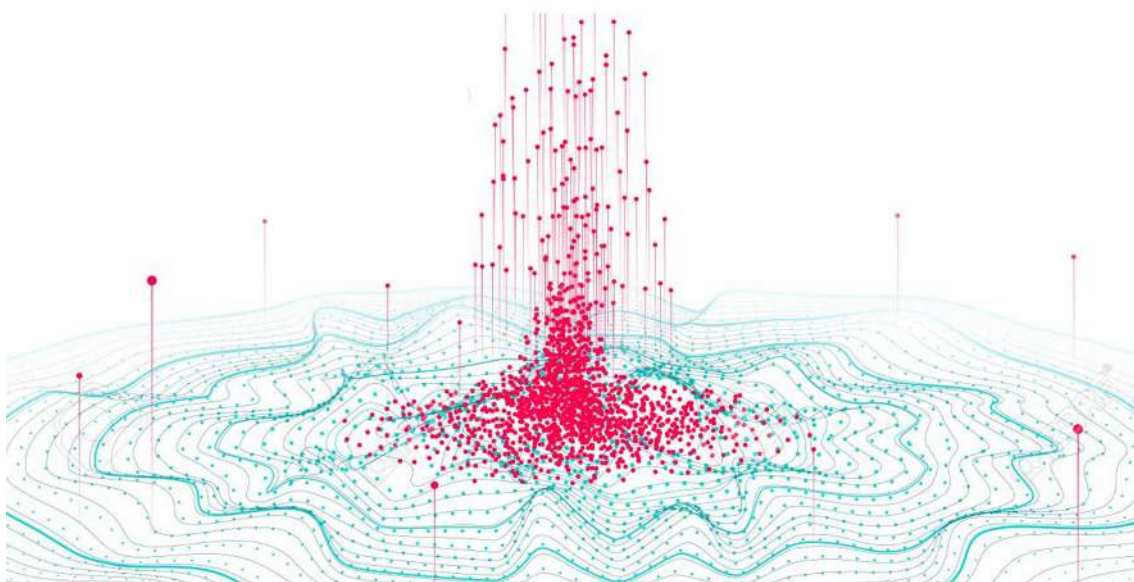


Datajournalisme **en pratique**



Laurence Dierickx / 2021-2022, quatrième édition
Université Libre de Bruxelles

Table des matières

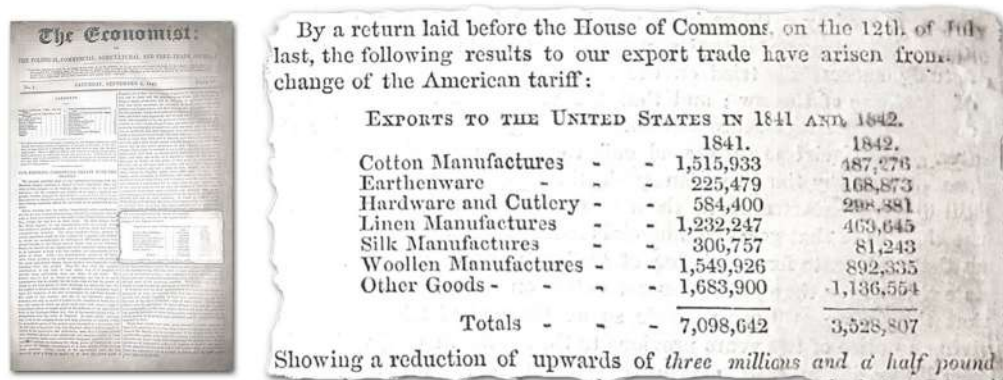
Introduction	4
1 Histoire du datajournalisme	6
1.1 Du journalisme de prédiction au journalisme de précision	7
1.2 Journalism assisté par ordinateur	9
1.3 Emergence du journalisme de données	12
1.4 Journalism computationnel, journalisme algorithmique	14
1.5 Typologie d'une approche par données dans le journalisme	21
1.6 Aperçu historique de la visualisation de données	22
1.7 Aperçu historique de la représentation cartographique	28
2 Contexte du datajournalisme	34
2.1 Au service de l'enquête journalistique	34
2.1.1 MP's expenses	34
2.1.2 WikiLeaks	35
2.1.3 The Migrant Files	36
2.1.4 The Panama Papers	37
2.1.5 Une femme tous les trois jours	39
2.1.6 How the virus got out	39
2.2 Dans les pratiques quotidiennes	39
2.2.1 Pratiques du datajournalisme en Belgique et en France	41
2.2.2 Le mythe de l'open data	44
2.3 Enjeux professionnels	44
2.3.1 L'objectivation chiffrée	46
2.3.2 La problématique de la qualité des données	48
2.4 Bonnes pratiques	50
2.4.1 Droit d'accès aux données publiques	51
2.4.2 Droits d'auteur	52
2.4.3 Normalisation des données	53
2.4.4 Modèles d'évaluation de la qualité des données dans un contexte journa- listique	55
2.4.5 Résoudre les problèmes de qualité des données	57
2.4.6 Visualisation de données	67

3 Outils du datajournalisme	78
3.1 Calculer	79
3.1.1 Formules mathématiques de base	80
3.1.2 Statistiques	81
3.1.2.1 Définitions	81
3.1.2.2 Statistiques descriptives	82
3.2 Rechercher	84
3.2.1 Formats et types de données	84
3.2.1.1 Bases de données tabulaires	84
3.2.1.2 Bases de données NoSQL (représentation graphique)	85
3.2.1.3 Bases de données relationnelles	85
3.2.1.4 Métadonnées	86
3.2.1.5 Conversion de fichiers	86
3.2.1.6 Types de données	87
3.2.2 Recherche booléenne	87
3.2.3 Bases de données en ligne	89
3.3 Récouter	91
3.3.1 Code HTML, CSS et code source	91
3.3.2 Scraper des données	94
3.3.2.1 Extraire des données d'un document PDF	94
3.3.2.2 Open Web Scraper pour Firefox	95
3.3.2.3 Chrome Web Scraper	98
3.3.2.4 OutWit Hub	99
3.3.2.5 Le package rvest de R	101
3.3.2.6 Outils en ligne	101
3.4 Vérifier	102
3.5 Nettoyer	104
3.5.1 Introduction à Open Refine	104
3.5.1.1 Organisation, tri et nettoyage des données	106
3.5.1.2 Exportation et historique des fichiers	112
3.6 Analyser	113
3.6.1 Introduction à Excel	113
3.6.1.1 Structure d'un tableau	113
3.6.1.2 Manipulations de base	114
3.6.2 Introduction à R	119
3.6.2.1 Interface de R Studio	119
3.6.2.2 Un super calculateur	120
3.6.2.3 Opérateurs logiques	121
3.6.2.4 Classes de données	121
3.6.2.5 Vecteurs, tableaux, listes, matrices	121
3.6.2.6 Scraper des données d'une page web	123
3.6.2.7 Ouvrir un fichier CSV, copier-coller un tableau	124

3.6.2.8	Analyse de données	125
3.6.3	Introduction à SQLite	128
3.6.4	Outils pour l'analyse du discours	129
3.7	Visualiser	130
3.7.1	Outils prêts à l'emploi	130
3.7.2	Visualisation de données avec R	132
3.7.3	Librairies JavaScript	134
3.7.3.1	Highcharts.js (package R : Highcharter)	134
3.7.3.2	Leaflet.js (package R : leaflet)	136
3.7.3.3	Tableaux de données interactifs (package R : DT)	138
3.7.3.4	Visualisation de réseaux, vis.js (package R : visNetwork)	138
3.8	Gérer	139
Bibliographie		140

Introduction

On retrouve les premières traces d'une approche par données dans le journalisme dès la fin du 19^e siècle. Par exemple, *The Economist* (illustration, ci-dessous)¹ a publié à la une de sa première édition, le 2 septembre 1843, un tableau de données relatives aux exportations de plusieurs entreprises. Le magazine britannique publiera sa première datavisualisation quatre ans plus tard. A la même époque, le même phénomène est observé dans le quotidien britannique *The Guardian*², qui adoptera pour devise, en 1921, "Les commentaires sont libres mais les faits sont sacrés" (Rogers 2013, McBride 2016), l'introduction de l'informatique a permis d'élargir le spectre des possibilités de traitement et du volume de données traitées.



By a return laid before the House of Commons, on the 12th. of July last, the following results to our export trade have arisen from the change of the American tariff:

EXPORTS TO THE UNITED STATES IN 1841 AND 1842.			
		1841.	1842.
Cotton Manufactures	- -	1,515,933	487,276
Earthenware	- -	225,479	168,873
Hardware and Cutlery	- -	584,400	298,381
Linen Manufactures	- -	1,232,247	463,645
Silk Manufactures	- -	306,757	81,243
Woollen Manufactures	- -	1,549,926	892,335
Other Goods	- -	1,683,900	1,136,554
Totals	- -	7,098,642	3,528,807

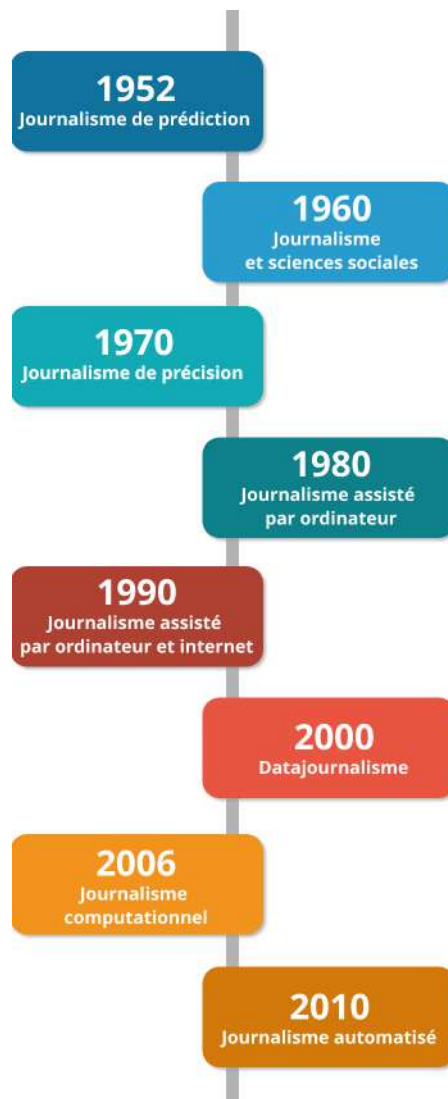
Showing a reduction of upwards of three millions and a half pound

Aujourd'hui, les pratiques du datajournalisme se fondent sur le principe de raconter une bonne histoire en s'appuyant sur des données en tant que matériel de base à l'information. Qu'il s'agisse de récolter, analyser ou traiter des jeux de données, la démarche fondamentale reste celle de n'importe quelle autre forme de journalisme. Mais les pratiques du datajournalisme requièrent aussi la maîtrise d'outils et de techniques spécifiques. Ces notes ont pour objet (1) de replacer le datajournalisme dans le contexte de son développement, (2) de présenter les différentes formes du datajournalisme, (3) d'aborder ses enjeux professionnels, (4) de fournir des outils et techniques opérationnels, tout en tenant compte de l'étendue des champs auxquels se rapporte le datajournalisme et de la diversité des techniques sur lesquelles reposent une approche par données dans le journalisme.

¹ Source : "Data journalism at The Economist gets a home of its own in print", Medium, <https://medium.economist.com/data-journalism-at-the-economist-gets-a-home-of-its-own-in-print-92e194c7f67e>

² Il s'agissait d'un listing dévoilant le nombre d'élèves et le coût de la scolarité dans les écoles Manchester (Rogers cité par Bounegru in Stray & al. 2013)

1 | Histoire du datajournalisme



Les premières traces de la rencontre entre informatique, données et journalisme remontent à 1952, époque à laquelle un programme informatique est développé pour le télédiffuseur américain CBS. Il a pour objectif de prédire le résultat des élections présidentielles : qui du républicain Eisenhower ou du démocrate Stevenson remportera le scrutin ? En choisissant Eisenhower, la machine ne se trompe pas (Cox 2000, Howard 2014). Nous sommes alors au début du développement de l'informatique. Les ordinateurs n'ont rien de commun avec les premiers ordinateurs personnels commercialisés au début des années 1980. Le Remington Rand UNI-

VAC¹ utilisé pour cette expérience a été commercialisé l'année précédente, il s'agissait du premier modèle d'ordinateur mis sur le marché (Chavand 2017)². La machine, qui comptait plus quatre mètres de long pour un peu plus de deux mètres de haut et un poids de huit tonnes, coûtait environ un million de dollars³. Elle était capable de traiter 10.000 opérations par seconde⁴.



FIGURE 1.1 – Source : <http://time.com/4271506/census-bureau-computer-history/>

1.1 Du journalisme de prédiction au journalisme de précision

En 1967, l'Américain Philip Meyer amorce une petite révolution dans les pratiques journalistiques en utilisant des outils informatiques pour asseoir une enquête menée auprès de la communauté afro-américaine, dans la foulée d'émeutes à Détroit, pour le journal *Detroit Free Press*. Cette méthode, issue des sciences sociales, donne lieu à une analyse des résultats sur un ordinateur IBM 360 à l'Université du Michigan (Cox 2000). Celle-ci révèle que, contrairement à l'hypothèse supposée, les personnes qui avaient fréquenté l'université étaient susceptibles de participer aux émeutes, de la même manière que les personnes ayant décroché de l'enseignement secondaire. Il décroche un Prix Pulitzer avec ce récit et marque ainsi les débuts de ce qu'il appellera, dans un livre éponyme, le journalisme de précision. Son intérêt n'est pas tant lié à

¹ *Universal Automatic Computer*

² L'ordinateur était capable de stocker une séquence en mémoire puis d'insérer cette séquence dans l'endroit approprié du programme, au besoin. En construisant une bibliothèque de séquences fréquemment utilisées, un programmeur pouvait ainsi écrire un programme complexe plus efficacement (Krusa et Sedek in Fuller 2008).

³ L'équivalent de plus de huit millions d'euros actuels.

⁴ Sources : "Nov. 4, 1952 : Univac Gets Election Right, But CBS Balks", *The Wired*, 11/04/2010, [urlhttps://www.wired.com/2010/11/1104cbs-tv-univac-election/](https://www.wired.com/2010/11/1104cbs-tv-univac-election/); et "The Story Behind America's First Commercial Computer", *The Time*, 31/03/2016, [urlhttp://time.com/4271506/census-bureau-computer-history/](http://time.com/4271506/census-bureau-computer-history/)

l'usage de l'informatique qu'à celui de l'application des méthodes de recherche en sciences sociales aux pratiques journalistiques. Meyer a toutefois appris Data-Text à Harvard (où il a également étudié les sciences sociales), un langage de haut niveau⁵ développé pour IBM (Cox 2000). Un mainframe IBM 360 avait été utilisé pour cette première expérience. L'ordinateur, qui lisait des cartes perforées, avait été lancé sur le marché en 1965. D'un volume de 8 m³, il permettait aux entreprises d'intégrer l'ensemble de leurs applications de traitement de données dans un seul système d'information⁶.

En 1968, c'est au tour du journaliste américain Clarence Jones d'exploiter les possibilités de l'informatique dans le cadre d'une enquête traitant de la corruption dans le système judiciaire du comté de Dade pour le *Miami Herald*. Maier (2000) souligne que cette approche est d'abord le fruit du hasard : le journaliste enquêtait sur un juge soupçonné de corruption mais celui-ci est mort alors que Jones recueillait des preuves. Il décida d'élargir son enquête pour retracer la manière dont chaque crime majeur avait été traité. Une telle investigation aurait nécessité deux années de travail. Une douzaine d'étudiants en droit furent embauchés et l'idée fut alors de stocker les données récoltées dans un ordinateur du même type que celui utilisé par Meyer. Face à la difficulté de traiter des données de type textuel, deux employés d'IBM furent détachés à la rédaction. 3.000 cas d'arrestations furent répertoriés, correspondant à 13.000 cartes perforées. Les résultats furent traités avec le langage de programmation de haut niveau COBOL (Cox 2000), créé en 1959 et qui est encore utilisé aujourd'hui même s'il n'est plus le plus répandu. Les résultats ont mis en lumière une série d'éléments : pour un taux de criminalité le plus élevé dans le pays, le taux d'arrestation dans ce comté était le plus bas ; la plupart des arrestations aboutissaient rarement à des peines d'emprisonnement ; plus d'un tiers des arrestations concernaient des jeunes âgés entre 17 et 20 ans ; s'ils représentaient 15% de la population, 73% des jeunes afro-américains avaient fait l'objet d'une arrestation.

Si cette enquête a eu peu d'impact en termes de reconnaissance du travail journalistique ou d'influence sur les politiques, il s'agissait du premier usage journalistique d'ordinateurs pour analyser des données publiques. David Burnham, journaliste au *New York Times* fait aussi partie de ces pionniers ayant fait usage de l'informatique pour analyser des données publiques. En 1972, il croise plusieurs types de données (statistiques des arrestations à New York City, rapports criminels et enregistrements du service de police) et conclut à une divergence entre le nombre de crimes signalés et le nombre d'arrestations effectuées. Il observe également un taux de criminalité par 1.000 habitants qui est de 328 fois supérieur dans le quartier de Harlem (DeFleur 1997, Cox 2000).

La première utilisation de documents publics publiés de manière électronique remonte, quant à elle, à 1978 (DeFleur 1997) lorsque Rich Morin et Fred Tasker, journalistes au *Miami Herald*,

⁵ En informatique, un langage de haut niveau désigne un langage de programmation proche du langage naturel et éloigné du langage binaire de l'ordinateur. Il permet donc de s'abstraire du fonctionnement de la machine, le code devant nécessairement être compilé – c'est à dire, être réinterprété pour être compris par la machine. Source : Techno-science.net, <http://www.techno-science.net/?onglet=glossaire&definition=11377>.

⁶ "Mainframe Introduction 2", archives IBM, URL : https://www-03.ibm.com/ibm/history/exhibits/mainframe/mainframe_intro2.html

entreprennent l'examen des relevés d'imposition dans le Comté de Dade. A l'aide d'un main-frame IBM et d'un logiciel SPSS – un logiciel d'analyse statistique et prédictive développé par IBM –, ils ont comparé le prix de vente des propriétés à celui de leur estimation. Les résultats ont montré que les biens les plus onéreux avaient été évalués à un taux effectif inférieur à ceux des biens vendus à bas prix.

Désignées par Meyer par le vocable "journalisme de précision"⁷, il y consacrera un livre dont la première édition est publiée en 1973. Elle sera rééditée en 1991 sous le titre "The New Precision Journalism". Il y plaide pour une intégration de l'informatique et des statistiques dans les pratiques journalistiques, constatant que l'introduction des ordinateurs a eu pour effet de mettre à la disposition des journalistes des volumes d'information de plus en plus importants. Il s'agit d'une approche scientifique, souligne-t-il, qui tranche avec le nouveau journalisme, popularisé par Breslin et Wolfe dans les années 1960 et caractérisé par une forme narrative empruntée à la fiction.

D'autres expériences croisant journalisme, informatique et données auront lieu tout au long des années 1970 aux États-Unis, où le phénomène reste circonscrit et le fait de quelques initiés, à commencer par Meyer. La raison principale, indique Maier (2000) est à trouver dans une résistance des journalistes à l'utilisation des nouvelles technologies. Mais cela s'explique aussi dans le temps nécessaire pour conduire une enquête et seuls les journalistes obstinés ont supporté ce processus. Un exemple illustre particulièrement ce propos : celui des journalistes Donald Barlett et James Steele, du *Philadelphia Inquirer*, qui ont dû retranscrire des informations provenant de plus de 10.000 documents et 20.000 pages de transcriptions, avec l'aide d'employés de bureaux, dans le cadre d'une enquête menée en 1973 sur les inégalités dans le système judiciaire.

Au-delà d'un appui à l'investigation journalistique, ces premières expériences témoignent des difficultés d'accès aux technologies informatiques, supposant que la rédaction concernée dispose de moyens pour s'offrir une technologie coûteuse à l'époque et que les journalistes disposent d'un niveau de compétence suffisant pour maîtriser les outils ou, à défaut, puissent bénéficier de l'appui d'experts. Le journalisme de précision peut être entendu comme "*une réaction aux insuffisances et aux faiblesses souvent prêtées au journalisme : dépendance aux communiqués de presse (plus tard qualifié de 'churnalism', ou journalisme prémâché), influence des sources d'autorité, etc.*" (Bounegru in Stray & al. 2013).

1.2 Journalisme assisté par ordinateur

L'arrivée sur le marché des premiers ordinateurs personnels, moins encombrants que leurs ancêtres et économiquement plus accessibles, couplée au développement de logiciels de tableurs permettront de changer progressivement la donne (Parasie 2013). Aux États-Unis, on parle alors de journalisme assisté par ordinateur (*computer assisted reporting*, CAR) pour dési-

⁷ Cette locution sera traduite en français par "journalisme scientifique".

gner ces pratiques qui concernent la collecte des données et leur analyse statistique. La question des compétences est importante dans ce développement mais il s'agit là de compétences générales plutôt axées sur la collecte d'information (Coddington 2014).

Le journalisme assisté par ordinateur se déploie dans le contexte de l'investigation journalistique (Gynnild 2013) et dans une diversité de formes (DeFleur 1997), mais il ne concerne toujours qu'une minorité d'acteurs (Parasie 2013). Ceux-ci, indique Meyer (cité par Lewis & Usher 2013) ont souvent acheté leur propre ordinateur avant que la technologie s'installe dans les rédactions. Au milieu des années 1980, constate Maier (2000), la majorité des journaux n'avaient toujours pas utilisé de bases de données en ligne, telles que Nexis⁸ et Dialog⁹. Malgré l'incroyable travail réalisé par quelques pionniers, indique Maier (2000b), l'industrie de la presse s'est montrée hésitante, voire résistante, quant à l'adoption de nouvelles technologies : de l'analyse de base de données à la publication numérique, l'utilisation de technologies informatiques dans les rédactions a souvent été considérée comme le fait d'élites technologiques.

Dans le cadre du développement du journalisme assisté par ordinateur dans les années 1980, deux noms font figure de cas d'école : Elliot Jaspin et Bill Dedman. Le premier travaille pour le *Providence Journal*. Il se positionne en précurseur dans l'usage de base de données informatiques. Parmi les enquêtes sur lesquelles il a travaillé, celle consacrée au décès d'enfants à la sortie des bus de ramassage scolaire. Il croise alors les données sur les chauffeurs de bus scolaires et celles sur les infractions à la circulation routière, puis les confronte aux dossiers judiciaires concernant les chauffeurs de bus en infraction, il découvre que certains chauffeurs sont des trafiquants de drogue. Cette découverte aura un effet immédiat : suite à la publication de son article, l'État réorganise les procédures de délivrance de permis pour les chauffeurs d'autobus scolaires (Cox 2000).

Une autre enquête de Jaspin a pour sujet la délivrance de prêts hypothécaires. En analysant des enregistrements sur bandes magnétiques, il découvre que les taux d'intérêt les plus bas sont accordés aux enfants de hauts fonctionnaires (Cox 2000). Jaspin est également considéré comme étant le premier journaliste à avoir démontré l'utilité de maîtriser le langage SQL (*Structured query language*) dans le cadre d'un travail journalistique (Cohen & al. 2011). Il fut aussi à l'origine du développement du "Nine-track-express" – un système imitant les fonctions d'un ordinateur mainframe, dont le prix de vente oscillait entre 9.000 et 12.000 dollars – (DeFleur 1997), ainsi que le premier journaliste américain à remporter un Prix Pulitzer pour une investigation s'appuyant sur les techniques du journalisme assisté par ordinateur (Maier 2000). Pour autant, il ne bénéficia pas du soutien immédiat de son journal. Meyer (1991) rapporte que Jaspin a été officiellement interdit de prendre des cours d'informatique, considérant qu'il n'avait

⁸ Nexis fournit des informations, notamment économique, issues d'une variété de source en ce compris la presse d'information internationale et la presse professionnelle. Cette base documentaire, mise à jour quotidiennement, est développée depuis le début des années 1980. Elle est toujours entretenue actuellement. Source : Université de Sheffield, URL : <https://www.sheffield.ac.uk/library/cdfiles/nexis>

⁹ Base de données de ressources documentaires, fondée en 1972 et rachetée en 2008 à Thomson Reuters par ProQuest. Sources : Summit (2002), "Press release : ProQuest Acquires Dialog", 2008, URL : <http://www.proquest.com/about/news/2008/ProQuest-Acquires-Dialog.html>

pas besoin de connaître quoi que ce soit à propos d'un ordinateur, le journal disposant d'un service IT.

Bill Dedman fait également de ces figures emblématiques ayant contribué à mettre en lumière les techniques d'enquête relevant du journalisme assisté par ordinateur. En 1988, il publie dans le *Atlanta Journal-Constitution* "La couleur de l'argent" (Berret & Phillips 2016b), qui lui valut, l'année suivante, un Prix Pulitzer. Dans cette enquête, le journaliste a croisé les données statistiques du bureau de recensement américain avec celui des données relatives aux prêts hypothécaires, démontrant une discrimination de la population afro-américaine dans l'octroi des prêts, y compris parmi les classes sociales plus aisées (Flew & al.). Pour Maier (2000), les pratiques relevant du journalisme assisté par ordinateur sont moins une nouvelle forme de journalisme qu'une extension de la tradition du journalisme d'investigation. Elles permettent de traiter des sources d'information disponibles dans un format numérique, de créer ses propres bases de données à partir d'un travail de récolte sur le terrain, d'analyser des jeux de données grâce à la puissance de calcul d'un ordinateur. Il s'agit donc de moyens pour les "*journalistes compétents de suivre les données, de creuser plus profondément, de déterrer les modèles sous-jacents et les faiblesses de la société*".

Le journalisme assisté par ordinateur ne concerne encore qu'une poignée de journalistes convaincus des bénéfices de cette approche par données. La question des compétences est également prégnante car l'outil informatique suppose un minimum de maîtrise. Cette double perspective présidera à la naissance de la NICAR (*National institute for computer assisted reporting*) en 1989, sous l'impulsion d'Elliot Jaspin dans le cadre d'un programme de la Missouri School of Journalism et de l'IRE (*Investigative Reporters and Editors*) (Cox 2010). L'organisation organise des conférences et des formations, gère des forums de discussion, ... Aujourd'hui encore, elle joue un rôle central en connectant et en formant les journalistes pratiquant le datajournalisme (Coddington 2014). Depuis sa création, la NICAR a formé des milliers de journalistes issus de plus de trente pays (Gynnild 2013).

A l'époque, les bases de données n'ont pas besoin d'être énormes pour être utiles. Un logiciel de gestion de bases de données offre plusieurs fonctionnalités en termes de tri, de calcul et d'analyse, lesquelles permettent de simplifier les tâches (Leonard 1992). Dans certains cas, il est pertinent de créer une base de données sur mesure, notamment lorsque l'information n'existe pas dans un format numérique. Les bases de données sont aussi utilisées dans une perspective documentaire, dans le cadre de pratiques devenues de plus en plus courantes. Considérées comme des bases de connaissance, elles permettent d'accéder à des documents officiels ainsi qu'aux archives des médias, via des services en ligne.

Les années 1990 connaîtront un tournant technologique majeur avec l'arrivée du World Wide Web dans le domaine public, le 30 avril 1993¹⁰. Au cours de cette décennie, l'utilisation de l'ordinateur dans le cadre des processus journalistique va varier en fonction de la complexité de la

¹⁰ "Naissance du web", CERN, URL : <https://home.cern/fr/topics/birth-web>

tâche à effectuer et va notamment concerner des dispositifs de collecte d'informations (Garri-son 2001). La diffusion de l'innovation technologique va toutefois dépendre tant de la forma-tion interne au sein des entreprises de presse que du soutien au développement de moyens permettant de créer une "masse critique d'utilisateurs" (Maier 2000). L'adoption des techno-logies du web est toutefois lente malgré le potentiel de ressources pour les journalistes (Bar-doel & Deuze 2001), tandis que les techniques d'enquête relevant d'une approche par don-nées vont bénéficier de nouveaux outils avec les logiciel SPSS (*Statistical Package for the Social Sciences*) et SAS (*Statistical Analysis System*), le second étant davantage accessible aux non-programmeurs.

Au milieu des années 1990, le nombre de projets journalistiques axés sur les bases de données augmente de manière considérable (Dagiral & Parasie 2012), gagnant également les entreprises de presse de taille moyenne ou de petite taille. L'utilisation de bases de données est subordon-née à une histoire journalistique, aux normes journalistiques et à la pratique essentielle des statistiques car elles permettent de comprendre les données. Dans le même temps, la quasi-totalité des rédactions ont été informatisées et l'ordinateur fait désormais partie du quotidien de l'activité journalistique, via le traitement de texte, la PAO et la consultation de bases de don-nées. Ces technologies font l'objet d'un usage individuel et collectif (Pélissier & Romain 1998).

1.3 Emergence du journalisme de données

Si les années 1980 furent celles des expérimentations et les années 1990 celles d'une intégration progressive d'internet dans les pratiques journalistiques, l'entame des années 2000 marque le tournant d'une troisième vague dans l'histoire du journalisme et des technologies de l'in-formation et de la communication : en à peine une décennie, la plupart des grands médias d'information se sont dotés d'un site web et les audiences en ligne se comptent en millions d'internautes. Au chevet de ces évolutions, les milieux académiques ont déjà commencé à s'in-téresser à l'émergence de nouveaux comportements et compétences au sein de la profession journalistique et les recherches traitant des médias et du journalisme en ligne (et de ses pra-tiques) tendent à se multiplier (Deuze 2003).

Sur le terrain, une nouvelle forme de journalisme liée à une approche par données va émer-ger : le datajournalisme ou journalisme de données, lequel s'inscrit dans la continuité du jour-nalisme assisté par ordinateur tout en s'en distanciant dans ses pratiques (Coddington 2014). Il fait référence au processus par lequel les journalistes utilisent des données numériques, lesquelles constituent le matériau de base à l'information (Gynnild 2013). S'il n'en réinvente pas les techniques, il compose "*avec les outils et pour les audiences de notre époque*" (Joannes 2007). Les contenus que les journalistes de données conçoivent "*sont désignés sous le terme de 'news applications' qui désignent aussi bien des cartes interactives, des infographies, des bases de données interrogeables et d'autres formes de présentation en ligne*" (Dagiral & Parasie 2013). Anderson & al. (2016) définissent le datajournalisme comme l'application de techniques sta-tistiques à l'analyse de sources telles que de bases de données informatiques, des enquêtes d'opinions ou de tout autre type d'enregistrement numérique. Pour Trédan (2014), le datajour-

nalisme s'appuie sur deux ressorts principaux : *"d'un côté, un savoir-faire technologique qui permet le développement des interfaces de visualisation des données; de l'autre, un savoir-faire dans l'identification et l'exploitation de données disponibles"*.

Le journalisme de données conserve le principe de subordination des données aux valeurs journalistiques, l'objectif étant de mettre les données au service du récit, et il maintient l'accent sur une activité éditoriale pilotée par le sens que l'analyse va donner aux données (Coddington 2014). Simon Rogers, fondateur du *Guardian Data Blog* aujourd'hui employé par Google, estime que le datajournalisme, c'est *"80% de transpiration, 10% de grandes idées et 10% de 'sortie' (output)"*. Mais, souligne-t-il, il ne faudrait pas réduire le journalisme de données aux seuls graphiques et visualisations. Le datajournalisme, c'est d'abord raconter une histoire de la meilleure manière possible. *"Parfois, cela peut être une visualisation ou une carte. (...) Parfois, publier les chiffres est suffisant"* (Rogers 2013). Steve Doig, figure emblématique du journalisme assisté par ordinateur des années 1990 aux États-Unis et professeur à l'Université du Massachusetts, indique que le journalisme de données, ce sont des techniques issues de sciences sociales à laquelle sont associés des *deadlines* et des outils informatiques devenus aujourd'hui plus accessibles et plus simples d'utilisation. Si les journalistes ne sont pas appelés à développer des compétences techniques spécifiques, il estime qu'ils doivent néanmoins apprendre à collaborer avec des développeurs et des designers ¹¹.

Bradshaw (2017) a défini dix principes de bases pour la pratique du journalisme de données. Si certains sont discutables (par exemple, lorsqu'il écrit qu'il ne faut ne compter sur des récits où les données existent et sont faciles à obtenir, cela s'oppose aux principes du journalisme d'investigation où il s'agit de rechercher ce qui n'est pas apparent; ou lorsqu'il affirme que partager le code d'applications développées "maison" ne sert pas à grand-chose si ce code n'est pas applicable dans d'autres contextes et/ou à d'autres domaines d'application), d'autres caractérisent très bien les enjeux d'une approche par données dans le journalisme. Il s'agit de ceux relatifs à l'appréhension des données en tant qu'objets d'exercice du pouvoir, de l'indépendance éditoriale et technologique, de l'objectivité – qui serait réalisée par la transparence – dans le choix des sources et la manière de concevoir des outils, et de se concentrer sur l'histoire humaine qui se trouve derrière les données ^{12 13}.

La paternité du terme "datajournalisme" est attribuée au journaliste-développeur Adrian Holovaty, (Howard 2014). D'abord responsable des innovations éditoriales au *Washington Post* et fondateur de EveryBlock, un projet de datajournalisme hyperlocal qui reposait sur l'extraction de données provenant de sites web institutionnels et de fichiers transmis par les autorités pour informer les habitants de Chicago à l'échelle de leur quartier (Parasie 2013). En 2005, il crée

¹¹ "Le datajournalisme, c'est de l'enquête, pas du code", Médiaculture.fr, Cyrille Franck, 27/03/2013, URL : <http://www.mediaculture.fr/le-data-journalisme-cest-de-lenquete-pas-du-code/>

¹² "10 principles for data journalism in its second decade", Paul Bradshaw, Medium, 14/09/2017, URL : <https://medium.com/thoughts-on-journalism/10-principles-for-data-journalism-in-its-second-decade-3b45e08a4793>

¹³ Voir aussi "Enjeux d'une approche par données" p.34

l'un des premiers *mashup* à l'aide de Google Maps, en proposant une cartographie du crime à Chicago. L'année suivante, il publie un manifeste appelant le journaliste à changer non pas sur le fond mais sur la forme, réclamant un journaliste de faits s'appuyant sur des chiffres et des données statistiques (Dagiral & Parasie 2012). Holovaty – diplômé de l'école de journalisme du Missouri – est impliqué dans la communauté des logiciels libres. Il est aussi le cocréateur de Django, un framework développé en langage Python. Conçu pour permettre aux programmeurs de respecter de courts délais de livraison, il a été imaginé à la fois pour un usage spécifique à un site web d'informations et pour soutenir des projets s'appuyant sur des bases de données informatisées (Dagiral & Parasie 2012). Pour Holovaty, faire du journalisme à l'aide de programmes informatiques n'est rien de plus qu'une manière différente d'accomplir des objectifs journalistiques (Powers 2012).

Pour Bounegru (in Stray & al. 2013), le journalisme de données traduit une démocratisation "massive" des ressources, outils, techniques et méthodologies utilisés précédemment par des spécialistes, qu'ils soient journalistes d'investigation, spécialistes des sciences sociales, statisticiens, analystes ou autres experts. Deux raisons permettent d'expliquer son développement : un accès accru à des volumes de données de plus en plus importants, et le développement de technologies permettant l'extraction de données ainsi que leur visualisation sous la forme de cartes ou graphiques interactifs (Henninger 2012).

1.4 Journalisme computationnel, journalisme algorithmique

L'évolution des technologies de l'information et de la communication a eu pour conséquence le développement de nouvelles pratiques journalistiques, désignées par la locution générique "journalisme computationnel" (*computational journalism*). Celle-ci fait référence à des pratiques hybrides — extraction et visualisations de données, génération automatique de textes et analyse de contenus — qui s'appuient sur des outils des sciences de l'informatique, des sciences sociales et des sciences humaines numériques pour collecter, vérifier, représenter, publier, diffuser des informations (Cohen & al. 2011). Le vocable fut utilisé pour la première fois en 2006, au Georgia Institute of Technology, à l'occasion d'un cours dispensé par le professeur Irfan Essa (Gynnild 2013). Il est défini comme étant une combinaison d'algorithmes, de données et de la connaissance des sciences sociales pour compléter la fonction de responsabilité du journalisme (Hamilton & Turner 2010). Il recouvre des pratiques aussi variées que sont l'exploration de données (*data mining*), l'extraction de données (*data scraping*), la fouille de textes (*text mining*), la manipulation et la visualisation de données, l'analyse statistique et la génération automatique de textes en langue naturelle. Lorsque ses processus se concentrent sur la seule génération automatique de textes, il est aussi appelé "journaliste algorithmique" (Graefe 2016).

L'ensemble des processus journalistiques sont concernés par les pratiques du journalisme computationnel, lesquelles respectent des valeurs fondamentales du journalisme telles que la précision, l'immédiateté et la vérifiabilité (Graefe 2016, Diakopoulos 2011). Leurs limites sont celles d'être très dépendantes d'une expertise du domaine d'application doublée d'une expertise

technique, qu'elle soit développée par le journaliste ou en collaboration avec Cdes professionnels qualifiés. Ces formes de journalisme seraient moins liées aux normes et pratiques professionnelles mais, tout comme le journalisme de données, elles constituent une réponse pertinente pour faire face à l'abondance d'informations (Coddington 2014, Flew & al. 2012) . Bozkowski (2004, cité par Anderson 2012) indique que le développement de ces nouvelles pratiques informationnelles s'explique à la fois par des facteurs organisationnels, des routines de travail et des représentations des utilisateurs qui façonnent des processus d'adoption de technologies qui résultent elles-mêmes des évolutions technologiques. Ce façonnage mutuel donnerait ainsi lieu à des produits éditoriaux distincts.

L'automatisation est l'une des principales composantes du journalisme computationnel, de la détection de *breaking news* ou de contenus à valeur informative ajoutée (Steiner & al. 2013, Leppänen & al. 2017) à la publication personnalisée, en passant par la sélection de l'information, le *fact-checking*, la hiérarchisation des contenus et la production automatisée de textes. Les processus d'automatisation prennent souvent la forme d'algorithmes, considérés comme l'un des éléments de la pensée computationnelle, laquelle consiste à comprendre comment fonctionne un programme informatique (Berret & Phillips 2016). Ces processus consistent en une abstraction d'une procédure itérative qui, dans le contexte journalistique, est utilisée pour prioriser, classer ou filtrer l'information (Diakopoulos 2015), prédire des demandes dans le contexte du "big data" où créer des contenus originaux (Napoli 2014). La croissance exponentielle des volumes de données disponibles ¹⁴.

La forme des applications automatisées, qui consistent en "des fenêtres sur les données d'une histoire" (Stray & al. 2013), varie en fonction de la problématique abordée. Elles permettent de prendre connaissance de données d'une manière vulgarisée ou d'interagir avec celles-ci. Elles ont pour point commun de s'appuyer sur des données publiques diffusées en temps réel. En ce sens, elles répondent à un critère temporel d'actualité et renvoient au processus de sélection de l'information, l'un des maillons essentiels de la chaîne journalistique. Un processus de sélection d'informations participe à des choix éditoriaux. Dans le cadre d'une application automatisée de production d'informations, celui-ci ne s'opèrerait pas de manière différente. Choisir un sujet, un angle et la manière dont sera traitée l'information : les routines du journalisme sont caractérisées par une série de choix. Comme le rappelle Gillespie (2013), les choix éditoriaux existent depuis que la presse existe et ceux-ci ne sont pas exempts de neutralité. Le parallèle est clairement établi avec un processus informatique ou algorithmique : il relève lui aussi de choix humains et donc de jugements de valeur (Kraemer & al. 2011).

L'automatisation de la chaîne de production a été souvent abordée sous l'angle de la production automatisée d'informations : en prenant en charge des tâches habituellement dévolues à des journalistes humains, ces processus informatisés sont susceptibles de remettre en question l'autorité du journaliste, son identité professionnelle mais aussi sa valeur ajoutée.

¹⁴ Pour Lewis & Westlund (2015), ce "grand moment de données" n'est pas seulement une transition technologique vers le déluge de données : il s'agirait plutôt d'un phénomène sociotechnique aux origines et implications culturelles, économiques et politiques.

Dans ce domaine, il convient de distinguer deux types de démarche : d'une part, ce que l'on appelle communément le robot journalisme, désignant des logiciels de génération automatique de textes en langue naturelle (GAT) ; et d'autre part, des applications web ou mobiles pilotées par des données dont le *storytelling* peut prendre des formes variées. Toutefois, la production automatisée d'informations est plus largement associée à la GAT, née dans le giron du traitement automatique du langage naturel (TAL) dont la genèse remonte à la fin des années 1940 (Bouillon 1998, Jones 2001). Longtemps confinée au champ de la recherche, elle s'est intéressée au traitement automatique des niveaux lexicaux, syntaxiques et sémantiques de la langue. Ses premières applications commerciales remontent au début des années 1990, dans le domaine de la rédaction automatique de bulletins météorologiques (Danlos & Pierrel 2000). Très vite, la GAT a été appliquée à une variété de domaines tels que les modes d'emploi de produits commerciaux et divers textes associés au monde de l'entreprise pouvant être standardisés (Dale 1995).

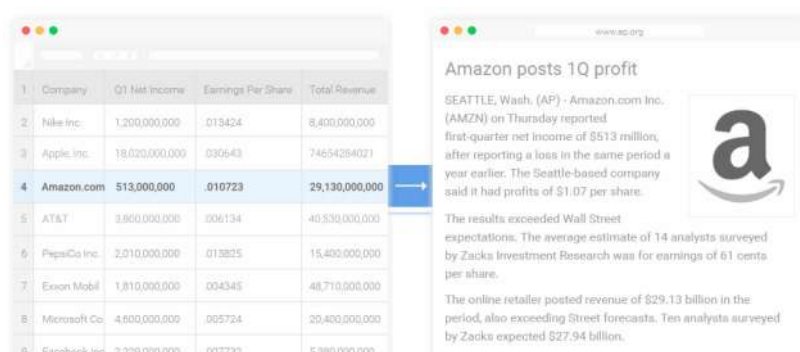


FIGURE 1.2 – Génération automatique de textes avec le logiciel WordSmith

La génération automatique de textes se développe dans le champ journalistique depuis 2007, d'abord aux États-Unis, et à partir de 2015 en Europe de l'Ouest, en Russie, en Chine et en Amérique latine. StatSheet est la première société à proposer ce type de services aux médias américains, dans le domaine des résultats sportifs. En 2010, elle lance un réseau de 345 sites couvrant les équipes de basket-ball universitaires du collège de la Division 1 de la NCAA et génère, chaque mois, plus de 15.000 textes (Van Dalen 2012). Elle change de nom en 2011 pour devenir Automated Insights. La société décrit ses activités comme suit : "*Notre technologie humanise de grands ensembles de données en y repérant les tendances et les idées clés. Elle décrit ces résultats un en anglais simple qui ne se distingue pas de celui produit par un écrivain humain*" (Dörr 2016). Automated Insights a produit, avec son logiciel WordSmith, 300 millions de contenus en 2013, un chiffre qui a dépassé le milliard l'année suivante (Hammond 2017), à une vitesse pouvant atteindre 2.000 textes générés par seconde. Un demi-million de tweets ont également été générés de manière automatique et 2.000 applications ont été alimentées par ses productions. Outre ses qualités de rapidité, le logiciel est multilingue et génère des visualisations de données.

En 2014, Automated Insights signe un partenariat avec l'agence de presse Associated Press, pour la rédaction automatisée de rapports trimestriels d'entreprises. WordSmith permet à l'agence d'étendre considérablement sa zone de couverture, passant d'une couverture de 300 à 3.000 sociétés. Les dépêches sont diffusées sans avoir fait l'objet d'une relecture humaine préalable, cette activité étant considérée comme trop consommatrice de temps (Dörr 2016). Toutefois, un contrôle humain demeure pour les entreprises nécessitant un traitement plus nuancé (Google, Coca Cola, American Airlines,...), tandis que 180 autres font l'objet d'un monitoring¹⁵. Le fait que ces articles contiennent moins de fautes d'orthographe et que le temps de traitement entre la réception des données et sa publication se situe entre une et dix minutes constituent deux autres avantages mis en avant (Linden 2017b).

Narrative Science, lancée en 2010, est l'autre leader du marché américain. La première version de son logiciel, StatsMonkey, traitait de résultats de matches de baseball et calculait les chances que l'équipe avait de gagner à mesure que le jeu progressait (Graefe 2016). Il s'agissait, au départ, d'un projet académique mené à la Northwestern University en partenariat avec la Medill School of Journalism (Carlson 2014). Kris Hammond, co-fondateur de Narrative Science, est connu pour ses déclarations choc : en 2011, il affirmait au *New York Times* qu'un "robot" gagnerait un Prix Pulitzer dans les cinq ans. L'année suivante, il déclarait que d'ici quinze ans, plus de 90% des contenus d'informations seraient automatisés (Hammond 2017). Mais il s'agissait moins de remplacer des journalistes que de faire face à la demande croissante de contenus originaux¹⁶. En 2012, la société intègre parmi ses clients le magazine économique *Forbes* pour la génération d'articles économiques (Graefe & al. 2015). Elle compte aussi parmi ses clients le site d'une chaîne sportive du groupe Fox et des sites spécialisés dans le sport local et l'information destinée aux jeunes.

Narrative Science et Automated Insights font partie de ces nouveaux acteurs arrivés sur le marché des médias, sans pour autant y être liés autrement que dans une relation fournisseur-client. En élargissant le champ de leurs services aux médias, ils participent à "un tournant quantitatif dans le journalisme" (Coddington 2014). Leurs logiciels sont capables de traiter n'importe quel type de récit dans une variété d'angles, à partir de données structurées. Les deux sociétés ont également pour point commun d'adosser leurs technologies au domaine de l'intelligence artificielle, par l'utilisation de techniques de *machine learning*. Toutes deux ont fait l'objet d'investissement importants depuis leurs débuts, ceux-ci se chiffrant en millions de dollars. Elles ont aussi pour autre point commun d'employer des ingénieurs et des linguistes. Chez Narrative Science, toutefois, une fonction de méta-journaliste (ou de méta-écrivain) a été créée. Celui-ci a pour mission de définir les cadres du récit ainsi que le langage descriptif approprié. Il intervient donc en amont de la génération automatique (Carlson 2014) avec le logiciel comme seul producteur final de l'information. A l'échelle mondiale, on dénombre moins d'une dizaine de sociétés spécialisées dans la génération automatique de textes en langage naturel (Dörr 2015). En Eu-

¹⁵ "Ap's 'robot journalists' are writing their own stories now, Miller Ross, *The Verge*, 29/01/2015, URL : <https://www.theverge.com/2015/1/29/7939067/ap-journalism-automation-robots-financial-reporting>

¹⁶ "L'actu automatique", Yves Eude, *Le Monde*, 15/11/2012, URL : http://www.lemonde.fr/actualite-medias/article/2012/11/15/l-actu-automatique_1790835_3236.html

rope, elles sont implantées en France (Syllabs, Labsense et YSEOP dont le siège social est établi aux États-Unis), au Royaume-Uni (Aria NLG), et en Allemagne (AEXEA, Restresto, Textomatic, Text-On et 2txt). Aucune des sociétés actives dans le secteur de la génération automatique de textes ne se revendique comme un média, pas plus qu'aucune n'estime "faire" du journalisme.

Créée en 2006, la start-up française Syllabs a fait ses débuts dans les milieux du marketing et de l'e-tourisme. Elle s'est ensuite orientée vers l'analyse sémantique et l'offre de services, tels que la valorisation d'archives et des revues de presse, destinés aux médias et éditeurs de contenus. Parmi ses premiers clients, *Les Échos*, *Slate.fr*, *La Tribune* et France Télévisions. Le 22 mars 2015, sa solution logicielle Data2Content génère 36.800 textes qui seront publiés en ligne sur le site du *Le Monde*, à l'occasion du premier tour des élections régionales. Pour Luc Bronner, directeur adjoint des rédactions du *Monde*, il s'agit de l'expérimentation de nouveaux outils susceptibles d'apporter un nouveau service aux lecteurs. Toutefois, il admet qu'en raison du volume de textes, ceux-ci sont plus facilement "*repérables sur les moteurs de recherche*" ¹⁷.

Les qualités de rapidité, de vitesse et de précision sont unanimement reconnues à ces technologies d'automatisation, qui permettent également d'étendre une zone de couverture médiatique à des domaines qui étaient peu – voire pas – couverts jusque-là, et de brasser de larges volumes de données de plus en plus disponibles et qu'il ne serait humainement pas possible de traiter sans ces technologies (Leppänen & al. 2017). Toutefois, si les expériences se multiplient, elles peuvent encore être considérées comme marginales ou expérimentales (Linden 2017), et cela pour deux raisons. La première est liée aux investissements financiers que cela représente : faire appel aux services d'une société spécialisée dans la génération automatique de textes ou mobiliser une équipe en interne pour développer une solution sur mesure présente un coût qu'il s'agit de mettre en balance avec les bénéfices attendus de l'opération. Dans un contexte de crise, seuls les médias les plus solides seront prêts à investir dans ces technologies.

Le deuxième argument est avancé par Graefe (2016) : les systèmes de production automatisés d'informations ne seraient pas davantage répandus dans les médias aujourd'hui car les rédactions ne disposent pas des ressources et compétences nécessaires pour développer des solutions en interne. "*Les développeurs dans les rédactions sont un peu comme l'eau dans le désert : rares et très recherchés*", affirmaient Gray & al. en 2013. Quant à ces professionnels hybrides, à mi-chemin entre journalisme et informatique, ils ne représentent encore qu'une minorité au sein de la profession, y compris aux États-Unis, qui fait figure de pionnier en la matière. Une seule expérience de journalisme automatisé y est attribuée à un profil de ce type : celle de Quakebot, un programme informatique développé en mars 2014 par Ken Schwencke pour le *Los Angeles Times*, et qui a pour objet d'alerter lorsqu'un tremblement de terre se produit dans la région. Si l'automatisation peut faire craindre des pertes d'emplois dans un secteur économiquement fragile, on constate également que le phénomène amène de nouvelles formes de travail (et de profils professionnels spécialisés). Celles-ci consistent à s'assurer de la qualité des

¹⁷ " Des robots au "monde" pendant les élections départementales? oui. . . et non", Luc Bronner, *Le Monde*, 23/03/2015, URL : <http://makingof.blog.lemonde.fr/2015/03/23/des-robots-au-monde-pendant-les-elections-departementales-oui-et-non/>

données qui alimentent les systèmes d'information, contrôler la qualité des productions automatisées, ou définir des modèles de textes standardisés en vue de leur automatisation (Diakopoulos 2019).

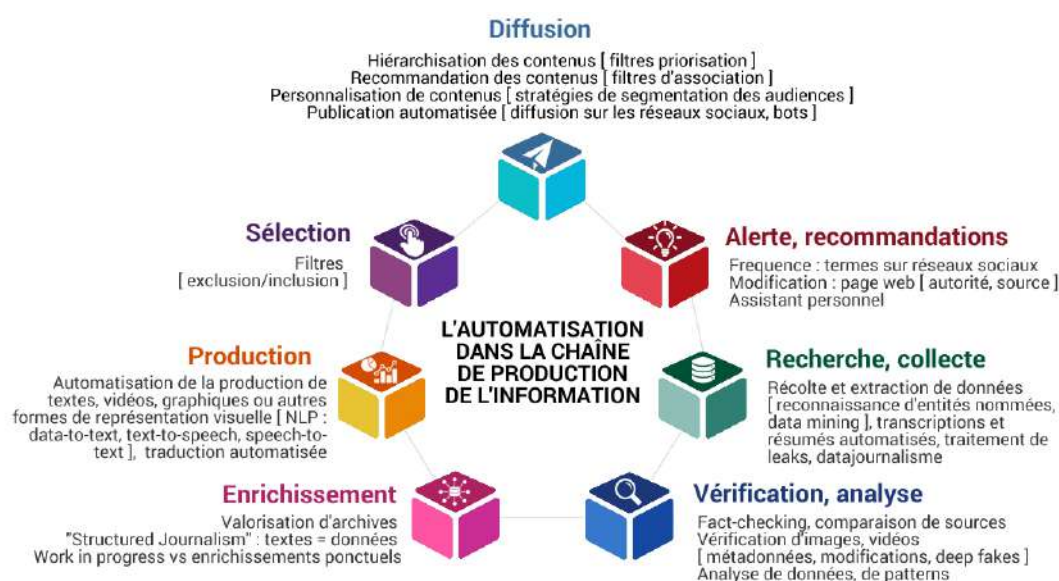


FIGURE 1.3 – L'automatisation dans la chaîne de production journalistique

L'ensemble des activités de la chaîne de production de l'information pourrait être entièrement automatisée, mais en théorie seulement. Leur point commun : s'appuyer sur des données, parfois disponibles en très grandes quantités. En pratique, de nombreux projets de datajournalisme incluent désormais l'automatisation¹⁸, la crise du Covid-19 ayant également inspiré de nombreux projets (Danzon-Chambeau, 2021)¹⁹. Ces projets reposent sur des algorithmes de machine learning (apprentissage par la machine) ou de deep learning (apprentissage profond). Il s'agit d'une approche informatique qui a pour objet de produire de la connaissance à partir de données observées, les algorithmes de l'IA se fondant sur des procédures probabilistes pour construire des modèles prédictifs (supervisés) ou descriptifs (non supervisés) (Lantz 2019). Selon le premier rapport Gartner sur les logiciels NLG publié en 2019, l'approche basée sur des systèmes à base de règles (qui imitent une forme rudimentaire d'intelligence dans un contexte limité grâce à une approche basée sur des modèles) continue de dominer malgré les promesses des techniques d'intelligence artificielle. Toutefois, ces deux types de technologie utilisent la connaissance pour générer du contenu (Reiter & Dale 1997). Ils sont également adaptés à leur domaine d'application car nous n'écrivons pas de la même manière sur les résultats sportifs que sur les marchés boursiers. Par conséquent, ce sont souvent des artefacts uniques et non reproductibles (Linden, 2017). Par ailleurs, les systèmes d'automatisation de la production automatisée d'information ne cessent d'évoluer. Avec le lancement de GPT-3 (Ge-

¹⁸ Lire aussi "Intelligence artificielle et journalisme : une course avec les machines", Laurence Dierickx, Equal Times, 2021, <https://urlz.fr/ghVf>

¹⁹ Voir "Covering COVID-19 with automated news", Tow Center for Digital Journalism https://www.cjr.org/tow_center_reports/covering-covid-automated-news.php

nerative Pre-trained Transformer 3), la société américaine OpenAI (co-fondée par Elon Musk en 2015) a élargi considérablement les horizons de la génération automatique de textes en langue naturelle. Il s'agit d'un modèle de langage autorégressif qui compte quelques 175 milliards de paramètres et est capable d'analyser 45 téraoctets de données stockées dans le cloud. Dans son corpus d'entraînement, on retrouve, entre autres, des articles de Wikipédia en anglais. GPT-3 est capable de traduire et d'écrire. Le système peut générer des échantillons d'articles de presse que des évaluateurs humains ont du mal à distinguer d'articles écrits par des journalistes. Début septembre 2020, le quotidien britannique The Guardian publiait un texte rédigé par GPT-3 : "Un robot a écrit entièrement cet article. Humains, avez-vous peur maintenant?"²⁰.

Pour une bibliographie complète à propos du journalisme algorithmique, rendez-vous sur le site du LaPIJ (Carnets n°2) : <https://lapij.ulb.ac.be/database/journalisme-algorithmique/>. Vous pouvez aussi le site <https://journodev.tech/>.

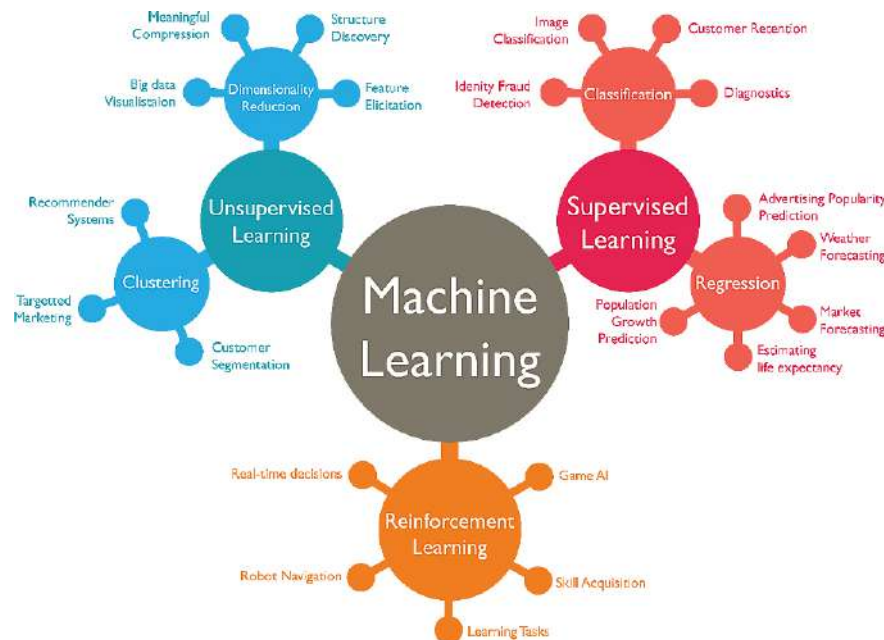


FIGURE 1.4 – Principales applications du machine learning. Source : "Qu'est-ce que le machine learning?", Medium, <https://urlz.fr/ghVB>

²⁰ Voir : <https://journodev.tech/eclairage-lecriture-humaine-depassee-par-lia/>

1.5 Typologie d'une approche par données dans le journalisme



L'approche par données dans le journalisme compte plusieurs avatars, lesquels vont du journalisme assisté par ordinateur à la rédaction automatique, en passant par le datajournalisme et le journalisme computationnel. Cette typologie est liée à l'histoire de cette pratique s'appuyant sur des données, l'usage de l'informatique et celui des outils et méthodes des sciences sociales. Coddington (2014) souligne que si les journalistes engagés dans les pratiques d'une approche par données semblent indifférents au fait de "classer" leur travail, il est utile pour le chercheur d'en dresser la typologie de manière à comprendre leurs formes et leurs implications dans les pratiques journalistiques. Si la frontière entre les genres peut s'avérer ténue, que ce soit dans leur définition ou dans le cadre des pratiques professionnelles, une typologie permet d'affiner ce qui les distingue. Dans celle qu'il propose, Coddington différencie le journalisme assisté par ordinateur du journalisme de données en termes de transparence, d'ouverture, et de connexion aux audiences avec lesquelles il entretient une relation active. Le journalisme computationnel serait plus opaque dans ses processus, tout en étant à même de gérer des volumes de données plus importants.

Les principales différences consistent dans l'enracinement du journalisme assisté par ordinateur aux méthodes issues des sciences sociales et son orientation vers le journalisme d'investigation; dans l'ouverture participative du journalisme de données; et dans les processus d'abstraction automatisés du journalisme computationnel. Ce qui distingue principalement le journalisme computationnel du journalisme de données consiste dans le fait que ce dernier repose principalement sur l'extraction d'informations pour produire des modèles informatisés. À l'inverse, le journalisme de données repose sur une analyse de données pour nourrir une information journalistique (Karlsen & Stavelin 2014, cité par Coddington 2014).

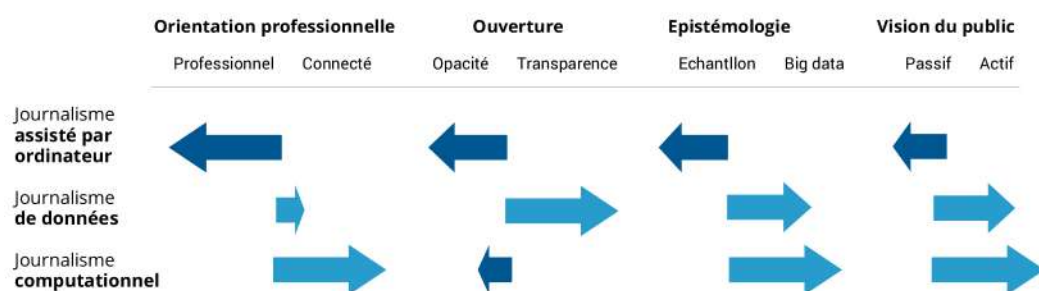
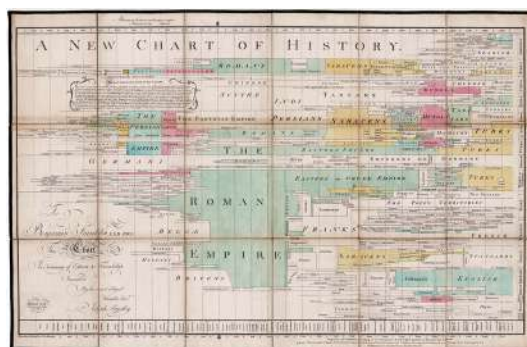


FIGURE 1.5 – Typologie visuelle d’une approche par données dans le journalisme (Coddington, 2015)

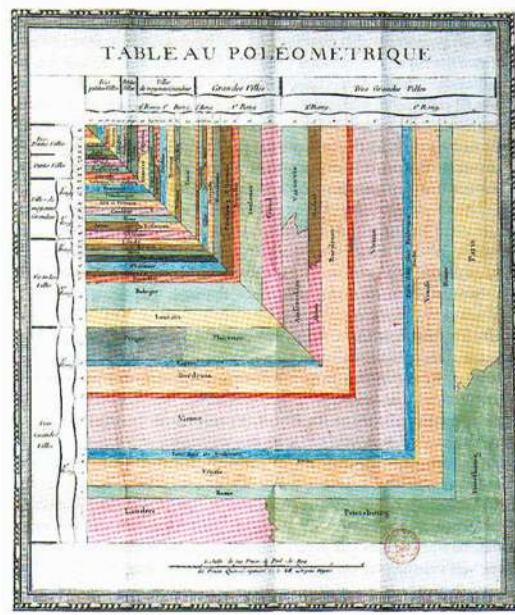
1.6 Aperçu historique de la visualisation de données

L’histoire de la représentation d’images mentales remonte à l’Antiquité – les Grecs avaient élaboré un système complexe de représentations mentales comme outils de mémoire, les Egyptiens cartographiaient les étoiles. Au Moyen-Âge, l’image comme "substitut du langage" se retrouve dans les vitraux et dans les manuscrits, où les enluminures servaient de point de repère dans le déroulement des textes. Certains manuscrits proposaient déjà une forme de "réalité augmentée" : les pointeurs représentés dans les marges seraient les ancêtres des pointeurs informatiques (flèche et main). D’autres annotations faisant référence à certaines sections du manuscrit se rapportent aux concepts actuels de navigation et d’hyperliens. En 1295, Raymond Lulle (France) préfigure les arbres de la connaissance en proposant un "arbre des vices et des vertus" (aussi appelé arbre des sciences). Un siècle plus tard, en 1370, Nicolas Oresme (France) représentera sous la forme graphique le rapport entre deux variables, préfigurant quant à lui les premiers graphiques en barres. Mais les visualisations de données quantifiées prendront leur essor avec le développement des statistiques, dès la fin du dix-huitième siècle.

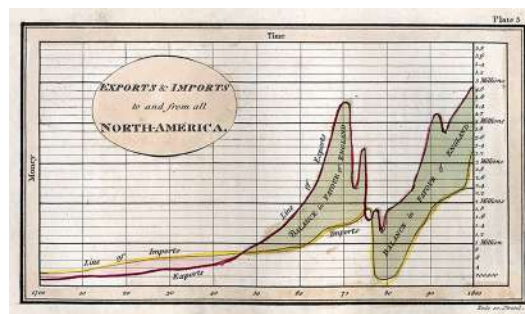
1765 Le britannique Joseph Priestley signe une première ligne du temps. Il en élaborera des dizaines tout au long de ses travaux de nature historique.



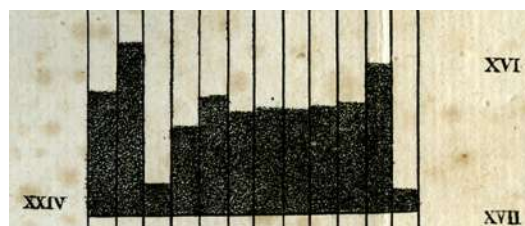
1782. Le mathématicien Charles de Fourcroy (FR) analyse la superficie de 200 villes.



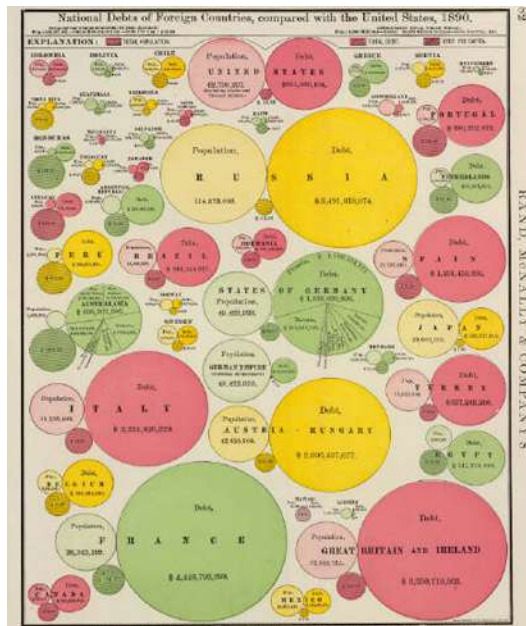
1786. William Playfair (RU) invente trois types de conception graphique : la série statistique sous forme de courbes, le graphique à barres et le graphique à secteurs.



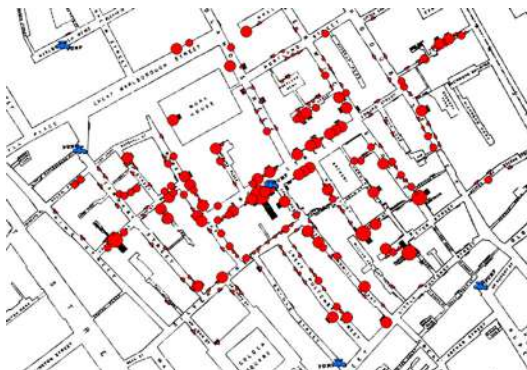
1829. André-Michel Guerry (FR) est à l'origine des premiers histogrammes. Il pratique aussi la visualisation de statistiques comparées (cartes).



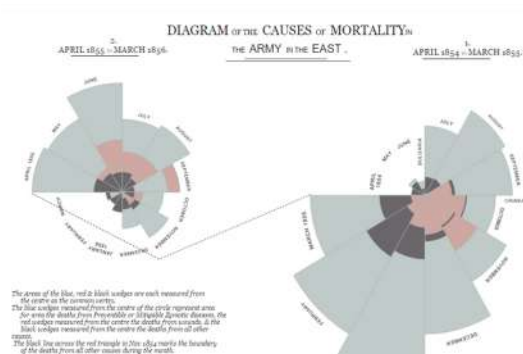
1830. Armand Joseph (FR), frère de Montizon, a l'idée d'une représentation par des points et des cercles.



1854. John Snow (RU), médecin, étudie les modes de propagation de l'épidémie de choléra à Londres.



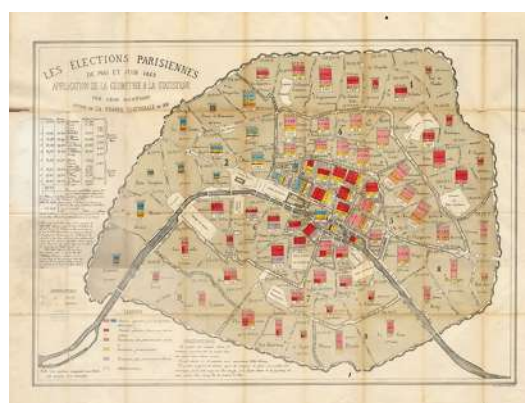
1857. Florence Nightingale (RU), pionnière de l'usage des statistiques dans le domaine de la santé, utilise des histogrammes circulaires pour illustrer les causes saisonnières de mortalité des patients de l'hôpital qu'elle gère. Elle réalisera aussi une étude statistique du système sanitaire en Inde, où elle contribuera à l'amélioration des soins médicaux.



1868. Emile Levasseur (FR), statisticien et géographe, 1868, travaille sur les premiers cartogrammes (statistiques figuratives).



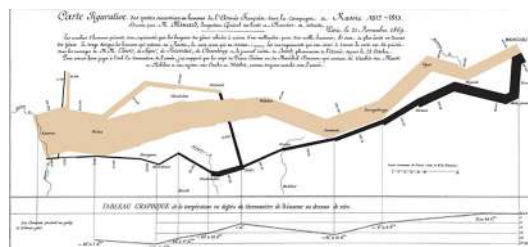
1869. Application de la géométrie à la statistique, par Léon Montigny (FR).



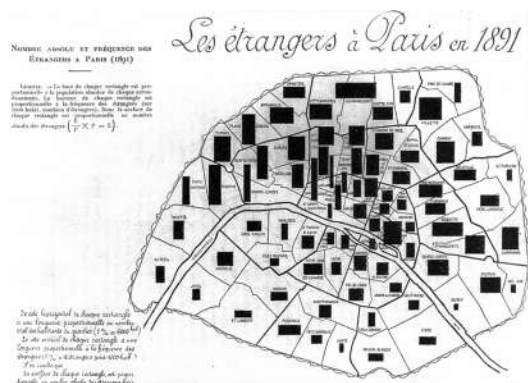
1874. Francis Walker (US), économiste, statisticien et personnalité académique, élabore les premières pyramides des âges. En 1881, il présida le Massachusetts Institute of Technology (MIT).



1889. Charles Joseph Minard (FR) établit une carte des pertes napoléoniennes lors de la campagne de Russie. Celle-ci comporte des informations sur le temps, la localisation, la distance, la température et le nombre de survivants.



1896. Jacques Bertillon (FR), statisticien et démographe, est à l'origine de nombreux travaux statistiques et cartographiques à propos de la ville de Paris.



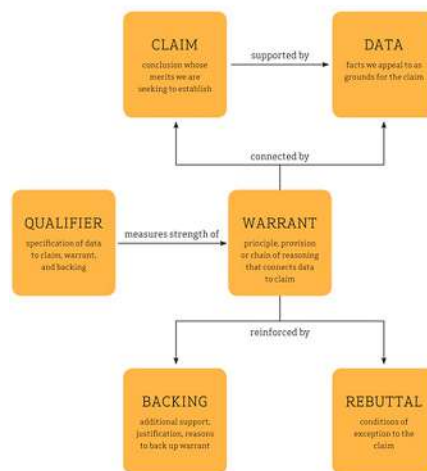
1914. Première édition de "Graphic Presentation", par Willard C. Brinton (USA). Cet ouvrage de 500 pages compile tous les types de graphiques et de techniques de présentation.

1920. L'Isotype (International System of Typographic Picture Education) est un langage visuel, à base de pictogrammes, développé par le philosophe autrichien Otto Neurath et le graphiste allemand Gerd Arntz. Naissance, en Allemagne, de la Gestalt ou philosophie de la forme. Le cer-

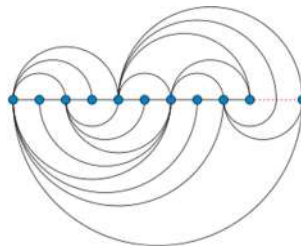
veau humain ne perçoit pas les taches de couleur et les formes comme des entités individuelles mais comme des agrégats. Aussi, les objets proches les uns des autres ont tendance à être perçus comme appartenant à un même groupe. Les lois de la Gestalt interagissent entre elles, de manière parfois contradictoire. Il s'agit des lois de la bonne forme (simple, symétrique), de la continuité et de la proximité, de la similitude, du destin commun et de la familiarité.

1934/1940. Le Cosmographe d'IBM permet de représenter de manière complexe des flux d'information dans le domaine des affaires (diagrammes de flux).

1958. Stephen Toulmin (R-U) propose un nouveau modèle graphique d'argumentation.



1964. Thomas Saaty (USA) invente le diagramme en arc.



1967. Le cartographe Jacques Bertin (FR) développe les règles de sémiologie graphique ("la graphique").

1977. John Tukey invente le diagramme en boîte (aussi appelé "boîte à moustaches") dans le cadre de la représentation graphique de données statistiques.

1985. Cleveland et MacGill cosignent un article sur la perception et les méthodes graphiques, dans le cadre de l'analyse scientifique. Ils constatent que le nombre de types de représentations graphiques ne cesse de croître rapidement.

Les représentations visuelles de données connaissent un important développement depuis les années 1980 – décennies de l'apparition de tableurs qui génèrent des graphiques automatiquement à partir de données encodées (Visical, Multiplan, Lotus, Excel). Les évolutions informatiques et technologiques vont de pair avec de nouvelles offres professionnelles en matière de traitement des données (Business Intelligence). En 1996, Ben Schneiderman propose une typologie de données induisant chacune des types de graphiques : les données temporelles, les données à une, deux ou trois dimensions, les données multidimensionnelles, les données hiérarchiques et les données interreliées.

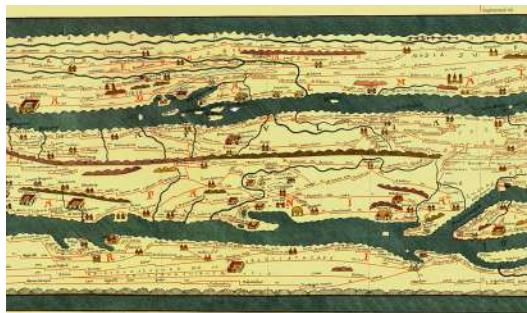
1.7 Aperçu historique de la représentation cartographique

L'histoire de la cartographie trouve ses origines dans l'Antiquité. Cette pratique – entre art et sciences - s'est développée à partir du 15^e siècle et a connu d'importantes évolutions au 19^e siècle, avec l'essor de la cartographie statistique : *"La carte est la représentation d'un espace. C'est la transcription dans une image de phénomènes localisés et des relations qui se développent entre ces phénomènes"* (Le Fur 2007). La carte n'est pas neutre : "elle résulte d'une série de choix et souvent exprime un point de vue". Une bonne carte *"demande le moins d'efforts dans un minimum de temps pour atteindre le but visé"* (Joly 1999). Elle apporte *"une réponse visuelle rapide et évidente. Une carte qui n'apporte pas de réponse visuelle instantanée est mal construite"* (Le Fur 2007).

Dans l'Antiquité, les premiers croquis cartographiques ont pour objectif la conservation de la mémoire des lieux et des itinéraires. Ils sont gravés sur des tablettes d'argile, à l'instar de la carte babylonienne du monde (circa 700-500 AJC). La carte d'Eratosthène est la première carte du "monde connu". Elle porte le nom du philosophe, mathématicien, poète et astronome grec Eratosthène de Cyrène (circa 276-194 AJC). Mais c'est la carte de Ptolémée (circa -150 AJC), du nom de l'astronome et astrologue grec Claude Ptolémée qui servira longtemps de référence.

La Table de Peutinger

Les empereurs romains établissent des cartes, qui consistent en des itinéraires routiers de l'empire à l'usage des armées. La Table de Peutinger (ou carte de Castorius), redécouverte en 1494, est une copie d'une carte romaine. Elle mesure 6 mètres de long et 30 cm de large.



Mapa Mundi, la carte du monde

La représentation du monde est en forme de T ("Carte en T") : la terre est circulaire et est symboliquement partagée en trois, évoquant la trinité catholique. Ci-contre, la Mapa Mundi du moine Beatus de Liébana (8e siècle). La Mapa Mundi d'Hereford (1280) est la plus ancienne carte imprimée du Moyen-Âge.



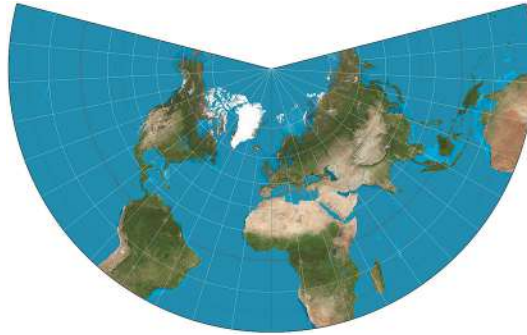
Les cartes marines

Si les premières cartes marines datent du 8e siècle, elles deviennent plus précises au 15e siècle, grâce à de nouveaux outils de mesure. C'est la période des grandes explorations. Ci-dessous, la projection de Mercator en 1569.



Les cartes détaillées et à grande échelle

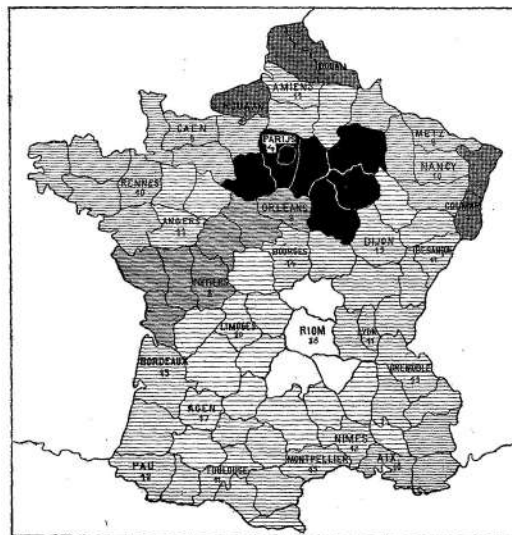
Les nécessités de la guerre et de l'administration exigent des cartes détaillées et à plus grande échelle. La carte de Cassini (ou carte de l'Académie) est la plus ancienne carte topographique de France (fin du 18e siècle). A la même époque, la projection de Lambert – du nom du mathématicien mulhousien Johann Heinrich Lambert - propose un système de projection conforme (les méridiens sont des droites et les parallèles des arcs de cercle).



La cartographie statistique

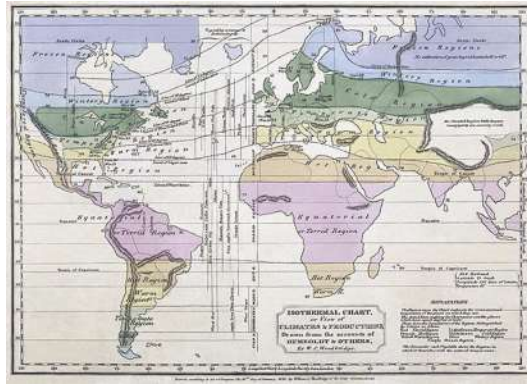
Avec l'émergence de la cartographie statistique, se développent de nouvelles formes de représentations. Les cartes servent à visualiser des données (liées aux populations) sur un territoire déterminé. La première carte démographique apparaît en Prusse en 1828. La première carte statistique moderne – sur l'instruction populaire en France – est présentée deux ans plus tôt, en France, par Charles Dupin (ci-dessous, une de ses premières cartes choroplèthes).

De 1845 à 1870, l'ingénieur français Charles Joseph Minard réalise une série de cartes figuratives dans lesquelles il s'essaie à l'adaptation de nombreux procédés graphiques. Les codes de la cartographie statistiques quantitative comprennent les cartes choroplèthes (cartes thématiques sur lesquelles les régions apparaissent en couleur), les cartes par points, les cartes isoplèthes (zones séparées par des lignes et points) et les cartogrammes.



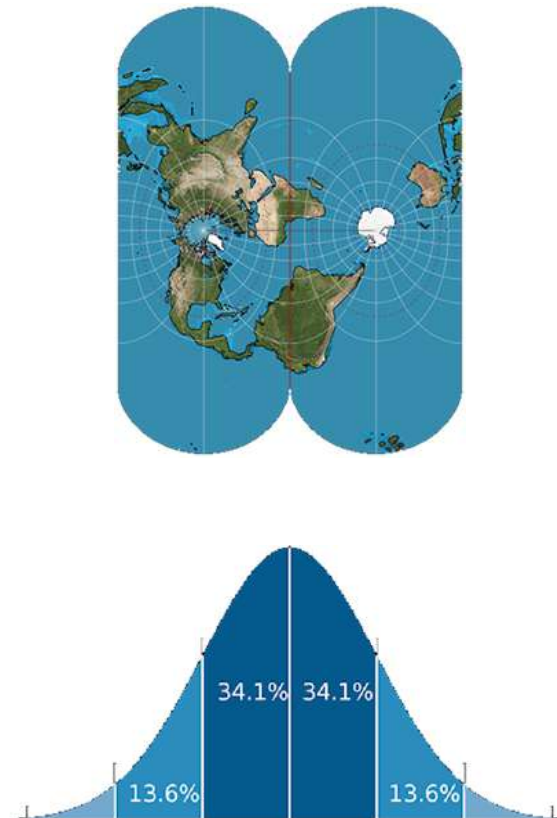
La carte climatique

La première carte climatique est présentée en 1817 par le naturaliste allemand Alexander von Humboldt. Il utilise des courbes isothermes pour indiquer les variations de température.



La courbe en cloche

Autre révolution du 19e siècle : la courbe en cloche du mathématicien allemand Carl Friedrich Gauss, est appliquée aux cartes – la terre y apparaît comme un ellipsoïde aplati aux pôles. Elle a donné son nom à la courbe de Gauss, utilisée en statistiques pour représenter une distribution (densité d'une mesure d'une série).



Théorie de la sémiologie graphique

Tandis que le développement de la photographie aérienne participe aux progrès enregistrés dans la réalisation de cartes, la pratique de la cartographie statistique se développe. En 1967, le cartographe français Jacques Bertin publie "La sémiologie graphique" (que l'on appellera plus tard "la graphique"), qui consiste en un catalogue de procédés graphiques organisés en sept variables rétinienne (ou visuelles) : orientation, forme, couleur, grain, valeur, taille et dimension.

Cartographie assistée par ordinateur

A la même époque, la cartographie assistée par ordinateur fait ses débuts. Selon Joly (1999), elle doit être considérée *"comme un maillon d'une chaîne continue d'opérations qui, partant d'une récolte de données, se poursuit par un traitement statistique ou mathématique (...) et aboutit à la visualisation et/ou à la mémorisation sous forme cartographique des résultats obtenus"*. Avec l'introduction de l'informatique, constate Le Fur (2007), *"la carte n'est plus un but mais un outil d'exploration des hypothèses et des données"*.

Mashups cartographiques

Avec le développement d'internet et des technologies de l'information, la représentation cartographique devient interactive. La première expérience de mashup cartographique, en datajournalisme, est menée en 2005 par le journaliste-développeur américain Adrian Holovaty, qui cartographie le crime à Chicago à l'aide du service Google Maps. Un mashup cartographique est *"un produit obtenu en prenant des données de géolocalisation, comme des adresses et des coordonnées, sur une carte et en les organisant par catégorie ou type d'information"* (Briggs 2014).

Un mashup cartographique est *"un produit obtenu en prenant des données de géolocalisation, comme des adresses et des coordonnées, sur une carte et en les organisant par catégorie ou type d'information"* (Briggs 2014)[20]. Il s'agit, pour Holovaty (à qui l'on attribue la paternité du terme "datajournalisme"), d'un second projet s'appuyant sur une exploration interactive de données : le premier s'appelait *Every Block* un projet de datajournalisme hyperlocal qui reposait sur l'extraction de données provenant de sites web institutionnels et de fichiers transmis par les autorités pour informer les habitants de Chicago à l'échelle de leur quartier (Parasie 2013)[35].

Tracking homicides in Chicago

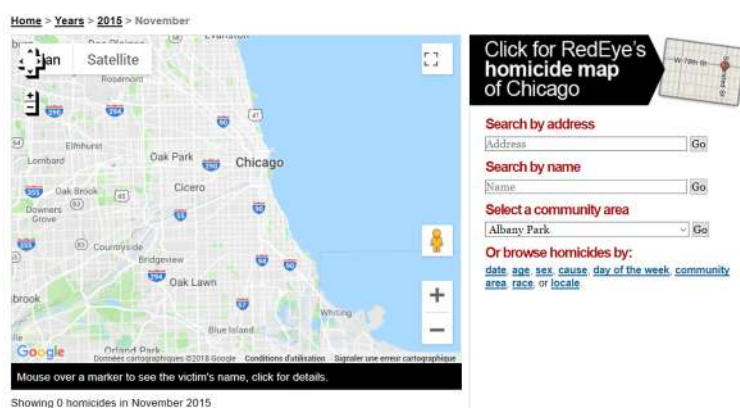


FIGURE 1.6 – Source : <http://homicides.redeyechicago.com/neighborhood/bridgeport/>

L'interface du projet "Chicagocrime" se compose d'une carte pouvant être explorée par l'utilisateur. Chaque "marqueur" fournit une information de contexte au lecteur²¹. Cette initiative était inspirée par la carte du crime de Los Angeles, initiative du journaliste-développeur Ken

²¹ Voir : <http://www.holovaty.com/writing/chicagocrime.org-tribute/>

Schwencke. Partiellement automatisée, cette application a pour objet de fournir aux journalistes une information de base, qu'ils traiteront plus en détail, de même qu'une information servicielle destinée directement aux lecteurs. Chaque zone renvoie vers une nouvelle page qui y est dédiée. Elle se présente sous la forme d'une ligne du temps interactive et comprend un texte court destiné à mettre les données en contexte.

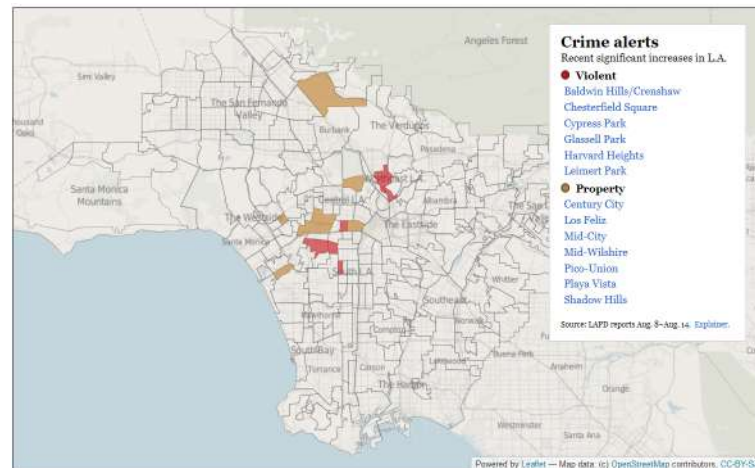


FIGURE 1.7 – Source : <http://maps.latimes.com/crime/>

Le récit cartographique peut être utilisé dans la presse locale, comme c'est régulièrement le cas à *L'Avenir* pour servir le propos d'une information servicielle (par exemple, la carte des zones de baignade en Wallonie). Mais elle peut aussi servir de base à une narration plus large s'appuyant sur des données géographiques. La cartographie peut également servir de fil conducteur à un récit interactif devant alors moins un outil de narration qu'un outil de représentation. Le KnightLab de la Northwestern University a développé, dans ce sens, l'outil StoryMap²².



FIGURE 1.8 – Source : <http://www.levif.be/actualite/international/les-attaques-de-daesh-en-europe-carte-interactive/game-normal-633971.html>

²² Voir <https://knightlab.northwestern.edu/projects/>

2 | Contexte du datajournalisme

Si elles prennent souvent la forme de représentations ou d'applications interactives, les expériences en matière de journalisme de données n'y sont pas cantonnées. Elles peuvent être envisagées comme un soutien à l'enquête journalistique et/ou comme un moyen de mobiliser de ressources et/ou des compétences dans le cadre de projets collaboratifs. Il serait impossible de répertorier toutes les expériences en matière de journalisme de données, tant elles sont nombreuses et variées. A titre d'exemple, le site dédié aux Data Journalism Awards, le premier concours international de datajournalisme initié par le Global Editors News Network en 2014, recensait plus de 2.000 projets portés par plus de 5.000 professionnels de 86 pays ¹.

2.1 Au service de l'enquête journalistique

De grandes enquêtes journalistiques collaboratives ont contribué à la popularisation du phénomène. Les projets ayant contribué à donner au genre ses lettres de noblesse ont été réalisés à partir de la fin des années 2000. Quatre d'entre eux sont particulièrement emblématiques de cette histoire récente.

2.1.1 MP's expenses

En 2009, le quotidien britannique *The Guardian* fait appel à ses lecteurs, dans une opération de crowdsourcing, pour examiner en ligne 460.000 notes de frais de parlementaires. 170.000 de ces notes de frais, transposées dans des logiciels tableurs (Excel et Google Spreadsheet), ont été passées en revue par 26.774 internautes.

La publication de ces résultats a donné lieu à une enquête du gouvernement sur les remboursements des députés, concluant que plus d'un million de livres sterling avaient été indûment payés. Quatre députés seront jugés pour fraude. Les bénéfices de cette opération furent multiples pour le quotidien, tant sur le plan de sa réputation que sur celui d'avoir pu gagner un temps non négligeable dans l'examen de ces notes de frais, grâce à la participation des lecteurs du journal. L'opération a également mobilisé des moyens importants sur le plan technique – un premier paquet de notes de frais avaient fait l'objet d'un tri préalable – mais n'aura nécessité qu'un faible investissement : 50 livres sterling, soit le prix d'un serveur web temporaire (Flew & al. 2010).

¹ URL : <http://community.globaleditorsnetwork.org/projects>

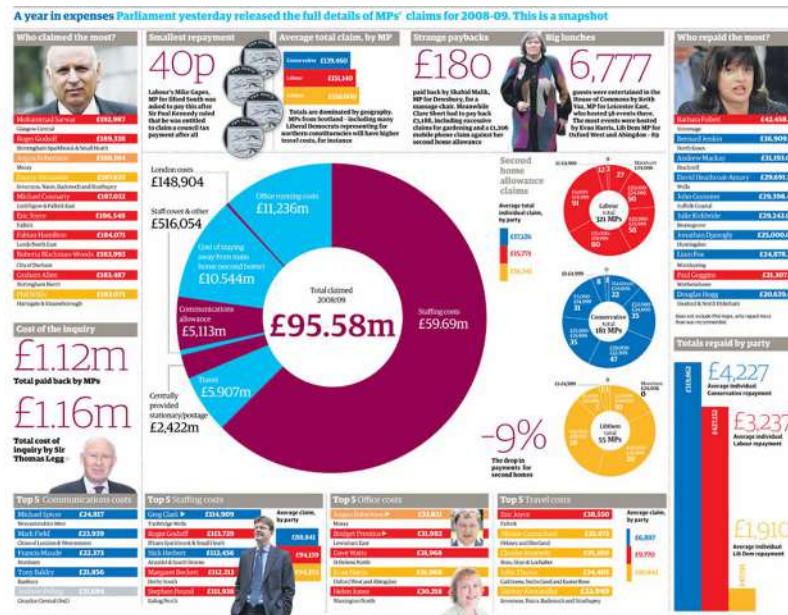


FIGURE 2.1 – Visualisation de données réalisée par *The Guardian* dans le traitement des données relatives aux dépenses des parlementaires Britanniques

2.1.2 WikiLeaks

En 2010 et 2011, WikiLeaks publie des dizaines de milliers de documents secrets sur les guerres en Afghanistan et en Irak. Derrière cette démarche, relève Dagiral & Parasie (2013), "*l'idée selon laquelle les journalistes traditionnels ne peuvent garantir au lanceur d'alerte que son identité ne sera pas révélée et que ses informations seront réellement diffusées*". Les médias s'empareront du sujet, à travers une première collaboration inédite entre WikiLeaks et cinq rédactions (*The Guardian*, *Der Spiegel*, *El Pais*, *Le Monde* et *The New York Times*). Ces masses de données ont fait l'objet de discussions internes à chaque média, décidant ce qu'il y avait lieu de publier et, le cas échéant, d'anonymiser pour protéger les noms de personnes pouvant être mise en danger (Howard 2014). Elles ont également été traitées sous la forme de visualisations et de cartes interactives.

Sur le plan technique, il a fallu mettre au point des solutions pour traiter des millions de lignes de données fournies de manière désordonnée dans des formats variés – soit un volume de 1.7 gigabytes de données – (Hernando 2017) – soulignant ainsi l'importance de l'input humain (Howard 2014, Gynnild 2013). Au *Monde*, la conception de l'interface de visualisation avait été externalisée auprès de petites sociétés innovantes (Linkfluence, spécialisée dans l'analyse sociale du web, et 22 Mars, editrice du *pure player* Owni.fr² - Trédan 2014). A l'inverse, *The Guardian* avait traité ces aspects en interne³.

² Lancé le 6 avril 2009 par la société 22 Mars, le Owni.fr était un projet de datajournalisme s'appuyant sur de l'open data. Le site, lauréat de deux Online Journalism Awards, a également collaboré avec WikiLeaks. Il fut placé en liquidation judiciaire le 21 décembre 2012, faute d'avoir pu trouver un modèle économique viable.

³ "WikiLeaks data journalism : how we handled the data", Guardian Data Blog, Simon Roger, 31/01/2011, URL : <https://www.theguardian.com/news/datablog/2011/jan/31/wikileaks-data-journalism>

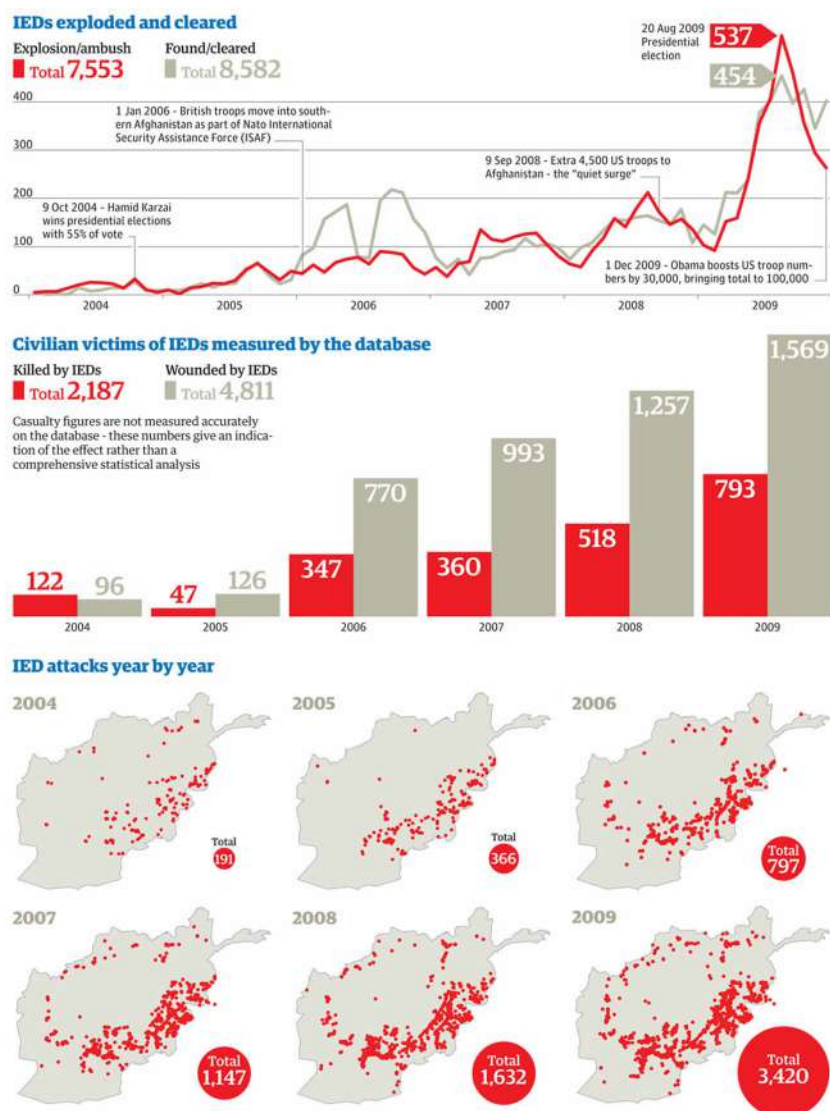


FIGURE 2.2 – Visualisation de données réalisée par *The Guardian* dans le traitement des données de WikiLeaks

2.1.3 The Migrant Files

Le projet "Migrant Files" avait pour double objectifs de commenter les flux migratoires vers l'Europe et de documenter le nombre de décès liés à ce phénomène. Il impliquait un consortium de journalistes issus de dix pays européens. Au-delà de la visualisation d'une carte interactive, réalisée à l'aide du SaaS (*software as a service*, un logiciel installé sur un serveur web) CARTO, l'expérience témoigne d'un input humain importante en matière de collecte d'informations. "Jusqu'alors, quasiment personne ne comptait ces décès. Une personne de l'administration française nous a même dit que ce n'était pas son problème car "une fois mort, ce n'est plus un migrant". En créant cette base de données et en la mettant à disposition dans un format facile à réutiliser, *The Migrants Files* a permis, à une échelle très modeste, d'influencer le regard des institutions sur le sujet. Lorsqu'un bateau faisait naufrage ou qu'un homme se faisait abattre par des gardes-frontières, c'était une anecdote. Avec des statistiques, ces morts peuvent être contextualisées.

sés. Ils ne sont plus des anecdotes, mais font partie d'un tout concret et mesurable. Et c'est parce qu'il est mesurable qu'il peut être l'objet d'une politique publique", comment Nicolas-Keyser Bril, l'un des journaliste ayant initié l'expérience⁴.

Faute de ressources financières suffisantes pour assurer sa pérennité, le projet a pris fin en 2016, au terme de trois années d'existence. Il avait récolté des bourses pour un montant total de 17.000 euros et avait été récompensé par deux prix de journalisme (Data Journalism Awards en 2014 et European press prize en 2015)⁵.



FIGURE 2.3 – Interface de l'application "The Migrant Files"

2.1.4 The Panama Papers

Les Panama Papers constituent, à bien des égards, une expérience inédite, tant par le volume de documents concernés soumis à l'analyse des journalistes, que par le nombre de professionnels mobilisés dans le cadre d'une fuite de documents sans précédents. 11,5 millions de documents numériques, d'un poids de 2.6 téraoctets, ont fait l'objet d'un travail d'enquête effectué par des centaines de journalistes provenant de plus de cent titres de presse différents à travers le monde, sous la coordination du Consortium international des journalistes d'investigation (ICIJ). Cette enquête transnationale a démontré comment des personnes évitaient de payer les taxes dans leurs pays respectifs en créant une société offshore à l'étranger (Iftikhar 2016, Heft & al. 2017).

Cette enquête de grande ampleur a été récompensée par un Prix Pulitzer, et un par un Data Journalism Awards. L'enquête a duré un an avant la révélation de ses résultats, en avril 2016. Pour Schwickerath & al. (2017), l'histoire des Panama Papers se lit comme le scénario d'un film d'espionnage, débutant en 2014 lorsqu'une source anonyme livre des documents provenant du cabinet d'avocats panaméen Mossack Fonseca au journal allemand *Süddeutsche Zeitung*.

⁴ "Se libérer par les données", Nicolas-Keyser Bril, 25/09/2017, URL : <http://blog.nkb.fr/se-liberer-par-les-donnees>

⁵ Source : <http://www.themigrantsfiles.com/>

Cette enquête de grande ampleur relève également de cette tradition du journalisme considérant son rôle comme celui de "chien de garde" de la démocratie.

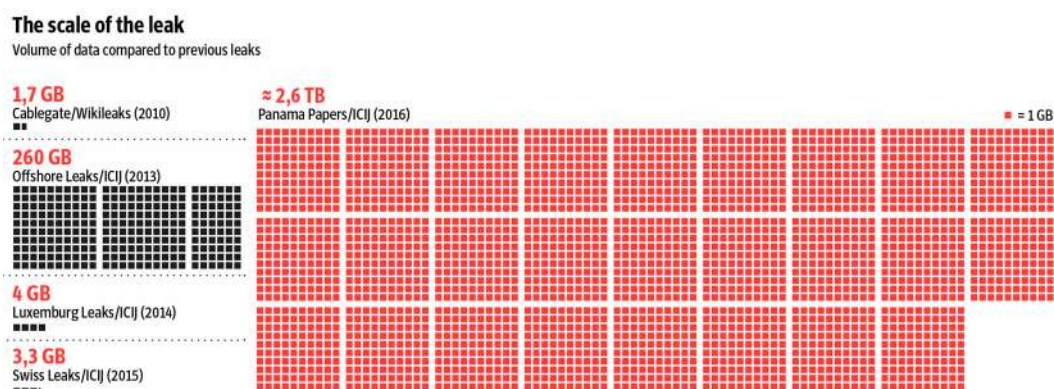


FIGURE 2.4 – Représentation du volume de données traitées dans les "Panama Papers" en comparaison avec les grandes "fuites" (*leaks*) ayant mobilisé un journalisme d'investigation collaboratif

L'utilisation de technologies informatiques fut cruciale dans le traitement de ces documents, notamment avec le logiciel Nuxi Investigator, utilisé principalement dans le cadre de l'investigation médico-légale, qui convertit des fichiers PDF et des images en informations "lisibles". Une base de données graphique (Neo4j) a également été constituée pour rendre accessible ces documents. *"Contrairement aux bases de données traditionnelles qui montrent des données dans une structure en forme de feuille, les bases de données graphiques permettent aux utilisateurs de visualiser les données en tant que réseau de nœuds et de connexions. Cela a mis en lumière les liens cachés entre les comptes et leurs propriétaires"* (Aguirre Hernando 2017). Mais ce ne fut pas le seul outil utilisé : l'ICIJ, qui présente sur son site les coulisses de l'enquête, explique que Apache Solr fut utilisé pour l'indexation des documents; Apache Tika, pour le traitement des documents; Tesseract, pour l'OCR-isation (laquelle a mobilisé de 30 à 40 serveurs temporaires); Project Blacklight, généralement utilisée par les bibliothécaires, pour l'interface utilisateur comprenant une fonction de recherche avancée par "facettes"; Linkurious, pour la visualisation des données; Talend, un logiciel d'extraction de données SQL; et Cypher, pour effectuer des requêtes plus complexes. La communication entre les journalistes a été assurée par la plateforme Global I-Hub⁶. Le traitement des Panama Papers a donc représenté un double défi : à la fois journalistique et technologique. Son impact fut également sans précédents. En Belgique, le fisc belge a pu récupérer plus de 8 millions d'euros suite à ces révélations⁷. L'affaire a également donné lieu à la création d'une Commission spéciale Panama Papers à la Chambre des représentants dont l'objet est d'établir des mesures politiques pour lutter contre la fraude fiscale.

⁶ "Wrangling 2.6TB of data : The people and the technology behind the Panama Papers", ICIJ, Mar Cabra and Erin Kissane, 25/04/2016, URL : <https://panamapapers.icij.org/blog/20160425-data-tech-team-ICIJ.html>

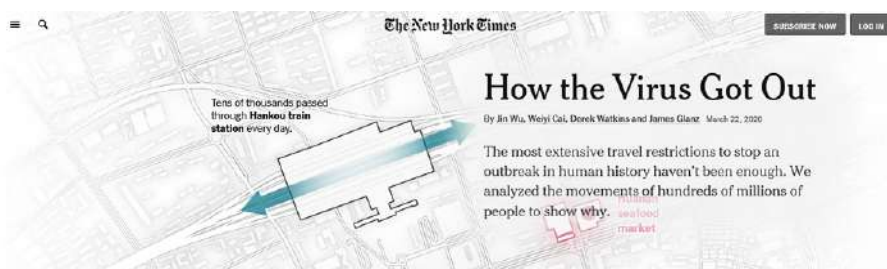
⁷ "Panama Papers : le fisc a récupéré plus de 8 millions d'euros", La Libre Belgique, agence Belga, 30/08/2017, URL : <http://www.lalibre.be/economie/conjoncture/panama-papers-le-fisc-a-recupere-plus-de-8-millions-d-euros-59a63e05cd706e263fad5bc0>

2.1.5 Une femme tous les trois jours

Le service "Interactive" de l'AFP publie régulièrement des longs formats s'appuyant sur des données⁸. Dans "Une femme tous les trois jours", un récit s'appuyant sur des données que des dizaines de journalistes ont récoltées et recoupées pour quantifier le phénomène des féminicides en France mais aussi pour l'expliquer au-delà des chiffres. Comme dans de nombreux projets de datajournalisme s'inscrivant dans le temps long de l'enquête, la méthodologie de collecte et de vérification des données est expliquée à la fin du long format. Voir https://interactive.afp.com/features/Une-femme-tous-les-trois-jours_597/

2.1.6 How the virus got out

Parmi les projets les plus récents, figure également le travail de la rédaction du New York Times, pour expliquer comment le Covid-19 s'est répandu depuis la Chine malgré les restrictions de voyage. Le récit a la particularité de mettre en avant les visualisations de données plutôt que le texte, et celles-ci proviennent d'une multiplicité de sources. Parmi celles-ci, figurent les données de Baidu et de deux sociétés de télécommunication chinoises, qui ont suivi le mouvement des téléphones portables. Celles-ci sont mentionnées et commentées à la fin du récit.



Par ailleurs, le site "Datajournalism.com" publie régulièrement des digests d'enquêtes journalistiques s'appuyant sur des données. Voir <https://datajournalism.com/read/longreads/covid-19-data-journalism>

2.2 Dans les pratiques quotidiennes

À côté de ces expériences qui mobilisèrent des moyens humains plus ou moins importants, le journalisme de données s'est installé de manière moins spectaculaire dans les pratiques journalistiques. Il est tantôt utilisé comme point de départ à un récit médiatique, tantôt comme objet d'exploration d'une information spécialisée ou hyper locale. De l'infographie commentée à des applications interactives complexes, en passant par un récit non illustré mais bel et bien piloté par une analyse de données, le datajournalisme est protéiforme. Si dans l'histoire de la profession, les journalistes ne se sont pas massivement appuyés sur une approche quantitative dans le cadre de leur travail (Linden 2017b), l'accès à des logiciels gratuits, à l'utilisation facile et intuitive, a ouvert de nouvelles opportunités.

⁸ <https://interactive.afp.com/>

Plusieurs grands médias – et plus seulement américains – disposent aujourd’hui de rédactions ou services dédiés au storytelling numérique. Certaines ont fait d’une approche par données leur spécialité. Journalistes et développeurs y sont amenés à travailler main dans la main, posant également la question des compétences que devraient avoir les journalistes attachés à ces services. Dans le monde anglo-saxon, les exigences sont élevées. Par exemple, une offre d’emploi publiée en 2015, *The Guardian* décrivait comme suit le profil d’un datajournaliste : compétences en matière de tableurs et de logiciels de bases de données tels que MS Excel et MS Access; expérience avec des bases de données telles que SQL Server, MySQL, Oracle ou PostgreSQL; connaissance basique des statistiques et des logiciels de statistiques tels que SPSS, R ou Stata; familiarité de base avec les technologies web HTML, CSS, Javascript, Python, Ruby, etc. avec le désir d’apprendre le code, une expérience avec au moins un langage de programmation constituant un avantage.

Un offre d’emploi a été publiée en 2017 pour le recrutement d’étudiants en datajournalisme pour *The Telegraph*. Elle indique que les candidats intéressés seront amenés à développer la capacité de nettoyer et d’analyser des données pour trouver des histoires intéressantes, et qu’ils doivent développer une connaissance des outils de collecte et de scraping de données ainsi qu’en matière de datavisualisation. Ces compétences font partie du programme enseignés aux candidats dans le cadre de leur master⁹.

Le datajournaliste "type" est un homme (57.5%), salarié à temps plein (64%), universitaire (96%), diplômé en communication ou journalisme (62%). Il produit essentiellement des contenus pour le web (77%) et compte de une à dix années d’expérience (78%), selon les résultats d’une enquête en ligne menée auprès de 181 datajournalistes de 43 pays¹⁰ (Heravi 2017). 86% des répondants ont déclaré se considérer comme des datajournalistes, 44% d’entre eux ont dit disposer de compétences à un niveau "intermédiaire" et 18% à un niveau "expert". L’auteur de l’étude indique à ce propos que les compétences liées aux aspects techniques en lien avec les données (analyse, statistiques, codage, ...) sont faibles : la moitié des répondants n’a pas été spécifiquement formée à la pratique du datajournalisme. Ces derniers estiment qu’il s’agit d’une pratique journalistique plus rigoureuse (69%), qui améliore la qualité du travail (91%) sans porter atteintes aux valeurs de la profession (83%).

Google News Lab et PolicyViz ont conduit une enquête à plus large échelle, dont la méthodologie s’appuie sur 46 entretiens exploratoires approfondis avec des journalistes et des éditeurs, et sur un questionnaire en ligne auquel ont répondu plus de 900 journalistes. Il en ressort que le datajournalisme est une pratique de plus en plus répandue, laquelle concerne quatre journalistes sur dix (42%). Mais elle fait aussi appel à des compétences spécialisées liées à la manipulation de données (analyse, traitement, nettoyage, visualisation...) qui ne sont pas facilement négociées par les journalistes, indiquent 53% des répondants. A ce premier obstacle s’ajoute

⁹ "Wanted : MA Data Journalism applicants to partner with The Telegraph", Online Journalism Blog, 19/10/2017, URL : <https://onlinejournalismblog.com/2017/10/19/wanted-ma-data-journalism-applicants-to-partner-with-the-telegraph/>

¹⁰ Principalement en Europe et aux États-Unis

celui de la pression du temps – 44% des répondants disent avoir besoin d'une semaine ou plus pour traiter un sujet – et celui d'un retour sur investissement qui ne va pas de soi ¹¹.

La pratique du journalisme de données, considérée comme une pratique d'excellence dans le journalisme (Craig & al. 2017), a peu évolué au cours de ces dernières années si l'on observe les projets de datajournalisme présentés aux Data Journalism Awards, créés en 2014 par le Global Editors News Networks et qui récompense chaque année des projets de datajournalisme du monde entier (suite à la fin des activités du GENN, ce prix de datajournalisme est devenu le "Sigma Awards" en 201) ¹². La révolution attendue n'aurait donc pas eu lieu et les pratiques en seraient au *status quo* (Loosen & al. 2017) : "*(le journalisme de données) couvre toujours la politique, il exige beaucoup de main d'œuvre et nécessite de grandes équipes, il est surtout pratiqué en presse écrite et utilise surtout des données publiques pré-traitées.*" La matérialité de son expression – visualisations de données et systèmes interactifs – n'a pas non plus évolué vers plus de sophistication. Les chercheurs estiment que le journalisme de données complète le journalisme traditionnel mais ne le remplace pas à grande échelle.

2.2.1 Pratiques du datajournalisme en Belgique et en France

Les pratiques du datajournalisme trouvent de plus d'échos au sein de la presse belge francophone, bien que cela reste encore modeste. La crise du Covid-19 a encouragé une pratique accrue d'outils de visualisation de données, qui sont le plus souvent des outils prêts à l'emploi (ils ne requièrent donc aucune expérience particulière en codage). Les journalistes présentant une filiation avec le journalisme de données sont loin de l'idéal de spécialisation porté dans l'espace anglo-saxon où l'on fait appel à des profils de plus en plus pointus, bien qu'il y soit longtemps resté activité de niche soutenue dans une poignée de rédactions, y compris jusqu'au milieu des années 2000 "Journalism by numbers", Emily Bell, Columbia Journalism Review, 05/09/2012, consulté le 09/09/2017, URL : http://www.cjr.org/cover_story/journalism_by_numbers.php. En Belgique francophone, il s'agit aussi d'une activité de niche, pratiquée essentiellement par quelques autodidactes passionnés. Comment expliquer cette situation? Par un manque d'intérêt des journalistes ou de formations spécialisées? Par des difficultés d'accès aux données? Ou par un manque de retour sur investissement pour des projets mobilisateurs de ressources? Chacun de ces arguments participe à expliquer cette situation de *statu quo*. Un récent intérêt marqué par plusieurs écoles de journalisme gagne également à populariser ces pratiques auprès des futurs journalistes.

On retrouve des traces de la matérialité des démarches relevant du journalisme de données, en Belgique, dans la base de données des *Data Journalism Awards* (Sigma Awards), le seul prix international de journalisme dédié à cette approche spécifique. Sur les 2.439 projets répertoriés en 2018, 25 sont Belges et la majorité d'entre eux émane de journalistes néerlandophones. A titre de comparaison, 86 projets sont Français, 43 sont Néerlandais, 106 sont Allemands, 263

¹¹ "The state of data journalism in 2017", Simon Rogers, Google, 18/09/2017, URL : <https://www.blog.google/topics/journalism-news/data-journalism-2017/>

¹² Voir le site : <https://sigmaawards.org/>

sont Britanniques et 507 sont Américains sur un total de 86 pays représentés.

L'existence du journalisme de données en Belgique francophone est donc longtemps restée moins matérielle que discursive. C'était d'ailleurs la conclusion de la seule recherche consacrée au phénomène, "*Waiting for datajournalism*" (De Maeyer & al. 2015). Celle-ci s'était appuyée sur vingt entretiens semi-structurés avec une diversité d'acteurs impliqués dans l'adoption et le développement du journalisme de données en Belgique francophone, en ce compris des responsables de la gestion des ressources humaines de rédactions et des personnes en charge d'organiser la formation des journalistes. La méthode avait été complétée par un corpus de 52 documents récoltés sur une période de deux ans, lesquels étaient relatifs à des programmes de formation, des articles de blogs et à un artefact "occasionnel" du journalisme de données. Les chercheurs avaient également participé à des événements relatifs à cette pratique, telles que des réunions du groupe belge de l'organisation internationale "Hacks/Hackers"¹³ et à des ateliers de journalisme de données, consistant en des observations informelles. Les résultats avaient démontré que la représentation du journalisme de données n'est pas partagée par l'ensemble des répondants, dont certains considéraient que la production d'une visualisation de données, c'est déjà du datajournalisme¹⁴. Parfois initiés aux pratiques du datajournalisme lors de formations de courte durée, le profil de la poignée de datajournalistes belges francophones est plutôt celui de journalistes formés sur le tas, présentant un intérêt personnel pour une approche par données.

Dans les rédactions, *Le Soir* dispose d'une équipe de trois journalistes travaillant notamment dans le cadre des grandes enquêtes collaboratives internationales de l'ICIJ (International Consortium of Investigative Journalim). A *L'Avenir*, la pratique du datajournalisme est intégrée dans les activités du "weblab", composé d'un journaliste et d'un développeur. Il s'agit là d'une approche essentiellement centrée sur un service aux lecteurs (par exemple, en proposant une carte interactive des eaux de baignade en Wallonie). A *L'Echo*, une cellule multimédia multidisciplinaire compte un journaliste de données. Cette cellule multimédia concentre ses activités, d'une manière plus large, sur le storytelling numérique. Le datajournalisme se pratique également de plus en plus au sein des différentes rédactions du pays et de plus en plus de journalistes se revendiquent "datajournalistes".

En Flandre, les pratiques du datajournalisme sont davantage encouragées que dans la partie francophone du pays, que ce soit en matière de formation continuée ou de soutiens financiers publics. Le journaliste Maarten Lambrechts fait figure de pionnier dans ce paysage médiatique. Se définissant lui-même comme journaliste de données, il a fait ses premières armes au quotidien *De Tijd*. Depuis le début 2017, il exerce cette activité, qu'il a élargie à celle de consultant en datavisualisations, en tant qu'indépendant : "*Aujourd'hui, je réalise toutes sortes de travaux ba-*

¹³ "Hacks and hackers est organisation américaine transnationale, qui a pour objet de réunir des journalistes et des programmeurs en vue de susciter des collaborations (Lewis & Usher 2014).

¹⁴ Cette vision réductrice pose raisonnablement la question de savoir si l'accessibilité d'outils de visualisation en ligne aboutit à considérer le journalisme de données non pas comme une démarche journalistique nourrie par l'analyse de données mais comme une forme de journalisme "gadgétisé".

sés sur la visualisation de données, mais aussi du journalisme de données. J'essaie donc de trouver des histoires dans de grands ensembles de données et des bases de données, puis je raconte ces histoires en combinant du texte avec des graphiques et des cartes" ¹⁵. Interviewé en 2016 à propos du nombre de journalistes de données en Belgique, il répondait : "J'aimerais avoir plus de collègues dans ce domaine, mais je suis certains que beaucoup de personnes qui travaillent dans celui de la business intelligence font des choses similaires. Peut-être qu'ils n'investissent pas autant dans le storytelling et dans la visualisation comme je le fais, mais il y a certainement d'autres profils que le mien avec des connaissances et compétences similaires. Seulement, ils ne sont pas dans les rédactions aujourd'hui" ¹⁶.

Mais l'épidémie de Covid-19 a bien changé la donne : tous les médias du pays publient chaque jour les données relatives à l'épidémie, que ce soit au niveau national ou international, et cela donne lieu à des récits où textes et graphiques se répondent, en vue de fournir une analyse approfondie de la situation. Toutefois, le récit est souvent moins piloté par les données que par des informations plus générales de contexte, les données jouant alors un rôle d'illustration du propos journalistique.

En France, plusieurs grands médias publient régulièrement des récits journalistiques s'appuyant sur des données. Aussi, les pratiques du journalisme de données y sont beaucoup plus répandues, que ce soit à l'échelon national (AFP, Le Monde, Libération, Médiapart,...) ou local (Le Parisien, Ouest France). Les pionniers du data journalisme en France étaient réunis sous la bannière du média en ligne OWNI, qui a officié de 2009 à 2012 (Nicolas Kayser-Bril, Pierre Romero...). Si à peine une vingtaine de datajournalistes étaient répertoriés, en France ¹⁷, au début des années 2010, la situation a bien évolué depuis. Certaines rédactions comptent des pôles dédiés à cette pratique, et l'on y retrouve des journalistes mais aussi des développeurs. Il existe également des agences spécialisées dans le traitement de données : WeDoData (visualisation de données), Syllabs et LabSense (service d'automatisation de traitement de données, génération automatique de textes, de nombreux partenariats avec des médias français pour la couverture des résultats électoraux). De plus, l'automatisation de la production d'informations s'invite de plus en plus souvent dans les rédactions, que ce soit en Belgique ou en France, pour traiter rapidement de larges volumes de données, mais aussi pour soutenir les journalistes dans leurs tâches quotidiennes en prenant en charge les aspects les plus routiniers et répétitifs (parexemple, le projet "Quotebot" à L'Echo, voir Dierickx 2020).

¹⁵ "Marteen Lambrechts, freelance data journalist", Glyn Moddy, Copybuzz, 2018, consulté le 23/04/2019, URL : <https://copybuzz.com/fr/copyright/humansofcopyright-maarten-lambrechts-freelance-data-journalist/>

¹⁶ "Data journalism blog : Maarten Lambrechts on the underused power of explorable explanations & more", Anastasia Valeeva, Open Belgium 2016, 14/09/2016, consulté le 23/04/2019, URL : <http://2016.openbelgium.be/blogpost/data-journalism-blog-maarten-lambrechts-underused-power-explorable-explanations-more>

¹⁷ Voir le projet Jourdain : <https://projetjourdain.org/les-datajournalistes-francais/>

2.2.2 Le mythe de l'open data

Le mouvement de l'open data a contribué à forger l'enthousiasme de journalistes embarqués dans des projets "pilotes" par des données. Toutefois, les données publiques ouvertes n'ont pas rencontré ces promesses et manquent souvent de pertinence pour conduire une enquête journalistique (Stoneman 2015, Goëta et Mabi 2014). Howard (2014) considère que les portails open data peuvent proposer de nombreuses données pertinentes mais que les données les plus sensibles, telles que celles relatives à l'argent public, s'y font encore trop rares. Le constat est partagé par Linden (2017), qui souligne que les données les plus utiles à un projet journalistique ne sont pas nécessairement accessibles en open data, et que cet accès serait davantage piloté par des raisons économiques et non démocratiques (Linden 2017b).

Stoneman (2015) estime que les sources de désenchantement de l'open data seraient à trouver dans la non-prise en compte du public des journalistes en tant qu'utilisateurs finaux, mais aussi dans la nature même des données, souvent obsolètes, incomplètes ou à valeur journalistique limitée. *"Même dans les meilleurs des cas, les open data ne répondent pas à leurs besoins"*, indique-t-il, soulignant que les journalistes privilégient d'autres voies pour collecter des données : soit en s'appuyant sur la législation relative au droit à l'information¹⁸, soit en les collectant eux-mêmes.

Courmont (2015) précise que *"toute donnée est le résultat d'un réseau sociotechnique. Il est impossible de dissocier la donnée de son usage, de détacher l'instrument technique de l'environnement social dans lequel il s'inscrit"*. Ajoutons qu'il n'est pas plus possible de dissocier la donnée de l'environnement social dans lequel elle est produite, renvoyant ainsi aux modèles de gouvernance des données et de leur corollaire, des politiques pertinentes et adaptées en matière d'open data public. Si la figure du journaliste est peu – voire pas du tout – mise en avant dans les politiques d'ouverture des données, la figure du développeur, elle, est omniprésente : c'est vers lui que sont destinés les encouragements à la réutilisation des données à travers le développement d'applications. L'enjeu est donc de proposer des outils simples aux journalistes pour que ceux-ci puissent étudier les données (Colpaert & al. 2013).

2.3 Enjeux professionnels

Les bases de données informatiques jouent aujourd'hui un rôle structurant dans le contexte journalistique. Elles sont devenues le matériau de base tant à un travail d'investigation qu'à des objets autonomes dans des pratiques désignées sous les vocables de "datajournalisme", "journalisme computationnel" ou "journalisme automatique". Ce phénomène a entraîné dans son sillage de nouvelles collaborations entre journalistes et informaticiens, voire l'émergence de nouveaux profils de journalistes technologues. Ces développements témoignent du fait, af-

¹⁸ En France et en Belgique, la Commission d'accès aux documents administratifs (CADA) est l'autorité administrative, indépendante et régionale, auprès de laquelle tout citoyen peut s'adresser pour obtenir des données spécifiques. Elle intervient en cas de refus d'accès ou de rectification d'un document par l'administration sollicitée

firmement Flichy et Parasie (2013) que les bases de données "*échappent de plus en plus aux intérêts des informaticiens. Celles-ci sont aujourd'hui au cœur des pratiques et des représentations d'individus qui rencontrent cette forme sociotechnique, alors même qu'ils s'inscrivent dans les mondes sociaux les plus variés*". Leur essor, poursuivent-ils, est lié à la fois aux transformations techniques et sociales.

Dans l'introduction au "Guide du datajournalisme", Bradshaw (2013) souligne que les données sont à la fois des sources d'information et des outils permettant de raconter une histoire. "*Comme n'importe quelle source, elles doivent être traitées avec scepticisme; et comme n'importe quel outil, nous devons prendre conscience de leurs limites et de leur influence sur la forme des histoires qu'elles nous permettent de créer*". Stray (2016) souligne que les enjeux d'une approche par données dans le journalisme portent sur la qualité de l'information numérique et de son corollaire, la fiabilité des sources d'information. Si les données constituent une source primaire, leur qualité ne peut être assurée entièrement dès lors qu'elles sont disponibles en ligne, évoluant de ce fait dans un contexte ouvert et non contrôlé (Boydens 2014). Stray ajoute que ces enjeux portent également sur le danger de l'utilisation de données incomplètes ou conflictuelles car si les chiffres rendent la réalité plus "objective", cette précision n'est pas toujours réalisée. Dans le cas où les données consistent en un enregistrement, un document, un artefact, elles ne sont donc pas à l'abri d'erreurs, souligne-t-il. Mais les erreurs ne concernent pas uniquement les données : une interprétation erronée et les erreurs statistiques sont susceptibles d'affecter des résultats (Stray 2016).

Daniel & al. (2010) ont identifié plusieurs enjeux spécifiques à la pratique du journalisme computationnel, qui constitue l'une des formes de l'approche par données dans le journalisme : des données de mauvaise qualité incluant des anomalies tels que les faux positifs, les doublons, le manque de sens, des données conflictuelles ou incomplètes; une mauvaise interprétation; l'introduction de biais dans les résultats. Ce qui implique que les journalistes doivent vérifier chaque jeu de données et chaque résultat, et apporter un contexte aux données. Ils soulignent également le danger d'un mauvais usage des données, ce qui suppose de bien identifier leur source et de s'assurer que celles-ci rencontrent bien les exigences de la protection de la vie privée.

Le journalisme de données implique que les organisations médiatiques se pensent et s'organisent de manière différente car le journalisme de données nécessite idéalement une approche multidisciplinaire où journalistes, programmeurs et designers sont appelés à travailler de concert. "*Les passerelles existent et sont nécessaires*", estime Alice Antheaume, qui relève que celles-ci doivent s'opérer "*dans le cadre de cultures différentes*". Se pose encore la question du modèle économique : la pratique du journalisme de données nécessite du temps, celui de l'enquête et/ou des activités liées à la collecte et au traitement des données; et des moyens, celui de sa mise en œuvre. Or, le retour sur investissement n'est pas toujours évident compte tenu du temps nécessaire à la mise en œuvre de tels projets.

2.3.1 L'objectivation chiffrée

La quantification, ce n'est pas quelque chose qui existe par nature. Si les données consistent en un enregistrement, un document, un artefact, elles ne sont pas à l'abri d'erreurs. Les données quantifiées représentent le monde mais cette représentation est fragile. En elle-même, une donnée n'est pas signifiante. Elle ne fera sens qu'à partir du moment où elle est analysée; ce qui implique un travail journalistique de mise en contexte. Toutes les analyses de données sont des interprétations. Ce travail d'analyse n'est pas exempt de subjectivité. De plus, en fonction du point de vue que l'on adopte, il existe plusieurs manières d'interpréter des données. Le fait de trouver une histoire dans un jeu de données consisterait ainsi en un acte de création. Stray met en garde contre les écueils de cette étape d'analyse : 1) des effets liés à la chance, au hasard ou au bruit peuvent obscurcir la relation entre deux variables ou donner une apparence de relation qui n'existe pas; 2) la nature de la cause peut être sujette à de multiples explications et la première explication qui fait sens n'est pas toujours celle à retenir. Aussi, les données peuvent-elles contenir des vérités mais aussi des contre-vérités (Howard 2014). Powell (2016), tout comme Stray (2016), préconise de les envisager avec suspicion. Si un algorithme d'exploration de données peut permettre de découvrir de nouvelles connexions entre de multiples variables ayant une très grande importance statistique en raison de l'énorme quantité de données analysées, les résultats peuvent être sans signification et n'ajouter aucune valeur à l'information. Des questions incorrectes peuvent donner lieu à des analyses biaisées, tout comme des données ou des procédures incorrectes (Latar 2015).

Toute activité d'analyse des données consistera en une représentation et elle est susceptible de comporter une marge d'erreur quant à la manière dont elle va être menée. Une représentation de données peut être trompeuse, en négligeant des aspects importants au profit d'éléments plus favorables à la démarche dans le cadre d'une analyse statistique. Hors de leur contexte ou prises de manières isolées, les données pourraient occulter une partie de la réalité ou signifier tout autre chose (Flew & al. 2010). Privilégier une échelle de temps sur une autre peut conduire à une analyse divergente : si le taux de chômage n'a jamais été au plus bas en dix ans, celui-ci est peut-être bien plus élevé qu'il y a quinze ou vingt ans. Une baisse annoncée du taux de chômage peut obscurcir le fait qu'il y ait eu davantage d'exclus ainsi qu'une augmentation de l'aide sociale. L'utilisation d'indices peut donner lieu à des résultats faussés par la prise en compte de valeurs disparates. Les données peuvent aussi renfermer des biais liés à leur sélection, voire avoir fait l'objet d'une manipulation intentionnelle de la part du producteur d'informations. La question de la fiabilité de la source des données se pose ici en filigrane. *"Si vous ne comprenez pas les données, d'où elles viennent et ce qu'elles représentent, vos conclusions peuvent être porteuses de biais. Les données financières sont spécialement sensibles à ce problème, mais cela peut intervenir dans n'importe quel type de données"* (McCallum 2012 :127).

Lorsqu'une analyse de données repose sur des statistiques issues de données empiriques récoltées via des méthodes d'enquête ou de sondage, il convient également de s'interroger sur la méthode de collecte des données, sur la taille de l'échantillon, sa représentativité et la marge d'erreurs des résultats obtenus : *"une enquête mal menée, un questionnaire mal rédigé ou un*

échantillon mal choisi peuvent aboutir à des résultats statistiques complètement erronés" (Foucart 2001). La quantification implique des choix complexes : comment quantifier des concepts abstraits tels que l'intelligence ou la qualité de vie? Les sondages fournissent des données quantifiées souvent avec une marge d'erreur, et les échantillons ne sont pas toujours représentatifs : dès lors, il faut toujours se poser la question de la méthodologie employée et replacer les résultats dans un contexte plus général. Par ailleurs, établir une relation entre deux variables ne permet pas nécessairement d'établir un lien de cause à effets : *"La causalité demande donc une description de la réalité à laquelle on se limite. Les interprétations d'une relation statistique ne sont pas des vérités objectives en ce sens que deux personnes peuvent en proposer des interprétations différentes puisqu'elle est interprétée nécessairement dans un contexte différent, ne serait-ce qu'à cause des personnalités différentes"* (Foucart 2001). De plus, les résultats seront souvent présentés en termes de probabilités.

La fonction des statistiques est de formuler de "bons" arguments et d'expliquer des différences comparatives : si des évidences pouvant être considérées de grande qualité sont nécessaires dans l'argument statistique, cela ne sera pas suffisant. Cela va dépendre de la qualité des données récoltées mais aussi de compétences dans la conception du modèle de récolte de données et dans la présentation des résultats où, là aussi, des facteurs subjectifs devront être pris en considération (Abelson 1995 :13-17). C'est là le propre d'une sociologie de la quantification, dont la visée est pourtant celle d'une connaissance "objective et neutre" (Desrosières 2008 :27). Dès lors, les données statistiques ne peuvent être abordées que de manière prudente et critique, bien qu'elles permettent d'identifier des tendances et des schémas importants (Lowrey 2019). De plus, il est toujours possible de faire mentir les chiffres selon ceux que l'on retient ou que l'on exclut ou selon de mauvaises données... (De Veaux & al. 2005, Huff 2010).

Toutes les formes de quantification transforment le monde (Desrosières 2008), et toutes les formes d'interprétation de ces quantifications dépendent toujours du point de vue adopté tant lors du traitement et de l'analyse des données que lors de la mise en récit médiatique. En ce sens, le concept de "réalité" est à géométrie variable : c'est là une autre limite du principe de l'objectivation chiffrée. Le traitement de données ne donne pas lieu à des acquis immuables et cela est d'autant plus vrai dans le contexte journalistique, où la description de la réalité s'inscrit dans le cadre mouvant d'une histoire se construisant de jour en jour, voire d'heure en heure. De plus, les domaines d'expertise peuvent également évoluer dans le temps, pouvant ainsi donner lieu à des interprétations différentes, voire divergentes. Une approche journalistique axée sur les données est fondée sur trois facteurs qui interagissent : la technologie pour collecter et traiter les données, l'analyse des données et le mythe selon lequel de grands ensembles de données aideront à atteindre les objectifs de vérité et d'objectivité (Sandoval-Martín & La-Rosa 2018). A cela, il faut également ajouter que les concepts et domaines d'application peuvent varier dans le temps : une indice de référence d'une valeur X peut voir cette valeur devenir Y quelques mois ou années plus tard. Une bonne connaissance du domaine d'application que l'on traite apparaît, dès lors, comme un préalable. Comment traiter la problématique de la qualité de l'air sans connaître et comprendre les concepts auxquels elle se rapporte? Comment traiter l'évolution

de valeurs boursières sans connaître à quoi se rapportent des indices boursiers et s'interroger sur ce sentiment de confiance qui influence les marchés boursiers? Dans de nombreux cas, cette compréhension passera par la sollicitation d'un expert.

2.3.2 La problématique de la qualité des données

En tant que terme technique, une base de données peut être définie comme une "collection de données qui sont, de manière critique, codées et organisées selon un schéma commun" (Dourish 2014). Une base de données empirique et la réalité qu'elle appréhende s'influencent mutuellement tout en se transformant à des rythmes différents : les concepts informatiques changent discrètement, interagissant avec une réalité normative en constante évolution. Par conséquent, leur représentation de la réalité observable sera toujours imparfaite (Boydens 1999). De nos jours, les bases de données jouent un rôle structurant dans le contexte journalistique, où l'on peut considérer que les faits sont "vrais" tant qu'ils existent dans une base de données (Anderson 2018 : 31-32). C'est pourquoi une approche journalistique axée sur les données doit également être approchée sous l'angle de la compréhension de ce qu'est une base de données et de la manière dont elle a été conçue en amont.

La nécessité de disposer de données structurées sera une pré-condition pour une utilisation journalistique. Cela suppose de traiter la source des données avec scepticisme (Bradshaw 2015), et de pouvoir bien en identifier l'origine (Dörr & Hollnbuchner 2016). Si l'on ne peut faire confiance au producteur ou au diffuseur des données, le risque est que l'on ne puisse pas davantage faire confiance aux données et donc à l'information. Cet aspect est d'autant plus important que l'un des plus grands défis actuels, pour les médias, est précisément de restaurer un rapport de confiance avec leurs audiences (Fink 2019). Comme n'importe quelle autre source d'informations, les données doivent pouvoir être vérifiées : il s'agit là de satisfaire le principe journalistique de recherche de la vérité, lequel suppose des activités de vérification (Cornu 2009 :78).

Il convient aussi de s'interroger aussi sur la manière dont les données ont été récoltées ainsi que sur leur mise en contexte (Bradshaw 2015). Les métadonnées attachées à un jeu de données donneront, à ce propos, des indications importantes pour aider les utilisateurs à facilement évaluer et comprendre les données (Shanks 1999). Toutefois, celles-ci ne seront pas nécessairement attachées à un ensemble de données, ce qui signifie que celui-ci ne pourra pas toujours être bien compris et que certaines ambiguïtés liées à la nature des variables ou de leurs attributs sont susceptibles d'apparaître. De plus, la mise en disponibilité de données ne signifiera pas que celles-ci seront les plus pertinentes pour un usage journalistique, pas plus qu'elle ne garantira qu'elles soient exemptes d'anomalies dans leurs valeurs (Casswell & Dörr 2017).

Les données sont devenues une matière première de l'information, que ce soit dans le cadre du traitement de l'actualité, d'un travail d'investigation ou du développement d'outils numériques dédiés à la mise en récit journalistique. Cette évolution des pratiques a entraîné dans

son sillage de nouvelles collaborations entre des journalistes et des programmeur. Cet essor, dans le monde du journalisme, est donc à la fois lié à des transformations techniques et sociales, attestant du fait que les bases de données "*échappent de plus en plus aux intérêts des informaticiens. Celles-ci sont aujourd'hui au cœur des pratiques et des représentations d'individus qui rencontrent cette forme sociotechnique, alors même qu'ils s'inscrivent dans les mondes sociaux les plus variés*" (Flichy & Parasie 2013).

La Data Science Organisation (Etats-Unis) définit une donnée comme un enregistrement tangible ou électronique de l'information brute, utilisé comme une base pour le raisonnement, la discussion, ou le calcul, et qui doit être traitée ou analysée pour avoir du sens. Une donnée peut être définie de deux manières : 1) un état de fait résultant de mesures et d'observations ; 2) un triplet composé d'une entité, d'un attribut et d'une valeur (e-a-v / exemple : Ville – Bruxelles - 1000).

Une donnée peut être considérée comme constituant le matériau de base (input) à partir duquel l'information est développée (output) (Dorn 1981, cité par Fox & al. 1994). L'information consiste en un processus de données donnant lieu à n'importe quel type de connaissance : c'est la partie non-redondante d'un message, le résumé pertinent des données (Redman, 1996). A l'inverse de l'entropie – principe défini dans la théorie de l'information de Shannon, qui consiste en une formule mathématique pour déterminer la quantité d'informations contenues ou délivrées dans une source d'information –, "c'est la partie du message qui informe". Dans le domaine de l'étude des bases de données informatisées, la qualité des données dépendra la qualité de l'information (Batini et al. 2009, Redman 1996).

Le concept de la qualité des données s'est développé dans le domaine de l'entreprise, où des données de mauvaise qualité peuvent être lourdes de conséquences financières (voir à ce propos les travaux de Thomas Redman et d'Isabelle Boydens). La qualité totale d'une base de données n'existe pas, en raison du caractère évolutif des données (avec le temps) et de l'absence de référentiel absolu pour comparer les données. Aussi, les indicateurs de qualité peuvent être différents d'une base de données à l'autre, dans la mesure où ces indicateurs sont déterminés par l'usage des données. La norme ISO (organisme international de standardisation) 9000 relative au management de la qualité dispose, en effet, que la qualité consiste dans l'aptitude d'un ensemble de caractéristiques intrinsèques à satisfaire les exigences ou besoins des utilisateurs (on parlera de *fitness for use*, Boydens & van Hooland 2011).

Les normes internationales actuelles en matière de "data quality" trouvent leur origine au début du 20^e siècle, dans la foulée de la révolution industrielle et du taylorisme (production de pièces normalisées à grande échelle). A l'époque, l'objectif qualitatif consiste à atteindre un seuil de précision compatible avec l'impératif d'interchangeabilité des pièces, tout en minimisant les coûts correspondants. Dès l'origine sont donc apparus quelques grands principes fondamentaux en matière de qualité des bases de données : l'idée de "one best" ("le meilleur relativement") indique que la perfection est une "non-valeur". Les enjeux de la qualité des don-

nées sont considérables, dès lors que les systèmes d'information sont des instruments d'action sur le réel (bases de données administratives, économiques ou militaires, par exemple). Les problèmes de qualité des données ne sont pas anodins, en particulier lorsque les données sont utilisées comme matière première pour les contenus d'actualité.

Le concept de qualité des données est à géométrie variable et va dépendre à la fois du domaine d'application et des usages qui sont faits des données. Par ailleurs, une donnée empirique n'est jamais qu'une photographie d'un état de fait à un moment donné. C'est la raison pour laquelle des données empiriques, issues d'observations, sont susceptibles d'évoluer dans le temps. La littérature scientifique propose différentes manières d'aborder le concept de la qualité des données, tant dans le contexte des systèmes d'information que dans celui des bases de données informatiques. Les caractéristiques de la qualité des données consistent en plusieurs "dimensions" complémentaires, lesquelles désignent un ensemble d'attributs caractérisant la qualité des données.

Les conditions d'utilisation de données dans des projets journalistiques vont se concentrer sur l'exactitude et la fiabilité des données, tout en insistant sur la nécessité de développer ces projets conformément aux principes éthiques régissant les pratiques journalistiques. Pour Bradshaw, il est utile de se demander si les données sont exactes et vérifiables. Il préconise que les journalistes s'interrogent sur la précision des données, sur la manière dont les données sont collectées, ainsi que sur celle de les replacer dans leur contexte. Il souligne également l'intérêt de travailler avec les données les plus pertinentes plutôt que de travailler avec les données les plus accessibles. Dans son guide consacré aux "mauvaises données", Quartz souligne que les journalistes sont de plus en plus confrontés aux données et celles-ci peuvent présenter des problèmes. Les connaître permet d'y apporter des solutions, pour autant que les problèmes identifiés puissent être résolus. Dans le cas contraire, elles ne pourront pas être utilisées. *Vous ne pouvez pas évaluer tous les jeux de données que vous rencontrez. Si vous essayez de le faire, vous ne verrez jamais rien. Cependant, en vous familiarisant avec les types de problèmes que vous rencontrerez, vous aurez une meilleure chance d'identifier un problème*"¹⁹. Ce guide explore ces différents problèmes, qui concernent tant la qualité formelle des données (valeurs manquantes, unités non spécifiées, étiquetage ambigu, ...) que les aspects liés aux dimensions de la qualité des données (marge d'erreur inconnue, agrégations calculées sur base de valeurs manquantes, trop grande granularité des données, échantillon partial...). Les aspects relatifs à la qualité des données recouvrent deux challenges : à la fois journalistiques et techniques, et cela est d'autant plus important dès lors que les données vont faire l'objet de traitement automatisés.

2.4 Bonnes pratiques

Le datajournalisme doit répondre à de nombreux défis : sur le plan de la collecte des données, il s'agit de s'assurer de la fiabilité de la source et du respect de la vie privée (dans certains cas, les

¹⁹ "The Quartz guide to bad data", url<https://github.com/Quartz/bad-data-guide>

données devront être anonymisées) mais aussi de la fiabilité des données (ce qui implique une vérification) ; sur le plan de l'analyse, il s'agit d'interroger correctement les données (ce qui implique de comprendre comment le jeu de données a été collecté, dans quels buts et de s'interroger sur les valeurs manquantes dans un jeu de données). De plus, si les données représentent un état de fait, les processus journalistiques impliquent des choix et des jugements éditoriaux. La manière l'information fait l'objet d'une sélection témoigne de ces choix et jugements éditoriaux. Des critères "objectivables" côtoient tout aussi bien des critères plus subjectifs liés à l'expérience, aux attentes, ou aux référents socioculturels du journaliste.

Sur le plan déontologique, Tom Kent, enseignant en journalisme à la Columbia University et journaliste à l'AP, préconise de se poser deux questions préalables à l'utilisation de donnée :

Les données d'origine sont-elles fiables ?

La source des données est-elle un ministère ou une agence du gouvernement ? Puisse-t-on dans les documents publics d'une entreprise cotée en bourse ? Dans ces cas, les données sont probablement fiables (il demeure nécessaire, bien sûr, de vérifier si leur transmission se fait toujours correctement). Les sources, cependant, ne sont pas toujours crédibles. Les données sur le football amateur peuvent être fournies par des parents qui assistent aux parties de leurs enfants, par exemple. Pouvez-vous faire confiance en tout temps à ce type de collecte de données ?

Avez-vous des droits sur les données ?

Vos fournisseurs de données ont-ils le droit de vous les faire parvenir ? Avez-vous le droit de les traiter et de les publier ? Si oui, vos droits s'étendent-ils sur l'ensemble des plateformes sur lesquelles vous diffusez ? Et ces droits sont-ils éternels ou limités dans le temps ? Ce n'est pas parce que l'on trouve des données sur un site internet que celles-ci sont forcément mises à disposition de manière libre par leur producteur ou diffuseur. La structure d'une base de données informatisées est couverte par le droit d'auteur, et certains flux de données doivent faire l'objet d'une souscription payante pour y avoir accès (notamment dans les domaines de la finance et du sport), engageant des coûts parfois très élevés pour un média d'informations. De plus, dès lors qu'une base de données contient des données à caractère personnel, protégées par la loi sur la vie privée, celles-ci ne peuvent pas être exploitées sans l'assentiment des individus concernés²⁰. La directive européenne RGPD a renforcé ce cadre récemment²¹. Cela implique également que toutes les données d'ordre privé diffusées en ligne ne sont pas exploitables sans l'assentiment explicite de la personne concernée.

2.4.1 Droit d'accès aux données publiques

La Directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013 modifiant la directive 2003/98/CE concernant la réutilisation des informations du secteur public ("Directive PSI") prévoit une série de dispositions en matière de communication des données publiques²².

²⁰ Voir <https://www.autoriteprotectiondonnees.be>

²¹ Voir <https://www.autoriteprotectiondonnees.be/reglement-general-sur-la-protection-des-donnees>

²² Voir : <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:FULL:FR:PDF>

Loi du 4 mai 2016, publiée au Moniteur Belge le 3 juin 2016, consiste en une transposition de cette directive²³. Toutefois, cela ne signifie pas pour autant que toutes les données publiques soient bien accessibles (ou "libérées") dans les faits. Par exemple, les données de l'Institut Royal Météorologique (IRM) ne sont toujours pas publiées en open data : l'IRM vend ses données et la mise en œuvre de cette politique aurait pour conséquences un important manque à gagner. A l'époque de l'adoption de cette loi, en 2016, le ministre compétent indiquait pourtant que *"Concrètement, chacun pourra réutiliser les informations du secteur public dans quelque but que ce soit, commercial ou non. Les entreprises publiques seront elles-aussi soumises à cette nouvelle réglementation, en ce qui concerne leurs missions de service public. Les concepteurs d'applications auront ainsi beaucoup plus facilement accès aux données du secteur public, par exemple aux horaires de la SNCB, au budget des dépenses de l'autorité fédérale ou aux prévisions météo de l'Institut royal météorologique (IRM). Les pouvoirs publics mettront ces données gratuitement à disposition. Seule une intervention dans les coûts de mise à disposition pourra être demandée, par exemple pour l'enregistrement sur support électronique. Seuls les pouvoirs publics qui doivent tirer des revenus de la mise à disposition de leurs documents, ainsi que les bibliothèques, les archives et les musées pourront facturer des tarifs plus élevés"*²⁴.

En Belgique, la Commission d'accès aux documents administratifs est un organisme de recours contre *"les manquements aux obligations de publicité active, de publicité passive et aux refus de rectifications d'initiative"* footnoteSource : <https://be.brussels/a-propos-de-la-region/commission-dacces-aux-documents-administratifs/cada>. L'organisation Transparencia offre, par ailleurs, une assistance pour saisir la CADA²⁵. Un tel organisme existe également en France, où le livre III du code des relations entre le public et l'administration reconnaît à toute personne le droit d'obtenir communication des documents détenus dans le cadre de sa mission de service public par une administration, quels que soient leur forme ou leur support. Ce droit s'exerce à l'égard de toutes les personnes publiques (l'État, les collectivités territoriales et leurs établissements publics) ainsi qu'à l'égard des organismes privés chargés d'une mission de service public. L'accès à certaines informations, par exemple les dossiers médicaux, les listes électorales ou les informations environnementales, obéit à des règles particulières, souvent plus libérales que le régime général. La loi prévoit toutefois quelques restrictions au droit d'accès, nécessaires pour préserver divers secrets, tel par exemple celui qui garantit dans l'intérêt des personnes le respect de la vie privée ou encore celui qui garantit dans l'intérêt de la concurrence le secret des affaires²⁶.

2.4.2 Droits d'auteur

La structure d'une base de données est protégée par le droit d'auteur. Par structure, il faut entendre la manière dont les données sont disposées et classées. Cette structure doit être originale, c'est-à-dire constituer une création intellectuelle propre à son auteur. Les données brutes

²³ http://tiny.cc/loi_open_data

²⁴ (Source : <https://www.decroo.belgium.be/fr/la-chambre-votera-demain-une-ambitieuse-loi-open-data>)

²⁵ <https://be.brussels/a-propos-de-la-region/commission-dacces-aux-documents-administratifs/cada>

²⁶ Voir : <http://www.cada.fr/>

ne sont donc pas, en elles-mêmes protégées. Cette protection garantit au producteur des données un monopole d'exploitation. En France, le Code de la propriété intellectuelle définit la notion de base de données comme un "recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen" (art. L.112-3 C.propr.intell.). Les bases de données sont protégées par le droit moral (droit intellectuel de l'auteur) et le droit patrimonial (droit d'exploitation commerciale ou non). Un producteur de données peut interdire l'extraction de tout ou partie du contenu de la base de données, ainsi que la réutilisation par la mise à disposition du public de tout ou partie du contenu. Par ailleurs, une collection de liens hypertextes peut être considérée, par la loi, comme étant une base de données. Des exceptions au droit d'auteur sont toutefois prévues par la loi, dont l'utilisation à des fins pédagogiques.

En Belgique, la structure originale d'une base de données, de même que les contenus d'une base de données, sont protégés par le droit d'auteur. *"Le rassemblement de ces données et leur organisation en un ensemble cohérent et susceptible d'une recherche d'informations précises, sont le fruit d'investissements considérables de la part des producteurs des bases données. Ces bases de données ont également, dans de nombreux cas, une valeur économique très importante. Ce sont les raisons qui justifient que les bases de données soient protégées par la loi²⁷".* Cela suppose que, dans certains cas, il faudra s'assurer de disposer de l'autorisation du producteur des données.

A propos des licences d'utilisation des données, celles estampillées Creative Commons, nées en 2002, définissent des licences ouvertes ainsi que les conditions de réutilisation. Six licences sont disponibles : <http://creativecommons.fr/> Il existe également d'autres formes de licence, telle que la licence ODBL relative à un type de licence ouverte pour les portails open data. Lorsque les données sont disponibles sur un portail de données ouvertes, il est essentiel de consulter la licence proposée ainsi que ses modalités de réutilisation.

2.4.3 Normalisation des données

Le terme norme est dérivé du nom latin "normat", qui signifie équerre, règle. Ce concept désigne littéralement la formule abstraite ou le type concret de ce qui doit être", Boydens (2012) ajoutant qu'une norme considérée comme principe de comparaison et d'évaluation "*renvoie à la notion de commune mesure*". Des données standardisées favorisent leur interopérabilité. Les normes se présentent comme objectives mais elles ne sont pas neutres : une norme consiste toujours en une convention fixée de manière arbitraire par une institution. Elle a pour caractéristiques de pouvoir être respectée ou violée.

Le domaine de l'informatique est régi par toute une série de normes, destinées à fournir une "commune mesure" à ses praticiens. Des organismes de standardisation élaborent et recommandent des normes : que ce soit sur le plan de la programmation et de ses bonnes pratiques via le W3C, un organisme de standardisation sans but lucratif chargé de promouvoir la com-

²⁷ Source : <https://economie.fgov.be/fr/themes/propriete-intellectuelle/droit-des-bases-de-donnees>

patibilité des technologies du web; ou sur celui de la définition de spécifications, de lignes directrices ou de caractéristiques destinés à assurer l'aptitude à l'emploi des matériaux, produits, processus et services (normes ISO). Comment standardiser? Il existe tout un éventail de normes ISO (Organisation internationale de normalisation), prévues pour un usage général dans les divers domaines scientifiques et techniques :

- **Références bibliographiques (ISO 690 :2010) :** donne des principes directeurs pour la rédaction des références bibliographiques, en organisant un ordre dans les mentions. Exemple pour un livre : NOM, Prénom. Titre. Édition, collection, année.
- **Représentation des pays (ISO 3166-3 :2013) :** énonce les principes pour une représentation des pays, BE pour Belgique, FR pour France...
- **Représentation des monnaies (ISO 4217 :2015) :** définit le code de trois lettres attribué aux devises dans le monde, EUR pour l'euro, USD pour le dollar américain...
- **Représentation normalisée de la localisation des points géographiques par coordonnées (ISO 6709 :2008) :** spécifie notamment la représentation des coordonnées, dont la latitude et la longitude, utilisées pour l'échange de données. La norme ISO 19101-1 :2014 définit le modèle de référence pour la normalisation dans le domaine de l'information géographique, lequel décrit la notion d'interopérabilité et établit les principes de base sur lesquels s'appuiera la normalisation.
- **Représentation de la date et de l'heure (ISO 8601 :2004) :** mode de représentation numérique de la date et de l'heure accepté à l'échelon international. Elle pour objet de lever l'ambiguïté d'interprétation lorsque les dates sont exprimées en chiffres : YYYY-MM-DD pour les dates, HH :MM :SS pour les heures, YYYY-MM-DD HH :MM :SS pour la date et l'heure.
- **Grandeurs et mesures (ISO 800001 :2009) :** informations générales à propos des grandeurs, des systèmes de grandeurs, des unités, des symboles de grandeurs et d'unités, et des systèmes cohérents d'unités. La norme s'appuie notamment sur le Système international de grandeurs (ISQ) et sur le Système international d'unités (SI). Elle concerne notamment le poids (μg , mg, g, kg, Mg), la longueur (nm, μm , mm, cm, dm, m, hm, km), la surface (mm^2 , cm^2 , dm^2 , hm^2 , km^2 , a28, ha29), la masse (mm^3 , cm^3 , dm^3 , m^3), la vitesse (m/s, km/h), le temps (s, m, h) et la température (ex. degrés Celsius = °C). La norme précise aussi les conventions de rédaction/d'encodage pour les chiffres : les décimaux sont séparés des unités par une virgule pour répondre aux conventions internationales.

ISO ne standardise pas tout, d'autres normes sont ainsi utilisées pour les :

- **Standards d'adressage :** chaque pays fixant ses règles, l'UPU (Universal Post Union) ras-

semble les informations relatives à ses membres et propose la norme S42, pour faciliter les échanges en uniformisant les pratiques d'encodage à partir de la nomenclature pré-nom – NOM – numéro de rue – type de rue – numéro – ville – région – code postal – pays.

- **Numérotations téléphoniques :** l'UIT/ITU (Union Internationale des Télécommunications) propose la recommandation E123 en vue d'harmoniser la nomenclature des numérotations téléphoniques. Elle prévoit deux modes de numérotations : national (dix chiffres séparés par des espaces, dont les deux premiers numéros correspondent au code territorial) et international (la numérotation débute par deux chiffres, correspondant à l'indicatif du pays, précédés du sigle "+", le troisième chiffre est celui du code territorial avec la suppression du premier 0 : + 32 2 222 22 22 pour Bruxelles, par exemple).
- **Formats de données numériques :** le W3C propose une série de recommandations par exemple sur l'utilisation de données tabulaires (format CSV), sur l'organisation de données dans le métalangage XML, sur l'usage des métadonnées et des linked data (RDF et OWL), sur le vocabulaire contrôlé SKOS (organisation des thésaurus) ou encore sur le langage de requête SPARQL (qui interagit avec le format RDF).

2.4.4 Modèles d'évaluation de la qualité des données dans un contexte journalistique

Ce modèle conceptuel d'évaluation pour l'utilisation de données dans un contexte journalistique a été défini à partir de recommandations et bonnes pratiques tant dans le domaine du journalisme que dans celui de la science des données. Il propose trois étapes : aux deux premières, il suffit de répondre par "vrai" ou par "faux" (modèle booléen), la troisième nécessite une démarche empirique et réflexive (Dierickx 2017). Les limites de ce modèle tiennent aux faits qu'une donnée de mauvaise qualité peut cohabiter avec une donnée correcte sans que cela génère des erreurs, et que la source des données peut ne pas contenir d'erreurs mais l'utilisateur n'en aura pas le sens attendu.

1ère étape : Indicateurs relatifs à la qualité formelle des données (challenge technique)

Axe	Evaluation
Documentaire	Mention de la licence d'utilisation, identifiant unique, présence de métadonnées, conformité du jeu de données aux métadonnées.
Encodage	Pas de problème d'encodage, pas de surcharge du code HTML, pas de données dupliquées.
Normatif	Application des standards (adressage, géolocalisation, unités de mesure, URL normalisés, métadonnées,...).
Sémiotique	Pas d'incohérences orthographiques, étiquetage des colonnes explicite et non ambigu, pas de valeurs manquantes.

Ces indicateurs visent à évaluer la qualité des données de manière formelle. Cela signifie que l'évaluation pourrait être faite sur un mode booléen, en répondant par "vrai" ou "faux". Le

nombre total d'indicateurs formels utilisés pour une évaluation peut varier d'un ensemble de données à un autre, car chaque domaine d'application a ses propres particularités. De plus, les anomalies formelles dans un ensemble de données pourraient faire l'objet d'interprétations. Par exemple, la valeur NULL peut être interprétée de plusieurs manières : l'information existe mais n'est pas connue, l'information n'est pas pertinente pour l'entité, l'information est pertinente mais n'existe pas pour l'entité, la valeur d'attribut est égale à zéro (Hainaut 2012). Des données de mauvaise qualité peuvent coexister avec des données correctes sans générer d'erreurs (Wang & Strong 1996).

2ème étape : dimensions de la qualité des données (challenge journalistique)

Dimension	Evaluation
Contextuelle	Exactitude (correction syntaxique), précision des valeurs (pas d'anomalies), actualité (date et fréquence de mise à jour), pertinence journalistique.
Intrinsèque	Provenance (source authentique), nombre de données approprié (pas de lignes vides), complétude (pas de valeurs manquantes).

Ces indicateurs de qualité visent à compléter l'aspect formel de l'évaluation de la qualité des données. Ils peuvent également être évalués sur un mode booléen (vrai ou faux). L'indicateur de "complétude" pourrait être plus difficile à évaluer en raison du problème de la valeur NULL. L'indicateur lié à la quantité appropriée de données pourrait également être difficile à évaluer, surtout si l'ensemble de données compte des centaines d'enregistrements. Il est également important de noter qu'une source primaire, qui désigne le responsable du contrôle et du suivi dans le temps de la qualité de l'information produite, ne garantit pas la qualité d'un ensemble de données mais seulement donne des indications à son sujet (Boydens 2014). La qualité d'un ensemble de données ne peut pas être seulement évaluée avec des indicateurs de type déterministe. Si les indicateurs empiriques sont subjectifs, ils ne peuvent être évités dans un contexte journalistique. Il appartient aux journalistes, ou à ceux qui sont impliqués dans un projet axé sur les données, de rester critiques vis-à-vis d'un jeu de données, y compris lorsqu'il provient d'autorités publiques ou lorsqu'il est signalé comme provenant d'une source authentique. De plus, les données peuvent contenir des vérités mais aussi des biais (Howard 2014). Ce niveau critique ne peut être objectivé sans un peu d'investigation.

3ème étape : modèle d'évaluation général

Le modèle d'évaluation général s'articule en quatre axes complémentaires, liés au principe de fiabilité mis en exergue dans les domaines du journalisme et des bases de données informatisées. Ils visent à s'assurer (1) de la fiabilité de la source des données, (2) de disposer des accès nécessaires à l'utilisation données, (3) de disposer des clés de lecture nécessaires à la compréhension des données et de leur contexte de production, (4) de présenter des caractéristiques pertinentes pour un usage journalistique.

Source

Le diffuseur des données en est-il le producteur et/ou la source authentique? Dans le cas où le diffuseur des données n'en est pas le producteur original et/ou la source authentique, quelle est la nature de sa relation avec le producteur original des données et/ou la source authentique? Le diffuseur, le producteur et la source authentique sont-ils dignes de confiance?

Accès

Les données sont-elles librement accessibles? Font-elles l'objet d'une licence permettant leur libre réutilisation? Sont-elles disponibles dans un format structuré?

Documentation

Les données sont-elles documentées par des métadonnées ou par tout autre type d'information permettant de comprendre la structure de la base de données, de lever les éventuelles ambiguïtés dans l'étiquetage des données, d'apporter une expertise pour comprendre à quoi correspondent les valeurs des données? Des éléments de contexte sont-ils apportés?

Pertinence journalistique

Les données présentent-elles une valeur ajoutée en termes journalistiques? En quoi le traitement des données fait-il sens?

2.4.5 Résoudre les problèmes de qualité des données

Le magazine américain Quartz a publié un guide visant à résoudre les problèmes de qualité formelle pouvant être rencontrés dans n'importe quel jeu de données. Cette section consiste en une traduction de la version originale du guide²⁸.

1. Problèmes que vos sources peuvent vous aider à résoudre

Valeurs manquantes

Méfiez-vous des valeurs vides ou « nulles » dans un jeu de données, à moins que vous ne soyez certain de savoir ce qu'elles signifient. Si les données sont annuelles, la valeur de cette année n'a-t-elle jamais été collectée? Si c'est un sondage, un répondant a-t-il refusé de répondre à la question? Chaque fois que vous travaillez avec des jeux de données qui comportent des valeurs manquantes, vous devriez vous demander : "Est-ce que je sais ce que signifie l'absence de cette valeur?" Si la réponse est non, vous devriez demander à votre source pourquoi ces valeurs sont manquantes.

Valeurs manquantes remplacées par zéro

Il y a pire qu'une valeur manquante : lorsque celle-ci est remplacée par une valeur arbitraire. Cela peut être le fait d'un humain qui n'a pas réfléchi aux implications de ce remplacement ou cela peut se produire dans le résultat d'un processus automatisé qui n'avait pas été paramétré pour gérer les valeurs nulles. Si vous voyez des valeurs "zéro" (ou nulles), demandez-vous si

²⁸ Voir : <https://github.com/Quartz/bad-data-guide>

elles sont vraiment nulles ou ce qu'elles peuvent signifier. Si vous n'en n'êtes pas certain, interrogez la source des données (idéalement, le producteur plutôt que le diffuseur car il s'agit de la source primaire des données). La même prudence doit être de rigueur pour toute autre valeur non numérique où un "0" peut être représenté d'une autre manière. Par exemple, une valeur fautive '0' pour une date est souvent affichée en tant que '1970-01-01T00 : 00 : 00Z' ou '1969-12-31T24 : 59 : 59Z' – ce format date de l'époque Unix pour les horodatages. En matière de géolocalisation, cela peut conduire à la représentation de localisations erronées telles que "0 °00'00.0" N + 0 °00'00.0 « E » ou simplement "0 °N 0 °E" qui est un point qui se trouve dans l'océan Atlantique, au sud du Ghana aussi appelé "Null Island".

Présence de valeurs suspectes

La feuille de calcul comporte des dates en 1900 ou 1904

Données manquantes mais existantes

Des données sont parfois manquantes dans un jeu de données. Il est toutefois possible de vérifier la complétude d'un jeu de données : par exemple, vous disposez d'un jeu de données à propos des Etats-Unis, vous pouvez donc contrôler que les 50 états y sont représentés. Si vous traitez des jeux de données à propos d'équipes sportives, assurez-vous d'y trouver le nom de tous les joueurs ou de toutes les équipes attendues dans le jeu de données. Faites confiance à votre intuition si vous pensez que des données sont manquantes et vérifiez auprès de votre source. Lignes ou valeurs dupliquées. Si la même ligne apparaît plus d'une fois dans votre jeu de données, vous devriez savoir pourquoi. Il peut s'agir de données dupliquées ou de lignes utilisant les mêmes identifiants uniques. Si vous n'en connaissez pas la raison, toute opération de calcul effectuée à partir de ces données sera erronée. Dans ce cas, demandez des précisions à votre source.

Incohérence orthographique

L'orthographe est l'un des moyens les plus évidents de savoir si les données ont été compilées à la main. Ne regardez pas seulement les noms des personnes, c'est souvent l'endroit le plus difficile pour détecter les fautes d'orthographe. Cherchez plutôt des endroits où les noms de villes ou les états ne sont pas cohérents. (« Los Angelos » est une erreur très commune.) Si vous les trouvez, vous pouvez être certain que les données ont été compilées ou éditées à la main : c'est donc une raison pour être sceptique. Les données qui ont été éditées à la main sont les plus susceptibles de comporter des erreurs. Cela ne signifie pas que vous ne devriez pas les utiliser mais que vous devriez peut-être corriger ces erreurs manuellement ou en tenir compte dans vos analyses.

Incohérence dans l'ordre des noms

Vos données comportent-elles des noms de pays du Moyen-Orient ou d'Asie de l'Est? Etes-vous certain que les noms de famille se trouvent toujours au même endroit? Les producteurs de données se trompent habituellement dans ce type de noms. Si vous travaillez avec une liste de noms étrangers, vous devriez au moins procéder à un examen superficiel pour vous assurer

de la correspondance ou compréhension entre les attributs (étiquettes) et valeurs.

Incohérence des formats de date

Quelle est la bonne date pour le mois de septembre? "10/9/15" ou "9/10/15"? La première date a été rédigée par un Européen et la seconde par un Américain : les deux sont donc correctes. Mais sans connaître l'historique des données, vous ne pouvez pas en être certain à coup sûr. Sachez d'où proviennent vos données et assurez-vous qu'elles ont toutes été créées par des personnes du même continent.

Type de valeurs non spécifié

Ni le terme "poids" ni le terme "coût" ne donnent d'informations sur l'unité de mesure utilisée dans un jeu de données. Ne soyez pas trop prompts à supposer que les données produites aux États-Unis sont en livres et en dollars. Les données scientifiques sont souvent métriques. Les prix étrangers peuvent être spécifiés dans leur monnaie locale. Si les données n'épellent pas leurs unités, revenez à votre source. Même si elle précise les unités utilisées, méfiez-vous toujours de significations qui ont pu évoluer dans le temps. Un dollar en 2010 n'est pas un dollar aujourd'hui.

Catégories mal choisies

Méfiez-vous des valeurs qui prétendent être seulement vraies ou fausses mais qui ne le sont pas vraiment. C'est souvent le cas avec les enquêtes où les refus ou les non-réponses sont également des valeurs valables et significatives. Un autre problème consiste dans l'utilisation de n'importe quel type d'autre catégorie. Les mauvaises catégories peuvent également exclure artificiellement les données. Cela arrive souvent avec les statistiques de la criminalité. Le FBI a défini le crime de « viol » de différentes façons au fil du temps. En fait, ils ont fait un travail si médiocre pour déterminer ce qu'est le viol que de nombreux criminologues affirment que leurs statistiques ne devraient pas être utilisées du tout. Une mauvaise définition peut signifier qu'un crime est comptabilisé dans une catégorie différente de celle que vous attendiez ou qu'elle n'a pas été du tout prise en compte. Soyez à l'affût de ce type problème lorsque vous travaillez sur des sujets où les définitions ont tendance à être arbitraires, comme en matière d'origine ethnique.

Étiquetage des colonnes ambigu

Qu'est-ce qu'une résidence? Est-ce là où quelqu'un vit ou où il paie ses impôts? Les étiquettes des colonnes (intitulé des variables) ne sont jamais aussi précises que nous le souhaiterions. Même si vous déduisez correctement ce que les valeurs sont censées signifier, cette ambiguïté pourrait facilement amener la personne qui collecte les données à entrer une valeur erronée.

Provenance non documentée

Les jeux de données peuvent être créés par une variété de types d'individus et d'organisations, y compris les entreprises, les gouvernements, les organisations à but non lucratif et les théoriciens du complot. Les données sont recueillies de différentes façons, y compris via des en-

quêtes, des capteurs et des satellites. Le fait de savoir d'où proviennent vos données peut vous donner une bonne idée de leurs limites. Les données d'enquête, par exemple, sont rarement exhaustives. Les capteurs varient dans leur précision. Les gouvernements sont souvent réticents à vous donner des informations impartiales. Les données provenant d'une zone de guerre peuvent comporter un fort biais géographique en raison du danger de traverser les lignes de bataille. Les données qui ont été écrites par un médecin peuvent être saisies par une infirmière. Chaque étape de la chaîne de collecte des données peut donner lieu à des erreurs. Vous devez savoir d'où proviennent les données que vous utilisez.

Données trop grossières

Les données dont vous disposez vous proposent des états et vous avez besoin de régions. Vous disposez d'une liste d'employeurs alors que vous avez besoin de celle des employés. Vos données sont ventilées par année mais vous auriez préféré par mois. Dans de nombreux cas, les données ne rencontrent pas nos objectifs. Si vous recevez des données trop grossières, vous devriez demander à votre source des données plus spécifiques. Par ailleurs, vous ne devriez jamais diviser une valeur annuelle par 12 puis la nommer "moyenne par mois". Sans connaître la distribution des valeurs, ce chiffre n'aura aucun sens.

Les totaux diffèrent des agrégats publiés

Imaginez que vous obteniez, après un long combat, une liste "complète" d'incidents liés à l'utilisation de la force par la police. Vous ouvrez le fichier et découvrez 2.467 lignes. Mais pas si vite! Avant de publier quoi que ce soit à partir de ce jeu de données, essayez de retrouver la dernière fois où le chef de la police a parlé du recours à la force par son ministère. Peut-être que dans une interview accordée six semaines plus tôt, il a parlé de « moins de 2.000 fois » ou qu'il a donné un nombre spécifique qui ne correspond pas à votre jeu de données. Ces types de divergence entre les statistiques publiées et les données brutes peut s'avérer une source d'informations importante. Souvent, la réponse expliquant cette divergence sera simple. Par exemple, les données que vous avez reçues ne couvrent peut-être pas la même période dont il parlait. Mais parfois vous les rattraperez dans leur mensonge. Dans les deux cas, vous devez vous assurer que les chiffres publiés correspondent aux totaux des données qui vous ont été fournies.

La feuille de calcul comporte 65.536 lignes

Le nombre maximal de lignes autorisées dans une feuille de calcul d'une ancienne version d'Excel était de 65 536. Si vous recevez un jeu de données avec ce nombre de lignes, vous avez probablement reçu un jeu de données incomplet. Les nouvelles versions d'Excel permettent 1.048.576 lignes, donc il est moins probable que vous utilisiez des jeux de données qui atteignent cette limite.

La feuille de calcul comporte 255 colonnes

L'application Numbers d'Apple ne peut gérer que des feuilles de calcul contenant 255 colonnes et elle tronque les fichiers comportant davantage de colonnes sans avertir l'utilisateur. Si vous recevez un jeu de données avec exactement 255 colonnes, demandez si le fichier a été ouvert

ou converti avec Numbers.

La feuille de calcul comporte des dates en 1900 ou 1904

Pour d'obscures raisons, la date par défaut d'Excel à partir de laquelle le logiciel compte toutes les autres dates est le 1er janvier 1900. Mais si vous utilisez Excel pour Mac, ce sera le 1er janvier 1904. Si vous repérez ces dates dans vos données, il s'agit probablement d'un problème.

Textes convertis en nombres

Tous les chiffres ne sont pas des nombres. Par exemple, un code peut être utilisé sans qu'il soit pour autant un nombre et qu'il signifie donc toute autre chose. Essayer de convertir ce type de données en nombres peut entraîner toute une série de problèmes, y compris si vous essayez de les convertir dans un autre format de fichier ou de les fusionner avec d'autres jeux de données.

Numéros stockés en tant que texte

Lorsque vous travaillez avec des feuilles de calcul, les numéros peuvent être stockés sous forme de texte avec un formatage indésirable. Cela arrive souvent lorsqu'une feuille de calcul est optimisée pour présenter des données plutôt que de les rendre disponibles pour une réutilisation. Par exemple, au lieu de représenter un million de dollars avec le nombre "1000000", une cellule peut contenir la chaîne "1,000,000" ou "1 000 000" ou "USD 1 000 000" avec le formatage des virgules, unités et espaces. Excel peut prendre en charge certains cas simples mais vous devrez souvent utiliser des formules pour supprimer les caractères indésirables, jusqu'à ce que les cellules soient suffisamment propres pour être reconnues comme des nombres. Une bonne pratique consiste à stocker des numéros sans formatage et à inclure des informations de contexte dans les étiquettes de colonnes ou les métadonnées.

2. Problèmes que vous pouvez résoudre

Problème d'encodage

Toutes les lettres sont représentées par des ordinateurs sous la forme de nombres. Les problèmes d'encodage sont des problèmes qui surviennent lorsque le texte est représenté par un ensemble spécifique de nombres (appelé "encodage"). Cela peut donner lieu à du texte illisible, parsemé de caractères spéciaux incompréhensibles. Votre source devrait être en mesure de vous indiquer le type d'encodage du jeu de données.

Les fins de ligne sont confuses/brouillées

Tous les fichiers de textes et de « données de texte », tels qu'au format CSV, utilisent des caractères invisibles pour représenter les extrémités des lignes. Les ordinateurs Windows, Mac et Linux sont historiquement en désaccord sur ce que devraient être ces caractères de fin de ligne. La tentative d'ouverture d'un fichier enregistré sur un système d'exploitation à partir d'un autre système d'exploitation peut parfois empêcher Excel ou d'autres applications d'identifier correctement les sauts de ligne. En règle générale, il est facile de résoudre ce problème en ouvrant simplement le fichier dans un éditeur de texte général et en le réenregistrant. Si le fichier est

exceptionnellement volumineux, vous devrez peut-être envisager d'utiliser un outil de ligne de commande ou d'obtenir l'aide d'un développeur.

Données au format PDF

Un nombre important de jeux de données n'est disponible qu'au seul format PDF. Pour extraire les données d'un document PDF, il existe plusieurs outils dont Tabula (gratuit). Toutefois, si vous disposez d'Adobe Creative Cloud, vous avez normalement accès à Acrobat Pro, qui dispose d'une fonctionnalité permettant d'exporter des tableaux vers Excel. L'une ou l'autre solution devrait être capable d'extraire la plupart des données tabulaires d'un PDF.

Données trop granulaires

C'est le contraire des données trop grossières. Dans ce cas vous avez des régions, mais vous voulez des états ou vous avez des mois mais vous voulez des années. Les données peuvent être agrégées en utilisant la fonction de tableau croisé dynamique d'Excel ou de Google Docs, en utilisant une base de données SQL ou en écrivant du code personnalisé. Les tableaux croisés dynamiques sont un outil fabuleux que tous les journalistes devraient apprendre, mais ils ont leurs limites. Pour les jeux de données exceptionnellement volumineux, il est préférable de solliciter un développeur.

Encodage humain des données

La saisie de données par des humains est un problème si courant que ses symptômes sont mentionnés dans au moins 10 des autres problèmes décrits dans ce document. Les erreurs sont humaines et sans procédure stricte de validation, des problèmes de toutes natures peuvent se poser. Surtout, méfiez-vous des données encodées par les utilisateurs qui n'ont sans doute pas la moindre idée de ce à quoi peut se rapporter le concept de qualité des données.

Données mélangées avec le formatage et les annotations

Les représentations complexes de données telles que HTML et XML permettent une séparation nette entre les données et leur mise en forme mais ce n'est pas le cas pour les formats tabulaires. Un des problèmes courants consiste à annoter ou décrire les données dans l'étiquetage des colonnes. Des lignes en tête de fichier peuvent être dupliquées ou le fichier peut comporter plusieurs feuilles qui comporteront plusieurs tables, lesquelles peuvent présenter un étiquetage différent. La solution consiste simplement à identifier le problème.

Agrégations calculées avec des valeurs manquantes

Imaginez un jeu de données avec 100 lignes et une colonne appelée "coût". Dans 50 lignes, la colonne des coûts est vide. Quelle est la moyenne de cette colonne? En général, si vous voulez calculer des agrégats sur des colonnes qui ne sont pas complètes, vous pouvez le faire en filtrant les lignes manquantes. Dans certains cas, la valeur manquante pourra être légitimement interprétée comme égale à zéro. Si vous n'en êtes pas certain, demandez à un expert. Sinon, cela peut conduire à une erreur d'analyse, une erreur que d'autres peuvent reprendre et ensuite transmettre.

Echantillon non aléatoire

Une erreur d'échantillonnage non aléatoire se produit lorsqu'une enquête ou tout autre jeu de données échantillonnées échoue intentionnellement ou accidentellement à couvrir l'ensemble de la population. Cela peut se produire pour une variété de raisons allant de l'heure de la journée à la langue maternelle du répondant. C'est une source fréquente d'erreur dans la recherche en sociologie. Cela peut aussi se produire pour des raisons moins évidentes, par exemple lorsqu'un chercheur pense avoir un ensemble de données complet et qu'il choisit de ne travailler qu'avec une partie de celui-ci. Si le jeu de données original était incomplet pour une raison quelconque, les conclusions tirées de leur échantillon seront incorrectes. La seule chose que vous pouvez faire pour corriger un échantillon non aléatoire est d'éviter d'utiliser ces données.

Marge d'erreur trop importante

La marge d'erreur est généralement associée aux données d'enquêtes. En règle générale, vous devriez être prudent dans l'utilisation de ce type de données, surtout si la marge d'erreur dépasse les 10%.

Marge d'erreur inconnue

Parfois, le problème n'est pas que la marge d'erreur soit trop grande, c'est que personne n'a jamais pris la peine de la comprendre. C'est un problème récurrent dans les sondages non-scientifiques. Sans calculer une marge d'erreur, il est impossible de savoir si les résultats sont exacts. En règle générale, chaque fois que vous avez des données provenant d'un sondage, vous devriez vous poser la question de la marge d'erreur. Si votre source ne peut pas vous le dire, ces données ne valent probablement pas la peine d'être utilisées pour une analyse sérieuse.

Echantillon biaisé

À l'instar d'un échantillon qui n'est pas aléatoire, un échantillon biaisé résulte d'un manque de soin dans la façon dont l'échantillonnage a été exécuté (ou d'une volonté de le déformer). Un échantillon pourrait être biaisé parce que le sondage a été réalisé sur Internet et que les personnes les plus pauvres n'utilisent pas Internet aussi fréquemment que les riches. Les enquêtes doivent être soigneusement pondérées pour s'assurer qu'elles couvrent des segments proportionnels de toute population qui pourraient fausser les résultats. Il est presque impossible de le faire parfaitement...

Données éditées manuellement

L'édition manuelle de données donne lieu quasi aux mêmes problèmes qu'un encodage humain des données sauf que cela arrive après coup. En fait, les données sont souvent éditées manuellement dans le but de les corriger. Les problèmes d'édition manuelle sont une des raisons pour lesquelles vous devez toujours vous assurer que la provenance de vos données soit bien documentée. Dans la mesure du possible, essayez d'obtenir le jeu de données original ou tout au moins sa version la plus ancienne.

L'inflation fausse les données

L'inflation monétaire signifie que l'argent change de valeur dans le temps. Il n'y a aucun moyen de dire si les chiffres ont été ajustés en fonction de l'inflation. Si vous obtenez des données et que vous n'êtes pas sûr qu'elles aient été ajustées, vérifiez auprès de votre source. Si ce n'est pas le cas, vous voudrez probablement effectuer cet ajustement. Des calculateurs d'inflation sont disponibles en lignes : <https://calculis.net/inflation>, <https://fxtop.com/fr/calculateur-inflation-entre-deux-dates.ph>

Des variations naturelles / saisonnières faussent les données

De nombreux types de données fluctuent naturellement en raison de certaines forces sous-jacentes. L'exemple le plus connu est celui de l'emploi fluctuant avec les saisons. Les économistes ont développé une variété de méthodes pour compenser cette variation. Les détails de ces méthodes ne sont pas particulièrement importants, mais il est important que vous sachiez si les données que vous utilisez ont été « désaisonnalisées ». Si ce n'est pas le cas et que vous voulez comparer le volume de l'emploi d'un mois à l'autre, vous voudrez probablement obtenir des données ajustées auprès de votre source (il est ici beaucoup plus difficile de les ajuster soi-même, contrairement à l'inflation).

Manipulation des périodes de temps

Une source peut accidentellement ou intentionnellement déformer la réalité en fournissant des données qui démarrent ou s'arrêtent à un moment précis. Par exemple, les données relatives à une « vague de criminalité nationale », qui furent diffusées en 2015, auraient dû être analysées sur une période plus longue : les journalistes auraient vu que les crimes violents étaient plus élevés aux Etats-Unis dix ans auparavant. Si vous disposez de données qui couvrent une période limitée, essayez d'éviter de commencer vos calculs dès la toute première période à laquelle commence le jeu de données.

Manipulation du cadre de référence

Les statistiques criminelles sont souvent manipulées à des fins politiques en les comparant à une année où le crime était très élevé. Cela peut être exprimé soit comme un changement (en baisse de 60% depuis 2004) ou via un indice (40, où 2004 = 100). Dans l'un ou l'autre de ces cas, 2004 peut ou non être une année appropriée pour la comparaison. Cela aurait pu être une année de criminalité anormalement élevée. Dans la mesure du possible, essayez de comparer les taux de plusieurs points de départ différents pour voir comment les chiffres changent.

3. Problèmes qu'un expert peut vous aider à résoudre

Source ou auteur non fiables

Parfois, les seules données dont vous disposez proviennent d'une source à propos de laquelle il est raisonnable de douter. Dans certaines situations, c'est très bien. Les seules personnes qui savent combien d'armes sont fabriquées sont les fabricants d'armes à feu. Cependant, si vous

ces données proviennent d'un fabricant d'armes douteux, vérifiez toujours avec un autre expert. Mieux encore, vérifiez vos données avec deux ou trois experts. Ne publiez pas de données provenant d'une source biaisée à moins d'avoir des preuves permettant de les corroborer de manière substantielle.

Opacité du processus de collecte des données

Il est très facile d'introduire de fausses suppositions ou des erreurs dans les processus de collecte des données. Pour cette raison, il est important que les méthodes utilisées soient transparentes. Il est rare que vous sachiez exactement comment un jeu de données a été recueilli mais les indications d'un problème peuvent inclure des nombres qui affirment une précision irréaliste ou des données qui sont trop bonnes pour être vraies. Parfois, l'histoire d'origine peut être louche : est-ce que tel ou tel universitaire a vraiment interviewé 50 membres de gangs actifs du côté sud de Chicago ? Si la façon dont les données ont été recueillies semble douteuse et que votre source ne peut vous fournir des données documentées, vous devriez toujours vérifier auprès d'un autre expert que les données auraient pu raisonnablement être collectées de la manière décrite.

Précision irréaliste des données

En dehors des sciences dures, peu de choses sont mesurées de manière routinière avec plus de deux décimales. Si un jeu de données prétend montrer les émissions d'une usine à la 7^e décimale. Cela ne constitue peut-être pas un problème en soi mais il est important d'être transparent au sujet des estimations.

Valeurs aberrantes inexplicables

Les valeurs aberrantes donnent lieu à des erreurs statistiques, notamment lorsque des moyennes sont utilisées. Un bon réflexe est d'examiner chaque nouveau jeu de données pour déterminer les valeurs les plus grandes et les plus petites, et vous assurer que celles-ci se trouvent dans une fourchette raisonnable. Si les données le justifient, vous pouvez également effectuer une analyse plus rigoureuse sur le plan statistique.

Un index masque la variation sous-jacente

Les analystes qui veulent suivre la tendance d'un problème créent souvent des indices de diverses valeurs pour en suivre l'évolution. Il n'y a rien d'intrinsèquement mauvais dans l'utilisation d'un index. Ils peuvent avoir un grand pouvoir explicatif. Cependant, il est important de se méfier des indices qui combinent des mesures disparates. Par exemple, l'indice des inégalités de genre des Nations Unies combine plusieurs mesures liées aux progrès des femmes vers l'égalité. L'une des mesures utilisées est la « représentation des femmes au parlement ». Deux pays dans le monde ont des lois qui imposent la représentation des deux sexes dans leurs parlements : la Chine et le Pakistan. En conséquence, ces deux pays obtiennent de bien meilleurs résultats grâce à cet indice.

Résultats piratés

Le piratage peut consister dans la modification intentionnelle des données ou de l'analyse statistique, ou encore dans le signalement d'une sélection de résultats statistiquement significatifs. Si vous tombez sur ce type de cas, il faut arrêter de collecter des données une fois que vous avez obtenu un résultat significatif, supprimer des observations pour obtenir un résultat significatif, ou effectuer de nombreuses analyses et ne rapporter que celles qui seront les plus significatives²⁹. Si vous souhaitez publier les résultats d'une étude, vous devez comprendre ce qu'est la valeur P, ce qu'elle signifie puis prendre une décision éclairée quant à la pertinence de l'utilisation des résultats.

Echec de la loi de Benford

La loi de Benford est une théorie qui dispose que les petits chiffres (1, 2, 3) apparaissent beaucoup plus fréquemment au début des nombres que les grands chiffres (7, 8, 9). En théorie, la loi de Benford peut être utilisée pour détecter des anomalies dans des pratiques comptables ou des résultats d'élections même si, dans la pratique, elle peut facilement être mal appliquée. Si vous pensez qu'un jeu de données a été créé ou modifié pour tromper, la loi de Benford est un excellent premier test, mais vous devriez toujours vérifier vos résultats avec un expert avant de conclure que vos données ont été manipulées.

Trop bon pour être vrai

Il n'existe pas de jeu de données global relatif à l'opinion publique. Personne ne connaît le nombre exact de personnes vivant en Sibérie. Les statistiques de la criminalité ne sont pas comparables d'un pays à l'autre. Méfiez-vous des données qui prétendent représenter quelque chose que vous ne pourriez pas parvenir à connaître. Ce ne sont pas des données. Il s'agit de l'estimation de quelqu'un et elle est probablement fausse. Mais aussi... cela pourrait être une bonne histoire : demandez à un expert de vérifier pour vous.

4. Problèmes qu'un développeur peut vous aider à résoudre

Données agrégées dans une mauvaise catégorie ou zone géographique

Parfois, vos données présentent un bon niveau de détail (ni trop grossier ni trop granulaire, mais elles ont été agrégées dans un groupe différent de celui que vous souhaitiez. L'exemple classique est celui des données agrégées via des codes postaux que vous préféreriez obtenir par quartier. Dans de nombreux cas, il s'agit d'un problème impossible à résoudre sans obtenir de votre source des données plus granulaires, mais parfois les données peuvent être mappées proportionnellement d'un groupe à l'autre. Cela ne doit être entrepris qu'avec une compréhension minutieuse de la marge d'erreur qui peut être introduite dans le processus. Si vous avez agrégé des données dans des mauvais groupes, demandez à un développeur s'il est possible de les regrouper.

²⁹ Voir : <http://fivethirtyeight.com/features/science-isnt-broken>)

Données dans un document scanné

Grâce aux législations réglementant l'accès aux données publiques, il est fréquent que les autorités soient obligées de vous fournir des données, même si elles ne le souhaitent pas vraiment. Une technique courante consiste alors à vous fournir des scans ou des photographies des pages de données. Il peut s'agir d'un fichier image (JPG, par exemple) ou, plus probablement, d'un fichier PDF.

Il est possible d'extraire du texte à partir d'images et de le restituer en données via un processus de reconnaissance optique des caractères (OCR ou OCR-isation). Si cette technique permet la précision dans la plupart des cas, cela va dépendre aussi de la nature du document. Chaque fois que vous utiliserez OCR pour extraire des données, vous devrez mettre en place une procédure permettant de valider la correspondance des données avec le fichier original. De nombreux sites web proposent logiciels de conversion. Il existe également des outils gratuits qu'un développeur pourra adapter à des documents spécifiques. Si vous en avez un sous la main, demandez-lui la meilleure stratégie à adopter pour le document dont vous disposez.

2.4.6 Visualisation de données

La visualisation de données a pour objectif de rendre compréhensible de grandes quantités de données qui ne pourraient pas l'être autrement et, partant, de permettre leur analyse. Cette discipline, qui allie art et fonctionnalités évidentes, est à mettre en relation avec le mode de perception humain : *"le cerveau n'enregistre pas seulement l'information que l'on voit. Il crée aussi des images visuelles mentales"* (Cairo 2012). Dès lors, les graphiques, tableaux et cartes ne sont pas seulement des outils destinés à être "vus" : ils doivent aussi être lus. Comment apprécier la valeur informative d'une visualisation ? *"Les graphiques ne doivent pas simplifier les messages. Ils doivent les clarifier"*. En datajournalisme, la visualisation de données remplit deux fonctions distinctes : celle d'un appui à l'analyse des données, pour repérer des tendances que l'on n'aurait pas pu observer autrement ; et celle d'un outil d'information, l'objet étant de présenter l'information aux audiences de manière simple et compréhensible au premier regard. Pour Jacques Bertin, les trois fonctions de la représentation graphique sont d'enregistrer l'information, de la communiquer en créant une image simple et mémorisable, et de la traiter. Il définit la représentation graphique comme *"la transcription (...) d'une 'information' connue par l'intermédiaire d'un système de signes quelconques. La représentation graphique est une partie de la sémiologie, science qui traite tous les systèmes de signes"*. Il identifie six variables visuelles (ou rétinienne) : la taille, la valeur, le grain, la couleur, l'orientation et la forme. A cela s'ajoutent quatre niveaux d'ordre de ces variables : semblables, différentes, ordonnées et proportionnelles. Avec ces variables indépendantes, *"la graphique offre, pour chaque information, un choix illimité de constructions"*. Il ajoute qu'un graphique, pour être lisible, ne devrait pas être en trois dimensions, et que l'efficacité d'un graphique se mesure à son temps d'observation : plus il sera court, plus vite l'information sera comprise [12].

Le choix d'une visualisation de données s'opère en fonction de l'échelle utilisée (Grosjean & Dommergues 2011) : un diagramme en barres ou circulaire pour une échelle nominale, un diagramme en barres pour une échelle ordinale, un histogramme pour une échelle d'intervalles

ou de rapports, des courbes pour les classes (variables regroupées comme par exemple, une tranche d'âges), un tableau à entrées multiples lorsque le graphique comporte plus de deux variables. Les statisticiens préconisent également le respect de la convention des trois-quarts pour la longueur des axes : la longueur de l'ordonnée (Y) correspond environ aux trois-quarts de la longueur de l'abscisse (X). Ils insistent également sur la présence d'un titre et de légendes pour les axes X et Y. De plus, la valeur maximale sur l'ordonnée (Y) ne devrait pas dépasser la donnée la plus élevée [52].

A l'image de la représentation géographique qui propose une "*certaines perception de la planète (...) et une certaine image du monde*"³⁰, une représentation graphique n'est pas neutre : il s'agit toujours d'un travail d'interprétation lié à une analyse des données. L'excellence graphique, selon Tufte (1983), consiste en "*une communication efficace d'idées quantitatives complexes*". Un graphique devrait idéalement rencontrer huit critères : montrer les données, induire le lecteur à penser à la substance plutôt qu'à la méthodologie, éviter de déformer ce que disent les données, présenter beaucoup de chiffres dans un petit espace, encourager l'œil à comparer différents types de données, rendre cohérents de grands jeux de données, servir un propos clair et raisonnable, être intégré dans la description statistique et verbale du jeu de données [117]. Tufte souligne que les visualisations consistent en des affichages symboliques qui révèlent les données. Ces affichages symboliques sont régis par une série de bonnes pratiques, relevant du domaine de la sémiologie graphique. Les visualisations de données s'y conforment (rédaction de titres et de légendes, identification non équivoque des variables), de manière à ce que l'information graphique puisse être comprise lorsqu'elle est détachée de son contexte.

Cleveland (1985) pose, quant à lui, quatre conditions pour qu'une visualisation de données soit efficace : l'utilisation d'éléments proéminents pour souligner les données et attirer l'attention sur leur représentation, l'utilisation d'une légende, l'absence d'efforts pour comprendre les données et le caractère itératif du graphique [26]. Pour Dona Wong (*The Washington Post*), un graphique d'information doit contenir trois éléments essentiels : un contenu riche, une invitation à la visualisation et une exécution sophistiquée. L'efficacité d'un graphique dépendra également de sa couleur, des typographies utilisées et du design... à condition qu'ils soient bien utilisés en adéquation avec l'information. Les bonnes pratiques sont, selon elle, une utilisation de la couleur et de la typographie pour accentuer le message-clé (et pas pour décorer!), une utilisation de comparaisons correctes (on ne compare pas des pommes à des poires!) ou encore celles de légendes pour identifier les variables utilisées. Mais de la qualité des données dépendra également la qualité de la visualisation.

"*La meilleure visualisation des données, indépendamment de son format et de sa présentation, indique-t-il, est celle qui permet de voir ce que les données ont à dire*", indique Yau (2013), étant entendu qu'une donnée constitue, en soi, un élément de représentation du monde réel. La visualisation de données implique un travail de recherche et de vérification des faits, mais aussi

³⁰ Lire "Data+Design", Infoactive et Donald W. Reynolds Journalism Institute, e-book, <https://infoactive.co/data-design>

sur des compétences graphiques. Car avant de commencer à travailler sur la visualisation, il s'agit de collecter les données et d'y détecter/corriger leurs éventuelles anomalies.

Mais en journalisme, souligne Alberto Cairo (2016), l'objectif du design est avant tout de donner sa place à l'information : "*La qualité de vos graphiques dépend fondamentalement de la qualité de votre reportage ou recherche, pas juste de quel bon designer vous êtes*". Ce qui implique la nécessité de clarté et de concision mais sans en faire trop ainsi que celle de bien structurer l'information. Si les valeurs du journalisme sont celles de la vérité, de la loyauté, de la vérification, de l'indépendance, de la critique, du sens à donner, de la compréhension et de la conscience professionnelle, Cairo estime que celles-ci ne sont pas exclusives à la profession : ces valeurs sont aussi citoyennes. Une infographie consiste en une représentation visuelle de l'information dans l'intention de communiquer un ou plusieurs messages spécifiques. Une data visualisation est un affichage de données qui engendre de l'analyse, de l'exploration et de la découverte. Une news application est une forme spéciale de visualisation qui permet aux gens de rapporter les données présentées à leur propre vie. Les cinq qualités d'une bonne visualisation de données : inspirer confiance, être fonctionnelle, être belle, donner du sens et du relief. Cairo insiste également sur sa nécessité d'être compréhensible et, lorsque le sujet le permet, de l'être peu importe la langue de l'utilisateur.

Pratiquement, le choix d'une visualisation de données sera toujours fonction du type de données qui sera traité : on privilégiera le graphique en barre pour comparer des variables (les barres doivent passer au mode horizontal dès que l'on a plus de 4 ou 5 variables, de manière à faciliter la lecture de la dataviz), le graphique en courbe montrera quant à lui une évolution dans le temps, le camembert montrera des proportions (mais il ne doit pas comporter trop de variables pour rester lisible). Il est donc important de bien tenir compte du type d'échelle de mesure utilisé. Pour une analyse plus complète, on utilise généralement une échelle d'intervalle ou de rapport.

ECHELLES DE MESURE

Echelle nominale : on nomme des catégories en utilisant des nombres nominaux (numéros sur des tenues de sport, codes postaux...)

Echelle ordinale : marque l'ordre (position d'arrivée dans une course)

Echelle d'intervalle : marque les quantités (degrés Celsius, dates du calendrier...), ne contient pas de 0 absolu

Echelle de rapports : marque la quantité, l'ordre de grandeur... On peut faire des multiplications et des divisions sans que le rapport des nombres soit modifié, en raison de la présence du 0 absolu (poids, âge)

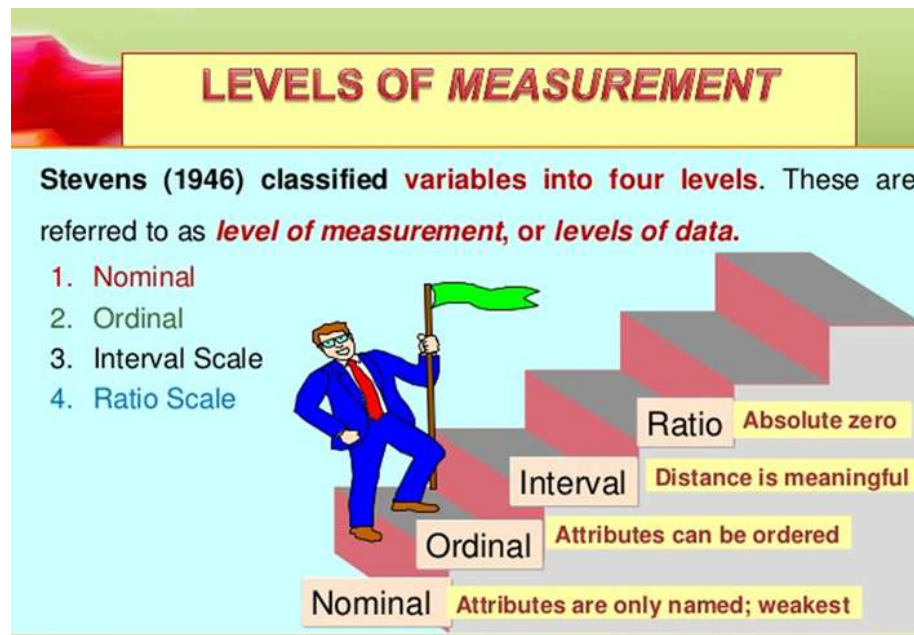


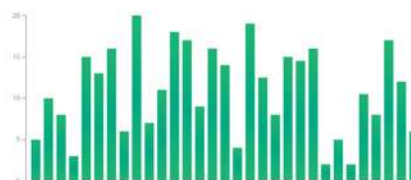
FIGURE 2.5 – Typologie des échelles de mesure établie par Stanley (1946), source : https://www.slideshare.net/amina_h/measurement-scales-38614023

Choisir la bonne visualisation

A. Echelle nominale

Diagramme en barres ou diagramme circulaire (en secteurs)

Données mesurées à l'aide d'une échelle nominale. Montre les proportions, les pourcentages et répond à la question "combien?". Les graphiques en barres utilisent des barres horizontales ou verticales, dont l'objet est de montrer des comparaisons numériques discrètes entre des catégories (variables catégorielles). Ils sont à distinguer des histogrammes, car il ne s'agit pas de développements continus avec intervalle. Lorsque le nombre de barre est trop élevé, rendant la lecture des étiquettes illisibles, on privilégiera les barres horizontales.



Les camemberts (ou graphiques en secteurs) affichent des proportions (pourcentages) entre des catégories (le total est de 100 %). S'ils sont efficaces pour montrer une distribution proportionnelle, l'affichage du nombre de valeurs est limité pour des raisons de lisibilité. Une alternative est alors d'utiliser des graphiques en barres empilées. Par ailleurs, ils ne permettent pas des comparaisons précises.

B. Echelle ordinale

Diagramme en barres

L'ordre est déjà défini, il est à respecter sur le graphique. Il s'agit d'une échelle d'intervalles ou de rapports.

Histogramme

Utilise les mêmes codes et convention qu'un diagramme en barres mais les barres sont "collées". Son objet est de montrer une distribution de valeurs sur un intervalle continu ou une période donnée. Il aide à donner une estimation de la concentration de valeurs et à identifier les extrêmes ou valeurs inhabituelles.



Courbe ou polygone

Affiche des valeurs quantitatives sur un intervalle ou une période continue. La ligne sera privilégiée lorsqu'il s'agit d'identifier des tendances et de voir comment les données ont évolué au fil du temps. Travailler avec plusieurs variables permet de pouvoir les comparer entre elles (plusieurs courbes) mais si elles sont trop nombreuses, cela va nuire à la lisibilité du graphique. Les graphiques par zone (de forme polygonale) en constituent une variante, la distance entre le point représentant la valeur et l'abscisse est colorée. Ils sont davantage utilisés pour montrer des tendances plutôt que pour montrer des valeurs spécifiques.



Classes

Variables regroupées : taille 32-34, de 20 à 25 ans... l'objectif est de réduire le nombre de modalités sur l'abscisse. Amplitude = intervalle à l'intérieur d'une classe.

Courbe des effectifs

Par exemple : le pourcentage d'hommes qui mesure au moins 1,79m.

Courbe des fréquences cumulées

Par exemple : le pourcentage cumulé de la taille d'un groupe d'hommes.

Trois types de courbes : courbe de Gauss (courbe normale, ci-dessous), courbe pointue (plus haute), courbe plate (plus écrasée).

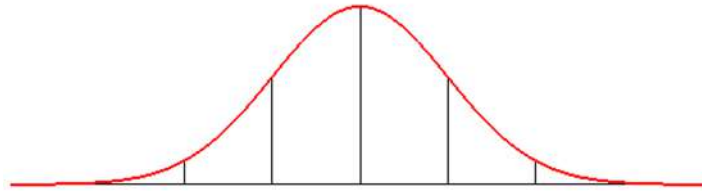
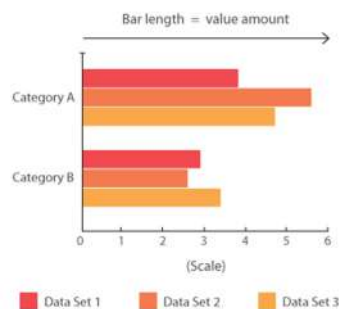
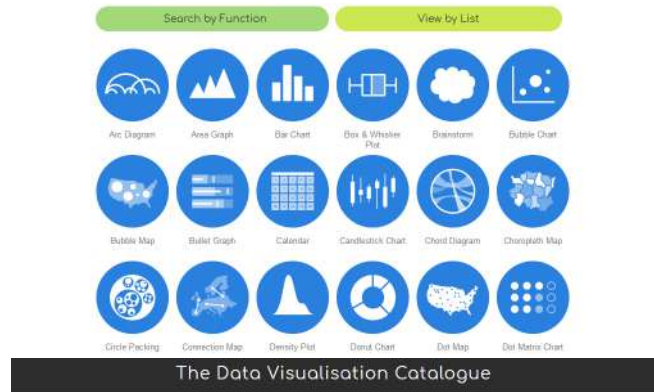


Tableau à entrées multiples

Qui comporte au moins deux variables (hommes et femmes, par ex.) Ici, comparaison de deux groupes comportant trois variables.



Il existe quantité d'autres formes de graphiques et le choix seront toujours fonction des variables et de leurs valeurs : les diagrammes en arbres représenteront des relations hiérarchiques, les cartogrammes vont permettre de représenter des variations de couleurs en fonction des variations de données, le graphique en donuts sera utilisé pour la présentation de résultats électoraux, les diagrammes de Venn serviront à visualiser des ensembles, les diagrammes de flux (diagramme de Sabkey ou dendogrammes) permettront d'établir des relations entre deux faits (par exemple, des sociétés établies dans un pays européen et les sommes d'argent qu'elles ont envoyé vers des paradis fiscaux), tandis que les graphiques en arbre ("treemaps") permettent de visualiser la proportion de différentes variables se rapportant à un seul phénomène. Par ailleurs, lors de l'analyse des données, les graphiques en points seront utiles pour établir des comparaisons, ou pour identifier des tendances ou des relations. Généralement, les outils de visualisation de données "prêts à l'emploi" proposeront automatiquement une représentation en fonction des données que vous lui fournissez (mais cela ne veut pas nécessairement dire que cette suggestion soit la plus pertinente). Le Dataviz Catalog présente un catalogue complet et documenté des différentes formes de visualisation de données, exemples et outils à l'appui : <https://datavizcatalogue.com/>, ou la "cheatsheet" <http://www.infographicsblog.com/chart-suggestions-a-thought-starter-andrew-abela/>



Area Graph

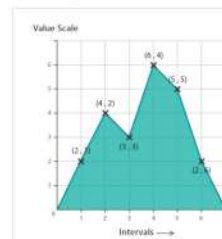
Description

Area Graphs are *Line Graphs* but with the area below the line filled in with a certain colour or texture. Area Graphs are drawn by first plotting data points on a Cartesian coordinate grid, joining a line between the points and finally filling in the space below the completed line.

Like *Line Graphs*, Area Graphs are used to display the development of quantitative values over an interval or time period. They are most commonly used to show trends, rather than convey specific values.

Two popular variations of Area Graphs are: grouped and *Stacked Area Graphs*. Grouped Area Graphs start from the same zero axis, while *Stacked Area Graphs* have each data series start from the point left by the previous data series.

Anatomy



Functions

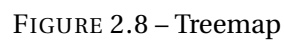
Patterns Data over time

FIGURE 2.6 – Source : <https://datavizcatalogue.com/>



FIGURE 2.7 – Diagramme de Sankey

WWW.TJG.BE/WAGENPAK



Eviter les pièges

Changer d'échelle de couleur peut modifier la perception que l'on peut avoir d'un phénomène observé. De la même manière, une graduation de l'axe Y doit toujours commencer par 0 pour ne pas donner un effet déformé de la réalité observée. Il en va de même pour l'axe des X, par exemple lorsque l'on observe une échelle de temps : deux temporalités peuvent donner lieu à des interprétations diamétralement opposées. Par ailleurs, il faut absolument éviter les représentations graphiques en trois dimensions car celles-ci faussent la perception de la valeur des données. Toujours au rayon des bonnes pratiques, ne comparer que ce qui est comparable (on ne compare pas des pommes et des poires). Et ce n'est pas parce que l'on peut observer une même tendance pour deux phénomènes différents qu'il y a corrélation entre les deux : le fait que la courbe de la consommation de margarine suit celle des divorces ne signifie pas que consommer de la margarine influe sur le divorce et vice-versa (Cairo 2019).

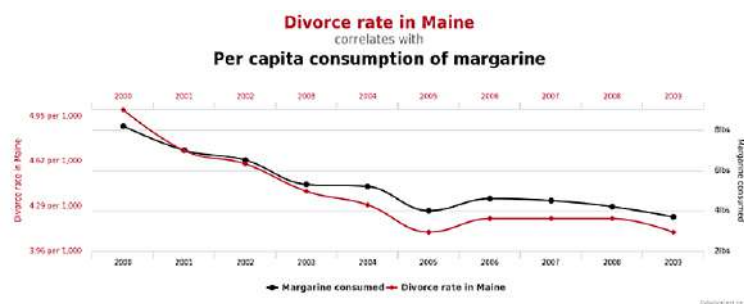


FIGURE 2.10 – Gare aux corrélations!

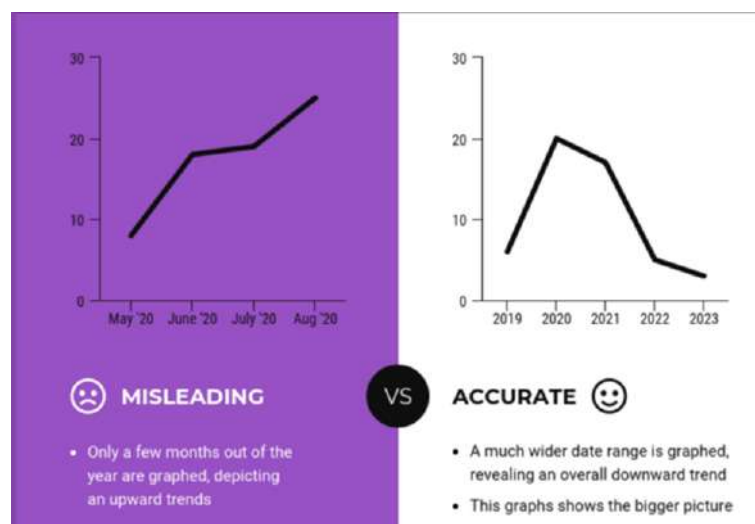


FIGURE 2.11 – L'ordonnée ne débute pas à 0.

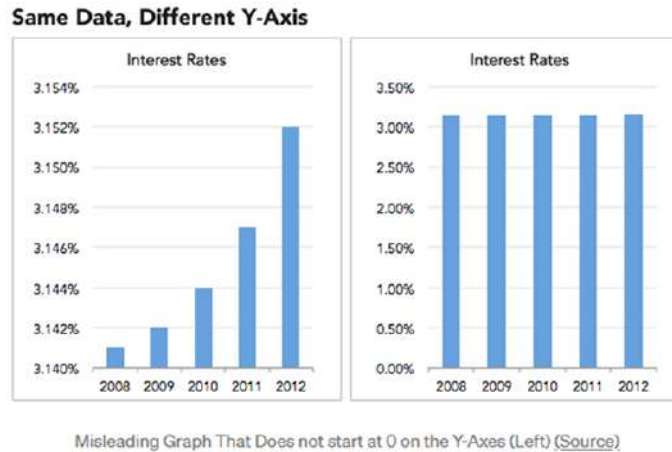


FIGURE 2.12 – Modification de l'échelle temporelle (abscisse)

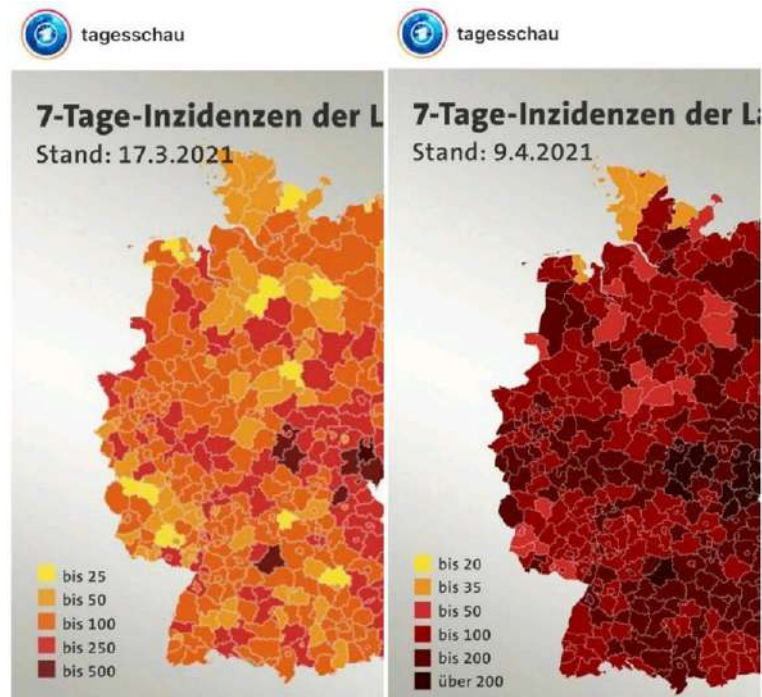


FIGURE 2.13 – Changement d'échelle de couleur.

Quelques bonnes pratiques

- Lisibilité : visualisation en deux dimensions
- Choix d'une visualisation de données en fonction de l'échelle de mesure utilisée
- Respecter la convention des pourcentages pour la longueur des axes
- Une visualisation comporte un titre et ses axes X et Y sont légendés
- La valeur maximum sur l'ordonnée (Y) ne doit pas trop dépasser la donnée la plus élevée
- Utiliser un système sobre, ne pas chercher à mettre en avant certaines données à l'aide

de couleur différente, par exemple (ce qui pourrait fausser la représentation ou accentuer un phénomène au détriment d'un autre)

- Présence d'un titre et de légendes pour les axes X et Y Doit comporter de l'information
- Doit être claire et lisible (diagrammes en barre : horizontal lorsque trop de variables ; en secteurs : trop de variables nuisent à la lisibilité)
- Doit être documentée
- Bonne pratique : jusque 100% et pas jusque 80% et la graduation commence à 0
- Pas une fin en soi : les chiffres peuvent être intégrés dans le texte
- La visualisation pour l'analyse ne sera pas forcément celle de la diffusion (ex. bubble charts)

Cartographie

En cartographie statistique, une carte choroplèthe permet une représentation géographique par zone de couleur dont l'intensité va dépendre de la valeur. Il existe également d'autres types de représentations cartographiques comme la carte isoplèthe, qui relie des valeurs de même caractéristique (utilisée en météorologie), ou la carte isotherme qui représente des variations de température. Les représentations géographiques vont dépendre de deux types de fichiers : celui des données (valeur des variables observées), et généralement un fichier geojson, qui reprend les coordonnées de contour de chaque zone géographique. Les données de géocodage (latitude et longitude) permettent de pointer un endroit précis sur une carte³¹.

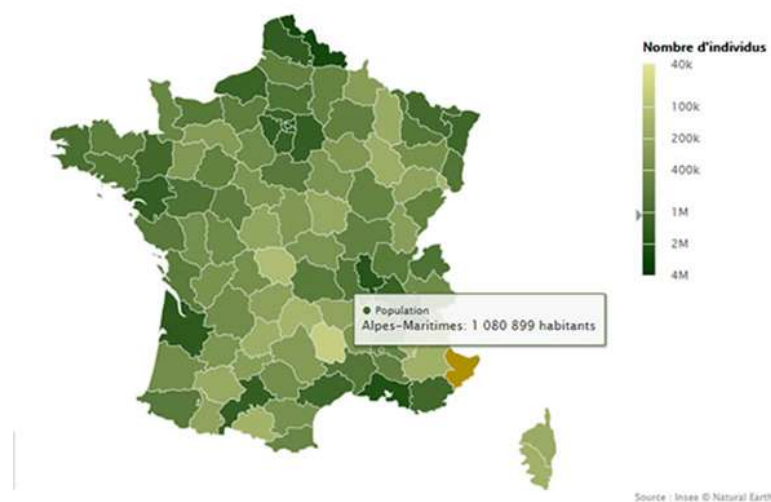


FIGURE 2.14 – Carte choroplèthe (population française).

³¹ Voir <https://geocode.localfocus.nl/>

3 | Outils du datajournalisme

Un processus journalistique piloté par des données comprend plusieurs étapes : la récolte des données, leur nettoyage, leur mise en contexte et leur combinaison (permet de confronter des données entre elles et d'infirmer / confirmer des hypothèses de travail). La dernière étape est la communication du travail journalistique, où la visualisation des données ne sera pas une fin en soi mais un outil journalistique qui sera privilégié pourvu d'être pertinent : il s'agit, avant tout, de faire sens.

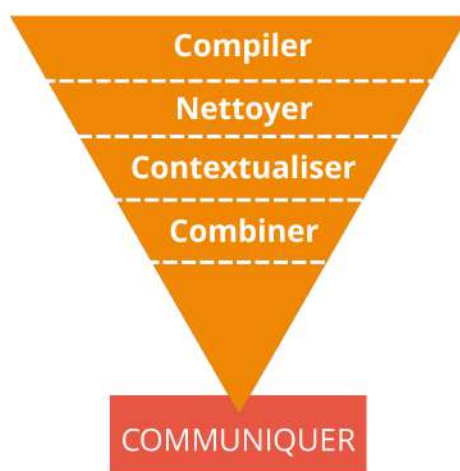


FIGURE 3.1 – Source : Online Journalism Blog (Paul Bradshaw) - <http://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/>

Au *Guardian Data Blog* par exemple, l'un des pionniers en matière de datajournalisme, le processus est sans cesse réadapté car de nouveaux outils et de nouvelles techniques en chassent régulièrement d'autres. "Certaines personnes pensent qu'il faut devenir une sorte de super hacker, écrire du code et manger du SQL au petit-déjeuner. Vous pouvez décider de suivre cette approche. Mais une grande partie de notre travail se fait simplement dans Excel. Tout d'abord, nous localisons les données ou nous les obtenons auprès de diverses sources, que ce soit des dépêches, des données gouvernementales, des études journalistiques, etc. Puis nous commençons à étudier ce qu'il est possible de faire avec ces données; devons-nous les recouper avec une autre base de données? Comment pouvons-nous illustrer leur évolution au fil du temps? Ces feuilles de calcul doivent souvent être nettoyées", expliquait Simon Roger, à l'époque où il travaillait au *Guardian*¹.

¹ Source : <http://jplusplus.github.io/guide-du-datajournalisme/pages/0303.html>

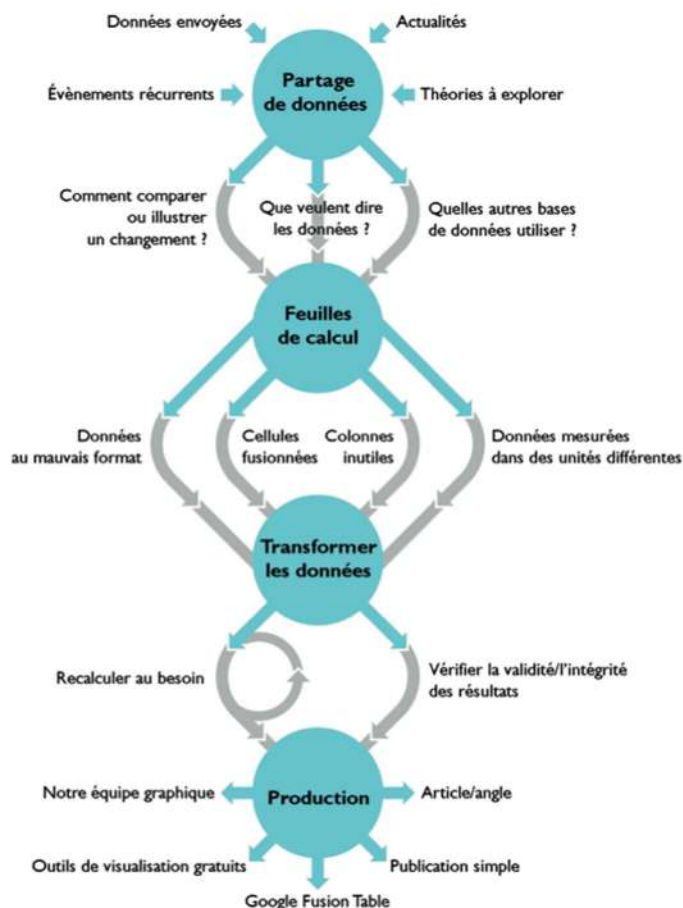


FIGURE 3.2 – Processus de datajournalisme au *Guardian* (2013)

Le plan de ce chapitre tient compte de toutes les tâches à effectuer dans le cadre d'un projet de datajournalisme : de la collecte à l'analyse en passant par la gestion des données.

3.1 Calculer

Il existe quantité d'outils gratuits ou payants susceptibles d'être mis au service d'une approche par données dans le journalisme. Certains nécessiteront des connaissances techniques avancées (dont la maîtrise d'un langage de programmation) tandis que d'autres restent accessibles moyennant l'apprentissage d'un logiciel ou du maniement d'un outil en ligne. Mais avant de se lancer dans un projet de datajournalisme, il est important de s'assurer de disposer des connaissances de base en mathématiques et en statistiques pour manipuler des données... et donc des chiffres.

3.1.1 Formules mathématiques de base

Convertir un nombre décimal en pourcentage² multiplier par 100 ou déplacer la virgule de deux caractères sur la droite.

Exemple : 0,888 = 8,88%

Convertir un pourcentage en nombre décimal

Diviser par 100 ou déplacer la virgule de deux caractères sur la gauche.

Exemple : 51,7% = 0,517

Pourcentage de X dans Y

$(X/Y) * 100$

Exemple : 3,5 hommes sur 10 = $(3/10) * 100 = 35\%$

Comparer l'écart en pourcentage entre X et Y

$(X/Y - 1) * 100$

Exemple : comparer l'écart entre 5 et 10 puis entre 10 et 5

$(5/10 - 1) = 1,5 * 100 = 150\%$ (X est 150% plus petit que Y)

$(10/5 - 1) = 1 * 100 = 100\%$ (X est 100% plus grand que Y)

Comparer une évolution en pourcentage entre X (nouvelle valeur) et Y (ancienne valeur)

$(X/Y - 1) * 100$ = calcul identique que pour comparer l'écart en pourcentage entre X et Y.

Calculer un taux sur 1000, sur 100, sur 10...

Par exemple, quel est le taux pour 1000 du nombre de personnes X sur une population Y?

Taux = $(X/Y) * UNITE = (X/Y) * 1000$

Calculer l'inflation des prix sur base de l'index des prix à la consommation

$(\text{Prix aujourd'hui} / \text{Prix hier}) = (\text{IPC aujourd'hui} / \text{IPC hier})$

Un produit X coûtait 10 euros en 2010, à l'époque l'IPC est de 101,7,6. En 2014, l'IPC était de 102,6. Quel était son prix en 2014?

$X (\text{inconnue} = \text{prix en 2010}) / 10 = \text{IPC 2014} / \text{IPC 2010} = 102,6 / 101,7 = (102,6/101,7) * 10 = 1,01 * 10 = 10,1 \text{ euros} \Rightarrow \text{il y a eu peu d'évolution.}$

Estimer le nombre de manifestants sur une place

1) Calculer la superficie de la place (L x l) en mètres carrés.

2) Estimer le nombre de personnes par mètre carré (un groupe dispersé = 1 personne par mètre carré).

3) Multiplier par le nombre de mètres carré. Pour une manifestation plus dense, multiplier la superficie par 1,3.

² Cette section est une adaptation de "Newsroom Math Cheat", Steve Doig, Mooc "Datajournalism, first steps, skills and tools, European Journalism Center 2014

3.1.2 Statistiques

Cette section constitue un résumé de "La statistique en clair" de François Grosjean et Jean-Yves Dommergues (éditions Ellipses, 2011)

3.1.2.1 Définitions

Statistiques

Ensemble des méthodes qui ont pour objet la collecte, le traitement et l'interprétation de données d'observation relatives à un groupe de personnes ou d'objets" (Grand dictionnaire de la psychologie, Larousse, 1991).

Variables

Caractéristique représentée par un nom ou un symbole. Variables catégorielles (qualitatives) : sexe, couleur, ... Variables quantitatives : valeurs numériques mesurables, par ex. poids, distance...

Variable quantitative discrète

Lorsque l'étendue des valeurs possibles est dénombrable (nombre de membres d'une famille, nombre de mots dans une phrase).

Variable quantitative continue

Lorsque les valeurs possibles ne sont pas parfaitement dénombrables (taille, poids, âge...)

Utilisation des variables dans le cadre d'une recherche

Variable indépendante (ou explicative) : ce qui est manipulé par le chercheur (ex. médicament – axe X, abscisse).

Variable dépendante (à expliquer) : mesure obtenue sur laquelle porte l'étude (ex. temps de guérison, axe Y, ordonnée)

Variable de contrôle : l'aspect qui doit être contrôlé pour ne pas interférer dans l'étude (par ex. âge)

Arrondi

3e chiffre ≥ 5 : 8,546 devient 8,55

3e chiffre < 5 : 8,332 devient 8,33

Fréquences

Fréquence absolue (n) : correspond au nombre d'objets dans une catégorie.

Voiture rouge : 5

Voiture verte : 4

Voiture bleue : 6

Voiture grise : 5

n = 20

Fréquence relative : rapport entre l'effectif d'une catégorie et l'effectif total.

Calcul de la fréquence de 13 mots de 50 syllabes dans une phrase : $(13/50) \times 100 = 26\%$

Echelles de mesure (rappel)

Echelle nominale : on nomme des catégories en utilisant des nombres nominaux (numéros sur des tenues de sport, codes postaux...)

Echelle ordinale : marque l'ordre (position d'arrivée dans une course)

Echelle d'intervalle : marque les quantités (degrés Celsius, dates du calendrier...), ne contient pas de 0 absolu

Echelle de rapports : marque la quantité, l'ordre de grandeur... On peut faire des multiplications et des divisions sans que le rapport des nombres soit modifié, en raison de la présence du 0 absolu. Ex. : taille, poids, âge... Une analyse sera plus complète en utilisant une échelle d'intervalle ou de rapport

Population

Ensemble complet d'éléments qui forme le champ d'analyse (individus, objets, événements).

Echantillon

Fraction représentative d'une population.

3.1.2.2 Statistiques descriptives

Permet de traiter les données, les organiser, les représenter sous la forme de tableaux ou de figure, les synthétiser. Mise en relation de deux ou plusieurs variables pour décrire l'intensité (force) et la forme d'une relation.

A. Mesures d'une tendance centrale

Calcul du mode

Dans une distribution, le mode (M_o) correspond à la valeur ou la catégorie qui possède l'effectif le plus élevé. Il est utilisé avec des données mesurées à l'aide d'une échelle nominale. L'effectif est le nombre de fois que l'on retrouve la même valeur dans une série donnée.

Exemple : 1 2 11 3 11 7 11 17 5 6 18 19

Le mode est 11 (valeur que l'on retrouve 3 fois)

Calcul de la médiane

La médiane (M_d) correspond à la valeur qui se trouve au centre d'une distribution de données. 3 étapes : ordonner les valeurs, trouver l'emplacement de la médiane à l'aide de la formule $M_d = (n+1) / 2$ (n = le nombre de valeur de la distribution) puis regarder la valeur correspondant à l'emplacement. Lorsque le résultat est décimal (5,5 par ex.), on fait la moyenne des deux valeurs situés au 5e et au 6e emplacement. De la même manière, il est possible de déterminer une classe médiane.

Calcul de la moyenne

Somme des valeurs divisée par leur nombre (ex : 38 variables ont un total de 333 = 333/38)

Moyenne pondérée : une moyenne par classe (car toutes les classes n'ont pas la même importance)

Mesures de dispersion

La dispersion est l'étalement ou la variation des données autour de la valeur centrale.

Indice de dispersion

Rapport construit à partir des données d'une variable catégorielle. k correspond au nombre de catégories, ni à l'effectif de chaque catégorie et n à l'effectif total.

$$D = \frac{k(n^2 - \sum n_i^2)}{n^2(k-1)}$$

Intervalle interquartile (IQ)

Comprend 50% des observations les plus au centre d'une distribution de données. Il est utilisé avec des variables quantitatives. Il représente l'intervalle entre le deuxième et le troisième quartile : IQ = Q3-Q1 (un quartile divise les observations en quatre parties égales).

Q1 = (n+1) (0.25) Q1= position du premier quartile

Q3 = (n+1) (0.75) Q3=position du troisième quartile

Ecart-type

S (échantillon) et O (population). Parmi les mesures de dispersion, c'est la plus utilisée. L'écart-type n'est pas à confondre avec la moyenne! Il correspond à la distance moyenne des valeurs par rapport à la moyenne de la distribution.

$$s = \frac{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{n-1}$$

Comparaison de variables différentes

Pour comparer des variables qui ont des unités de mesure différentes (poids exprimé en kilos, taille en cm), il faut procéder à une normalisation, que l'on appelle valeur réduite (Z). Pour obtenir une valeur réduite, il faut connaître la moyenne et l'écart-type. Les valeurs réduites servent notamment à comparer le comportement de personnes, d'objets, d'événements sur deux ou plusieurs variables.

$$z = \frac{x - \bar{x}}{s}$$

Corrélation Il s'agit de l'intensité (de la force) de la relation entre une variable x et une variable y. Par exemple, l'âge d'une personne et son degré de dextérité. Cette corrélation peut s'observer

sur un graphique en points. La deuxième étape consiste à obtenir le coefficient de corrélation, qui s'étend de +1 (corrélation positive maximale) à -1 (corrélation négative maximale). Il existe deux manières de la calculer : le coefficient de corrélation de Pearson et le coefficient de corrélation de Spearman. Attention : une erreur souvent commise consiste à penser qu'il existe une relation de cause à effet directe entre la variable x et la variable y. Il faut s'assurer que la corrélation n'est pas liée au hasard (et donc, qu'elle est significative), que la corrélation n'est pas due à d'autres facteurs, et déterminer s'il existe une relation causale directe. Tout cela requiert non seulement une réflexion poussée mais aussi de disposer d'éléments suffisants pour la mener. Ce n'est pas parce qu'une population donnée consomme beaucoup d'alcool et que la même population présente un taux élevé du cancer du côlon qu'il y a forcément un lien de cause à effet!

3.2 Rechercher

La recherche de données peut se passer de plusieurs manières : via des requêtes sur des moteurs de recherche ou sur des bases de données de type open data... mais elle peut aussi relever d'un travail d'enquête où les données seront récoltées auprès de sources de manière plus "classique".

3.2.1 Formats et types de données

Une donnée se présente sous une forme structurée ou non-structurée. Une approche simpliste, permettant de faire la différence entre ces deux types de données, dispose que toutes les données pouvant être contenues dans un système de gestion de bases de données (SGBD) sont structurées (indexées, prêtes à l'emploi). Une autre approche spécifie qu'une "donnée structurée est tout sauf textuelle", mais celle-ci ne tient pas la route dans la mesure où un langage naturel est structuré (orthographe, grammaire, ponctuation). Dans une autre approche, une donnée est dite non-structurée si et seulement si sa structure ne peut être expliquée de manière rationnelle (suivant cette logique, les images et vidéos sont donc de bons exemples de données non-structurées). Il existe trois systèmes de stockage et d'organisation de données. Il existe également des données semi-structurées, dont les fichiers informatiques n'ont pas fait l'objet d'une indexation (Bierbbat et Lutz 2015).

3.2.1.1 Bases de données tabulaires

Il s'agit de données tabulaires (organisées en tableau). Le logiciel le plus connu est Excel. Un format standard de données organisées de manière tabulaire est le .csv (Comma-Separated Values), qui permet de représenter des données tabulaires en une suite de caractères séparés par une virgule. Chaque ligne de texte correspond à une ligne du tableau.

Nom,Ville,Code
Dupont,Bruxelles,1000
Durant,Namur, 500
est l'équivalent de :

Nom	Ville	Code
Dupont	Bruxelles	1000
Durant	Namur	5000

3.2.1.2 Bases de données NoSQL (représentation graphique)

Les données sont organisées dans des méta-balises. Ces bases de données ne sont pas destinées à être interprétées par un être humain mais par des machines (machine readable).

On retrouve dans cette catégorie deux formats principaux :

XML (Extensible Markup Language)

Langage de balisage permettant de structurer les données. Exemple d'application : le flux RSS (Really Simple Syndication / Rich Site Summary) d'un site récupéré par un agrégateur ou un réseau social. Se présente comme un fichier HTML mais le balisage est très différent (ainsi que la finalité).

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>Exemple de flux RSS</title>
    <link>http://xml.toutestfacile.com</link>
    <description>Exemple de flux RSS pour XML Facile!</description>
    <language>fr-fr</language>
    <managingEditor>webmaster@toutestfacile.com</managingEditor>
    <pubDate>Mon, 02 Feb 2009 10:00:00 GMT</pubDate>
    <item>
      <title>Exemple du jour</title>
      <description>Exemple du jour pour XML Facile!</description>
    </item>
  </channel>
</rss>
```

JSON (JavaScript Object Notation)

Permet la représentation de données structurées, dans un fichier qui pourra être interprété par du langage JavaScript.

```
{
  "menu": {
    "id": "file",
    "value": "File",
    "popup": {
      "menuitem": [
        { "value": "New", "onclick": "CreateNewDoc()" },
        { "value": "Open", "onclick": "OpenDoc()" },
        { "value": "Close", "onclick": "CloseDoc()" }
      ]
    }
  }
}
```

Traduction en XML :

3.2.1.3 Bases de données relationnelles

Les bases de données relationnelles font référence à des bases de données organisées dans des tableaux à deux dimensions (tables) qui entretiennent des relations entre eux. On parle d'un modèle entités-associations. Sur le web, le logiciel de gestion de base de données la plus

```

<menu id="file" value="File">
  <popup>
    <menuitem value="New" onclick="CreateNewDoc()" />
    <menuitem value="Open" onclick="OpenDoc()" />
    <menuitem value="Close" onclick="CloseDoc()" />
  </popup>
</menu>

```

largement utilisé est PhpMyAdmin. Lorsque l'on interroge une base de données, cela s'appelle une requête. Une requête est introduite via un langage de programmation compatible avec la base de données, généralement du SQL (Structured Query Language) ou MySQL (web).

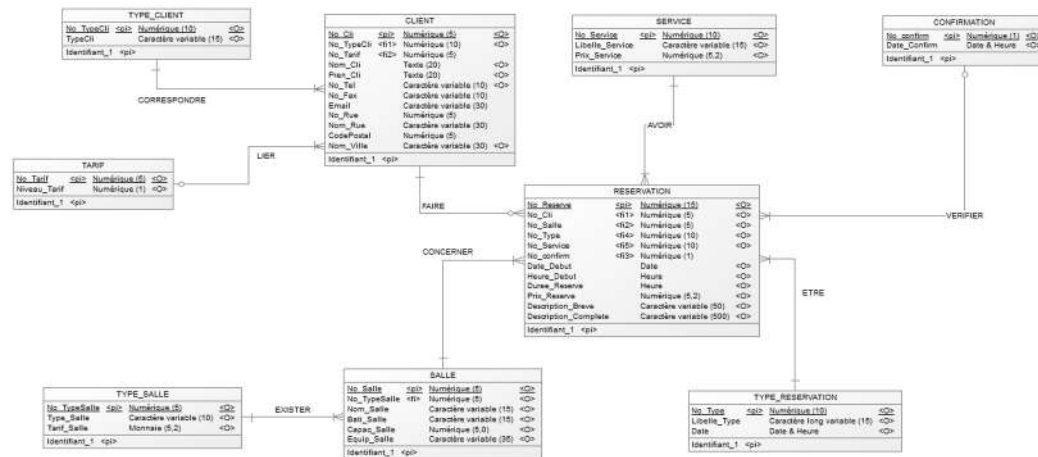


FIGURE 3.3 – Exemple du schéma logique d'une base de données relationnelle

Les informations diffusées sur le web sont généralement stockées dans des bases de données, un principe sur lequel reposent les systèmes de gestion de contenus (Content Management System, CMS) dont WordPress ou Drupal.

3.2.1.4 Métadonnées

Les métadonnées sont des informations sur les données. Sur un portail open data, leur objet est donc de les documenter. Elles ne sont pas toujours disponibles et, parfois, les métadonnées renseignent des structures de données qui ne sont pas conformes avec les jeux de données proposés en téléchargement. Les métadonnées peuvent être disponibles dans une variété de formats : xls, csv, xml... La norme Dublin Core est un format de métadonnées descriptives, beaucoup utilisée dans les bibliothèques³. Cette norme comporte quinze métadonnées descriptives relatives au titre, à la description, à l'auteur, à l'éditeur ou encore à la date de création du document. Voir <http://www.dublincore.org/>

3.2.1.5 Conversion de fichiers

On peut à peu près convertir tous les formats de fichier en ligne, du WAV au MP3, du PDF au DOCX ou au XLS, du MySQL au XLS, du JSON au CSV... La plupart de ces convertisseurs sont

³ Voir sur le site de la BNF : http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_dublin_core.html

	id	date	pm10	pm10be	pm25	pm25be	black	blackbe	ozone	ozonebe	azote
<input type="checkbox"/>	8	2017-02-02 13:24:04	19	19	17	15	0.9	1	NULL	NULL	NULL
<input type="checkbox"/>	7	2017-02-01 12:42:55	17	21	16	17	0.8	0.9	NULL	NULL	NULL
<input type="checkbox"/>	3	2017-01-26 19:31:29	18	20	13	14	0.9	1.4	NULL	NULL	NULL
<input type="checkbox"/>	4	2017-01-29 13:05:18	15	17	13	14	0.6	0.7	NULL	NULL	NULL
<input type="checkbox"/>	5	2017-01-30 13:39:32	9	8	8	5	0.5	0.7	NULL	NULL	NULL
<input type="checkbox"/>	6	2017-01-31 16:05:04	21	23	18	18	0.6	0.9	NULL	NULL	NULL
<input type="checkbox"/>	9	2017-02-03 18:46:36	16	13	9	7	0.7	0.9	NULL	NULL	NULL
<input type="checkbox"/>	11	2017-02-18 16:34:01	30	31	25	24	2.1	2.4	NULL	NULL	NULL

FIGURE 3.4 – Exemple d'une base de données PhpMyAdmin

gratuits et la requête standard pour les trouver est : "convert FORMAT to FORMAT".

3.2.1.6 Types de données

On distingue les données qualitatives continues (qui peuvent prendre n'importe quelle valeur dans un ensemble de valeurs) des données qualitatives discrètes (qui ne peuvent prendre qu'un nombre limité de valeurs). On distingue également les données quantitatives (ou catégorielles), dont les modalités ne peuvent être ordonnées (le sexe, la couleur des yeux); des données quantitatives ordinales, dont les modalités sont ordonnées selon un ordre logique (voir Bierbbat et Lutz 2015). Lorsque l'on manipule des données, on se retrouve confrontés face à différents types de données : des chaînes de caractères (string), des données numériques, des dates, des booléens (vrai ou faux),... Lorsque l'on est amené à traiter un ensemble de données, il est important de bien définir le type de données pour chaque variable utilisée.

3.2.2 Recherche booléenne

Rechercher une information sur internet – que ce soit via un moteur de recherche ou via une base de données en ligne – est un processus qui suppose que l'on ait préalablement défini les "bons" mots clés, ceux qui donneront accès à l'information recherchée. Introduire des mots clés les uns à la suite des autres est la méthode la plus simple (et la plus couramment utilisée) mais elle présente le risque de générer un important nombre de résultats qui ne s'avèreront pas forcément pertinents (en recherche documentaire, on appelle cela du bruit).

Une des manières d'éliminer ce bruit et de faire le tri en amont est de recourir aux fonctions de recherche avancées proposées par le moteur de recherche ou la base de données. Généralement, ils proposent de désigner des mots obligatoires, d'en exclure d'autres ou encore de déterminer une temporalité (par exemple, du 01/01/2012 au 30/12/2012). Une autre manière est d'utiliser les opérateurs booléens, issus de la théorie des ensembles développée au 19e siècle

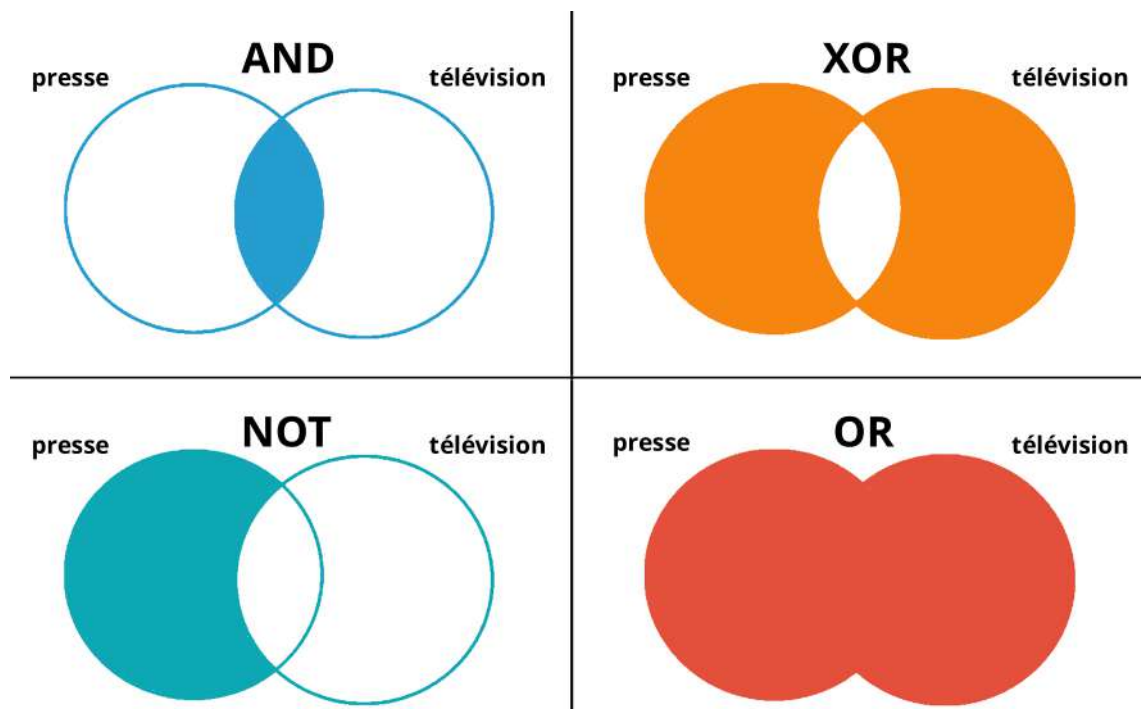
par le mathématicien britannique Georges Boole.

Formuler des équations de recherche

La combinaison d'opérateurs booléens permet de trier les résultats d'une recherche en les affinant en fonction de l'information recherchée.

Opérateur	Signification	Fonction	Exemple
AND	et (tous les mots)	Inclusion	presse AND télévision
NOT	sauf (à l'exception de)	Exclusion	presse NOT magazine
OR	ou (l'un ou l'autre mot)	Union	presse OR quotidiens
XOR	ou (mais pas les résultats comportant les deux mots)	ou exclusif	presse XOR radio
*	troncature, renverra dans les résultats tous les mots commençant, par exemple, par télé : télévision, télévisuelle, télévisuel, télégénique, téléphone, télégraphe...	joker	presse AND télé*
#	masque, renverra dans les résultats les variantes orthographiques, par exemple en cas de pluriel : quotidien et quotidiens	remplacement d'un seul caractère	quotidien#
" "	expression (groupe de mots)	expression	"presse en ligne"

Les résultats obtenus par ces combinaisons peuvent être visualisés dans des diagrammes.



L'intérêt de la recherche booléenne est de pouvoir effectuer des équations complexes qui permettront de préciser votre recherche. Les parenthèses sont utilisées pour procéder à une recherche par sous-ensembles.

Exemples

(presse AND magazine) NOT (audiovisuel AND quotidien*)

(presse OR magazine) AND (quotidien* XOR journaliste*)

La formulation suivante est également possible :

((presse OR magazine) AND (quotidien* XOR journaliste*) NOT emploi)

Astuce : dessinez vos ensembles avant de vous lancer dans votre recherche pour mieux visualiser le champ de celle-ci.

A propos des opérateurs de Google

Google propose ses propres opérateurs pour affiner ses recherches.

location : Belgique - pour une recherche géographique

source : lesoir - pour une recherche sur la source

allintext :presse quotidien - pour une recherche incluant tous les termes de la requête

presse filetype pdf - pour une recherche limitée aux documents pdf

link :lesoir.be - pour une recherche portant sur le lien

Liste des opérateurs Google : <https://support.google.com/websearch/answer/2466433?hl=fr>

Aller plus loin

"Méthodologie pour rechercher sur internet" sur le portail français Eduscol : <https://eduscol.education.fr/numerique/dossier/competences/rechercher/methodologie>

"Recherches d'infos" : <http://universite.online.fr/supports/recherche/pdf.htm>

3.2.3 Bases de données en ligne

Cette section présente une liste non-exhaustive de base de données en ligne proposant des jeux de données susceptibles d'être utilisés dans un contexte journalistique. Les bases de données nationales ou régionales ne figurent pas dans cette liste mais il est important de retenir que les gouvernements ainsi que de nombreuses institutions publiques proposent leur propre portail open data. La qualité des données y sera à géométrie variable : dans de nombreux cas, les données récoltées devront faire l'objet de nettoyage. Par ailleurs, ce n'est pas parce qu'un jeu de données est en ligne qu'il est nécessairement mis à jour.

1. Bases de données internationales

Plusieurs organisations internationales proposent des bases de données traitant de sujets divers, et dont la qualité est généralement acceptable.

Europe

Eurostat

Eurostat est une direction générale de la Commission européenne chargée de l'information statistique à l'échelle communautaire. Son siège se trouve au Luxembourg : <http://ec.europa.eu/eurostat/fr/data/database>

Portail open data de la Commission européenne

<https://data.europa.eu/euodp/fr/data/>

Monde

Banque mondiale

Accès aux statistiques de données de développement dans le monde : <https://donnees.banquemondiale.org/>

Explorer la base de données : <http://databank.banquemondiale.org/data/databases.aspx>

Institutions

Unicef : <https://data.unicef.org/>

Base de données de l'organisation mondiale de la santé : <http://www.who.int/gho/database/fr/>

API

Une API (application programming interface, interface de programmation) est un ensemble de définitions et de protocoles qui facilite la création et l'intégration de logiciels d'applications. De nombreux jeux de données sont disponibles via ce type d'interface (Twitter, Airbnb, Amazon...), et ils sont généralement disponibles au format JSON.

API Directory (recense plus de 24.000 API disponibles en ligne) : <https://www.programmableweb.com/category/all/apis>

De nombreuses organisations ainsi que des développeurs partagent également leurs données sur Github (les formats CSV et JSON y sont les plus courants), un service d'hébergement et de gestion de logiciels. Par exemple, toutes les données de OurWorldInData s'y trouvent en téléchargement : <https://github.com/owid/covid-19-data/tree/master/public/data>

2. Bases de données nationales

Les autorités nationales mettent également de nombreux jeux de données disponibles en licence ouverte. Statbel (institut de statistiques belge) : <https://statbel.fgov.be/fr>

INSEE (institut de statistiques français) : <https://www.insee.fr/fr/accueil>

Sciensano (données santé belges) : <https://epistat.wiv-isp.be/covid/>

Portail open data Belgique : <https://data.gov.be/fr/>

Portail open data France : <https://www.data.gouv.fr/fr/>

En ce qui concerne les portails de données open data, ceux-ci sont répertoriés par Opendatasoft (qui commercialise des solutions pour le partage de données) à cette adresse : <https://www.opendatasoft.com/fr/blog/2015/11/02/comment-avons-nous-liste-plus-de-2500-portails-de-donnees-ouvertes-pour-la-communaute-open-data>

3.3 Récolter

Lorsque l'on recherche des données sur un sujet, celles-ci ne sont pas forcément disponibles dans un format ouvert et réutilisable. La plupart du temps, elles seront accessibles depuis une page web ou via un fichier au format PDF. Cette section présente différentes techniques d'extraction de données, en ce compris le "web scraping" qui consiste en plusieurs techniques d'extraction de contenus qui vont généralement s'appuyer sur le code HTML (en ce compris les identifiants et classes de balises) de manière à cibler le contenu que l'on souhaite récupérer. Les techniques de récolte présentées sont toutes accessibles et gratuites. Elles ne nécessitent pas la connaissance d'un langage de programmation pour scraper des données (comme R ou Python, les plus utilisés pour l'extraction).

3.3.1 Code HTML, CSS et code source

Le mode de fonctionnement d'internet s'appuie sur une relation client-serveur : le mode de communication, à travers le réseau, comprend un client qui envoie des requêtes (client). Plusieurs langages vont contribuer à façonner une page web et à permettre les échanges entre le client et le serveur.



FIGURE 3.5 – Relation client-serveur

Une page web est une ressource du World Wide Web (www) consultée par des visiteurs à l'aide d'un navigateur web (Internet Explorer, Firefox/Mozilla, Safari, Google Chrome...) Elle comporte essentiellement du texte, des liens hypertextes mais aussi des images, des sons et/ou de la vidéo. Son format est généralement du HTML (Hyper Text Markup Language).

La représentation visuelle d'une page web à l'écran est toujours l'interprétation d'un code par un navigateur. Tous les navigateurs n'interprètent pas le code de la même manière, c'est pourquoi il peut exister des différences dans l'affichage d'une page web selon que l'on surfe sur tel ou tel navigateur. On dit d'une page dont le code a été travaillé de manière à assurer le même rendu sur tous les types de navigateur qu'elle a été optimisée.

Le code HTML n'a cessé d'évoluer depuis sa création mais s'est aujourd'hui stabilisé :

- 1990/2 : création du langage HTML par Sir Tim Berners-Lee, physicien britannique, inventeur du World Wide Web (WWW), au CERN (Genève). Le HTML est une déclinaison du langage de balisage SGML (apparu au début des années 1980 mais dont la genèse remonte à la fin des années 1960). Sa particularité est de permettre les hyperliens, grâce

auquel on navigue dans un site ou sur internet ⁴.

- 1994 : HTML 2
- 1997 : implémentation du HTML dynamique
- 1999 : HTML 4.01
- Début 2000 : développement du XHTML, identique au HTML 4 mais propose des règles strictes de balisage.
- 2006 : début des travaux pour le développement du HTML 5
- 2009 : abandon du XHTML
- Aujourd'hui : le HTML5 est devenu un standard, il est notamment caractérisé par le fait de séparer la structure de la forme (style).

Les grands principes du code HTML :

- le code définit la structure logique d'une page;
- le code est lu ligne par ligne, de haut en bas;
- une page est composée de balises commençant par < et se terminant par > et la plupart des balises ont des attributs de mise en forme;
- une balise ouverte doit toujours, sauf exception, être refermée : <> </>;
- le code doit être simple, pertinent et compris par celui qui s'en sert (la seule manière fiable d'éviter les erreurs!);
- une page web doit être la plus légère possible (poids du fichier) pour s'afficher rapidement quels que soient la connexion internet et le navigateur utilisés;
- les commandes HTML ne sont pas sensibles à la casse (majuscule ou minuscule) bien que le code soit généralement écrit en lettres minuscules (le XHTML strict impose les lettres minuscules).

Le code HTML peut être compris comme un enchaînement de boîtes qui se superposent et/ou s'imbriquent. Sa syntaxe repose sur des balises, qui définissent quel type de boîte le navigateur va interpréter. Le code HTML compte de nombreuses balises sémantiques dont l'objet est de donner du sens au codage du document : <header>, <footer>, <article>, <section>, ... Les balises se rapportant aux informations destinées au navigateur ainsi qu'aux moteurs de recherche et réseaux sociaux contiennent des informations invisibles pour l'utilisateur. Il s'agit notamment des métadonnées, qui comportent différentes informations à propos de la page (titre, description, catégorie, mots-clés, image...).

Ressources pour l'apprentissage du code HTML

- Codecademy : <https://www.codecademy.com/fr/tracks/web>
- Open Classrooms : <https://openclassrooms.com/courses/apprenez-a-creer-votre-site-web-avec-html5-et-css3>
- W3Schools : <https://www.w3schools.com/html/>
- HTML starter kit, Boilerplate : <https://html5boilerplate.com/>

⁴ Lire : "Histoire d'internet" <http://histoire-internet.vincaria.net/>

Structure de base d'un document HTML

<!DOCTYPE html> — dit au navigateur qu'il s'agit d'un document HTML

<html lang="fr"> – ouvre la balise HTML et précise la langue

<head> – boîte dont les informations sont destinées au navigateur

<meta /> — balise non fermante (ou autofermante) pour les méta informations (description, icône,...)

<link> — balise autofermante pour lier un fichier CSS ou JavaScript

</head>

<body> — boîte qui contient le code visualisé à l'écran

<header> — En-tête du site

<h1>Titre de niveau 1 (grand titre)</h1>

</header>

<article id="identifiantunique"> </article>

<section class="repetition"> </section>

<p class="repetition">Paragraphe</p>

 Retour à la ligne

<hr/> Ligne horizontale

<ul id="monidunique"> — (liste à puces, liste numérotée =)

Liste 1

Liste 1

<table>

<tr> Ligne

<td> Colonne 1 </td>

<td> Colonne 2 </td>

</tr>

</table>

<footer>

Informations de pied de page

</footer>

</html> —sauf exception, une balise ouverte = une balise fermée

Si le code HTML définit la structure d'une page, sa mise en forme va être assurée par la CSS (Cascading Style Sheet) ou feuille de style. L'apprentissage du HTML passe donc également par celui de la CSS. Ceux-ci vont interagir entre eux via un système d'identifiants ou de classes (id, class) attribués à des balises HTML.

Le **code source** est un texte qui représente les instructions de programme telles qu'elles ont été écrites par un programmeur. Le code source se matérialise souvent sous la forme d'un ensemble de fichiers textes. Le code source est généralement écrit dans un langage de programmation permettant ainsi une meilleure compréhension par des humains. Une fois le code

source écrit, il permet de générer une représentation binaire d'une séquence d'instructions — code binaire — exécutables par un microprocesseur.

Afficher le code source d'une page web

Cliquer sur "Afficher / code source" dans son navigateur ou cliquer simultanément sur les touches "CTRL + u" (ou "POMME + u").

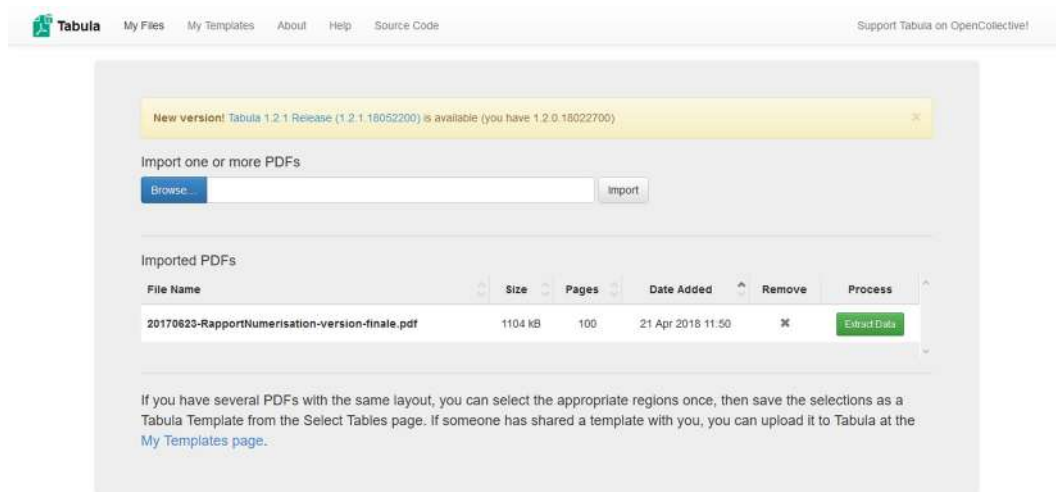
3.3.2 Scraper des données

Dans la plupart des cas, il est possible de récupérer un tableau figurant sur une page HTML par un simple copier-coller. Il existe également plusieurs outils de web scraping (présentés ci-dessous).

3.3.2.1 Extraire des données d'un document PDF

Le format PDF (Portable Document File) permet une extraction rapide (par copier-coller ou par conversion, au format Excel (plusieurs convertisseurs en ligne proposent des services de conversion, de manière gratuite ou payante).

Tabula permet d'extraire des données de documents PDF (logiciel à télécharger) : <http://tabula.technology/>



Le logiciel se lance dans la fenêtre d'un navigateur, proposant de télécharger le fichier dont on souhaite récupérer le tableau. Il conserve en mémoire les documents qu'il a déjà traités. Une fois le document PDF téléchargé, une nouvelle fenêtre va s'ouvrir permettant de visualiser toutes les pages du document. Il s'agit ensuite de sélectionner la page qui nous intéresse pour en extraire les données. Lorsque la page s'affiche dans la fenêtre principale, il faut sélectionner le tableau que l'on souhaite extraire avant de cliquer sur le bouton vert "Preview and extract data". Les données vont alors s'afficher sous la forme d'un tableau. Plusieurs formats d'exportation sont proposés (CSV, TSV, Json...) et il est également possible de copier le tableau.

Tabula My Files My Templates About Help Source Code Support Tabula on OpenCollective!

20160607_EtudeChomage_B_AL_FR_P... Templates Clear All Selections Autodetect Tables Preview & Export Extracted Data

9

10

2.1 Le chômage complet indemnisé (données administratives)

2.1.1 Evolution du chômage complet indemnisé

La méthode décrite dans la partie 1 nous permet d'obtenir des séries de données mensuelles portant sur le chômage complet indemnisé des demandeurs d'emploi indemnisés par le biais d'un régime d'assurance-chômage ou par le biais d'un régime d'aide sociale, ou sur le chômage complet indemnisé total. Grâce à ces chiffres mensuels, il est possible de calculer des moyennes trimestrielles et annuelles.

Tableau 4
Le chômage complet indemnisé :
assurance-chômage, aide sociale et total (chiffres absolus)

	Assurance-chômage				Aide sociale				Chômage complet indemnisé total			
	Belgique	Allemagne	France	Pays-Bas	Belgique	Allemagne	France	Pays-Bas	Belgique	Allemagne	France	Pays-Bas
2007	448 982	797 002	1 728 596	214 603	54 670	2 368 122	1 484 172	339 645	503 652	3 165 124	3 212 568	554 526
2008	423 358	726 514	1 688 715	173 745	56 643	2 097 167	1 422 344	310 953	480 000	2 823 581	3 111 059	484 698
2009	454 964	967 302	1 991 928	224 553	62 744	2 031 083	1 472 325	318 104	517 708	2 988 385	3 464 253	542 658
2010	460 709	880 021	2 042 901	266 778	65 629	1 982 917	1 545 523	348 044	526 338	2 862 938	3 588 423	614 822
2011	445 412	727 690	2 043 559	259 837	64 767	1 923 833	1 581 374	364 504	510 178	2 651 523	3 624 932	624 341
2012	441 406	747 555	2 135 737	303 187	65 160	1 836 333	1 639 261	371 694	506 575	2 583 888	3 774 997	674 881
2013	457 785	802 514	2 208 820	393 649	67 281	1 819 583	1 774 887	400 978	525 066	2 622 098	3 983 707	794 626
2014	458 642	774 718	2 250 704	438 045	69 579	1 799 583	1 869 619	427 671	526 221	2 574 301	4 120 322	865 716
2015	417 432	720 778	2 415 702	429 992	79 675	1 772 417	1 899 415	442 944	497 107	2 493 195	4 315 117	872 936

Tableau 5
Evolution du chômage complet indemnisé :
assurance-chômage, aide sociale et total ; base = 2007

	Assurance-chômage				Aide sociale				Chômage complet indemnisé total			
	Belgique	Allemagne	France	Pays-Bas	Belgique	Allemagne	France	Pays-Bas	Belgique	Allemagne	France	Pays-Bas
2007	1	1	1	1	1	1	1	1	1	1	1	1
2008	0.98	0.91	0.95	0.80	0.98	0.83	0.84	0.85	0.97	0.91	0.95	0.89
2009	1.03	1.16	1.12	1.26	1.03	0.83	0.84	0.85	1.03	1.16	1.12	1.26
2010	1.04	0.91	1.12	1.26	1.04	0.83	0.84	0.85	1.04	0.91	1.12	1.26
2011	0.99	0.91	1.12	1.26	0.99	0.83	0.84	0.85	0.99	0.91	1.12	1.26
2012	0.98	0.93	1.12	1.26	0.98	0.83	0.84	0.85	0.98	0.93	1.12	1.26
2013	1.02	0.91	1.12	1.26	1.02	0.83	0.84	0.85	1.02	0.91	1.12	1.26
2014	1.03	0.91	1.12	1.26	1.03	0.83	0.84	0.85	1.03	0.91	1.12	1.26
2015	0.93	0.91	1.12	1.26	0.93	0.83	0.84	0.85	0.93	0.91	1.12	1.26

Repeat this Selection

Tabula My Files My Templates About Help Source Code Support Tabula on OpenCollective!

20160607_EtudeChomage_B_AL_FR_P... Export Format: CSV Export Copy to Clipboard

Is the extracted data incorrect? You can revise your selected cells or try an alternate extraction method.

Revise Selected Cells Data has been extracted from the cells you selected in the previous step. You can revise your selection(s) to add or remove cells.

Choose Alternate Extraction Method The current preview uses the Stream extraction method. If the data is not mapped to the correct cells, try the Lattice method instead.

Stream Lattice

Preview of Extracted Tabular Data

	Assurance-chômage				Aide sociale				Chômage complet indemnisé total			
	Belgique	Allemagne	France	Pays-Bas	Belgique	Allemagne	France	Pays-Bas	Belgique	Allemagne	France	Pays-Bas
2007	448 982	797 002	1 728 596	214 603	54 670	2 368 122	1 484 172	339 645	503 652	3 165 124	3 212 568	554 526
2008	423 358	726 514	1 688 715	173 745	56 643	2 097 167	1 422 344	310 953	480 000	2 823 581	3 111 059	484 698
2009	454 964	967 302	1 991 928	224 553	62 744	2 031 083	1 472 325	318 104	517 708	2 988 385	3 464 253	542 658
2010	460 709	880 021	2 042 901	266 778	65 629	1 982 917	1 545 523	348 044	526 338	2 862 938	3 588 423	614 822
2011	445 412	727 690	2 043 559	259 837	64 767	1 923 833	1 581 374	364 504	510 178	2 651 523	3 624 932	624 341
2012	441 406	747 555	2 135 737	303 187	65 160	1 836 333	1 639 261	371 694	506 575	2 583 888	3 774 997	674 881
2013	457 785	802 514	2 208 820	393 649	67 281	1 819 583	1 774 887	400 978	525 066	2 622 098	3 983 707	794 626
2014	458 642	774 718	2 250 704	438 045	69 579	1 799 583	1 869 619	427 671	526 221	2 574 301	4 120 322	865 716
2015	417 432	720 778	2 415 702	429 992	79 675	1 772 417	1 899 415	442 944	497 107	2 493 195	4 315 117	872 936

3.3.2.2 Open Web Scraper pour Firefox

Il s'agit d'un "add-ons" (module additionnel) à ajouter au navigateur via le menu "Extensions et thèmes". Cet utilitaire est intégré dans les outils de développement de Firefox (pour l'afficher, touche F12)⁵.

1) Installer l'extension

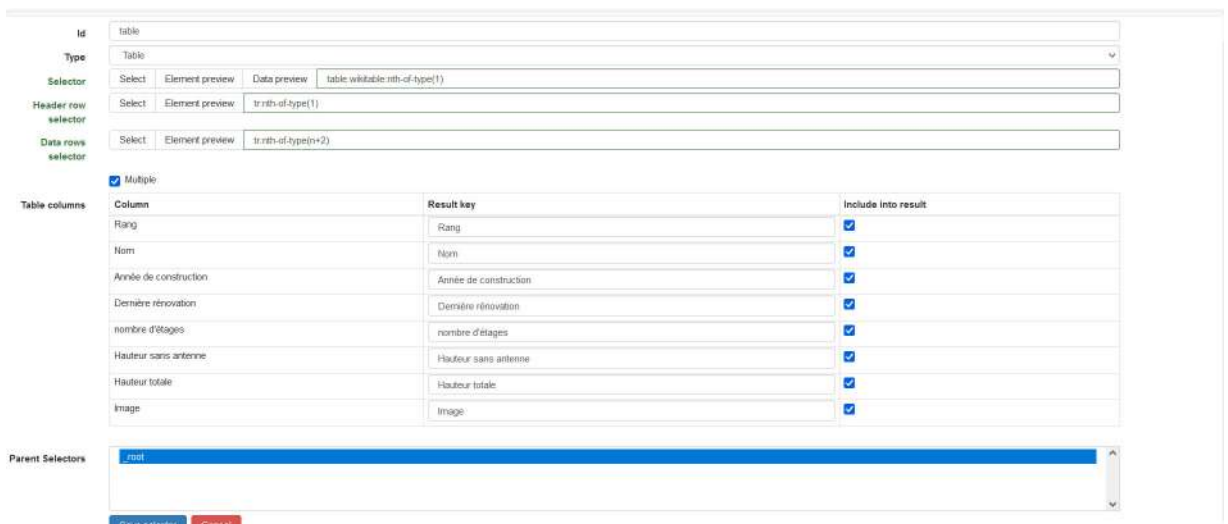
2) Se rendre sur la page à scraper et cliquer sur la touche F12 (ici, exemple avec les plus hauts gratte-ciels de Bruxelles sur Wikipedia : https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles

3) Cliquer sur l'icône "Web Scraper" (petit cercle bleu sur la droite), puis cliquer sur "Create new sitemap"

⁵ Voir la documentation complète : <https://webscraper.io/documentation/open-web-scraper>



4) Donner un nom à sa sitemap et renseigner l'URL, puis cliquer sur "Create Sitemap". Une nouvelle fenêtre apparaît, cliquer sur "Create new selector". Cliquer sur "Select" à chaque étape pour sélectionner les éléments de la page à scraper, ne pas oublier de valider la sélection et de sauvegarder le sélecteur. "Data preview" permet de prévisualiser le tableau de données.



5) Cliquer sur "Scraper" puis sur "Export data as csv"

Rang	Nom	Année de construction	Dernière rénovation	nombre d'étages	Hauteur sans antenne	Hauteur totale	Image
1	Tour du Midi	1961	1996	38	150 m[1]	168 m[1]	
2	Tour des Finances	1968	2008	36	145 m[1]	174 m[1]	
3	Tour UP-site	2010-2014	-	42	142 m	142 m	
4	Tour Rogier	2003	2006	38	135.9 m	145 m	
5	Iris Tower	2020	-	32	137	137 m	
6	Madou Plaza Tower	1965	2005	33	128.15 m	135 m	
7	Proximus Tower I	1994	1996	32	102 m	134 m	
8	Tour Astro	1972	2016	33	107 m	107 m	
9	North Galaxy A	2002-2004	aucune	29	107 m	107 m	
10	North Galaxy B	2002-2004	aucune	29	107 m	107 m	
11	World Trade Center 3	1983	2000	28	105 m	105 m	
12	Proximus Tower II	1994	1996	32	102 m	102 m	
13	Manhattan Center	1972	aucune	30	102 m	102 m	
14	World Trade Center 1 et 2	1971(1) 1974(2)	1973(1) 1974(2)	28	102 m	102 m	

W Liste des plus hauts gratte-ciel

https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles

133 %

Inspecteur Console Débugueur Réseau Éditeur de style Performances Mémoire Stockage Accessibilité Applications Web Scraper AdBlock Plus

Sitemap tours Create new sitemap

_root

Selectors
Selector graph
Edit metadata
Scrape
Browse
Export Sitemap
Export data as CSV

ID	type	Multiple	Parent selectors	Actions
table	SelectorTable	yes	_root	Element preview Data preview Edit Delete

Add new selector

F6	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	web-scraper-order,web-scraper-start-url,Rang,Nom,Année de construction,Dernière rénovation,nombre d'étages,Hauteur sans antenne,Hauteur totale,Image																	
2	1629536693-187	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	13	Manhattan Center	1972	aucune	30	102 m	102 m	102 m								
3	1629536693-177	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	3	Tour UP-site	2010-2014	-	42	142 m	142 m	142 m								
4	1629536693-184	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	10	North Galaxy 8	2002-2004	aucune	29	107 m	107 m	107 m								
5	1629536693-178	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	4	Tour Rogier	2003	2006	38	135.9 m	145 m	145 m								
6	1629536693-194	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	20	Tour Louise	1963-1965	-	23	90 m	90 m	90 m								
7	1629536693-197	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	23	Ellipse building	2004-2006	-	21	80 m	80 m	80 m								
8	1629536693-189	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	21	Tour ITT	1968-1971	aucune	25	102 m	102 m	102 m								
9	1629536693-195	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	1	Blue Tower	1976	1993 - 1997	25	88 m	88 m	88 m								
10	1629536693-191	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	17	Covent Garden	2003-2004	-	26	100 m	100 m	100 m								
11	1629536693-188	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	14	World Trade Center 1 et 2	1971(1) 1974(2)	1973(1) 1974(2)	28	102 m	102 m	102 m								
12	1629536693-175	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	11	Tour du Midi	1961	1996	38	150 m[1]	168 m[1]	168 m[1]								
13	1629536693-180	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	6	Madou Plaza Tower	1965	2005	33	120.15 m	135 m	135 m								
14	1629536693-176	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	2	Tour des Finances	1968	2008	36	145 m[1]	174 m[1]	174 m[1]								
15	1629536693-183	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	9	North Galaxy A	2002-2004	aucune	29	107 m	107 m	107 m								
16	1629536693-181	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	7	Proximus Tower I	1994	1996	32	102 m	134 m	134 m								
17	1629536693-193	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	19	Tour Zénith	2007-2009	-	23	95 m	95 m	95 m								
18	1629536693-182	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	8	Tour Astro	1972	2016	33	107 m	107 m	107 m								
19	1629536693-179	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	5	Iris Tower	2020	-	32	137	137 m	137 m								
20	1629536693-196	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	22	Tour TBR	Années 1970	-	22	84 m	84 m	84 m								
21	1629536693-186	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	12	Proximus Tower II	1994	1996	32	102 m	102 m	102 m								
22	1629536693-185	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	11	World Trade Center 3	1983	2000	28	105 m	105 m	105 m								
23	1629536693-190	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	16	Brusila	1967-1970	-	36	100 m	100 m	100 m								
24	1629536693-192	https://fr.wikipedia.org/wiki/Liste_des_plus_hauts_gratte-ciel_de_Bruxelles	18	Tour Möbius	2018-2021	-	18	98 m	98 m	98 m								

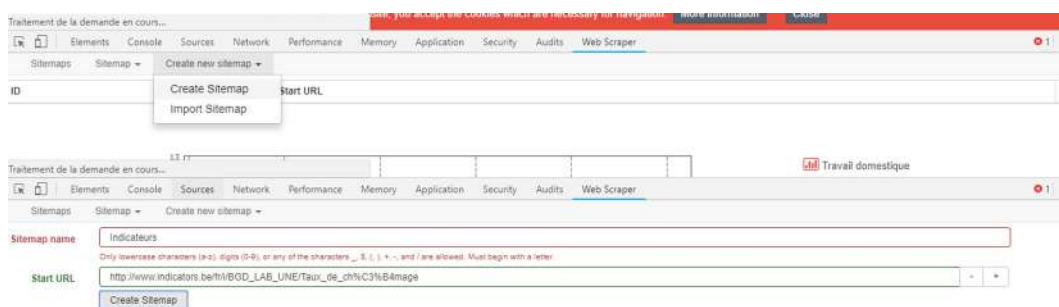
3.3.2.3 Chrome Web Scraper

Comme son nom l'indique, Chrome Web Scraper est une extension du navigateur Chrome (Google) dédié à l'extraction de données.

- 1) Installer l'extension : <https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehlkli>
- 2) Cliquer sur "Options" (menu hamburger) / plus d'outils / outils de développement ou presser la touche F12 du clavier.



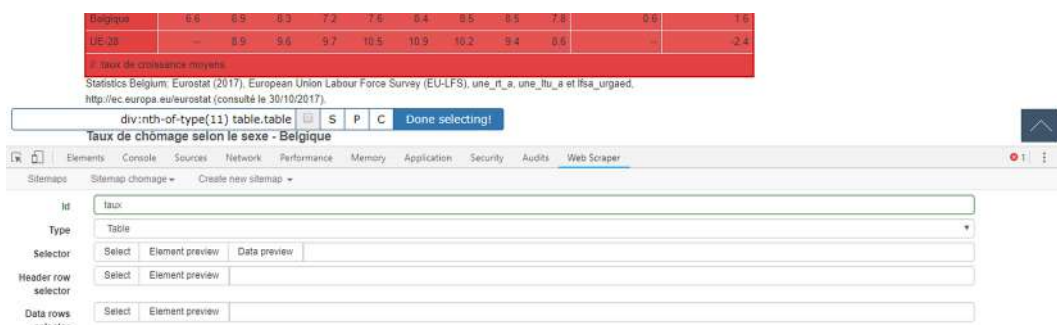
- 3) Dans l'onglet "Web Scraper", cliquer sur "Create Sitemap". Nommer la sitemap et la créer



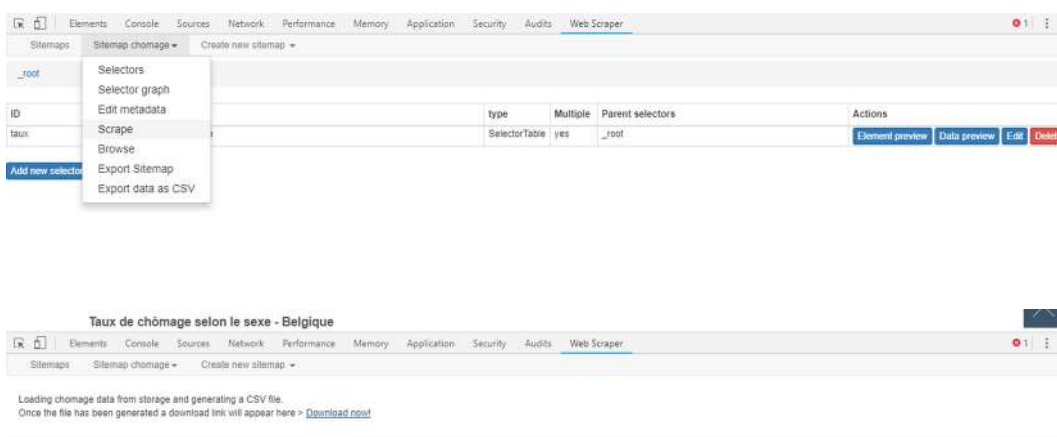
- 4) Cliquer sur "Add selector". Dans la fenêtre "Selector" donner un identifiant (id) au sélecteur et définir le type de données (table pour un tableau). Dans le champ Selector, cliquer sur "Select" puis cliquer sur le tableau que l'on souhaite récupérer et terminer l'action en cliquant sur "Done Selecting" (tous les champs seront remplis automatiquement). Le bouton "Data Preview" permet de visualiser les données qui seront extraites. Cocher la case "multiple" pour sélectionner les variables à scraper puis cliquer sur "Save Selector".

- 5) Afficher la liste déroulante sous l'onglet de la sitemap que l'on a créé puis cliquer sur "Scrape" puis sur "Start Scraping". Le tableau scrapé va s'afficher automatiquement.

- 6) Pour exporter les données en CSV, afficher la liste déroulante sous l'onglet de la sitemap que l'on a créé puis cliquer sur "Export data as CSV". Un lien de téléchargement "Download



Now" apparaît ensuite, il faut cliquer dessus.

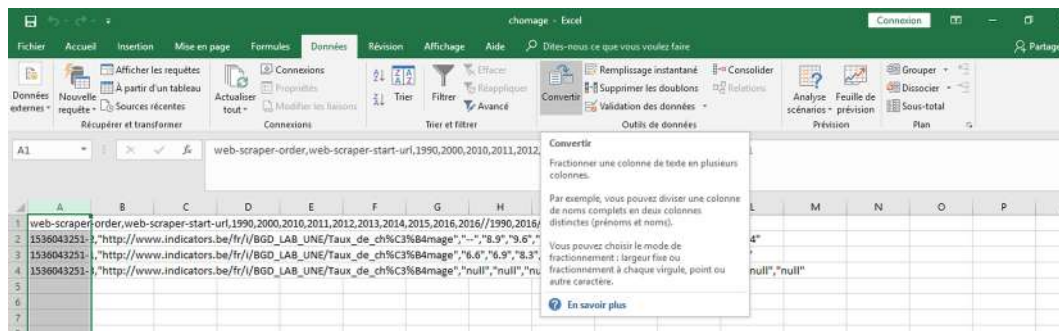


7a) Ouvrir le document sauvegardé dans Excel. Comme il s'agit d'un format CSV avec une virgule comme séparateur, le résultat ne se présentera pas sous la forme d'un tableau. Pour afficher le tableau, sélectionner la première colonne, celle qui comporte toutes les données, puis se rendre dans l'onglet "Données". Cliquer sur "Convertir". Dans l'assistant de conversion, choisir "Délimité". A l'étape 2, cocher uniquement la case "virgule", suivre la procédure jusqu'au bouton "Terminé". Le tableau qui va s'afficher peut nécessiter un peu de nettoyage pour un affichage correct.

OU 7b) Ouvrir le fichier CSV dans un éditeur de code (Notepad++, Bracket, Sublime...) et avec la fonction "Rechercher/Remplacer", remplacer toutes les virgules par un point virgule. Lorsque le fichier sera ouvert en Excel, il sera automatiquement proposé de choisir quel est le délimiteur du tableau. Il faut alors choisir "point-virgule". Le tableau qui va s'afficher peut nécessiter un peu de nettoyage pour s'afficher correctement.

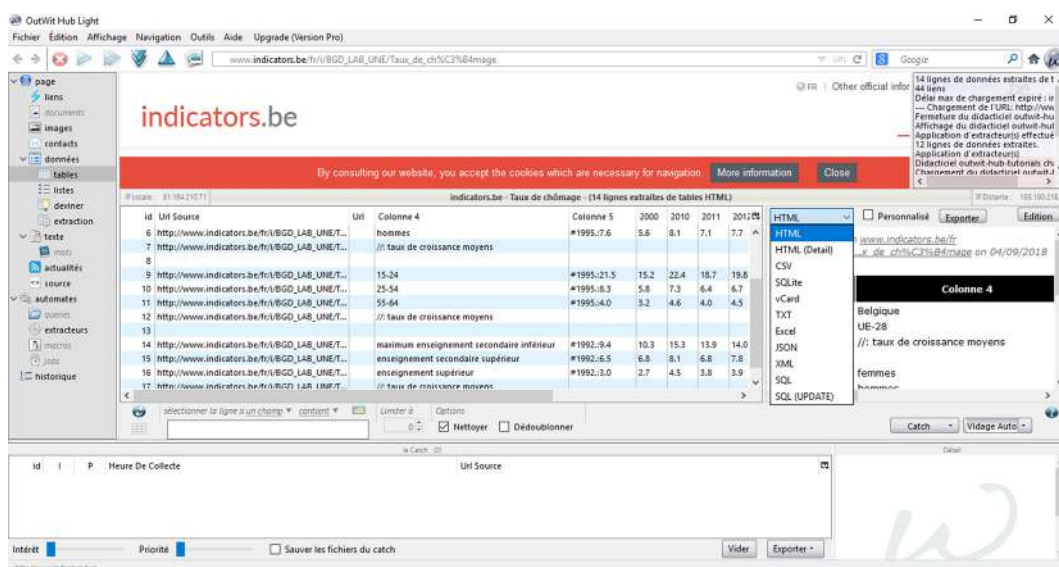
3.3.2.4 OutWit Hub

Outwit Hub est un logiciel de scraping disponible en version gratuite (limité à 100 lignes de données) et en version payante (89 euros la licence). La force de ce logiciel réside dans sa simplicité d'utilisation et son interface intégrée de prévisualisation : <https://www.outwit.com>



Plusieurs tutoriels sont également disponibles : <http://www.outwit.com/support/help/tutorials/>

- 1) Indiquer la page web à scraper dans le navigateur Outwit Hub
- 2) Dans la colonne de droite, sélectionner le type de données à extraire (par exemple, les tableaux : table). Toutes s'afficheront automatiquement dans la fenêtre principale. Un bouton latéral "Export" (sur la droite de l'écran) va permettre d'exporter les données dans le format souhaité (HTML, TXT, CSV, Excel, JSON, XML...).




```

1 #install.packages("rvest")
2 library(rvest)
3
4 jo <- read_html("https://fr.wikipedia.org/wiki/Sportifs_les_plus_m%C3%A9dailles_aux_jeux_olympiques")
5
6 table <- jo %>%
7   html_table(fill = TRUE)
8 View(table)
9
10 first_table <- table[[1]]
11 medailles <- as.data.frame(first_table)
12 str(medailles)
13
14 write.csv(first_table, 'medailles.csv')
15
16 #Documentation https://www.rdocumentation.org/packages/rvest/versions/0.3.6

```

3.3.2.5 Le package rvest de R

Rvest est un package pour le langage R (que l'on exécute via R Studio, voir infra) qui permet de récolter des données sur le web de manière rapide et efficace. En effet, quelques lignes de code suffisent pour récupérer un tableau de données dans un format qui permettra son analyse et/ou sa visualisation. La logique de récupération des données s'appuie sur des noeuds (ou nodes) en HTML. Il s'agit donc bien d'identifier les balises dans lesquelles se trouvent le contenu que l'on souhaite scraper. Voir également infra "Introduction à R", page 115.

Voir l'exemple complet : <https://github.com/laurence001/datajournalisme-R/blob/main/scraping>
 Documentation : <https://www.rdocumentation.org/packages/rvest/versions/0.3.6> - http://edutechwiki.unige.ch/fr/Web_scraping_avec_R

3.3.2.6 Outils en ligne

On citera encore trois outils en ligne dédiés au scraping de données qui ne nécessitent pas une connaissance du code HTML ou d'un langage de programmation. Tous trois fournissent une vidéo de présentation de leurs services.

- ParseHub - consiste en une application à télécharger (il existe également une version payante, la version gratuite est limitée à 200 pages et les données sont conservées pendant 14 jours) <https://www.parsehub.com/>
- Webrobots - ici aussi, une version gratuite est proposée. L'application s'installe via Chrome Web Store.
- Octoparse - également disponible en version payante et gratuite, il s'agit d'une application à télécharger : <https://www.octoparse.com/>
- Scraper API (payant) : <https://www.scraperaapi.com/documentation/>

Comment choisir entre tous ces outils ? C'est une question plutôt subjective qui va dépendre du type et du volume de données à traiter, ainsi que de son appétence pour l'une ou l'autre interface... et de son budget. Lorsque l'on maîtrise R, il s'agit toutefois de la solution la plus pratique pour ensuite travailler directement sur ses données.

3.4 Vérifier

Le fact-checking des données récupérées en ligne est indispensable dans toute démarche journalistique car on n'est jamais certain que tout ce que l'on trouve sur le web, un univers ouvert et non contrôlé, est authentique. A côté des techniques "classiques" de vérification, plusieurs outils peuvent vous y aider⁶.

Recherche d'images et recherche inversée

Google : <https://www.google.fr/imghp?hl=fr>

TinEye : <https://tineye.com/>

Voir les métadonnées d'une image <http://exif.regex.info/exif.cgi>

Réseaux sociaux

Comment détecter des faux comptes Twitter? Plusieurs indications : le nombre de personnes suivies, le nombre de tweets, l'activité générale du compte

Twitter Audit permet d'évaluer le nombre de faux comptes suivant un compte donné : <https://www.twitteraudit.com/>

Recherche avancée sur Twitter : <https://twitter.com/search-home>

Twitter Explorer (TwXplorer) : <https://twxplorer.knightlab.com/>

Visible Tweets (pour voir les mots-clés le plus fréquemment utilisé par un compte Twitter / utilisateur) : <http://visibletweets.com>

Statistiques d'un compte Twitter, Tweet Stats : <http://www.tweetstats.com/>

Recherche sur les comptes Facebook

1) Se connecter et passer son compte en langage anglais / US

2) Se rendre sur le site : <http://graph.tips/>

Web : auteur, historique, trafic

A qui appartient un site web? <http://whois.net/> - <http://who.is/>

Retrouver l'historique d'un site web via Way Back Machine : <https://web.archive.org/>

Informations sur le trafic d'un site web

Alexa : <http://www.alexa.com/> Similar Web : <https://www.similarweb.com/>

Informations sur une vidéo Date d'upload d'une vidéo YouTube : <https://citizenevidence.org/2014/04/01/how-to/>

Extraire métadonnées de vidéos (YouTube) : <https://citizenevidence.amnestyusa.org/>

InVid plugin : <http://www.invid-project.eu/tools-and-services/invid-verification-plugin/>

Vérifier une rumeur (hoax)

HoaxSlayer : <http://hoaxslayer.com/>

Hoaxbuster : <http://www.hoaxbuster.com/>

⁶ Voir aussi "Outils de fact-checking en ligne" : <http://www.ohmybox.info/online-fact-checking/>

Snopes : <http://www.snopes.com/>

TruthOrFiction : <https://www.truthorfiction.com/>

Navigation anonyme sécurisée

Tor (ordinateur de bureau) : <https://www.torproject.org/>

Tor (Android) : <https://openclassrooms.com/courses/protegez-l-ensemble-de-vos-communications-sur-internet-1/orbot-pour-utiliser-tor-sur-vos-appareils-android>

Anonymox (Firefox) : <https://addons.mozilla.org/fr/firefox/addon/anonymox/>

Anonymox (Chrome) : <https://addons.mozilla.org/fr/firefox/addon/anonymox/>

Sécurité (RSF) : <https://rsf.org/fr/actualites/comment-se-proteger-contre-la-surveillance-en-ligne>

Recherche bibliographique

Bibliothèque nationale de France (catalogue) : <http://catalogue.bnf.fr/index.do>

Bibliothèque royale de Belgique (catalogue) : <http://opac.kbr.be/index.php?lang=FR>

Persee.fr : <http://persee.fr/>

Europeana : <http://europeana.eu>

Analyse du discours : tag clouds

Tag Crowd : <http://tagcrowd.com/>

Word Clouds : <http://www.wordclouds.com/>

Wordle : <http://wordle.net>

Détection de plagiat

Plag Spotter : <http://www.plagspotter.com/>

Copyscape : <http://www.copyscape.com/>

Plagium : <http://www.plagium.com/index.cfm?language=fr>

Plag Tracker : <http://www.plagtracker.com/fr/>

Grille d'évaluation des contenus des documents web :

<http://www.emse.fr/spip/IMG/pdf/CIDE.pdf>

Base de connaissances générale :

<http://www.wolframalpha.com>

Alertes en temps réel :

<https://visualping.io/> (page web)

3.5 Nettoyer

Les données ne nous parviennent pas toujours de manière bien formatée, claire, lisible et bien organisée. C'est pourquoi avant de procéder à leur analyse, un nettoyage est souvent nécessaire. Les opérations de data cleaning (ou cleansing) sont aussi très importantes car elles permettent d'harmoniser les valeurs (dates, distances..) et d'éliminer doublons et cellules vides du tableau. Le nettoyage permet encore de procéder à des recherches/remplacement de textes, de corriger les erreurs orthographiques, de changer la casse du texte (majuscule, minuscule), de supprimer les espaces et caractères indésirables,.. Si ces opérations de nettoyage peuvent être effectuées dans n'importe quel de tableurs, certains outils traitent spécifiquement de cet aspect.

3.5.1 Introduction à Open Refine

OpenRefine est un data quality tool offrant de nombreuses possibilités pour nettoyer et corriger des tableaux de données. Il permet l'ouverture de nombreux formats, y compris des formats conçus pour être machine readable (XML, JSON) – le cas échéant, il transforme ces données automatiquement en tableau. Autre avantage du logiciel : il conserve l'historique des modifications (utile pour faire marche arrière) sans altérer les données de base soumises au nettoyage. Ce logiciel open source a été initialement développé par Google (Google Refine) avant que la société américaine abandonne le projet.

Télécharger Open Refine : <http://openrefine.org/download.html>

Manuel de référence : "Using OpenRefine", Ruben Verborgh et Max De Wilde (2013)
<http://openrefine.org/>

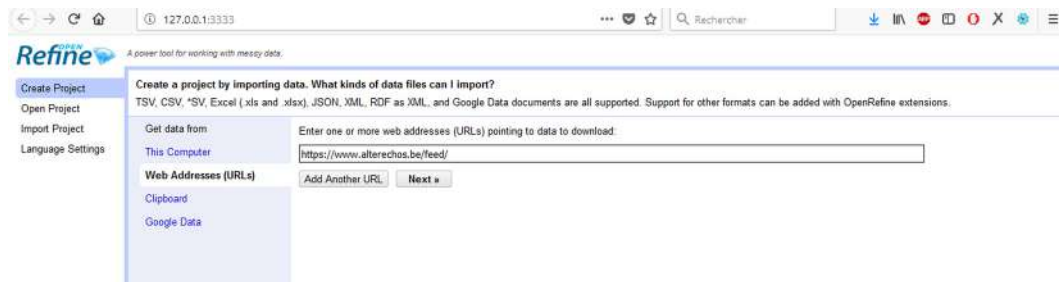
Wiki : <https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine/fr> - <https://github.com/OpenRefine/OpenRefine/wiki>

Utilisation du logiciel

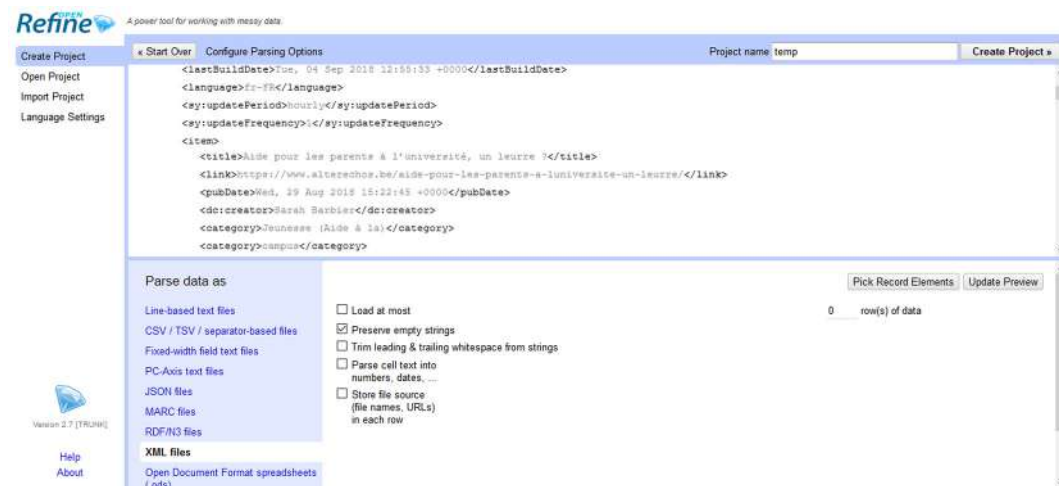
Lancer Open Refine. Le logiciel s'ouvre dans un navigateur web mais aucune connexion internet n'est requise. Le traitement de larges volumes de données peut s'avérer problématique en termes de consommation de ressources. Dans certains cas, il faudra allouer davantage de mémoire au logiciel. Procédure : dans le répertoire Open Refine, éditer le fichier openrefine.14j.ini (avec un éditeur de texte comme Notepad, Wordpad ou Notepad++). Trouver la ligne débutant par -Xmx qui montre la mémoire allouée par défaut à 1024M. Remplacer cette valeur : 2048M (par exemple).

A. Ouvrir un fichier XML (exemple : flux RSS)

- 1) Sélectionner "Web Adress" et insérer le lien du fichier XML puis cliquer sur le bouton "Next"
- 2) Sélectionner "XML File" dans les options et choisir le sélecteur XML qui contient le premier



groupe d'éléments (généralement "item" dans un flux RSS) puis cliquer sur le bouton "Create Project"

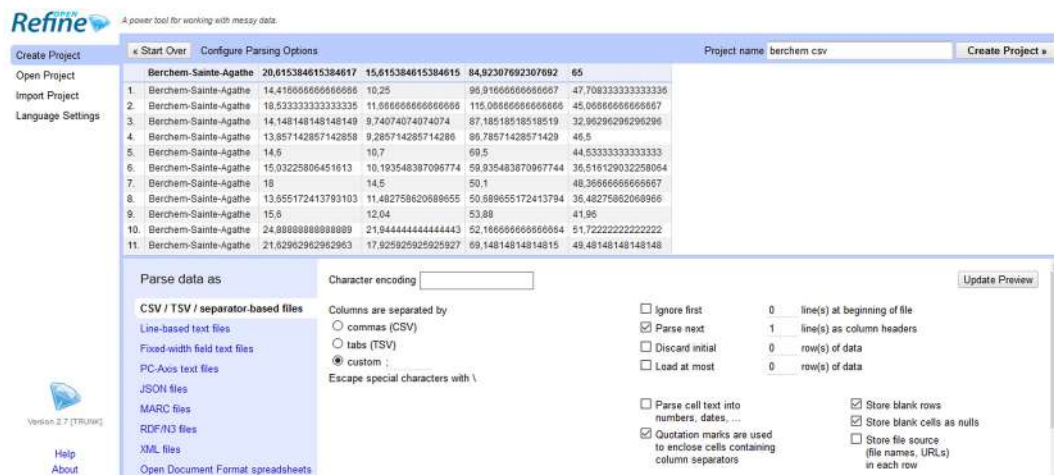


3) Le tableau des données s'affiche, cliquer sur "Create Project"



B. Ouvrir un fichier CSV

Sur l'écran d'accueil, sélectionner "This Computer". Via le bouton "Parcourir" sélectionner le fichier puis cliquer sur le bouton "Next". Cette opération concerne tous les types de fichiers : Open Refine accepte plusieurs formats de données (JSON, RDF, WML, CSV). Dans le cas d'un CSV, il faut choisir le délimiteur de colonnes : généralement, il s'agit d'une virgule mais si le fichier CSV provient de la conversion d'un fichier Excel dans le logiciel, le séparateur sera un point-virgule (par défaut, OR proposera la virgule). L'option "Columns are separated by" permet de choisir de quel délimiteur il s'agit. Dans le cas d'un point virgule, il faut choisir "custom" comme délimiteur et encoder le;



Problème d'encodage

Lors de la prévisualisation du tableau de données, il peut arriver que les caractères accentués s'affichent de manière étrange : il s'agit d'un problème d'encodage (le fichier a probablement été encodé en ISO 8858-1, qui ne prend pas en charge les caractères accentués. Il faut dès lors cliquer sur le champ "Character encoding" et choisir dans la liste UTF-8.



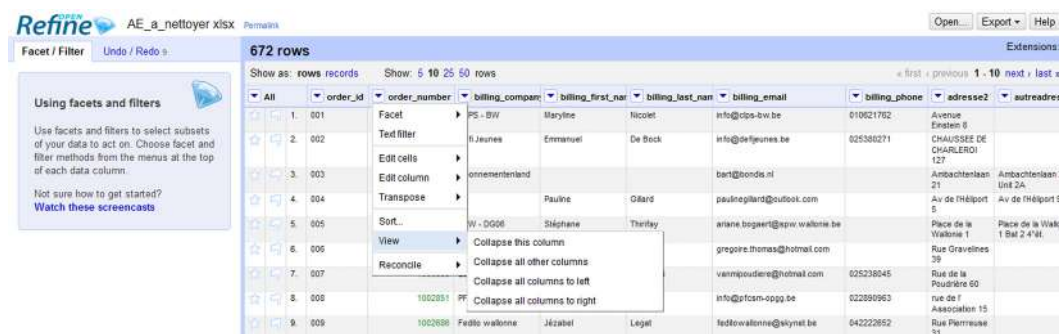
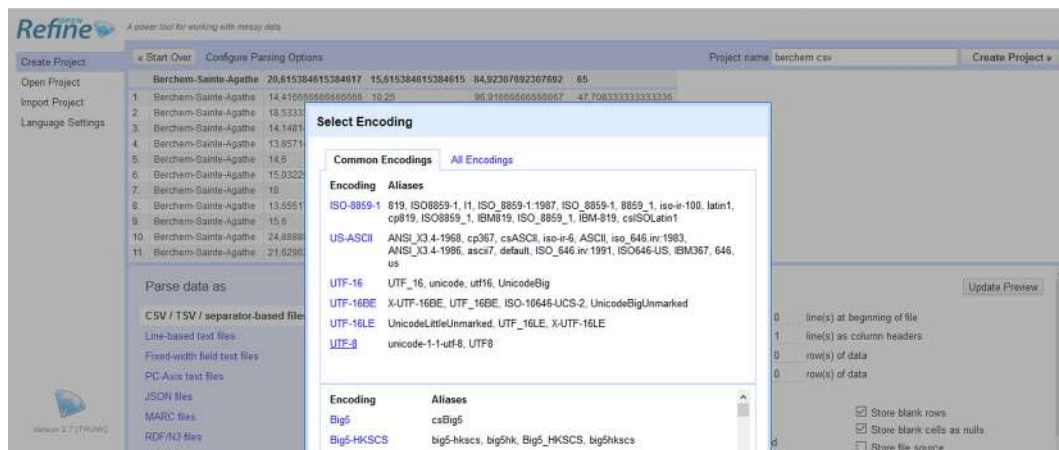
3.5.1.1 Organisation, tri et nettoyage des données

Le nombre de lignes affichées dans le tableau dépendra toujours du nombre d'enregistrements. L'affichage par défaut est celui des colonnes déployées (expanding). L'option "Edit column" permet des opérations sur la colonne, parfois nécessaires pour bien organiser ses données. C'est via cette commande que l'on supprime les colonnes vides, pour optimiser l'affichage de son tableau. Les options de contenu des cellules permettent de filtrer ou éditer les contenus.

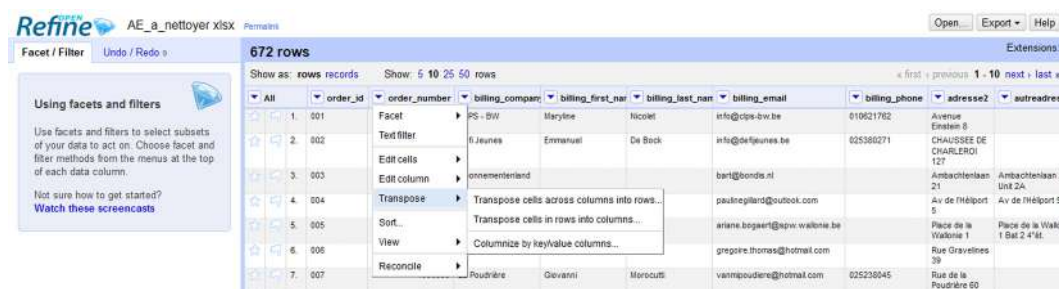
Ajuster l'affichage des colonnes

Lorsque le tableau comporte un nombre trop important de colonnes, l'option "View" permet d'en masquer certaines pour se concentrer sur celles qu'il s'agira de nettoyer : cliquer sur la petite flèche sous l'étiquette de la colonne pour afficher les options.

Editer une colonne



L'option "Edit column" permet des opérations sur la colonne, parfois nécessaires pour bien organiser ses données. C'est via cette commande que l'on supprime les colonnes vides, que l'on scinde des colonnes en deux, que l'on déplace, ajoute ou supprime une colonne.



Lorsque plusieurs types de valeurs sont contenus dans une colonne, il est possible de scinder cette colonne ("Split in several columns"). Il s'agit des fonctions avancées d'OR (voir manuel d'utilisation), qui comprennent encore d'autres possibilités liées à l'usage d'expressions régulières.

Editer le contenu d'une cellule

Pour modifier le contenu d'une cellule, placer le curseur de la souris sur la cellule et cliquer sur "Edit".

Type de données d'une cellule

Pour définir ou modifier le type de données, répéter la même opération que précédemment et

Facet / Filter Undo / Redo 672 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows » first < previous 1 - 10 next > last »

Using facets and filters

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email	billing_phone	address
1.	001	1000990	CLPS - BW	Maryline	Nicolet	info@dps-bw.be	010621762	Avenue Einstein 8

sélectionner le type souhaité (texte, nombre, booléen, date).

Refine AE_a_nettoyage.xlsx Permalink Open... Export... Help

Facet / Filter Undo / Redo 672 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows » first < previous 1 - 10 next > last »

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email	billing_phone	address
1.	001	1000990	CLPS - BW	Maryline	Nicolet	info@dps-bw.be	010621762	Avenue Einstein 8
2.	002	1000085	Déi Jeunes			info@délijeunes.be	025380271	CHAUSSEI CHARLERI 127
3.	003	101629819	Abonnementerland					Ambschten 21
4.	004	1003039						Av de l'Héli 9

Data type: text

CLPS - BW

Apply Apply to All Identical Cells Cancel

Editer les cellules d'une colonne

Cliquer sur la flèche sous l'étiquette de la colonne et sélectionner "Edit cells". Les modifications les plus courantes sont la conversion vers un type de données, le nettoyage du code HTML ("Unescape HTML entities") et la modification du type de caractère (majuscules, minuscules ou titre - dans ce cas, chaque premier caractère d'une cellule prendra une minuscule).

Refine AE_a_nettoyage.xlsx Permalink Open... Export... Help

Facet / Filter Undo / Redo 672 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows » first < previous 1 - 10 next > last »

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email	billing_phone	address
1.	001	1000990	Facet	ryline	Nicolet	info@dps-bw.be	010621762	Avenue Einstein 8
2.	002	1000085	Text filter	manuel	De Bock	info@délijeunes.be	025380271	CHAUSSEI CHARLERI 127
3.	003	101629819	Edit cells					Ambschten 21
4.	004	1003039	Edit column					Av de l'Héli 9
5.	005	1003077	Transpose					Place de la Wallonie 1
6.	006	1003075	Sort...					Rue Gravel 39
7.	007	1000588	View					Rue de la Poudrière 6
8.	008	1002851	Reconcile	Hassane	Moussa		045	Association
9.	009	1002686	Fedto wallonne	Jézabel	Legat		952	Rue Pierre 31
10.	010	1000762	Cpas Watermal Botsfort	Cécile	Philippot		820	Rue du Los 69

Transform...

Common transforms

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- Blank out cells

OR permet de faire du "facetting", opération de filtrage qui peut donner lieu à une meilleure compréhension des données ou à éliminer les doublons (via "Customize facet" qui permet aussi de repérer les cellules vides). L'option "Text filter" permet de filtrer par mot clé.

Trier les données

Facet / Filter
Undo / Redo

Refresh
Reset All
Remove All

20,615384615384617
change reset

No numeric value present

20,615384615384617
change reset

No numeric value present

Sort by 20,615384615384617

Sort cell values as

☒ text
☐ case-sensitive

☐ numbers

☐ dates

☐ booleans

Position blanks and errors

Valid values

Errors

Blanks

Drag and drop to re-order

☒ a - z
☐ z - a

OK
Cancel

11 rows

« first « previous
1 - 10
next » last »

153846	84,9230769230769	85
96,91666666666667	47,70833333333336	
115,06060606060666	45,06060606060667	
87,18518518518519	32,96296296296290	
96,78571428571429	46,5	
69,5	44,53333333333333	
59,935493870967744	36,516129032258064	
50,1	48,38060606060667	
50,689655172413794	36,46275862068966	
53,88	41,96	
52,166666666666684	51,72222222222222	

Tri basé sur la fréquence des données

L'option "Text Facet" permet d'opérer un tri sur les données. En cliquant sur le terme choisi dans la colonne de droite, seul le tableau de ces données sera affiché.

Facet / Filter Undo / Redo 9

Refresh Reset All Remove All

523 choices Sort by: name count Cluster

Abonnementenland 2 edit exclude

2 matching rows (672 total)

Show as: rows records Show: 5 10 25 50 rows

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email
3.	003	1016298619	Abonnementenland			bart@bondis.nl
539.	539	1016490206	Abonnementenland	Bart	Van Eijk	bart@bondis.nl

Rechercher-remplacer

OpenRefine permet de procéder à des remplacements de textes, dans le cadre d'un nettoyage de données. Pour ce faire, cliquer sur "Edit cells/Transform" Il faut ensuite composer une expression régulière (Regex). Dans Open Refine, on parle de Google Refine Expression Language (GREL).

Expression régulière pour le remplacement

value.replace("texte à remplacer", "texte de remplacement")

Refine

Facet / Filter Undo / Redo 9

Using facets and filters of your data to act on filter methods from the of each data column. Not sure how to get started? Watch these screencasts

Expression Language: General Refine Expression Language (GREL)

value.replace("eu", "euxai") No syntax error

Preview History Starred Help

row	value	value.replace("eu", "euxai")
1.	CLPS - BW	CLPS - BW
2.	Défi Jeunes	Défi Jeuxaines
3.	Abonnementenland	Abonnementenland
4.	null	Error: replace expects 3 strings, or 1 string, 1 regex, and 1 string
5.	SPW - DG06	SPW - DG06
6.	null	Error: replace expects 3 strings, or 1 string, 1 regex, and 1 string

On error: ☒ keep original ☐ set to blank ☐ store error ☐ Re-transform up to 10 times until no change

OK Cancel

Text transform on 17 cells in column billing_company: grel:value.replace("eu", "euxai") Undo

672 rows

Stars et flags

La première colonne permet de "flaguer" ou "étoiler" une ligne (peut être utile pour faire du repérage et/ou pour trier dans le cadre d'un profiling des données).

Refine AE_a_nettoyex.xlsx Permalink Flag row 4 Undo Open... Export Help

Facet / Filter Undo / Redo 14

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

672 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email
1.	001	1000960	CLPS - BW	Maryline	Nicolet	info@clps-bw.be
2.	002	1000085	Défi Jeunes	Emmanuel	De Bock	info@defijeunes.be
3.	003	1016298619	Abonnementerland			bart@bondis.nl
4.	004	1003039		Pauline	Gillard	paulinegillard@outlook

Exemple d'opération : supprimer des lignes "flaguées".

Cliquer ensuite sur : "Facet/By Flag"

Refine AE_a_nettoyex.xlsx Permalink Open... Export Help

Facet / Filter Undo / Redo 14

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

672 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email
Transform		1000960	CLPS - BW	Maryline	Nicolet	info@clps-bw.be
Facet		1000085	Défi Jeunes	Emmanuel	De Bock	info@defijeunes.be
Edit rows		1016298619	Abonnementerland			bart@bondis.nl
Edit columns		1003039		Pauline	Gillard	paulinegillard@outlook

Cliquer ensuite sur "Edit rows/Remove matching rows"

Refine AE_a_nettoyex.xlsx Permalink Open... Export Help

Facet / Filter Undo / Redo 15

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

672 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email
Transform		1000960	CLPS - BW	Maryline	Nicolet	info@clps-bw.be
Facet		1000085	Défi Jeunes	Emmanuel	De Bock	info@defijeunes.be
Edit rows						
Edit columns						
View						

Remove 672 rows Undo

0 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 0 next > last »

	order_id	order_number	billing_company	billing_first_name	billing_last_name	billing_email
--	----------	--------------	-----------------	--------------------	-------------------	---------------

Répéter/annuler une action

Il est toujours possible de répéter ou d'annuler une action : OR les conserve toutes en mémoire, y compris les actions effectuées lors de précédentes ouvertures de fichiers.



3.5.1.2 Exportation et historique des fichiers

OR permet d'exporter le tableau de données dans plusieurs formats : CSV (tab-separated value ou comma-separated value), HTML, Excel et ODF (Open Office).



Chaque fichier ayant été traité dans OR sera conservé. Lors de la prochaine ouverture du logiciel, le fichier traité apparaîtra dans la liste des projets créés. Pour travailler dans un projet déjà traité, il ne faut plus sélectionner "Create Project" mais "Open Project".



3.6 Analyser

3.6.1 Introduction à Excel

Un tableur est un logiciel informatique qui permet l'organisation de données en tableaux ainsi que leur manipulation. Le plus populaire d'entre eux est Excel, développé par Microsoft (format propriétaire, suppose l'achat d'une licence). Open Office, logiciel open source, dispose d'un tableur. Il existe plusieurs autres tableurs : <http://fr.wikipedia.org/wiki/Tableur>

La manipulation d'Excel, sur les plans statistique et mathématique, suppose l'application de formules (fonctions). Le logiciel dispose de nombreux raccourcis qui permettent notamment de : calculer la somme d'une colonne, convertir un nombre décimal en pourcentage, trier une liste par ordre alphabétique, trouver la valeur maximale et la valeur minimale... Il peut réaliser des opérations arithmétiques complexes. Il traite en outre différents formats de données (.xls, .csv...).

Excel permet également la génération de graphiques, ce qui est particulièrement utile pour analyser les données : lorsque l'on se retrouve face à un tableau de données comportant des centaines d'entrées, un graphique permet de visualiser les tendances, de faire émerger des concepts ou des informations rapidement. Excel propose différents types de graphiques : diagramme en barres ou circulaire, courbes... (voir aussi "Comment choisir la bonne visualisation?" p.70).

3.6.1.1 Structure d'un tableau

Les données sont organisées sous la forme de lignes et de colonnes. Les colonnes (A,B,C,...) listent les variables. La première ligne de la colonne est toujours utilisée pour les noms des variables. Les variables contiennent plusieurs lignes = plusieurs attributs.

Présentation incorrecte

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Grèce	France	Hongrie	Tchéquie	Suède	Suisse	Autriche	Norvège	Royaume Uni	Espagne	Italie	Pologne	
2	Journalistes	11	29	20	20	25	35	32	32	29	35	33	42	
3	tions (presse	16	28	28	29	36	39	42	42	42	43	43	46	
4														
5														
6														

Présentation correcte

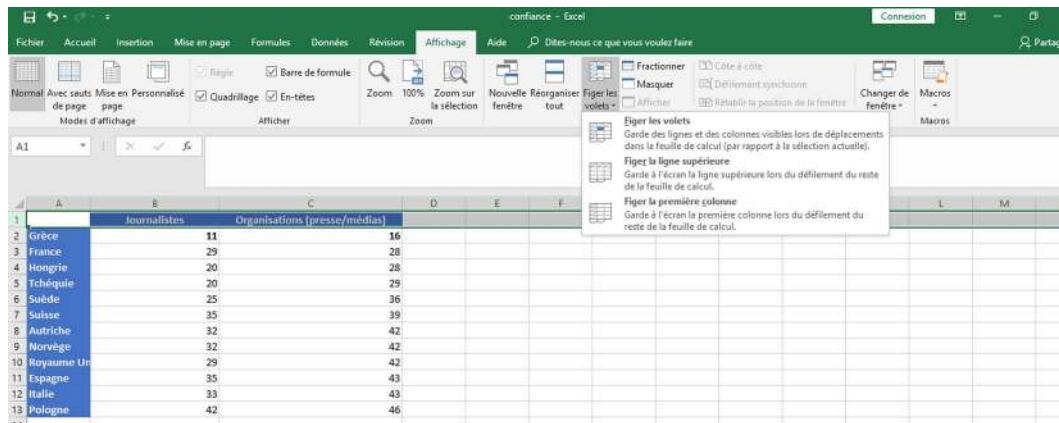
	A	B	C	D
1		Journalistes	Organisations (presse/médias)	
2	Grèce	11	16	
3	France	29	28	
4	Hongrie	20	28	
5	Tchéquie	20	29	
6	Suède	25	36	
7	Suisse	35	39	
8	Autriche	32	42	
9	Norvège	32	42	
10	Royaume Uni	29	42	
11	Espagne	35	43	
12	Italie	33	43	
13	Pologne	42	46	
14				

3.6.1.2 Manipulations de base

Affichage

Améliorer la présentation en figeant la première ligne de titre (possibilité également de figer une colonne) :

Affichage / figer les volets



Raccourci pour descendre à la dernière ligne du tableau :

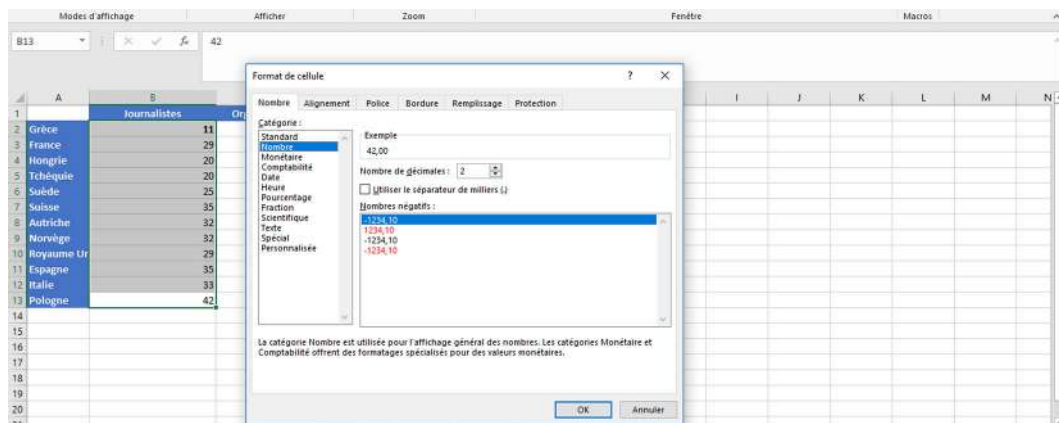
CTRL + flèche vers le bas

Remonter à la première ligne du tableau :

CTRL + flèche vers le haut

Formater une colonne en fonction de sa valeur (date, nombre, texte...) :

Sélectionner la colonne, clic droit de la souris et choisir "format de cellule" (pour un nombre, on peut choisir le nombre de décimales affichées; pour une date, on peut choisir la présentation).



Trier et filtrer des données :

Cliquer sur l'icône de tri A-Z (menu "Affichage"). Tri par ordre décroissant ou croissant : positionner le curseur dans la cellule de la colonne que l'on veut trier. Accueil : cliquer sur le filtre

A-Z (ordre croissant ou décroissant). Pour trier en fonction de deux variables : Données / Trier // Choisir une variable puis ajouter un niveau.

	A	B	C
1		Journalistes	Organisations (presse/médias)
2	Grèce	11	16
3	France	29	28
4	Hongrie	20	28
5	Tchéquie	20	29
6	Suède	25	36
7	Suisse	35	39
8	Autriche	32	42

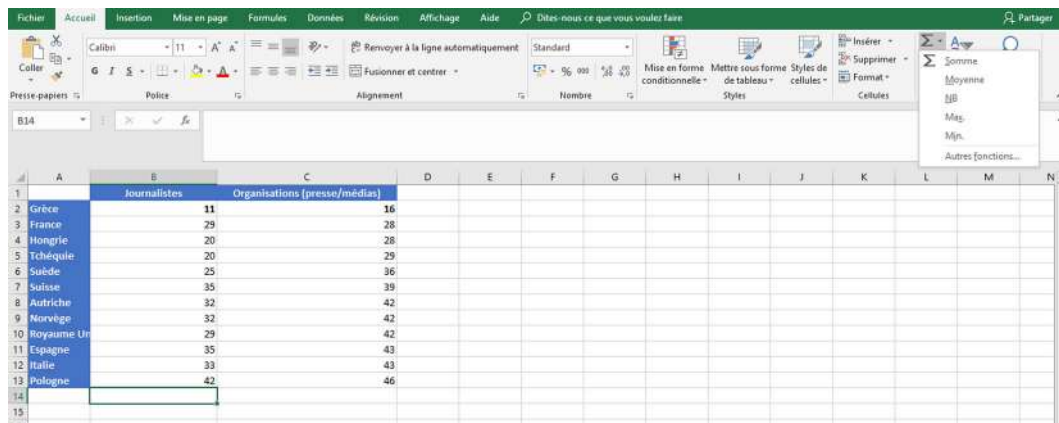
	A	B	C
1		Journalistes	Organisations (presse/médias)
2	Grèce	11	16
3	France	29	28
4	Hongrie	20	28
5	Tchéquie	20	29
6	Suède	25	36
7	Suisse	35	39
8	Autriche	32	42
9	Norvège	32	42
10	Royaume Uni	29	42
11	Espagne	35	43
12	Italie	33	43
13	Pologne	42	46

Un filtre peut être affiné en cliquant sur "Filtre textuel".

	A	B	C
1	Pays	Journalistes	Organisations (presse/médias)
7	France	29	28
14	France	29	28

Opérations

Calculer une moyenne, trouver le minimum et le maximum, calculer la valeur de l'écart-type : Sélectionner la cellule sous la colonne ou dans celle se trouvant à côté de la dernière ligne (selon les valeurs visées par l'opération) et cliquer sur la flèche à côté de l'icône "Somme" dans le menu "Affichage". Il s'agit de raccourcis pour des calculs pouvant être effectués dans la barre d'état d'Excel. Pour d'autres opérations statistiques, il faut cliquer sur "Autres fonctions". Pour calculer un pourcentage, il faut cliquer sur % (ne pas indiquer le symbole du pourcentage dans la colonne!).



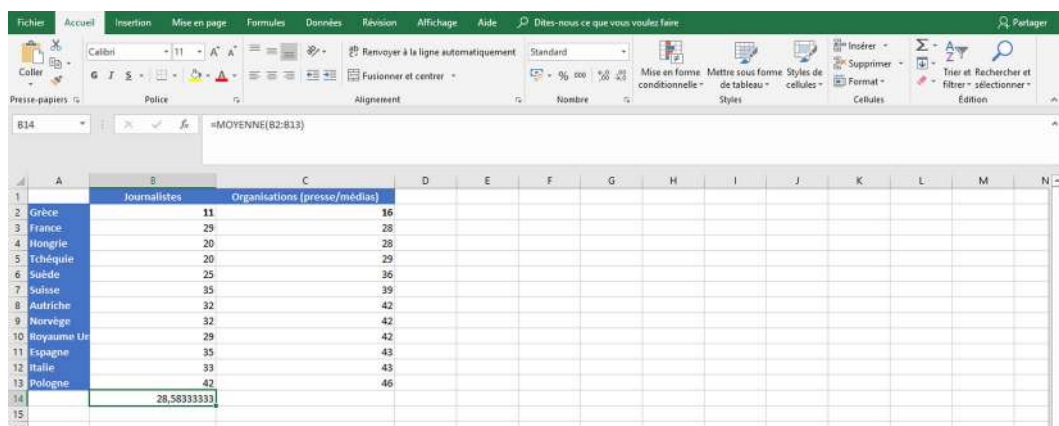
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Journalistes	Organisations (presse/médias)											
2	Grèce	11	16											
3	France	29	28											
4	Hongrie	20	28											
5	Tchéquie	20	29											
6	Suède	25	36											
7	Suisse	35	39											
8	Autriche	32	42											
9	Norvège	32	42											
10	Royaume-Uni	29	42											
11	Espagne	35	43											
12	Italie	33	43											
13	Pologne	42	46											
14														
15														

Les formules pour les calculs d'une moyenne, de la valeur médiane et de l'écart-type sont formulées comme suit :

=MOYENNE(A1 :A110)

=MEDIANE(A1 :A7)

=ECARTYPE(A2 :A11)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Journalistes	Organisations (presse/médias)											
2	Grèce	11	16											
3	France	29	28											
4	Hongrie	20	28											
5	Tchéquie	20	29											
6	Suède	25	36											
7	Suisse	35	39											
8	Autriche	32	42											
9	Norvège	32	42											
10	Royaume-Uni	29	42											
11	Espagne	35	43											
12	Italie	33	43											
13	Pologne	42	46											
14		=MOYENNE(B2:B13)												
15		28,58333333												

Les formules sont en français ou en anglais, selon la langue du logiciel :

=AVERAGE : moyenne

=MEDIAN : médiane

=MIN : minimum

=MAX : maximum

Une somme peut être obtenue via les formules :

= cellule1+cellule2 (=D2+E2) => pour deux cellules en particulier

= SUM(D2 :G2) => pour toutes les cellules entre D2 et G2 incluses.

Calculer un pourcentage :

= cellule1/cellule2*100 => application de la règle de trois

Calculer un taux, par 100.000 habitants par exemple : = nom-de-la-colonne-total / nom-de-la-colonne-population *100000

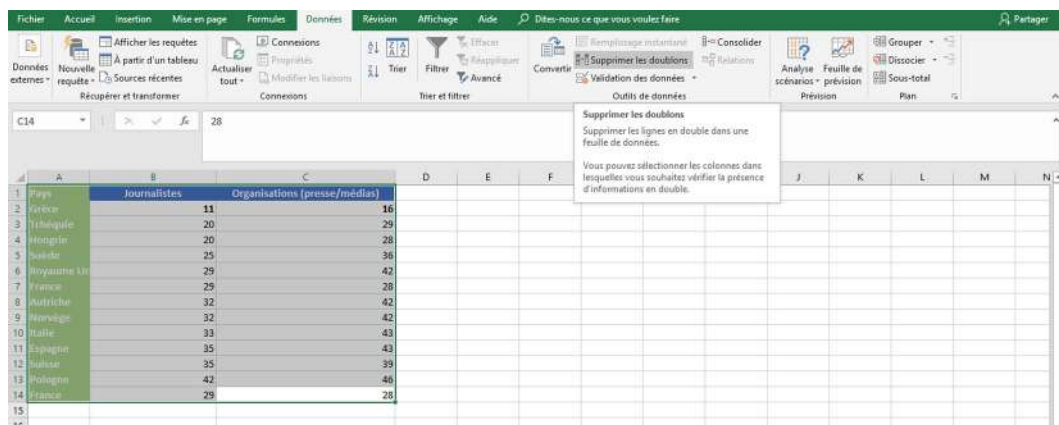
Calculer le nombre de jours séparant deux dates :

= date2-date1 (doivent être formatées en date! important! cf format de cellule au clic droit de la souris)

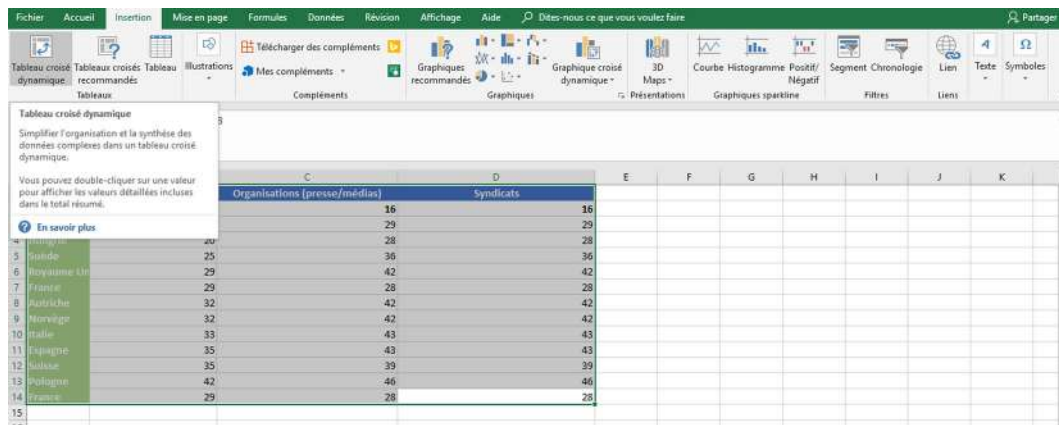
= (date2-date1)/365,25 nombre d'années

Astuce pour répéter une même opération pour l'ensemble des colonnes : étirer la colonne pour reproduire l'opération. On peut également double-cliquer dans le coin inférieur droit (croix) pour remplir instantanément toute la colonne. De cette manière, de nombreuses opérations arithmétiques peuvent être effectuées.

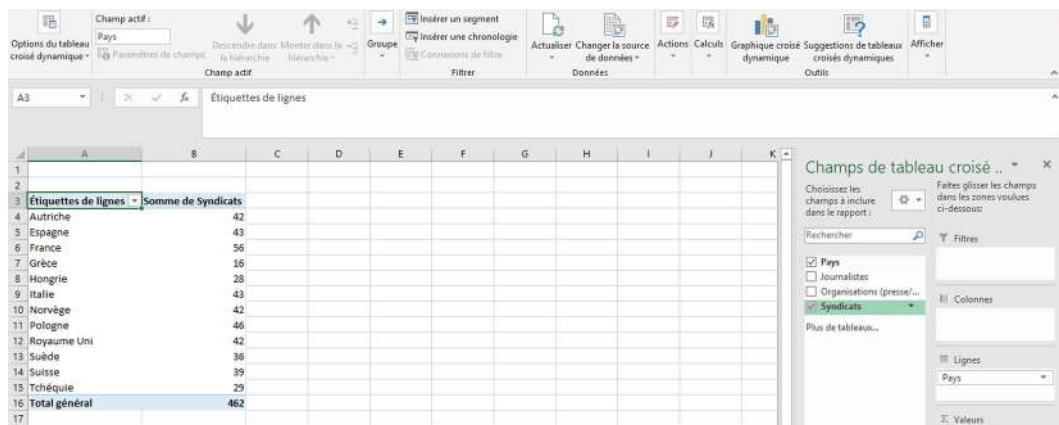
Supprimer des doublons : Données / Suppression de doublons



Créer un tableau croisé dynamique : Excel peut aider à l'analyse de jeux de données complexes grâce à la création de tableaux croisés dynamiques, qui « permet de synthétiser, analyser, explorer et présenter une synthèse des données d'une feuille de calcul ou d'une source de données externe. Il se sert de données bidimensionnelles pour créer un tableau à trois dimensions, à partir de conditions multiples possédant des points d'intersection ». Pour créer un tableau croisé dynamique, cliquer sur une cellule de la feuille de calcul puis sur "Insertion" / "Tableau croisé dynamique"



Excel va créer une nouvelle feuille de calcul dans laquelle il faut indiquer, dans la colonne de droite, les champs que l'on souhaite isoler.



Pour visualiser les données d'un tableau Excel, cliquer sur "Insertion" / "Insérer graphique". Excel vous proposera le graphique optimal pour votre visualisation mais libre à vous de choisir celle qui vous paraîtra la plus pertinente. A noter que pour visualiser des tendances, dans le cadre d'une analyse de données, on utilisera davantage le graphique en points ou à bulles (mais ce ne sera sans doute pas la datavisualisation la plus efficace pour l'utilisateur final, lecteur/internaute).

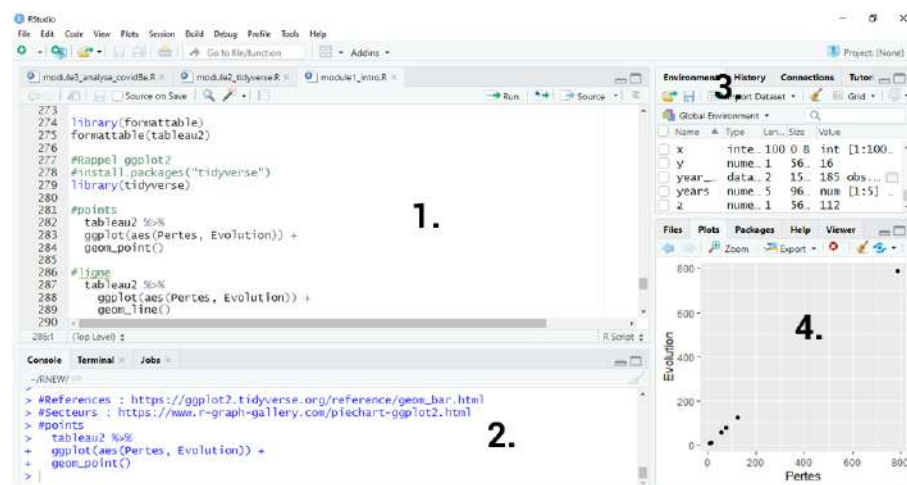
Aller plus loin

- Aide sur les procédures Excel :
<http://office.microsoft.com/fr-be/excel-help/>
- Explorer les fonctions Excel sur OpenClassRooms :
<http://fr.openclassrooms.com/informatique/cours/analysez-des-donnees-avec-excel/les-fonctions-d-excel>
- Tutoriel – mise en forme statistique :
<http://education.francetv.fr/videos/tutoriel-excel-mise-en-forme-de-statistiques-1-v106485>

3.6.2 Introduction à R

R est un langage de programmation spécifiquement dédié au traitement de statistiques. R Studio⁷ est un logiciel libre (environnement de programmation) permettant le traitement de données en R. Son interface se découpe en plusieurs parties, la fenêtre réservée à l'éditeur de code (sur la gauche de l'écran) étant la plus importante. Comme n'importe quel autre langage de programmation, R permet de réaliser des opérations arithmétiques complexes et d'une manière rapide et comporte des variables (le principe d'une variable est de stocker des informations) et des fonctions (le principe d'une fonction est de pouvoir répéter des opérations sans devoir les reprogrammer à chaque fois, il s'agit donc d'un petit programme informatique dans la mesure où une fonction est composée d'une séquence d'instructions). L'objet de cette section est de présenter les principales fonctionnalités de R et de proposer des outils pour celles et ceux souhaitant aller plus loin dans cet apprentissage. Les opérations avec R sont réalisées avec le logiciel R Studio⁸. Plusieurs médias utilisent R que ce soit pour l'analyse ou la visualisation de données, comme par exemple à la BBC où une équipe de datajournalistes utilise R, depuis un certains temps, pour l'analyse de données complexes et reproductibles ainsi que pour construire des prototypes⁹.

3.6.2.1 Interface de R Studio



L'interface comporte quatre parties :

- (1) Exécution du code (script).
- (2) Console (permet l'exécution du code).
- (3) Accès à l'environnement (liste des objets enregistrés) et à l'historique.
- (4) Accès à la bibliothèque, à l'écran d'aide et aux visualisations.

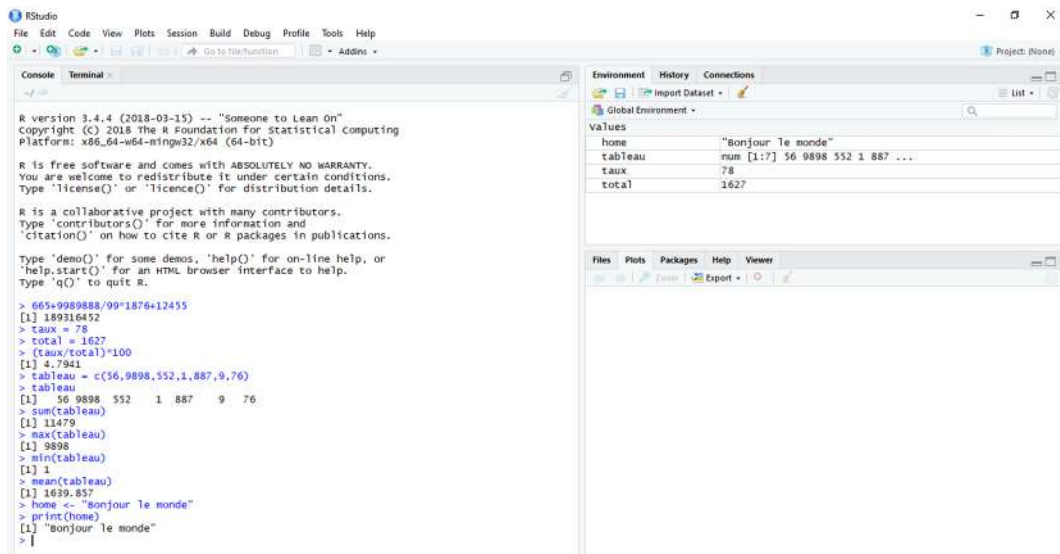
⁷ <https://www.rstudio.com/>

⁸ <https://rstudio.com/products/rstudio/download/>

⁹ Voir <https://medium.com/bbc-visual-and-data-journalism/how-the-bbc-visual-and-data-journalism-team-works-with-graphics-in-r-ed0b35693535>

3.6.2.2 Un super calculateur

Si des opérations arithmétiques peuvent être introduites manuellement, celles-ci peuvent également être effectuées en utilisant des variables. Une variable a pour fonction de stocker une valeur. Il existe différents types de valeurs : des valeurs numériques, des chaînes de caractères (string), des dates (ex. : `date <- as.Date("2012-08-08")`). Les valeurs peuvent être stockées de manière unique (une variable = une valeur) ou elles peuvent être contenues dans un tableau de valeurs soit à entrée simple (que l'on appelle un vecteur) soit à deux dimensions (que l'on appelle un tableau).



The screenshot shows the RStudio interface. The console pane on the left displays the following R code and its output:

```
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 665+9989888/99*1876+12455
[1] 189316452
> taux = 78
> total = 1627
> (taux/total)*100
[1] 4.7941
> tableau = c(56,9898,552,1,887,9,76)
> tableau
[1] 56 9898 552 1 887 9 76
> sum(tableau)
[1] 11479
> max(tableau)
[1] 9898
> min(tableau)
[1] 1
> mean(tableau)
[1] 1639.857
> home <- "Bonjour le monde"
> print(home)
[1] "Bonjour le monde"
> |
```

The environment pane on the right shows the following variables:

Variable	Value
home	"Bonjour le monde"
tableau	num [1:7] 56 9898 552 1 887 ...
taux	78
total	1627

Les manipulations sur les données numériques et sur les chaînes de caractères sont similaires à celles effectuées dans n'importe quel autre type de langage de programmation. Par exemple, pour compter le nombre de caractères dans une chaîne : `nchar("combien de signes dans ma chaîne")` OU `machaine <- "combien de signes dans ma chaîne" nchar("machaine")`. R est dit "case sensitive", c'est à dire qu'il est sensible à la casse. Dans le cas présent, il reconnaîtra bien la variable "machaine" mais pas "MaChaine", par exemple. Pour supprimer une variable : `rm(variable)` ex : `rm(machaine)`

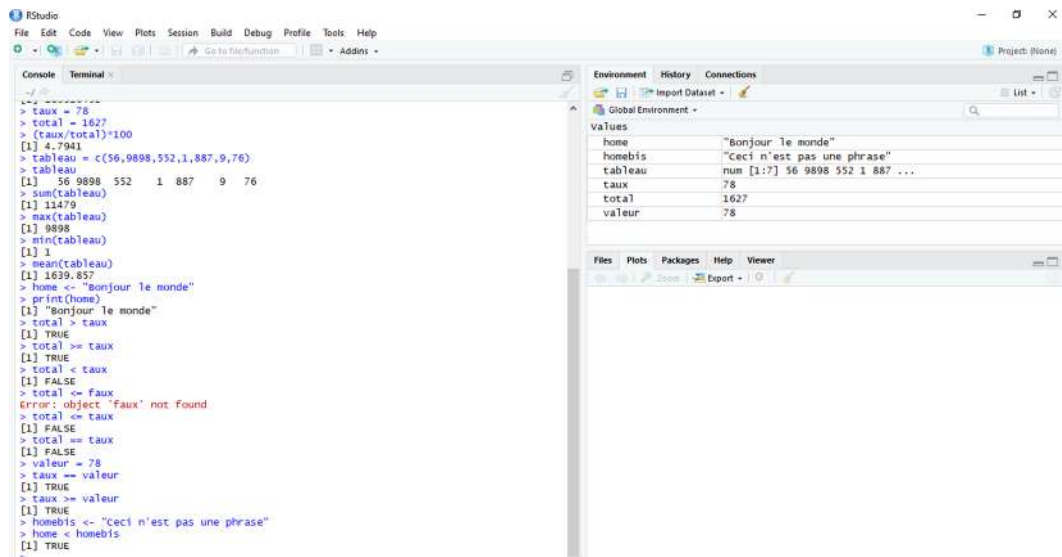
Quelques opérations utiles sur des chaînes de caractères :

- `str_lenght` : longueur de la chaîne
- `str_to_upper` : convertir une chaîne de caractères en capitales
- `str_to_lower` : convertir une chaîne de caractères en lettres minisucules
- `str_trim` : supprime les espaces blancs

On appelle un langage de programmation de haut niveau un langage qui sera proche du langage naturel et éloigné du langage binaire de l'ordinateur (constitué de 0 et de 1). A contrario, un langage de bas niveau sera proche de celui de l'ordinateur. Un langage de haut niveau devra être traduit dans un langage binaire pour être correctement interprété par l'ordinateur. Cette conversion est appelée la compilation.

3.6.2.3 Opérateurs logiques

Les opérateurs logiques permettent d'établir des comparaisons entre deux variables. Il s'agit également d'un concept qui se retrouve dans tous les langages de programmation. Le résultat d'une comparaison sera toujours "vraie" ou "fausse" ("true" ou "false"). On compare toujours des valeurs de même nature : un nombre avec un nombre, une date avec une date, une chaîne de caractères avec une autre chaîne de caractères.



Il est également possible de comparer les valeurs de deux tableaux, par exemple leur moyenne :

```
x = c("44,55,66")
y = c("77,88,2,2,44,99")
mean(x) > mean(y)
```

3.6.2.4 Classes de données

- Factor : `is.factor(data) / as.factor(data)`, facteur (peu de valeurs identiques, exemple : sexe, F ou M)
- Numeric : `is.numeric(data) / as.numeric(data)`, numérique, chiffres (le plus courant)
- Integer : `is.integer(data) / as.integer(data)`, nombres entiers relatifs
- Date : `is.Date(data) / as.Date(data)`, date
- Character : `is.character(data) / as.character(data)`, chaînes de caractères

`is(...)` : vérifie si la donnée est de telle classe

`as(...)` : attribue une nouvelle classe à la donnée

`class(data)` : vérifie la classe de la donnée

3.6.2.5 Vecteurs, tableaux, listes, matrices

Un vecteur peut être comparé à une seule ligne d'un tableau Excel. Pour encoder un vecteur :

```
x <- c(7,8,10,2,45)
```

```
rm(x) #supprimer la variable x (ici, un vecteur)
```

`sort(x)` #tri croissant du vecteur
`rank(x)` #donne la place du tri dans le vecteur (2=7, 3=8, 4=10, 1=2, 5=45)
`order(x)` #donne la place de la valeur dans le vecteur par rapport tri (4=2, 1=2, 2=7, 3= 8, 5=45)
`Ctrl+L` #vider la console
`Ctrl + Enter` #exécuter le code
`View(x)` #visualisation du vecteur
`mean(x)` #moyenne du vecteur x
`median(x)` #médiane du vecteur x
`min(x)` #valeur minimale du vecteur x
`max(x)` #valeur maximale du vecteur x
`summary(x)` #sommaire (maximum, minimum, médiane, moyenne, distribution)
`length(x)` #longueur du vecteur x
`is.null(vecteur)` #y a-t-il une valeur NULL dans le vecteur?
`sum(x)` #additionne toutes les valeurs du vecteur x
`ls()` #affiche toutes les variables enregistrées dans l'environnement
 Si les valeurs du vecteur sont un facteur, `levels(x)` affiche les valeurs

Plusieurs manipulations sont possibles telles que remplacer une valeur par une autre, incrémenter de x unités chaque valeur du vecteur, faire le total ou bien la somme du vecteur. Un tableau est composé d'au moins deux vecteurs. Un tableau consiste en une collection de vecteurs. Les exemples suivants montrent, pas à pas, les manière de procéder en passant d'opérations sur un vecteur simple à des opérations sur une matrice. La fonction `data.frame` permet de créer un tableau mais aussi de donner un titre aux colonnes. La fonction `rownames` va permettre de donner un intitulé (étiquette) aux lignes.

```

> vecteur1 <- c(55,88,99,99)
> vecteur1
[1] 55 88 99 99
> vecteur2 <- c(66,88,45,12)
> vecteur2
[1] 66 88 45 12
> tableau
      vecteur1 vecteur2
1          55         66
2          88         88
3          99         45
4          99         12
> tableau <- data.frame(vecteur1=vecteur1, "Taxe"=vecteur2)
> names(tableau)
[1] "vecteur1" "Taxe"
> tableau
  vecteur1 Taxe
1       55  66
2       88  88
3       99  45
4       99  12
> dim(tableau)
[1] 4 2
> # 4 lignes et 2 colonnes (dimension)
Error: unexpected '=' in '-'
> rownames(tableau) <- c("Janvier", "Février", "Mars", "Avril")
> tableau
      Quantité Taxe
Janvier      55  66
Février      88  88
Mars         99  45
Avril        99  12
  
```

Un tableau peut comporter énormément de lignes. Pour en afficher les premières lignes : `head(variable)`
 - si l'on souhaite afficher les trois premières lignes : `head(variable, n=3)`. Il s'agit de la même logique pour afficher les dernières lignes du tableau sauf que cette fois, on procédera comme suit : `tail(variable)`. Si l'on veut connaître la position précise, par exemple la valeur de la pre-

mière ligne et de la deuxième colonne : `variable[1,2]` soit `variable[ligne, colonne]`. Pour l'affichage d'une colonne en particulier, par exemple la deuxième : `variable[,2]`. Il est également possible de sélectionner une colonne en fonction de son nom (étiquette) : `variable[,c("titredelacolonne")]`. Parmi les autres formes d'affichage, on notera encore celui d'une colonne sous la forme d'une ligne : `variable[,numerodelacolonne]` soit `tableau[,2]` ; et pour afficher la colonne sous la forme d'un tableau.

Les valeurs peuvent également être stockées dans des listes, qui peuvent contenir n'importe quel type de données. La création d'une liste est simple : `list(1,2,3)` ; elle peut également être créée en tant que vecteur `list(c(1,2,3))` ; ou encore être créée à partir d'un vecteur : `variableduvecteur <- list(x)`. Cette fonctionnalité sera toutefois moins utilisée dans le cadre d'analyses de données dans un contexte journalistique, les opérations sur les valeurs et tableaux permettant l'essentiel des opérations nécessaires. Enfin, une dernière manière de stocker un ensemble de données consiste dans la création de matrices. Comme les vecteurs, les matrices peuvent être additionnées entre elles, multipliées ou vérifiées avec des opérateurs booléens (vrai ou faux). Pour créer une matrice (pour information) : `variable <- matrix(1 :10, nrow=2)` cela correspond à `1 :10` = valeurs de 1 à 10 et `nrow=2` = sur deux rangées.

Fichier attaché : **intro.r**

<https://github.com/laurence001/datajournalisme-R/blob/main/intro.r>

3.6.2.6 Scraper des données d'une page web

R comporte des dizaines de packages, qui consistent en des librairies gratuites de code. L'un d'entre eux a pour objet de scraper des données d'une page web : "rvest" (voir aussi page 95). Pour l'utiliser il faut l'installer si ce n'est déjà fait : `install.packages("rvest")`
Il faut ensuite appeler la librairie : `library(rvest)`

L'extraction des données se base sur les balises HTML, et sur les identifiants (id) et classes (class) CSS. Il est donc important d'examiner préalablement la structure de la page que l'on souhaite scraper, et en particulier les éléments à extraire.

Les fonctionnalités principales de rvest sont :

- `html()`
- `html_nodes()`
- `html_text()`, `html_attrs()`, `html_tag()`
- `html_table`

Pour récupérer tout le code d'une page :

```
page <- read_html("http://www.irceline.be/fr/qualite-de-lair/mesures/dioxyde-dazote/last-14-days")
```

La limite tient dans le fait que certaines pages web soient surchargées de code HTML et que des opérations de nettoyages soient plus simples à effectuer via d'autres solutions comme OR ou un tableur.

Plusieurs exemples de web scraping sont disponibles sur R Blog (anglais) : <https://www.r-bloggers.com/rvest-easy-web-scraping-with-r/> et ici en français http://edutechwiki.unige.ch/fr/Web_scraping_avec_R

A titre d'information, il est possible de scraper en PHP suivant cette même logique avec PHP Simple HTML Dom Parser : <http://simplehtmldom.sourceforge.net/>

Voir aussi : Extraire et nettoyer des données d'un fichier PDF (tuto en anglais) : <https://medium.com/@CharlesBordet/how-to-extract-and-clean-data-from-pdf-files-in-r-da11964e252e>

Fichier attaché : scraping.r

<https://github.com/laurence001/datajournalisme-R/blob/main/scraping>

3.6.2.7 Ouvrir un fichier CSV, copier-coller un tableau

Pour ouvrir un fichier CSV dans R, il faut d'abord s'assurer que le séparateur utilisé est une virgule et pas un point-virgule. Il est possible de préparer son fichier très rapidement en ouvrant son fichier CSV dans un éditeur de code (Notepad++, Sublime, Bracket...) en utilisant la fonctionnalité "rechercher / remplacer". Le package utilisé est readr : `library(readr)`

Pour ouvrir le fichier :

Il est également possible d'ouvrir un fichier directement depuis Github. Pour ce faire, se rendre sur le lien du fichier et cliquer sur "raw" en haut à droite du fichier. Une nouvelle fenêtre s'ouvre et c'est l'URL de cette fenêtre qu'il s'agit de copier : `mesdata <- read_csv("fichier.csv")`

Pour exporter un fichier CSV : `write_csv(mesdata, "https://raw.githubusercontent.com/curran/data/pages/worldFactbook/GDPPerCapita.csv")`

Il est possible qu'un fichier comporte des valeurs NULL. Dans ce cas, l'export sera : `write_csv(mesdata, "fichier.csv", na="")`

Plusieurs manipulations peuvent avoir lieu une fois le fichier importé, comme le renommage des colonnes (attributs des variables) : `colnames(mesdata) <- c("titre1", "titre2", "titre3")`

Dans la même logique, il est également possible d'ouvrir un fichier Excel dans R Studio. Toutefois, cette opération est beaucoup plus difficile et il est conseillé de convertir ce fichier en CSV. Pour info, le package utilisé est readxl : `library(readxl)` - Enfin, tout autre type de base de données peut être importé dans R Studio, le package utilisé va dépendre du format de la base de données. Des requêtes en SQL peuvent aussi être effectuées (voir la section "Introduction à SQLite", p.128). Attention, il faut bien définir le chemin du fichier à ouvrir, par exemple (pc) : `C:/folder/folder/Desktop/file.csv`

La librairie utilisée pour coller un tableau que l'on a copié par ailleurs (fichier, page HTML) est : `library(datapasta)` et `library(tibble)`. Le tableau peut être copié comme vecteur, comme `data.frame` etc. Les explications relatives à la manière de procéder se trouvent dans ce tutoriel en ligne : <https://cran.r-project.org/web/packages/datapasta/vignettes/how-to-datapasta.html>

Enfin, la librairie `rjson` permet d'ouvrir un fichier json.

```

11
12 library("rjson")
13 epistat_json <- fromJSON(file = "https://epistat.sciensano.be/Data/COVID19BE_CASES_AGESEX.json")
14 cov_json <- lapply(epistat_json, function(x) {
15   x[sapply(x, is.null)] <- NA
16   unlist(x)
17 })
18 cov_json <- as.data.frame(do.call("cbind", cov_json))
19 df_json <- cov_json %>%
20   t() %>%
21   as.data.frame(stringsAsFactors = F)
22
23 ?str
24 str(df_json)

```

3.6.2.8 Analyse de données

Tidyverse est un package comportant plusieurs librairies pour l'analyse des données. Les plus fréquemment utilisés sont les packages `dplyr` (manipulations) et `ggplot2` (visualisation). La plupart des fonctions des extensions du tidyverse acceptent des `data frames` en entrée, mais retournent un objet de classe `tibble` (pouvant être ensuite converti en `dataframe`). Les principales fonctions pour la manipulation sont : `filter`, `select`, `arrange`, `mutate` (pour la création d'une nouvelle variable avec des données dérivées, par exemple) et `slice` (pour sélectionner des lignes du tableau en particulier).

```

16 #Filtres
17 select_data <- gapminder %>%
18   select(year, country, region, life_expectancy, population) %>%
19   arrange(country)
20
21 filter_data <- select_data %>%
22   select(year, country, life_expectancy) %>%
23   filter(country %in% "France")
24
25 year_data <- select_data %>%
26   filter(year == 2016) %>%
27   select(country, life_expectancy) %>%
28   arrange(desc(life_expectancy))
29
30 taux_esp <- gapminder %>%
31   filter(year == 2015 & country %in% c("France", "Sweden")) %>%
32   select(country, life_expectancy)
33

```

Fichier attaché : intro-tidyverse.r

<https://github.com/laurence001/datajournalisme-R/blob/main/intro-tydiverse.r>

Documentation packages Tidyverse

<https://www.tidyverse.org/packages/>

Aller plus loin

- R for journalists : <https://learn.r-journalism.com/en/>
- R for beginners : <https://www.statmethods.net/r-tutorial/index.html>
- Packages de R : <https://www.computerworld.com/article/2921176/business-intelligence/great-r-packages-for-data-import-wrangling-visualization.html>



ggplot2

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. [Go to docs...](#)



dplyr

dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges. [Go to docs...](#)



tidyr

tidyr provides a set of functions that help you get to tidy data. Tidy data is data with a consistent form: in brief, every variable goes in a column, and every column is a variable. [Go to docs...](#)



readr

readr provides a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes. [Go to docs...](#)



purrr

purrr enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors. Once you master the basic concepts, purrr allows you to replace many for loops with code that is easier to write and more expressive. [Go to docs...](#)



tibble

tibble is a modern re-imagining of the data frame, keeping what time has proven to be effective, and throwing out what it has not. Tibbles are data.frames that are lazy and surly: they do less and complain more forcing you to confront problems earlier, typically leading to cleaner, more expressive code. [Go to docs...](#)



stringr

stringr provides a cohesive set of functions designed to make working with strings as easy as possible. It is built on top of stringi, which uses the ICU C library to provide fast, correct implementations of common string manipulations. [Go to docs...](#)



forcats

forcats provides a suite of useful tools that solve common problems with factors. R uses factors to handle categorical variables, variables that have a fixed and known set of possible values. [Go to docs...](#)

3.6.3 Introduction à SQLite

SQL est un langage de requête utilisé pour interroger des bases de données relationnelles. MySQL consiste en une de ses variantes, le langage étant ici utilisé dans le contexte du web pour faire interagir une base de données (PHPMyAdmin) avec une interface (page HTML codée en PHP) sur le plan des contenus. C'est la manière dont fonctionnent les systèmes de gestion de contenus. Dans WordPress, par exemple, chaque nouveau contenu ajouté (article, page, long format...) est stocké dans une base de données. Chaque fois qu'un utilisateur va afficher une page réalisée avec WP, le contenu sera appelé directement dans la base de données : contenu, structure de l'interface et style de l'interface sont donc les trois variables qui vont permettre l'affichage d'une page.



SQLite consiste en une autre variante de SQL sauf qu'ici, le système de requête n'est pas basé sur un serveur (comme SQL) mais sur un fichier¹⁰. **Requêtes**

- Sélectionner tout le tableau :
`SELECT * FROM input`
- Sélectionner une valeur particulière :
`SELECT * FROM input WHERE Ville = 'Ukkel'`
- Sélectionner les valeurs uniques d'une seule colonne (pour supprimer les doublons) :
`SELECT DISTINCT Ville FROM input`
- Trier le tableau pour l'affichage d'une colonne sous la forme d'une liste alphabétique descendante (DESC, ASC pour ascendante) :
`SELECT * FROM input ORDER BY Ville DESC`
- Afficher les valeurs d'une seule seule par liste alphabétique descendante :
`SELECT DISTINCT Ville FROM input ORDER BY Ville DESC`
- Sélectionner deux colonnes en particulier avec un tri alphabétique ascendant sur une des deux colonnes :
`SELECT Ville,07/09 FROM input ORDER BY Ville ASC`
- Sélectionner toutes les valeurs se terminant par une chaîne de caractère précise, par exemple toutes les villes dont le nom contient ou se termine par "beek" (le caractère % signifie qu'il peut y avoir des lettres avant et qu'il peut y avoir des lettres après) :
`SELECT * FROM input WHERE Ville LIKE '%eek%'`
- Sélectionner les valeurs contenues en X et Y dans une colonne (range) :
`SELECT * FROM input WHERE date1 BETWEEN 20 AND 40`
- Même action mais avec un tri ascendant de la date et une valeur plus grande que 40 :

¹⁰ Voir, pour davantage de précisions techniques "SQLite : une approche SQL lightweight" : <https://www.smalsresearch.be/sqlite-une-approche-sql-lightweight/>

```
SELECT * FROM input WHERE date1 > 20 ORDER BY date1 ASC
```

— Limiter le nombre de lignes affichées :

```
SELECT * FROM input LIMIT 5
```

Documentation SQLite et tutos : <http://www.sqlitetutorial.net/>

3.6.4 Outils pour l'analyse du discours

Les techniques d'analyse de données peuvent également s'appliquer aux données textuelles, par exemple pour une analyse de sentiments sur un sujet en particulier abondamment commenté sur les réseaux sociaux; ou encore pour épinglez les termes les plus fréquemment utilisés dans le discours d'un politique (nuages de mots). Des ressources et codes (en R, via R Studio et Iramuteq) sont disponibles sur cette page : <https://journaldev.tech/outils-pour-lanalyse-du-discours/>. Plusieurs outils en ligne permettent également de réaliser des nuages de mots :

- WordArt : <https://wordart.com/>

- Jason Davies WordCloud : <https://www.jasondavies.com/wordcloud/>

- Word It Out : <https://worditout.com/word-cloud/create>

- Introduction au text mining (R) : <https://www.tidytextmining.com/index.html>

3.7 Visualiser

Pour visualiser des données, les outils sont de deux types : des outils en ligne (gratuits ou payant) dont le code sera intégré dans une page web et des bibliothèques JavaScript qui vont servir à réaliser ses propres datavisualisations. Il existe une grande variété d'outils disponibles, cette section vous en donne un aperçu sans pour autant prétendre à l'exhaustivité : il s'agit donc ici d'une sélection au regard des usages et besoins journalistiques. A noter également que, chaque année, des outils disparaissent et d'autres apparaissent : c'est là la principale limite des services proposés en ligne, dont on ignore tout de la pérennité. De plus, certains outils gratuits aujourd'hui peuvent devenir payants demain, et inversement. En matière de choix des couleurs, certaines règles particulières doivent être observées : elles doivent présenter un degré suffisant de luminosité, elles ne doivent pas obscurcir l'information et être distinguées facilement ¹¹. Data Color Picker est un outil qui permet de sélectionner une palette de couleurs pour des dataviz, dans une perspective UI (user interface).

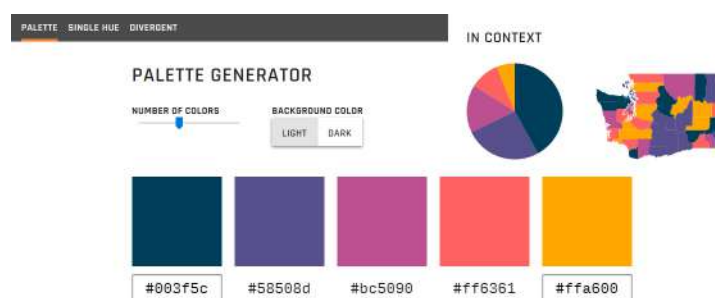


FIGURE 3.6 – Exemple d'utilisation d'une palette de couleur UI. Source : <https://learnui.design/tools/data-color-picker.html>

Dans un contexte journalistique, il convient également de respecter les bonnes pratiques en matière de sémiologie graphique car la vocation première de ces représentations visuelles n'est pas esthétique : elles doivent faire sens. N'oublions pas les trois fonctions de la représentation graphique définies par Jacques Bertin : enregistrer l'information, la communiquer en créant une image simple et mémorisable, et la traiter.

3.7.1 Outils prêts à l'emploi

Les exemples ont été développés à partir du jeu de données relatif à l'évolution des faillites en Belgique, au format CSV (source INASTI) :

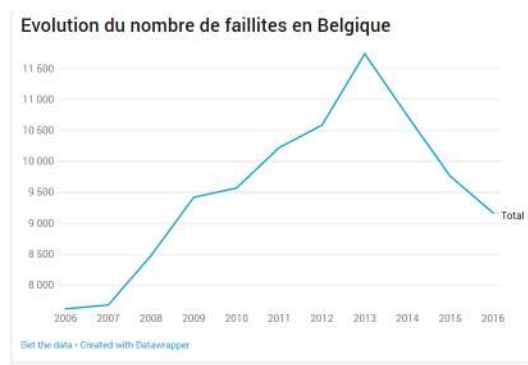
Année,Total
2006,7616
2007,7680
2008,8476

¹¹ Voir : <https://blog.graphiq.com/finding-the-right-color-palettes-for-data-visualizations-fcd4e707a283>

2009,9420
2010,9570
2011,10224
2012,10587
2013,11740
2014,10736
2015,9762
2016,9170

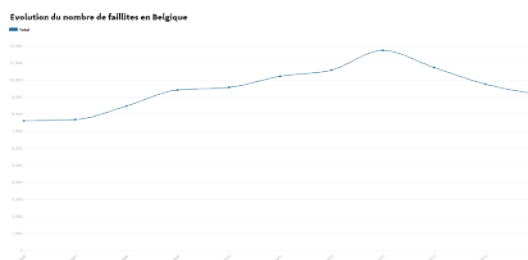
Datawrapper

Propose de la réalisation de graphiques ainsi qu'un outil de cartographie. Tient également compte du daltonisme (color blind check disponible) : <https://www.datawrapper.de/>



Flourish

Permet de créer des "story" à partir de plusieurs visualisations de données ainsi que d'exporter l'image du graphique avec un fond transparent au format PNG (permet de placer son propre fond) : <https://flourish.studio/>



Dataviz catalogue (ressources et liens utiles) <https://datavizcatalogue.com/>

Les outils de cartographie sont un peu moins nombreux. Jusqu'à son passage à des formules payantes assez onéreuses, Google Maps était largement utilisé. Parmi les alternatives, citons notamment :

- Carto : <https://carto.com>
- MapBox : <https://www.mapbox.com/>

Voir également les outils "data" sur [journodev.tech](http://www.journodev.tech/tools/) : <http://www.journodev.tech/tools/>

3.7.2 Visualisation de données avec R

La visualisation de données dans RStudio est possible via deux packages : ggplot (qu'il ne faut pas appeler) et ggplot2 (qu'il faut appeler via la fonction library). Le premier donnera une vue des données assez basique (pour l'analyse) tandis que le second permettra de réaliser des visualisations beaucoup plus élaborées (c'est, par exemple, le package utilisé par le Washington Post pour ses graphiques interactifs). L'utilisation la plus simple est la suivante : plot(variable) - pour appeler ggplot2 : require(ggplot2)

plot(x) #Visualiser un vecteur (manière simple)

ggplot2 permet donc des visualisations plus élaborées. Le principe d'une visualisation consiste à définir le type de visualisation souhaité, ainsi que les données à afficher via la fonction aes (aesthetic) :

geom_point()

geom_line()

geom_histogram()

geom_bar()

Exemple de code pour un simple graphique en ligne :

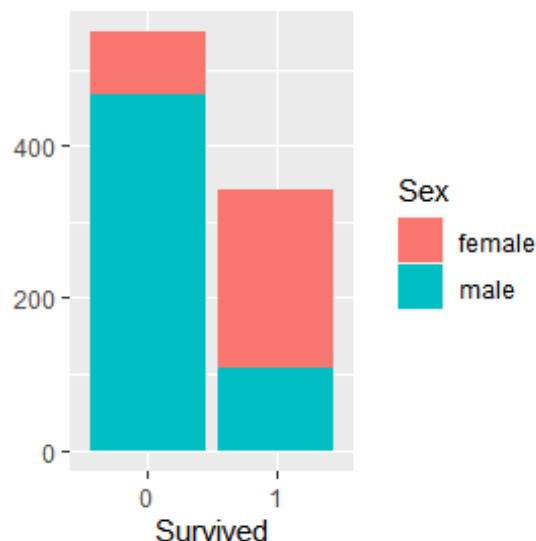
tableau %>% #pipe : pour poursuivre l'exécution du code

ggplot(aes(variable1, variable2)) + ## pour poursuivre l'exécution du code avec ggplot2

geom_line()

Exemple de code pour un graphique en barres empilées :

titanic %>% ggplot(aes(Survived, fill = Sex)) + geom_bar()

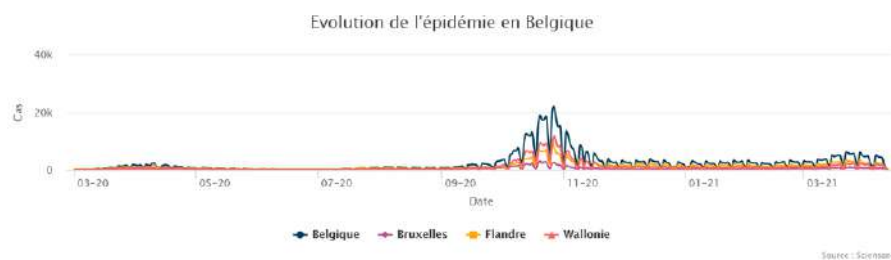


Tutoriel en ligne : <http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html>
et ici <http://www.storybench.org/getting-started-data-visualization-r-using-ggplot2/>

De nombreuses librairies JavaScript sont également disponibles sous la forme de package pour R. Généralement, leur mode de fonctionnement est semblable à celui du code en JS, à ceci près que la nomenclature du code a été adaptée pour R. C'est le cas du packager Highcharter (issu de la librairie Highcharts, lire infra). De nombreux types de graphiques interactifs sont disponibles et il est également possible de réaliser des cartes interactives (dans ce cas, un fichier geojson est requis). Une visualisation générée avec Highcharter peut être sauvegardée dans un format image ou sous la forme d'un fichier HTML. La logique de codage avec Highcharter est similaire à celle de ggplot2, à ceci près que "aes" devient ici "hcaes".

Exemple : évolution du nombre de cas de Covid-19 en Belgique

```
all %>%
hchart( 'spline', hcaes(x = Date, y = Cas, group = Région) )
hc_colors(c("#003f5c", "#bc5090", "#ffa600", "#ff6361")) %>%
hc_xAxis(dateTimeLabelFormats = list(month = "%m-%y")) %>%
hc_tooltip(dateTimeLabelFormats = list(day = "%d-%m-%y")) %>%
hc_title(text = "Evolution de l'épidémie en Belgique") %>%
hc_credits(text = "Source : Sciensano", href = "https://epistat.wiv-isp.be/covid/", enabled =
TRUE)
```



Fichier attaché : intro-highcharter.r

<https://github.com/laurence001/datajournalisme-R/blob/main/intro-highcharter.r>

Ressources

Documentation package Highcharter : <https://www.rdocumentation.org/packages/highcharter/versions/0.8.2>
https://rstudio-pubs-static.s3.amazonaws.com/304105_70f2ad540827454e934117e3d90f6c1a.html

Tutoriel : <https://jkunst.com/highcharter/>

3.7.3 Bibliothèques JavaScript

Les éléments interactifs d'une page web sont souvent le résultat d'une programmation en langage JavaScript (JS). De nombreux "frameworks" ont été développés pour ce langage, leurs fonctionnalités étant très variées. Par exemple : Bootstrap.js pour le design, Highcharts.js et D3.js pour les graphiques, Leafletmap.js pour la cartographie... *"Un framework est un espace de travail modulaire qui consiste en une collection de fichiers, qui contiennent des classes et des fonctions, et de conventions permettant le développement rapide d'applications. Il fournit suffisamment de briques logicielles et impose suffisamment de rigueur pour pouvoir produire une application aboutie et facile à maintenir. Ces composants sont organisés pour être utilisés en interaction les uns avec les autres"*¹². Les frameworks présentés ci-dessous nécessitent des connaissances de base en HTML. Ils présentent l'avantage de pouvoir réaliser rapidement des visualisations (graphiques et cartes) interactives de bonne qualité. Chacune de ces bibliothèques trouve un équivalent dans R. Si seulement trois bibliothèques sont présentées ci-dessous, il est bon de savoir que la bibliothèque D3.js est également beaucoup utilisée dans les projets de datajournalisme (mais elle est beaucoup plus complexe à prendre en mains et nécessite de solides bases en programmation javascript).

3.7.3.1 Highcharts.js (package R : Highcharter)

Highcharts est une bibliothèque JavaScript que l'on peut intégrer dans ses pages web pour réaliser des graphiques interactifs. Cette intégration est un peu plus complexe que celle de la bibliothèque Leaflet.js (voir point suivant) mais elle n'est pas du tout insurmontable à condition de rester précis et rigoureux : en principe, un graphique comportant trois variables peut être réalisé en moins d'un quart d'heure. Une connaissance plus poussée de la bibliothèque permet de travailler davantage sur le style des graphiques via du code CSS.

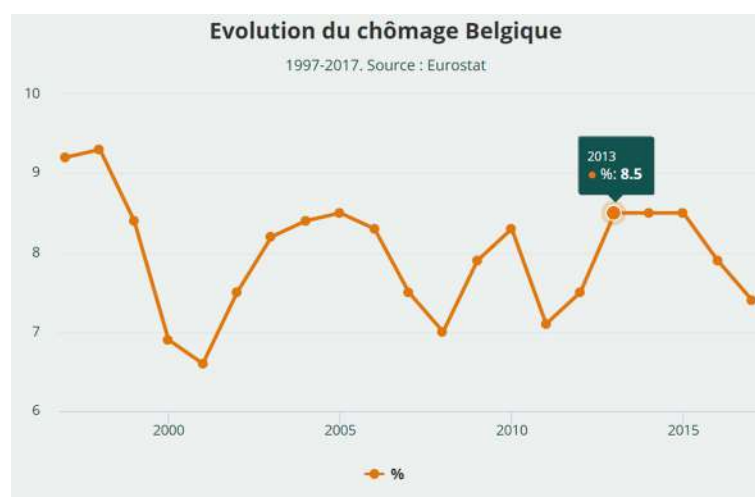


FIGURE 3.7 – Exemple de customisation de Highcharts, source : <https://www.alterechos.be/4erevolution/episode2>

¹² Source : <http://www.techno-science.net/?onglet=glossaire&definition=1471>

Ce tutoriel propose de suivre pas à pas la création d'un graphique, en utilisant le framework Boilerplate (ce qui permet de ne pas devoir coder entièrement la page HTML en disposant des fichiers de base à l'édition dans un éditeur de code de type Notepad++, Bracket ou Sublime).

1) Télécharger Boilerplate : <https://html5boilerplate.com/>

2) Ouvrir le fichier index.html, placer le lien jQuery dans la balise <head>

```
<script src="https://ajax.googleapis.com/ajax/libs/jquery/1.11.2/jquery.min.js"></script>
```

3) Placer le lien Highcharts dans la balise <head>

```
<script src="https://code.highcharts.com/highcharts.js"></script>
```

4) Créer un conteneur dans la balise body avec un identifiant unique

```
<div id="container" class="highcharts-container"> </div>
```

C'est dans ce conteneur que s'affichera la visualisation.

5) Se rendre sur <https://www.highcharts.com/docs/chart-and-series-types/chart-types> et choisir le type de graphique que l'on souhaite.

6) Se rendre sur les références de l'API pour copier les paramètres du graphique <https://www.highcharts.com/docs/getting-started/how-to-set-options>

Modèle de base (graphique en barres)

Script à placer au-dessus du conteneur.

```
<script>
$(function() {
var chart1 = Highcharts.chart({
chart : {
renderTo : 'container',
type : 'bar'
},
title : {
text : 'Fruit Consumption'
},
xAxis : {
categories : ['Apples', 'Bananas', 'Oranges']
},
yAxis : {
title : {
text : 'Fruit eaten'
}
},
},
```

```

series : [{
  name : 'Jane',
  data : [1, 0, 4]
}, {
  name : 'John',
  data : [5, 7, 3]
}]
});
});
</script>

```

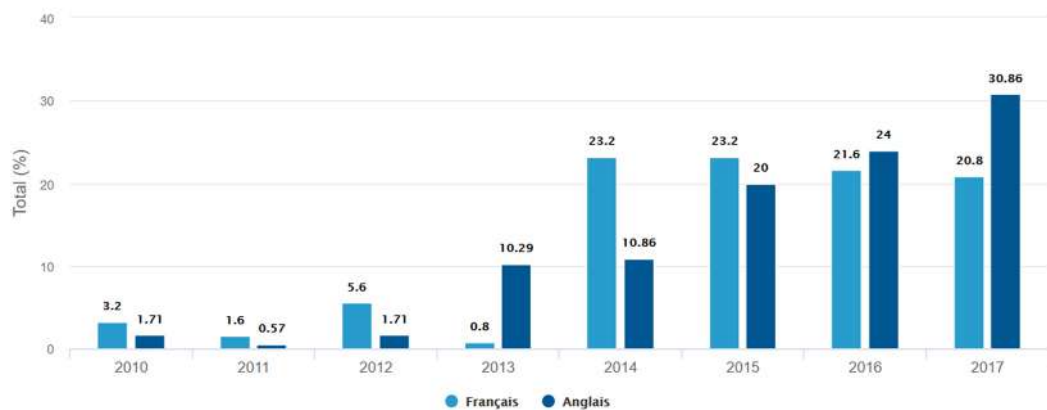


FIGURE 3.8 – Exemple de réalisation avec Highcharts

Exemple de graphiques et code source <http://www.ohmybox.info/exemple-hicharts/> (mot de passe : web)

Par ailleurs, Highcharts comporte une extension Highmaps qui permet de réaliser des cartes choroplèthes. Voir sur cette page un exemple de code source : <https://journodev.tech/une-carte-choroplethe-avec-highmaps-js/>

3.7.3.2 Leaflet.js (package R : leaflet)

Leafletmap est une bibliothèque permettant de réaliser, simplement, des cartes interactives. Ce tutoriel propose de suivre pas à pas la création d'une carte simple, en utilisant le framework Boilerplate (ce qui permet de ne pas devoir coder entièrement la page HTML en disposant des fichiers de base à l'édition dans un éditeur de code de type Notepad++, Bracket ou Sublime).

1) Télécharger Boilerplate : <https://html5boilerplate.com/>

2) Télécharger Leaflet <http://leafletjs.com/>

Placer le répertoire dézippé dans le répertoire Boilerplate.

OU passer directement au point 3 et suivre la seconde option (CDN)

3) Ouvrir le fichier index.html, placer le lien vers jQuery, la CSS et le JS

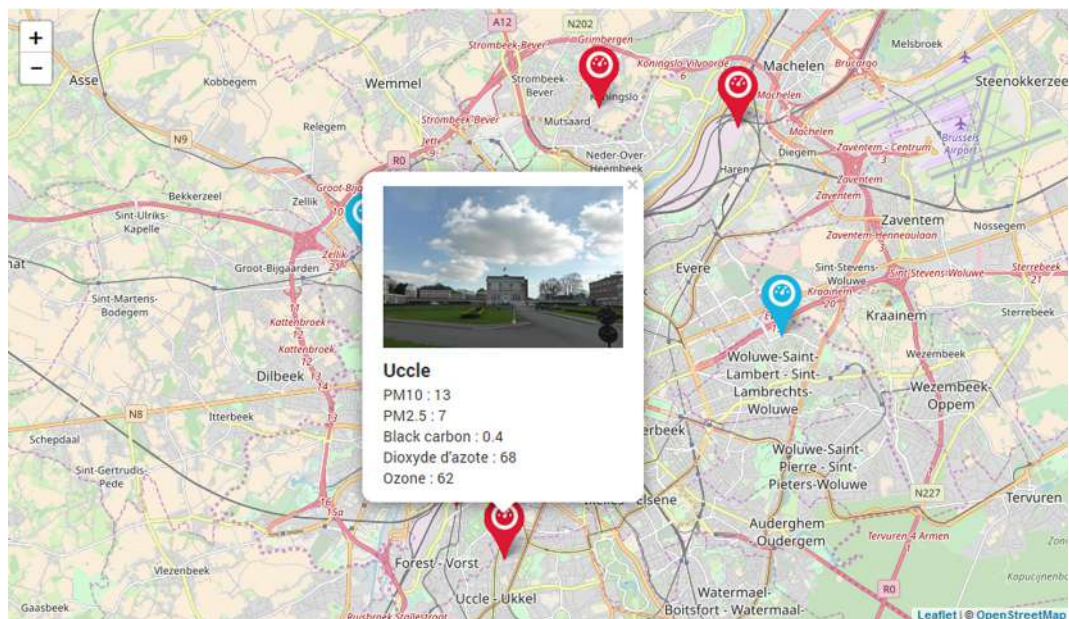


FIGURE 3.9 – Exemple de réalisation avec Leaflet map, la "Pollucarte" de l'application Bxl'air bot : <http://bxlairbot.be/carte.php>

<link rel="stylesheet" href="leaflet/leaflet.css" />

<script src="leaflet/leaflet.js"></script>

OU si on ne souhaite pas une installation en local, copier-coller les lignes d'accès aux fichiers CSS et JS de Leaflet (CDN) :

<link rel="stylesheet" href="https://unpkg.com/leaflet@1.3.4/dist/leaflet.css" crossorigin="" />

<!-- Make sure you put this AFTER Leaflet's CSS -->

<script src="https://unpkg.com/leaflet@1.3.4/dist/leaflet.js" crossorigin=""></script>

4) Créer un div "mapid" et ouvrir le fichier CSS, en lui donnant une hauteur (height) OU lui donner directement une hauteur dans la balise <div id="mapid" style="height:450px;"></div>

5) Créer le JavaScript en ouvrant main.js

Choisir un fond de carte (attention, tous ne vont pas nécessairement fonctionner en raison des dalles non prévues pour la zone géographique ou de la nécessité de disposer d'une clé API) :

<https://leaflet-extras.github.io/leaflet-providers/preview/>

Pour indiquer la latitude et la longitude, voir latlong.net

Ressources

Starter Guid : <http://leafletjs.com/examples/quick-start/>

Exemple fichier HTML : <http://bl.ocks.org/awoodruff/e9739a6719e0604eef58>

Tuto : <https://maptimeboston.github.io/leaflet-intro/>

Un marqueur personnalisé? Suivez le guide : <http://leafletjs.com/examples/custom-icons/>

3.7.3.3 Tableaux de données interactifs (package R : DT)

Deux bibliothèques JavaScript permettent de réaliser des tableaux de données interactifs soit en intégrant les éléments de la bibliothèque dans une page HTML, soit en travaillant sous R Studio avec le package R ad hoc.

DataTables

La documentation de la librairie est disponible à l'adresse suivant : <https://datatables.net/manual/>

Un exemple du code en R (très court, package DT) et en HTML (beaucoup plus long) est disponible sur cette page : <https://datatables.net/manual/>

Commune	Province	Cas
Bruxelles	Bruxelles Capitale	2191
Anderlecht	Bruxelles Capitale	1707
Schaerbeek	Bruxelles Capitale	1469
Molenbeek-Saint-Jean	Bruxelles Capitale	1355
Ixelles	Bruxelles Capitale	868
Uccle	Bruxelles Capitale	783
Forest	Bruxelles Capitale	660
Jette	Bruxelles Capitale	657
Saint-Gilles	Bruxelles Capitale	578
Evere	Bruxelles Capitale	519

Affichage de l'élément 1 à 10 sur 19 éléments (filtré de 582 éléments au total) Précédent 1 2 Suivant

FIGURE 3.10 – Exemple de réalisation avec DataTable

JSspreadsheet

JSspreadsheet (anciennement JExcel) est une librairie très puissante qui permet de réaliser des tableaux de données interactifs avec un nombre important d'options de configuration (tri, filtre, mise en forme...). La documentation de cette librairie est disponible sur cette page (de nombreux codes sources sont disponibles) : <https://bossanova.uk/jspreadsheet/v3/> (le package R correspondant est "excelR").

3.7.3.4 Visualisation de réseaux, vis.js (package R : visNetwork)

Vis.js permet de créer des cartographies de réseaux interactives ainsi que des lignes du temps. Il s'agit d'une bibliothèque de visualisation dynamique basée sur un navigateur. Elle est conçue pour être facile à utiliser, gérer de grandes quantités de données dynamiques et permettre la manipulation et l'interaction avec les données. Celle-ci se compose de : DataSet, Timeline, Network, Graph2d et Graph3d. Le principe de base d'une visualisation en réseau s'articule autour de "nodes" (noeuds du réseau) et de "edges" (relations entre les noeuds). Démon : <https://visjs.github.io/vis-network/examples/>

Voir aussi la visualisation de réseaux sur R : https://github.com/laurence001/datajournalisme-R/blob/main/intro-flux_bubbles_reseaux.R

3.8 Gérer

Gérer ses données devient une nécessité lorsque l'on collectionne les jeux de données et que ceux-ci sont sujets à diverses transformations. Une bonne gestion permet de retrouver ses données facilement (et de ne pas les égarer). Une manière de gérer ses jeux de données consiste à créer un tableau Excel avec les champs suivants : identifiant, nom du jeu de données, description, date de création, auteur, source.

Les principes d'une bonne gestion de ses jeux de données consistent à :

- Documenter ses jeux de données : de quoi s'agit-il? de quand datent-elles?
- Bien nommer ses jeux de données, ex : dataset_lannion_111215.xls
- Conserver une trace de chaque transformation de données. Ex : dataset_lannion_111215_v1.xls (gestion du versionnage)
- Procéder à des sauvegardes régulières : les données numériques sont fragiles et s'altèrent avec le temps. Par ailleurs, les supports (disques durs, clés USB) sont sensibles aux variations de température et d'humidité. Il est conseillé de procéder à un back-up complet de ses données au moins tous les deux ans (copie d'un disque dur vers un autre disque dur). Une copie peut également altérer la qualité d'un fichier (deux copies sont donc recommandées).
- Stocker ses données dans un cloud? Ce système de stockage n'est pas recommandé pour plusieurs raisons : qualité du prestataire (pérenne? gratuit avec la garantie que le service ne deviendra pas payant?), risques de hacking (accès à vos données par des tiers, qui pourraient les manipuler ou les supprimer)
- Stocker ses données dans un seul répertoire (méthode, organisation)

Identifiant unique	Nom du fichier	Titre du jeu de données	Description	Date de mise à jour	URL	Source
ID_001	mesdata_09_10_2017.csv					
ID_002	mesautresdata_11_10_2017.csv					

FIGURE 3.11 – Modèle de tableau pour la gestion de données

Un planning de gestion peut aider à considérer tous ces aspects. Un tel document comprend :

- Une description de la méthode de collecte des données
- Une description des jeux de données collectés (informations descriptives et contextuelles)
- De quels types de données s'agit-il? Quel format?
- Des considérations légales (propriété intellectuelle, données à caractères personnelles, anonymisation nécessaire pour exploitation cf Wikileaks, type de licence cf Creative Commons etc.) ou éthiques liées à l'usage ou l'exploitation des jeux de données (éventuelles autorisations)

— La manière dont les données sont stockées et préservées

Une bonne stratégie de gestion de données permet de s’y retrouver plus facilement (et donc, de gagner du temps!). Une bonne documentation permet de gagner du temps, surtout si le projet de datajournalisme s’inscrit dans la durée.

Ressources pour le data management

UK Data Service : <https://www.ukdataservice.ac.uk/manage-data>

Datalib : <http://datalib.edina.ac.uk/mantra/> (cours gratuit en ligne)

Note pour le nommage du fichier

Mot-clé explicite + date de l’enregistrement, pas d’espace blanc, pas de caractère accentué et si possible pas de majuscule.

Exemple : mesautresdata_11_10_2017.csv

Bibliographie

ABELSON, R. P. *Statistics as principled argument*. Psychology Press, 1993.

AGUIRRE HERNANDO, C. Backstage to the panama papers : big data analytics and collaborative journalism. *POLIS : journalism and society at the LSE* (2017).

ALBERTO, C. *The Functional Art : An introduction to information graphics and visualization*. Voices That Matter. Pearson Education, New York City, New York, 2012.

ANDERSON, C. W. Towards a sociology of computational and algorithmic journalism. *New Media & Society* 15, 7 (novembre 2013), 1015–1021.

ANDERSON, C. W. *Apostles of certainty : Data journalism and the politics of doubt*. Oxford Studies in Digital Politics. Oxford University Press, 2018.

ANDERSON, C. W., DOWNIE, L., AND SCHUDSON, M. *The news media : What everyone needs to know*. Oxford University Press, 2016.

ANNE, L. F. *Pratiques de la cartographie*. 128. Géographie. Armand Colin, 2007.

ANTHEAUME, A. *Le Journalisme numérique*. Nouveaux débats. Presses de Sciences Po (PFNSP), Paris, 2013.

BATINI, C., AND SCANNAPIECO, M. *Data quality : concepts, methodologies and techniques*. Data-Centric Systems and Applications. Springer, Cham, 2006.

BERRET, C., AND PHILLIPS, C. The computational turn : thinking about the digital humanities. *Columbia Journalism School* (2016), 97.

BERRET, C., AND PHILLIPS, C. Teaching data and computational journalism. *Columbia Journalism School* (2016b).

BERTIN, J. *Sémiologie graphique : les diagrammes, les réseaux, les cartes*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, 2013.

BOUILLON, P. *Traitement automatique des langues naturelles*. Champs Linguistiques. De Boeck Supérieur, 1998.

BOYDENS, I. *Informatique, normes et temps : évaluer et améliorer la qualité de l'information : les enseignements d'une approche herméneutique appliquée à la base de données "LATG" de l'O.N.S.S*. Academia Bruylant, Louvain-la-Neuve, 1999.

- BOYDENS, I. L'océan des données et le canal des normes. *Les Annales des Mines*, 67 (juillet 2012), 22–29.
- BOYDENS, I. Dix bonnes pratiques pour améliorer et maintenir la qualité des données. (En ligne, site consulté le 12/03/2016). <https://www.smalsresearch.be/dix-bonnes-pratiques-pour-ameliorer-et-maintenir-la-qualite-des-donnees/>.
- BOYDENS, I. Open data et e-government, 2014.
- BOYDENS, I., AND VAN HOOLAND, S. Hermeneutics applied to the quality of empirical databases. *Journal of Documentation* 67, 2 (2011), 279–289.
- BRADSHAW, P. Data journalism. In *Ethics for Digital Journalists : Emerging Best Practices*, L. Zion and D. Craig, Eds. London, Routledge/Taylor & Francis, 2015, pp. 202–219.
- BRIGGS, M. *Manuel de journalisme web : Blogs, réseaux sociaux, multimédia, info mobile*. Eyrolles, 2014.
- CAIRO, A. *The truthful art : data, charts, and maps for communication*. New Riders, 2016.
- CAIRO, A. *How charts lie : Getting smarter about visual information*. WW Norton & Company, 2019.
- CARLSON, M. The robotic reporter. *Digital Journalism* 2, 4 (2014), 1–16.
- CASWELL, D., AND DÖRR, K. Automated journalism 2.0 : Event-driven narratives : From simple descriptions to real stories. *Journalism Practice* 12, 4 (2018), 477–496.
- CHAVAND, F. *Des données à l'information*. ISTE Editions, 2017.
- CLEVELAND, W. S., AND CLEVELAND, W. S. *The elements of graphing data*, vol. 2. Wadsworth Advanced Books and Software Monterey, CA, 1985.
- CODDINGTON, M. Clarifying journalism's quantitative turn. *Digital Journalism* (2014), 1–18.
- COHEN, S., AND ALLI. Computational journalism. *Communications of the ACM* 54, 10 (novembre 2011), 66–71.
- COHEN, S., HAMILTON, J. T., AND TURNER, F. Computational journalism. *Communications of the ACM* 54, 10 (2011), 66–71.
- COLPAERT, P., JOYE, S., MECHANT, P., MANNENS, E., AND VAN DE WALLE, R. The 5 stars of open data portals. In *Proceedings of the 7th International Conference on Methodologies, Technologies and Tools Enabling e-Government* (2013), pp. 61–67.
- CORNU, D. *Journalisme et vérité : l'éthique de l'information au défi du changement médiatique*. Le champ éthique. Labor et Fides, 2009.
- COURMONT, A. Open data et recomposition du gouvernement urbain : de la donnée comme instrument à la donnée comme enjeu politique. *Informations sociales*, 5 (2015), 40–50.

- COX, M. The development of computer-assisted reporting. *Informe presentado en Association for Education in Journalism and Mass Communication*. Chapel Hill, EEUU : Universidad de Carolina del Norte (2000).
- CRAIG, D., KETTERER, S., AND YOUSUF, M. To post or not to post : Online discussion of gun permit mapping and the development of ethical standards in data journalism. *Journalism & Mass Communication Quarterly* 94, 1 (2017), 168–188.
- DAGIRAL, E., AND PARASIE, S. Data-driven journalism and the public good : "computer-assisted-reporters" and "programmer-journalists" in chicao. *New Media Society* 15, 6 (novembre 2012), 853–871.
- DAGIRAL, E., AND PARASIE, S. Des journalistes enfin libérés de leurs sources? promesse et réalité du "journalisme de données". *Sur le journalisme* 1, 2 (2013), 52–63.
- DALE, R. An introduction to natural language generation. *European Summer School in Logic, Language and Information, ESSLLI'95* (1995).
- DANIEL, A., AND FLEW, T. The guardian reportage of the uk mp expenses scandal : A case study of computational journalism.
- DANIEL, A., FLEW, T., AND SPURGEON, C. The promise of computational journalism. In *Media, Democracy and Change : Refereed Proceedings of the Australian and New Zealand Communications Association Annual Conference* (Canberra, 2010), Australia and New Zealand Communication Association, pp. 1–19.
- DANLOS, L., AND PIERREL, J.-M. C. P. *Ingénierie des langues*. in IC2 : information, commande, communication. Hermès Science publications, 2000.
- DE MAEYER, J., LIBERT, M., DOMINGO, D., HEINDERYCKX, F., AND LE CAM, F. Waiting for data journalism : A qualitative assessment of the anecdotal take-up of data journalism in french-speaking belgium. *Digital Journalism* 3, 3 (2015), 432–446.
- DE VEAUX, R. D., HAND, D. J., ET AL. How to lie with bad data. *Statistical Science* 20, 3 (2005), 231–238.
- DEFLEUR, M. H. *Computer assisted investigative reporting : Development and methodology*. Mahwah, NJ : Lawrence Erlbaum Associates, 1997.
- DESROSIÈRES, A. *Pour une sociologie historique de la quantification : L'argument statistique*. Collection Sciences sociales. Presses des Mines, Paris, 2008.
- DEUZE, M. The web and its journalisms : Considering the consequences of different types of newsmedia online. *New Media & Society* 5, 2 (2003), 203–230.
- DEUZE, M., AND BARDOEL, J. Network journalism : Converging competences of media professionals and professionalism. *Australian Journalism Review* 23, 1 (2001), 91–103.
- DIAKOPOULOS, N. A functional roadmap for innovation in computational journalism. *White paper* (2011).

- DIAKOPOULOS, N. Algorithmic accountability : Journalistic investigation of computational power structures. *Digital Journalism* 3, 3 (2015), 398–415.
- DIAKOPOULOS, N. *Automating the news : How algorithms are rewriting the media*. Harvard University Press, 2019.
- DIERICKX, L. News bot for the newsroom : How building data quality indicators can support journalistic projects relying on real-time open data. In *Global Investigative Journalism Conference 2017 Academic Track* (2017), Investigative Journalism Education Consortium.
- DIERICKX, L. The social construction of news automation and the user experience. *Brazilian Journalism Research* 16, 3 (2020), 432–457.
- DOMMERGUES, J.-Y., AND GROSJEAN, F. *La statistique en clair*. Ellipse, 2011.
- DÖRR, K. N. Mapping the field of algorithmic journalism. *Digital Journalism* 4, 6 (2016), 700–722.
- DÖRR, K. N., AND HOLLNBUCHNER, K. Ethical challenges of algorithmic journalism. *Digital Journalism* 5, 4 (2016), 404–419.
- DOURISH, P. No sql : The shifting materialities of database technology. *Computational Culture*, 4 (2014).
- FERNAND, J. *La cartographie*. Que sais-je? Presses Universitaires de France, Paris, 1999.
- FINK, K. The biggest challenge facing journalism : A lack of trust. *Journalism* 20, 1 (2019), 40–43.
- FLEW, T., SPURGEON, C., DANIEL, A., AND SWIFT, A. The promise of computational journalism. *Journalism Practice* 6, 2 (2012), 157–171.
- FLICHY, P., AND PARASIE, S. Sociologie des bases de données. *Réseaux* 2, 3 (2013), 178–179.
- FOUCART, T. L'interprétation des résultats statistiques. *Mathématiques et sciences humaines. Mathematics and social sciences*, 153 (2001), 21–28.
- FOX, C., LEVITIN, A., AND REDMAN, T. The notion of data and its quality dimensions. *Information processing & management* 30, 1 (1994), 9–19.
- FULLER, M. *Software Studies : A Lexicon*. EBSCO ebook academic collection. The MIT Press, 2008.
- GARRISON, B. Diffusion of online information technologies in newspaper newsrooms. *Journalism* 2, 2 (2001), 221–239.
- GILLESPIE, T. Relevance of the algorithms. In *Media Technologies : Essays on Communication, Materiality, and Society*, G. Tarleton (in), B. P.J., and F. K.A., Eds., Inside Technology. MIT Press, Cambridge, Massachusetts, 2014, pp. 167–194.

- GOËTA, S., AND MABI, C. L'open data peut-il (encore) servir les citoyens? *Mouvements*, 3 (2014), 81–91.
- GRAEFE, A. Guide to automated journalism. *Tow Center for digital journalism* (janvier 2016).
- GRAEFE, A., HAIM, M., HAARMANN, B., AND BROSIUS, H.-B. Perception of automated computer-generated news : credibility, expertise, and readability. *11th Dubrovnik Media Days, Dubrovnik* (2015).
- GYNNILD, A. Journalism innovation leads to innovation journalism : The impact of computational exploration on changing mindsets. *Journalism* (2013), 1–18.
- HAINAUT, J. *Bases de données. Concepts, utilisation et développement*. Dunod, 2012.
- HAMILTON, J. T., AND TURNER, F. Accountability through algorithm : Developing the field of computational journalism. In *Report from the Center for Advanced Study in the Behavioral Sciences, Summer Workshop* (2009), pp. 27–41.
- HAMMOND, P. From computer-assisted to data-driven : Journalism and big data. *Journalism* 18, 4 (2017), 408–424.
- HEFT, A., ALFTER, B., AND PFETSCH, B. Transnational journalism networks as drivers of europeanisation. *Journalism* (2017), 1464884917707675.
- HENNINGER, M. Data-driven journalism. *Reassessing Journalism* 157 (2013), 157–184.
- HERAVI, B. R. The state of data journalism global. In *Proceedings of the European Data and Computational Journalism Conference* (2017), pp. 5–7.
- HOWARD, A. The art and science of data-driven journalism (report). *Tow Center for Digital Journalism* (2014).
- HOWARD, A. The art and science of datajournalism.
- HUFF, D. *How to Lie with Statistics*. W. W. Norton, 2010.
- IFTIKHAR, H. The biggest leak : the panama papers. *POLIS : journalism and society at the LSE* (2016).
- JOANNES, A. *Bases de données et visualisation de l'information*. CFPJ Editions, Paris, 2010.
- JONES, K. S. Natural language processing : a historical review. *University of Cambridge* (2001), 2–10.
- KARLSEN, J., AND STAVELIN, E. Computational journalism in Norwegian newsrooms. *Journalism practice* 8, 1 (2014), 34–48.
- KRAEMER, F., VAN OVERVELD, K., AND PETERSON, M. Is there an ethics of algorithms? *Ethics and information technology* 13, 3 (2011), 251–260.

- LANTZ, B. *Machine learning with R : expert techniques for predictive modeling*. Packt publishing ltd, 2019.
- LATAR, N. L. The robot journalist in the age of social physics : the end of human journalism? In *The New World of Transitioned Media*, G. Einav, Ed. Springer, Cham, 2015, pp. 65–80.
- LEONARD, T. Databases in the newsroom : Computer-assisted reporting. *Online* 16, 3 (1992), 62–65.
- LEPPÄNEN, L., MUNZERO, M., SIRÉN-HEIKEL, S., GRANROTH-WILDING, M., AND TOIVONEN, H. Finding and expressing news from structured data. In *Proceedings of the 21st International Academic Mindtrek Conference* (2017), ACM, pp. 174–183.
- LEWIS, S. C., AND USHER, N. Open source and journalism : Toward new frameworks for imagining news innovation. *Media Culture Society* 35, 5 (2013), 4–9.
- LEWIS, S. C., AND USHER, N. Code, collaboration, and the future of journalism. a case study of the hacks/hackers global network. *Digital Journalism* 2, 3 (2014).
- LEWIS, S. C., AND WESTLUND, O. Big data and journalism : Epistemology, expertise, economics, and ethics. *Digital Journalism* 3, 3 (2015), 447–466.
- LINDEN, C.-G. Algorithms for journalism. *The Journal of Media Innovations* 4, 1 (2017b), 60–76.
- LOOSEN, W., REIMER, J., AND DE SILVA-SCHMIDT, F. Data-driven reporting : An on-going (r) evolution? an analysis of projects nominated for the data journalism awards 2013-2016. *Journalism DOI : 10.1177/1464884917735691* (2017).
- LOWREY, W., BROUSSARD, R., AND SHERRILL, L. A. Data journalism and black-boxed data sets. *Newspaper Research Journal* 40, 1 (2019), 69–82.
- LUTZ, E.-M., BIERNAT, E., AND LECUN, Y. *Data science : fondamentaux et études de cas. Eyrolles edition* (2015).
- MAIER, S. R. The digital watchdog's first byte : Journalism's first computer analysis of public records. *American Journalism* 17, 4 (2000), 75–91.
- MAIER, S. R. Digital diffusion in newsrooms : The uneven advance of computer-assisted reporting. *Newspaper Research Journal* 21, 2 (2000b), 95–110.
- MARK, B. *Manuel de journalisme web : Blogs, réseaux sociaux, multimédia, info mobile*. Eyrolles, 2014.
- MCBRIDE, R. E. D. The ethics of data journalism. *Digital Commons University of Nebraska-Lincoln* (2016), 1–44.
- MCCALLUM, E. Q. *Bad data handbook : Cleaning up the data so you can get back to work*. O'Reilly Media, 2012.

- MEYER, P. *Precision journalism : A reporter's introduction to social science methods*. Indiana University Press, Bloomington, 1973.
- MEYER, P. *The new precision journalism*. Indiana University Press, Bloomington, 1991.
- NAPOLI, P. M. Automated media : An institutional theory perspective on algorithmic media production and consumption. *Communication Theory* 24, 3 (2014), 340–360.
- NATHAN, Y. *Data visualisation : De l'extraction des données à leur représentation graphique*. Eyrolles, 2013.
- PARASIE, S. Des machines à scandale. *Réseaux*, 2 (2013), 127–161.
- PÉLISSIER, N., AND ROMAIN, N. Journalisme de presse écrite et nouveaux réseaux d'information. *Les Cahiers du Journalisme* 5 (1998), 54–71.
- POWELL, A. Making and measuring news : data and algorithms in journalism. *POLIS : journalism and society at the LSE* (2016).
- REDMAN, T. C. *Data Quality for the Information Age*. Artech House Telecommunications Library. Artech House, 1996.
- REITER, E., AND DALE, R. Building applied natural language generation systems. *Natural Language Engineering* 3, 1 (1997), 57–87.
- ROGERS, S. *Facts are sacred*. Faber & Faber, London, 2013.
- SANDOVAL-MARTÍN, M. T., AND LA-ROSA, L. Big data as a differentiating sociocultural element of data journalism : the perception of data journalists and experts. *Communication & Society* 31, 4 (2018), 193–209.
- SHANKS, G., AND CORBITT, B. Understanding data quality : Social and cultural aspects. In *Proceedings of the 10th Australasian Conference on Information Systems* (1999), vol. 785, Victoria University of Wellington, New Zealand.
- STONEMAN, J. Does open data need journalism?
- STRAY, J. The curious journalist's guide to data. *Tow Center for Digital Journalism* (2016).
- STRAY, J., BOUNEGRU, L., CHAMBERS, L., AND KAYSER-BRIL, N. *Guide du datajournalisme : Collecter, analyser et visualiser les données*. Eyrolles, Paris, 2013.
- SUMMIT, R. Reflections on the beginnings of dialog : the birth of online information access. *Dialog Corporation History* (2002).
- THOMAS, S., SETH, V. H., AND ED, S. MJ no more : Using concurrent Wikipedia edit spikes with social network plausibility checks for breaking news detection. In *Proceedings of the 22nd international conference on World Wide Web companion* (2013), International World Wide Web Conferences Steering Committee, pp. 791–794.

TRÉDAN, O. Quand le journalisme se saisit du web : l'exemple du datajournalism. In *Changements et permanences du journalisme*, F. Le Cam and D. Ruellan, Eds. L'Harmattan, Paris, 2014, pp. 199–214.

TUFTE, E. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 2001.

VERBORGH, R., AND DE WILDE, M. *Using OpenRefine*. Community experience distilled. Packt Publishing, 2013.

WAND, Y., AND WANG, R. Y. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM* 39, 11 (1996), 86–95.

WONG, D. M. *The Wall Street Journal guide to information graphics : The dos and don'ts of presenting data, facts, and figures*. WW Norton, 2010.