



Le Data Journalismisme Lab

Une **expérience** de **data journalismisme** à Bordeaux

menée par
l'Institut de Journalisme de Bordeaux Aquitaine (IJBA)
et AEC (L'agence des initiatives numériques en Aquitaine)

Modus Operandi

Edith Rémond et Suzanne Galy

Juillet 2012

1 Le Data Journalism Lab s'inscrit dans une histoire et une réflexion	p.3
2 Neuf mois de gestation	p.4
3 Des objectifs et un programme	p.6
1/ Les objectifs du Data Journalism Lab	p.6
2/ Les acteurs de l'open data et les experts	p.7
3/ Les missions à mener	p.7
4/ Le calendrier des productions	p.8
4 Premiers écueils, premiers réajustements	p.9
1/ Pas de données ne vaut pas mieux que trop de données	p.9
2/ Réunir les acteurs des projets pour créer une dynamique collaborative	p.9
5 Une dernière ligne droite pour boucler et gérer les frustrations	p.11
6 Les belles potentialités de cette nouvelle forme de journalisme... et ses limites	p.14
7 Un retour d'expérience riche en enseignements	p.15
1/ Les remarques des étudiants	p.15
2/ Les conclusions provisoires de l'équipe d'encadrement	p.16
8 Annexes :	p.20
1/ Ressources à consulter avant le lancement du Data Journalism Lab	p.21
2/ Les data visualisations de référence des étudiants de l'IJBA	p.22
3/ Les étapes et les acteurs d'une production de data journalisme	p.23
4/ Les 12 projets et les 12 équipes de travail du Data Journalism Lab	p.24
5/ Note juridique : données publiques, Open data et licences d'utilisation	p.25
6/ I want to make a data visu – Document établi par des étudiants du Data Journalism Lab et réalisé par Pascal Bonnefon	p.27
7/ Outils pour la data visualisation	p.28

Le Data Journalisme Lab est un module pédagogique lancé en janvier 2012 à l'initiative de l'IJBA (école de journalisme reconnue par la convention collective) et d'AEC, agence des initiatives numériques en Aquitaine.

L'objectif était de travailler sur les données publiques locales libérées pour construire de l'information en s'adressant à l'intelligence visuelle des lecteurs (infographies, data visualisations...) et d'initier les étudiants à la culture de l'innovation et au travail en mode collaboratif.

Dans l'esprit de l'open data, nous avons voulu travailler dans la transparence et avons mis en ligne tous les documents produits dans le Data Journalisme Lab et bien sûr les productions sur <http://www.datajournalismelab.fr>

1/ Le Data Journalisme Lab s'inscrit dans une histoire et une réflexion

Le Data Journalisme Lab s'est substitué à un module pédagogique antérieur, *les Débats numériques*, lui aussi initié et encadré par l'IJBA et AEC en partenariat avec *Le Monde*, qui de 2008 à 2011 a permis aux étudiants d'organiser des débats publics sur les enjeux sociétaux des mutations numériques : <http://www.ijba.u-bordeaux3.fr/Les-d%C3%A9bats-num%C3%A9riques/>

Le dernier débat numérique traitait des « données numériques entre usage commercial et usage citoyen » : <http://www.ijba.u-bordeaux3.fr/le-debat-numerique-edition-11-les-donnees-numeriques-entre-usage-commercial-et-usage-citoyen.html> Dans la phase de préparation, la faiblesse des connaissances des étudiants sur la libération des données publiques, l'appropriation et le traitement que la presse pouvait imaginer a encouragé les formatrices à réfléchir à une initiative pédagogique palliant ces lacunes inacceptables pour de jeunes journalistes.

Dans le même temps, les attentes – et à certains égards l'attentisme – de la presse française sur le journalisme de données était évoquées dans de nombreuses conversations ou tables rondes avec des dirigeants d'entreprises ou des professionnels en situation de responsabilités.

La place et le rôle des écoles de journalisme étaient eux aussi interpellés et de nombreux enseignants plaidaient, à l'image de Marc Mentré, que « les formations au journalisme doivent devenir des laboratoires » et qu'il est impératif de « placer l'innovation au cœur des cursus » : http://www.cahiersdujournalisme.net/cdj/pdf/22_23/11_MENTRE.pdf

Enfin l'IJBA menait chaque année depuis 2007, à l'initiative d'Edith Rémond, des blogs de quartier avec ses étudiants (<http://www.ijba.u-bordeaux3.fr/prod-ecrit.html>) et savait combien l'expression multimédia est un terrain pédagogique riche pour la formation des journalistes mais aussi combien ils peinent à inventer des formats et des modes de traitement nouveaux de l'information.

Dans une ingénierie pédagogique innovante, le Data Journalism Lab devait donc partir de sources d'information nouvelles et mettre en action des compétences et des acteurs issus d'univers et de cultures professionnels différents pour inventer ensemble de nouveaux modes de traitement de l'information.

2/ Neuf mois de gestation

De mai 2011 à janvier 2012, la préparation du Lab a porté sur le programme bien sûr mais aussi sur les objectifs, les acteurs et les moyens. Elle est racontée sur <http://www.datajournalismelab.fr/mai-2011/>

Une fois arrêtés les objectifs,

- Trier des données, les comprendre et les analyser, confronter des jeux de données complémentaires,
- Interroger les politiques publiques,
- Enquêter en amont et en aval des données utilisées,
- Travailler sur une production commune avec des statisticiens, des web designers, des graphistes, des développeurs...
- Raconter des histoires en s'adressant à l'intelligence visuelle du lecteur (data visualisation, graphiques, tableaux, infographies, schémas...),
- Produire et diffuser au terme de l'expérimentation, à destination des professionnels des médias, un *Modus Operandi* de data journalisme.

nous avons formalisé le *process*,

- **Compiler** et **trier** la masse gigantesque de données,
- **Se documenter et enquêter** sur le sujet et l'angle retenu,
- **Organiser** les données pertinentes et non pertinentes,
- **Traduire** la donnée brute en statistiques et selon différentes unités de valeur pour les comparer. Recalculer si nécessaire,
- **Analyser** les données, donner du sens, angler à nouveau l'information. Évaluer la pertinence du résultat obtenu,
- **Publier** sur le web la donnée rendue informative et éventuellement interactive par le travail graphique et le développement informatique.

et nous avons identifié les compétences requises,

- journalistes
- graphistes
- développeurs
- statisticiens
- cartographes
- démographes

fréquenté les acteurs de l'open data, notamment dans les open data bar d'AEC⁽¹⁾, et beaucoup consulté les sites de la presse anglo-saxonne.

Puis nous nous sommes formées, début janvier 2012, à Paris avec Caroline Goulard de *Dataveyes*, pour prendre en main des logiciels gratuits et disponibles en ligne de traitement des données et de data visualisation.

Suzanne Galy a également participé à la même période à *Hack The Press* à la Cantine de Paris.

Très vite il nous est apparu que les moyens matériels et les compétences d'une école de journalisme ne suffiraient pas à financer et réaliser notre projet.

Nous avons donc fait appel à deux écoles qui se sont associées au Data Journalism Lab :

- l'Ecole de COMMUNICATION visuelle (ECV) de Bordeaux <http://www.ecv.fr/?gclid=CJmP4K-kj7ACFVMetAodaWZTqQ> qui nous a délégué quatre étudiants graphistes regroupés dans un collectif de jeunes créatifs <http://burgercom.fr/> : Adrien Colombié, Yohan Benazzouz, Martin Caro et Rémy Gendre,
- l'EPITECH Aquitaine <http://www.epitech.eu/ecole-informatique-bordeaux-aquitaine/> qui a présenté le Lab à ses étudiants. Six d'entre eux se sont manifestés auprès de nous : Maxime Chaisse-Leal, Colin Delhalle, Raphaël Estrach, Babacar Gueye, Baptiste Rio, Quentin Rivière.

Cependant ni l'une ni l'autre des deux écoles n'ont intégré le Data Journalism Lab dans leur programme de formation. Les étudiants qui se sont lancés dans cette aventure l'ont donc assumé en plus de leurs charges pédagogiques, de leurs stages et/ou de leur contrat d'alternance.

Un développeur web professionnel, David Bruant, nous a donc rejoints pour étoffer une équipe trop restreinte qui ne pourrait manifestement pas nous permettre d'aboutir sur tous nos projets. Très investi dans des démarches de production collaborative en lien avec le mouvement Open Source, David Bruant a accepté de travailler bénévolement pour le Data Journalism Lab et nous a sauvé la mise.

(1) Les open data bar d'AEC étaient des ateliers visant à la réutilisation des données publiques locales libérées.

Nous avons également constitué une équipe d'encadrement du Lab au sein de laquelle les compétences requises étaient présentes.

Outre les deux coordonnatrices :

- **Suzanne Galy**, rédactrice en chef à AEC
- **Edith Rémond**, enseignante à l'IJBA et responsable du master professionnel

un informaticien et un journaliste graphiste ont accepté des responsabilités dans le projet :

- **Pascal Bonnefon**, graphiste web et journaliste indépendant plus particulièrement chargé d'encadrer et de coordonner le travail des étudiants graphistes de l'ECV,
- **Olivier Pichon**, développeur et chef de projet web au sein de la société de services informatique groupe-SII avec laquelle nous avons signé un contrat de prestation de service. Dans le Lab, Olivier Pichon était le référent des développeurs mais il s'est très vite investi dans les sujets et auprès des étudiants en journalisme de l'IJBA.

Pendant que nous réfléchissions à notre projet, le quotidien *Sud-Ouest*, en la personne de Guillaume Vasse, directeur du développement numérique et directeur marketing du groupe, nous a contactées pour témoigner de son intérêt et du souhait du journal d'être partenaire du Data Journalism Lab.

3/ Des objectifs et un programme

Avant de partir en congés de Noël, les étudiants ont trouvé sur leur messagerie un document de références⁽²⁾ à consulter pour ne pas aborder à la rentrée le Data Journalism Lab en ignorant tout de ce genre journalistique et/ou de l'open data : <http://www.datajournalismelab.fr/quelques-references/>

Et le jeudi 12 janvier, à 14 heures, les 36 étudiants de première année de master de l'IJBA ont découvert :

1/ Les objectifs du Data Journalism Lab :

- Utiliser des données publiques locales concernant le territoire aquitain libérées par la CUB, le Conseil Général de la Gironde, le Conseil Régional d'Aquitaine et d'autres organismes publics ou privés, sous des formats plus ou moins structurés, pour construire de l'information,
- Trier des données, les comprendre et les analyser, confronter des jeux de données complémentaires,
- Interroger les politiques publiques,

(2) Voir en annexe 1 page 21 : Ressources à consulter avant le lancement du Data Journalism Lab.

- Enquêter en amont et en aval des données utilisées,
- Travailler sur une production commune avec des statisticiens, des web designers, des graphistes, des développeurs...
- Raconter des histoires en s'adressant à l'intelligence visuelle du lecteur (data visualisation, graphiques, tableaux, infographies, schémas...)
- Produire et diffuser pendant et au terme de l'expérimentation, à destination des professionnels des médias, un Modus Operandi de data journalisme.

2/ Les acteurs de l'open data et les experts qu'ils allaient rencontrer :

le jeudi 19 janvier

Les domaines de compétences de la CUB, du Conseil général de la Gironde et du Conseil régional d'Aquitaine, leur implication et leurs choix stratégiques dans la libération des données publiques :

- Sylvain Maire, chargé de mission Fing / animateur dynamique open data en Aquitaine
- Frédéric Théodore, DSI adjoint Communauté Urbaine de Bordeaux (CUB)
- Pascal Romain, DSI Conseil Général de la Gironde

les jeudi 26 janvier ; 2, 9 et 16 février ; 8 et 15 mars de 14h à 16h

- 26 janvier : Olivier Pichon, Ingénieur groupe-SII et les stagiaires d'Epitech
- 2 février : Alexandre Bertin, économiste statisticien AEC
- 9 février : Julien Gelly, cartographe
- 16 février : Marine Llopart, démographe
- 8 et 15 mars : Pascal Bonnefon, journaliste graphiste et les stagiaires de l'ECV

3/ Les missions à mener

- Par groupes de trois étudiants, présenter et analyser un exemple de data journalisme : <http://www.datajournalismelab.fr/les-analyses/>⁽³⁾
- Nourrir le blog datajournalisme.fr, work in progress du Data Journalismisme Lab.
- Et bien sûr – et surtout – produire par groupes de trois journalistes + un graphiste + un développeur des data visualisations informatives sur un sujet de leur choix.

(3) On trouvera en annexe 2 page 22, la liste des data visualisations présentées et analysées au cours de ces exposés.

4/ Le calendrier des productions

Pour les conférences de rédaction, qui avaient lieu le jeudi en fin d'après-midi, la promotion a été divisée en deux demi-groupes G1 et G2 avec, dans la phase préparatoire, le calendrier suivant :

	G1	G2
Identification des sujets, des données et des angles	jeudi 26 janvier	jeudi 2 février
Remise d'un pitch à Edith Rémond : sujet, angle, jeu de données utilisées, première ébauche visuelle	lundi 30 janvier	lundi 6 février
Présentation des enquêtes, des reportages et des objectifs de la production	jeudi 9 février	jeudi 16 février
Propositions de mises en forme visuelles et graphiques	jeudi 8 mars	jeudi 15 mars

Une journée de regroupement pré production de l'ensemble de la promotion était ensuite prévue à l'IJBA le vendredi 23 mars.

Et la finalisation des data visualisations devait se dérouler en session par demi-groupe :

- du lundi 26 mars au vendredi 30 mars de 9h30 à 18h – G1 en six sous-groupes de trois étudiants
- du lundi 2 avril au vendredi 6 avril de 9h30 à 18 h – G2 en six sous-groupes de trois étudiants

Enfin, le 12 avril en fin d'après-midi sur le plateau public de l'IJBA étaient prévus une projection et un débat public autour des productions du Data Journalism Lab qui seraient mises en ligne le jour même sur le site de l'école. Il était également convenu que les meilleures productions pourraient être mises en ligne sur le site du journal *Sud-Ouest* mais celui-ci ne s'est jamais manifesté par la suite dans le Lab.

Après un cours introductif <http://www.datajournalismelab.fr/cours-dintroduction/> et une présentation du partenariat AEC par Suzanne Galy, une première conférence de rédaction nous a fait mesurer l'ampleur de la tâche !

4/ Premiers écueils, premiers réajustements

Dès les premiers tours de table, quelques projets d'étudiants nous ont mises en difficultés. Nous avons beau leur répéter qu'il fallait se plonger dans les données libérées et identifier des sujets sur lesquels ils devraient ensuite réfléchir à des angles, plusieurs parmi eux, en bons journalistes à l'ancienne, commençaient par énoncer un sujet sur lequel ils aimeraient travailler sans se soucier de l'existence de données libérées – ou tout simplement accessibles - sur ce sujet.

1/ Pas de données ne vaut pas mieux que trop de données

Un groupe, par exemple, s'est cramponné pendant un mois à un sujet qui semblait le passionner : « les pratiques mortuaires en France ». Jusqu'à ce qu'ils réalisent enfin, en voyant avancer les autres groupes, qu'ils prenaient du retard et piétinaient, que ce sujet n'était pas à sa place dans le Data Lab, faute de données disponibles identifiées. D'autres projets initiaux sur le prix de l'eau, l'évolution de la population bordelaise par quartier ou les subventions aux associations ont connu la même infortune.

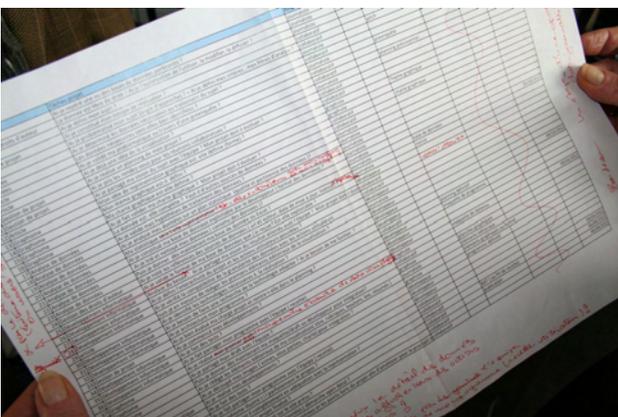
D'autres, peu sensibles à l'open data, ont résisté à nos conseils en nous expliquant qu'ils allaient trouver les données, même si elles n'étaient pas libérées. Sans bien mesurer que les fichiers PDF, pages html, docs word, etc. nécessitent, pour produire une data visualisation, un temps de retraitement très long pour les transformer en tableurs réutilisables par des développeurs et un travail assez fastidieux qui risque, de surcroît, de générer des erreurs de saisie.

A l'inverse un trop-plein de données peut aussi être source de difficultés. Lorsqu'il a eu fait le deuil de son sujet sur les pratiques mortuaires, le même groupe s'est emballé pour un projet sur la réussite scolaire et s'est attelé à un travail de titan pour nettoyer des données cette fois pléthoriques, choisir un périmètre, angler le sujet et le mettre en scène. Un autre groupe qui voulait absolument travailler sur les crimes et délits en France s'est confronté au même écueil.

2/ Réunir les acteurs des projets pour créer une dynamique collaborative

Malgré la création de groupes sur Facebook et l'utilisation intensive des réseaux sociaux par tous les étudiants impliqués, nous nous sommes rapidement rendu compte que, si nous voulions avancer et produire, il fallait asseoir les journalistes, les graphistes et les développeurs dans un lieu de travail confortable pendant au moins deux jours afin de créer une véritable dynamique collaborative.

Nous avons donc organisé les 10 et 11 mars dans l'incubateur de l'Auberge numérique d'AEC un week-end d'accélération : <http://www.datajournalismelab.fr/les-photos-du-datajournalismelab-au-travail/> pour produire « la tête dans le guidon » dans une dynamique collaborative, découvrir le métier des collaborateurs et comprendre clairement le circuit d'une production de data journalisme.



Ces deux journées, dont la majorité des étudiants avaient dans un premier temps boudé le projet, ont été déterminantes dans leur implication et leur intérêt pour le Data Journalism Lab, qui ne se sont plus démentis. En travaillant aux côtés des étudiants développeurs et des étudiants graphistes, ils ont pris la mesure de l'intérêt de ce chantier, de son caractère innovant et porteur d'une réflexion journalistique et esthétique qu'ils étaient les premiers à expérimenter dans une école de journalisme. Le 11 mars au soir, ils s'étaient approprié le défi.

Les premières ébauches visuelles y furent pour beaucoup. Mais il avait fallu auparavant finir de nettoyer les données, constituer les bases à exploiter, imaginer et concevoir l'interactivité de chaque production, réfléchir aux enquêtes de terrain et rédiger déjà quelques éléments de textes.

Au cours du week-end, plusieurs acteurs locaux du numérique et journalistes professionnels sont venus observer ce laboratoire en action et faire part de leur intérêt pour ce projet. Les experts statisticiens et cartographes se sont assis auprès des groupes qui les sollicitaient pour leur apporter leur aide et répondre à leurs questions pratiques. Et l'équipe d'encadrement a profité d'une courte pause pour mettre en forme un tableur de gestion d'un projet de data journalisme⁽⁴⁾ : qui fait quoi et dans quel ordre pour mener à bien un projet de production de data visualisation ?

<http://www.datajournalismelab.fr/les-preconisations/>

Ce fut le premier enseignement du Data Journalism Lab : il fallait clarifier pour les étudiants le process d'une production. Et mettre l'accent sur le rôle et les attributions du chef de projet qui, au fil du temps, nous apparaissait de plus en plus important.

5/ Une dernière ligne droite pour boucler et gérer les frustrations

A peine un mois plus tard chaque groupe de journaliste était libre de tout autre engagement pédagogique pendant une semaine seulement pour boucler son projet. Les développeurs et les graphistes, eux, travaillaient à distance, et surtout la nuit, pour procéder aux derniers réajustements et mettre la dernière main aux productions⁽⁵⁾. Dans le meilleur des cas. Car une minorité de productions en étaient réellement au bouclage. Les autres étaient encore entre les mains des développeurs ou n'avaient fait l'objet que d'esquisses graphiques très approximatives.

(4) Voir en annexe 3 page 23, le tableur de gestion d'un projet de data journalisme.

(5) Voir en annexe 4 page 24, les 12 projets et le nom des étudiants des 12 équipes du Data Journalism Lab.

Après les espoirs et les ambitions qui avaient émergé pendant le week-end d'accélération où, dans l'enthousiasme et l'émulation des groupes de professionnels si différents, tout semblait possible, la semaine de bouclage fut celle des renoncements et de la pression qu'impose une date de livraison.

Déception de certains face à des réalisations graphiques bien en deçà de ce qui avait été défini oralement, comme par exemple pour le groupe qui travaillait sur les performances des grimpeurs du Tour de France dont le visuel n'était pas à la hauteur de la qualité des informations obtenues auprès d'un chercheur et de la réflexion menée par les auteurs. Frustration pour le groupe qui travaillait sur les crimes et délits qui devait, dans la hâte, renoncer à des pans entiers de l'interactivité de son projet parce que le développeur n'avait plus assez de temps à leur consacrer. Dépit pour le groupe « réussite scolaire » qui comprenait qu'il serait le seul à n'avoir pas de data visualisation à présenter le jour J parce qu'il avait pris du retard et que son graphiste les avait « plantés » après leur avoir assuré qu'il était sur le point de livrer. Et puis il fallait aussi, pour certains, mener encore des enquêtes de terrain, des interviews. Et pour tous trouver les titres, réfléchir à l'habillage, écrire et réécrire les textes d'accompagnement, les légendes et appliquer les préconisations sur les conditions juridiques de publication des données, des visuels et les licences d'utilisation⁽⁶⁾...

Et, content ou pas, préparer le topo pour la présentation publique du 12 avril.

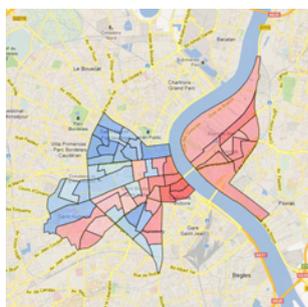
Ce jour là, dans l'amphi public de l'IJBA, 11 productions de data journalisme furent dévoilées au public attentif qui avait répondu présent :



Communauté Urbaine de Bordeaux : Votre ville doit-elle se serrer la ceinture ?

<http://www.datajournalismelab.fr/wp-content/uploads/dettes/index.html>

Le niveau d'endettement de chaque commune de la CUB.



Législatives à Bordeaux : quelle chance pour Alain Juppé dans la deuxième circonscription ?

<http://www.datajournalismelab.fr/wp-content/uploads/legislatives/index.html>

Analyse comparative des résultats par bureau de vote aux élections législatives à Bordeaux, 2ème circonscription en 1997, 2002, 2004 et 2007.

(6) Voir en annexe 5 pages 25 et 26 la note juridique : « Données, publiques, open data et licences d'utilisation.



Le comparateur des fragilités sociales : vit-on vraiment mieux en Gironde qu'en France ?

<http://www.datajournalismelab.fr/wp-content/uploads/visagesocialgironde/index.html>

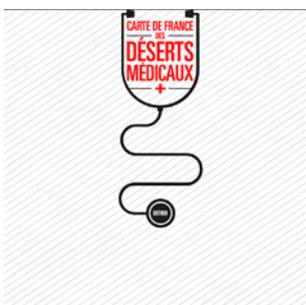
Comparaison de l'Indice de Santé Sociale (ISS) en Gironde et en France à partir d'indicateurs portant sur la santé, le logement, la qualification et l'emploi, le nombre de bénéficiaires du RMI, du RSA ou du minimum vieillesse.



L'IVG : plus de méthodes, moins de choix

http://www.datajournalismelab.fr/wp-content/uploads/visu_IVG/index.html

Une étude comparative des méthodes d'interruption de grossesses qui interroge le discours sur l'alternative de « l'IVG en ville ».



La carte de France des déserts médicaux

<http://www.datajournalismelab.fr/wp-content/uploads/desertsmedicaux/index.html>

Une carte de France en anamorphose qui présente pour chaque région (sauf la Corse faute de données libérées) le temps de parcours pour accéder à un hôpital et à un médecin généraliste.



L'évolution des crimes et délits en France : à quoi a servi la politique sécuritaire de Nicolas Sarkozy?

<http://www.ecl-autoecole.fr/criminalite/www/>

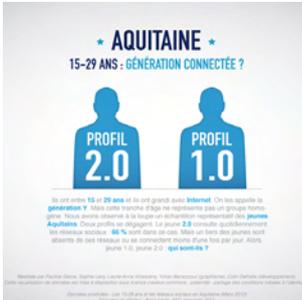
Une comparaison nationale et régionale du nombre de crimes et de délits, classés par catégories, en 2002, 2007 et 2010.



Arrêt sur images : à quoi servent les 56 caméras de sécurité à Bordeaux ?

http://www.datajournalismelab.fr/wp-content/uploads/visu_camera/index.html

Une carte des caméras de vidéosurveillance de la ville de Bordeaux et le « tableau de chasse » de chacune en matière d'infractions.



Aquitaine – 15-29 ans : génération connectée ?

<http://www.datajournalismelab.fr/wp-content/uploads/generationy/index.html>

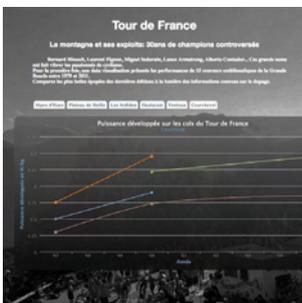
Deux profils de jeunes Aquitains en fonction de leur usage des réseaux sociaux.



La France sous les projecteurs : fréquentation des cinémas en France

<http://www.datajournalismelab.fr/wp-content/uploads/cinema/index.html>

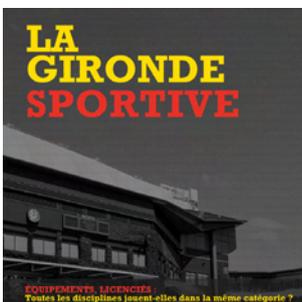
Dans les grandes villes de France, le nombre d'entrées au cinéma par habitant ainsi que le nombre de fauteuils de cinéma par habitant, le prix moyen d'un ticket d'entrée, le salaire médian et le nombre de multiplexes.



Tour de France – La Montagne et ses exploits : 30 ans de champions controversés

<http://www.datajournalismelab.fr/wp-content/uploads/tourdefrance/index.html>

Dans des cols légendaires du Tour, la puissance développée par des coureurs qui le sont tout autant, Bernard Hinault, Jan Ullrich, Lance Armstrong et bien d'autres, comparée sur un même graphique.



La Gironde sportive – Equipements, licenciés : toutes les disciplines jouent-elles dans la même catégorie ?

<http://www.datajournalismelab.fr/wp-content/uploads/equipementsportifs/index.html>

Comparaison par sport du nombre de licenciés et du nombre des équipements girondins et podium des gagnants.

6/ Les belles potentialités de cette nouvelle forme journalistique... et ses limites

Ces douze visualisations témoignent de la richesse de ce nouveau genre.

Il permet de traiter toutes sortes de sujets : politiques, sociaux, économiques, culturels, sportifs...

Il peut être appliqué à tous les territoires : local, départemental, régional, national et international.

Il peut prendre des formes démonstratives comme le sujet sur l'IVG qui est statique mais sert implacablement un propos cohérent. Il peut aussi emprunter la voie de la comparaison pour mettre en lumière les spécificités d'une situation locale - c'est le cas par exemple du sujet sur les fragilités sociales - ou pour mettre en compétition des champions qui n'ont pas exercé leurs talents à la même période, comme les coureurs du Tour de France. Il peut être plus simplement, mais avec autant de force, révélateur quand il met en relation des données qui n'ont jamais été comparées comme l'emplacement des caméras de surveillance et leurs « performances » ou les équipements sportifs et les pratiquants sur la même zone géographique.

La forme visuelle donne aux informations un impact important. Imaginons le sujet sur la dette des communes de la CUB traité en texte ! Même avec quelques camemberts, il serait indigeste. L'interactivité, la liberté donnée à l'internaute de naviguer à son gré, d'interroger le sujet à partir de ses propres questionnements, confère à ces produits journalistiques un intérêt et un attrait évidents.

Mais la rançon des formidables capacités du data journalisme est l'exigence décuplée de rigueur qu'il impose à ses auteurs. Et en cela nos productions ne sont pas exemptes de faiblesses. La majorité de nos visuels sont accompagnées de textes explicatifs qui visent souvent à relativiser le propos ou à en expliquer la complexité. Mais combien d'internautes en prendront-ils vraiment connaissance ? Et combien se contenteront plutôt d'un visuel qui frappe l'esprit ? Notre travail souffre aussi de la faiblesse des moyens dont nous disposons. Nous aurions souhaité, pour beaucoup de productions, approfondir davantage... Nous nous sommes heurtées enfin à la question des données « propres » ou non. Le groupe qui travaillait, par exemple, sur les équipements sportifs en Gironde, a du « nettoyer » le tableau des équipements sportifs mis en ligne par le Conseil Général qui comportait, à l'évidence, de nombreuses erreurs. Mais s'il est simple de constater qu'un gymnase n'est pas un terrain de golf et qu'une erreur s'est glissée dans une ligne, combien de chiffres utilisés ont-ils été exploités dans nos productions sans que nous puissions garantir leur pertinence ? Cette question de la fiabilité des données, qui requiert des heures et des heures de vérification, nous semble la plus cruciale et la plus fastidieuse dans le travail du journaliste qui s'empare des données.

7/ Un retour d'expérience riche en enseignements

1/ Les remarques des étudiants

Le 3 mai, à leurs retours de stages en entreprises, les étudiants ont été invités à un debriefing du module pédagogique consacré au data journalisme.

Ils avaient tous conscience d'avoir beaucoup appris en matière de :

- maîtrise des données chiffrées et du tableur Excel,
- travail et organisation en équipe,
- impérieuse nécessité d'une cohérence entre sujet, angle et sources d'information,
- compréhension du travail de développeur et de celui de graphiste, même s'ils regrettaient pour certains d'avoir du mal à évaluer le temps nécessaire au développement en fonction du projet,
- concessions et réajustements permanents d'un projet journalistique en fonction des données mais aussi de la quantité de travail qu'impacte chaque idée dans la chaîne de production (développeur et/ou graphiste),
- nécessité d'enquêter sur le sujet : le traitement et la sélection des données ne sont pas suffisants pour être pertinents journalistiquement,

Ils prônaient pour la plupart :

- une intervention des experts au plus près de chaque projet journalistique. Les interventions théoriques, pourtant accompagnées de visuels et d'exemples, n'avaient manifestement eu de l'intérêt que pour celles et ceux qui avaient par la suite fait appel à la compétence d'un des experts pour répondre à leurs interrogations et travailler ensemble sur leur projet,
- une organisation qui permette aux journalistes, graphistes et développeurs de partager des temps de travail communs et qui inclue davantage de développeurs et de graphistes,
- un week-end data de lancement et plus généralement un travail en sessions plutôt que sur un rythme hebdomadaire.

Certains parmi eux avaient eu l'idée de produire un document⁽⁷⁾ librement inspiré des affiches « I want to make... » de Canal + qui témoigne du sentiment de complexité que leur a inspiré le travail dans le data lab : <http://www.datajournalismelab.fr/i-want-to-make-a-data-visu/>

(7) Voir en annexe 6 page 27, *I want to make a data-visu*.

2/ Les conclusions provisoires de l'équipe d'encadrement

Avant d'engager un projet de production de data journalisme, quelques compétences **préalables** sont nécessaires :

- une maîtrise des sources, ce qui signifie une connaissance des pourvoyeurs de données, des formats de données et la capacité à identifier les « bonnes données » ou les données « propres »,
- une connaissance approfondie des formats informatiques des bases de données et de leur réutilisation,
- une initiation solide à Excel,
- une connaissance, voire une maîtrise, de nouvelles compétences techniques : statistique, cartographie, langages informatiques (html, CSS, Javascript / qu'est-ce qu'une API ?), et quelques outils de visualisation disponibles sur le web⁽⁸⁾,
- une sensibilisation au cadre juridique de la data visualisation en journalisme de données⁽⁹⁾.

Nous avons, en cours d'expérience, produit les documents sur les outils de visualisation et le cadre juridique pour combler les lacunes des étudiants. Ils n'y ont pas prêté grande attention. Il nous reste à trouver un moyen de leur faire comprendre la nécessité de maîtriser ces savoirs austères et techniques avant de « jouer » au journaliste de données et d'éprouver les joies de la concrétisation visuelle de leur projet.

Au terme de cette expérience, nous avons quelques convictions sur le data journalisme :

a/ Pour une data visualisation informative, le meilleur principe est « **less is more** ». La simplicité est un pari gagnant :

- la problématique centrale ou « message essentiel » doit tenir en une ligne,
- le nombre de données traitées doit être raisonnable, voire limité, pour ne pas perturber le message informatif,
- le graphisme doit servir l'information et écarter le « délire créatif » pour la beauté de l'art,
- pour mettre en adéquation moyens (temps) et objectifs (information), il faut souvent renoncer au tout dynamique pour choisir plus de statique.

(8) Voir en annexe 7 pages 28, 29 et 30, le document qui a été fourni aux étudiants de l'IJBA.

(9) Voir annexe 5 page 25.

b/ Les productions de data journalisme imposent **un modèle vertueux de travail collaboratif**

- les membres de l'équipe de production doivent poursuivre et réinterroger sans cesse un but commun : la data visualisation doit servir l'information et répondre aux questions du public cible,
- ils doivent développer un savoir-faire mais aussi un savoir être dans lequel sont importantes les qualités relationnelles de bienveillance et de convivialité,
- chaque membre doit reconnaître les « compétences en miroir » : les capacités des uns complètent les limites et résistances des autres. Ils sont complémentaires mais à égalité. Chacun doit trouver sa place pour jouer son rôle,
- pourtant un chef de projet doit émerger pour coordonner et adopter des outils collaboratifs de gestion de projet, pour planifier les différentes étapes et déterminer les moments d'intervention des différents acteurs,
- « 100 fois sur le métier remettez votre ouvrage » et gardez courage ! Le processus de création est itératif : le data journalisme développe une culture de la « bidouille » et de l'éternel recommencement,
- il est préférable que les membres de l'équipe travaillent ensemble plutôt qu'à distance ou qu'au moins des moments de regroupements réguliers soient programmés. Si on est beaucoup sur son écran en data journalisme, il est indispensable que les acteurs du projet s'assoient côte à côte et se parlent.

c/ L'enseignement du data journalisme dans une école telle que l'IJBA, s'il ne prétend pas rendre tous les étudiants opérationnels sur ce format, a un mérite exceptionnel : c'est la meilleure **école de la rigueur journalistique**. La compatibilité sujet-angle-données-public doit être parfaite pour que la production fonctionne.

Il a aussi le mérite d'exposer les étudiants à des enseignants qui ne sont pas journalistes et à ouvrir leur horizon alors que le modèle de formation français est extrêmement corporatiste. En cela il prépare les étudiants à l'évolution de leur métier tel qu'il est étudié dans la synthèse des réflexions des groupes de travail « métiers » publié par l'observatoire des métiers en 2011 qui s'interroge « Quelles activités ? Quelles compétences pour demain ? » <http://www.metiers-presse.org/?uid=63>

d/ Mais l'expérience du Data Journalismisme Lab a également mis en évidence le **coût important** de ces productions. Quatre graphistes et six développeurs pour douze productions, c'était à l'évidence insuffisant. Et nous ne remercierons jamais assez David Bruant, qui a accepté de travailler le week-end et la nuit pour nous bénévolement parce que cette aventure innovante et l'enthousiasme de cette équipe le séduisaient.

Mais dans la vraie vie, le coût approximatif de production en développement aurait été d'à peu près 1000 € pour la production sur les élections législatives et de 500 € pour celle sur l'endettement des communes de la CUB.

Pour le graphisme et le développement de la production sur les caméras de surveillance à Bordeaux, il aurait fallu compter environ 3000 €. Le budget d'une production plus statique comme celle sur l'IVG aurait atteint tout de même 1500 €.

Si nous restons convaincues que ce laboratoire a toute sa place dans une formation telle que celle dispensée à l'IJBA et que ses enseignements sont riches, nous cherchons encore le modèle économique qui nous permettra de poursuivre l'aventure. Et nous partageons en cela les difficultés de la presse française à se lancer dans le data journalisme.

Les deux coordinatrices du Lab :

Suzanne Galy

Rédactrice en chef

AEC

suzanne.galy@aecom.org

Edith Rémond

Responsable du master

IJBA

edith.remond@ijba.u-bordeaux3.fr

Table des annexes

1/ Ressources à consulter avant le lancement du data journalisme Lab	p. 21
2/ Les data visualisations de référence des étudiants de l'IJBA	p. 22
3/ Les étapes et les acteurs d'une production de data journalisme	p. 23
4/ Les 12 projets et les 12 équipes de travail du Data Journalismisme Lab	p. 24
5/ Note juridique : Données publiques, Open data et licences d'utilisation	p. 25
6/ I want to make a data-visu – Document établi par des étudiants du Data Journalismisme Lab et réalisé par Pascal Bonnefon	p. 27
7/ Outils pour la data visualisation	p. 28

Annexe 1

2012 – M1 – DATAJOURNALISME LAB en partenariat avec AEC et Sud-Ouest

Ressources à consulter avant le lancement le 12 janvier 2012

Edith Rémond

1. Qu'est-ce que le data journalisme ?

- D'abord cette vidéo de 13 minutes dans laquelle Nicolas Kayser Bril explique pourquoi les journalistes doivent se former au data journalisme : <http://www.youtube.com/watch?v=vGDc55edfsA>
- Une présentation certifiée Guardian : <http://prezi.com/vzindkg-k8l5/data-journalism-master/>
- Le compte-rendu d'un atelier data journalisme organisé en octobre 2011 par Etalab : <http://www.jeromecukier.net/blog/2011/10/14/open-data-and-data-journalism/>

2. Expériences de data visualisation :

- Le site d'owni : <http://owni.fr/>
- La rubrique data du Guardian : <http://www.guardian.co.uk/data>
- Le site de visualisation de l'actualité : <http://www.actuvisu.fr/>
- Un work in progress sur l'atelier « Visualisons rennes » : <http://blog.visualisons-rennes.fr/>
- Le site de Dataveyes, start-up de visualisations interactives de données : <http://dataveyes.com/> et le blog de Dataveyes : <http://blog.actuvisu.fr/>
- <http://developmentseed.org/projects/>
- Une présentation de Nicolas Kayser Brill avec notamment des outils gratuits : http://whatdatacan-doforyou.cloudcontrolled.com/#screen_jlist_investigate

3. Les données locales à partir desquelles nous pourrions produire :

- Sur le site d'AEC <http://www.aecom.org/Soutenir-vos-projets/L-Auberge-Numerique/L-OpenData-LAB>, vous trouverez un Tableur recensant les jeux de données libérées en Aquitaine, classés par thèmes (Culture, éducation, environnement, solidarité, sports/loisirs, transport, eau, équipements, urbanisme, tourisme)
- Le catalogue des données de la Gironde (Conseil Général – portail mutualisé avec la Région Aquitaine) : <http://datalocale.fr/dataset/listing-dataset>
- Le portail de la Communauté urbaine de Bordeaux : <http://data.lacub.fr/index.php>
- Le catalogue de données de PIGMA (Plateforme de l'Information Géographique Mutualisée en Aquitaine) : <http://cartogip.fr/spip.php?rubrique22> qui contient des données d'origines très diverses issues par exemple du recensement de la population 2007 (INSEE), de Pôle Emploi, de la direction générale des Impôts ou du répertoire des entreprises SIRENE.
- Sur demandes spécifiques si nécessaire, les données recensées par SIRTAQUI (Système d'Information Régional touristique d'Aquitaine) qui ne libère pas des données à proprement parler mais les rend accessibles sur convention.

Annexe 2

Les data visualisations de référence des étudiants de l'IJBA

- Atlas des micro-partis 2009
- Ce qu'ils savent / What they know
- Contrefaçon : la liste noire
- ? Interactive : Where & How We Live ?
- La carte des homicides à New York
- La carte d'un monde d'espions
- La France carcérale
- La puissance des coureurs sur le Tour de France
- Mémorial des morts aux frontières de l'Europe
- Who owes what to whom?/Dette de l'Eurozone : qui doit quoi à qui ?

Annexe 3 Le tableur de gestion d'un projet de data journalisme

Fonction	Tâches projet
Données	Ai-je repéré une ou des bases de données pertinente(s) ?
Gestion des droits d'auteur	La donnée est-elle libre de droit ? Ai-je l'autorisation de l'utiliser, la modifier, la diffuser ?
Article	Ai-je défini mon sujet ?
Gestion de projet	Ai-je constitué mon équipe de réalisation ?
Article	Ai-je problématisé mon sujet (angle d'approche) ? Ai-je défini mes critères , mes filtres d'analyse ?
Données	Ai-je déterminé les données nécessaires au traitement du sujet ?
Données	Les données nécessaires sont-elles bien identifiées ?
Article	Ai-je validé ma problématique au travers d'une préenquête ?
Article	Ai-je déjà une idée de la visualisation ?
Gestion de projet	Ai-je réalisé un planning prévisionnel de mon projet ?
Article	Ai-je rédigé mon projet (pitch) ?
Gestion de projet	Ai-je pris les rendez-vous pour les réunions collaboratives ?
Article	Ai-je partagé ma vision du sujet avec l'équipe ? Mon projet doit-il évoluer ?
Graphisme	Mon idée de représentation graphique est- elle définie (forme) ?
Graphisme	Mon thème graphique est-il défini ?
Graphisme	Ma charte graphique est-elle définie ?
Graphisme	Ai-je fait valider ma charte graphique ? Mon projet doit-il évoluer ?
Analyse de données	La donnée est- elle complète ? Me faut-il d'autres sources de données ?
Analyse de données	Ai-je validé mes hypothèses de traitement des données ?
Analyse de données	Ai-je analysé mes données sur un plan statistique (qualité et pertinence des données) ?
Analyse de données	Ma statistique est-elle cohérente avec mon sujet et mon angle ?
Analyse de données	Ai-je constitué ma base de données exploitable par le développeur (format des données) ?
Analyse de données	Ai-je trié et nettoyé mes données (contenu des données) ?
Analyse de données	Puis-je communiquer dès maintenant ma base de données ?
Gestion des droits d'auteur	Ai-je noté les références des sources de données : url et auteur, date de mise à jour, licence d'exploitation (notice) ?
Développement Informatique	Ai-je validé le format de ma base de données que j'ai reçue ? Est-elle exploitable ?
Développement Informatique	Analyse du besoin par le développeur : ai-je partagé ma vision du résultat attendu avec l'équipe ? Mon projet doit-il évoluer ? (users' story)
Développement Informatique	Ai-je défini l'interactivité du projet et les actions utilisateurs possibles ?
Développement Informatique	Ai-je prévu le détail des données que je vais afficher lors les actions utilisateurs ?
Article	Suis-je en mesure de rédiger les premiers éléments éditoriaux de la visualisation ?
Développement Informatique	Ai-je identifié les technologies nécessaires à la réalisation ?
Développement Informatique	Ai-je besoin de faire une recherche de la technologie adaptée ? Ai-je besoin de me former ?
Développement Informatique	Ai-je architecturé mon programme ?
Gestion de projet	Ai-je bien évalué ma charge projet et rentre-t-elle dans le planning (planning définitif) ?
Développement Informatique	Ai-je commencé mes développements (maquette de programme) ?
Développement Informatique	Ai-je fait des tests ?
Développement Informatique	Ai-je partagé mon programme avec l'équipe (recette) ?
Graphisme	Ai-je réalisé ma première ébauche de data visu (draft) ?
Graphisme	Ai-je partagé mon draft avec l'équipe ?
Graphisme	Puis-je donner dès maintenant ma maquette/draft au développeur pour intégration (planches graphiques) ?
Article	Ai-je rédigé les éléments éditoriaux de ma visualisation (titre, rubriques, boutons, bio, résumés, etc.) ?
Graphisme	Ai-je découpé mes images (icones, sous parties, charte) pour intégration (éléments graphiques au format) ?
Développement informatique	Ai-je réalisé mes développements (programme définitif) ?
Développement informatique	Ai-je fait des tests ?
Développement informatique	Ai-je fait valider par l'équipe le fonctionnement de mon programme avec l'équipe (recette utilisateur) ?
Bouclage	le graphisme illustre t il parfaitement le propos ?
Bouclage	le propos est il complètement en adéquation avec la représentation ?
Bouclage	Ai-je validé mon projet finalisé (BAT ou PV de recette) ?
Publication	Ai-je publié mon projet sur le site prévu ?
Publication	Ai-je libéré mes bases de données (démarche open data) ?
Bilan projet	Ai-je analysé les difficultés rencontrées et les points forts du projet afin d'être meilleur pour le prochain sujet ?

Annexe 4 Les 12 projets et les 12 équipes de travail du Data Journalism Lab

	Sujets et Journalistes	Développeur	Graphiste
1 G1	Visage social Gironde - Aymeric Parthonaud - Emmanuel Grabey - Charlotte Jousserand	Raphaël Estrach	Yohan Benazzouz
2 G1	Caméras surveillance - Marc Bourreau - Jean-Baptiste Bourgeon - Alix Mounou	Babacar Gueye	Adrien Colombié
3 G1	La réussite scolaire - France - Nicolas Canderatz - Audrey Chabal - Elodie Cabrera	Baptiste Rio	Remy Gendre
4 G1	Criminalité en Gironde - Marion Aquilina - Sophie Boutboul - Laetitia Volga	Baptiste Rio	Remy Gendre
5 G1	Elections Législatives 2ème - Anthony Jolly - Maxence Kagni - Pierre Garrat	David Bruant	Martin Caro
6 G1	Equipements sportifs - Clément Chaillou - Antoine Huot de St Albin - Boris Jullien	Babacar Gueye	Adrien Colombié
7 G2	Sujet : Cinéma - Olivier Mary - Elsa Landard - Adrian de San Isidro	Raphaël Estrach	Yohan Benazzouz
8 G2	Tour de France - Julien Chabrout - Cyril Domanico - Clément Lannuque	Quentin Riviere	Martin Caro
9 G2	Avortement - Manon Barthelemy - Joël Le Pavous - Rémy Demichelis	Maxime Chaisse-Leal	Adrien Colombié
10 G2	Déserts médicaux-France - Cécile Andrzejewski - Aurélie Simon - Julian Colling	Colin Delhalle	Remy Gendre
11 G2	L'endettement des communes de la CUB - Aurore Jarnoux - Clémence Bohême - Lélia de Matharel	Maxime Chaisse-Leal Davide Bruant	Martin Caro
12 G2	Génération Y - Sophie Lévy - Pauline Gleize - Laurie-Anne Virassamy	Colin Delhalle	Yohan Benazzouz

Annexe 5 Données publiques, open data et licences d'utilisation Note juridique

La production de data visualisations implique l'utilisation de données. Ces données sont protégées par le droit et, souvent, des licences d'utilisation leurs sont attribuées. Ces licences encadrent les possibilités d'utilisation des données par des libertés et des limites.

L'open data

Le mouvement Open Data prône une ouverture large, massive et immédiate des données, notamment des données publiques. L'objectif est de favoriser les réutilisations de données pour générer de nouvelles données et de nouveaux services. Au final, il s'agit de voir émerger des débouchés économiques basés sur l'utilisation des données libérées.

Fonctionnant à l'aide de licences d'utilisation (Étalab, ODBL, CC0, etc.), l'Open Data s'insère dans le mouvement du « Libre » qui prône l'utilisation libre des créations intellectuelles (logiciels, écrits, musiques, dessins, données, etc.) : libertés de copie, d'utilisation, de modification et de diffusion.

Appliqué aux données, le mouvement du Libre permet d'utiliser les données pour les modifier, générer des données « dérivées » et les diffuser librement. Les données dérivées sont des données créées ou obtenues à l'aide de différentes données sources.

NB : les licences d'utilisation de données ne prévoient pas de transfert de propriété. Ainsi, vous ne pouvez pas vous considérer comme propriétaire des données que vous utilisez. En revanche, vous êtes propriétaire des données dérivées que vous générez vous-même.

Les licences relatives aux données

Les données diffusées sans aucune licence affiliée sont soumises au droit commun : utilisation des données de façon très limitée. Pour les utiliser sans risques juridiques, il faut alors obtenir une autorisation auprès de leur producteur.

Certaines licences peuvent se ressembler mais ne contiennent pas forcément les mêmes préconisations. Il faut rester vigilant et bien lire le contenu de chaque licence pour comprendre quelles libertés sont offertes.

NB : Qu'elle que soit la licence qui accompagne les données libérées, il est obligatoire d'indiquer dans la data visualisation réalisée les noms de chacun des producteurs des données, le lien hypertexte vers le site source (ou catalogue) et la date de leur mise à jour.

Les licences Open Data : elles offrent de très grandes libertés d'utilisation et permettent de générer des données dérivées.

- L'ODBL – Open Data Base Licence : okfn.org
- L'ODC – Open Data commons : opendatacommons.org
- La CC0 – Creative Commons Zero : creativecommons.org/publicdomain/zero/0.1
- La licence Étalab – Licence Ouverte-Open Licence : etalab.gouv.fr, sribd.com/fullscreen/69252168

Les licences françaises : ces licences sont applicables en France. Plus particulièrement, la licence Étalab a été rédigée pour être utilisée selon les principes de l'Open Data.

- La licence Étalab** – Licence Ouverte-Open Licence : etalab.gouv.fr, sribd.com/fullscreen/69252168
- La licence APIE (Agence du patrimoine immatériel de l'État) : economie.gouv.fr/apie

- La licence IP (informations publiques) : www.rip.justice.fr/licence-type
- Les licences particulières : licences spéciales rédigées par les producteurs eux-mêmes. Elles sont généralement basées sur des licences qui existent déjà. Cependant, elles présentent des différences : plus de libertés... ou plus de contraintes et de limites. Il faut être vigilant !

Le choix d'une licence – préconisations

Il est préférable d'utiliser des licences Open Data sans les modifier. Elles fonctionnent toutes selon les mêmes principes et sont compatibles entre elles (interchangeables).

La licence Étalab a été rédigée dans ce but : elle respecte le droit commun, applique les vertus de l'Open Data et s'avère compatible avec les licences internationales les plus utilisées (ODBL et ODC). Elle est facile à comprendre et à appliquer. Enfin, elle est disponible en français et en anglais.

Lorsque l'on utilise des données libérées ou que l'on génère des données dérivées, il faut toujours indiquer leur origine : nom du ou de leurs producteurs, lien hypertexte vers le site source (ou catalogue) et date de dernière mise à jour (généralement, il s'agit de leur date de libération). Ces indications peuvent être mises en annexe de la base de données créée, à la suite des données, dans un cadre attaché aux données ou dans un fichier joint.

Les data visualisations

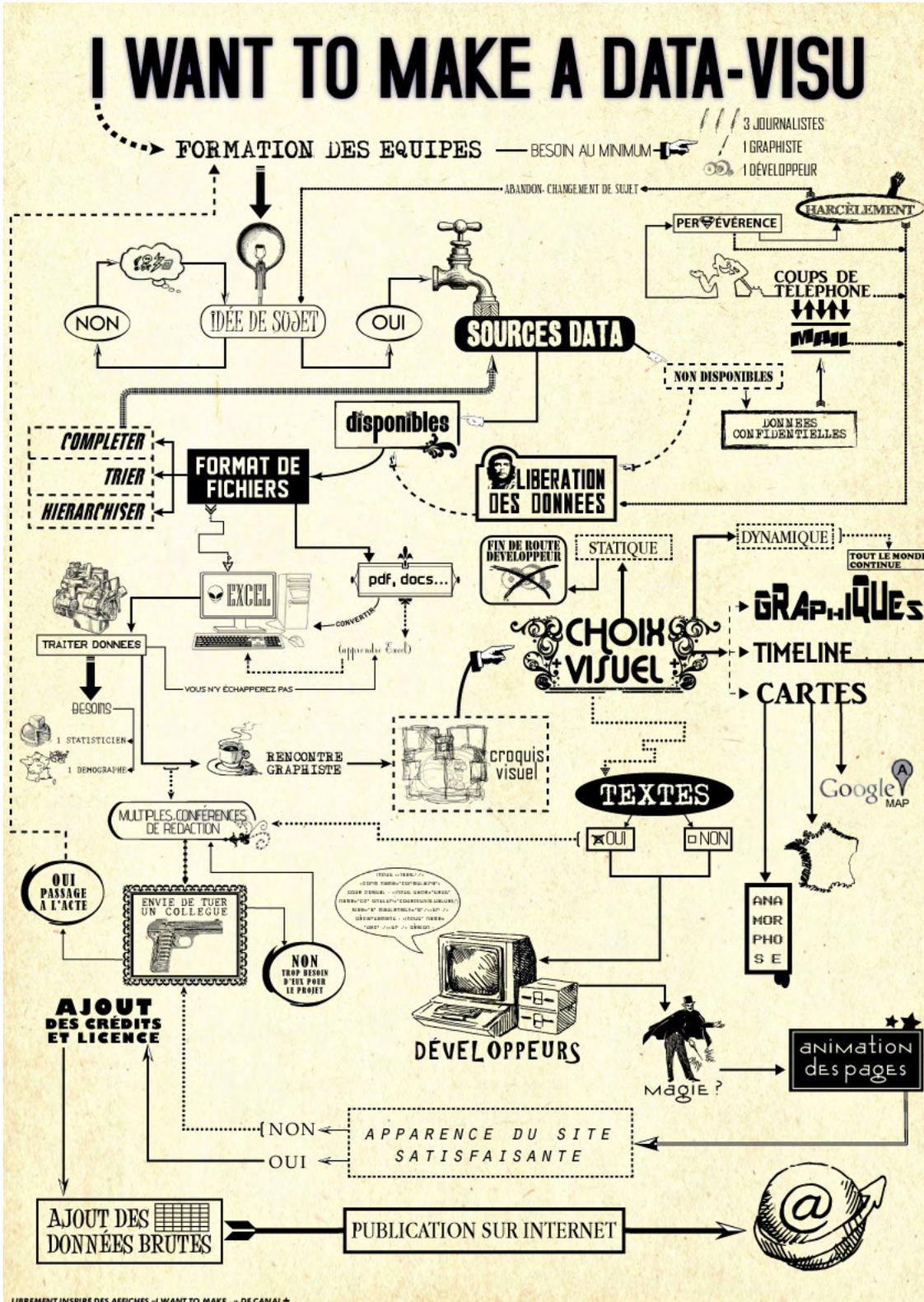
Les data visualisations ne sont pas des données... se sont des graphiques utilisant les données pour les « mettre en scène » et permettre leur interprétation. En cela, elles sont des créations dérivées basées sur les potentiels d'exploitation de données libérées.

Souvent, une data visu est conçue à partir de données libérées soumises à une licence d'utilisation. Il est essentiel d'identifier la licence afférente pour savoir qu'elles libertés d'utilisation et de modifications sont offertes... et n'utiliser que des données librement réutilisables. Toutes les données soumises à une licence Open Data entrent dans ce cadre : libertés d'utilisation permettant de générer des données dérivées et/ou des data visus.

Les data visus représentent des données... mais ne sont pas les données elles-mêmes ! Ainsi, si l'on souhaite les mettre à disposition de manière libre, il faut les soumettre à une licence issue du mouvement du Libre.

NB : chacun est peut utiliser la licence libre qu'il souhaite. Cependant, si l'on veut être fidèle aux mouvements du Libre et de l'Open Data, la licence CC By-SA peut être utilisée (creativecommons.org/by-sa/3.0/deed.fr) : libertés d'utilisation, d'adaptation et de modification des data visus en obligeant à indiquer son auteur. De plus, toute création dérivée devra être diffusée sous les mêmes conditions d'utilisation.

Annexe 6



Outils pour la data visualisation

Vous avez des données à explorer ? Voici quelques outils qui pourront vous être utiles pour les transformer en informations et en graphiques attrayants.

Source : <http://www.lemondeinformatique.fr/actualites/lire-22-outils-gratuits-pour-visualiser-et-analyser-les-donnees-1ere-partie-47241-page-3.html>

1/ Nettoyage de données, analyse statistique, outils et services de visualisation

- Cometdocs.com pour transformer un fichier .pdf en fichier .Excell

Google Docs intègre une fonction « feuilles de calcul » qui permet d'importer certaines parties de bases de données en ligne : inscrire dans la première cellule la formule ImportHTML suivie entre parenthèses et entre guillemets de l'adresse du site web, du numéro de la table et des numéros des colonnes à sélectionner, eux aussi entre guillemets.

- Google Refine

Il ressemble à un tableur pour examiner à la fois les données numériques et alphanumériques, mais à l'inverse du tableur, il ne permet pas d'effectuer des calculs. Comme Excel, il peut importer et exporter dans différents formats, incluant les fichiers tabulés, textes, Excel, XML et JSON.

Refine intègre plusieurs algorithmes retrouvant les mots orthographiés différemment mais qui devraient en fait être regroupés. Il y a aussi des options pour passer rapidement en revue les données numériques. Ces fonctionnalités peuvent pointer des anomalies pouvant résulter d'erreurs de saisie, telles que 800 000 dollars à la place de 80 000 dollars pour un salaire par exemple, ou mettre à jour d'autres incohérences. Inconvénient, si le jeu de données est volumineux, son examen peut prendre un certain temps. A noter que Refine propose aussi des outils de tri et de filtre. Niveau de compétences requis : débutant avancé. La connaissance des concepts d'analyse de données est plus importante que la prouesse technique. Les utilisateurs avancés d'Excel qui ont l'habitude du nettoyage de données devraient être à l'aise avec cet outil.

Google Refine fonctionne sur Windows, Mac OS X et Linux.

2/ Outils de développement, SIG, analyse de données temporelles, nuages de mots, visualisation de données relationnelles

Ces outils offrent différentes options de visualisation. Certains se cantonnent aux graphiques conventionnels (représentations sectorielles, histogrammes...), mais la plupart proposent un éventail de choix supplémentaires tels que les Treemap pour afficher les données hiérarchisées ou les nuages de mots. Quelques-uns disposent aussi de représentations géographiques.

- Google Fusion Tables : simple à utiliser et personnalisable

C'est l'une des plus simples façons de transformer des données en graphiques. On met en ligne son fichier (dans différents formats) et on choisit comment l'afficher : sous forme de tableau, de carte, de bargraphe, de camembert, de « heat map », de diagramme de dispersion, d'historique, d'animation... C'est assez personnalisable. Il est notamment possible de changer les icônes des cartes et le style des fenêtres d'information. Fusion Tables comporte aussi des fonctions d'édition de données, quoi que cela devient vite fastidieux dès que l'on commence à devoir modifier au-delà que quelques cellules. On peut aussi faire des jointures de table (important quand les données à afficher se trouvent dans différentes tables), filtrer, trier, ajouter des colonnes et commenter les données.

- Google Chart Tools : des graphiques statiques et interactifs

Contrairement à Google Fusion Tables qui est comme une application à part entière pour stocker des données en ligne et générer graphiques et cartes, Chart Tools est conçu pour visualiser des données résidant ailleurs, par exemple sur un site web ou au sein de Google Docs. Il offre à la fois les API Chart utilisant une simple requête URL vers un serveur de graphiques Google pour créer une image statique, et l'API Visualization qui accède à une bibliothèque JavaScript pour bâtir des graphiques interactifs. Avec celle-ci, on n'échappe pas à l'écriture de code. Google donne des informations (taille des données, compétences requises...) pour déterminer quelle option retenir.

Pour les graphiques statiques les plus simples, un assistant fournit quelques exemples. Il va jusqu'à aider à saisir les données ligne par ligne, quoi qu'à partir d'une certaine taille, il est plus logique de formater celles-ci dans un fichier texte.

Niveau de compétences : débutant avancé et expert.

Fonctionne sur tout navigateur.

En savoir plus : http://code.google.com/apis/chart/image/docs/making_charts.html

- Many Eyes : facile d'accès et bien documenté

Le projet Many Eyes d'IBM est l'un des pionniers de la visualisation de données sur le web. Il est très facile à utiliser et très bien documenté, incluant des suggestions sur le type de représentations qu'il convient d'utiliser suivant les cas. Many Eyes comprend plus d'une douzaine d'options de rendu, depuis les graphiques sectoriels, jusqu'aux nuages de mots, en passant par les treemaps, les diagrammes relationnels, plots, ainsi que quelques cartes géographiques.

Il faudra ouvrir un compte gratuit pour mettre ses données en ligne. Le formatage est basique : pour la plupart des visualisations, les données doivent être dans un fichier texte avec séparation par tabulateurs, comportant des têtes de colonnes sur la première ligne. Les résultats offerts apparaissent bien plus sophistiqués que ce que l'on aurait pu attendre au regard des efforts déployés pour les créer. La liste de visualisations possible s'accompagne d'explications pour déterminer celles qui sont les plus appropriées suivant le contexte. Inconvénient : vos jeux de données apparaissent publiquement sur le site Many Eyes et peuvent être facilement téléchargées, partagées, republiées et commentées par d'autres, ainsi que le projet soutenu par IBM encourage à le faire. Autre limite : la personnalisation est limitée, de même que la taille du fichier, qui ne peut pas dépasser 5 Mo.

Niveau de compétences : débutants.

Fonctionne sur les navigateurs Java et ceux pouvant afficher en Flash.

En savoir plus : sur Many Eyes

- Tableau Public : personnalisable et interactif

Transforme les données en diverses visualisations, simples ou plus complexes. Les champs peuvent être glissés/déplacés sur l'espace de travail et le logiciel peut alors suggérer un type de visualisation, puis personnaliser tous les éléments : étiquettes, infobulles, taille, filtres interactifs, légendes... L'un des intérêts de l'outil réside dans les différentes façons d'afficher des données interactives sur lesquelles un filtre de recherche peut agir sur de nombreux graphiques, diagrammes et cartes. Les tables sous-jacentes peuvent également être jointes. Et une fois que vous savez comment fonctionne le logiciel, le maniement de son interface « drag and drop » se fait bien plus rapidement que de coder en JavaScript ou en R. Ce qui incite à essayer davantage de scénarios avec les données. En outre, on peut facilement effectuer des calculs sur les données au sein du logiciel.

Dans cette version gratuite du logiciel de BI de Tableau Software, les données doivent résider sur le site de Tableau. La mise à jour vers la version desktop coûte environ 1 000 dollars. Par ailleurs, sans surprise, les fonctionnalités gratuites ont malgré tout un coût : le temps d'apprentissage comparé à, par exemple, Fusion Tables. Même avec l'interface drag and drop, cela prend plus d'une heure ou deux pour savoir se servir du logiciel, à moins de réaliser d'abord des graphiques simples.

Niveau de compétences : débutant avancé ou intermédiaire.

Fonctionne sur Windows 7, Vista, XP, 2003, Server 2008 et 2003.

En savoir plus : quelques vidéos de formation

