SOCIETE FRANÇAISE DE STATISTIQUE

Les cafés de la statistique

"La statistique éclaire-t-elle les questions de société" ?

Soirée du 14 avril 2015

Le journalisme de données

Synthèse des débats [*]

Depuis quelques années, un certain nombre de journalistes se rassemblent autour de l'expression "Journalisme de données" pour promouvoir un renouvellement de leur activité professionnelle par l'exploitation et la mise à disposition du public de données statistiques. Beaucoup d'entre eux appartiennent à des médias sur Internet. Dans certains cas, il s'agit d'un effort pour améliorer l'utilisation des données chiffrées traditionnelles par les médias : mieux transmettre les messages importants, notamment à l'aide d'infographies pertinentes, mais aussi critiquer les usages abusifs des données dans l'espace public. C'est ainsi qu'on voit se multiplier dans les journaux et les radios, ou sur des sites spécialisés, les rubriques qui décortiquent l'usage des chiffres par les hommes politiques. Dans d'autres cas, le journaliste recherche l'information dont il a besoin dans des bases de données, ou la recueille avec la participation de son public ; après quoi il la traite et la met en forme lui-même. C'est une nouvelle forme d'investigation journalistique qui apparaît ainsi, dans laquelle le journaliste a forcément recours à des outils statistiques.
Quelles sont les potentialités de ce mouvement, et quels en sont les risques ? Comment statisticiens et journalistes peuvent-ils coopérer pour que le public dispose d'une information dont il puisse pleinement tirer parti, en en comprenant bien l'origine et la portée ?

Invité :
Alexandre Léchenet, journaliste-web à Libération

[*] Tant l'exposé liminaire que le contenu des échanges sont structurés en quelques thèmes, sans suivre l'ordre chronologique. Par ailleurs, l'identité des intervenants n'était pas toujours connue et l'on a choisi de ne pas attribuer nominativement les propos. Au reste, ceux-ci ont été reconstitués à partir des notes du secrétariat sans reprendre leur formulation détaillée. Pour retracer le débat, les thèmes sont souvent introduits sous forme d'une question : ce qui vient ensuite n'est pas la seule réponse de l'invité, mais l'ensemble des contributions des participants.

Exposé introductif :

Alexandre Léchenet explique qu'il fait du data journalisme à Libération depuis un mois après une expérience analogue de quelques années au journal Le Monde, expérience qui nourrira pour l'essentiel son propos.

Qu'est-ce que le data journalisme[1] ? Selon la définition qu'on en donne, on peut considérer que cela existe depuis longtemps. Il s'agit en première approximation d'utiliser les moyens informatiques pour faire du journalisme, c'est-à-dire collecter, exploiter, interpréter et mettre en forme de l'information. On peut se référer au *computer-assisted reporting* (CAR) http://en.wikipedia.org/wiki/Computer-assisted_reporting aux États-Unis[2]. En juillet 1967, lors des émeutes à Détroit, Philip Meyer, journaliste d'investigation qui étudie les sciences sociales, décide d'appliquer leurs méthodes au journalisme. Il fait un sondage sur la population des manifestants et en tire des articles et une matière qui nourrit des travaux de chercheurs[3]. Il s'agissait de journalisme assisté par ordinateur. Ainsi, l'informatique crée de nouveaux moyens de faire du journalisme. Depuis 2009, le développement de l' « open data »[4] ajoute une dimension : voir le Guardian et son blog ; voir aussi le site OWNI[5] en France.

Le data journalisme regroupe aujourd'hui des développeurs, des journalistes et des graphistes et permet de faire de l'interactif « joli » (celui qui suscite un « Waouh ! quand on le découvre...). Alexandre Léchenet a travaillé au site OWNI pendant six mois en 2011. Le terme de data journalisme peut s'appliquer aussi bien à la matière rédactionnelle qu'au graphisme en ligne ou à l'infovisualisation ou encore s'étendre aux « quizz »[6] journalistiques proposés aux lecteurs. Il y a donc une très vaste acception du terme mais il s'agit dans tous les cas de travailler sur des données informatiques[7].

L'invité explique qu'il est entré au Monde en septembre 2011 avec un bagage de formation en mathématiques, en informatique et en communication. L'un de ses premiers travaux de data journalisme a porté sur les dépassements des honoraires des médecins. Les données n'étaient pas accessibles sous forme de fichier, mais le site http://ameli-direct.ameli.fr/ permet de connaître les

---

[1] NDR : on retient dans la suite l'expression « data journalisme » (qu'en toute rigueur il faudrait écrire en anglais : data journalism) pour désigner le « journalisme de données ».

[2] Selon Wikipédia, le terme « reporting » peut désigner une technique informatique de préparation de rapports, consistant à extraire des données pour les présenter dans un rapport plus facilement lisible, voire pédagogique ou de vulgarisation (affichable ou imprimable).

[3] NDR : Il établit notamment qu'il n'y a pas de corrélation entre statut économique et participation aux émeutes, et que les personnes d'origine étrangère n'ont pas joué un rôle majeur dans ces émeutes.

[4] Wikipédia : Une **donnée ouverte** (open data) est une donnée numérique d'origine publique ou privée. Elle peut être notamment produite par une collectivité, un service public (éventuellement délégué) ou une entreprise. Elle est diffusée de manière structurée selon une méthodologie et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière.

[5] NDR : OWNI pour « objet web non identifié » a été un site Internet français de distribution libre créé en 2009 et placé en liquidation judiciaire fin 2012. Ses archives sont restées en ligne : http://owni.fr/

[6] Jeux à base de questions pour tester ses connaissances ou ses compétences.

[7] NDR : Wikipédia a donné un temps la définition suivante : « Le journalisme de données est une nouvelle technique journalistique qui consiste à analyser des données complexes (par exemple des statistiques sportives) ou à extraire des informations pertinentes de quantités importantes de données…. La question de la visualisation des données est également un aspect important de ce type de journalisme."

tarifs des médecins individuellement. Les tarifs de consultation pour chaque médecin parisien ou des neuf autres plus grandes villes du pays ont donc été récupérés systématiquement. Un « script »[8] a été utilisé pour visiter les pages de chacun des médecins de ces dix villes et recueillir son nom, son adresse et les tarifs pratiqués. À Paris, 3 000 tarifs de médecins ont ainsi été récupérés. Cela a permis un dialogue serré avec le directeur de la sécurité sociale, puisqu'il est apparu que le tarif de base était dépassé en moyenne de 14 € et que certains honoraires allaient jusqu'à cinq ou huit fois le tarif de base. Une double page a été publiée dans Le Monde et des pages ainsi que des cartes ont été mises en ligne sur le site Internet du quotidien. On pouvait observer que les dépassements d'honoraires se concentraient auprès des hôpitaux, notamment dans l'Ouest parisien, cela étant à mettre en relation avec les pratiques de consultation privée de médecins hospitaliers à l'hôpital. Le journal ayant appelé des médecins pour avoir leurs commentaires, ceux-ci ont fait part des tolérances existant en matière de dépassements d'honoraires ; ces derniers doivent être pratiqués « avec tact et mesure », sans autre précision. Tout cela a débouché fin 2012, sous l'égide de la ministre Marisol Touraine, sur des accords précisant la notion de « tact et mesure » et indiquant qu'il fallait comprendre que le dépassement ne devait pas excéder 150 % du tarif remboursable. À l'époque, ces chiffres n'étaient pas connus ; la possession des données a permis de faire des analyses et de la visualisation et de produire des articles et des pages sur Internet. Ce genre d'investigations permet la sortie de nouvelles informations et la mise en lumière de phénomènes intéressants.

Le data journalisme peut aussi contribuer à de salubres démystifications. Ainsi, le « fact checking » ambitionne de vérifier les informations ou affirmations factuelles lancées dans l'espace public, notamment par le personnel politique. C'est ainsi par exemple qu'en matière de délinquance on peut entendre ou lire n'importe quoi autour des chiffres disponibles. Il est alors utile de rappeler que lorsque la collecte des informations est faite par la police ou la gendarmerie, des biais sont possibles. Ou que les comparaisons dans le temps appellent des précautions compte tenu de l'évolution des nomenclatures ou des modes de collecte. C'est l'occasion de faire de la pédagogie auprès des lecteurs.

Il faut avoir en tête aussi qu'aux yeux des lecteurs l'image l'emporte toujours sur les mots. D'où la place croissante donnée à l'infographie. Après la récente catastrophe survenue à l'avion de la GermanWings, beaucoup de médias ont utilisé des graphiques pour expliquer le déroulement des derniers instants de vol. Il existe une base de données suisse qui répertorie chaque accident depuis les débuts de l'aviation commerciale. Grâce à cette base, on a pu illustrer l'évolution du nombre de morts par accident d'avion depuis 1914. Certes le nombre d'accidents a été élevé en 2014 mais c'est la baisse du nombre de morts qui domine.

De plus en plus de bases de données sont accessibles et donc de plus en plus d'informations disponibles. Mais ces informations sont-elles pertinentes ? Le journaliste doit se poser la question. D'autant plus que publier des chiffres et des graphiques, ça fait sérieux, ça fait bien et que du même coup naît le risque d'une insuffisante rigueur dans l'information donnée au lecteur, qui ne doit en aucun cas être trompé. La fiabilité de la source de l'information doit par conséquent être vérifiée et, chaque fois que possible, l'information doit être recoupée avec d'autres. Dans le cas de l'open data, on peut supposer que les institutions qui donnent accès à leurs données se sont assurées qu'elles ne contiennent rien d'inexact ni rien de dérangeant. Il n'en est pas de même pour les sources auxquelles le journaliste a accès sans l'accord de leur propriétaire et cela doit amener à beaucoup de prudence dans l'utilisation des informations qu'elles contiennent. Il existe désormais des noms

[8] NDR : Un script sert principalement à lancer et coordonner l'exécution de programmes.

célèbres illustrant l'accès non souhaité par leurs propriétaires à de grands ensembles de données et leur exploitation journalistique : qui ignore l'existence de WikiLeaks et de son porte-parole Julian Assange ? ou de Edward Snowden, qui a révélé le contenu de plusieurs programmes de surveillance de masse de la National Security Agency aux Etats-Unis ? ou encore de Hervé Falciani, qui a livré aux autorités fiscales françaises les archives numérisées de la banque suisse HSBC de fin 2006 à début 2007 ? Le journal Le Monde a pu avoir accès ensuite à un extrait de ces archives, retravaillées par les services des impôts. Dans ce dernier cas, il fallait être capable d'exploiter les bases de données récupérées, représentant 120 000 lignes… Le fisc français ne s'intéressait, pour sa part, qu'aux 3 000 lignes exploitables (sur 6 000) concernant des Français. L'ampleur de la tâche a conduit à une collaboration journalistique internationale sous l'égide du Consortium international de journalistes d'investigation (ICIJ). Le détail de cette investigation est retracé dans le document auquel donne accès le lien mentionné en annexe.

Enfin, le traitement visuel de l'information doit éviter bien des écueils. Un exemple est donné par un lien en annexe : si on veut illustrer par des cercles l'importance relative de deux grandeurs, le rapport de ces deux grandeurs doit s'appliquer à l'aire des cercles correspondants et non à leur diamètre. La pédagogie n'est pas qu'à usage externe, elle doit s'exercer aussi au sein des médias.

Débat :

L'animatrice, au moment d'ouvrir le débat, signale la présence parmi les participants de membres de l'association Pénombre[9]. Elle mentionne que selon Pénombre la bonne infographie n'est pas l'apanage de tous les journalistes ; certains pensent qu'il est plus important d'attirer l'œil du lecteur que de se montrer rigoureux dans la représentation graphique…

1 – Data journalisme et déontologie

Quel est le statut du journaliste ? Aux yeux de l'invité, le journaliste est théoriquement le professionnel qui dispose d'une carte de presse, c'est-à-dire qu'il est employé à titre principal par une entreprise de presse et qu'il tire l'essentiel de ses revenus d'une entreprise de presse. Mais il existe aussi des journalistes occasionnels. Un blogueur est-il un journaliste ? Il est difficile de se prononcer sur des cas limites. De plus en plus de gens publient sur les réseaux sociaux. Si quelqu'un prend des photographies ou un film d'un accident et les diffuse avec des commentaires, ce quelqu'un fait assurément acte de journalisme mais pas acte de journaliste.

Quelle que soit la démarche du data journaliste, qu'il exploite des bases de données disponibles en lien avec un sujet déterminé ou qu'il fabrique lui-même ces bases de données, il doit évidemment faire preuve d'esprit critique sur la fiabilité de ces données. La base du journalisme est de recouper les sources et d'en contrôler la fiabilité. L'invité donne un exemple d'une base de données constituée par recoupement de multiples informations : il a pu établir une série, commençant en 2005, du nombre des décès d'immigrés en route vers l'Europe.

Qu'en est-il en ce qui concerne l'identification des personnes, s'enquiert un participant ? Le principe est que le journaliste ne rediffuse pas les données personnelles, à moins qu'il ne s'agisse de

[9] L'association Pénombre offre un espace de réflexions et d'échanges sur l'usage du nombre dans le débat public. L'attention se porte sur la qualité des informations chiffrées et les enjeux de l'usage qui en est fait. Pénombre cherche à relier les questions de méthode et de présentation, le pain quotidien des producteurs de données, avec les enjeux politiques et sociaux du recours à l'information chiffrée, qui concernent les « utilisateurs » de chiffres. http://www.penombre.org/

personnalités politiques ou de grande notoriété prises en flagrant délit de contradiction entre leurs déclarations publiques et ce qui était constaté dans la base de données. On en a vu quelques exemples à propos de comptes détenus chez HSBC. Mais ce qui animait les data journalistes dans cette affaire, en tout cas ceux du Monde, c'était de mettre en évidence l'existence d'un système de fraude fiscale monté de toutes pièces par cette banque. Ils y sont parvenus en traitant des informations individuelles, en faisant des rapprochements de fichiers pour rattacher plusieurs comptes à une même personne, en utilisant le fichier des personnes recherchées, en exploitant les comptes rendus d'entretien des cadres de la banque avec leurs clients, etc. Bien sûr, les données individuelles remises par H. Falciani à l'administration fiscale étaient couvertes par le secret fiscal mais, Le Monde ayant pu y avoir accès, il a considéré qu'il était en droit d'exploiter le fichier qu'il détenait, y compris en ce qui concerne les données individuelles.

Un autre exemple d'utilisation de données personnelles a été la détermination des primes au mérite des policiers : ce travail a été fait en 2013 et a permis de répondre par des statistiques anonymes à des questions relatives à la répartition des primes selon les activités des policiers (travail de bureau ou sur le terrain). L'avantage de ne délivrer aucune donnée personnelle est qu'un travail interactif devient possible avec le lecteur.

Soit, demandent des participants, mais quelle est l'habilitation juridique des journalistes à conserver des données personnelles ? Et, en matière de déontologie, y a-t-il une organisation collective des journalistes ou bien s'agit-il pour chacun de veiller à un bon comportement personnel ? Y a-t-il des contrôles ? L'invité reconnaît ne pas maîtriser la matière juridique ; il voit les questions relatives au droit de la presse avec l'avocat de son journal. Sur le plan technique, les ordinateurs sont sécurisés, aucune information n'est déposée dans le « cloud ». Toutes ces questions de déontologie sont intéressantes mais il faut bien voir qu'elles sont traitées en fonction des cultures locales. Lors d'une conférence aux États-Unis, l'invité a entendu un avocat expliquer que si des données individuelles concernant des Européens venaient en possession de journalistes, ils ne devaient faire autre chose que les jeter car inutilisables à ses yeux compte tenu des lois de l'Union européenne. On en est aux balbutiements dans tout cela.

Quand les e-mails de salariés de Sony ont été récupéré par des « hackers » et mis en ligne, les comportements des journalistes ont été très variés, depuis la non utilisation totale par principe jusqu'à une utilisation plus politique ou « people » ; il n'y a pas de philosophie partagée, ni de comités de déontologie. Et en effet il peut y avoir des risques de dérapages.

Dans l'affaire Swiss Leaks, observe un participant, il a bien fallu que les dizaines de journalistes qui ont travaillé de manière coordonnée dans plus de quarante pays partagent quelques règles déontologiques ? Apparemment, cela s'est fait au fil du travail. La déontologie n'était pas formalisée mais elle s'est élaborée lors des échanges entre les rédactions. Après avoir préparé la base de données, le consortium a pris contact avec des journalistes jugés fiables dans différents pays, par exemple ceux du Guardian. Le Monde avait posé comme seules conditions qu'aucune accessibilité publique à ce fichier ne soit consentie et qu'il soit systématiquement mentionné comme source de l'information.

Un autre participant évoque la mode consistant à établir des palmarès. L'hebdomadaire Le Point établit ainsi celui des hôpitaux à partir des données du PMSI[10] qu'il a pu récupérer. Or, l'association

[10] Programme médicalisé des systèmes d'information. C'est un dispositif faisant partie de la réforme du système de santé français et ayant pour but la réduction des inégalités de ressources entre les établissements de santé (ordonnance du 24 avril 1996 sur la réforme de l'hospitalisation). Afin de mesurer l'activité et les ressources des établissements, il est nécessaire de disposer d'informations quantifiées et standardisées, qui alimentent le PMSI.

Pénombre a regardé les choses de près et a soulevé beaucoup de questions au sujet de ce palmarès. Le souci déontologique et méthodologique affiché par certains data journalistes ira-t-il jusqu'à les conduire à critiquer leurs confrères ? L'invité pense qu'il vaut mieux que la critique émane de Pénombre que de confrères, qui pourraient être soupçonnés de jalousie ou de concurrence déloyale. Il est toujours possible aussi de s'adresser des courriers entre confrères. Un contre-exemple toutefois : alors qu'Éric Zemmour est présenté comme journaliste, certaines de ses affirmations ont été dénoncées comme des approximations ou des contrevérités par d'autres journalistes.

2 – Enjeux de société

Un participant se demande pourquoi la télévision est tellement en avance sur la presse écrite en ce qui concerne l'infographie. Il se réfère à l'émission quotidienne « 28 minutes » sur Arte, qui inclut quelques minutes de vérification des dires des politiques, ou encore à l'émission « Le dessous des cartes », émission de douze minutes avec des visuels remarquables à son avis. L'invité fait observer en premier lieu que la vérification par les faits dans l'émission « 28 minutes » est l'adaptation de quelque chose qui existe dans Libération depuis cinq ans. S'agissant de « Le dessous des cartes », il note que la presse écrite, contrairement à la télévision, n'est pas obligée de produire des images, ce qui la met en position défavorable. Il y a en outre chez les lecteurs des journaux un attachement à l'écrit plus prononcé que chez les téléspectateurs. Quoique désavantagé aussi en termes de belles images par rapport à la télévision, l'Internet permet de s'adapter aux lecteurs : ainsi, la revue « 60 millions de consommateurs » s'est inspirée sur son site de l'article du Monde sur les dépassements d'honoraires médicaux.

Un participant se déclare pleinement d'accord pour la critique des affirmations chiffrées. Mais d'où sort l'information ? Lui-même n'a jamais obtenu de réponse de la part des journaux ou de responsables politiques quand il posait la question. Comment éviter que des « chevaliers blancs » auto-proclamés ne lancent des contre informations erronées elles aussi ? Quelqu'un fait-il la police là-dessus ? Et puis, pratiquée systématiquement (et à la limite à chaud dans certaines émissions), la vérification par les faits ne risque-t-elle pas de tourner au gadget ou, pire, de laisser au lecteur ou au téléspectateur - ou plus gravement encore au citoyen - le sentiment qu'on ne peut se fier à personne ? D'autant plus, observe un autre participant, que l'information télévisuelle passe vite et laisse davantage d'impressions que de souvenirs précis.

La manière dont le data journaliste travaille est donc d'une grande importance pour la bonne santé de l'opinion publique. Concrètement, comment les choses se passent-elles ? Le data journaliste est-il un fouineur qui détecte des sujets à explorer et soumet l'idée à son rédacteur en chef ou bien est-ce ce dernier qui détermine les thèmes à creuser ? Les deux situations se rencontrent ; dans le cas des dépassements d'honoraires des médecins, l'initiative est venue du journaliste. Dans d'autres cas, c'est le rédacteur en chef qui donne des pistes de recherche en fonction de l'actualité, par exemple le décès de la doyenne de l'humanité ou l'ouverture du festival de Cannes (à propos duquel on a examiné le palmarès depuis 1946 avec l'espoir – déçu – d'en déduire la probabilité que tel ou tel film gagne !). Le data journaliste ne va pas toujours dans la direction souhaitée par son rédacteur en chef : l'invité s'est refusé pour sa part à produire un papier sur les évolutions comparées du nombre de chômeurs et du nombre de manifestants dans les mobilisations du monde du travail, au motif que l'existence d'une corrélation négative entre ces deux évolutions ne signait pas pour autant une relation de causalité entre elles.

Les journalistes étant constamment dans l'urgence, ce qui peut entraîner des erreurs, qu'en est-il des validations internes ? Quel est le rôle du rédacteur en chef ? Le comportement des rédacteurs en

chef, indique l'invité, est très variable. Certains vont très loin dans le détail des vérifications. Mais ce qui domine est le désir des data journalistes, de l'invité en particulier, de faire du bon travail, de la bonne pédagogie, et de délivrer une information contrôlée. De ce point de vue, les statistiques des institutions publiques sont réputées solides ; mais quand il y a eu collecte par d'autres organismes comme des organisations non-gouvernementales, les précautions sont plus grandes car il y a des risques de pièges. L'invité donne l'exemple d'une journaliste américaine qui disposait d'un fichier des agressions sexuelles dans un pays d'Afrique. Elle a établi une carte à partir de ces données et l'a publiée. Il en est résulté force protestations pour dénoncer les nombreux biais dont la collecte avait été affectée. Cette journaliste a repris entièrement son article à la lumière des informations complémentaires ainsi recueillies.

3 – Data journalistes et statisticiens

Un participant est frappé par le fait que l'invité n'a pas prononcé le mot statistique ou le mot statisticien, alors que tout son travail est du travail de statisticien : collecte de données, traitement et exploitation de ces données puis commentaires interprétatifs. Quelle relation y a-t-il entre les journalistes et les statisticiens ? Les organes de presse ont-ils des statisticiens dans leurs équipes ? A contrario, l'infographie de la presse fait-elle sentir ses effets sur les pratiques des statisticiens publics ?

L'invité ne se sent pas statisticien et donc évite de prononcer le mot. Il pense qu'il serait intéressant d'avoir dans les organes de presse davantage de gens férus en statistique. Selon les sujets il faut vérifier beaucoup de choses pour prétendre faire un travail de statisticien. Il donne l'exemple des arrêts cardiaques constatés à Paris : si un fichier des arrêts cardiaques était disponible en open data, le data journaliste aurait positionné les événements sur une carte, aurait repéré la proximité de ces décès avec les gares et s'en serait tenu là. L'Inserm[11], lui, a pensé qu'il y avait là un vrai sujet d'étude.

Il n'y a pas de statistique au programme des écoles de journalisme (ni d'ailleurs de cours de journalisme au programme des écoles de statisticiens !). Pourtant, le rapprochement entre les deux disciplines serait fructueux. Au demeurant, ce rapprochement existe, indiquent plusieurs participants : alors qu'il y a quarante ans les journalistes faisaient faire leurs calculs par les statisticiens (exemple du Nouvel Observateur avec le prix d'un Français), ils les font maintenant eux-mêmes. De leur côté, les statisticiens publics se rapprochent du monde des médias : ainsi de l'Observatoire français des drogues et des toxicomanies, qui produit un rapport très travaillé par son service de presse, ou de l'Institut national d'études démographiques (Ined), qui en la personne de Michel-Louis Lévy a mis au point un quatre-pages mensuel (« Population et Sociétés ») d'excellente vulgarisation scientifique, ou encore de l'Insee qui a détaché quelques temps certains de ses agents au Monde et à Alternatives économiques.

Pour sa part, l'invité fait état de travaux menés en commun avec des statisticiens : il a pu lui-même travailler pendant une demi-journée avec un doctorant en sciences politiques versé en statistiques pour dresser des cartes du vote Front national dans trois départements et il a pu observer à cette occasion que les chercheurs ne sont pas dans la même temporalité que les journalistes ! Allant plus loin, il se demande pourquoi il n'y aurait pas de statisticiens journalistes ? Il faudrait à ses yeux

[11] Institut national de la santé et de la recherche médicale. Créé en 1964, l'Inserm est un établissement public à caractère scientifique et technologique, placé sous la double tutelle du ministère de la santé et du ministère de la recherche. Seul organisme public de recherche français entièrement dédié à la santé humaine, il s'est vu confier, en 2008, la responsabilité d'assurer la coordination stratégique, scientifique et opérationnelle de la recherche biomédicale.

développer dans les rédactions une compétence comme celle qui a permis aux USA de mettre au point des modèles de probabilité sur les sports puis sur les élections[12]. Malheureusement, dans les écoles de journalisme, l'appétit pour le data journalisme – qui fait largement appel à des techniques statistiques - n'est pas encore marqué. Il faut sans doute y voir un effet de l'orientation scolaire qui se fait en France sur le critère des mathématiques et opère dès la seconde le tri entre les littéraires et les scientifiques. De fait, les écoles de journalisme accueillent pour l'essentiel des rebelles à la chose mathématique. Dans ses fonctions d'enseignant, l'invité a du mal à faire utiliser les fonctionnalités des fichiers Excel par ses élèves. Une simple capture d'écran peut être pour eux une difficulté. Quoique ne se sentant pas lui-même un spécialiste des chiffres, l'invité est parfois considéré comme tel dans son univers professionnel.

Bien que l'invité ne soit pas statisticien de formation, il faut se féliciter de deux choses, estime un participant : d'une part, la rapidité enviable avec laquelle le data journaliste qu'il est produit de l'information et, d'autre part, la nature de ses investigations : ainsi, qu'il s'agisse des dépassements d'honoraires des médecins ou des primes de certains fonctionnaires, il produit une information que jamais la statistique publique n'est parvenue à publier !

4 – Un métier qui émerge

Les échanges de la soirée permettent d'esquisser une typologie des actions de data journalisme :

- le journalisme d'investigation, qui repose sur l'exploitation de bases de données préexistantes (accessibles en open data ou obtenues par divers canaux) ou élaborées par le journaliste lui-même. Les exemples cités du dépassement des honoraires médicaux, de l'affaire HSBC ou des décès d'immigrants dans leur cheminement vers l'Europe appartiennent à cette catégorie. Même pour les sources facilement accessibles, le data journaliste rencontre des difficultés car il veut toujours aller plus loin : ce sont les données brutes qui l'intéressent, parmi lesquelles il cherchera à détecter les situations extrêmes. Par exemple, Pôle emploi diffuse le nombre moyen de chômeurs par agent mais les journalistes ont voulu avoir accès à l'intégralité des données pour examiner la situation par agence. Il leur a fallu saisir pour cela la Commission d'accès aux documents administratifs (Cada)[13] ; le ministère a fini par céder mais en diffusant à tous les journalistes un fichier PDF... L'exploitation de ce fichier a néanmoins été possible et les journalistes ont pu établir que les zones les plus touchées par le chômage (notamment dans le Nord) manquaient de conseillers. Dans le même esprit, le data journaliste souhaite avoir les données brutes des fichiers de l'Education nationale – quel que soit par ailleurs l'intérêt des analyses que ce dernier publie – afin de débusquer par exemple les classes comptant trop d'élèves.

D'une manière générale, le data journaliste cherche des données lisibles et surtout qui collent à l'actualité. Ce n'est pas toujours palpitant : à l'annonce d'Hillary Clinton se déclarant candidate aux élections présidentielles aux États-Unis 576 jours avant l'élection[14], le data journaliste reprendra

---

[12] Allusion au travail de Nate Silver avant les dernières élections présidentielles américaines : voir le blog http://fivethirtyeight.com/

[13] La Commission d'accès aux documents administratifs est une autorité administrative indépendante et consultative chargée de veiller à la liberté d'accès aux documents administratifs. Son rôle est principalement de rendre des avis sur le refus opposé par l'administration aux demandes de communication des particuliers, des entreprises ou des associations. Sa saisine est obligatoire avant tout recours contentieux. Elle conseille les administrations sur le caractère communicable de document et peut être consultée par le gouvernement ou proposer des modifications sur des textes législatifs ou réglementaires. Elle informe le public sur le droit d'accès.

[14] Cela a pu apparaître beaucoup mais Barack Obama s'était déclaré plus tôt encore.

toutes les candidatures selon leur date avant celle des élections mais pour s'apercevoir qu'il n'y a pas de lien entre cette durée et le résultat de l'élection.

Dans ces investigations, note un participant, il est rare que les journalistes exploitent véritablement les statistiques publiques disponibles ; la plupart reprennent telles quelles les publications officielles plutôt que de retravailler les données. L'invité ne partage pas ce point de vue : tout un travail est fait sur les données disponibles, par exemple sur les indicateurs concernant les lycées, de manière à aider les parents à trouver le meilleur lycée autour de chez eux ;

- le fact checking, par lequel le data journaliste va chercher la vérification d'un chiffre avancé comme argument dans le débat. À Libération, on fait ainsi de la désintoxication avec le but de pousser les personnages politiques à cesser de dire n'importe quoi. La fréquence des vérifications utiles ne signifie pas que tous les politiques seraient des gens peu sérieux, mais elle inspire l'idée que la communication politique peut pourrir le débat ; c'est manifeste sur des questions comme la délinquance ;

- la pédagogie, consistant à aller plus loin que la simple vérification des faits et, comme le font les « décodeurs » du Monde, à approfondir les recherches pour mieux éclairer le lecteur ou à pratiquer une infographie de qualité. L'invité évoque l'exemple d'un Insee Première (dont l'embargo laissait un peu de temps au journaliste pour travailler le sujet) relatif aux entreprises sous contrôle de l'État[15]. Le fichier Excel accompagnant la publication fournissait les données brutes ; donc, au lieu de pratiquer une simple capture d'écran, l'invité a pu refaire un graphique et l'« éditorialiser », c'est-à-dire l'affecter d'un titre qui parle aux lecteurs et ne se contente pas d'annoncer le seul contenu du graphique.

Bien sûr, il y a des recouvrements entre ces trois catégories d'actions et le data journalisme est un métier en construction qui se dotera peu à peu de ses propres règles. Peu de journalistes l'exercent pour le moment en France : une dizaine au sens strict ; 50 à 75 au maximum si l'on inclut les journalistes qui font des classements d'hôpitaux ou de maisons de retraite. A l'avenir, un minimum de compétences dans le traitement des fichiers informatiques et en statistique sera requis.

Que penser de la rentabilité du data journalisme ? Les coûts engagés sont-ils justifiés par les résultats obtenus ? demande une participante. Cette question pose le problème du retour sur investissement. Globalement, le data journalisme donne surtout une image innovante aux médias qui le pratiquent. Si on veut déboucher sur une information très précise, le data journalisme demande plus de travail que le journalisme classique. L'invité cite l'exemple d'une investigation approfondie sur les emprunts toxiques : la carte de leur répartition sur le territoire qui en a résulté a été fortement consultée par les lecteurs pour leur univers de proximité, par les journalistes de la presse quotidienne régionale qui ont publié de nombreux articles sur le sujet et par les maires ; parmi ces derniers, certains ont pu vérifier la nature de l'endettement de leur commune. Monétiser les visites des lecteurs sur les sites Internet sera-t-il demain le modèle économique dominant ?

Rebondissant sur les propos de l'invité qui laissent entendre que le soin (et donc le temps) consacré aux sujets jugés sérieux est plus important que pour les investigations plus banales, un participant voit là un risque que « la mauvaise statistique chasse la bonne », c'est-à-dire que le souci de rentabilité ne privilégie les exploitations rapides aux méthodes peu assurées par rapport aux exploitations statistiquement irréprochables mais forcément longues.

[15] Insee Première n° 1541 d'avril 2015.

Dans un avenir proche, deux projets de loi vont probablement changer le cadre de travail des data journalistes. Dans le domaine de la santé, des avancées intéressantes sont prévues avec l'accès aux bases de données SNIIRAM[16] et PMSI mais une disposition du projet de loi soulève une inquiétude, à savoir que tout projet d'étude sera soumis à validation par des experts. Le collège d'experts acceptera-t-il toujours de délivrer les données aux journalistes, quel que soit le sujet envisagé par ces derniers ?

Le projet de loi sur le renseignement est inquiétant en raison des boîtes noires qui pourront être installées chez les fournisseurs d'accès Internet (FAI). On nous affirme qu'il n'y aura pas de surveillance généralisée des navigations et des communications car seules les métadonnées seront collectées et non pas le contenu des échanges. C'est déjà beaucoup. Et puis, le but étant de cibler et d'identifier de potentiels terroristes par « datamining »[17] et par « machine learning »[18], est-il possible de faire des repérages de cette nature sans se référer à une population réputée sans danger ? L'inquiétude porte aussi sur ce qui pourra être consulté sur Internet ou sur la confidentialité des communications téléphoniques. Ou encore sur la nature des actions justifiant une surveillance. Les lanceurs d'alerte seront-ils amenés à s'autocensurer, au motif qu'on pourrait leur reprocher d'être une menace pour les intérêts économiques de la nation ? Un participant rappelle que, au début de la décennie 70, c'est un projet de loi relatif à la mise en place d'un répertoire national d'identification des personnes physiques qui a donné naissance à la loi sur l'informatique et les libertés. À l'époque, certains s'étaient demandé si ces grandes bases de données d'identification ne devraient pas être munies de moyens de sabordage, comme il existe paraît-il des chambres à mines dans les ponts, dans lesquelles on peut disposer des explosifs si une menace ennemie apparaît.

Pour aller plus loin, les liens indiqués ci-dessous orientent vers des documents de nature à compléter la réflexion sur le data journalisme :

1 – sur le travail concrètement réalisé dans l'affaire HSBC :
http://data.blog.lemonde.fr/2015/02/09/comment-nous-avons-travaille-avec-les-donnees-de-swissleaks/

2 – pour se lancer dans le data journalisme : http://jplusplus.github.io/guide-du-datajournalisme/

3 – pour méditer sur les pièges de l'infographie :

https://mobile.twitter.com/thedomstrat/status/517047382534418432

https://mobile.twitter.com/decodeurs/status/517228610051395584

[16] Système national d'information inter-régimes de l'assurance maladie. Créé en 1999 par la loi de financement de la sécurité sociale, le SNIIRAM est une base de données nationale dont les objectifs sont de contribuer à une meilleure gestion de l'assurance maladie et des politiques de santé, d'améliorer la qualité des soins et de transmettre aux professionnels de santé les informations pertinentes sur leur activité. Son périmètre, ses finalités, son alimentation et l'accès aux données sont définis dans un arrêté du ministère des affaires sociales et de la santé transmis pour avis à la Commission nationale de l'informatique et des libertés (CNIL) puis publié au Journal officiel.

[17] Le « data mining » ou fouille de données est un terme générique englobant différents outils facilitant l'exploration et l'analyse des données contenues dans une base contenant de grandes quantités d'informations.

[18] Wikipédia : L'apprentissage automatique (*machine learning* en anglais), un des champs d'étude de l'intelligence artificielle, est la discipline scientifique concernée par le développement, l'analyse et l'implémentation de méthodes automatisables qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques.

http://mobile.lemonde.fr/les-decodeurs/article/2014/06/30/mefiez-vous-des-cercles-proportionnels_4447953_4355770.html?xtref=acc_dir