communication et réputation en ligne d'un maire urbain

Jusqu'à la fin du XXe siècle, c'est principalement à travers la presse, la radio et la télévision que le pouvoir se donne à voir et se met en scène (Balandier, 2006)[1]. Le développement de l'Internet public et du Web au début des années 1990 vient recomposer les logiques et le sens de l'information et de la communication, mais aussi l'analyse de leurs effets et l'interprétation de leur signification[2]. Conçu initialement pour faciliter les échanges interindividuels et non pour adresser des messages à une masse de récepteurs, Internet marque l'avènement d'un média d'une nouvelle nature, à bien des égards incomparable aux médias de masses du XX^{ème} siècle (Cardon, 2010, p. 9). S'y confondent désormais la sphère de la conversation interpersonnelle ordinaire des interactions privées, et l'espace professionnalisé de l'information publique, des industries culturelles et de la politique. Le Web ouvre dans mes années 1990 l'espace public aux individus et ce faisant, il recompose les formes mêmes de la démocratie moderne (présupposé égalitaire, coopération, auto-organisation, participation, consensus...). Au fil des échanges profanes, « la société démocratique sortirait de l'orbite de la politique représentative » (Cardon, 2010, p. 99). Dans cette « démocratie électronique », les publics citoyens s'émanciperaient du paternalisme infantilisant des professionnels et des experts. Et cet espace public soit disant « ouvert, décentralisé et renouvelé » (Coleman et Blumler, 2009) fait alors l'objet de tentatives plus ou moins efficaces de reconquêtes de la part des institutions classiques : médias de masse, industries culturelles, entreprises et partis politiques. Rapidement pensé comme un projet d'émancipation politique, Internet est donc désormais réinvesti par l'ensemble des acteurs centraux et institutionnels de la vie politique des démocraties représentatives. Tous les partis sont aujourd'hui visibles et actifs en ligne. Les net-campagnes se systématisent dès 2002 en France et la « cyberisation » des organisations politiques et partisanes semble s'accélérer à chaque nouveau scrutin[3]. La plupart des travaux disponibles se focalisent néanmoins encore sur la manière dont l'Internet affecte les acteurs de la compétition électorale et concentrent donc l'attention sur les périodes de campagne et les acteurs militants, collectifs ou partisans qui les animent, au détriment des entrepreneurs politiques au travail et des temps politiques ordinaires (Nicot, 2012). De fait moins nombreuses sont les enquêtes consacrées à l'usage du Web par les élus et les entrepreneurs politiques durant leur mandat et la manière dont les formes renouvelées de communication en ligne affectent le travail politique de représentation « hors campagne » (Chibois, 2014 ; Marques, Aquino et Miola, 2014 ; Norton, 2007)[4], alors que les

1 Les médias de masse se sont alors imposés comme les relais incontournables entre le personnel politique et les citoyens dans la consolidation des régimes démocratiques, contribuant à une redéfinition des règles du jeu du débat public, de la conquête et de l'exercice du pouvoir (Bailey, 1971), autour des principes généraux de publicité et de transparence.

2 Les sciences sociales s'interrogent très rapidement sur les ressorts de la « société de l'information » et les effets des médias sur les comportements des publics, spectateurs plus ou moins passifs et variablement sensibles aux messages diffusés par les outils de propagande et de publicité. Mais elles questionnent également à d'autres niveaux les transformations médiatiques des règles et des formes de la compétition politique (personnalisation et désacralisation du pouvoir, politique-spectacle, réduction du discours politique à la « petite phrase », peopolisation…), de l'action publique (démocratie de l'opinion, construction des problèmes publics, théorie de l'agenda gouvernemental…) et du travail politique (dépendance accrue des acteurs politiques aux communicants professionnels et à leurs outils – coaching, relooking, media training – et aux commentateurs – sondeurs, journalistes et politologues).

3 La nature conversationnelle des formes politiques du Web aurait *a priori* limité l'investissement partisan institutionnel (Blondeau-Coulet et Allard, 2007). De fait, l'Internet politique est longtemps resté un Internet militant (Granjon, 2001), reposant principalement sur les sites des groupes activistes périphériques, les blogs et pages Facebook des personnalités politiques, les sites de campagne des candidats où sont valorisés l'individualisation de l'expression et l'échange conversationnel. Cette soit-disant « Web-incompatibilité » des organisations politiques, trop souvent réduits au centralisme et à l'immobilisme sclérosé, est néanmoins aujourd'hui contestée par les faits et dans les travaux de science politique (Gibson et Ward, 2009 ; Greffet, 2012).

4 Il est certes artificiel de distinguer le « temps de campagne » et le « temps du gouvernement » dans le continuum indifférencié du travail politique indistinctement « électoraliste » et « administratif » (Marrel et Payre, 2006). Il s'agit ici de cibler par convention les périodes d'activité politiques les moins affectées par l'inévitable anticipation

pratiques des élus et de leurs entourages contribuent à la restructuration « par le bas » des règles du jeu politique.

Or la communication politique en ligne dépasse largement le cadre et le moment d'intensité particulier auquel correspond la campagne électorale. Une fois élus, le député ou le maire semblent désormais[5] maintenir une importante activité de communication numérique en ligne sur leur activité de représentation, via les sites institutionnels de leurs collectivités, leurs blogs, leurs pages Facebook et leurs comptes Twitter. On émet donc une première hypothèse selon laquelle cette activité de valorisation et de légitimation par la mise en récit et la mise en scène de l'action politique, occupe une place croissante dans les collectifs de travail politique, voire dans l'emploi du temps de l'élu lui même lorsque celui-ci investit personnellement et réellement l'échange conversationnel.

De fait, l'e-marketing politique, public ou territorial devient un champ consistant (Harfouche, 2012), avec ses professionnels, ses consultants, ses formations payantes, la rationalisation de ses outils, des recommandations sur l'usage par « l'élu 2.0 » des notices *Wikipedia*, des sites internet, de *Facebook*, du microblogging *Twitter*, de *Foursquare* et de *Google+*, de *Flickr*, *Dailymotion* ou *Youtube* et des diverses applications pour mobiles[6]. Les équipes politiques prennent conscience de l'importance de leur image numérique et dédient des agents spécialisés à la communication en ligne sur l'activité et les prises de positions de leurs leaders (Carton, 2014). L'influence croissante du Web et des réseaux sociaux en politique poussent notamment à focaliser l'attention sur les enjeux de la mise en scène du travail politique et de sa réception dans ce nouvel espace Web-médiatique.

3 Hypothèses sur l'écho Web-médi@tique de l'activité de l'élu

L'enquête s'inscrit dans le cadre d'un projet de recherche plus large engagé sur les configurations de travail politique local (collectivisation, apprentissages, confusions politico-administratives, effets de trajectoires...) (Demazière et Le Lidec, 2014) et l'examen de l'agenda et de l'emploi du temps des membres d'exécutifs locaux participant directement à la production de l'action publique (Godmer et Marrel, 2014a ; Lefebvre, 2014). Il s'agit moins ici d'objectiver le quotidien des acteurs politiques, que d'analyser le jeu médiatique du « faire savoir » autour de l'action politique et à partir de l'agenda passé et à venir de l'élu, le travail de mise en scène de l'action politique, l'occupation du Web, l'allumage de contre-feux, le contrôle de l'image (Marrel, 2016). Les chargés de communication des politiques et les médias classiques communiquaient relativement peu sur l'emploi du temps des élus et sur leurs agendas. L'idéal de transparence (J Meijer et al., 2012) conduit désormais certains acteurs à jouer le jeu de la communication sur leur activité, souvent pour mieux dissimuler ce qui restera, quoi qu'il arrive, de l'ordre du secret d'alcôve. Le média Web vient potentiellement recomposer ce jeu autour de la mise en scène du travail politique, les stratégies de communication, la gestion de la transparence, le rapport au secret (Florini, 1998), jusqu'à l'organisation même de l'agenda possiblement affecté de manière rétroactive par d'éventuels calculs sur l'augmentation du risque médiatique (ne plus rencontrer tel type

---

des échéances électorales, que ce soit pour le renouvellement du mandat en cours, ou bien dans la perspective de la conquête d'un autre trophée politique cumulable ou conduisant à une « reconversion ».

5 Les travaux portant sur les années 2000 ont tendance à montrer qu'à l'Assemblée nationale « les sites ou blogs des députés n'ont qu'un rôle limité dans les stratégies électorales, que ce soit comme instrument de communication partisane ou comme outil de propagande et de mobilisation » (Nicot, 2012, p. 105) et que localement « internet n'est guère une voie de communication privilégiée par les acteurs partisans, bien que d'usage croissant » (Bué, 2011, p. 229).

6 Voir par exemple : http://www.ideose.com/pierre-guillou-web-politique-propositions-devenir-depute-20/#propositions_facebook.

d'interlocuteur susceptible de ternir l'image de l'élu, ou de perturber la mise en œuvre d'une politique…).

L'enquête est construite sur une hypothèse de représentation des différents flux de communication politique autour de l'activité d'un maire urbain. L'action politique est a priori décomposée en trois catégories :

1. activités et discours secrets, théoriquement non publicisés ;
2. activités et discours confidentiels, en principe faiblement publicisés ;
3. activités et discours publics, fortement publicisés.

Ces types d'activités et de discours sont susceptibles d'être médiatisés par plusieurs types d'acteurs et de cinq manières a priori :

1. marginalement par certaines fuites plus ou moins contrôlées au sein de l'équipe ;
2. massivement par la communication politique maîtrisée de l'élu et de son entourage ;
3. massivement par les journalistes des médias institutionnels (presse, radio, télé et Web) ;
4. par les messages et commentaires des divers acteurs experts et/ou partisans de l'espace politique autour de l'élu ;
5. par les messages et commentaires de citoyens plus ou moins ordinaires, militants ou leaders d'opinion.

L'ensemble de ces actions de médiatisation de l'activité et des paroles du maire nourrissent par hypothèse six catégories de médias sur le Web :

1. les sites et blogs officiels des institutions politiques et des collectivités publiques où circule la communication institutionnelle ;
2. les sites et les pages Web des médias de presse écrite, audio-visuelle ou Web ;
3. les blogs personnels de journalistes, d'experts, de personnalités politiques ou de citoyens plus ou moins ordinaires ;
4. les médias sociaux et en particulier *Facebook* et *Twitter* ;
5. des forums plus ou moins spécialisés ;
6. les sites d'hébergement de vidéos, comme *YouTube*.

L'ensemble des textes et des images publiés sur ces supports par tous les acteurs pré-cités, concernant l'action du maire et quels qu'en soient la véracité ou la polarité, constituent ce qu'il convient de nommer l'écho Web-médiatique de l'activité politique. C'est ce corpus théorique qu'il s'agit ici dans un premier temps d'identifier, d'extraire et de compiler.

4 Actualisation d'un dispositif de recherche

Nous avons mené une première expérience au printemps 2015, sur l'écho Web-médiatique de Cécile Helle, maire de la ville d'Avignon sur quatre semaines consécutives entre le 30 mars et le 16 avril (Marrel, Labatut et El Bèze, 2015). La recherche incluait d'une part un opérateur humain réalisant le traitement manuellement, et d'autre part un logiciel développé par nos soins et effectuant une partie du traitement de façon automatique. L'un de nos objectifs était en effet d'étudier la possibilité d'automatiser ce type de tâche, afin de pouvoir ensuite l'étendre à des données de plus grande taille (période plus longue, ciblage de plusieurs personnes, etc.).

L'approche manuelle, était destinée à servir de base de référence pour évaluer l'automatisation. La première étape consistait à utiliser le moteur de Google pour rechercher le patronyme complet de l'élue Cécile Helle. La fonctionnalité « Custom range » de ce moteur de recherche nous avait permis de préciser explicitement les bornes chronologiques, et donc de filtrer les documents renvoyés, vraisemblablement à partir de leur date de publication. Les 150 résultats enregistrés sur la machine personnelle d'un des auteurs avaient été triées et filtrés manuellement selon leur pertinence (65, soit 43 % concernaient bien l'élue et étaient dans la période) puis

catégorisés selon le type de contenus, les supports observés, les sources recensées, la nature des interactions politiques commentées. La taille réduite de la municipalité et de la communauté politique locale et leur présence réduite en ligne n'ont pas réussi à eux seuls à expliquer la faiblesse des résultats (20 événements politiques détectés au final). Basée sur le seul algorithme Google, la démarche d'extraction était fragile, adaptant les résultats à l'historique de l'utilisateur, laissant de côté l'essentiel de la presse locale et la quasi totalité des contenus Facebook et Twitter.

L'élaboration d'un outil automatisé de détection et d'extraction a ensuite soulevé d'autres limites, les plus importantes étant le bridage du moteur Google lorsqu'on y accède programmatiquement ; le filtrage sur le format éliminant d'office les documents audio-visuels et les textes en PDF ; et la performance insatisfaisante des outils de détection des entités nommées pour le français. Cependant, l'outil n'était pas tellement moins performant qu'un opérateur humain, puisque sur les 70 résultats récupérés, il identifiait des documents décrivant 13 événements politiques sur la période, contre 20 manuellement. En définitive, automatisé ou non, la démarche ne récupérait pas une partie assez significative de l'écho Web-médiatique du travail politique. L'essentiel du problème consistait alors à élargir le nombre et la diversité des résultats.

La reprise de cette tâche de recherche au printemps 2017 vise à corriger certains aspects de cette première expérimentation. Le terrain d'enquête est d'abord élargi de manière comparative à la situation de trois maires urbains en France métropolitaine. Afin de neutraliser certains effets de contexte, nous avons choisi de travailler sur les échos Web-médiatiques de trois femmes socialistes à la tête de trois villes de tailles différentes : Cécile Helle à Avignon, Martine Aubry à Lille et Anne Hidalgo à Paris. L'extraction des corpus des publications relatives à chacune d'entre elles est ensuite limitée à une semaine du lundi 6 au dimanche 12 mars 2017, période choisie pour le caractère a priori dense et routinier de son activité, loin des échéances administratives et budgétaires de fin d'année civile, à bonne distance des vœux du mois de janvier, entre deux périodes de vacances scolaires, sans journée de commémoration spécifique et assez éloignée des rendez-vous électoraux nationaux de mai et juin 2017. Autant d'événements saisonniers à éviter parce que susceptibles d'introduire trop de spécificité dans l'activité politique au sein de la municipalité.

*

L'outil automatique a été amélioré de différentes manières, décrites ci-dessous. En particulier, il fait maintenant appel à quatre moteurs de recherche (et non plus seulement Google), et est capable de traiter plus efficacement les pages Web hébergées par une dizaine de sites de presse nationale et locale. Il a de plus été complété d'une fonctionnalité permettant d'effectuer une recherche sur le média social Facebook. Les outils de détection d'entités nommées en langue française ont été actualisés et sont plus nombreux. En raison de ces améliorations, mais aussi et surtout de l'élargissement de la recherche elle-même, la quantité de données à traiter est bien plus conséquente que lors de notre première expérience. Pour cette raison, nous avons réduit l'approche manuelle à l'évaluation des résultats renvoyés par les moteurs de recherche. L'enquête a enfin permis d'obtenir une information de référence sur l'activité de l'une des trois maires étudiées : Anne Hidalgo à Paris a en effet accepté que ses équipes nous communiquent le contenu de son agenda professionnel pour la semaine du 6 au 12 mars 2017, permettant ainsi de disposer d'une liste fiable des événements (publics ou non, mais non-confidentiel) planifiés sur la période.

5 Description de l'outil de fouille et d'extraction automatique

Le traitement effectué par notre outil se compose de plusieurs étapes, décrites dans la Figure 1, pour ce qui est de la recherche sur le Web (i.e. hors médias sociaux). Tout d'abord, l'utilisateur doit spécifier sa requête, qui prend la forme du nom de la personne ciblée, et de la période temporelle visée. Une recherche Web est alors effectuée afin d'obtenir les adresses des documents supposés pertinents pour cette personne et cette période. Ces adresses sont ensuite filtrées selon le caractère exploitable (ou pas) des formats des documents auxquels elles renvoient. Puis, les adresses Web restantes sont traitées afin de récupérer le code source HTML des pages correspondantes, duquel sont extraits les contenus textuels proprement dits. Une autre étape de filtrage tire parti de ce contenu pour écarter certaines pages. On détecte alors les entités nommées présentes dans ces contenus, et on réalise un troisième filtrage sur cette base. On identifie ensuite les événements présents dans les pages restantes. Un regroupement automatique des pages est alors réalisé, afin de mettre en évidence le fait que le même événement peut apparaître dans plusieurs documents. En ce qui concerne le traitement des médias sociaux, les fonctionnalités qu'ils fournissent permettent d'éviter les étapes de filtrage par format et d'extraction du contenu textuel, mais le reste du processus est le même.

Cette section décrit en détail comment ces différentes opérations sont réalisées et les choix techniques que nous avons été amenés à faire lors de leur implémentation. Le code source de l'outil, en Java, est disponible librement sous GPL sur le service GitHub[7].

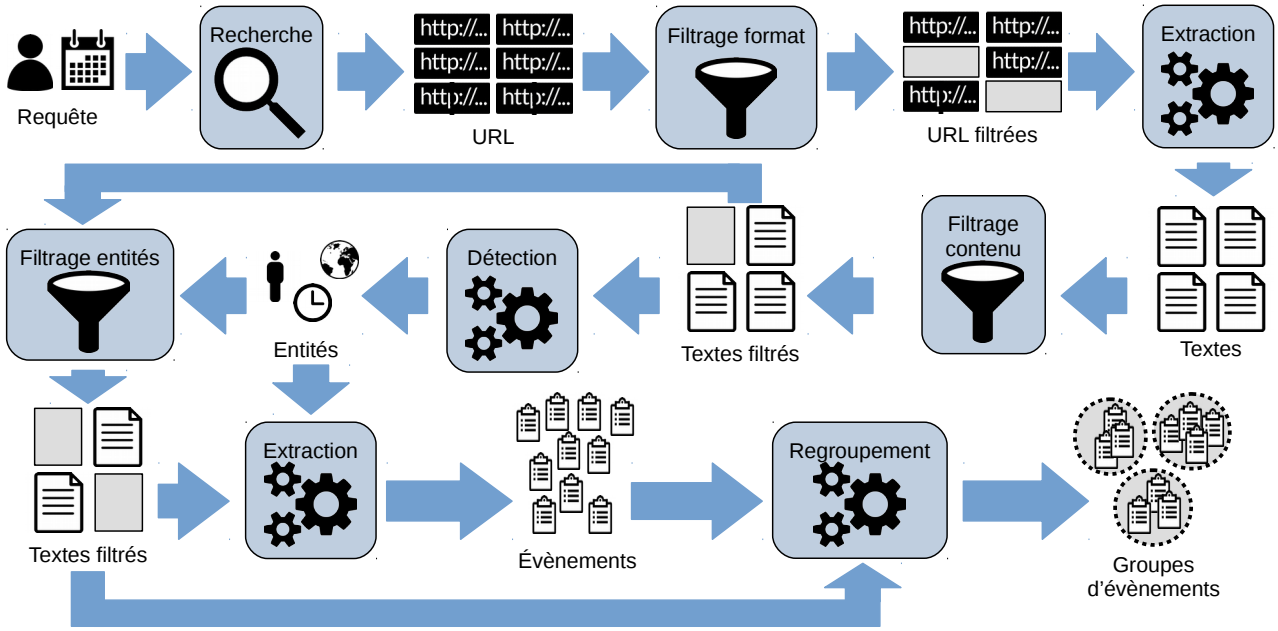

Figure 1. Représentation du traitement effectué sur les pages Web et les réseaux sociaux numériques. Chaque étape est représentée en bleu et est caractérisée (outre le traitement réalisé) par les données qu'elle prend en entrée et celles qu'elle produit en sortie.

5.1 Fouiller le Web

La première étape consiste à invoquer un *moteur de recherche Web*. Il s'agit d'un service en ligne capable de traiter une requête et de renvoyer un certain nombre d'URL (adresses Web) correspondant à des documents jugés pertinents. Les principaux moteurs offrent une API[8], un

7 https://github.com/CompNet/TranspoloSearch
8 Application Programming Interface

dispositif permettant d'effectuer des recherches de façon programmatique (i.e. par le biais d'un logiciel).

L'examen des principaux moteurs accessibles gratuitement, nous a conduit dans un premier temps à en sélectionner 4 pour la recherche : Bing, Google, Qwant et Yandex[9]. Il est à noter que les API de certains d'entre eux n'offrent qu'un accès limité aux fonctionnalités du service, par rapport à une utilisation manuelle, alors que d'autres, au contraire, les enrichissent d'options supplémentaires. Ainsi, les 4 API permettent de concentrer la recherche sur les sites hébergés dans un pays donné, et rédigées dans une langue spécifique. Il est également possible de restreindre la recherche aux pages d'un site particulier. Par contre, Bing et Qwant ne permettent pas de spécifier explicitement la période ciblée lors de la recherche. Pour contourner ce problème, nous avons effectué une recherche distincte pour chaque jour de la période, en incluant littéralement la date de ce jour dans la requête envoyée via l'API. De plus, l'API de Qwant n'est quasiment pas documentée, ce qui rend son utilisation difficile. En mode d'utilisation gratuit, Google limite le nombre de résultats renvoyés à 100, et Bing limite la période d'utilisation à un mois et 1000 recherches.

L'intérêt d'effectuer une recherche basée sur plusieurs moteurs est multiple. Tout d'abord, ceux-ci reposent sur des approches différentes, et sont donc susceptibles d'amener des résultats différents et possiblement complémentaires. Une comparaison nous permettra de déterminer si certains moteurs sont moins informatifs que d'autres, et peuvent donc être écartés, ou si au contraire il y a effectivement complémentarité entre eux. De plus, comme nous l'avons mentionné, ils sont sujets à des limitations techniques différentes, et la combinaison de leurs résultats peut permettre de limiter l'impact de ces limitations.

5.2 Filtrer par format

La deuxième étape consiste à filtrer les URL renvoyées par les moteurs de recherche, de manière à évacuer celles qui ne sont pas exploitables. Dans cette première version, et ce pour des raisons de difficulté technique, nous ignorons les documents qui ne sont pas des pages Web : PDF, fichiers texte, images, vidéos, etc. Il y a donc ici une limitation de l'outil : un opérateur humain serait capable d'exploiter ces données, et il y a donc une perte d'information potentielle. Celle-ci pourrait être résolue de façon relativement simple pour les différents formats encodant du texte (PDF, MS Word, texte brut, etc.)[10]. En revanche, le verrou concernant les données multimédia (images, vidéos, audio) ne semble pas pouvoir être cassé à court terme. Le développement de technologies telles que le Web sémantique, la reconnaissance de visages, le résumé textuel d'images et la transcription automatique de flux audio n'est pas encore assez avancé pour le permettre.

5.3 Récupérer les documents

Les pages restant après le filtrage sont ensuite téléchargées grâce aux adresses Web renvoyées par le moteur de recherche lors de la première étape. À cette occasion, il est possible que certains serveurs Web ne répondent pas, et que les pages correspondantes soient donc considérées

9 Baidu (http://www.baidu.com/) est un moteur chinois écarté pour des raisons linguistiques, parce que la documentation est exclusivement en chinois. Bing (http://www.bing.com/) est le moteur développé par la firme Microsoft. DuckDuckGo (https://duckduckgo.com/) st en réalité un méta-moteur, dans le sens où il intègre des résultats provenant de nombreuses sources, y compris d'autres moteurs de recherche (notamment Bing et Yandex). Il a été cependant écarté parce que les fonctionnalités proposées par une API très récente et encore en développement sont trop limitées pour notre usage. Google (https://www.google.com) est, de très loin, le moteur de recherche le plus utilisé en France (JDN, 2015). Qwant (https://www.qwant.com/) est un moteur de recherche français, soutenu par la Banque européenne d'investissement. Yahoo! (https://yahoo.com) était initialement un moteur propriétaire mais depuis 2011, il repose complètement sur Bing, c'est pourquoi nous l'avons aussi écarté. Yandex (https://www.yandex.com/) est un moteur russe très populaire dans certaines régions linguistiques (notamment russophones et turcophones).

10 De la même façon qu'un utilisateur humain a besoin d'un logiciel approprié pour lire un document MS Word, un logiciel a besoin d'une bibliothèque logicielle appropriée pour accéder au contenu d'un tel fichier.

comme inaccessibles. Cependant, le même problème se poserait pour un opérateur humain. Sur ce point, notre outil a plus de chance d'accéder à l'information, car il est conçu pour solliciter de façon répétée un serveur qui ne répondrait pas : si l'indisponibilité est temporaire, la page sera récupérée après plusieurs tentatives. Un opérateur humain n'aurait pas la même persistance.

Les pages Web téléchargées sont enregistrées localement afin de pouvoir être traitées rapidement, et de permettre un accès rapide dans le futur. Cependant, les pages récupérées ne sont pas constituées uniquement de contenu pertinent : on y trouve également des liens vers d'autres sites, des menus, des mentions légales, de la décoration, etc. Il est donc nécessaire d'identifier quelle est la partie de la page sur laquelle l'extraction sémantique à suivre devra se concentrer. Nous appellerons cette partie le *contenu* dans le reste de notre description.

Nous avons développé un extracteur automatique permettant de traiter ce problème de façon générique, dans le sens où il peut traiter une grande diversité de structures de pages Web. Cependant, il est susceptible de commettre des erreurs sur certaines pages, voire de ne pas récupérer toutes les données qu'elles contiennent. C'est notamment le cas pour des sites codés *manuellement*, i.e. qui ne sont ni maintenus par des professionnels (journaux, entreprises, institutions, etc.), ni hébergés sur des plate-formes spécialisées (WordPress, Blogger, etc.). Ces sites ne respectent pas toujours les bonnes pratiques et les standards du Web, et peuvent tromper notre logiciel là où un humain serait plus robuste. Cependant, il faut noter qu'ils constituent une minorité des sources renvoyées par les moteurs de recherche.

Pour obtenir des résultats plus fiables, nous avons identifié lors de nos expérimentations un certain nombre de sites particulièrement pertinents, pour lesquels nous avons développé des extracteurs spécifiques. Ceci implique d'identifier manuellement la position du contenu pertinent dans le code source HTML des pages Web du site ciblé, et d'adapter ensuite l'algorithme de manière à cibler cette position. L'opération ne doit être effectuée qu'une seule fois par site, et permet par la suite de traiter efficacement toute page issue de ce site Web-là. Cependant, le développement de ce type d'extracteur est coûteux, car il implique un long travail de programmation manuel. Pour cette raison, nous nous sommes concentrés sur des sites de presse qui revenaient souvent dans les résultats obtenus pour les personnes ciblées dans ce travail, à savoir : La Provence, La Voix du Nord, Le Figaro, Le Monde, Le Parisien, Le Point, L'Express et Libération.

Cette étape inclut également un nettoyage du texte récupéré, afin d'en supprimer les éventuels caractères exotiques (notamment les caractères non-latins) susceptibles de causer des problèmes lors des étapes suivantes, notamment la reconnaissance d'entités nommées.

5.4 Fouiller les médias sociaux

Les moteurs de recherche classiques ne couvrent pas l'intégralité du Web. En particulier, tous les principaux médias sociaux leur interdisent désormais (il faut noter que cela n'a pas toujours été le cas) d'indexer les contenus qu'ils hébergent (ou alors seulement très partiellement). Ceci pose un problème important, car les médias sociaux constituent aujourd'hui un canal de communication très populaire auprès d'une grande partie de la population, et sont devenus incontournables dans la vie politique.

Pour effectuer une recherche sur un média social, il est désormais nécessaire de passer par l'API fournie par la plate-forme concernée, si celle-ci existe. Il faut souligner que ces API sont elles-mêmes extrêmement limitées. Celle de Facebook, par exemple, ne permet plus de réaliser des recherches par mot-clé sur l'intégralité de la plate-forme (c'était auparavant le cas), du type de celles que l'on effectue avec Google sur le Web. Cette API n'autorise que la récupération de données publiées sur des pages que l'utilisateur doit identifier *a priori*. On est donc loin d'une approche exploratoire qui consisterait à retrouver, par exemple, tous les posts écrits par n'importe quelle personne sur un sujet donné : au contraire, on doit d'abord identifier l'auteur avant de préciser la nature de la recherche (ex. période ciblée, mot-clé, etc.). Sur d'autres plate-formes telles

que Twitter, la restriction ne porte pas sur ce type de recherche, qui est autorisée, mais sur l'exhaustivité des résultats : ceux-ci ne sont pas complets, et la sélection des tweets renvoyés se fait de façon opaque, dans un but affiché de personnalisation, selon un certains nombre de critères relatifs à la localisation de l'utilisateur effectuant la requête, sa langue, etc[11].

Malgré ces limitations, nous avons décidé d'inclure dans notre outil une étape de recherche sur Facebook. Dans un premier temps, nous avons ciblé la page de la personnalité visée par la requête originale, en se concentrant sur la période spécifiée par l'utilisateur. Nous récupérions ainsi tous les posts publiés par la personne sur la période donnée. Nous élargissions ensuite le traitement, en sollicitant les commentaires laissés par d'autres utilisateurs en réponse à ces posts. Enfin, nous identifions ces commentateurs, et récupérions tous les posts qu'ils avaient eux-mêmes écrits sur la période ciblée. Mais devant la rareté des résultats obtenus pour certaines des personnes ciblées, nous avons décidé dans un second temps d'élargir encore notre approche. Pour ce faire, nous avons identifié manuellement les principaux soutiens et opposants de la personne ciblée, et leur avons appliqué exactement le même traitement. Les posts supplémentaires obtenus ont été rajoutés à ceux déjà récupérés pour la personne ciblée.

Grâce à l'API de Facebook (mais cela resterait vrai pour d'autres médias sociaux), les textes obtenus sont directement utilisables, et ne nécessitent pas tout le traitement précédemment décrit pour la recherche Web afin d'extraire le contenu pertinent des pages Web (filtrage sur le format, et extraction du contenu textuel).

5.5 Filtrer par contenu

À ce stade du traitement, nous disposons du contenu textuel associé à la recherche sur le Web et à celle sur les médias sociaux. Nous tirons parti de ce contenu pour réaliser un filtrage dans le but d'écarter les textes non-pertinents, en nous basant successivement sur plusieurs aspects.

Le premier est la langue. En effet, si tous les moteurs de recherche Web utilisés peuvent être paramétrés pour cibler une langue donnée, en pratique certains d'entre eux renvoient quand même de nombreuses pages rédigées dans d'autres langues. En ce qui concerne les médias sociaux, l'API ne donne aucun contrôle sur la langue des posts et des commentaires obtenus. Notre outil détecte automatiquement la langue des textes traités, et ne garde que ceux rédigés en français. Le reste du traitement n'est effectué que sur ceux-ci.

Le deuxième aspect est l'apparition explicite du nom de la personne ciblée dans le texte. On cherche seulement le patronyme, car la présence d'une expression comportant exactement le prénom et le nom semble trop restrictif : un article peut très bien porter sur la personne ciblée tout en ne mentionnant que son nom de famille, et sans jamais mentionner son prénom. Bien sûr, cette approche est plus sensible à la présence d'homonymes, mais elle permet de ne pas écarter des résultats potentiellement intéressants. Ce filtre-là n'est pas appliqué aux posts que la personne ciblée a directement publiés sur les médias sociaux, mais il l'est à ceux produits par les personnes considérées dans notre recherche étendue (soutiens et opposants).

5.6 Détecter les entités nommées

L'étape suivante consiste à détecter dans chaque article les mentions d'*entités nommées*. Au sens strict, une entité nommée est un concept du monde réel caractérisée par un nom propre : personne, entreprise, lieu, etc. Le but de l'opération est d'identifier les références à ces entités dans les articles traités. Le traitement repose sur des outils issus du TALN (Traitement Automatique du Langage Naturel). La version actuelle du logiciel permet de détecter les personnes, lieux, organisations (entreprises, institutions, associations...), fonctions (maire, directeur…), productions

11 À noter que certains services payants permettaient jusqu'à il y a peu d'accéder à des résultats exhaustifs, mais qu'ils ne sont plus proposés.

(œuvres intellectuelles), réunions (congrès, meetings…) ainsi que les dates (qui ne sont pas des entités nommées au sens strict).

La principale limitation ici vient du peu d'outils existant pour le traitement du français (là où les anglophones sont bien pourvus). En effet, idéalement, la détection d'entités nommées devrait être complétée de plusieurs traitements pour extraire toute l'information disponible. Tout d'abord, il serait nécessaire de détecter les *situations anaphoriques*, i.e. les cas où une expression textuelle qui n'est pas un nom propre fait référence à une mention d'entité apparue auparavant dans le texte. La tâche de *résolution de coréférences* permet de déterminer à quelle mention cette expression fait référence[12]. Enfin, la désambiguïsation de mentions consiste à déterminer à quelle entité exactement un groupe de mentions fait référence[13]. La résolution de ces limitations est envisageable à court ou moyen terme, car plusieurs méthodes existent déjà pour l'anglais, et les travaux pour les adapter au français (ou développer des solutions spécifiques à cette langue) sont en cours. En l'état, notre outil fonctionne selon une approche à base d'ensemble, c'est à dire qu'il applique plusieurs détecteurs possédant des caractéristiques différentes, et qu'il combine leurs résultats[14].

5.7 Filtrer par entité

Une fois les entités identifiées, nous pouvons les utiliser pour réaliser un dernier filtrage. Il est en particulier possible de vérifier que les dates mentionnées dans le contenu sont bien comprises dans l'intervalle spécifié pour la recherche. Quand ce n'est pas le cas, nous pouvons considérer le contenu comme non-pertinent et l'écarter. On pourrait généraliser l'approche aux autres types d'entités nommées, si l'on désire par exemple concentrer la recherche sur les relations entre la personne ciblée et certaines entités.

5.8 Extraire les événements

Nous faisons l'hypothèse simplificatrice qu'un article décrit un seul événement. La sémantique associée à cet événement est décrite par l'ensemble des entités détectées dans le contenu textuel. Les différents types d'entités correspondent aux dimensions utilisées pour caractériser l'événement : *où* (lieux où l'événement s'est déroulé), *quand* (dates de l'événement) et *qui* (personnes et organisation ayant participé à l'événement). Les autres types d'entités nommées sont optionnels et servent seulement à préciser l'événement. Bien entendu, un article peut en réalité décrire plusieurs événements. Notre hypothèse simplificatrice est dictée par la limitation relative à la détection des entités nommées, que nous avons mentionnée précédemment. Le fait d'être pour l'instant dans l'incapacité de traiter les anaphores nous empêche de descendre à un niveau de granularité plus fin tel que le paragraphe ou la phrase.

12 Par exemple, considérons le texte suivant : "François Hollande a déclaré à la presse ne pas en savoir plus sur cette affaire. Le président s'est ensuite envolé vers Paris pour participer au conseil des ministres." On voudrait déterminer automatiquement que non seulement "Le président" est une anaphore grammaticale, mais qu'en plus elle fait référence à "François Hollande".

13 Par exemple, on veut déterminer si l'expression "Michael Jordan" fait référence au fameux basketteur (https://fr.wikipedia.org/wiki/Michael_Jordan) ou au pionnier de l'apprentissage automatique (https://fr.wikipedia.org/wiki/Michael_Jordan_(informaticien)).

14 Cette approche utilisée précédemment (Atdağ et Labatut, 2013 ; Labatut, 2013) repose sur les détecteurs OpenCalais (http://www.opencalais.com/), un service Web proposé par Thomson Reuters et existant en version gratuite (mais limité en termes de nombre de requêtes par jour) ; OpeNER (http://www.opener-project.eu/), un autre service Web, développé dans le cadre d'un projet européen, HeidelTime (https://github.com/HeidelTime/heideltime), qui est spécialisé dans la détection de dates et autres entités temporelles ; Nero (https://nero.irisa.fr/) un outil général de détection d'entités nommées ; et TagEn (http://gurau-audibert.hd.free.fr/josdblog/wp-content/uploads/2008/03/TagEN.tar.gz), un autre outil général.

5.9 Regrouper les documents

Il est possible qu'un même événement soit décrit par plusieurs sources distinctes. En pratique, sur les cas traités ici, c'est même très fréquent. Pour affiner les résultats produits par notre outil, nous exécutons une dernière étape dans notre traitement, qui consiste à regrouper par similarité les contenus restant. Nous adoptons pour cela une approche classique en recherche d'information, basée sur le score *tf-idf* et la distance cosinus (Manning, Raghavan et Schütze, 2008). En résumé, chaque contenu est caractérisé par la fréquence d'apparition des mots qui le composent, et les contenus présentant des distributions lexicales proches sont considérés comme similaires. Un algorithme standard de classification hiérarchique ascendant est ensuite appliqué à ces mesures de similarité. La meilleure partition de l'ensemble des contenus est finalement obtenue en sélectionnant la coupe de Silhouette (Rousseeuw, 1987) optimale dans le dendrogramme produit. Les documents groupés ensemble sont supposés décrire le même événement.

*

Notre travail d'analyse de l'écho Web-médiatique a été effectuée en deux phases. Dans la première, nous avons considéré l'activité de Cécile Helle à Avignon, Martine Aubry à Lille et Anne Hidalgo à Paris, afin de réaliser une étude comparative. Cette phase exploratoire éclaire l'instabilité du comportement des moteurs. Dans la seconde phase, nous avons tiré parti des résultats obtenus à la première phase pour affiner la recherche en se concentrant sur A. Hidalgo. Sur la base de ces deux expérimentations, nous discutons ensuite des limitations de notre outil. Enfin, nous proposons une première interprétation des résultats renvoyés par notre outil lors de la seconde phase.

6 Première phase exploratoire : instabilité des données comparatives

Lors de cette première phase, nous avons effectué une première recherche le 20 juillet 2017 pour chacune des trois élues, en appliquant exactement la même procédure : utilisation du nom complet (prénom et nom de famille), et restriction à la période ciblée (du lundi 6 au dimanche 12 mars 2017). Cette recherche a été réitérée à l'identique après l'été, le 12 septembre 2017, dans le but d'évaluer la stabilité de nos résultats. Les deux recherches ont donc été menées plus de 4 mois après la période ciblée. Comme expliqué plus loin, des recherches supplémentaires ont ensuite été réalisées sous différentes modalités, afin de compléter nos résultats et d'étayer leur interprétation. Les moteurs de recherche Web fournissent l'essentiel de l'information traitée par notre outil, il est donc important d'étudier leur comportement. Pour cette raison, nous nous concentrons dans un premier temps sur les pages Web obtenues en combinant les 4 moteurs précédemment mentionnés.

6.1 Une instabilité temporelle

La comparaison des listes des URL *filtrées* (texte indisponible, langue étrangère, absence du patronyme, format PDF) obtenues lors des recherches de juillet et de septembre montre une grande instabilité dans le temps. La Table 1 contient les nombres de pages filtrées renvoyées uniquement lors de la première itération, uniquement lors de la deuxième, et enfin lors des deux itérations. La deuxième itération donne d'abord plus de résultats que la première (+34 % en moyenne). Il y a peu de chances pour qu'à plus de quatre mois de distance cette inflation entre deux sondages soit due à l'apparition de nouvelles informations ou de nouveaux commentaires, mais d'avantage à l'instabilité générale du Web et au éventuelles mutation des algorithmes de recherche. Cette dimension temporelle mériterait d'être approfondie : il serait par exemple intéressant d'étudier comment la restitution des moteurs de recherche vis-à-vis d'une période temporelle fixée change au cours du temps. Cela permettrait par exemple d'étudier l'évolution de l'écho Web-médiatique d'une

semaine de travail politique régulièrement à plusieurs moments, suivant les événements, afin de mesurer d'éventuelles effets de buzz, d'oubli ou de surgissement.

Les variations sont également importantes d'une élue à l'autre, entre C. Helle à Avignon pour laquelle on ne retrouve en septembre que 27 % des URL identifiées en juillet, et A. Hidalgo à Paris, pour laquelle cette part s'élève à plus de 73 %. Plus les moteurs identifient de contenus relatifs à l'élue, moins cette instabilité temporelle semble grande.

Élue	20/07/17 seule	20/07 et 12/09	12/09/17 seule	Total	Taux de stabilité (%)
Helle	8	3	22	33	27,3
Aubry	47	78	88	213	62,4
Hidalgo	43	119	91	253	73,5

Table 1. Comparaison des nombres de pages Web obtenues à l'issue de l'interrogation des moteurs et après filtrage, pour les deux itérations de la recherche (juillet et septembre 2017).

6.2 La variation des mobiles du filtrage automatique des URL

Nous nous penchons maintenant sur les données issues de la seconde itération. La Table 2 contient les nombres de pages Web avant et après filtrage, ainsi que des informations démographiques sur les villes associées aux 3 élues ciblées. Les moteurs de recherche Web proposent une masse presque équivalente de résultats bruts non filtrés pour ces trois élues, entre 465 pour C. Helle à Avignon et 613 pour A. Hidalgo à Paris. Mais le filtrage donne au final des résultats très variables : entre 34 % d'URL conservées pour l'exploitation sur le cas d'A. Hidalgo et seulement 5,4 % pour C. Helle.

Élu(e)s	URL non filtrées	URL filtrées exploitables	%	Ville	Population municipale	Population aire urbaine
Helle	465	25	5,4	Avignon	92 209	518 981
Aubry	580	166	28,6	Lille	233 897	1 182 127
Hidalgo	613	208	33,9	Paris	2 220 445	12 475 808
Ries	583	19	3,3	Strasbourg	276 170	773 447

Table 2. Comparaison des nombres de pages Web avant et après filtrage, pour les trois élues ciblées et Roland Ries, pour la seconde itération de la recherche (septembre 2017).

La Figure 2 décompose le filtrage, en distinguant les différentes causes amenant l'outil à écarter une page. On voit ainsi qu'une proportion non-négligeable des URL est inaccessible (le serveur Web ne répond pas à nos requêtes), écrite dans une langue autre que le français, ou correspond à un document qui n'est pas exploitable en raison de son format. Mais dans tous les cas considérés, la cause principale de filtrage est l'absence du patronyme de la personne ciblée dans le texte extrait de la page Web.

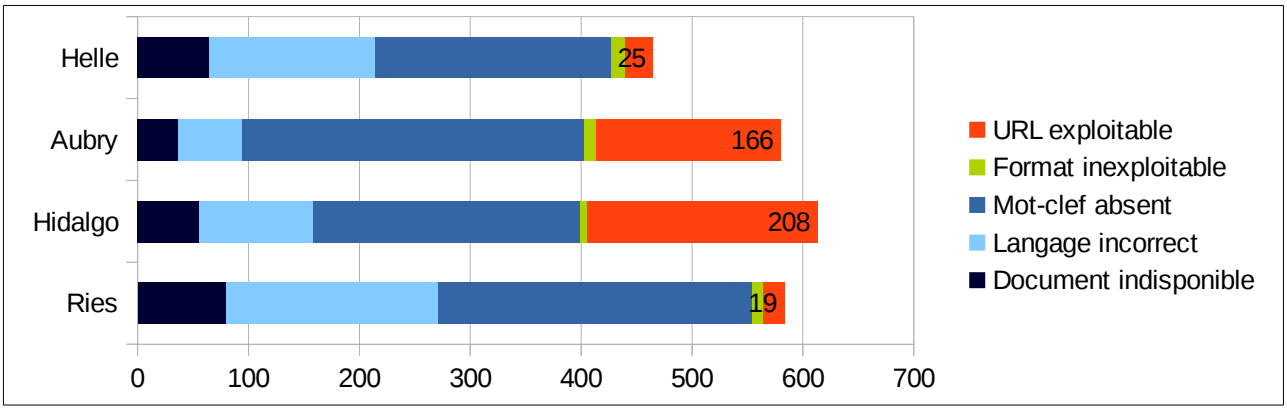

Figure 2. Distribution du filtrage des pages Web obtenues lors de la deuxième itération de la recherche (Septembre 2017), pour les 3 élues ciblées + Roland Rie entre le 6 et le 10 mars 2017.

La taille de la ville et les volumes d'activités publiques et de publicité associées semblent un premier facteur déterminant de l'ampleur de l'écho Web-médiatique donné au travail édilitaire, ainsi que des capacités de notre outil à les détecter. Mais de multiples facteurs peuvent affecter la recherche automatique. Pour explorer cette hypothèse, nous avons réalisé lors de la deuxième itération une recherche supplémentaire portant sur Roland Ries, maire socialiste de Strasbourg, une ville de taille équivalente à Lille (Table 2 & Figure 2). Alors que les moteurs ramènent autant de résultats bruts que pour Aubry (583 contre 580), le taux d'URL conservées après filtrage plafonne à Strasbourg à 3,3 %, sous le niveau d'Avignon. La zone frontalière avec l'Allemagne et la nature du patronyme recherché introduit notamment et vraisemblablement un taux d'erreur linguistique important.

6.3 Les compléments de fouille dans la presse locale et les médias sociaux

La première enquête menée en 2015 avait montré que les moteurs de recherche utilisés ne proposent qu'une infime partie des publications de la presse locale, pourtant indexée. Nous avons donc effectué une autre série de recherches pour chacune des 3 élues ciblées, en ciblant uniquement les sites Web des trois principaux journaux locaux présents sur ces trois territoires : Le Parisien, La Voix du Nord, et La Provence. Nous avons obtenu par ce biais 7 nouveaux articles pour C. Helle à Avignon, 14 pour M. Aubry à Lille et 47 pour A. Hidalgo à Paris. Ce complément documentaire confirme les effets de distorsion liés à la faiblesse de l'espace Web-médiatique avignonnais et à l'inverse à la surexposition médiatique parisienne.

Les résultats obtenus via l'API de Facebook montrent une distorsion inversement proportionnelle à la taille de la ville : c'est la maire d'Avignon qui publie le plus avec 9 posts, contre 6 pour la maire de Paris et aucun pour celle de Lille. Il faut cependant noter que dans ce dernier cas, la maire de Lille et son équipe ont été complètement inactives sur ce média pour une longue période incluant la semaine que nous ciblons.

Au final, le corpus comparatif apparaît comme incohérent et il est difficile d'imputer ses variations aux propriétés individuelles, pratiques ou contextuelles bien différentes des trois maires observées sur la même période, ou bien alors aux difficultés des algorithmes de recherche à générer des résultats comparables. Les fragilités observées lors de cette première phase exploratoire nous ont conduit à approfondir le cas d'Anne Hidalgo, pour laquelle le Web renvoie un nombre plus conséquent de résultats.

7 Deuxième phase ciblée sur Anne Hidalgo : évaluation manuelle

Anne Hidalgo est la seule des trois maires sollicitées qui a accepté de nous livrer le contenu de son agenda professionnel pour la semaine étudiée, du 6 au 12 mars 2017. Cette demande consistait à récupérer l'intégralité des événements planifiés dans l'agenda électronique de la maire, avec l'ensemble des informations associées (jour, heures de début et fin, lieu, objet, interlocuteurs et commentaires…), afin de disposer d'une base de référence reflétant une grande partie de l'activité de l'élue et à laquelle comparer les événements de l'activité plus ou moins publics, publicisés, commentés et récupérés par notre outil.

7.1 Un corpus Hidalgo limité au Web

Il s'agit ici en principe de circonscrire le corpus de l'ensemble des pages Web filtrées et des posts Facebook compilés pour la période du lundi 6 au dimanche 12 mars 2017, dans le but de tester les capacités de détection des entités nommées de l'outil à partir d'une analyse manuelle de référence. Nous considérons donc d'abord les résultats obtenus sur la semaine du lundi 6 au dimanche 12 mars, combinant la recherche neutre (i.e. sur le Web entier) et la recherche ciblant spécifiquement *Le Parisien*. Mais en ce qui concerne les médias sociaux, nous complétons les posts et commentaires de la page publique Facebook d'A. Hidalgo par ceux récupérés pour six membres du conseil municipal, répartis sur l'échiquier politique, disposant d'un minimum de notoriété, et utilisant ce média : Jean-Louis Missika (PS), Ian Brossat (PCF), Christophe Najdovski (EELV), Nathalie Kosciusko-Morizet (Prog), Claude Goasguen (LR), Brigitte Kuster (LR). Le recensement est cependant décevant. Sur la période, seules 3 des 7 personnes testées postent et ils postent peu : la maire 5 fois, Goasguen 5 fois et Nadjovski 1 seule fois. Les 11 documents compilés ne donnent pas une vision significative de l'écho de l'activité politique du maire sur les médias sociaux et cela nous conduit à renoncer en l'état à leur exploitation.

Pour ce qui est du Web, nous récupérons 817 URL, dont 284 se révèlent être, après filtrage, des pages exploitables[15]. La Table 3 détaille la distribution des URL en fonction de leur filtrage.

Anne Hidalgo	Web	leparisien.fr	Total
Page indisponible	63	45	108
Langue étrangère	117		117
Mot-clef manquant	256	45	301
Format inexploitable	7		7
URL exploitables	**237**	**47**	**284**
Total	680	137	817

Table 3. Distribution, en termes de filtrage, des URL obtenues pour A. Hidalgo en combinant les deux recherches réalisées sur elle lors de la deuxième itération (recherche générale et recherche ciblant *Le Parisien*).

Le regroupement automatique (clustering) effectué sur ces pages nous permet d'identifier 129 thèmes. La plupart d'entre-eux ne concerne qu'une seule (67%) ou 2 pages (22 %), seuls 6 sont repris par plus de 10 pages, le plus fréquent apparaissant dans 24 pages.

15 La recherche sur l'ensemble du Web donne dans un premier temps 680 URL. Seules 237 sont filtrées comme exploitables. La recherche ciblant spécifiquement le site du journal *Le Parisien* compilent 137 résultats, dont seulement 47 subsistent après filtrage.

7.2 Des moteurs aux résultats hétérogènes

Une première remarque concerne la grande hétérogénéité des résultats des 4 moteurs utilisés, illustrée par la Table 4. D'un côté, Yandex et Google renvoient respectivement moins de 9% et 20% des pages subsistant après le filtrage, alors que de l'autre, Bing et Qwant atteignent respectivement plus de 96% et 66%. Il est à noter que, de façon générale, ces deux derniers renvoient bien plus de résultats, en raison du fait que leur API ne permet pas de spécifier explicitement la période visée par la recherche. En effet, comme expliqué précédemment, pour contourner cette limitation, nous incluons explicitement une date dans les requêtes soumises à ces moteurs, et nous traitons donc autant de requêtes qu'il y a de jours dans la période ciblée.

Total URL exploitables	**284**	**%**
Bing	249	96,4
Google	49	19,0
Qwant	172	66,6
Yandex	23	8,9

Table 4. Proportion d'URL exploitables renvoyées par les 4 moteurs de recherche Web utilisés dans l'outil. Une même URL peut être renvoyée par plusieurs moteurs différents.

La Table 5 montre les classements obtenus par les 25 premiers résultats de Google auprès des 4 moteurs considérés. On peut observer la faible correspondance entre les classements, ce qui illustre encore la variété de fonctionnement de ces algorithmes, ainsi que les biais possibles introduits par la nécessité de traiter chaque journée séparément pour Bing et Qwant.

Spécification des bornes de la période 06/03/2017 – 12/03/2017		Compilation des 7 recherches par jour			
Google	**Yandex**	**Bing**		**Qwant**	
Rang	**Rang**	**Jour**	**Rang**	**Jour**	**Rang**
1	9	7	1	7	1
2		4	1	4	1
3		7	4	7	5
6	7				
7					
8		3	34	3	16
9					
10		3	2	3	2
11					
12					
14					
15		7	13		
16	6				
17		7	38		

18					
20		7	20		
21	24				
23		2	30	2	30
24		2	10	2	10
25		2	3	2	5

Table 5. Comparaison des rangs des URL renvoyées par les 4 moteurs de recherche Web considérés. Les deux premiers autorisent de spécifier explicitement la période temporelle ciblée. Ce n'est pas le cas des deux derniers, pour lesquels cette limitation a été contournée en réalisant plusieurs requêtes incluant chacune explicitement la date d'un jour de la période, aboutissant à autant de classements distincts. Si la même URL apparaît en réponse à plusieurs de ces requêtes, nous conservons ici son meilleur classement.

7.3 L'évaluation manuelle des résultats Hidalgo

Devant cette hétérogénéité, nous avons décidé de réaliser une évaluation manuelle du filtrage effectué par l'outil. Pour cela, chaque page subsistant à l'issue du filtrage a été ouverte, examinée visuellement et catégorisée. Ce processus comporte 4 étapes, dont les effets sont décrits par la Table 6. Il commence par l'élimination d'une minorité d'URL indisponibles ou dont l'accès est payant (1,8%), puis de celles renvoyant à des pages qui traitent d'homonymes ou simplement hors-sujet (4,6%). Il s'agit ensuite d'éliminer les quelques doublons (3,2%), i.e. les pages associées à des URL différentes mais ayant le même contenu textuel. Celles-ci proviennent essentiellement de la recherche ciblant *Le Parisien*[16]. Leur faible nombre montre d'ailleurs la nécessité d'un traitement spécifique des sites de presse locaux. Un filtrage plus important est ensuite opéré sur les dates de publication et la période du déroulement de l'événement dont il est question dans le texte ouvert par l'URL. L'analyse visuelle élimine ici une part importante des pages pourtant sélectionnées par notre outil, puisque 85 liens (soit 30 %) concernant bien l'activité d'Anne Hidalgo sont néanmoins considérés comme « hors-période » et donc inexploitables. Le corpus final pertinent se réduit finalement à 172 URL renvoyant à des pages Web évoquant bel et bien l'activité de l'élue entre le 6 et le 12 mars 2017.

Anne Hidalgo	**Filtres manuels**	**Corpus filtré = 284**
Indisponible	5	112
Homonyme / erreur	13	
Doublons	9	
Hors périodes	85	
Corpus final		**172**

Table 6. Résultat du traitement manuel des pages issues du filtrage automatique.

Comme indiqué par la Table 7, sur ce corpus final filtré manuellement, les performances des moteurs restent dans les mêmes proportions que celles observées précédemment : augmentation légère pour Google et Yandex, diminution de quelques points pour Bing et Qwant.

16 Parmi les 47 références valides récupérées en traitant spécifiquement Le Parisien, on ne repère que 7 doublons avec celles obtenues lors de la recherche sur tout le Web. On observe également une situation de « triplon » : le site contient des pages listant ses propres papiers, et ces pages-ci sont elles-mêmes indexées par les moteurs de recherche.

Total URL filtrée manuellement	172	%
Bing	147	94,0
Google	43	27,5
Qwant	107	68,4
Yandex	17	10,9

Table 7. Proportion d'URL exploitables renvoyées par les 4 moteurs de recherche Web après filtrage automatique et manuel (à comparer à la Table 4, qui ne comporte pas le filtrage manuel).

8 Les limites de l'outil automatique

L'évaluation expérimentale de notre outil a mis en évidence un certain nombre de limitations, que ce soit dans le traitement de la recherche sur le Web ou de celle dans les médias sociaux. En ce qui concerne les premières, il y a deux problèmes principaux : la pertinence des pages obtenues à l'étape initiale, et l'extraction du texte qu'elles contiennent.

8.1 La fragile pertinence des pages identifiées

Les moteurs de recherche qui constituent l'étape initiale de notre processus renvoient de nombreux résultats non-pertinents relativement à nos besoins informationnels. Nous avons observé essentiellement trois types d'erreurs : hors-sujet complet, hors-sujet relatif et hors-période.

Nous considérons comme *hors-suje*t un article qui ne traite ni de la personne visée par la recherche, ni même de thèmes qui lui sont connectés. En l'absence d'informations détaillées sur la façon dont les index des moteurs utilisés sont construits, il est difficile d'expliquer la présence de ces pages dans les résultats qu'ils renvoient. La plupart de ces articles ne contiennent pas explicitement le patronyme de la personne ciblée par la recherche, et sont donc écartés lors de notre étape de filtrage à base de contenu. Les homonymes posent cependant problème, car ils parviennent à passer nos filtres. Une solution serait d'avoir recours à des méthodes de désambiguïsation d'entités nommées. La marge d'amélioration paraît faible, mais ce genre d'erreur reste de toute façon très minoritaire. Tous les moteurs étant configurés pour ne renvoyer que des documents écrits en français et publiés sur des sites français, nous considérons également comme hors-sujet les pages rédigées dans une autre langue (de nombreuses pages en allemand ont ainsi été renvoyées pour R. Ries, en raison de l'existence de nombreux homonymes allemands ou autrichiens). Notre étape de filtrage à base de contenu est capable de les détecter efficacement.

Le *hors-sujet relatif* touche les pages qui traitent d'un sujet proche de la personne visée et de son activité, mais sans jamais la mentionner directement. Ce type de résultat correspond généralement à des pages issues de sites de presse et prenant la forme soit un article contenant des liens vers d'autres articles relatifs, soit d'une liste thématique de résumés d'articles. Dans les deux cas, le nom de la personne recherchée est susceptible d'apparaître, provoquant l'indexation de la page par le moteur de recherche, sans que son contenu traite pour autant de cette personne. Le premier cas pourrait être traité en améliorant la phase d'extraction de texte, sur laquelle nous revenons plus loin puisqu'elle constitue le second problème majeur de notre outil en ce qui concerne la recherche Web. Quant au second cas, il serait possible de détecter automatiquement ces listes et de les écarter automatiquement en en exploitant les particularités du code HTML.

Les pages *hors-période* traitent bien de la personne visée par la recherche, mais pas pour la période visée. Le problème peut s'expliquer en partie par le fait que certains moteurs de recherche ne permettent pas de spécifier de contraintes sur la période visée. Cependant, les moteurs qui possèdent cette fonctionnalité commettent eux-aussi ce type d'erreur. Il est difficile de filtrer automatiquement les articles selon un critère temporel, car ils ne contiennent souvent pas explicitement de date. L'information la plus intéressante semble être la date de publication, qui fait

partie des méta-données souvent associées à l'article. Cependant, cette date n'est pas toujours disponible, et son extraction requiert la plupart du temps un traitement spécifique, que nous n'avons pas pu implémenter pour tous les cas de figure possibles.

8.2 La difficile extraction des textes

Le second problème majeur de la recherche Web est l'extraction du contenu textuel pertinent à partir du code HTML de la page. Cette tâche, qui paraît triviale à un être humain, est difficile à automatiser en raison de la complexité et de la diversité que permet le code HTML. Celui-ci contient bien plus d'information que celle qui apparaît au lecteur humain, ce qui vient masquer le contenu pertinent à l'outil automatique chargé de son extraction. On se retrouve alors face à deux types d'erreurs, qui ne sont pas mutuellement exclusives : récupérer seulement une partie du contenu textuel pertinent, et récupérer du contenu non-pertinent.

Une *extraction incomplète* a pour conséquence que les textes ne décrivent que très partiellement les événements, ce qui dégrade la suite du traitement. Cette situation survient rarement quand l'accès au contenu de l'article est sujet à un abonnement payant. Le principal problème vient de la tâche d'extraction du texte. Nous avons développé un certain nombre d'extracteurs spécifiquement adaptées à des sites bien particuliers, et traitons le reste des pages rencontrées avec une version générique, qui utilise un ensemble de règles pour tenter d'identifier efficacement quelle section du code HTML contient le texte pertinent. Le problème d'incomplétude se produit justement avec cet extracteur, quand celui-ci ne parvient pas à traiter correctement une page qui ne se conforme pas aux règles prédéfinies.

On pourrait croire *a priori* que récupérer trop de contenu (i.e. plus que le seul texte pertinent) ne constitue pas une difficulté réelle, mais en réalité il s'agit d'un problème plus important que le précédent. En effet, le texte non-pertinent récupéré peut rendre complètement inexploitable une page initialement pertinente, en dominant par son volume le texte pertinent. La plupart des pages contiennent des menus, des publicités, des liens vers des partenaires, des articles ou d'autres parties du site, qui mentionnent souvent des entités nommées. L'inclusion de ce texte lors de la phase d'extraction entraîne ainsi plus tard la détection d'un grand nombre d'entités qui ne sont généralement aucunement liée au thème de l'article proprement dit, et brouillent donc l'identification des événements qu'il décrit.

8.3 L’inaccessibilité des médias sociaux

L'extraction des données spécifiquement issues des médias sociaux souffre de limitations engendrées par leurs API. Tout d'abord, celles-ci sont beaucoup moins stables que celles proposées par les moteurs de recherche, et sont régulièrement modifiées, ce qui peut rendre obsolète les outils les utilisant. Deuxièmement, elles sont prévues pour être utilisées dans le cadre d'applications destinées aux terminaux mobiles et présentent donc un nombre limité et bien spécifiques de fonctionnalités avec lesquelles il faut composer pour parvenir à ses fin. Enfin, et c'est le point le plus important, ces API offrent un accès extrêmement restreint aux utilisateurs désireux d'accéder programmatiquement aux données hébergées par les médias sociaux.

À l’heure actuelle, sur Facebook, comme sur Twitter, on ne peut ainsi plus effectuer de recherche globale à partir d’un mot-clé (i.e. chercher ce mot dans tout le contenu public hébergé par le service). Dans le cadre de ce travail, le meilleur substitut que nous avons trouvé à été de d'effectuer une recherche ciblant sur Facebook uniquement les posts publiés par une personne donnée sur une période donnée. On n'a donc pas la possibilité de faire une recherche exploratoire comme c'est le cas sur le Web : il faut avoir identifié au préalable des utilisateurs ou pages Facebook servant de point de départ. Dans le cas d'A. Hidalgo, nous avons ainsi ciblé sa page ainsi que celles de ses principaux soutiens et opposants. Il est clair que cette contrainte introduit des biais très fort dans la recherche.

Un autre souci technique avec Facebook est la distinction qui est faite dans l'accès programmatique aux posts, en fonction du fait que ceux-ci sont publiés sur un compte d'utilisateur ou une page Facebook. Un compte utilisateur est relié à une personne physique, elle est personnelle, alors qu'une page peut représenter aussi bien une institution, une entreprise ou une personne publique. Un compte utilisateur donné peut ainsi contrôler plusieurs pages. Une application peut accéder à tous les posts publics publiés sur une page, mais pas à ceux publiés sur un compte utilisateurs (alors qu'il est possible de les lire manuellement, en naviguant sur le site Facebook). Or, certaines personnalités politiques utilisent une page, mais d'autres utilisent un compte personnel (et certains les deux simultanément) : notre outil ne peut accéder qu'aux premiers. Cette limitation empêche aussi d'explorer profondément les réactions du public aux posts publiés par les personnes étudiées : il est possible d'obtenir les commentaires émis directement en réponse de leurs posts, mais pas d'aller analyser leurs propres posts, qui sont publiés sur leur compte personnel.

9 Analyse manuelle des contenus de l'écho Web-médiatique Hidalgo

Cette exploration des potentialités et des limites actuelles de notre outil de reconstitution automatique de l'écho Web-médiatique du travail politique édilitaire ouvre de multiples perspectives. Outre les ajustements liés aux limites identifiées ici, il s'agit de poursuive le développement de l'analyse sémantique afin d'automatiser sinon accompagner les analyses manuelles proposées ci-dessous.

9.1 Détection manuelle des thèmes et types d'interactions médiatisées

Qu'est-ce qui est médiatisé dans l'activité politique de l'élue durant la semaine observée, d'après le corpus recensé ? Parmi les 172 liens du corpus final, l'analyse manuelle permet d'identifier 39 objets distincts, renvoyant à des événements, des secteurs d'activités et des types d'interaction plus ou moins différents. Ceux-ci sont décrits dans les Tables 8 et 9. Comparons ces objets à ceux détectés automatiquement via le regroupement automatique dans les 284 pages filtrées. En se limitant aux 172 pages traitées manuellement, on obtient 55 groupes qui préservent essentiellement les événements les plus fréquents. Le recouvrement avec les objets manuel est de 0,87 en termes de NMI[17], ce qui dénote d'une très grande proximité entre les deux partitions.

Les 39 objets peuvent d'abord être répartis en 18 « secteurs d'action publique ou politique ». Leur examen montre un écho Web-médiatique principalement centré sur l'actualité relative d'abord au sexisme (7 événements), à l'occasion de la journée de la femme du 8 mars, ensuite à la compétition électorale (5 événements), en cette période d'ajustement des alliances politiques autour des candidats à l'élection présidentielle de mai 2017, et enfin à l'urbanisme, au transport, au sport et au social (3 événement chacun), dans le cadre notamment des compétences d'aménagement du maire. Cinq autres objets renvoient à deux secteurs singuliers classés à part : « la corruption » et « l'injure », lorsque la lecture des textes dévoile des mises en causes nominales de l'élue pour des malversations supposées, voir des injures à caractère sexuel, sur des forums notamment.

Secteurs (18)	***n***
Sexisme	7
Compétition électorale	5
Social	3
Sport	3

Interactions (12)	***n***
Cérémonie	9
Annonce de politique publique	7
Déclaration partisane	6
Engagement partisan	3

17 La NMI est une mesure permettant de comparer deux partitions (Fred et Jain, 2003) : 0 signifie qu'elles sont indépendantes, et 1 qu'elles sont exactement identiques.

Transports	3
Urbanisme	3
Climat	2
Culture	1
Économie	1
Fiscalité	1
International	1
Justice	1
Mémoire	1
Santé	1
Sécurité	1
Université	1
Corruption	3
Injure	2
Total	**39**

Communiqué	2
Discours de politique publique	2
Condamnation	1
Interview	1
Publication	1
Objet mise en cause	4
Objet injure	2
Objet trajectoire	1
Total	**39**

Table 8. Secteurs d'action publique ou politique (à gauche) et du type d'interaction (à droite) caractérisant les 39 objets identifiés parmi les pages subsistant après le filtrage manuel.

Une seconde classification des 39 objets peut être proposée en fonction du type d'interaction mobilisant l'élue face à ses interlocuteurs (Godmer et Marrel 2014). Les médias semblent principalement publiciser sur cette période les cérémonies, les annonces de politiques publiques et les déclarations partisanes. Pour 7 cas, l'élue est également l'objet d'une mise en cause directe, d'une injure ou d'un rappel de sa trajectoire biographique.

9.2 Identification des événements du buzz sur la semaine politique du maire de Paris

Comment ces 39 objets sont-ils médiatisés, avec quelle ampleur et par qui ? L'étude de la fréquence des publications relatives à un même objet, qui est détaillée dans la Table 9, permet de préciser la nature de l'écho Web-médiatique du travail de l'élue sur la période. Les objets donnent d'abord lieu à une couverture médiatique théorique de 4,3 publications par événement. Celle-ci n'est cependant pas régulièrement répartie et d'importantes distorsions apparaissent entre les 5 objets donnant lieu à au moins 10 publications distinctes (et jusqu'à 33), 6 autres publiés entre 3 et 8 fois, 6 autres publiés 2 fois et les 22 derniers qui ne produisent d'une seule publication. Au final, les 5 événements les plus diffusés, commentés et relayés concernent :

1. ***La vie politique nationale*** : la prise de position partisane de la maire de la capitale de la France, dans la campagne présidentielle, pour le candidat socialiste Benoît Hamon, qui donne lieu à 33 publications différentes, du fait notamment des relais médiatiques locaux et nationaux de cet engagement à caractère national dans la séquence politique disposant de la plus grande couverture médiatique ;
2. ***L'urbanisme parisien*** :l'annonce d'une série de mesure de propreté dans le cadre des politiques municipales, qui occasionne 27 publications, pour un enjeu très localisé, mais populaire ;
3. ***La culture nationale*** : le soutien de la maire de Paris au vaste projet d'investissement du milliardaire Bernard Arnaud et de LVMH autour du Musée des arts et des traditions populaires, qui produit 20 publications en ligne, pour un événement de niveau national très relayé dans le monde de la culture ;

4. ***La justice*** : l'annonce de la condamnation en appelle d'Anne Hidalgo pour diffamation à l'encontre de l'architecte Jean-François Cabestan, qui donne lieu à 17 publications qui reflète l'importance médiatique de la personne du maire de Paris et des poursuites judiciaires qui pourraient affecter son image ;
5. ***Les transport et la sécurité routière*** : l'appui de la maire de Paris à une « campagne mondiale » de sensibilisation associant de nombreuses personnalités médiatiques, largement diffusé et donnant lieu ici à 10 publications recensées.

Viennent ensuite 4 événements avec au moins 5 médiatisations : le déplacement du maire avec le Premier ministre Emmanuel Valls à Barcelone pour un match Barça-PSG, deux déclarations critiques à l'égard du programme du candidat Macron, et une petite polémique sur la présence de Nicolas Sarkozy à Barcelone. Si l'on rassemble l'ensemble des publications relatives au ralliement d'Anne Hidalgo à la candidature de Benoît Hamon et ses suites immédiates en terme de déclarations dans le courant de cette semaine, c'est clairement l'action partisane et électorale nationale qui domine l'écho Web-médiatique et l'image de l'élue parisienne entre le 6 et le 12 mars dernier.

	Événements	Interactions	Secteurs	*f*
1	Évoquée comme soutien déclaré à Hamon, contrairement à Delanoë soutenant Macron	Engagement partisan	Compétition électorale	33
2	Annonce 10 mesures sur la propreté à Paris	Annonce PP	Urbanisme	27
3	Soutien au projet de Musée des arts et traditions populaires de Bernard Arnaud (LVMH)	Cérémonie	Culture	20
4	Condamnée en appel pour diffamation d'un architecte	Condamnation	Justice	17
5	Soutien campagne sécurité routière	Cérémonie	Transports	10
6	Assiste au match Barça-PSG	Cérémonie	Sport	8
7	Critique du programme Macron sur le social	Déclaration partisane	Social	7
8	Critique du programme Macron sur la taxe d'habitation	Déclaration partisane	Fiscalité	5
9	Évoquée dans la polémique concernant Sarkozy au match Barça-PSG	Cérémonie	Sport	5
10	Ouverture d'un centre d'accueil de réfugiés	Cérémonie	Social	3
11	S'engage à respecter l'accord de Paris avec M. Bloomberg	Annonce PP	Climat	3
12	Annonce d'un engagement à la diminution des ondes électromagnétiques	Annonce PP	Santé	2
13	Annonce hausse des amendes de stationnement	Annonce PP	Transports	2
14	Assiste à une cérémonie concernant les JO et l'ESS	Cérémonie	Sport	2
15	Évoquée comme épouse de JM Germain, nommé porte-parole de B. Hamon	Engagement partisan	Compétition électorale	2
16	Projet Sea bubble (voiture volante)	Annonce PP	Transports	2
17	Réaction à l'enfermement du maire de Dakar SG de l'AIMF	Communiqué	International	2
18	Annonce des nouvelles fontaines du rond-point des Champs Élysées	Annonce PP	Urbanisme	1
19	Annonce mesures de lutte contre le dérèglement climatique	Annonce PP	Climat	1
20	Annonce veto à la candidature d'Audrey Azoulay	Déclaration partisane	Compétition électorale	1
21	Appelle à ne pas faire élire Fillon	Déclaration partisane	Compétition électorale	1
22	Assiste à la soirée des Prix du Grand Paris	Cérémonie	Économie	1
23	Assiste à une cérémonie à Colombey	Cérémonie	Mémoire	1
24	Critique du programme macron sur les retraites	Déclaration partisane	Social	1
25	Déclaration sur l'affaire Baupin à la matinale Europe1	Déclaration partisane	Sexisme	1
26	Diffusion le 8 mars d'un enregistrement d'un message pour les femmes russes	Discours PP	Sexisme	1
27	Discours à l'Agora de Dauphine	Discours PP	Université	1
28	Évoquée comme soutien déclaré à Hamon, contrairement à Villani qui soutient Macron	Engagement partisan	Compétition électorale	1
29	Évoquée dans des allusions sexuelles	Objet injure	Injure	1

30	Évoquée pour son action comme adjointe de Delanoë	Objet trajectoire	Sexisme	1
31	Injuriée par le maire du Plessi-Robinson	Objet injure	Injure	1
32	Interviewée pour la Lycéenne Maif Run	Interview	Sexisme	1
33	Lance la campagne de l'ONU concernant les femmes élues locales	Cérémonie	Sexisme	1
34	Mise en cause pour des malversations supposées 1	Objet mise en cause	Corruption	1
35	Mise en cause pour des malversations supposées 2	Objet mise en cause	Corruption	1
36	Mise en cause pour des malversations supposées 3	Objet mise en cause	Corruption	1
37	Mise en cause pour la politique concernant les salles de shoot	Objet mise en cause	Sécurité	1
38	Publication d'une tribune féministe	Publication	Sexisme	1
39	Soutien à l'entrepreneuriat des femmes	Communiqué	Sexisme	1

Table 9. Liste des 39 objets identifiés dans les pages filtrées automatiquement puis manuellement.

L'examen des 5 premiers du classement des 4 moteurs de recherche Web, décrit dans la Table 10, montre ici plus de cohérence : les événements qui font le « buzz » remontent partout en première ou deuxième position sur la semaine, parfois plusieurs fois dans le Top-5. Seule l'inauguration du projet culturel LVMH n'y figure pas. Plus étonnant : deux événements n'ayant donné lieu qu'à une seule publication, donc peu relayées figurent en 3ème et 4ème rangs dans nos moteurs : l'un sulfureux, puisqu'il s'agit de l'évocation d'une injure prononcée par un maire de l'agglomération parisienne, l'autre engagé et fortement lié à la journée mondiale des femmes du 8 mars, concernant la publication d'une tribune féministe.

Top5	Bing	Google	Qwant	Yandex
1	Annonce 10 mesures sur la propreté à Paris	Annonce 10 mesures sur la propreté à Paris	Annonce 10 mesures sur la propreté à Paris	Annonce 10 mesures sur la propreté à Paris
2	Évoquée comme soutien déclaré à Hamon, contrairement à Delanoë qui soutient Macron	Évoquée comme soutien déclaré à Hamon, contrairement à Delanoë qui soutient Macron	Évoquée comme soutien déclaré à Hamon, contrairement à Delanoë qui soutient Macron	Condamnée en appel pour diffamation d'un architecte
3	Injuriée par le maire du Plessi-Robinson	Annonce 10 mesures sur la propreté à Paris	Injuriée par le maire du Plessi-Robinson	Publication d'une tribune féministe
4	Assiste au match Barça-PSG	Publication d'une tribune féministe	Assiste au match Barça-PSG	Annonce 10 mesures sur la propreté à Paris
5	Condamnée en appel pour diffamation d'un architecte	Condamnée en appel pour diffamation d'un architecte	Condamnée en appel pour diffamation d'un architecte	Annonce 10 mesures sur la propreté à Paris

Table 10. Liste des 5 premiers événements recensés par chacun des 4 moteurs de recherche Web.

9.3 Analyse des supports de publication et des types d'auteurs

L'analyse manuelle permet d'identifier 5 types de supports et 4 types d'auteurs différents, listés dans la Table 11. Parmi les supports, les sites de presse écrite représentent plus de 71% des résultats valides. Avec les sites de télévisions et de radios, les supports Web des médias institutionnels représentent plus de 80% des sources recensées par les moteurs de recherche. Les blogs ne comptent que pour 18% et les forums pour moins d'1%.

5 types de supports	%
Site de presse	71,5%
Blogs	18%
Site de TV	6,4%

4 types d'auteurs	%
Journaliste	95,4%
Militant	2,3%
Entreprise	1,2%

Site de radio	3,5%
Forum	0,6%

Individu	1,2%

Table 11. Types de supports (à gauche) et d'auteurs (à droite) identifiés parmi les pages issues du filtrage manuel.

Les auteurs des messages compilés dans l'écho Web-médiatique sont principalement des journalistes (95%). Les militants, les acteurs économiques et les individus isolés sont très marginaux dans la diffusion de messages concernant la maire de Paris.

9.4 La comparaison entre l'écho Web-médiatique et l'agenda de la maire

L'agenda électronique de la semaine de travail politique observé a été transmis par le chef de cabinet du maire, dans un format texte, après élimination des éléments confidentiels. Nous l'avons recomposé sous forme de tableau analytique (Table 12), dans lequel nous proposons une comparaison des 30 événements planifiés dans l'emploi du temps du maire avec les 39 événements de l'écho médiatique recensés sur le Web. On observe immédiatement que seules 11 actions programmées par le maire figurent sous une forme médiatisée sur le Web. Il s'agit des événements délibérément construits sur la période par l'équipe politique et les journalistes comme des rendez-vous médiatiques à publiciser : déclaration lors de la matinale d'Europe1, interview, conférence de presse, cérémonie liée à la culture, aux Jeux Olympiques ou aux Accords de Paris sur le climat. Le déplacement pour un match du PSG à Barcelone est d'autant plus commenté qu'Anne Hidalgo accompagne le Premier ministre Manuel Vals et que l'événement donne lieu à un incident diplomatique avec l'ancien Président Sarkozy. En revanche, 19 engagements de l'élue ne laissent apparemment aucune trace en ligne. Il s'agit la plupart du temps de rendez-vous en tête à tête ou de réunions d'équipe, de conseils ou de thématiques, peu ou pas médiatisés et dont il n'est sans doute pas étonnant de ne retrouver aucun écho sur le Web[18].

L'écho Web-médiatique dépasse par ailleurs assez largement l'action engagée sur la période. Entre le 6 et le 12 mars, la maire de Paris est citée ou interpellée sur pas moins de 13 objets dont l'origine est antérieure au travail politique de la semaine en cours. Il s'agit soit de « vieilles histoires » comme lorsqu'on évoque l'ancienneté de son engagement féministe comme adjointe de Bertrand Delaoë, d'engagements ou d'affaires datant de plusieurs mois comme les injures du maire du Plessi-Robinson, la cérémonie à Colombey ou l'ouverture d'un centre de réfugié, soit encore d'annonces ou de prises de positions remontant à quelques jours comme la déclaration du maire sur la diminution des ondes électromagnétiques du 3 mars ou encore le soutien qu'elle apporte vers le 5 mars au candidat B. Hamon face à B. Delanoë engagé auprès d'E. Macron.

18 Plus étonnant est le silence de notre outil de recensement concernant les probables publications sur la visite officielle du maire d'Erevan le mardi 7 mars et celle du président Arménien le lendemain, qui occupe le maire, son équipe et une partie du conseil de 11h00 à 13h30. Vérification faite, il s'agit d'un défaut de compilation lors de la dernière extraction, lié à l'instabilité des retour des moteurs dans le temps. Un rapide passage en revue des 43 URL proposées par les moteurs le 20 juillet, mais disparus dans les résultats du 12 septembre, montre en effet que deux publications évoquent bien cette visite officielle, l'une sur le site de l'Ambassade d'Arménie en France, l'autre sur un blog francophone de la diaspora arménienne. Ce cas met en question la fiabilité de l'outil reposant sur les 4 moteurs. Mais d'autre côté, il ne semble ne pas y avoir eu d'autre publications de cet événement dans les médias institutionnels français que les moteurs programmés ont vraisemblablement fidèlement recensés.

Agenda professionnel d'Anne Hidalgo 30 événements				Écho Web-médiatique 39 événements			
Dates/heures	**Événements**	**Interlocuteurs**	**Lieu**	**Objet**	**Date événement ou publication**	**Source**	**f**
< 06/03/2017				Assiste à une cérémonie à Colombey	< 06/03/2017	Blogue	1
				Évoquée comme épouse de JM Germain, nommé porte-parole de B. Hamon		Site de presse	2
				Évoquée pour son action comme adjointe de Delanoë		Site de presse	1
				Mise en cause pour des malversations supposées 3		Blog	1
				Mise en cause pour la politique concernant les salle de shoot		Site de presse	1
				Injuriée par le maire du Plessi-Robinson	04/06/16	Site de presse	1
				Annonce veto à la candidature d'Audrey Azoulay	01/11/16	Site de presse	1
				Ouverture d'un centre d'accueil de réfugiés	10/11/16	Site de presse	2
						Site de TV	1
				Annonce mesures de lutte contre le dérèglement climatique	01/02/17	Blog	1
				Critique du programme macron sur le social	05/02/17	Blog	1
						Site de presse	6
				Critique du programme macron sur les retraites		Site de presse	1
				Évoquée comme soutien déclaré à Hamon, contrairement à Delanoë qui soutient Macron		Blog	1
						Site de presse	11
						Site de TV	2
				Annonce d'un engagement à la diminution des ondes électromagnétiques	03/03/17	Blog	2
Lundi 6 mars	2 événements						
8h30	Rendez-vous	SG	Bureau maire				
9h00	Réunion hebdomadaire	SG + Cabinet	Salon du 1er				
Mardi 7 mars	7 événements						
8h00	Matinale Europe 1	Journaliste	Europe 1	Déclaration sur l'affaire Baupin à la matinale Europe1	07/03/17	Site de presse	1
9h30	Visite Tremplin		Stade Jean Bouin				

11h00	Interview téléphonique pour Reuters sur le Women4climate	Journaliste		Soutien à l'entrepreneuriat des femmes	07/03/17	Blog	1
11h30	Rendez-vous	Cédric Villani + comité de soutien de l'Union Math. Internat.	Bureau maire				
13h30	Déjeuner	Maire d'Erevan	Salon du 1er				
15h00	Conseil de Juridiction	Conseillers et adjoints	Palais de justice				
17h30	Débat avec les étudiants de l'Université Paris Dauphine	Etudiants	Paris Dauphine	Discours à l'Agora de Dauphine	07/03/17	Site de presse	1
				Critique du programme macron sur la taxe d'habitation		Blog	3
						Site de presse	2
				Évoquée dans des allusions sexuelles		Forum	1
				Publication d'une tribune féministe		Site de presse	1
				Réaction à l'enfermement du maire de Dakar SG de l'AIMF		Site de presse	2
Mercredi 8 mars	10 événements						
9h00	Réunion de l'exécutif municipal	Adjoints, maires d'arrondissement...	Salle de commission n°1				
11h00	Rendez-vous	Ambassadeur de Grande Bretagne	Bureau maire				
12h00	Visite officielle (signature d'un parchemin)	Président de la République d'Arménie	Bureau maire				
12h30	Réception officielle	Président de la République d'Arménie	Salle des fêtes				
13h00	Coktail	Président de la République d'Arménie	Salle des fêtes				
13h15	Départ officiel	Président de la République d'Arménie					
14h15	Conférence de presse Musée des Arts et Traditions Populaires	Président de la République FH	Verrière du Jardin d'Acclimatation	Soutien au projet de Musée des arts et traditions populaires de Bernard Arnaud (LVMH)	08/03/17	Blog	5
						Site de presse	14
						Site de radio	1
17h40	Départ pour Barcelone		CDG	Assiste au match Barça-PSG			
20h45	Match Ligue des Champions : FC Barcelone / PSG		Barcelone			Blog	2
						Site de presse	5

						Site de radio	1
				Évoquée dans la polémique concernant Sarkozy au match Barça-PSG		Blog	1
						Site de presse	3
						Site de TV	1
2h50	Retour de Barcelone avec les joueurs		CDG				
				Appelle à ne pas faire élire Fillon	08/03/17	Site de radio	1
				Diffusion le 8 mars d'un enregistrement d'un message pour les femmes russes		Site de presse	1
				Évoquée comme soutien déclaré à Hamon, contrairement à Delanoë qui soutient Macron		Site de presse	17
						Site de TV	1
				Interviewée pour la Lycéenne Maif Run		Blog	1
				Mise en cause pour des malversations supposées 1		Blog	1
Jeudi 9 mars	5 événements						
10h00	Rendez-vous	FH et Mickael Bloomberg	Palais de l'Élysée	S'engage à respecter l'accord de Paris avec M. Bloomberg	09/03/17	Site de presse	3
11h30	Réunion sur le plan budgétaire pluriannuel 2018-2020	Adjoint, conseillers...	Bureau de la Maire				
13h00	Déjeuner	Adjoint	Salon du 1er				
14h30	Interview téléphonique	Journaliste	Bureau de la Maire	?			
17h30	Visite inaugurale de la Maison de santé Pluri-professionnelle		Porte de Vanves, 14e				
				Assiste à la soirée des Prix du Grand Paris	09/03/17	Blog	1
				Condamnée en appel pour diffamation d'un architecte	09/03/17	Blog	3
						Site de presse	13
						Site de TV	1
Vendredi 10 mars	5 événements						
8h48	Accueil du Professeur Muhammad Yunus		Bureau de la Maire	Assiste à une cérémonie concernant les JO et l'ESS	10/03/17	Site de presse	2
9h15	Passage à la mobilisation pour des J.O. inclusifs		Auditorium de l'Hôtel de Ville				

11h30	Lancement de la campagne mondiale sécurité routière		Siège parisien de la FIA	Soutien campagne sécurité routière	10/03/17	Blog	1
						Site de presse	9
15h15	Rendez-vous	Guillaume Pepy	Bureau de la Maire				
17h00	Rendez-vous	Préfet de Paris Ile-de-France	Bureau de la Maire				
				Évoquée comme soutien déclaré à Hamon, contrairement à Delanoë qui soutient Macron	10/03/17	Site de radio	1
				Évoquée comme soutien déclaré à Hamon, contrairement à Villani qui soutient Macron		Site de radio	1
				Mise en cause pour des malversations supposées 2		Blog	1
Samedi 11 mars	0 événements						
Dimanche 12 mars	1 événements						
12h10	Départ pour Chicago		CDG				
				Annonce 10 mesures sur la propreté à Paris	12/03/17	Blog	4
						Site de presse	17
						Site de radio	1
						Site de TV	5
>12/03/2017				Annonce des nouvelles fontaines du rond-point des Champs Élysées	2018	Site de presse	1
				Annonce hausse des amendes de stationnement	2018	Site de presse	2
				Projet Sea bubble (voiture volante)	2020	Site de presse	2

Table 12. Détail de la correspondance entre l'agenda et l'écho Web-médiatique.

Dans l'écho médiatique, le plus intéressant concerne les 15 publications concernant vraisemblablement la période observée ne correspondant à aucune base dans l'agenda professionnel communiqué par l'équipe politique du maire : 4 le mardi 7 mars (sur la critique du programme de Macron, la publication d'une tribune féministe ou encore la réaction de la maire à l'annonce de l'incarcération du maire de Dakar, par ailleurs secrétaire général de l'AIMF, qu'elle préside) ; 5 le mercredi 8 mars (sur deux messages féministes, son soutien à Hamon encore une fois et une critique du candidat Fillon) ; 2 le jeudi 9 mars (dont sans surprise sa condamnation en appel pour diffamation qui fait le buzz avec 27 publications le jour même) ; 3 le vendredi 10 mars (sur son divorce avec Delanoë soutenant Macron) ; et enfin 1 le dimanche 12 mars (sur l'annonce très relayée des dix mesures pour la propreté de Paris). Sur cette dernière le silence de l'agenda est suspect. De fait, l'annonce est faite dans les colonnes du JDD et donc sans doute quelques jours avant pour l'agenda, peut-être à l'occasion de l'interview téléphonique du 9 mars qui ne correspondait encore à rien de précis dans l'écho-médiatique.

Enfin, trois publications ne renvoient à rien dans l'agenda de la semaine puisqu'elles font références à des annonces passées pour des projets futurs : une nouvelle hausse des amendes de stationnement et de nouveaux aménagements urbains pour le rond-point des Champs-Élysées pour 2018 et un projet de voiture volante (Sea Bubble) pour 2020.

Ces distorsions observées dans l'analyse relèvent de certaines limites de l'ouillage disponible en l'état pour compiler toute l'information circulant en ligne sur l'activité d'une personnalité publique. Elles révèlent surtout certaines dynamiques propres au temps médiatique, ses retards, son élasticité. Seuls deux événements du top5 des faits les plus médiatisés sont dans l'agenda : la conférence de presse au Musée des Art populaire et celle sur la campagne de sécurité routière. Les 3 autres objets les plus diffusés relèvent moins d'interaction politique programmables dans l'emploi du temps que de déclarations, discours ou prises de position : 1) d'une position politique adoptée quelques jours avant l'observation et ses conséquences évoquées et commentées tout au long de la semaine ; 2) la condamnation en appel et 3) les dix mesures de propreté.

10 Conclusions et perspectives

Le chantier de l'équipement automatique de l'analyse sociologique de l'écho Web-médiatique du travail politique doit être poursuivi afin d'affiner les conclusions. Nonobstant l'incertitude relative des résultats générés par les moteurs de recherche et en l'absence des éléments circulant sur les médias sociaux, l'enquête invite d'ors et déjà à ajuster les hypothèses formulées au point 3 sur la structure de la médiatisation politique 2.0. En particulier, la diversité hypothétique des publications attendues, de leurs sources et de leurs supports n'est pas réellement présente dans les résultats recensés et appelle de nouveaux développements et expérimentations.

Un certain nombre de limitations de l'outil pourraient en effet être résolues de différentes manières. Pour ce qui concerne la pertinence des résultats renvoyés, deux approches complémentaires sont possibles : améliorer le filtrage, ce qui réduirait le nombre de faux positifs, et augmenter le nombre d'URL pertinentes renvoyées, ce qui diminuerait le nombre de faux négatifs. Pour améliorer le filtrage, nous pourrions tirer parti de l'utilisation conjointe de plusieurs moteurs de recherche, en donnant plus de poids aux pages renvoyées par plusieurs d'entre eux, aux détriment de celles renvoyées potentiellement accidentellement par un seul moteur. L'augmentation du nombre d'URL pertinentes pourrait passer par la généralisation de l'approche itérative utilisée dans ce travail. On effectue une première recherche, dans laquelle on identifie les sources qui semblent particulièrement pertinentes et/ou prolixes, et on complète nos résultats avec des recherches supplémentaires ciblant ces sites-là (comme nous l'avons fait ici avec la presse locale).

Pour ce qui concerne l'extraction du contenu textuel des pages Web, la solution la plus évidente consiste à définir plus d'extracteurs spécifiques, par exemple en listant tous les titres de

presse présent sur le Web et en les traitant tous un par un. Cependant, ce travail est coûteux en temps, et peu robuste : il suffit qu'un site change sa mise en page (ce qui se produit régulièrement) pour rendre l'extracteur correspondant obsolète. L'extension des règles de notre extracteur générique semble plus appropriée, mais il est impossible de prévoir tous les cas possibles. Une autre solution consisterait à adopter une approche à base d'apprentissage artificiel, i.e. un logiciel qui s'entraînerait à réaliser la tâche d'extraction. Cependant, pour obtenir des résultats concluant, il faudrait pour se faire constituer un corpus conséquent, et l'annoter manuellement.

Il est très difficile de proposer des solutions aux limitations relative aux médias sociaux, car elles dépendent directement de l'API qu'ils imposent. Et le but de ces sociétés commerciales est clairement, pour l'heure, de reprendre le contrôle des données qu'elles hébergent afin de les exploiter au mieux. Une solution technique serait de contourner l'API imposée, et de créer un extracteur spécifique pour chaque média social visé. Cependant, outre les grandes difficultés techniques et le coût élevé de cette approche, elle est explicitement interdite dans les conditions générales d'utilisation de ces services. Une autre solution est de négocier directement des accords avec les médias sociaux concernés, mais ceux-ci disposent de leurs propres centres de recherches, et sont réticents à la collaboration scientifique (hormis avec les institutions prestigieuses). Dans l'immédiat, la meilleure solution semble être d'étendre l'outil pour exploiter au mieux l'information disponible, qui est essentiellement non-textuelle : nombre de partages d'un post, de commentaires, de likes, etc.

Enfin, l'outil pourrait également être étendu en approfondissant l'analyse du contenu textuel qu'il peut d'ores et déjà obtenir. Ceci passe par l'intégration d'outils d'analyse linguistique plus poussés, tels que la détection et la résolution de coréférences, mais aussi l'analyse de sentiment, qui pourrait permettre d'identifier la polarité d'un texte pour déterminer s'il est positif ou négatif relativement à la personne ciblée.

11 Références bibliographiques

AGGARWAL, C.C., ZHAI, C. (dirs.), 2012, *Mining Text Data*, New York, Springer.

ALDRIN, P., HUBÉ, N., OLLIVIER-YANIV, C., UTARD, J.-M. (dirs.), 2014, *Les mondes de la communication publique: légitimation et fabrique symbolique du politique*, Rennes, France, Presses universitaires de Rennes, 189 p.

ATDAĞ S., LABATUT V., 2013, « A comparison of named entity recognition tools applied to biographical texts », *2nd International Conference on Systems and Computer Science*, p. 228-233.

BAILEY F.G., 1971, *Les règles du jeu politique: étude anthropologique*, traduit par COPANS J., Paris, Presses universitaires de France, 254 p.

BALANDIER G., 2006, *Le pouvoir sur scènes*, Paris, Fayard, 248 p.

BLONDEAU-COULET O., ALLARD L., 2007, *Devenir média: l'activisme sur Internet, entre défection et expérimentation*, Paris, France, Éd. Amsterdam, impr. 2007, 389 p.

BUÉ N., 2011, « Le web partisan dans une ville moyenne. Une ressource d'usage limité », dans GREFFET F. (dir.), *Continuerlalutte.com*, Presses de Sciences Po (Académique), p. 215-230.

CADIOU S., 2009, *Le pouvoir local en France*, Grenoble, Presses universitaires de Grenoble, impr. 2009, 206 p.

CARDON D., 2010, *La démocratie Internet: promesses et limites*, Paris, France, Seuil, DL 2010, 101 p.

CARTON M., 2014, « Clémence Pène, e-déaliste », *Les Inrocks*.

CHIBOIS J., 2014, « Twitter et les relations de séduction entre députés et journalistes. La salle des Quatre Colonnes à l'ère des sociabilités numériques », *Réseaux*, *188*, 6, p. 201-228.

COLEMAN S., BLUMLER J.G., 2009, *The Internet and democratic citizenship: theory, practice and policy*, Cambridge, Royaume-Uni, Etats-Unis, ix+220 p.

DEMAZIÈRE, D., LE LIDEC, P. (dirs.), 2014, *Les mondes du travail politique: les élus et leurs entourages*, Rennes, Presses universitaires de Rennes, 264 p.

FLORINI A., 1998, « The end of secrecy », *Foreign Policy*, 111, p. 50-63.

FRED A.L.N., JAIN A.K., 2003, « Robust data clustering »,.

FROCHOT D., MOLINARO F., 2010, *E-réputation : suivre, soigner, défendre l'image de l'élu local sur le Net*, Territoriales éditions, Voiron, 88 p.

GIBSON R.K., WARD S.J., 2009, « Parties in the Digital Age. A Review Article », *Representation*, *45*, 1, p. 87-100.

GODMER L., MARREL G., 2014a, « La production de l'agenda. Comment se fabrique l'emploi du temps d'une vice-présidente de conseil régional. », dans DEMAZIÈRE D., LE LIDEC P. (dirs.), *Les mondes du travail politique. Les élus et leurs entourages*, PUR, Rennes, p. 37-52.

GODMER L., MARREL G., 2014b, « Que font vraiment les professionnels de la politique ? L'agenda électronique et l'emploi du temps d'une élue régionale », dans MAZEAUD A. (dir.), *Pratiques de la représentation politique*, PUR, Rennes, p. 139-162.

GRANJON F., 2001, *L'Internet militant: mouvement social et usages des réseaux télématiques*, Rennes, France, Éd. Apogée, 189 p.

GREFFET, F. (dir.), 2012, *Continuerlalutte.com*, Paris, France, Presses de Sciences Po, 313 p.

HARFOUCHE A., 2012, « The same wine but in new bottles. Public e-services divide and low citizens' satisfaction: an example from Lebanon », dans *Technology Enabled Transformation of the Public Sector: Advances in E-Government*, IGI Global, p. 267-300.

JDN, 2015, « Parts de marché des moteurs de recherche en France », *Journaldunet.com*.

LABATUT V., 2013, « Improved Named Entity Recognition Through SVM-Based Combination »,.

LAGROYE J., 1985, « La légitimation », dans *Traité de science politique*, p. 395-467.

LEFEBVRE R., 2014, « Les élus comme entrepreneurs de temps : les agendas des cumulants. », dans DEMAZIÈRE D., LE LIDEC P. (dirs.), *Les mondes du travail politique. Les élus et leurs entourages*, PUR, Rennes, p. 53-70.

MANNING C.D., RAGHAVAN P., SCHÜTZE H., 2008, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge University Press.

MARQUES F.P.J.A., AQUINO J.A. DE, MIOLA E., 2014, « Congressmen in the age of social network sites: Brazilian representatives and Twitter use », *First Monday*, *19*, 5.

MARREL G., 2016, *Gouverner par le temps : Sociologie politique des agendas personnels et des emplois du temps d'acteurs publics*, Accreditation to supervise research, Université d'Avignon.

MARREL G., LABATUT V., EL BÈZE M., 2015, « Le Web comme miroir du travail politique quotidien ? Reconstituer l'écho médiatique en ligne des événements d'un agenda d'élu », Aix-en-Provence, juin 2015.

MARREL G., PAYRE R., 2006, « Temporalités électorales et temporalités décisionnelles. Du rapport au temps des élus à une sociologie des leaderships spatio-temporels », *Pôle Sud*, *25*, 2, p. 71-88.

MEIJER A.J., CURTIN D., HILLEBRANDT M., 2012, « La gouvernance ouverte: relier visibilité et moyens d'expression », *Revue internationale des sciences administratives*, *78*, 1, p. 13-32.

NG V., 2010, « Supervised Noun Phrase Coreference Research: The First Fifteen Years », *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1396–1411.

NICOT A.-L., 2012, « Les sites internet des députés, terrains annexes de lutte partisane », dans *Continuerlalutte.com*, Presses de Sciences Po, Paris, p. 95-105.

NORTON P., 2007, « Four Models of Political Representation: British MPs and the Use of ICT », *The Journal of Legislative Studies*, *13*, 3, p. 354-369.

PROULX S., 2015, « La sociologie des usages, et après ? », *Revue française des sciences de l'information et de la communication*, 6.

ROGINSKY S., PERRIER V.J., 2014, « La fabrique de la communication des parlementaires européens : "Tweet ton député" et les "ateliers du député 2.0" », *Politiques de communication*, *3*, 2, p. 85-124.

ROUSSEEUW P.J., 1987, « Silhouettes: A graphical aid to the interpretation and validation of cluster analysis », *Journal of Computational and Applied Mathematics*, *20*, Supplement C, p. 53-65.

THÉVIOT A., 2016, « Les data : nouveau trésor des partis politiques ? : Croyances, constitutions et usages comparés des données numériques au Parti Socialiste et à l'Union pour un Mouvement Populaire », *Politiques de communication*, *6*, 1, p. 137-166.