Media Laboratory (IMI)

Annexe 3

Démocratiser et favoriser la pratique du datajournalisme (II): recommandations en vue d'un développement d'outils techniques

Lena Würgler
Nathalie Pignard Cheynel
Andrew Robotham

Académie du journalisme et des médias

Neuchâtel, mars 2021

Résumé

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 favorisant la pratique du datajournalisme dans les rédactions.

Ce document se base sur les résultats du document *Démocratiser et favoriser la pratique du datajournalisme (I): un état des lieux*. Ces recommandations ont été élaborées par l'Académie du journalisme et des médias (UNINE) en collaboration étroite avec le Distributed Information Systems Laboratory (LSIR) de l'EPFL dans le but d'établir une feuille de route pour les développements d'outils et de technologies et en vue de l'affectation des ressources y relatives.

Sur la base de l'état des lieux, nous émettons les recommandations suivantes :

- La construction d'une infrastructure permettant, avec un minimum de développement, d'offrir des tableaux de bord thématiques à destination des journalistes non-spécialistes des données;
- La mise à disposition en libre accès de la technologie accompagnée d'une feuille de route permettant à tout média de créer facilement des tableaux de bord sur d'autres thématiques;
- La réalisation d'un projet pilote en partenariat avec le partenaire média (*Heidi.news*) via un premier tableau de bord thématique consacré à l'environnement et au climat;
- L'intégration d'une dimension didactique favorisant la prise en main de l'outil par des journalistes non spécialistes des données;
- La création d'un outil de dataification des déclarations publiques (permettant une approche data sur des enjeux d'actualité et des controverses).

Recommandations

Sur la base de la revue de la littérature et de l'état des lieux du datajournalisme en Suisse (recherche existante et entretiens exploratoires) présentés dans le document *Démocratiser et favoriser la pratique du datajournalisme (I): un état des lieux*, les chercheurs du projet MediaLaboratory ont formulé les recommandations suivantes à destination des développeurs des futurs outils de data journalism.

Pistes rejetées

Parmi les premières pistes envisagées pour le développement technique figurait l'idée de développer des outils de datavisualisation "génériques" destinés aux journalistes non-spécialistes du journalisme de données. Nous recommandons d'écarter cette piste. Pour plusieurs raisons.

Premièrement, il existe déjà de nombreux outils similaires, offrant une gamme très large de fonctionnalités et auxquelles la plupart des rédactions souscrivent déjà (Datawrapper, Tableau, Genial.ly, etc.). Le plus emblématique – et largement utilisé dans les rédactions – est Datawrapper. Mais ces outils ne sont utilisés que par un petit nombre de personnes dans les rédactions (les plus spécialisées). Ils sont rarement pris en main par les autres journalistes par manque de temps (et lorsque c'est le cas, cela nécessite un temps de prise en main à chaque fois).

Deuxièmement, il convient de noter que dans la mesure où chaque éditeur ou rédaction possède sa propre charte graphique, les graphiques « prêts-à-l'emploi » proposés par des institutions officielles sont généralement retravaillés pour correspondre aux notamment aux besoins de marque du titre. Si ce constat vaut tout particulièrement pour les médias d'information de moyenne et de grande taille. Pour celles-ci, il ne fait donc pas de sens de simplement proposer d'autres outils de visualisation produisant des graphique « prêts à l'emploi », d'autant plus que cela reviendrait à reproduire un type de service déjà largement rempli par les institutions étatiques. Enfin, des visualisations accessibles et utilisables par tous les titres fait disparaître toute dimension d'exclusivité chère aux médias et perdent ainsi de leur intérêt aux yeux des différentes rédactions. En somme, il apparaît très difficile de produire des visualisations « universelles » qui puissent répondre aux attentes de divers journalistes ou rédactions. En outre, l'exemple de Q à la *NZZ* a démontré que tout outil de visualisation, pour être effectivement utilisé, se devait de rester extrêmement simple et fonctionnel, tout en étant très intégré aux outils et interfaces du média. Ce constat oriente notre recommandation vers le développement d'une architecture générique et un système automatisant la collecte des données, mais pouvant être repris et adaptés aux besoins spécifiques des médias, notamment pour intégrer des éléments d'identité visuelle (voir le point "dashboard" ci-dessous).

Il n'est pas pertinent non plus de développer des outils de scraping ou de nettoyage de données puisque, là encore, il en existe déjà. De plus, leur utilisation implique que les journalistes ont déjà un set de données ou savent déjà où se trouvent les informations recherchées, ce qui n'est souvent pas le cas. Ce type d'outil semblent plus utiles à des journalistes spécialisés dans le datajournalisme, et qui maîtrisent déjà un minimum le codage informatique. Les journalistes non spécialisés semblent donc avoir besoin de sets de données pertinents, nettoyés – donc « fiables » – mais exploitables selon leurs besoins, leurs idées.

Ces différents constats ont amené les membres du projet MediaLaboratory à réfléchir à des outils qui répondent aux besoins suivants:

- Ils s'adressent à des journalistes n'étant pas des spécialistes du journalisme de données (il convient peut-être de rappeler que ce critères figuraient dans parmi les objectifs originaux du projet);
- Ils sont faciles d'accès et ne requiert pas une longue période de prise en main

- Ils effectuent un premier travail de sélection et de « prémâchage » des données afin permettre aux professionnels de surmonter les difficultés les plus courantes rencontrées lors de la récolte ou le traitement de données;
- Ils s'intègrent facilement au système technique des rédactions et leur permettent d'intégrer leur identité visuelle, laquelle garantit une certaine originalité et exclusivité du contenu proposé.

Pistes envisagées

En prenant en compte ces différentes observations, plusieurs pistes de développement du projet ont alors été discutées :

1. Création d'un « dashboard » universel sur le climat

Il a été envisagé de produire une forme de « dashboard » sur un thème particulier, sur lequel les journalistes pourraient trouver les données structurées et nettoyées les plus pertinentes concernant le thème, au format « brut », et avec la possibilité de les « prévisualiser » avec une option dédiée. Ce scénario implique, d'abord, de garantir que les données mises à disposition soient effectivement structurées, nettoyées et fiables. Il implique aussi d'effectuer un choix pertinent et justifié des données et des sources utilisées. Soit un important travail de « tri » et de vérification des données au départ. De plus, il est important de tenir compte du fait que la valeur informationnelle des données peut varier selon les professions (développeurs, journalistes, graphiste). Il serait donc fondamental que les besoins des journalistes soient inclus dans le choix des données et des visualisations. Enfin, cette piste de projet implique aussi que les données soient mises à jour régulièrement et que leur dernière mise à jour soit indiquée sur le dashboard. Il faut donc que la mise à jour soit automatisée et simultanée pour l'ensemble des données utilisées, soulevant par là un défi technique important.

De plus, il est nécessaire de s'assurer de la pérennité des données. D'un côté, elles doivent être accessibles sur le long terme. Cela signifie qu'il faut impérativement que le site où se trouve les données ne soit pas modifié. D'un autre côté, il faut que ces données soient régulièrement mises à jour par l'institution qui les produit. Il est par exemple plus probable que les données d'une institution étatique soient mises à jour sur le long terme que celles produites par un projet de recherche universitaire conduit sur un nombre restreint d'années.
Ce type d'outil a l'avantage de laisser une certaine flexibilité aux journalistes : ils peuvent utiliser soit les données « brutes », soit les (pré-) visualisations proposées. La présentation des données sous forme de Dashboard leur offre aussi plus de choix en termes de sujet puisqu'ils peuvent se concentrer sur l'un ou l'autre des volets, selon le sujet traité. Cet outil officierait alors comme une sorte de plateforme de service public sur un thème donné.

Le thème envisagé est le « climat », a priori au niveau suisse et cantonal, mais des comparaisons internationales peuvent faire sens pour certains sujets. Si ce type d'outil peut sembler "pré-mâcher" le travail des journalistes, il offre l'avantage de leur fournir des datavisualisations "clés en main" ayant nécessité en amont une grande maîtrise statistique de la part des développeurs qui vont les structurer, les nettoyer et, en partie, les (pré-)interpréter.

2. Création d'un moteur de recherche de « déclarations publiques »

La deuxième piste envisagée, apparue dans une discussion avec un datajournaliste d'une rédaction romande, est celle d'un « moteur de recherche » de citations de personnalités publiques. Il s'agirait alors plutôt d'une base de données sur laquelle les journalistes pourraient aller retrouver des propos tenus par des politiciens dans les médias ou durant des séances parlementaires. L'avantage de ce scénario est que la mise à jour des données n'impliquerait pas de devoir refaire toute la base de données à chaque fois. Il suffirait d'ajouter les nouveaux éléments à ceux déjà stockés. Le second avantage est que ce type d'outil ne « prémâche » pas du tout le travail journalistique. Il aide juste à la recherche d'informations. Le désavantage se situe plutôt en termes de visualisations : il faudrait permettre de « visualiser » les résultats d'une recherche par mots-clés, sous une forme ou une autre (par ex. « nuages de mots » ou « nombre d'occurrence selon l'années/le mois » ou « co-occurrences les plus fréquentes »). Enfin, ce type d'outil exige un développement technique et informatique important en amont, ainsi qu'un grand travail de rassemblement de données. L'accessibilité et le format des données (audio, vidéo, écrite, pdf) constitue une difficulté supplémentaire, ainsi que la nécessité d'avoir les données les plus exhaustives possible.

3. Un dashboard « Heidi.news » avec prototype de recherche lexicale

La troisième piste mêle les deux premières, tout en revoyant les ambitions de l'une et de l'autre à la baisse. Il s'agirait bel et bien de produire un « dashboard » sur un thème, en l'occurrence le climat, mais à destination d'un seul média, à savoir *Heidi.news*. Une collaboration étroite avec le média partenaire peut permettre de mieux orienter et cibler tant les données utilisées que les visualisations produites, voire l'interprétation des données. Le lieu d'hébergement de la plateforme reste à définir avec le média partenaire. Une possibilité serait qu'elle soit hébergée par *Heidi.news*, mais propose des liens vers les données brutes de l'EPFL qui seront, elles, accessibles et téléchargeables à tout autre journaliste. La plateforme se diviserait ainsi en deux niveaux : le premier niveau serait composé des bases de données constituées, le second niveau des visualisations créées spécifiquement pour *Heidi.news*. Graphiquement, les visualisations produites répondraient toutefois à la charte *Heidi.news*. Tout autre médias intéressés pourrait utiliser non pas les visualisations mais les données.

Dans ce « dashboard » serait intégrée une partie consacrée aux citations de politiciens sur le climat, non exhaustives et ciblées sur des documents écrits (presse, voire PV de séances). Il s'agirait d'un « prototype » d'outils de datajournalisme « discursif ». Le projet devrait permettre d'explorer des pistes de sujets et de visualisation à partir de ce type de données, éventuellement dans la perspective d'un projet futur.

C'est cette troisième piste qui a été privilégiée à ce stade (26.01.2021). Elle permet au projet de se construire autour de plusieurs « sous-projets » ayant un rapport avec le thème central. Le projet consisterait donc bel et bien en une plateforme dédiée au climat, à laquelle peuvent avoir accès l'ensemble des journalistes (et du public), construite sous forme de différents « volets » alimentés par différents membres du projet. Chaque volet comprend un ou plusieurs sets de données brutes (nettoyées, structurées), accessibles à tout le monde, et des

visualisations conçues pour *Heidi.news* qui permettent d'observer une ou plusieurs formes de visualisations envisageables des données brutes.

4. Les étudiants : des bêta-testeurs

Le projet MediaLaboratory intègre deux étudiants de l'Académie du Journalisme et des médias qui souhaitent réaliser leurs travaux de master en utilisant des outils de data journalism. Ils officient comme bêta-testeur non seulement des futurs outils développés mais aussi de la démarche dans son ensemble (développement d'outils de datajournalisme à l'usage des journalistes non spécialisés). Les premiers mois durant lesquels ils ont participé au projet ont déjà permis de relever quelques défis probablement courants. Tout d'abord, ils sont journalistes. Ils recherchent donc des sujets « originaux », « inédits » et qui les intéressent. Le (vaste) thème du « climat » a été choisi. Une fois le thème établi, les étudiants ont été amenés à trouver des bases de données sur lesquelles travailler. Ils ont vite été confrontés à d'importantes questions : quelles données utiliser ? Où peut-on trouver des données ? Sont-elles originales ? Plusieurs propositions de bases de données ont été analysées. Chacun des deux étudiants s'est tourné vers des bases de données « scientifiques », produites par des Universités. Ils ont vite constaté que ces données étaient trop précises, trop spécifiques pour être traitées sous la forme d'une production médiatique « grand public ». De plus, il leur manquait une dimension « historique » et le potentiel d'être mises à jour régulièrement.