Media Laboratory (IMI)

Annexe 2

Démocratiser et favoriser la pratique du datajournalisme (I): un état des lieux

Lena Würgler
Académie du journalisme et des médias

Neuchâtel, février 2021

Préambule

La crise du Covid-19 a obligé la plupart des rédactions à revoir, en un laps de temps très court, non seulement leurs priorités et leur organisation, mais également leur pratique. Pendant plusieurs mois, l'attention médiatique s'est—presque de force—focalisée sur la pandémie, sur la connaissance du virus, sa propagation, ses conséquences, ses victimes. L'arrêt soudain, en Suisse comme ailleurs, d'une frange importante de l'activité économique, sociale et culturelle ne laissait plus beaucoup de marge de manœuvre aux journalistes, notamment dans le choix des sujets à couvrir. La plupart des médias ont donc consacré une bonne partie de leurs ressources et de leur énergie à couvrir « en direct » le développement de la pandémie. Cette couverture impliquait généralement de connaître les chiffres concernant la propagation du virus et ses conséquences.

De gré ou de force, même les rédactions sans expérience préalable se sont aventurées dans la pratique du datajournalisme, certes souvent de manière à la fois improvisée et basique. Certaines en avaient l'habitude, d'autres moins. Dans tous les cas, il s'est rapidement avéré que les données à disposition comportaient certaines limites : les différentes méthodes de comptage et les différents délais de mise à jour selon les pays, cantons, institutions, ainsi que la multiplicité des sources d'information et leur absence de coordination, ont agi comme un révélateur des difficultés de la pratique du datajournalisme et des éventuelles lacunes en la matière au sein des rédactions.

Introduction

Le présent document s'insère dans le contexte du projet MediaLaboratory, mené conjointement entre l'EPFL l'Académie du journalisme et des médias (AJM) et financé par L'Initiative for Media Innovation (IMI). L'objectif du projet consiste à développer des technologies et pratiques associées qui pourraient favoriser le recours au datajournalisme dans les rédactions. Ce document "d'état des lieux", élaboré par l'Académie du journalisme et des médias (UNINE), servira notamment à formuler des recommandations en vue du développement des outils et technologies prévu par le projet, qui seront développées par le Distributed Information Systems Laboratory – LSIR de l'EPFL . Concrètement, il offre une revue de la littérature portant sur l'exercice du journalisme de données ainsi qu'une exploration des pratiques en la matière en Suisse entre fin 2019 et début 2020.

Le projet part du principe qu'un outil « ne sera pas utilisé s'il ne rencontre pas les dimensions cognitives (…) et normatives (…) de la culture du journalisme (Singer 2003) » (Dierickx 2019, 155). L'objectif de ce document consiste donc à identifier ces dimensions afin d'orienter le développement d'outils de datajournalisme dans leur direction.

Nous verrons notamment que la pratique se divise clairement entre deux façons de pratiquer le datajournalisme : la première est centrée sur la recherche d'information, repose sur un travail collaboratif, demande du temps et est réalisée par des journalistes spécialisés. La deuxième se concentre sur les visualisations, repose sur un travail individuel, peut être réalisée rapidement et par des journalistes ayant des compétences techniques et statistiques mêmes rudimentaires. À ces deux façons de faire correspondent deux « communautés » relativement hermétiques. Émerge aussi la distinction entre une pratique du datajournalisme « à l'ancienne », « artisanale » et une pratique « avancée », (partiellement) automatisée, qui semble agir comme un idéal à la fois désirable et inatteignable. Nombreuses recherches ayant été menées suggèrent également que cette forme de journalisme, telle qu'elle s'imagine, nécessite d'importants moyens et se heurte souvent à des réalités telles que le manque de moyens à disposition, des problèmes organisationnels ainsi que des difficultés d'accès aux données.

Dans cette enquête exploratoire sur la pratique en Suisse, basée sur six entretiens semi-directifs avec des journalistes et responsables de rédactions, nous faisons l'état des lieux, décrivant notamment comment le datajournalisme se pratique à la *NZZ*, chez Tamedia, à la RTS, au sein des (futures) rédactions (partiellement) fusionnées que sont *Le Temps* et *Heidi.news*, ainsi que dans une forme embryonnaire au sein du groupe ESH Médias. Enfin, nous identifions les cinq freins suivants à la pratique du datajournalisme en Suisse, à son développement ou à sa démocratisation:

- Un manque de temps et de ressources;
- L'indisponibilité des données;
- L'inaccessibilité des données;
- Une incompréhension des données;
- Un manque de culture des données au sein des rédactions.

Datajournalisme: une revue de la littérature

Le journalisme de données n'est pas une pratique foncièrement nouvelle. Elle s'insère dans la continuité du Computer-assisted Journalism (CAR) et de l'infographie (Knight 2015, 56). Depuis quelques années toutefois se développe un journalisme fondé sur les big Data, soit des sets de données très importants (Lewis 2015, 322). La « naissance » du journalisme de données est ainsi généralement établie en 2010, lorsque le *Guardian*, le *New York Times* et Der *Spiegel*, ont publié les Afghan War Documents. « La publication de ces rapports peut être considérée comme le point culminant du journalisme de données ou, du moins, comme son apparition dans les répertoires de médias réputés » (Hahn et Stalph 2018, 3).

Nouveaux sujets et économie de moyens

Dès le départ, le datajournalisme est perçu comme une façon nouvelle, pour les journalistes, d'obtenir de potentiels sujets. Comme le soulignent Aitamurtro et al. (2011, 3), « Les journalistes voient les données comme une façon de trouver des histoires cachées – des histoires qui n'auraient autrement pas été racontées »[1]. Tandoc et Oh (2017, 1008) soulignent également que l'accès à de larges sets de données offre aux journalistes la possibilité de s'affranchir de l'actualité et de ne pas rester simplement réactifs aux événements extérieurs. Il permet donc aux journalistes de prendre l'initiative de leurs sujets. Actuellement, de nombreux médias misent sur le datajournalisme. En témoignent la constitution d'équipes spécialisées en la matière dans plusieurs médias, l'engagement de datajournalistes dans de nombreuses rédactions ainsi que l'inclusion de cours de datajournalisme dans les cursus de formation au journalisme [E1][2].

Du point de vue des entreprises médiatiques, le datajournalisme représente également une solution potentielle pour répondre aux principaux défis auxquels elles sont confrontées, notamment la chute des revenus publicitaires et la baisse subséquente et concomitante des ressources. Pour Lewis et Westlund (2015, 450), le datajournalisme est apparu au moment où « de nombreuses organismes de presse traditionnels luttent pour trouver leur voie face à la fragilisation de l'autorité professionnelle, des modèles d'affaire, et des logiques traditionnelles de production et de distribution médiatique ».

Pour certaines entreprises médiatiques, le datajournalisme est conçu comme un outil d'automatisation de tout — ou d'une partie — du travail journalistique. Des algorithmes sont conçus afin de produire en quelques minutes une grande quantité d'articles, tâche irréalisable humainement. A titre d'exemples, Tamedia en Suisse a développé l'outil Tobi pour produire des milliers d'articles différents en quelques minutes au sujet de résultats de votations (Plattner et Orel 2019), Quotebot est utilisé par le journal belge L'Echo pour rédiger automatiquement les « tops et les flops » de la bourse (Dierickx 2019) et SudPresse a développé un outil permettant de générer automatiquement des articles sur l'ensemble des matchs de l'Union belge de football, même dans les ligues inférieures [E2]. Ces outils, dits de data-verbalisation, ont systématiquement été développés en collaboration avec des sociétés externes, respectivement Automated Insights pour Tamedia (Plattner et Orel 2019),

[1] Nous avons traduit les citations issues de la littérature anglophone afin d'offrir une lecture plus fluide et une meilleure compréhension.
[2] Les six entretiens réalisés dans le cadre de cet état des lieux sont référencés [E1], [E2], …, [E6].

Syllabs à L'Echo (Dierickx 2019, 157) et Lab-Sense chez SudPresse [E2]. « Pour les entreprises de presse, le recours à ces technologies présente plusieurs avantages : rapidité d'exécution, couverture en temps réel, production de contenus à grande échelle, possibilité de générer différentes formes de représentations visuelles à partir d'un même jeu de données, multilinguisme et extension de leurs zones de couverture médiatique » (Dierickx 2019, 155). Globalement, selon Dierrickx (2019, 155), « le phénomène de la production automatisée d'informations implique que de nouveaux acteurs, issus du monde de la technique, prennent désormais une part active à un processus éditorial ».

Collaborations avec des entités externes

Aitamurto et al. (2011, 3) relèvent d'ailleurs que le datajournalisme a tendance à se pratiquer de plus en plus sous forme de collaboration entre des médias et des entités externes. « Le scenario futur le plus probable verra les organisations médiatiques exister dans un état de symbiose avec d'autres fournisseurs de services ». C'est le cas dans les exemples de data-verbalisation cités plus haut, où les entreprises médiatiques collaborent avec des sociétés qui, à l'origine, n'ont aucun lien avec le journalisme. Des collaborations existent aussi avec d'autres plateformes spécialisées dans la constitution de bases de données ou encore avec des sociétés ou organismes spécialisés dans le datajournalisme, comme WedoData, Correctiv ! ou encore la Shared Data Unit de la BBC. En termes d'usages, Aitamurto et al. (2011, 16) estiment que d'autres organisations externes aux médias vont se charger de la récolte et de la mise à disposition de données - tâche dont les médias n'ont pas les moyens de s'occuper -, pendant que les rédactions vont, elles, se concentrer sur l'analyse de ces données en fonction des besoins journalistiques, à l'aide d'outils libres d'accès et faciles à utiliser. Des plateformes telles que Lobbywatch sur les liens d'intérêts des parlementaires, Glamos sur la surveillance des glaciers, ou encore l'IMTAG (Indicateur des Mouvements du Trafic Aérien à Genève) sur le suivi des avions qui passent par l'aéroport de Genève, en sont des exemples suisses. L'ensemble de ces plateformes ont servi de bases pour des articles journalistiques[3], bien qu'elles n'aient pas, à l'origine, été créées spécifiquement dans ce but.

Le projet *Media Laboratory* s'insère dans cette tendance générale, puisqu'il est développé par une institution externe non spécialisée dans le journalisme (LSIR / EPFL) et qu'il a pour but de fournir aux journalistes de faciliter l'accès à et la pratique du datajournalisme. Son indépendance le différencie toutefois des formes de collaborations observées dans les exemples de data-verbalisation, alors que son but premier — être utile aux journalistes — le différencie des projets citoyens tels que Lobbywatch. Sa forme s'apparente donc plutôt à celle de projets tels que WeDoData ou de la Shared Data Unit. Les différentes organisations qui participent directement ou indirectement à la pratique du datajournalisme ont été résumées dans un tableau [Fig.1]. Le projet MediaLaboratory s'insère donc dans la seconde catégorie.

[3] Exemples:
1) https://www.heidi.news/climat/la-fonte-des-glaciers-alpins-se-poursuit-de-plus-en-plus;
2) https://www.heidi.news/economie/les-interets-croises-d-isabelle-chevalley-et-de-l-homme-d-affaires-jurg-staubli;
3) https://www.swissinfo.ch/fre/politique/cou-cou-_a-gen%C3%A8ve-un-robot-signale-les-avions-des-dictateurs/42544160.

Figure 1 : typologie de quelques exemples de collaborations en datajournalisme

Nom de l'outil	Créateur	Type de relation avec média(s)	But(s)	Public cible
Tobi	Automated Insights	Collaboration étroite	Outil utile à une rédaction	Tamedia (mandataire)
Shared Data Unit	BBC	Service « public », service à disposition des médias	Fournir des données (outils) aux rédactions	Rédactions multiples
Lobbywatch	Lobbywatch	Indépendance	Fournir des informations sur les liens d'intérêt des parlementaires	Public en général

Deux approches opposées

D'un point de vue de la méthode journalistique, le datajournalisme ne constitue pas une forme de journalisme clairement définie. Tout d'abord, Parasie (2015) a démontré qu'il existait deux approches différentes du travail journalistique à partir de données. L'une, appelée *data-driven approach*, consiste à utiliser des bases de données comme point de départ d'une recherche. C'est à partir des données que les journalistes trouvent et conçoivent un sujet. L'autre approche, appelée *story-driven approach*, prend le chemin inverse : le journaliste réalise un sujet que les données viennent ensuite illustrer, compléter, confirmer ou appuyer.

Cette distinction analytique s'avère pertinente pour rendre compte des différentes pratiques de datajournalisme existantes, qui se rapprochent tantôt plutôt de l'une ou de l'autre selon les journalistes ou selon les rédactions. « Il y a deux façons de débuter un sujet de journalisme de données. Dans la première option, un set de données fournit des informations supplémentaires pour un sujet qui avait déjà été découvert par la rédaction. Dans la seconde, un set de données sert de point de départ pour tout le sujet » (Aitamurto et al. 2011, 11). Dans ce dernier cas, la démarche est considérée comme s'apparentant au journalisme d'enquête (Bradshaw 2018, 19). Les données représentent alors une source parmi d'autres de pistes de sujet à traiter, mais ne contiennent pas encore, en elles-mêmes, une histoire. D'où la formulation régulièrement mentionnée par les professionnels selon laquelle « les données ne sont pas l'histoire ». Autrement dit, une « donnée n'a pas de valeur journalistique en soi ». Pour Paul Bradshaw (2018, 22-23), le journalisme a, dans le cas présent, « un langage curieux et trompeur : nous parlons de 'trouver des histoires', alors que ce qu'on entend vraiment est que l'on trouve des pistes pour des histoires. (…) souvent, les chiffres n'ont pas de sens en eux-mêmes. Il est important d'établir le contexte qui donnent du sens à ces chiffres».

La distinction entre ces deux approches permet également de mieux saisir la confusion existante concernant la définition du journalisme de données, qui est tantôt considéré comme un processus particulier (l'analyse de sets de données comme point de départ d'une enquête), tantôt comme un produit spécifique (une production journalistique contenant des visualisations de données) (Weber et al. 2018, 191). Comme le souligne Knight (2015, 59),
« (…) pour certains observateurs, le journalisme de données est fondamentalement la production de graphiques d'information (…). Pour d'autres, l'accent est mis sur d'importantes sources de données, souvent acquises par des fuites ou des demandes de transparence, dont l'analyse étendue et complexe est considérée comme essentielle, reliant ainsi le datajournalisme à la pratique du journalisme d'investigation ».

C'est dans cette dernière perspective que s'insèrent les leaks tels que les Offshore Leaks, les Football Leaks, les Panama Papers, les Dubai Papers, etc. De plus en plus régulièrement, des consortiums internationaux de journalistes d'enquête reçoivent en effet des fuites de terabytes de données non structurées (Plattner et al. 2016, 4), qui leur servent de point de départ à de vastes enquêtes internationales.

Les leaks : formats variés, équipes spécialisées

Ces leaks révèlent aussi que, dans le terme datajournalisme, la notion même de donnée porte à confusion. S'il renvoie la plupart du temps à l'idée d'une donnée chiffrée, de statistiques, ce n'est pas une condition nécessaire. Les projets de data-verbalisation en sont un exemple. Mais les exemples les plus emblématiques sont, justement, ces méga-leaks qui ont marqué la sphère médiatique ces dernières années et qui représentent une évolution fondamentale dans la pratique du métier, ouvrant la voie à des collaborations inter-médiatiques et transfrontalières. Dans ces grandes enquêtes, les données étaient à la fois très hétérogènes et non structurées. L'exemple des Panama Papers, l'une des plus grandes enquêtes réalisées par un consortium de journalistes international, l'a montré : les données utilisées étaient alors très disparates, contenant des emails, des PDF, ainsi que d'autres types de fichiers. Le traitement de ces données, en plus de requérir une plateforme cryptée d'accès pour les journalistes impliqués, a nécessité le développement du moteur de recherche Blacklight, lui aussi crypté, permettant l'exploration des données récoltées via une recherche par mots clés (Plattner et al. 2016, p.4). Le terme datajournalisme englobe donc plusieurs approches différentes, plusieurs types d'analyse, plusieurs types de données et ne se limite pas à la visualisation de sets de données chiffrées.

Loin de constituer une pratique rapide et évidente, le datajournalisme, en tant que processus ou méthode particuliers — et non de produit —, nécessite donc un investissement financier important pour les entreprises médiatiques, sans garantie d'un retour sur investissement (Aitamurto et al. 2011, 14). Ce coût ne semble toutefois pas les décourager puisqu'elles semblent prêtes à investir d'importantes sommes non seulement dans l'automatisation de certaines tâches récurrentes ou l'écriture d'articles personnalisés à faible plus-value journalistique, mais aussi dans des unités spécialisées dans le datajournalisme, regroupant des datajournalistes, des développeurs et des graphistes. Ces unités, sur lesquelles se sont concentrées la plupart des recherches académiques sur le datajournalisme, se développent dans les grandes structures médiatiques, telles que le *New York Times*, le *Washington Post*, le *Guardian*, le *Zeit*, le *Spiegel*, la *NZZ*, Tamedia ou encore la RTS/SRF. « Les équipes se

composent généralement d'un mélange de compétences en journalisme, en développement web, en analyse de données, en visualisation et en statistiques » (Aitamurtoet al. 2011, 12).

Les unités de datajournalisme réunissent donc souvent plusieurs professionnels de plus en plus spécialisés dans leur propre domaine (Stalph 2018, 7; Stalph, 2020, 7). Les grandes enquêtes de datajournalisme sont l'œuvre de ces unités spécialisées, travaillant par projets.

Au quotidien : sources officielles, données chiffrées et prétraitées

Ces grandes collaborations internationales, si elles font grand bruit, restent toutefois exceptionnelles. D'un point de vue analytique, le datajournalisme s'oriente en réalité plus souvent vers la visualisation que l'enquête. Dans une analyse de journaux anglais, Knight a montré que le datajournalisme, dans sa forme « quotidienne », était largement superficiel, fondé sur des données de sources institutionnelles et peu spectaculaires. Elle en conclut que « dans le quotidien des rédactions, le décryptage des données n'est pas devenu plus marquant, plus important, que d'autres formes de journalisme » (Knight 2015, 70). Elle observe que, pour les journalistes, la forme compte parfois autant, voire plus, que le fond : « la présentation des informations sous forme de données peut avoir son propre intérêt, quelle que soit la valeur réelle de l'information ou son impact sur la société » (70).

De son côté, Stalph (2018) a mené une analyse des productions de datajournalisme du *Guardian*, du *Zeit*, du *Spiegel* et de la *NZZ*. S'intéressant également aux formes « courantes », « quotidiennes » de datajournalisme plutôt qu'aux exemples emblématiques et prestigieux, l'auteur montre que la plupart des productions quotidiennes de datajournalisme sont fondées sur des données prétraitées, rarement issues d'un travail collaboratif. Selon lui, ces résultats indiquent que « la collecte et l'analyse de ses propres données n'est pas (encore) faisable pour un datajournalisme quotidien, car le temps et les ressources sont des facteurs limitants » (2018, 1347). En termes de format, les innovations sont rares et la plupart des visualisations se composent de bar charts ou de cartes, rarement interactifs (Stalph 2018, 1338). En somme, la forme la plus aboutie de datajournalisme, à savoir la constitution d'une base de données par des journalistes, son analyse et son utilisation comme point de départ d'un sujet, reste très rare pour les journalistes non spécialisés.

Concrètement, le traitement et la visualisation automatisés de données se concentrent actuellement sur certains sujets spécifiques, tels que les votations, les élections, les résultats de la bourse ou les résultats sportifs. « La nécessité de disposer de données structurées répondant à des exigences qualitatives, tant sur le plan technique que sur le plan journalistique, explique pourquoi les domaines d'application couverts sont actuellement limités au sport, à l'économie, aux résultats d'élections ou à l'environnement » (Dierickx 2019, 154-55). Les élections/votations constituent des moments privilégiés d'innovation en matière de visualisations automatisées. Les données étant simples et structurées, l'essentiel du travail consiste à automatiser leur récolte et leur représentation et à conceptualiser des formats innovants de visualisation.

« Au *Washington Post*, quelqu'un m'a dit : 'En 2016, on a perdu les élections'. Ce qu'il entendait, c'est que le *New York Times* avait alors vraiment fait de meilleures visualisations (…). Il y a vraiment une compétition entre les titres pour la meilleure visualisation. En somme, les votations sont des stimulateurs d'innovation » [E1]. La recherche de Stalph (2018, 1345) a également révélé que les élections constituent le sujet le plus traité en datajournalisme.

En résumé, il ressort de la revue de littérature que, si le datajournalisme en tant que produit particulier (visualisations) est fréquent, le datajournalisme comme processus (enquête) demeure rare. Ainsi, le premier est pratiqué à plus large échelle, par des journalistes tant généralistes que spécialisés, alors que le second reste l'apanage des datajournalistes ou des unités de datajournalisme.

Les promesses du journalisme de données

Lors de son apparition, le datajournalisme a rapidement été perçu comme chargé de promesses : d'un côté, il devait permettre de produire des articles à moindre coûts. D'un autre, il devait permettre aux journalistes de « se libérer de leurs sources » (Parasie et Dagiral 2013). Enfin, il devait permettre une forme « d'objectivation » de la réalité, inspirée des sciences dites « dures », alors que le journalisme s'inspirait jusqu'alors essentiellement des sciences sociales (Parasie 2015). Sans être complètement infirmées, ces promesses se sont avérées moins fructueuses que ne le laissaient penser les espoirs placés dans le datajournalisme.

Concernant la production d'articles à moindre coût, Parasie (2015, 14) a montré que les ressources nécessaires à la récolte, l'élaboration, l'analyse et l'interprétation des données peuvent représenter un coût financier élevé. Le même chercheur conteste aussi la deuxième promesse : si le journalisme de données permet effectivement de s'affranchir des déclarations de sources dites « autorisées », il reste dépendant de l'obtention de bases de données, qui sont la plupart du temps fournies par les institutions publiques (Parasie et Dagiral 2013, 61). La dépendance vis-à-vis de ces institutions persiste donc et l'absence de données concernant d'autres organisations puissantes, notamment dans le milieu économique, rend plus compliquée la surveillance de tout un pan de la société.

Enfin, le datajournalisme comportait aussi la promesse d'une forme d'objectivation de la réalité. Lewis et Westlund (2015, 449) soulignent qu'il existe une « croyance largement répandue selon laquelle de grandes bases de données offrent une forme supérieure d'intelligence et de connaissance qui peuvent générer des éclairages qui auraient été impossibles par le passé, avec une aura de vérité, d'objectivité et de justesse ». Bien qu'encore très prégnante, cette promesse est également contestée. Il apparaît plutôt que la configuration même des données, leur interprétation et leur représentation visuelle impliquent différents choix cognitifs et formels qui auront un impact sur ce que disent ou montrent finalement les données. « Les visualisations de données sont souvent représentées dans les discours publics comme des preuves objectives de faits. Toutefois, une visualisation n'est qu'une forme de traduction de la réalité, tout comme d'autres médias, dispositifs de représentation, ou modes ou représentations » (Kosminsky et al. 2019, 43). Comme le résument Kennedy et al. (2016, 722), « en elles-mêmes, les visualisations ne sont pas neutres : elles sont ancrées dans des conventions spécifiques, des pratiques et

philosophies historiques ». D'ailleurs, ils démontrent que les visualisations répondent à une série de conventions formelles que sont (a) une perspective bidimensionnelle ; (b) des formes et des lignes géométriques ; (c) l'inclusion de sources de données et (d) une présentation épurée. Cette dernière convention a un effet persuasif,

> « comme si les processus de visualisation des données n'étaient pas soumis à des décisions difficiles. Ce sentiment de simplicité masque la complexité des données et de leurs visualisations et contribue à donner l'impression que, dans la visualisation, nous pouvons 'voir' les données directement : les voici, claires et nettes » (Kennedy et al., 2017, 729)

Tandoc et Oh (2017) se sont penchés sur les productions du Data Blog du *Guardian*. Ils notent que la plupart (80%) des productions analysées ne contiennent aucune source humaine ni de commentaires de journalistes expliquant ce que signifient ces données. Ce faisant, le Data blog laisse supposer que les données parlent d'elles-mêmes. Or, selon Tandoc et Oh (2017, 1011), les données sont toujours subjectives. Une forte dépendance dans les données peut donner une « illusion » d'objectivité mais occulte le fait que « la récolte, l'analyse et la publication de données peuvent aussi être sujettes à manipulation ». Pour Lewis et Westlund (2015, 453), la donnée « ne représente pas une vérité objective. Comme Gitelman (2013) et d'autres l'ont souligné, la donnée brute est un oxymoron. Les chiffres issus de grandes données - même s'ils sont énormes, robustes et fortement corrélés - doivent encore être interprétés ».

Le risque principal de la pratique est ainsi une forme de fétichisation de la donnée, qui conduit à penser les données comme dénuées d'intervention humaine, et reflétant ainsi objectivement la réalité. Les données parleraient d'elles-mêmes. Or, la méthode même de récolte des données, à la source, peut avoir une influence sur leur sens et leur interprétation, comme l'a montré l'expérience de la crise sanitaire du covid-19, durant laquelle les différentes méthodes de récoltes de données par différents pays, voire cantons ou institutions, ont rendu les comparaisons internationales et intercantonales pratiquement impossibles. Ou, du moins, elle a montré que ces facteurs doivent être pris en compte dans l'interprétation et la visualisation des données. Ainsi, dans cet article de Swissinfo du 26 mars 2020, les journalistes mentionnent plusieurs fois les limites des comparaisons entre les différentes entités prises en considération :

> « Mais il est difficile de se livrer à ce type de comparaison, car chaque pays a sa méthode pour compter les cas et pour faire les tests. (…). Cependant, en raison des différences cantonales et de la difficulté de recueillir les chiffres des hôpitaux, il est impossible d'avoir des statistiques exactes. »

Cet exemple n'est pas unique. En pratique, relève Kosminsky (2019, 51), « la plupart des visualisations représentent des données qui sont déjà une représentation imparfaite et incomplète du monde ». En somme, le choix même des variables utilisées conditionne déjà la représentation de la réalité. Dans un deuxième temps, leur transformation en visualisation s'apparente seulement à une représentation possible parmi d'autres. Il y a donc une forme de double filtre entre la réalité et sa représentation. Ce constat n'implique pas que les datajournalistes créent ou inventent une réalité mais suppose de concevoir les données comme une forme de représentation possible, parmi d'autres, des faits du monde. « Sinon,

le risque est de sous-estimer leur influence sur le message des visualisations et de succomber à ce que nous appelons la croyance a priori - la considération non critique des visualisations comme une représentation suffisante de la réalité » (2019, 51). Kosminsky invite donc à systématiquement être attentif « au contexte dans lequel les données ont été collectées, représentées, encadrées et visualisées, et des éventuels biais qui ont pu influencer leur représentation » (2019, 52)

Les freins au journalisme de données

Les promesses rattachées au journalisme de données ne se concrétisent donc que partiellement dans la pratique, notamment celle selon laquelle il devait permettre de simplifier et accélérer le travail des journalistes. Cette « désillusion » s'explique par la présence de plusieurs « obstacles » pratiques très concrets, liés d'un côté aux caractéristiques des données, de l'autre aux compétences professionnelles.

Du côté des données, Aitamurto souligne que les freins à l'utilisation du datajournalisme dans les rédactions commencent déjà lors de la phase de récolte. Il note que « l'un des plus grands [défis] est d'obtenir les données » (Aitamurto 2011, 13), notamment en raison du refus des agences gouvernementales de fournir les données demandées. En second lieu, il est nécessaire que le format des données les rende exploitables. Dans le cas contraire, « l'analyse et le nettoyage des données peut exiger beaucoup de ressources de la part des institutions médiatiques » (Aitamurto 2011, 14). Valeeva (2017) constate également que la disponibilité, la forme et l'état des données compliquent souvent la pratique du datajournalisme, obligeant parfois les journalistes à compiler laborieusement des données manuellement. Pour la chercheuse, « les problèmes de standardisation et d'unification des données, ainsi que leur disponibilité, tout comme l'utilité (ou non) des portails de données, n'aident pas les journalistes à faire des sujets à partir de données. Au contraire, ils en constituent des obstacles importants » (Valeeva 2017, 12).

Les journalistes auraient donc besoin de trouver de meilleurs moyens pour obtenir, vérifier et nettoyer les données. « Pour raconter de meilleures histoires, les journalistes ont d'abord besoin de meilleures données : des ensembles de données granulaires et utiles, dans un format lisible par une machine » (Valeeva 2017, 25). L'enjeu principal lié à la qualité des données initiales concerne la pertinence et la validité des conclusions qu'il est possible d'en tirer. Si les journalistes utilisent des données dans le but de rendre public des faits encore inconnus, et que seule l'analyse des données leur permet de révéler, ils doivent s'assurer que leurs affirmations sont bel et bien correctes. Comme l'a observé Parasie (2015), « le désordre et les imprécisions des données ont également soulevé des préoccupations d'ordre éthique. » (Parasie 2015, 371)

Cependant, le frein le plus important, selon Aitamurto (2011, 2), est à chercher du côté des professionnels : « Le principal obstacle qui empêche les journalistes de lancer des projets de données semble donc être un manque de connaissances sur la façon de travailler avec les données ». D'où la régulière et presque systématique collaboration de journalistes avec des spécialistes des données. Mais ces formes de collaborations entre journalistes et techniciens ne sont, selon Lewis et Westlund (2015, 456), « ni faciles, ni largement institutionnalisées actuellement ». Or, pour les auteurs, toute tentative d'implémentation du datajournalisme doit viser à combler ces décalages de compétences.

Un état des lieux du datajournalisme en Suisse

La revue de littérature proposée ci-dessus offre un panorama de la recherche sur le journalisme de données. Cependant, il n'offre aucune visibilité de l'état actuel des pratiques de journalisme dans les rédactions en Suisse. C'est pourquoi nous avons mené une série de six entretiens exploratoires ayant pour but de dégager des premières réponses aux questions suivantes, issues de la littérature :

- Dans quelle mesure les « obstacles » observés se retrouvent-ils dans le quotidien des médias suisses romands ?
- Quels médias/journalistes pratiquent le datajournalisme en Suisse et sous quelle forme ?
- Les médias romands sont-ils intéressés par cette forme de journalisme ?
- Sous quelle forme le datajournalisme est-il déjà pratiqué dans les rédactions?
- Quelles sont les attentes qu'en ont les rédactions qui le pratiquent ou qui souhaiteraient le pratiquer?
- Quels obstacles empêchent les rédactions de pratiquer le datajournalisme avec succès ?

Pour tenter de répondre (sommairement et superficiellement) à ces questions, six entretiens ont été menés entre le 1er octobre 2020 et le 15 février 2021. Le choix des interviewés visait à varier les profils afin d'obtenir une perspective aussi large que possible. Les entretiens ont été menés avec : (1) une chercheuse, Rahel Estermann, qui réalise une thèse sur la communauté des datajournalistes en Suisse, (2) un responsable du développement d'un média en Belgique, (3) une journaliste locale suisse romande, (4) un rédacteur en chef d'un journal local suisse romand, (5) un datajournalist d'un médias suprarégional suisse romand et (6) une journaliste d'un média en ligne suprarégional suisse romand. Il s'agit d'entretiens approfondis d'environ deux heures, semi-directifs, orientés autour de trois grands thèmes : 1) les pratiques actuelles de datajournalisme, 2) les facteurs limitants la pratique du datajournalisme, 3) les potentiels besoins en la matière.

Jusqu'à présent, les études menées sur le datajournalisme en Suisse se sont concentrées sur les journalistes et unités spécialisées en la matière. A notre connaissance, ces recherches sont celles de Florian Stalph (2018, 2020) et Rahel Estermann, dont le travail de doctorat est en cours. Le premier a mené des entretiens avec des membres des cellules data du *Guardian*, du *Spiegel* et de la *NZZ*. La seconde a réalisé des observations et des entretiens auprès de journalistes de la communauté hack/hackers et au sein de l'unité Storytelling de la *NZZ*.

Les éléments de réponse ci-dessous sont issus des six entretiens ainsi que des recherches susmentionnées concernant spécifiquement la pratique du datajournalisme en Suisse.

Une poignée de journalistes spécialisés

Le datajournalisme reste encore, en Suisse, une forme de journalisme pratiquée essentiellement par des journalistes spécialisés. Ces journalistes se réunissent dans une

forme de communauté informelle que Rahel Estermann a étudiée dans le cadre de sa thèse de doctorat. Ses observations et les entretiens qu'elle a menés lui ont permis de relever une différence de culture entre les journalistes appartenant à cette communauté et ceux non spécialisés dans les données. La culture des journalistes de données, essentiellement collaborative et transparente, entre en conflit avec celles des journalistes 'traditionnels', plus individuelle, concurrentielle et plus discrète concernant les sources utilisées.

Le travail collaboratif est l'un des grands changements induits par le journalisme de données. Cette culture collaborative se fonde, selon Rahel Estermann, sur deux aspects culturels : d'abord sur l'idée que le partage d'information est bénéfique à tout le monde et que l'expérience des autres datajournalistes permet à chacun de continuer à se former en continu [E1]. Ainsi, à l'opposé du journalisme traditionnel, qui reste une pratique largement individuelle, les journalistes de données privilégient le travail collaboratif, qu'il soit avec d'autres journalistes ou avec des professionnels spécialisés dans l'analyse de données ou le développement web, le design ou l'infographie. La culture de la donnée – celle des développeurs ou hackers – plus ouverte et plus transparente, influence ainsi la pratique des datajournalistes.

Contenu contextuel versus anecdotique

Un autre fondement du datajournalisme est l'interactivité (Lewis et Westlund, 2015, p.450), soit la possibilité, pour le public, d'interagir avec (ou dans) la visualisation produite. Rahel Estermann a observé que la recherche d'interactivité est désormais supplantée, chez les datajournalistes suisses, par une volonté de « prendre le lecteur par la main » pour comprendre les données. Ils privilégient les « productions +explicatives », qui consacrent une large part à la mise en contexte. Ils décrivent leur travail comme étant « contextuel », par opposition aux contenus « anecdotiques » réalisés par leurs confrères.

« Ce qui réunit aussi les datajournalistes, c'est que beaucoup d'entre eux ne veulent plus communiquer sur le monde comme avant. La *NZZ* propose par exemple des 'Erklärer Stücke' qui donnent diverses informations sur un même sujet, sur le contexte, sur l'évolution de la question traitée, etc. Il y a une volonté d'expliquer les choses avec un contexte plus large. Mais cela entre en contradiction avec le besoin médiatique habituel de 'titres' et de délais rapides. Les Datajournalistes se voient plus comme des scientifiques. Ce qu'ils disent souvent, c'est : 'Jusqu'à maintenant, le journalisme a écrit des anecdotes. Maintenant, nous voulons écrire des 'connexions' (Zusammenhänge) » [E1].

Une observation confirmée par Rinsdorf et Boers (2016, 2) selon qui la pratique du datajournalisme peut être vue comme « faisant partie d'un changement général dans le journalisme, passant d'une focalisation sur l'actualité et les scoops à des informations de fond, ainsi qu'à l'explication des tendances actuelles ». Les datajournalistes ont aussi pour habitude d'accompagner leurs productions de « méta-récits » concernant l'obtention et l'analyse des données (Weber et al. 2018, 204).

Conflits de cultures

Les membres de la « communauté » des datajournalistes partagent ainsi une forme de sous-culture journalistique commune, inspirée du monde de l'informatique, pour laquelle ils ressentent un fort sentiment d'appartenance. Selon Florian Stalph (2020, 11), cette sous-culture « transcende l'espace institutionnel et agit comme une force unificatrice inter-organisationnelle : cela remet en question les structures organisationnelles historiques des salles de rédaction (...) qui doivent maintenant s'adapter à la sous-culture du journalisme de données, ce qui pourrait remettre en question une culture organisationnelle actuellement dominante et conduire à un conflit normatif » . Autrement dit, la culture des datajournalistes se distingue de celle des rédactions traditionnelles, conduisant dans certains cas à des divergences de conception du métier difficilement conciliables. Dans son observation de la communauté des datajournalistes suisses, Rahel Estermann observe d'ailleurs que la plupart de ses membres se sentent quelque peu « exclus » de leur rédaction, incompris par leurs collègues, en raison notamment de leur grande fascination pour les données [E1].

La perspective des datajournalistes, proche de celle des hackers, s'éloigne ainsi de la perspective traditionnelle des journalistes. Si bien que, lorsqu'une entreprise médiatique souhaite implémenter le datajournalisme dans sa ou ses rédactions, elle peut vite être confrontée soit à des réticences, soit à un rejet ou à de la méfiance. « Les différents agents des organisations médiatiques peuvent interpréter le phénomène du big-data de façon différente et, en conséquence, l'aborder de diverses manières : en lui résistant, en s'y adaptant, en intervenant contre lui ou en le refaçonnant » (Lewis et Westlund 2015, 450-451). L'implémentation du datajournalisme dans des rédactions ayant leurs propres habitudes, routines et pratiques nécessite donc de tenir compte des réactions observées parmi les collaborateurs.

Les divergences de perspective ne s'observent pas seulement entre datajournalistes et non-datajournalistes, mais aussi entre les membres d'une même unité data. Les points de vue d'un designer, d'un informaticien ou d'un journaliste concernant le traitement d'un même set de données peuvent conduire à des désaccords quant à la manière de les utiliser. Cependant, Rahel Estermann note que, comme ces unités sont intégrées dans des structures médiatiques, « c'est finalement la vision journalistique qui prévaut. Le critère selon lequel il faut avant tout trouver une 'bonne histoire' reste dominant ».

Les unités data en Suisse

Malgré ces différences culturelles avec leurs confrères non spécialisés, la plupart des datajournalistes de la communauté hack/hackers suisse travaillent dans les rédactions de médias traditionnels. A l'instar de leurs pairs européens, ils sont le plus souvent intégrés à des cellules dédiées au datajournalisme ou à la visualisation, dont l'existence atteste, en soi, de l'intérêt démontré par les médias pour cette nouvelle pratique journalistique. Cependant, le rôle et la structure de ces unités varient fortement d'un média à l'autre, du moins en Suisse.

L'unité Storytelling de la NZZ

A la *NZZ*, une équipe de journalistes, développeurs et graphistes composent l'unité storytelling. Depuis sa naissance en 2012 jusqu'à 2015, l'unité s'appelait « Data NZZ » et était composée d'un à deux datajournaliste(s) selon les périodes. En 2015, elle n'était composée plus que d'un journaliste qui collaborait avec l'unité *Interactives teams*, elle-même composée d'un développeur et de deux designers. Fin 2015, avec l'arrivée d'un nouvel éditeur, l'unité Storytelling a été créée en absorbant et agrandissant l'ancienne équipe avec des infographistes, des datajournalistes, des développeurs interactifs et des développeurs de frontend (Stalph 2020, 7). L'unité Storytelling est conçue autour de deux concepts :

> « D'abord, ils produisent leurs propres histoires, principalement des 'pièces explicatives sur des sujets complexes qui ont une valeur concrète pour le lecteur et l'aident à comprendre' (DE5). L'équipe soutient également d'autres desks au quotidien, en apportant des idées de sujet ou en développant des formats de narration pour des sujets en cours de réalisation (DE5) (…). Deuxièmement - et cela semble être une caractéristique unique dans les organisations examinées - l'équipe de Storytelling agit comme un facilitateur technologique pour l'ensemble de la rédaction dans le but d'équiper d'autres reporters internes avec des outils leur permettant de produire eux-mêmes des sujets et des éléments visuels fondés sur des données »[4] (Stalph 2020, 8).

En l'occurrence, la *NZZ* a développé une boîte à outils, appelée Q, développée et rendue disponible publiquement en 2017. Q propose des fonctions d'exploration et de visualisation de données, destinées à permettre à tous les membres de la rédaction de pratiquer du datajournalisme. Q permet de créer des graphiques, d'effectuer des recherches à partir de l'outil lui-même, l'automatisation de la mise à jour, la recherche de nouveaux outils et l'adaptation du mode selon les compétences (expert vs débutant).

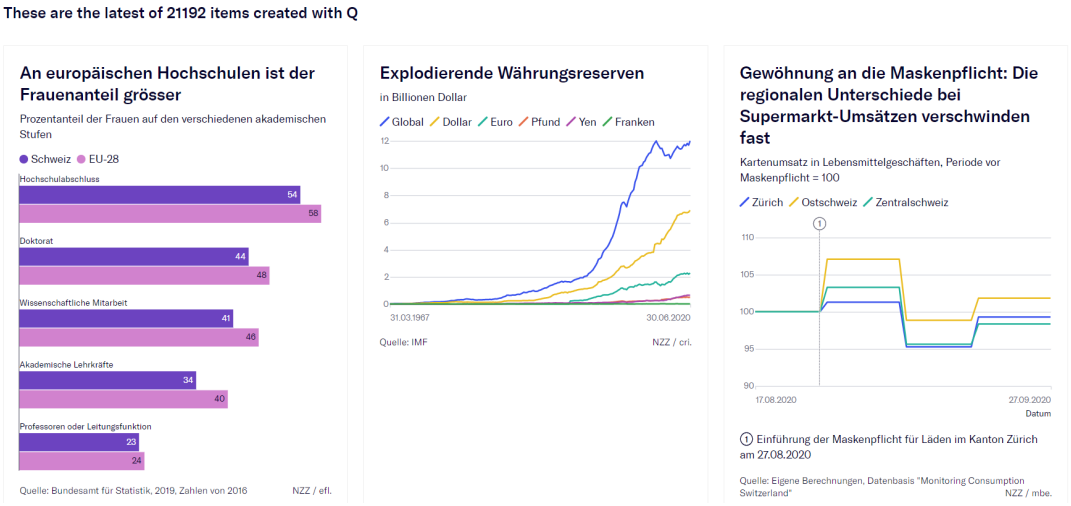
D'après un membre de l'équipe interrogé par Stalph (2020, 8), « ce kit personnalisé, fabriqué en interne, s'est avéré plus utile que d'autres outils tiers qui 'sont souvent un peu trop complexes ou offrent trop d'options, de sorte que nous ne pouvons garantir aucune cohérence. Plus fondamentalement, les équipes n'ont plus à se familiariser avec différents outils, mais seulement à une plateforme intégrée' » (DE5). Stalph (2020, 10) considère que l'équipe Storytelling officie comme « teacher et helper », permettant de favoriser la réalisation de sujets fondés sur des données ou des visuels. Un des objectifs de Q était, justement, de pouvoir décharger l'équipe Storytelling de tâches répétitives et basiques, pour lui permettre de se concentrer sur des sujets et des visualisations plus complexes, irréalisables avec Q [E1].

Les observations de Rahel Estermann, conjuguées à la lecture d'articles de blog écrits par les membres de l'unité Storytelling, montrent que l'implémentation de l'outil a demandé beaucoup d'efforts, à la fois techniques et humains. Techniquement, il a notamment fallu trouver des solutions pour que les visualisations puissent être directement exploitées dans le print à partir de Q, sans passer entre les mains d'un designer. Selon les membres de la

[4] Le code DE5 n'est pas à confondre avec E5. Le premier réfère à l'entretien avec un Data Editor tel qu'il est référencé par dans Stalph (2020), selon sa convention de notation. Le deuxième, indiqué entre crochets [E5], réfère au 5e entretien mené dans le cadre de ce projet.

l'équipe Storytelling, l'outil « Q-to-Print », développé pour permettre ce transfert, « a demandé des changements de processus dans le quotidien de beaucoup de gens, ainsi qu'une communication habille et claire » (Wiederkehr, 2019). L'outil devait en outre, impérativement, rester aussi simple que possible, pour être exploitable par le plus grand nombre de journalistes. L'équipe de développement a donc éliminé certaines pistes afin de garantir cette simplicité. Les graphiques, notamment, restent très basiques [fig.1].

Figure 2: Exemples de graphiques réalisés avec l'outil Q de la NZZ



Ensuite, les développeurs de l'outil ont dû consacrer du temps et de l'énergie à faire connaître l'outil aux collaborateurs, à développer des tutoriaux, organiser des workshops, proposer des formations individuelles, etc. Ils ont également mis en place ce qu'ils appelaient des « Quesdays », soit des jours durant lesquels tous les journalistes de la rédaction devaient obligatoirement créer une visualisation pour leur article, afin de se familiariser avec les fonctionnalités de Q. Enfin, l'équipe Storytelling a également effectué un important travail de contrôle de qualité, en vérifiant chacune des publications utilisant Q. Toutes ces mesures visaient non seulement à permettre aux journalistes de s'approprier l'outil, mais aussi à amoindrir leurs potentielles réticences à devoir modifier leur routine. Actuellement, le groupe médias *NZZ* (NZZ Mediengroup) envisage d'étendre l'utilisation de cet outil à l'ensemble de ses titres, soit également aux titres locaux [E1].

Tamedia

Chez Tamedia, ce sont des outils d'écriture automatique ou d'archivage et de recherche de données qui ont été développés. L'outil Tobi, a permis aux rédactions du groupe de produire 39'996 articles concernant des votations, adaptés selon la langue (2 variables), les préférences de vote des lecteurs (23+1 variables) et la commune d'habitation des lecteurs (2222 variables) (Platner et Al., 2018, p.1). L'algorithme avait été développé par l'entreprise Automated Insight et adapté aux besoins du groupe. Pour que la forme des articles soit la plus 'humaine' possible, et contienne aussi des éléments d'analyse, cinq journalistes expérimentés ont pré-écrit des briques de texte utilisables par l'algorithme. « Nous avons essayé d'éviter les répétitions et voulions faire en sorte que nos textes soient plus qu'une simple énumération de pourcentages, rangs, et nombres de votants. » (Plattner et al, 2018,

p.2). L'équipe de développement a donc créé une base de données Excel contenant plus de 300'000 cellules, avec 137 variables par commune. A terme, Tamedia envisage d'utiliser le même type de data-verbalisation pour d'autres sujets, par exemple les résultats sportifs ou des résultats économiques.

Entre 2016 et 2018, Tamedia a également développé l'outil Tadam, abréviation de Tamedia Data Mining Project. Il vise à constituer une base de données réunissant différents types de documents récoltés par les journalistes de Tamedia durant leur travail quotidien. Son but principal est de permettre ensuite une recherche par mots clés dans l'ensemble de ces documents, dans trois langues. Il se fonde sur le constat que « les groupes de presse sont confrontés à un monde de plus en plus complexe, qui produit littéralement un flot permanent de données. Dans le même temps, la plupart d'entre eux se débattent avec des ressources limitées et souvent en déclin pour la production d'un journalisme de fond » (Platner et Orel, 2016, p.1). A nouveau, l'outil a été conçu à partir d'un software développé par une société tierce, Expert System, qui œuvrait jusque-là essentiellement pour des compagnies d'assurance ou des agences d'état. La base de données est constamment nourrie par deux types de sources : des sources statiques, à savoir des documents insérés par les journalistes, et des sources dynamiques, c'est-à-dire des informations récoltées automatiquement à partir de certains sites internet cibles. Tadam a été prévu de manière à permettre une recherche par mots clés dans tous types de documents et dans n'importe quelle langue. Un système de traduction permet ainsi d'effectuer une recherche en français et d'obtenir des résultats en allemand. Comme dans le cas de l'implémentation de Q à la *NZZ*, celle de Tadam chez Tamedia comporte, selon les développeurs, un défi majeur : celui de son acceptation et de son utilisation effective par les professionnels :

> « Travailler avec un système comme Tadam nécessite non seulement une formation courte, mais aussi un accompagnement individuel étroit, afin de guider le changement de mentalité qu'il requiert. (...) Et le changement d'attitude, qui consiste à passer des habitudes de loup solitaire à une collaboration transparente, est un grand pas pour de nombreux journalistes. Afin de faciliter ce changement, un travail de persuasion patient est nécessaire. Le management intermédiaire doit être impliqué pour mener à bien ce changement » (Platner et Orel, 2016, p.5)

Les outils Tadam et Tobi restent toutefois limités à l'usage des rédactions du groupe et ne sont pas accessibles à d'autres journalistes/rédactions.

Tamedia compte aussi plusieurs datajournalistes, engagés soit dans sa cellule-enquête, soit dans un titre particulier. Depuis 2016, le groupe comprend notamment l'Interaktive Team, qui crée des sujets interactifs de différents genres. Dans leur cas, l'accent est avant tout mis sur la recherche de formats innovants de représentation. Pour la crise du Covid-19, l'équipe a créé un tableau de bord sur « Les chiffres les plus récents sur l'épidémie de Covid-19 en Suisse »[5]. Ils possèdent aussi un page consacrée aux « chiffres les plus récents sur le réchauffement climatique »[6]. D'autres formats narratifs et visuels ont été conçus selon les

[5] « Die neusten Zahlen zur Corona-Pandemie » (https://interaktiv.tagesanzeiger.ch/2020/covid-19-ausbruch-im-vergleich/)
[6] « Die neusten Daten zum Klimawandel », (https://interaktiv.tagesanzeiger.ch/2020/daten-zum-klimawandel-in-der-schweiz/)

projets, du scrolling au BD-reportage en ligne, en passant par des Quiz interactifs[7]. Côté romand, le data-scientist Duc-Quang Nguyen travaille pour les titres régionaux 24 Heures et la Tribune de Genève. Il a notamment conçu une représentation géométrique des cantons suisses pour suivre l'épidémie de covid-19 [fig.2] ou encore créé un format interactif sur l'exploitation des sols en Suisse [fig.3].

Figure 3 : carte "en tuiles" (tilemap) des cantons suisses développée par Duc-Quang Nguyen pour 24 Heures (https://blog.datawrapper.de/tilemap-of-swiss-cantons/)

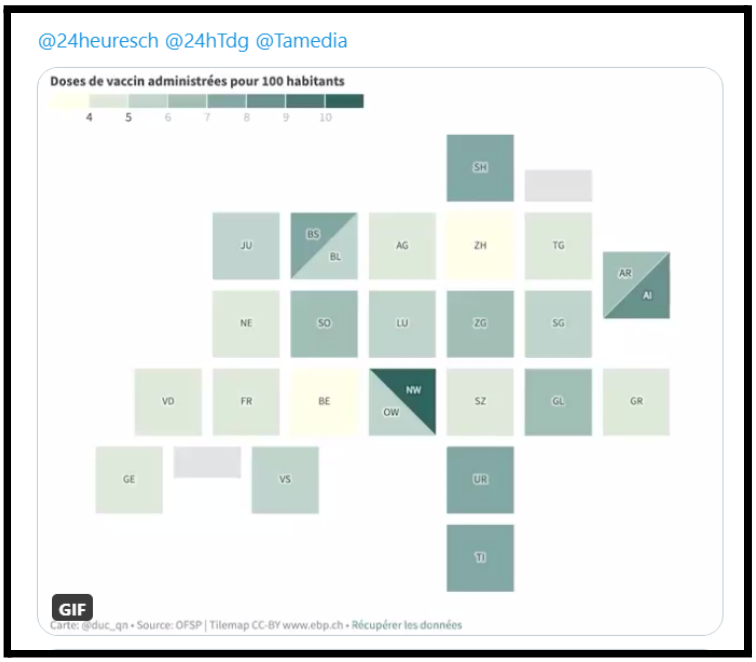
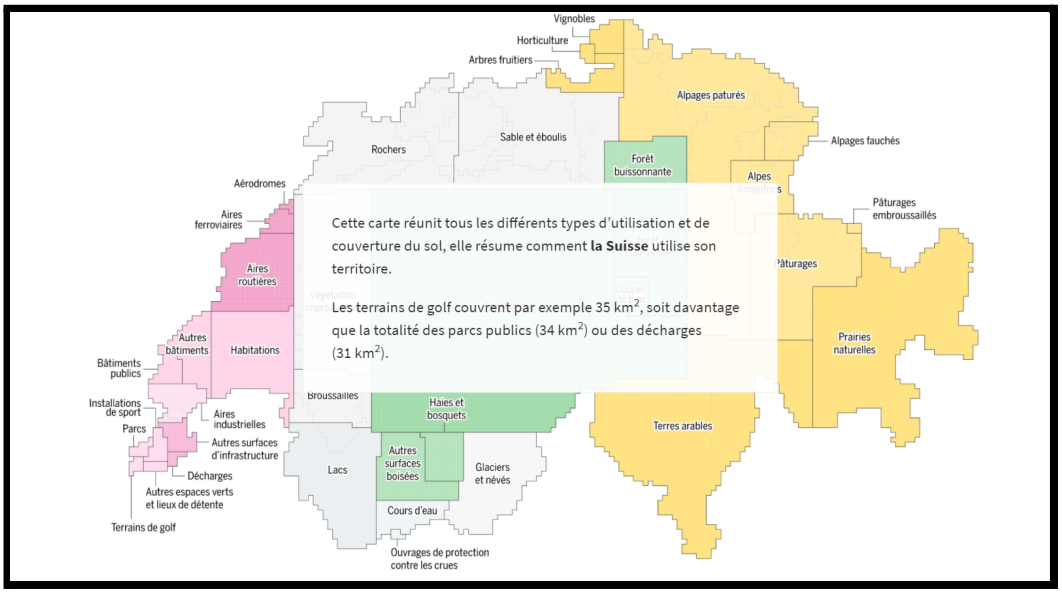


Figure 4 : carte interactive de l'occupation du sol en Suisse créée par Duc-Quang Nguyen pour 24 Heures (https://interactif.24heures.ch/2018/utilisation-du-territoire/?openincontroller)



[7] Voir notamment https://hccd.hypotheses.org/1406

Ainsi, chez Tamedia, le datajournalisme englobe plusieurs démarches différentes et se répartit dans plusieurs unités différentes, tantôt focalisées plutôt sur l'enquête, tantôt plutôt sur la création de formats innovants, tantôt sur la création d'outils simplifiant le travail quotidien des journalistes.

Deux versions à la RTS

De son côté, la RTS possède depuis octobre 2018 une équipe spécialisée dans le datajournalisme, la cellule « data-innovation », qui produit ses propres recherches à partir de données. Elle officie de manière relativement indépendante à la rédaction. L'équipe comprend trois journalistes (Marc Renfer, Tybalt Felix, Valentin Tombez), travaille par « projets » et s'inscrit dans une démarche d'enquête. « Nous effectuons un travail transversal et transmédia à l'intention des différents départements, parfois de notre propre initiative, parfois sur sollicitation de collègues de la radio ou de la télévision », explique Tybalt Felix[8].

Elle a par exemple réalisé une enquête sur les biens et transactions immobilières du Groupe Rolex à Genève, basée sur des données du registre foncier genevois[9] et ayant abouti à une carte interactive en trois dimensions de la ville de Genève [fig.4], ou sur le temps de parole des parlementaires à Berne, à partir des procès-verbaux de session, qu'il leur a fallu convertir en durée en fonction des débits de parole particuliers de chaque parlementaire[10].
Dans la plupart des exemples étudiés, les journalistes consacrent un espace au dévoilement de la « boîte noire », c'est-à-dire à expliquer quelles sources ont été utilisées, comment elles ont été analysées, etc. Cette démarche s'intègre dans une volonté de transparence plus grande, inspirée de la culture informatique.

Depuis plusieurs mois, une bonne partie du travail de la cellule se concentre sur le covid-19 et l'évolution de la pandémie. Un fil Twitter appelé @covid-explorer a notamment été créé. Il s'agit d'un « Bot interactif créateur de visualisations de données sur l'épidémie de

---

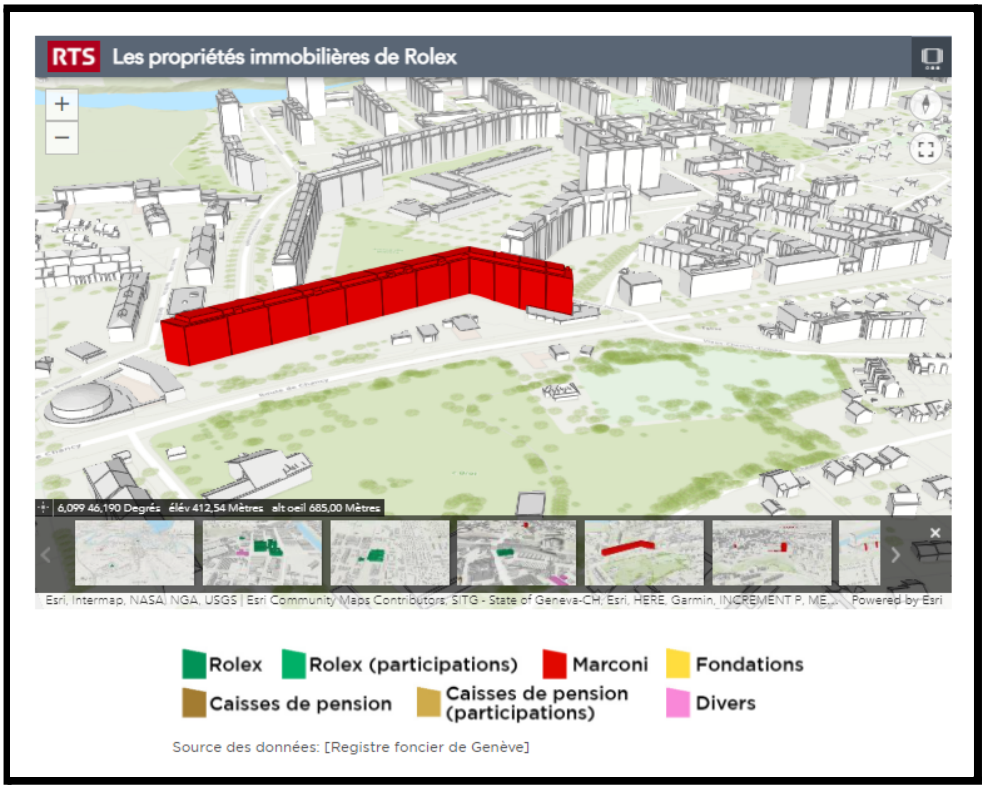
[8]-https://www.lesmediasfrancophones.org/nos-actualites/la-rts-met-en-place-une-cellule-de-journalisme-de-donnees

[9] « Cette enquête data basée sur l'analyse des centaines de transactions foncières a révélé que la très secrète Rolex était un acteur majeur de l'immobilier genevois. Une carte interactive 3D permet d'explorer les biens et de réaliser l'ampleur de ces investissements, qui frôle le milliard de francs. » (https://swisspressaward.ch/fr/user/c00029218/)

[10] Les journalistes expliquent leur méthode dans leur sujet : « Comme il n'existe aucun décompte officiel des temps de parole au Parlement, l'analyse a été réalisée à partir des procès-verbaux de chaque intervention au Conseil national. Afin d'obtenir un examen des débats et non des aspects protocolaires, les nombreuses prises de parole des présidents et vice-présidents de la Chambre n'ont pas été comptabilisées. Les départs et arrivées de parlementaires en cours de législature ont été pris en compte.
Les procès-verbaux nous ont permis d'additionner toutes les interventions par parlementaire, que nous avons converties en nombre de caractères. Ensuite, nous avons chronométré le débit de parole de chaque parlementaire sur une intervention de quelques minutes, la plus neutre possible. Le temps de parole a alors été calculé avec le débit de parole et la somme des interventions.
Cette méthode, conseillée par des linguistes, a été privilégiée par rapport aux durées des vidéos des interventions disponibles sur le site du Parlement, car le découpage de celles-ci est souvent très approximatif. »
(https://www.rts.ch/info/suisse/9523150-dixneuf-minutes-par-heure-le-temps-de-parole-des-femmes-au-parlement.html)

#Covid19 »[11]. Les visualisations sont, elles, créées avec l'outil opensource Datawrapper. Les journalistes de la cellule « data-innovation » maîtrisent donc le langage informatique et l'exploitent pour leurs projets d'enquête ou pour transmettre les résultats de votations via un algorithme, mais le travail de visualisation s'effectue, lui, avec des outils open source.

Figure 4: extrait de la carte interactive produite par la cellule data-innovation de la RTS sur les biens immobiliers de Rolex à Genève. (https://swisspressaward.ch/fr/user/c00029218/)



En parallèle, des recherchistes de la RTS pratiquent une autre forme de datajournalisme, à savoir sans programmation informatique : ils constituent des bases de données « à la main », en partant d'une hypothèse préétablie. Ils effectuent par exemple ce qu'ils appellent la « tournée des cantons » [E5] : ils contactent l'ensemble des cantons pour obtenir les chiffres cantonaux concernant certaines questions et, à partir des réponses, construisent une base de données ad hoc. Un même type de « sondage » a par exemple été effectué auprès de l'ensemble des communes romandes, afin de produire un sujet sur la longévité des syndics communaux suisses romands[12]. Les deux types de démarches (utilisation du code informatique ou constitution de bases de données 'à la main') se côtoient donc dans la

[11] https://twitter.com/covid_explorer

[12] Dans leur article, les auteurs précisent leur méthodologie« Pour établir ce classement, la RTS a contacté par courriel et par téléphone un total de 769 communes. Ont été prises en compte, toutes les communes des six cantons romands, de même que celles appartenant aux arrondissements administratifs du Jura bernois et de Bienne pour le canton de Berne. » https://www.rts.ch/info/suisse/8427270-vingtdeux-elus-communaux-romands-sont-au-pouvoir-depuis-plus-de-20-ans.html.

rédaction de la RTS. Dans les deux cas, les journalistes travaillent par « projet » et le processus d'enquête demande du temps et, donc, de l'argent.

La RTS fait partie de l'EIC (European Investigative Consortium), un consortium international de journalistes d'enquête. Ses datajournalistes sont ainsi régulièrement impliqués dans de grandes enquêtes internationales, comme les Football Leaks. Il leur arrive aussi régulièrement de recevoir d'importants sets de données de sources anonymes, qu'ils utilisent ou envisagent d'utiliser comme sources de sujets [E5]. A termes, ils souhaiteraient développer un outil permettant de croiser les données de l'ensemble de ces leaks.

Quelques journalistes chez Heidi.News/Le Temps

Heidi.news ne compte pas de journalistes ou de cellule qui se focalisent exclusivement sur l'analyse de données. Mais la rédaction compte quelques journalistes qui entretiennent une relation étroite avec l'univers des données, dont Sarah Sermondadaz, responsable du « flux science » de la rédaction, Kylian Marcos, journaliste 'nouveaux formats', Florent Hiard, journaliste spécialisé sur le thème du climat, ou encore Annick Chevillot, responsable du « flux santé ». En parallèle, *Heidi.news* travaille avec Lukas Stöcklin, un ingénieur médias responsable de la technologie. Actuellement, *Heidi.news* se trouve en pleine fusion avec le journal Le Temps, dont l'ancien datajournaliste a démissionné et qui recherche un remplaçant. Dès lors, les journalistes des deux rédactions ne peuvent se prononcer sur les pratiques qui auront cours ou non dans le futur.

A l'heure actuelle, chez *Heidi.news*, « tout le monde fait du datajournalisme, mais avec des outils grand public » [E6]. L'équipe utilise donc principalement des outils tels que Datawrapper. Souvent, les journalistes se tournent vers les mêmes bases de données pour produire de nouveaux sujets, par exemple sur le climat, réitérant des démarches qui pourraient probablement être automatisées.

Forme « embryonnaire » à Arcinfo

Si la plupart des médias supra-régionaux, comme la *NZZ*, *Le Temps*, le *Tages Anzeiger* ou la RTS comptent une unité data ou, du moins, un ou plusieurs datajournalistes, cela n'est pas le cas des médias locaux ou régionaux appartenant à de plus petits groupes de presse. C'est d'ailleurs la raison pour laquelle le groupe *NZZ* réfléchit actuellement à implémenter Q dans ses rédactions locales. Car l'intérêt existe aussi chez elles. La rédaction en chef du journal régional *Arcinfo*, propriété du groupe ESH Médias, affiche par exemple un fort désir d'utiliser le datajournalisme. Plusieurs journalistes ont d'ailleurs suivi une formation en la matière en 2017, mais celle-ci n'a pas abouti à la production concrète de sujets de datajournalisme par la suite, faute de temps à disposition. Pour y parvenir, le titre collabore depuis peu avec un recherchiste externe, mandaté pour créer des bases de données ad hoc, « solides et inattaquables » [E4] sur certains sujets. Son engagement a été décidé suite au constat que plusieurs sujets n'avaient pas été traités par la rédaction, faute de temps et de ressources suffisantes, mais aussi de compétences nécessaires. Le projet consiste à mettre en place une collaboration entre le recherchiste et les journalistes de la cellule-enquête du titre, les seconds s'appuyant sur le travail du premier comme base pour leurs enquêtes.

En dehors de cette collaboration toute récente, la pratique du datajournalisme reste toutefois « embryonnaire » au sein de la rédaction :

> « Il y a des moments clés où on a des données, par exemple quand les primes d'assurance maladie sont publiées, ce genre de chose. Dans ces cas, on a besoin de faire des extractions ou alors de faire des visualisations, mais on le fait un peu 'à l'ancienne' : on met des chiffres dans un tableau, Excel ou Google doc, et on en sort des choses » [E4]

Certains journalistes de la rédaction utilisent occasionnellement des données chiffrées dans leurs articles. Toutefois, il s'agit souvent de données « pré-traitées », produites par exemple par l'Office fédéral de la statistique (OFS), et utilisées sans travail d'analyse complémentaire. L'utilisation de données comme point de départ de sujets journalistiques est pratiquement inexistante dans le journal.

> « On a réalisé que, par rapport aux forces à disposition, par rapport au temps qu'on peut accorder à un sujet, ces visualisations allaient surtout être utiles à montrer, à démontrer quelque chose, à démontrer un propos, mais elles n'allaient pas forcément être exploitées comme une sorte de terreau duquel on peut faire sortir un sujet » [E3]

Le plus souvent, lorsque des données sont utilisées, elles proviennent donc d'entités externes, essentiellement étatiques, et visent à créer des visualisations pour compléter des articles déjà conçus. Les sets de données sont donc déjà « tout faits » [E4]. Concrètement, les journalistes utilisent alors des logiciels de visualisation, comme Infogram, Datawrapper ou Tableau pour produire des visualisations. L'infographiste de la rédaction est quant à lui régulièrement mandaté pour produire des infographies ou statiques interactives avec l'outil Genial.ly.

Dans certaines situations, le journal a cherché à produire des contenus plus originaux et complexes. Par exemple, lors des dernières élections cantonales en 2017, il a développé une carte interactive sur laquelle l'internaute pouvait déplacer son curseur et découvrir quel(le)s candidat(e)s avaient été élu(e)s dans chaque commune. Pour la réaliser, la rédaction avait dû collaborer avec l'équipe technique du groupe ESH Medias. Mais ce genre de projet reste très rare, car il demande d'importantes ressources.

> « C'était un travail absolument dantesque de faire cette carte, de la faire construire par l'équipe technique. Il lui a fallu en tout cas dix jours pour la mettre en place, et il fallait qu'on leur ait livré toutes les données avant. (…). Souvent, c'est des projets qui demandent énormément de travail au service technique, beaucoup de travail en amont pour réunir les données, et aussi beaucoup de travail pendant la journée d'élection pour tout actualiser » [E3].

Les entretiens montrent donc que des projets a priori assez simples peuvent, pour une rédaction comme celle d'Arcinfo, rapidement prendre des proportions importantes. C'est pourquoi, pour l'instant, l'approche data-driven, consistant à partir d'une base de données pour traiter un sujet, n'est pratiquement jamais utilisée.

> « Ça, on n'a pas les forces. En tout cas, on ne prend pas le temps d'aller dans des bases de données et de se dire : bon, on va essayer de trouver un sujet avec ces chiffres… ça, on ne le fait pas. C'est toujours la même chose : au niveau des effectifs, on préfère avoir quelqu'un qui peut produire de la matière quotidienne, pour l'instant en tout cas » [E3.2]

Ainsi, dans la rédaction d'Arcinfo, la pratique du datajournalisme consiste presque exclusivement à produire des visualisations à partir de sets de données existants, dans le but de compléter des articles. L'approche est donc essentiellement *story-driven*. L'engagement d'un recherchiste externe pour récolter des ensembles de données démontre toutefois une volonté d'utiliser plus régulièrement l'approche data-driven. Actuellement, cette collaboration est en phase de test et sera renouvelée en fonction des résultats obtenus.

Médias et covid

Avant la crise du Covid-19, la rédaction d'Arcinfo n'avait donc que peu l'habitude de récolter et analyser des données. La pandémie l'a dès lors confrontée à toute une série d'obstacles auxquels elle n'était pas habituée. Les principaux problèmes provenaient alors de la qualité des données.

« On s'est rendu compte que c'était moins un problème de visualisation ou de traitement de données que de pertinence des données. Parce qu'il faut aller les chercher, il y a plusieurs sources pour le même indicateur, mais elles ne sont pas toujours d'accord entre elles, avec un décalage dans la mise à jour, avec des méthodologies qu'on ne connaît pas toujours. Enfin, c'est vraiment compliqué » [E5]. Les mêmes constats ont été faits dans d'autres médias, avec des différences entre différents pays ou cantons qui rendaient les données difficilement comparables entre elles [E1]. « Les données du Covid ont montré que tout n'était pas comparable, que certains 'trous' s'expliquaient surtout par des changements dans la méthodologie, etc. » [E4]

A certains moments, même les chiffres officiels de l'OFSP se sont avérés décalés par rapport à la réalité, comme l'a expliqué Marc Renfer sur Info Verso le 4 avril 2020 :

> « D'abord on a tous tournés notre tête vers l'OFSP qui est quand même l'Office fédéral de la santé publique, et donc l'organe le plus officiel et national à ce sujet. Et puis, peu à peu, les cantons ont commencé à publier chacun de leur côté leurs propres chiffres, et là on a commencé à se rendre compte que les chiffres cantonaux, si on les agrégeait, si on les prenait tous, et bien ils étaient plus [+] à jour que les chiffres de l'OFPS, qui a eu un peu de peine à traiter toute cette information aussi rapidement. Donc on a vu des sites se mettre en place qui suivaient les chiffres cantonaux et donc pendant plusieurs jours, il y a eu un assez grand décalage entre ces chiffres-là, cantonaux, et les chiffres de l'OFSP. Il y a eu

> je pense jusqu'à 50% de différence sur les chiffres des décès, certains jours. »[13]

Le site du chercheur bernois Daniel Probst a donc servi de base de données de référence pour plusieurs médias pendant quelques jours, car mis à jour plus fréquemment et plus rapidement que celle de l'OFSP.

Dans la plupart des titres, comme chez Arcinfo, *Heidi.news* ou la RTS, la récolte des données s'est donc d'abord faites « à la main ». Les données étaient quotidiennement récoltées et entrées dans une base de données par un/e journaliste. A l'instar de celle de la RTS, la rédaction de *Heidi.news* a désormais automatisé la récolte. Mais elle se retrouve confrontée à la même difficulté concernant le nombre de vaccinations par canton et espère à nouveau pouvoir automatiser la récolte [E6].

13 https://www.rts.ch/play/radio/six-heures-neuf-heures-le-samedi/audio/info-verso-quelles-donnees-pour-quels-graphiques?id=11183028

Freins à la pratique du datajournalisme dans les rédactions de Suisse romande

Les entretiens menés avec des journalistes de différentes rédactions montrent, de manière générale, que les rédactions sont confrontées aux mêmes « obstacles » que ceux relevés dans la littérature scientifique : absence, indisponibilité, inexploitabilité, impertinence et/ou incompréhension des données, manque de compétences à l'interne.

Manque de temps et de ressources

Au niveau des médias locaux, comme l'a souligné Rahel Estermann [E1], les journalistes manquent déjà souvent de ressources pour traiter les sujets quotidiens. Ils n'ont dès lors pas de temps à « perdre » à rechercher des sujets dans des bases de données, d'autant moins lorsqu'ils n'ont pas la garantie d'y trouver des pistes exploitables. Le simple fait de devoir utiliser un nouvel outil comme Q, pourtant conçu dans un but de simplification du travail, peut être perçu comme une contrainte supplémentaire par les professionnels. « Les journalistes locaux sont déjà tellement chargés qu'ils pourraient exploser si on leur demandait de rajouter des visualisations à tout ce qu'on leur demande déjà de faire » [E1]. Comme l'indique cette journaliste :

« Si tu dois te plonger dans les chiffres, rien que pour comprendre où aller chercher les données, pour les regarder, pour en tirer des conclusions, faire réagir des gens, etc. on n'a pas les forces de production (…). C'est un travail assez lourd. A part de manière sporadique, on n'a simplement pas les forces quoi » [E3.1]

Rahel Esterman note qu'au fond, peu de journalistes constituent eux-mêmes leurs bases de données, que cela soit 'à la main' ou via du web scraping (extraction web). « Cela demande beaucoup de temps, non seulement pour les extraire, mais aussi pour les nettoyer, les organiser, etc. Donc souvent, les visualisations sont basées sur des données étatiques » [E1]. La chercheuse observe que seules des unités spécialisées, comme la Data Team de la SRF, ont le temps de collecter leurs propres données.

(In)disponibilité des données

Un autre frein essentiel est l'indisponibilité des données. En effet, il arrive souvent que des journalistes aient envie de traiter un sujet via des données mais que celles-ci soient simplement inexistantes, non accessibles ou seulement partiellement accessibles.

Plusieurs exemples ont été donnés durant les entretiens. Par exemple, un journaliste désirait depuis plus d'un an prendre le temps de comparer les tarifs pratiqués par les différents vétérinaires d'un canton. Or, aucune base de données existantes ne le permet. Il aurait alors été nécessaire de constituer une base de données manuellement, ce que la rédaction n'a pas eu le temps de faire jusqu'à présent [E4]. En revanche, elle s'est penchée sur les liens d'intérêt des élus du canton. Pour ce sujet, elle a mandaté un recherchiste qui s'occupe de compiler les informations, en multipliant les sources et les types de sources (documents, appels téléphoniques, etc.) afin de constituer une base de données exploitable par les journalistes. [E4]. Une mission de plusieurs semaines que seuls de grands médias comme la RTS ont le temps de réaliser.

Dans d'autres cas, les données n'existent simplement pas. Désireuse de connaître les chiffres concernant le harcèlement de rue, une journaliste s'est retrouvée confrontée à l'absence de statistiques en la matière, le harcèlement de rue étant comptabilisé dans une catégorie d'infraction plus large [E3.2]. Le phénomène du harcèlement de rue en tant que tel reste alors impossible à estimer.

Inaccessibilité des données

Dans d'autres situations, les données recherchées existent quelque part – ou du moins les journalistes pensent qu'elles doivent exister – mais elles se révèlent inaccessibles. Un interviewé indique notamment qu'il est impossible de connaître la taille (nombre d'employés) des entreprises implantées dans le canton, alors que ce genre d'information est disponible en France. « Il existe bien un registre des entreprises, mais pas de registre par taille d'entreprise. Tout ça, ça reste confidentiel. Il y a un manque de données criant de ce côté-là (…). En Suisse, tu sens bien qu'il y a une culture du secret et un libéralisme qui fait qu'on sait très peu de choses sur les entreprises et qu'on doit se fier aux chiffres 'macro' de la Confédération, qu'on doit croire sur parole (…). Mais pour les chiffres sur les entreprises, les chiffres d'affaire ou ce genre de chose, il y a un manque » [E4]

Enfin, il arrive également que des données existent, mais que les journalistes ne soient pas au courant de leur existence ou ne parviennent pas à les trouver. Cela a été le cas sur le thème du Chlorothalonil, un fongicide récemment interdit en Suisse et en Europe. Après la publication de quelques articles à ce sujet dans d'autres médias, un journal local a voulu connaître la situation dans son canton. Mais il n'a pas trouvé les données.

« On avait vu qu'il y avait pas mal de régions, d'autres cantons qui s'en plaignaient mais on n'avait pas les résultats pour le nôtre. C'était la fin de l'été et on a laissé un peu filé (…). Et derrière, tu as un sujet de la RTS qui dit que dans notre canton, le taux de Chlorothalonil qui se déverse dans le lac est délirant. On ne l'a pas vu venir, alors que les données devaient être quelque part ! » [E4.].

Le premier obstacle auquel sont confrontés les journalistes est donc l'absence, l'indisponibilité ou l'inaccessibilité des données concernant les sujets qu'ils souhaitent traiter.

Inexploitabilité des données

Dans d'autres situations, les journalistes savent que les données existent et comment les obtenir. Mais cela ne résout pas tous les problèmes. En effet, il arrive souvent que les formats mêmes des données posent problème, empêchant l'analyse, compliquant l'interprétation, rendant les comparaisons difficiles ou la mise à jour automatique impossible. Dans certains cas, le problème est que les différentes données nécessaires à l'analyse se trouvent sur des fichiers hébergés à différents endroits, dans des formats divers et variés (ex. : texte, Excel, Powerpoint, pdf, jpeg) et qu'il est donc presque impossible d'automatiser leur analyse ou de croiser les documents.

« Il n'y en a pas beaucoup de bases de données qu'il est possible de mettre à jour automatiquement (…). Par exemple sur le Chlorothalonil, il y a peut-être dix stations d'épuration qui ont un fichier, toujours le même, toujours au même format, toujours au même endroit, mais il suffit que la quinzième soit juste un pauvre PDF dans un communiqué (…) et tu es alors obligé d'aller le chercher à la main régulièrement » [E3].

Sur le sujet des mandats des élus, un journaliste déclare qu'« il y a tellement de sources différentes : c'est de la source PDF, comme c'est parfois sur des sites internet, comme c'est parfois des coups de fil, du déclaratif auprès de chargés de communication… » [E4]. Il se peut aussi que les nomenclatures utilisées par les statisticiens (par exemple sur les importations et exportations) soient difficilement compréhensibles. « Du coup, tu n'arrives pas à extraire clairement les produits, tu ne sais pas de quoi on parle » [E4]. L'opportunité de croiser plusieurs sets de données nécessite aussi que la structure des différentes bases de données soient similaires [E2]. En somme, pour être exploitées, les données doivent répondre à de nombreux facteurs qui dépassent leur seule qualité.

Parfois, elles s'avèrent simplement incomparables entre elles. La crise du Covid l'a montré : les tentatives de plusieurs médias de réaliser des visualisations à partir de données produites par plusieurs institutions différentes a rendu très ardue leur exploitation ; chaque pays, chaque canton, voire chaque hôpital, a utilisé des méthodologie différentes, comptabilisé certains facteurs mais pas d'autres, compté les « cas » selon des critères différents, mis à jour leurs données selon des temporalités différentes, ce qui empêchait toute comparaison réellement correcte.

Les journalistes citent d'autres exemples comparables où les méthodes de comptabilisation de certains faits diffèrent d'une institution à l'autre. Concernant la criminalité, deux villes romandes comptabilisaient par exemple différemment les occurrences, l'une selon les « cas », l'autre selon les « infractions ». Or, un cas peut être composé de plusieurs infractions, si bien qu'une des deux villes semblait injustement plus « criminogène » que l'autre. Dans d'autres situations encore, ce sont les formats de présentation des données qui rendent leur exploitation plus compliquée. Il arrive en effet que des données soient présentées sous forme de texte continu ou sous forme de PDF et que leur exploitation nécessite donc un transfert systématique d'un format X vers le format Excel. Une tâche chronophage, rébarbative et durant laquelle surviennent régulièrement des erreurs.

En somme, la qualité des données a d'importantes implications sur la (non-)pratique du datajournalisme. Des difficultés dans l'obtention, la préparation ou le nettoyage des données constituent des facteurs limitant la pratique du datajournalisme.

Incompréhension des données

Même lorsque les données sont accessibles, structurées et disponibles dans un fichier Excel, les journalistes peuvent rencontrer des difficultés pour savoir quelle conclusion en tirer. Dans certaines situations, il arrive en effet qu'une « augmentation » d'une variable X s'explique pour d'autres raisons qu'une réelle augmentation de ladite variable. Par exemple, même si des données indiquent une augmentation des cas de « harcèlement », cette augmentation peut s'expliquer par le fait que les cas sont simplement plus souvent dénoncés aujourd'hui. Idem pour les « cas de covid » : les présenter sans les mettre en relation avec le nombre de tests effectués représente peu d'intérêt.

Dans de nombreuses situations, les journalistes peinent à savoir comment interpréter les données dont ils disposent. L'analyse passe donc par la consultation de spécialistes et il est possible que le sujet perde soudainement de l'intérêt. Le travail d'interprétation peut aussi exiger un important travail de réflexion, de consultation d'experts, de vérification, pour lequel

les journalistes manquent parfois de temps. Ils ont bien conscience du fait que, même lorsque des données semblent indiquer une certaine tendance, celle-ci peut être faussée par la méthodologie de récolte des données. « Une augmentation d'une valeur peut être liée au simple fait d'une augmentation des dénonciations par exemple » [E3.2]. Durant une recherche menée par des journalistes, beaucoup de précautions apparaissent durant les discussions quant aux conclusions qu'il est possible de tirer à partir de données. Des journalistes ont par exemple obtenu des données inédites concernant les retraits de permis de voiture. Nous retranscrivons ici leur discussion concernant l'usage à en faire:

> J1 : J'ai obtenu des données sur les retraits de permis de conduire de chaque commune. Et on voit que, par exemple à Genève, il y a beaucoup plus de retraits de moto qu'ailleurs… Ou encore qu'il y a beaucoup plus de retraits dans les communes rurales qu'en ville…
> J2 : Est-ce qu'il y a les raisons de ces retraits ?
> J1 : Oui, il y a les raisons. Il y en a trois principales : il y a l'alcool, les excès de vitesse et le comportement au volant. On voit que le canton de Berne envoie beaucoup plus de gens en « examen psychologique » que les autres communes. (…).
> J3: Je me demande par contre si cela ne serait pas mieux de mesurer les données selon le nombre de propriétaires de voiture par commune plutôt que d'habitants par commune. Cela fait plus de sens. Si on mesure le taux de retraits de permis « par tête », c'est un peu difficile d'en tirer des conclusions parce qu'il y a beaucoup de gens en ville qui n'ont pas de permis.
> J4 : Et c'est par 1'000, par 100'000 habitants ? Parce que selon, les petites communes risquent d'être vraiment surreprésentées.

Ce dialogue montre bien que « les chiffres ne parlent pas d'eux-mêmes » et que les journalistes doivent réfléchir à la manière dont ils les traitent et les interprètent, afin d'éviter les conclusions erronées. Il en va pour eux de la crédibilité de leur travail. Les « classements » par cantons ou par communes sont des sujets très fréquents dans les médias romands. Ils se fondent souvent sur des statistiques déjà disponibles, voire récoltées par les journalistes. Il apparaît toutefois régulièrement que les facteurs explicatifs d'un classement sont plus complexes que ne le laissent supposer les chiffres[14].

Absence de « culture de la donnée »

Enfin, le dernier obstacle découle des précédents. En effet, les entretiens montrent que les journalistes ont conscience des difficultés que représentent l'obtention, l'analyse et l'exploitation des données, ainsi que des risques inhérents de commettre des erreurs d'interprétation. Dès lors, hormis pour les journalistes spécialisés, beaucoup de

[14] Un exemple est celui d'un article du Matin Dimanche de février 2018, intitulé « Le délai d'attente moyen pour une place en crèche est de sept mois ». Le chapeau de l'article mentionne d'importantes différences intercantonales : « Selon nos calculs, Genève est le canton où les délais sont les plus longs contrairement à Neuchâtel ». Mais l'apparente inégalité se voit rapidement rectifiée : « Ce n'est pas parce qu'une commune est plutôt basse dans notre recensement que les structures sont forcément plus adaptées qu'ailleurs. ». L'exemple est révélateur d'une « réalité » souvent plus complexe que ce que laissent supposer les chiffres obtenus.

professionnels préfèrent éviter le datajournalisme en raison d'un sentiment de manque de compétences en la matière. « Il y a très peu de gens qui sont à l'aise avec les chiffres ou à l'aise avec les outils d'analyse. Donc il y a cet appétit qui manque » [E1]. Toute volonté de favoriser la pratique du datajournalisme dans une rédaction implique donc, d'abord, de réussir à convaincre les journalistes de l'intérêt de l'approche.

> « La plupart des journalistes ne comprennent pas le potentiel du datajournalisme. Ils sont tous contents d'avoir une carte interactive sur laquelle tu peux cliquer et voir un chiffre pour chaque commune mais pour moi, ce n'est pas du datajournalisme mais de l'infographie, qui existait déjà dans les journaux papier. Les grands projets de datajournalisme nécessitent peut-être des sacrifices de la part de la rédaction, c'est-à-dire peut-être licencier deux personnes » [E2].

Dès lors, les potentiels, tant journalistique que technique et économique du datajournalisme doivent, au préalable, être expliqués aux journalistes. En somme, il est nécessaire de d'abord faire mûrir de tels projets dans leur esprit pour les faire accepter et de les former à l'emploi des outils – comme l'ont fait Tamedia et la *NZZ* – pour envisager une implémentation réussie. En somme, tout outil de datajournalisme doit converger avec la culture journalistique. Et inversement.

Dans la plupart des projets de développement d'outils de datajournalisme, la mise en relation entre les développeurs de l'outil et les journalistes est passée par une médiation, des efforts de formation, d'explication, etc. Ces efforts semblent fondamentaux pour faire accepter tout nouvel outil appelé à modifier les routines professionnelles des journalistes.

Conclusion

Les entretiens ont montré que la pratique du datajournalisme se divise en Suisse clairement entre deux approches différentes : d'un côté, l'approche data-driven est essentiellement pratiquée par des journalistes spécialisés, qui partagent une sous-culture commune et travaillent presque systématiquement dans des cellules dédiées. Ils travaillent en général par projet et de manière collaborative. Il s'agit d'une forme de datajournalisme valorisée, car apparentée au journalisme d'enquête, mais qui, dès lors, reste exceptionnelle. D'un autre côté s'observe une forme plus « quotidienne » de datajournalisme, pratiquée dans toutes les rédactions, même locales, plus proche de l'approche story-driven, où les données sont utilisées pour illustrer des sujets en cours. Il arrive en effet souvent que des journalistes utilisent des sets de données, notamment produits par l'Office fédéral de la statistique, pour illustrer ou appuyer leur propos. Il s'agit donc essentiellement d'un travail de visualisation, pour lequel la plupart des rédactions utilisent des outils courants tels que Datawrapper, Infogram, Genial.ly ou Tableau.

En résumant, les journalistes non-spécialisés ont tendance à pratiquer du « general data journalism » et les journalistes spécialisés du « investigative data journalism », selon la distinction proposée par Uskali and Kuutti (2015, 85).

Plus encore, mêmes les approches data-driven peuvent être distinguées en sous-catégorie. En effet, même les journalistes spécialisés ne pratiquent pas tous le data journalism de la même manière. Certains – en l'occurrence, dans les cas observés, souvent des « recherchistes » - utilisent des méthodes de récolte de données « à l'ancienne ». Partant d'une hypothèse ou question de départ (par ex. : « Quels sont les liens d'intérêts des élus du canton », ou « Combien de temps les élus locaux restent-ils en poste ? »), les journalistes récoltent les informations et construisent les bases de données « à la main ». Dans ces cas, il est nécessaire de consulter plusieurs sources différentes et de rassembler des données aux formats parfois très disparates. D'autres journalistes, plus proches de l'univers des informaticiens, utilisent des algorithmes pour « extraire » des données à partir de sites web. Certains, enfin, mettent l'accent sur la récolte de bases de données existantes mais encore inexploitées ou inédites. L'automatisation de la récolte et de la mise à jour des données constituent un idéal commun à l'ensemble des journalistes interviewés, mais paraît pour beaucoup inatteignable, à l'inverse du data journalism manuel pour lequel le seul obstacle envisagé est le manque de ressources.

Les entretiens montrent également que la plupart des journalistes ont conscience des biais liés à l'exploitation et à l'interprétation de données. Les journalistes connaissent les promesses associées au datajournalisme, mais leur expérience les a conduits à les relativiser, soit parce qu'il leur est arrivé d'aboutir à des conclusions erronées, soit parce qu'ils ont été confrontés à des difficultés d'interprétation qui ont nécessité un important travail supplémentaire, rendant caduque la promesse d'une représentation factuelle de la réalité. Parfois, l'obtention même de données s'est avérée problématique. Le manque de ressources, de compétences et de temps constitue également des freins régulièrement mentionnés. Un rédacteur en chef voit même « le manque d'intérêt » des journalistes et le « manque de compétences en la matière » dans la rédaction comme les freins les plus essentiels à la pratique du datajournalisme.

Les quelques exemples existants de tentatives d'implémentation d'outils datajournalistiques dans des rédactions révèlent que le développement de l'outil ne constitue qu'une étape préalable. Dans chaque exemple, les développeurs ont dû consacrer un temps important à la sensibilisation et à la formation des professionnels. Une étape nécessaire à laquelle le projet MediaLaboratory ne pourra échapper.

Références

Aitamurto, Tanja, Esa Sirkkunen, et Pauliina Lehntonen. 2011. *Trends In Data Journalism*. Helsinki: Next Media Finaland.

Bradshaw, Paul. 2018. « Zeroes and Ones: Investigating with Data ». P. 19-29 in *Digital Investigative Journalism*, édité par O. Hahn et F. Stalph. Cham: Springer International Publishing.

Dierickx, Laurence. 2019. « Information automatisée et nouveaux acteurs des processus journalistiques ». *Sur le journalisme, About journalism, Sobre jornalismo* 8(2):154-67. doi: 10.25200/SLJ.v8.n2.2019.408.

Hahn, Oliver, et Florian Stalph, éd. 2018. *Digital Investigative Journalism: Data, Visual Analytics and Innovative Methodologies in International Reporting*. 1st ed. 2018. Cham: Springer International Publishing : Imprint: Palgrave Macmillan.

Keller, Léna. 2019. « La RTS met en place une cellule de « journalisme de données » ». *Les Médias Francophones Publics*. Consulté 16 mars 2021 (https://www.lesmediasfrancophones.org/nos-actualites/la-rts-met-en-place-une-cellule-de-journalisme-de-donnees).

Kennedy, Helen, Rosemary Lucy Hill, Giorgia Aiello, et William Allen. 2016. « The work that visualisation conventions do ». *Information, Communication & Society* 19(6):715-35. doi: 10.1080/1369118X.2016.1153126.

Knight, Megan. 2015. « Data Journalism in the UK: A Preliminary Analysis of Form and Content ». *Journal of Media Practice* 16(1):55-72. doi: 10.1080/14682753.2015.1015801.

Lewis, Seth C., et Oscar Westlund. 2015. « Big Data and Journalism: Epistemology, Expertise, Economics, and Ethics ». *Digital Journalism* 3(3):447-66. doi: 10.1080/21670811.2014.976418.

Parasie, Sylvain. 2015. « Data-Driven Revelation?: Epistemological Tensions in Investigative Journalism in the Age of "Big Data" ». *Digital Journalism* 3(3):364-80. doi: 10.1080/21670811.2014.976408.

Plattner, Titus, et Didier Orel. 2019. « Addressing microaudiences at scale: How Tamedia generated about 40,000 articles in five minutes to report on Swiss popular vote results at the municipalities level ». P. 1-2 in *communication présentée à Computation+ Journalism Conference, Miami University, Floride*.

Rinsdorf, Lars, et Raol Boers. 2016. « The need to reflect: Data journalism as an aspect of disrupted practice in digital journalism and in journalism education ». in *Promoting understanding of statistics about society. Proceedings of the roundtable conference of the International Association of Statistics Education (IASE). Berlin: ISI/IASE*.

Stalph, Florian. 2018. « Classifying Data Journalism: A Content Analysis of Daily Data-Driven Stories ». *Journalism Practice* 12(10):1332-50. doi: 10.1080/17512786.2017.1386583.

Stalph, Florian. 2020. « Evolving Data Teams: Tensions between Organisational Structure and Professional Subculture ». *Big Data & Society* 7(1):205395172091996. doi: 10.1177/2053951720919964.

Tandoc, Edson C., et Soo-Kwang Oh. 2017. « Small Departures, Big Continuities?: Norms, Values, and Routines in *The Guardian* 's Big Data Journalism ». *Journalism Studies* 18(8):997-1015. doi: 10.1080/1461670X.2015.1104260.

Valeeva, Anastasia. 2017. *Open Data in Closed political system: Open data investigative journalism in Russia*. Oxford: Reuters Institute for the Study of Journalism.

Weber, Wibke, Martin Engebretsen, et Helen Kennedy. 2018. « Data stories: Rethinking journalistic storytelling in the context of data journalism ». *Studies in Communication Sciences* 2018(1):191-206.