GUIDE DU DATAJOURNALISME

Collecter, analyser et visualiser les données

Sous la direction de

Jonathan Gray
Liliana Bounegru
Lucy Chambers

et de

Nicolas Kayser-Bril
pour la version française

EYROLLES

© Groupe Eyrolles, 2013, pour l'édition en langue française
ISBN 978-2-212-13685-2

Sommaire

Préface **1**
Contributeurs 2
Ce qu'est ce livre (et ce qu'il n'est pas) 4

1 Introduction **7**
Qu'est-ce que le datajournalisme ? 8
Pourquoi les journalistes doivent utiliser des données 10
Pourquoi le datajournalisme est-il important ? 12
Quelques exemples du genre 19
Le datajournalisme en perspective 23

2 Dans la salle de rédaction **29**
Datajournalisme à la BBC 30
Comment fonctionne l'équipe des applications d'information du *Chicago Tribune* 33
Dans les coulisses du Guardian Datablog 35
Datajournalisme au Zeit Online 39
Comment recruter un hacker 43
Aller chercher les talents dans les hackathons 46
Suivre les flux financiers : datajournalisme et collaboration internationale 49
Nos histoires sont du code 52
Kaas & Mulvad : contenu semi-fini pour groupes d'influence 55
Créations d'applis à Rue89 59
Modèles économiques de datajournalisme 61

3 Études de cas **65**
Le fossé des opportunités 66
Une enquête de neuf mois sur les fonds structurels européens 68
Aspirer les données d'Ameli 71
Contrôler les dépenses publiques avec OpenSpending.org 72
Une pige de « scraping olympique » 76
Hack électoral en temps réel (Hacks/Hackers Buenos Aires) 79
Crowdsourcing : l'accès à la TNT dans le sud-est de la France 82
Le hackathon Mapa76 84

La couverture des émeutes au Royaume-Uni par le Guardian Datablog 87
Évaluer les écoles de l'Illinois 90
Contrôler les factures d'hôpitaux 92
Le Véritomètre 95
Le téléphone omniscient 97
Quel modèle de voiture ? Taux d'échec au contrôle technique 99
Le subventionnement des bus en Argentine 100
Comment Regards Citoyens a créé NosDéputés.fr, base de données de l'activité parlementaire 104
Le grand tableau des élections 107
Crowdsourcing du prix de l'eau 108

4 Obtenir des données 111
Guide de référence rapide 112
Votre droit d'accès aux données publiques 118
Le wobbing, ça marche ! 122
Recueillir des données sur le Web 127
Le Web comme source de données 133
Le crowdsourcing de données au Guardian Datablog 139
Utiliser et partager des données : la loi, les petits caractères et la réalité 141

5 Comprendre les données 145
Se former aux données en trois étapes simples 146
Notions de base pour travailler avec des données 151
Histoires de données 155
Les datajournalistes parlent de leurs outils préférés 157
Utiliser la visualisation pour faire parler les données 164

6 Publier des données 175
Présenter des données au public 176
Concevoir une application d'information 180
Applications d'actualité chez ProPublica 183
La visualisation, meilleur outil du datajournaliste 185
Utiliser la visualisation des données pour raconter des histoires 191
Différents graphiques pour différents angles 200
Visualisation de données maison : nos outils préférés 205
Comment nous publions nos données au *Verdens Gang* 211
Données publiques sur les réseaux sociaux 214
Impliquer les gens autour de ses données 217

À propos des directeurs d'ouvrage 220
À propos des coordinateurs du projet 220

La couverture des émeutes au Royaume-Uni par le Guardian Datablog

Au cours de l'été 2011, l'Angleterre a été touchée par une vague d'émeutes. À l'époque, des politiciens avaient prétendu que ces actions n'avaient aucun lien avec le taux de pauvreté et que les pilleurs étaient de vulgaires criminels. De plus, le Premier Ministre et les principaux politiciens conservateurs avaient accusé les réseaux sociaux d'avoir provoqué les émeutes, suggérant que ces plates-formes avaient attisé la violence et que les émeutiers s'étaient organisés *via* Facebook, Twitter et Blackberry Messenger (BBM). Certains ont appelé à bloquer temporairement l'accès aux réseaux sociaux. Comme le gouvernement n'a pas ouvert d'enquête sur les raisons des émeutes, *The Guardian*, en collaboration avec la London School of Economics, a réalisé un projet révolutionnaire intitulé « Reading the Riots » (« Lire les émeutes ») pour répondre à ces questions.

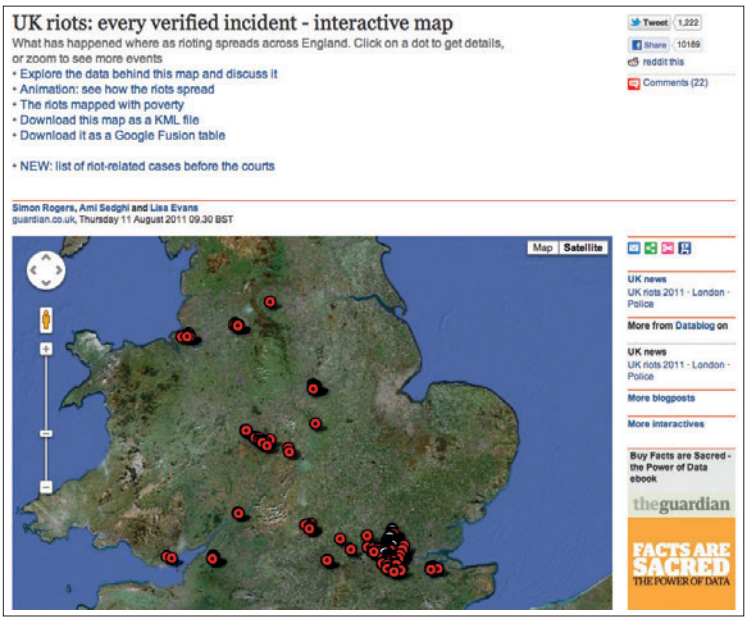

Figure 3-9. *Les émeutes au Royaume-Uni : tous les incidents vérifiés* (The Guardian)

Le journal a ainsi fait la part belle au datajournalisme pour permettre au public de mieux comprendre qui avait participé aux émeutes et pourquoi. Par ailleurs, le journal a collaboré avec une autre équipe d'universitaires, dirigée par le professeur Rob Procter de l'université de Manchester, pour mieux comprendre le rôle des réseaux sociaux, que *The Guardian* lui-même avait largement utilisés pour couvrir les émeutes. L'équipe de Reading the Riots était dirigée par Paul Lewis, le rédacteur des projets spéciaux du *Guardian*.

Au cours des émeutes, Paul avait rapporté des nouvelles du front dans plusieurs villes d'Angleterre (notamment par le biais de son compte Twitter, @paullewis). Cette seconde équipe a travaillé sur 2,6 millions de tweets concernant les émeutes, fournis par Twitter. Le principal objet de ce travail sur les réseaux sociaux était de comprendre comment les rumeurs circulaient sur Twitter, quelle fonction jouaient les différents utilisateurs/acteurs dans la propagation des informations, dans quelle mesure la plate-forme avait été utilisée pour inciter à la violence, et d'observer de nouvelles formes d'organisation.
En matière de datajournalisme et de visualisation de données, il est utile de distinguer deux périodes-clés : la période des émeutes à proprement parler et la façon dont les données ont permis de raconter leur déroulement en temps réel ; puis une période d'étude plus intense avec deux équipes universitaires collaborant avec *The Guardian* pour recueillir des données, les analyser et produire des rapports détaillés. Les résultats de la première phase du projet Reading the Riots ont été publiés au cours d'une semaine de couverture intense au début du mois de décembre 2011. Ci-après figurent quelques exemples-clés de l'utilisation du datajournalisme au cours de ces deux périodes.

Phase 1 : les émeutes en temps réel

Avec des cartes simples, l'équipe de datajournalisme du *Guardian* a cartographié l'emplacement des émeutes confirmées, puis en recoupant des données sur le taux de pauvreté avec la localisation des émeutes, elle a commencé à déconstruire le mythe politique qui voulait que les émeutes n'aient aucun lien avec la pauvreté. Ces deux exemples utilisent des outils de cartographie prêts à l'emploi, et le second combine des données géographiques avec une autre base de données afin d'établir des recoupements.
Concernant l'utilisation des réseaux sociaux (en l'occurrence, Twitter) au cours des émeutes, le journal a créé une visualisation des hashtags liés aux émeutes publiés au cours de cette période, qui soulignait que Twitter avait principalement été utilisé pour répondre aux émeutes plutôt que pour planifier des pillages, le hashtag #riotcleanup (une campagne spontanée visant à nettoyer les rues après les émeutes) présentant le plus fort pic de trafic au cours de cette période.

Phase 2 : décrypter les émeutes

Quand le journal a rapporté les résultats de plusieurs mois d'enquête intensive et de collaboration étroite avec deux équipes universitaires, deux visualisations sont ressorties du lot et ont beaucoup fait parler d'elles. La première est une courte vidéo présentant le recoupement des endroits où des émeutes se sont produites avec les adresses des émeutiers, ainsi qu'une sorte d'itinéraire des émeutes. Pour ce faire, le journal a travaillé avec un spécialiste de la cartographie des transports, ITO World, afin de modéliser l'itinéraire le

plus vraisemblablement emprunté par les émeutiers pour aller commettre leurs pillages, établissant des parcours différents selon les villes, avec parfois de longues distances parcourues.

La seconde visualisation s'intéresse à la propagation des rumeurs sur Twitter. En accord avec l'équipe universitaire, sept rumeurs ont été sélectionnées pour analyse. L'équipe universitaire a ensuite recueilli toutes les données liées à chaque rumeur et a conçu un système de codage pour classer les tweets selon quatre critères : les gens qui répétaient simplement la rumeur (affirmation), qui la réfutaient (négation), qui la remettaient en question (question) ou qui la commentaient (commentaire). Toutes les tweets ont été codés en trois exemplaires et les résultats ont été analysés par l'équipe interactive du *Guardian*. L'équipe du *Guardian* a détaillé le développement de cette visualisation sur son site.

Ce qui est si frappant dans cette visualisation, c'est qu'elle montre de façon percutante quelque chose de très difficile à décrire, à savoir la nature virale des rumeurs et leur cycle de vie. Le rôle des médias dominants est manifeste dans certaines de ces rumeurs (par exemple en les réfutant ou en les confirmant rapidement), de même que la nature corrective de Twitter lui-même. Non seulement cette visualisation permettait d'enrichir le storytelling, mais elle donnait une bonne idée de la façon dont les rumeurs circulaient sur Twitter, apportant des informations utiles pour gérer de futurs évènements.

Ce qui apparaît clairement dans ce dernier exemple, c'est la puissante synergie entre le journal et l'équipe universitaire qui a permis d'analyser 2,6 millions de tweets de manière détaillée. Bien que l'équipe universitaire ait développé des outils sur mesure pour cette analyse, elle travaille maintenant à les rendre plus largement accessibles à quiconque souhaiterait les utiliser pour ses propres analyses. Combinés au mode d'emploi fourni par *The Guardian*, ces outils constitueront une étude de cas utile démontrant comment de telles analyses et visualisations des réseaux sociaux peuvent être utilisées pour raconter des histoires.

Farida Vis, université de Leicester

Le design d'informations au service du datajournalisme

Créée par une journaliste et un graphiste, l'agence WeDoData aide les rédactions à produire de nouvelles narrations autour de la data. Récit des coulisses d'un projet réalisé en 2013 pour France Télévisions à l'occasion de la journée de la femme…

Nous avions imaginé, avec le service Nouvelles Écritures de la chaîne publique, le Pariteur[1], une application permettant à chacun de découvrir la différence de salaire existant avec quelqu'un du sexe opposé effectuant le même métier, au même âge et dans la même région.

Objectif : sortir de la classique statistique rappelée chaque année, « les Françaises gagnent en moyenne 20 % de moins que les hommes ». Mais pour cela, il fallait pouvoir accéder à la base complète des salaires français. Une seule institution la possède : l'Insee, à travers les DADS (Déclaration annuelle de données sociales), document rempli chaque année par les entreprises privées, ainsi que par les administrations et les établissements publics, indiquant pour chacun de leurs salariés sa catégorie socioprofessionnelle et le montant des salaires perçus.

Une fois l'Insee convaincu de la pertinence de cette démarche, restait à imaginer une interactivité qui fasse oublier à l'internaute qu'il allait plonger dans une base de données qui, sur tableur, représentait plus de 300 000 lignes.

Inspirés par l'application du Slavery Footprint, nous avons imaginé un questionnaire par étape, simple et très graphique, permettant à chacun de présenter son profil de salarié sans avoir l'impression de compléter un formulaire administratif. Le fil d'Ariane en haut de l'appli est l'un des points forts de cette « user experience » : fonctionnant comme un rébus graphique, il permet à chacun de modifier un paramètre en un simple clic et sert également de synthèse de la recherche en cours.

Dans l'esprit du livre *Paris vs New York* (de Vahram Muratyan, éditions 10/18), le graphisme avait un rôle fédérateur dans ce projet visant un très grand public, surfeurs du Net ou pas, donc plus ou moins adeptes des applications interactives. À base de pictos et d'une interface épurée, il permettait de plonger dans une narration sur les inégalités *a priori* anxiogène, mais que nous avons tenté de rendre ludique et revendicative.

Karen Bastien, WeDoData

1 : http://appli-parite.nouvelles-ecritures.francetv.fr

Évaluer les écoles de l'Illinois

Chaque année, la Commission de l'éducation de l'État de l'Illinois publie des rapports d'évaluation de ses écoles, à savoir des données sur le profil démographique et les performances de toutes les écoles publiques de l'Illinois. C'est une base de données massive – cette année, la publication faisait 9 500 *colonnes* de large. Le problème quand on dispose d'autant de données, c'est de choisir ce que l'on veut présenter. Comme pour n'importe quel projet de logiciel, le plus dur n'est pas de *développer* le logiciel, mais de développer le *bon* logiciel.

Nous avons travaillé avec les journalistes et les rédacteurs préposés à l'éducation pour sélectionner les données significatives. Beaucoup de données peuvent sembler intéressantes, mais un journaliste vous dira qu'elles sont en fait trompeuses ou erronées.
Nous avons également enquêté auprès des personnes ayant des enfants en âge d'aller à l'école dans notre salle de rédaction. Au passage, nous en avons appris beaucoup sur nos utilisateurs et l'ergonomie de la version précédente de notre site.

Figure 3-10. *Rapports d'évaluation 2011 des écoles de l'Illinois* (The Chicago Tribune)

Nous voulions répondre à deux types d'utilisateurs et de cas d'utilisation différents :

- les parents souhaitant connaître les performances de l'école de leurs enfants ;
- les parents souhaitant déménager, comme la qualité des écoles environnantes a souvent une grande influence sur cette décision.

La première version du site a demandé environ six semaines de travail à deux développeurs. Notre mise à jour de 2011 n'a demandé que quatre semaines. (Il y avait de fait trois personnes qui travaillaient sur la dernière itération, mais aucune à plein temps.)
Un élément-clé de ce projet était le graphisme de l'information. Nous présentons beaucoup moins de données que ce qui est disponible, mais cela représente tout de même *beaucoup* de données, et c'était un véritable défi que de les rendre digestes. Par chance, nous avons pu nous adjoindre les services d'un collègue du service de graphisme – un designer spécialisé dans la présentation d'informations complexes. Il nous a beaucoup appris sur la conception de graphiques et plus généralement, nous a guidés pour concevoir

une présentation lisible sans sous-estimer la capacité ou le désir du lecteur de comprendre les chiffres.

Le site a été conçu en Python et Django. Les données sont stockées dans MongoDB – les données des écoles sont hétérogènes et hiérarchiques, ce qui convient mal à une base de données relationnelle, autrement, nous aurions probablement utilisé PostegreSQL.

Sur ce projet, nous avons expérimenté pour la première fois le framework d'interface utilisateur Bootstrap de Twitter, et nous étions satisfaits des résultats. Les graphiques sont affichés avec Flot.

L'application héberge également les nombreux articles que nous avons écrits sur les performances des écoles. En ce sens, elle agit comme une sorte de portail ; quand un nouvel article sort sur le sujet, nous le plaçons au sommet de l'application, accompagné d'une liste des écoles concernées par l'histoire. Et quand un nouvel article paraît, les lecteurs de www.chicagotribune.com sont redirigés vers l'application, pas vers l'article.

D'après les premiers retours, nos lecteurs adorent l'application. Le feedback que nous avons reçu a été largement positif (ou au moins constructif !), et le nombre de visites a explosé. En prime, ces données conservent leur intérêt pendant toute l'année, alors même si nous nous attendons à voir les visites décroître à mesure que les articles disparaissent de la page d'accueil, notre expérience passée nous a prouvé que les lecteurs utilisaient l'application tout au long de l'année.

Voici quelques idées-clés que nous avons retenues de ce projet.

- Les graphistes sont vos amis. Ils savent comment rendre digestes des informations compliquées.
- Demandez de l'aide dans la salle de rédaction. C'est le deuxième projet pour lequel nous avons mené une enquête et des entretiens internes, et c'est une bonne manière d'obtenir l'avis de gens qui, comme notre public, viennent d'horizons différents et ne sont généralement pas très à l'aise avec les ordinateurs.
- Montrez votre travail ! Beaucoup de gens nous ont demandé à obtenir les données que l'application utilisait. Nous avons rendu une bonne partie des données accessible au public par le biais d'une API, et nous publierons bientôt tout ce que nous n'avons pas pensé à inclure initialement.

Brian Boyer, *The Chicago Tribune*

Contrôler les factures d'hôpitaux

Les journalistes d'investigation de *California Watch* ont été avertis de la possible existence d'une vaste escroquerie au programme fédéral Medicare, qui rembourse les frais médicaux des Américains de plus de 65 ans, au sein d'une grande chaîne d'hôpitaux californienne.

La fraude en question, baptisée *upcoding* (modification des codes de diagnostic), consistait à rapporter des affections plus graves qu'elles ne l'étaient en réalité pour obtenir un meilleur remboursement. Mais une source-clé dans l'affaire était un syndicat en lutte contre la direction de la chaîne d'hôpitaux, et l'équipe de *California Watch* savait qu'une vérification indépendante était nécessaire pour que l'histoire soit crédible.

Fort heureusement, le ministère de la Santé californien dispose d'archives publiques contenant des informations très détaillées sur tous les cas traités dans les hôpitaux de l'État. Les 128 variables comprennent jusqu'à 25 codes de diagnostic issus du manuel intitulé *Classification statistique internationale des maladies et des problèmes de santé connexes* (couramment appelé CIM-9), publié par l'Organisation mondiale de la santé. Les patients ne sont pas identifiés par leur nom, mais d'autres variables rapportent l'âge du patient, le mode de paiement de ses frais hospitaliers et le nom de l'hôpital qui l'a traité. Les reporters ont compris qu'avec ces données, ils pouvaient vérifier si les hôpitaux de la chaîne rapportaient certaines pathologies rares à des taux significativement plus élevés que ce que l'on retrouvait dans d'autres hôpitaux.

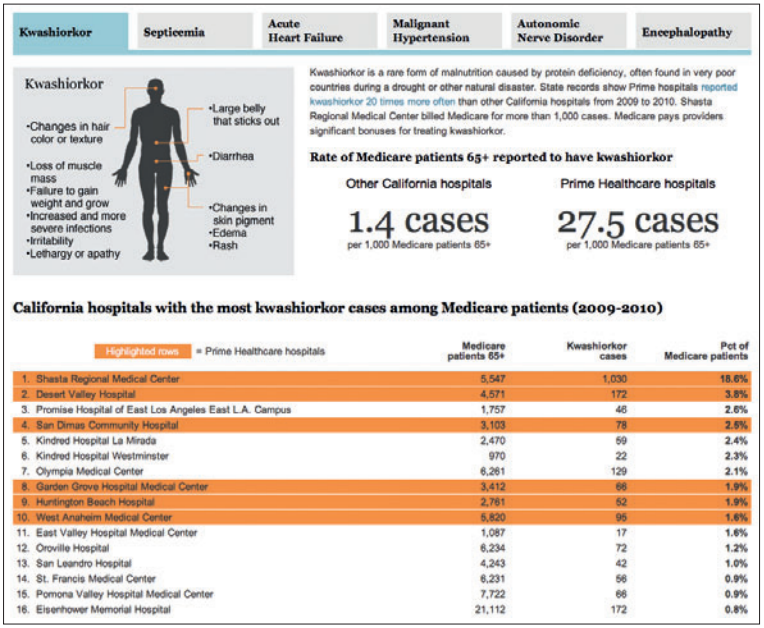

Figure 3-11. *Kwashiorkor* (California Watch)

Les bases de données étaient vastes : pratiquement quatre millions de dossiers par an. Les reporters voulaient étudier l'équivalent de six ans d'archives pour voir comment les tendances évoluaient. Ils ont obtenu les données auprès de l'agence étatique ; elles sont arrivées sur des CD-Rom qui ont facilement été copiés sur un ordinateur de bureau. Le

reporter en charge de l'analyse des données s'est servi d'un système appelé SAS. Cet outil très puissant permet d'analyser plusieurs millions de dossiers et est utilisé par de nombreuses agences gouvernementales, y compris le ministère de la Santé californien, mais il est coûteux – le même type d'analyse aurait pu être réalisé à l'aide de n'importe quel autre système de base de données, comme Access ou la suite open source MySQL.

Une fois les données obtenues et les programmes écrits pour les analyser, il était relativement facile de déterminer les tendances suspectes. Par exemple, une des allégations rapportait que la chaîne signalait divers degrés de malnutrition à des taux bien plus élevés que ce que l'on constatait dans d'autres hôpitaux. À l'aide de SAS, l'analyste des données a extrait des tableaux de fréquences présentant le nombre de cas de malnutrition rapportés chaque année par les 300 et quelques unités de soins intensifs de Californie. Les tableaux de fréquences bruts ont ensuite été importés dans Excel pour une inspection plus fine des tendances de chaque hôpital ; la capacité d'Excel à trier, filtrer et calculer des taux à partir de chiffres bruts a permis de faire ressortir facilement les tendances.

Un des exemples les plus frappants était le signalement d'une affection appelée kwashiorkor, un syndrome de déficience en protéines constaté presque exclusivement chez des nouveau-nés affamés dans des pays en voie de développement frappés par la famine. Pourtant, les hôpitaux de la chaîne diagnostiquaient près de 70 fois plus de cas de kwashiorkor chez des personnes âgées que la moyenne des hôpitaux de Californie.

Pour d'autres histoires, des techniques similaires ont été employées afin d'examiner les taux rapportés d'affections telles que les septicémies, les encéphalopathies, l'hypertension artérielle maligne et les atteintes du système nerveux autonome. Une autre analyse se penchait sur les allégations prétendant que la chaîne admettait dans ses urgences des pourcentages inhabituels de patients Medicare, dont la source de paiement des frais hospitaliers est plus sûre que pour bien d'autres patients.

Pour résumer, il nous a été possible de raconter ces histoires en utilisant des données pour vérifier indépendamment les allégations de sources susceptibles d'avoir des intentions inavouées. Ces histoires sont également un bon exemple de la nécessité d'avoir des réglementations strictes en matière d'archives publiques ; si le gouvernement demande aux hôpitaux de consigner ces données, c'est pour que ce type d'analyses puisse être réalisé, que ce soit par le gouvernement, des universitaires, des enquêteurs ou même des journalistes citoyens. Le sujet de ces histoires est important parce qu'il examine dans quelle mesure des millions de dollars d'argent public sont dépensés à bon escient.

Steve Doig, Walter Cronkite School of Journalism,
Arizona State University

Le Véritomètre

Nombre de chômeurs, taux de croissance, coût de la délinquance ou encore bénéficiaires du RSA : depuis ces dernières années, les chiffres ont envahi les discours des hommes politiques. Chacun donne son estimation, renforce son argumentaire à grand coup de statistiques « officielles ». Le terrain politique est devenu un champ de bataille de données. À l'approche de la présidentielle, il nous[6] a semblé utile, chez OWNI, de débroussailler ce champ de bataille et donner des clés de compréhension.

Le Véritomètre : vérifier et donner du sens

Une fois ce grand principe posé, nous nous sommes concentrés sur deux objectifs principaux.

1. Faire du factchecking (vérification de l'information), de manière ouverte. Nous voulions éclairer sur la « manipulation » des chiffres sans pour autant nous placer en juges, détenteurs de la vérité absolue des données, contrairement d'ailleurs à ce que le nom de l'application – Véritomètre, un choix finalement assez marketing – laissait supposer.
2. Permettre aux internautes de questionner eux-mêmes les discours des politiques et leur donner accès à des séries de données vérifiées sur les grandes thématiques de la présidentielle (économie, sécurité, éducation, etc.).

D'un point de vue éditorial, nous avons donc fixé rapidement un certain nombre de règles pour notre factchecking. Nous ne vérifiions que les données chiffrées ou chiffrables (par exemple « il y a trois fois plus de chômeurs en France qu'en Allemagne »), nous ne factcheckions pas le futur (comme les prévisions de croissance) et les vérifications se faisaient à partir de sources officielles. Nous nous sommes volontairement limités aux six principaux candidats de la présidentielle (Hollande, Sarkozy, Le Pen, Mélenchon, Joly et Bayrou) lors de leurs interventions médiatiques importantes.

Enfin, nous appliquions une marge d'erreur fixe : plus de 10 % de marge = incorrect (-1), entre 5 et 10 = imprécis (0), moins de 5 = correct (1). La note ainsi obtenue nous permettait de donner un classement des candidats, mis à jour en permanence.

6 : Le concept de l'application et ses règles ont été imaginés par 3 journalistes (Sylvain Lapoix, Nicolas Patte et Marie Coussin). Les vérifications étaient faites par cette équipe et 2 autres journalistes (Pierre Lebovici et Grégoire Normand). 2 développeurs (Tom Wersinger, James Lafa), 2 designers (Loguy, Marion Boucharlat) et une chef de projet (Anne-Lise Bouyer) ont travaillé sur la réalisation de l'application Véritomètre.

Un mot-clé : adaptabilité

L'application, pour répondre à ces objectifs, devait permettre de stocker un grand nombre de données, mais aussi de les mettre à jour facilement, d'en ajouter et de les visualiser : nous ne pouvions pas proposer au grand public une application basée sur un ensemble de tableaux. Nous avions les mêmes contraintes pour les pages « textes » contenant les interventions des politiques : autonomie, ajout et mise à jour facile, rendu graphique clair et ergonomique, etc.

Avec toutes ces contraintes, impossible de demander l'aide d'un développeur à chaque fois pour rentrer une donnée ou pour rentrer un discours : nous devions être autonomes, une fois l'application lancée. De plus, nous voulions aussi pouvoir réutiliser la base de données ainsi constituée, indépendamment du projet Véritomètre.

À cela s'ajoutaient les exigences propres à notre partenaire, I>Télé, qui souhaitait pouvoir montrer l'application à la télévision (notamment les graphiques, en plein écran). Un journaliste devait être présent en plateau et faire la démonstration des vérifications en direct, en naviguant dans l'application avec un iPad pour afficher les graphiques sélectionnés.

Les choix techniques

Une fois les fonctionnalités et contraintes établies, nous avons travaillé avec les développeurs et le directeur artistique d'OWNI pour imaginer comment y répondre.

Pour la structure de l'application, le choix s'est porté sur WordPress. Tant pour des raisons de souplesse que par automatisme chez OWNI. Nous aurions tout aussi bien pu utiliser un autre moteur de blog ou un site avec une interface d'administration propre et fonctionnelle.

Pour l'insertion des graphiques, nous avons choisi de les intégrer directement *via* une base MySQL et de créer des graphiques avec la librairie HighCharts. Ce sont les journalistes qui intégraient les données et sélectionnaient le type de graphique le plus pertinent. HighCharts offrait un rendu fluide et simple, tout en ayant un petit effet au survol des données : indispensable pour la télévision.

Entrer directement les données dans la base de données MySQL *via* l'interface PHP MyAdmin nous a demandé de révolutionner quelque peu notre approche de la donnée. Là où nous récupérions un tableau « propre » de l'Insee ou d'un organisme quelconque, nous devions le démanteler intégralement pour isoler la donnée, lui donner un identifiant et lui lier ensuite les autres informations (catégorie, source, année, etc.).

Ce travail de conception était passionnant car nous avions construit l'application de A à Z, en partant de nos usages et de nos capacités techniques. Nous avons réellement pris conscience du travail de datajournalisme en équipe : expliquer les contraintes journalistiques aux développeurs, qui cherchent des solutions techniques pour les résoudre.

Le rôle d'un chef de projet, bien que nous ayons tous déjà conçu des applications dans le pôle datajournalisme, s'est avéré également indispensable pour un projet de cette taille, pour gérer les allers-retours avec I>Télé et suivre toute la réalisation du projet.

Les enseignements

Tant la conception que l'animation de l'application pendant presque quatre mois ont été riches d'enseignements pour l'équipe et pour notre pratique du datajournalisme. Se plonger dans les données pendant tout ce temps nous a rappelé à quel point il s'agit d'un sujet d'études mouvant, lié lui aussi à des choix politiques ou à des temporalités. Il n'y a pas de « vérité » dans les données, elles doivent subir un traitement journalistique comme n'importe quelle source. C'est pourquoi lors du débat de l'entre-deux tours, nous n'avons publié « en direct » (sur Twitter) que des vérifications que nous avions déjà faites : pas question d'en faire de nouvelles sans avoir le temps d'étudier les sources et résultats. Heureusement pour nous, les politiques se répètent beaucoup.

Marie Coussin, AskMedia

Le téléphone omniscient

La plupart des gens n'ont qu'une vague idée de ce que l'on peut effectivement faire avec les données de leur téléphone portable ; il existe peu d'exemples concrets. C'est pour cette raison que Malte Spitz, du parti des Verts allemand, a décidé de publier ses propres données. Pour accéder aux informations, il a dû intenter un procès au géant des télécommunications Deutsche Telekom. Ces données, contenues dans un énorme document Excel, constituent les fondations de la carte interactive du Zeit Online. Chacune des 35 831 lignes de la feuille de calcul représente un transfert d'informations depuis le téléphone portable de Spitz au cours d'une période de six mois.

Prises individuellement, ces données sont inoffensives. Mais vues dans leur ensemble, elles offrent ce que les enquêteurs appellent un profil : une vision claire des habitudes et des préférences d'une personne, et de fait, de sa vie. Ce profil révèle quand Spitz sort de chez lui, quand il prend le train et quand il se trouve dans un avion. Il montre qu'il travaille principalement à Berlin et dans quelles autres villes il se déplace. Il permet de déterminer quand il est éveillé et quand il dort.

Deutsche Telekom gardait déjà une partie des données de Spitz privées, notamment le numéro des personnes qu'il appelait et qui le contactaient. Ces informations étaient non seulement susceptibles de porter atteinte à la vie privée de beaucoup de personnes, mais elles auraient également – même si les numéros avaient été cryptés – révélé beaucoup trop de choses sur Spitz (bien entendu, des agents gouvernementaux auraient accès à ces informations).

Nous avons demandé à Lorenz Matzat et Michael Kreil d'OpenDataCity d'explorer les données et de trouver une solution de présentation visuelle. « Au début, nous avons utilisé

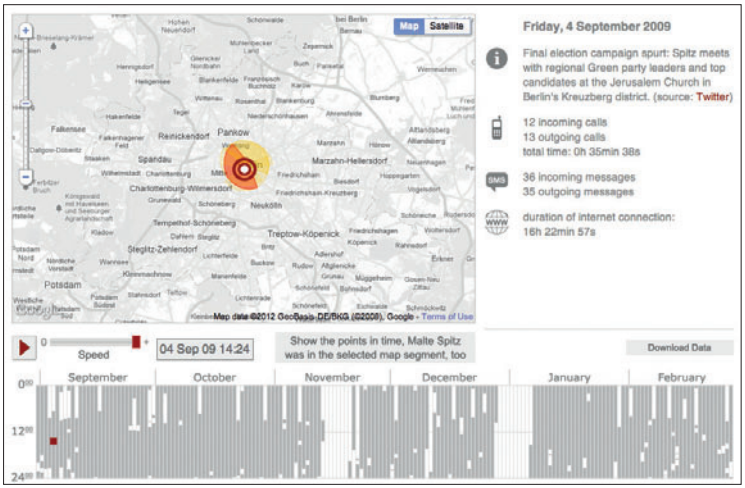

Figure 3-12. *Le téléphone omniscient*

des outils comme Excel et Fusion Tables pour comprendre les données nous-mêmes. Puis nous avons commencé à développer une interface cartographique pour permettre au public d'interagir avec les données de manière non linéaire », dit Matzat. Enfin, pour illustrer le niveau de détail que l'on peut extraire de ces données, celles-ci ont été enrichies d'informations publiquement disponibles sur les déplacements de Spitz (Twitter, billets de blog, informations du parti telles que le calendrier public disponible sur son site web). C'est ce genre de processus que tout bon enquêteur emploierait vraisemblablement pour profiler une personne sous surveillance. Avec les équipes de graphisme et de R&D internes du Zeit Online, ils ont finalisé une superbe interface pour explorer les données : en appuyant sur le bouton Lecture, vous partez en voyage dans la peau de Malte Spitz.

Après un lancement du projet couronné de succès en Allemagne, nous avons remarqué que nous recevions beaucoup de trafic de l'étranger, et nous avons décidé de créer une version anglaise de l'application. Après avoir reçu le prix Grimme Online allemand, le projet a été récompensé par un prix de l'Online News Association[7] en septembre 2011, le premier attribué à un site web d'information allemand.

Toutes les données sont disponibles dans une feuille de calcul Google Documents Lisez l'histoire sur le Zeit Online.

Sascha Venohr, Zeit Online

7 : http://journalists.org/2011/09/25/2011-online-journalism-award-winners-announced/

À propos des directeurs d'ouvrage

Jonathan Gray est directeur Politique et idées à l'Open Knowledge Foundation (http://okfn.org/), une organisation à but non lucratif dédiée à la promotion de l'open data, de l'open content et du domaine public, dans une très grande variété de domaines. Il est à l'origine de nombreux projets à l'OKFN, notamment « OpenSpending.org », qui illustre sur une carte les dépenses publiques dans le monde entier, et de « Europe's Energy », qui replace les objectifs énergétiques européens dans leur contexte. Il est par ailleurs chercheur en philosophie et histoire des idées au Royal Holloway, à l'université de Londres. Pour en savoir plus sur lui : jonathangray.org

Liliana Bounegru est rédactrice pour DataDrivenJournalism.net et chef de projet en datajournalisme à l'European Journalism Centre (www.ejc.net/). Dans ce cadre, elle coordonne les Data Journalism Awards et participe au travail éditorial du *Data Journalism Handbook*. Liliana est aussi chercheuse à l'université d'Amsterdam, où elle travaille sur la Digital Methods Initiative et sur le projet collaboratif EMAPS (*Electronic Maps to Assist Public Science*), dirigé par le Français Bruno Latour, sociologue des sciences et anthropologiste. Liliana est titulaire d'un master en Nouveaux médias et culture numérique et d'un master recherche en Études des médias à l'université d'Amsterdam. Elle anime un blog dédié à ces sujets, lilianabounegru.org

Lucy Chambers dirige l'unité Savoir *(Knowledge Unit)* à l'Open Knowledge Foundation. Elle est aussi chef de projet de School of Data, et a été *community manager* pour OpenSpending, Data-Driven Journalism et Spending Stories.

Nicolas Kayser-Bril a dirigé la version française du *Data Journalism Handbook,* le *Guide du datajournalisme*. À 27 ans, c'est l'un des pionniers du journalisme de données en France. Après avoir mis en place le pôle Datajournalisme chez OWNI à Paris en 2010, il a fondé avec Pierre Romera une société spécialisée, Journalism++, en 2011. Il a travaillé notamment avec WikiLeaks autour des *warlogs* irakiens et avec *60 Millions de Consommateurs* pour une enquête sur le prix de l'eau en France. Il intervient régulièrement dans les écoles de journalisme et participe à de nombreuses conférences professionnelles en Europe.

À propos des coordinateurs du projet

L'**European Journalism Centre** (www.ejc.net/) propose des formations destinées à renforcer la qualité de la couverture médiatique des affaires européennes et à apporter un support stratégique aux médias européens.

L'**Open Knowledge Foundation** (http://okfn.org/) aspire à un monde dans lequel le savoir « ouvert » serait omniprésent – à la fois en ligne et hors ligne – et promeut le savoir ouvert pour son potentiel à générer un impact social de grande envergure.