

Datajournalisme : Introduction

Lundi 4 octobre

Alice Palussière

Journaliste

palussierealice@gmail.com

Ana Lutzky

Journaliste

analutzky@gmail.com

*Design Informationnel et
Journalisme Transmédia (DIJT)*

Rédactrice en chef adjointe au #data

Ana Lutzky



analutzky@gmail.com



[@anouchka](https://twitter.com/anouchka)



[@anouchk](https://github.com/anouchk)



Ana#9955



Journaliste indépendante et
formatrice

Référente “Fil rouge” du Master
Journalisme Eco Data Investigation

CFJ — SCIENCES
PO LYON

Alice Palussière



Passée par



palussierealice@gmail.com



@journalice



Journalice#9606

Programme

Semestre 1
(4/10 au 15/10)

Lundi 4 octobre

- Le data journalisme, c'est quoi?
- Hands-on : Tableurs et autres spreadsheets
- Bonus : intro à la dataviz

Vendredi 8 octobre (AM)

- Bonus : intro à la dataviz
- Datawrapper
- Flourish

Lundi 11 octobre (AM)

- Tableau
- Infogram

Lundi 11 (PM), mercredi 13 (AM), vendredi 15 octobre

- Projet final en équipe



Avant toute chose

Créez-vous des comptes.

DATAWRAPPER

<https://www.datawrapper.de/>

FLOURISH

<https://flourish.studio/>

INFOGRAM

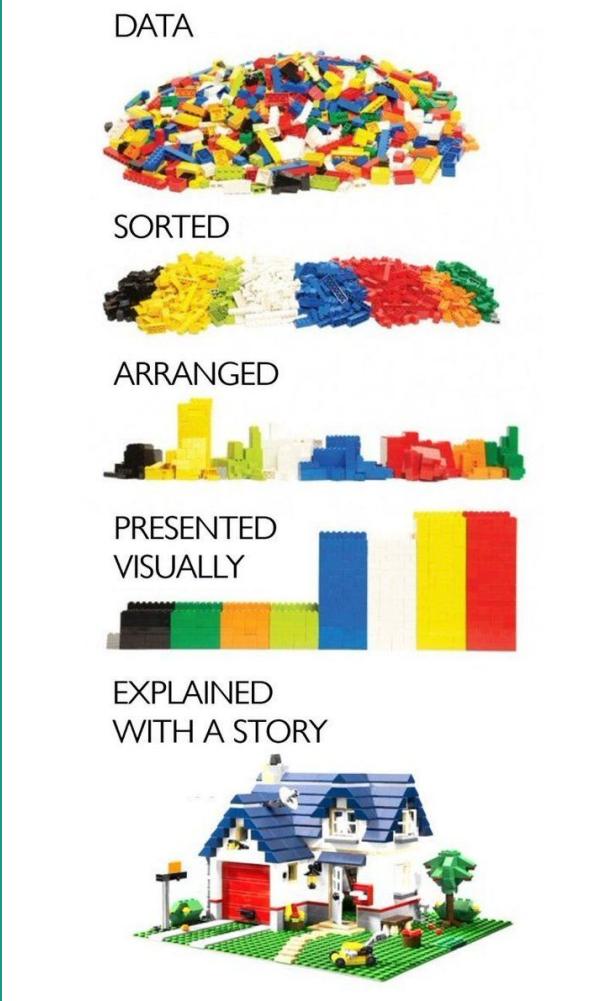
<https://infogram.com/>

TABLEAU

<https://www.tableau.com/academic/students>

Datajournalisme : quelle serait votre définition?

Le journalisme de données



Les définitions des experts



Paul Bradshaw
@paulbradshaw

Award-winning data journalist; MA Data Journalism + MA Multiplatform and Mobile Journalism bit.ly/1LFDK3T Books: leapub.com/u/paulbradshaw

Birmingham, UK [onlinejournalismblog.com](#)
A rejoint Twitter en février 2007

11,7 k abonnements 27,6 k abonnés

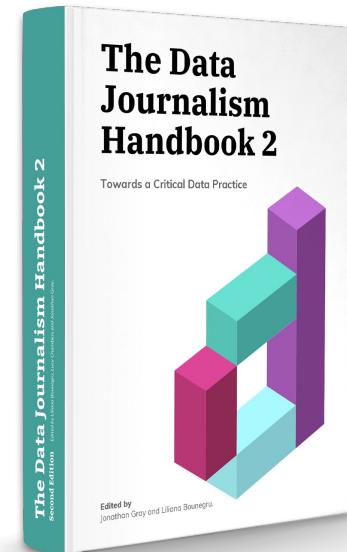
Suivre

“Data can be the source of data journalism, or it can be the tool with which the story is told—or it can be both”.

Les données peuvent être la source du journaliste ou l'outil avec lequel il raconte son histoire – ou les deux.

— Paul Bradshaw (Birmingham City University),
Data Journalism Handbook

<https://datajournalism.com/read/handbook/two>





"Au sens large, le journalisme de données c'est l'ensemble de processus de création d'un récit qui n'aurait pas pu être fait sans ordinateur. Tout comme le journalisme donne une voix aux sans-voix, le datajournalisme mesure ce qui n'est

pas mesuré. Les institutions officielles délaisse trop de problèmes qui nécessitent une approche systématique pour mieux les cerner et les comprendre. Le datajournalisme est là pour ça".

[Le site de l'agence Journalisme++](#)

Jan 7, 2016 10:36 AM



Cédric Motte

"C'est la capacité à trouver dans des chiffres de belles histoires à raconter. Ou de moches histoires, c'est selon. Mais en tout cas, des histoires. La visualisation est la cerise sur le gâteau mais n'est pas le but du journalisme de données ! "

[Découvrez le site de Cédric](#)

"Data journalism is the practice of finding stories in numbers and using numbers to tell stories"
(Howard, Art and Science, 2014).

"En général quand j'essaie d'expliquer aux gens ce que je fais, je finis par sortir mon téléphone et leur montrer directement".

*Jan Diehm, journalist engineer,
The Pudding*

"Le datajournalisme peut embarquer un ensemble très vaste : travailler sur des données pour nourrir une investigation purement textuelle, porter à la connaissance du public des informations d'intérêt général dans une visualisation interactive, ou encore faire du décryptage citoyen d'algorithme."

Alberto Cairo, professeur de dataviz à l'université de Miami et auteur d'ouvrages dont le très bon "How Charts Lie".

In Data Journalism, Tech Matters Less Than the People

Ben Casselman, an economics reporter, uses a programming language called R and works with vast data sets. But he says interviews still make for the best stories.

“Certaines personnes pensent que le “data journalisme” signifie regarder une feuille de calcul jusqu’à ce qu’un récit ou une information apparaisse comme par magie, mais en réalité, cela ne se produit quasiment jamais”.



The New York Times

Ben Casselman



Econ/business/data reporter for @nytimes. Formerly: @fivethirtyeight, @WSJ.

“Les meilleures informations naissent presque toujours à la faveur de discussions avec des personnes, qu'il s'agisse d'experts ou de simples personnes touchées par les questions sur lesquelles nous écrivons”.

La définition du datajournalisme

- Permet au lecteur de trouver des **informations qui le concernent personnellement**
- **Révèle une histoire** *remarquable* jusqu'alors passée inaperçue
- Aide le lecteur à **mieux comprendre un problème complexe**

Des exemples d'enquêtes réalisées avec des données

// Du flair : une enquête d'un étudiant en journalisme sur les précaires de la vigne

Saisonniers : les précaires de la vigne

Les étiquettes et le prestige des grands crus bordelais cachent une autre réalité. Celle de travailleurs pauvres qui (sur)vivent dans l'ombre des châteaux.

La Gironde est le premier vignoble de France. Mais passé la grille des châteaux, l'envers du décor révèle un spectacle beaucoup moins réjouissant. La vigne a fait germer une population de précaires : les saisonniers, ces hommes et ces femmes employés dans les vignobles pendant les pics d'activité. Petites mains au dur labeur, ils ne travaillent souvent que quelques mois dans l'année et subsistent le reste du temps grâce aux aides sociales.

Le RSA dans les campagnes, face cachée de la viticulture

<https://bit.ly/3zOMkN9>

// Du participatif très malin : les frais de mandat des sénateurs argentins, par La Nacion

Gastos del Senado 2013 (Terminado y Datos Abiertos)



Con más de 3000 documentos sobre los gastos del Senado argentino de 2013, este segundo proyecto de VozData invita a clasificar las rendiciones de cuentas de senadores y de las distintas dependencias de la Cámara alta.

- Ver video TUTORIAL;
- Testimonio de M. Baron (Dir. Legislativo);
- Testimonios de colaboradores de Gastos del Senado 2010-2012.
- DATOS ABIERTOS / OPEN DATA

- Ver notas sobre gastos del Senado

Ranking de adjudicatarios [ver mas](#)

1	Honorble Senado de la ...	\$ 21.168.515
2	Servicio de comedor - Ho...	\$ 1.377.129
3	Telecentro S.A.	\$ 501.724

Ranking de rubros [ver mas](#)

1	Viáticos y gastos	\$ 9.706.417
2	Caja chica	\$ 4.909.171
3	Gastos de protocolo y cer...	\$ 4.371.391

[Acerca](#) | [Preguntas frecuentes](#) | [Aviso legal](#) | [Contacto](#)

[Equipos](#) | [Login](#)

Liberá un documento

¡Ya revisamos 3049 de 3049 documentos!

Gastos del Senado procesados: \$ 31.271.382

Compartir

0 [weet](#)
[Compartir](#)

Enviar por e-mail
[enviar](#)

Ranking de usuarios [ver +](#)

Por cantidad de documentos revisados

	berbas	1601
	agustinlorenzo93	1000
	MICMSR	700
	jaelariadna	558

<https://bit.ly/3iyGOJ1>

// Du tact : aborder les violences obstétricales en évoquant les statistiques des maternités

LES DÉCODEURS

Partage



Episiotomie, césarienne, allaitement : comment accouche-t-on en France en 2017 ?

Si les taux d'épisiotomies et de césariennes diminuent, les facteurs de risques chez les mères tels que l'obésité ou le tabagisme augmentent, d'après une étude de l'Inserm et de la Drees.

Par Laura Motet et Anne-Aël Durand

Publié le 12 octobre 2017 à 14h09 - Mis à jour le 14 mai 2018 à 15h25 • ⏲ Lecture 7 min.

<https://bit.ly/3EY3SKG>

// De la tenacité : l'enquête du Boston Globe sur la pédophilie dans l'Eglise primée par le prix Pulitzer



La cellule investigation du Boston Globe, nommée Spotlight, a fait une recherche méthodique dans un registre annuel de l'Église qui liste les prêtre transférés ou absents sans raison explicite. Qu'ont fait les journalistes ? Ils ont construit une base de données !

// Panama Papers : faire bouger les lignes



Des millions d'euros récupérés
Monde : 1,15 milliard d'euros
**France : 126 millions d'euros +
50 redressements** (issus de 519
contrôles fiscaux)



The screenshot shows the Le Monde website's header with the logo and navigation menu (Home, ACTUALITÉS, ÉCONOMIE, VIDÉOS, OPINIONS, CULTURE, M LE MAG). Below the header, the article title is 'EVASION FISCALE • PANAMA PAPERS « Panama Papers » : cinq ans après, des milliards récupérés et plusieurs condamnations'. The text discusses the ongoing investigations and convictions following the operation. The author is Will Fitzgibbon, Anne Michel, Maxime Vaudano, and Jérémie Baruch. It was published on April 8, 2021, and last updated on April 9, 2021. A pink curved arrow points from the text above to this article.

// Qui a même été adapté en
série par Steven Soderbergh
(The Laundromat - La laverie ; Netflix)





**DATA+
LOCAL**

Travail collaboratif

Le Parisien



Économie

#TransparenceCHU : comment nous avons enquêté sur les liens entre labos et médecins

Quinze titres de presse quotidienne régionale, dont Le Parisien, publient ce vendredi une enquête sur les liens entretenus entre les laboratoires et les professionnels de la santé.

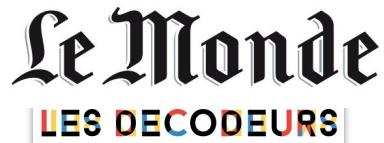


Enquête sur les liens entretenus entre les laboratoires et les professionnels de la santé

Les rédactions qui l'utilisent en France et dans le monde

Datajournalisme : dans quels médias?

Presse écrite et web en France



la montagne

Les Echos



le dauphiné
libéré

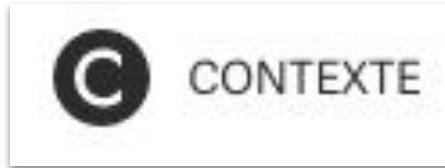


franceinfo:



dataspot
Le Télégramme

Agences de presse et agences spécialisées

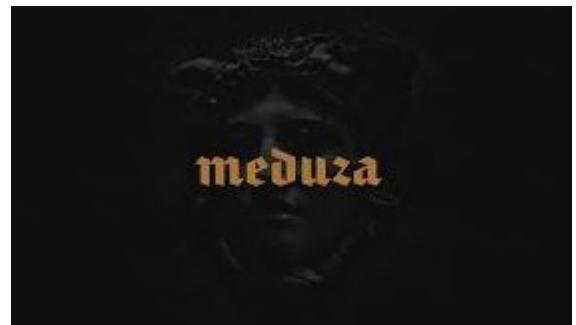


WEDODATA



BRONX.

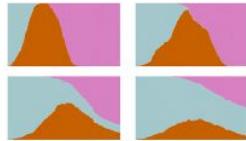




Des enjeux multiples

- Débusquer **fake news** et fausses déclarations (**Fact checking**)
- Trouver de **nouveaux angles** (ex: réseaux sociaux & manifestations..)
- Gestion de l'open data (mise à disposition de données pour les citoyens, et non pas une sélection communicationnelle de ces données par les administrations!)
- Une façon de **rester indépendant.e** et de moins dépendre du storytelling des communicants (aller chercher ses données à son propre rythme...)
- **S'extraire de l'actu chaude** (mais chronophage)
- Faire partie d'une communauté (entraide, partage de connaissances...)
- Intérêt visible lorsqu'il permet de faire des "**gros coups**" qui vont **booster la fréquentation du site/ l'audience** : ex les « Panama Papers »

Une animation graphique devenue un blockbuster



Santé

Pourquoi des épidémies comme
celle du coronavirus se propagent
de manière exponentielle et
comment “aplatir la courbe”?

Par [Harry Stevens](#) Le 17 mars 2020

<https://wapo.st/3D1zCgf>

- [The backstory behind The Washington Post's most-read article](#)
- [The Most-Viewed Washington Post Article Ever | Stats + Stories](#)
- [Le code en D3.js](#)

Tout est donnée

- Nous vivons aujourd’hui dans un monde numérique dans lequel **pratiquement tout peut être** (et est *de fait*) **décrit par des chiffres**. Tout ce qui peut être enregistré, consigné… est “donnée”.
- Analyse des tendances, des changements climatiques, évolutions sociopolitiques, résultats électoraux, discours : **tout est potentiellement exploitable**, autant de données qui peuvent, une fois mises en forme et présentées, constituer un **angle** ou un **récit**.
- En utilisant les données (pas exploitables telles quelles) pour relier les points entre eux, les **datajournalistes donnent sens** à de petits éléments d'information qui, pris isolément, ne seraient pas pertinents.



Les **éléments de base de la profession** (chercher, collecter et publier des informations) restent les mêmes.

L'exemple des balles de tennis

Que peut-on en dire des données sur cette photo?

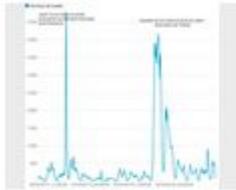


- Utilisées pour le tennis = sport
- Couleur : jaune
- Taille : 6,35 cms
- Nombre : 21
- Etat : neuf
- Valeur : 7 €
- Marque : Technifibre

etc...

➡ Tout cela constitue des “données” exploitables

Récupérer des données des réseaux sociaux : La loi ORE (Parcoursup) contestée en 2018



Mobilisation étudiante : le mouvement sur Twitter se cristallise autour du hashtag #Tolbiac

Publiée le 26/04/2018 - dépêche n° 584729



Mobilisation étudiante : ce que nous apprend le hashtag #NonALaSelection sur Twitter

Publiée le 06/04/2018 - dépêche n° 583592

Plusieurs types de données...

L'Open data

Données publiées par des institutions publiques ou privées et sont partagées dans le but d'être réutilisées.

“Open” : implique que ces données soient à la disposition de tous.

Elles sont accompagnées d'une licence qui encadre la manière dont on a le droit de les utiliser.

Les données “sauvages”

Informations non disponibles et non visualisables par le grand public (que les journalistes doivent informer), hors Open Data donc.

On peut parler de données « sauvages » que le journaliste doit se procurer à sa manière en scrappant, en “off”...).

Les données “manuelles”

Données récoltées par ses propres moyens.

Ex : *Le Hindustan Times* à Delhi a mis en place des capteurs de pollution atmosphérique pour pallier la piétre qualité des mesures officielles.

Ou simplement à l'aide d'un sondage ou d'un questionnaire en ligne. Ou même avec un téléphone!

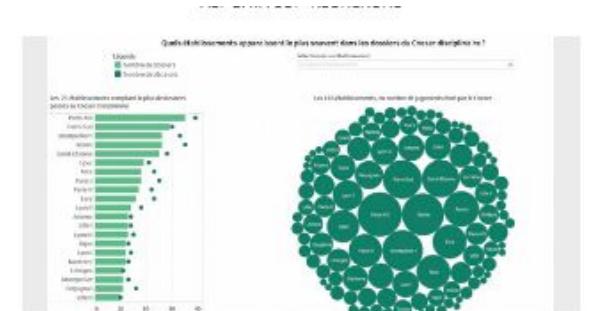
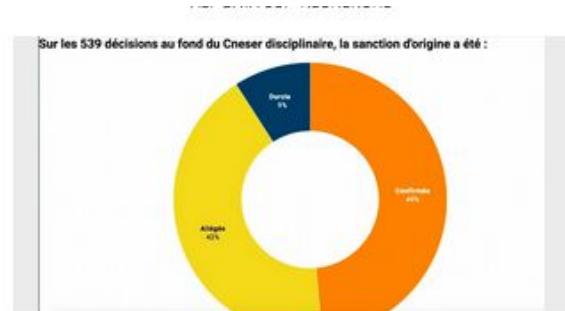
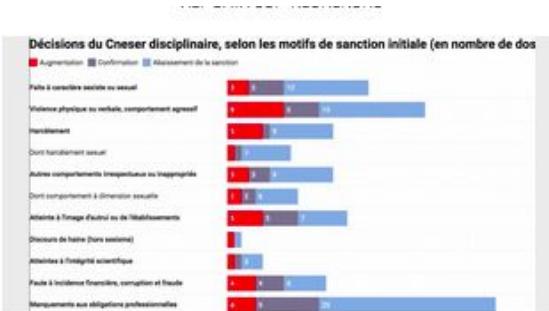


Attention à toujours recouper ses données



Texte-> tableur. L'exemple des décisions de justice / d'instances disciplinaires

Enquête AEF sur l'enseignement supérieur



1 jeu de données “maison”, 5 papiers publiés

- Qui sont les 14 membres du Cneser disciplinaire depuis 2011 ?
- Qui sont les personnes jugées, pour quels faits et dans quels établissements ?
- 42 % des décisions allègent la sanction de première instance
- Dans un tiers des cas concernant les enseignants-chercheurs, la relaxe est prononcée
- Plus d'une décision sur deux a été cassée par le Conseil d'État

Le jeu de données :

<https://www.data.gouv.fr/en/datasets/decisions-du-cneser-disciplinaire-2008-2019/>



833 lignes et 13 variables :

1. Numéro et date du BO
2. Dossier enregistré sous le(s) numéro(s)
3. Année de la décision
4. Date de la décision
5. Établissement concerné (le nom a été conservé tel qu'au moment des faits)
6. Personne visée (étudiant, maître de conférences, professeurs des universités...)
7. Type de fait incriminé (voir plus bas)
8. Décision
9. Principaux éléments fondant la décision
10. Nature de la décision (décision rendue au fond, sursis à exécution, désistement d'appel...)
11. Lien vers la décision

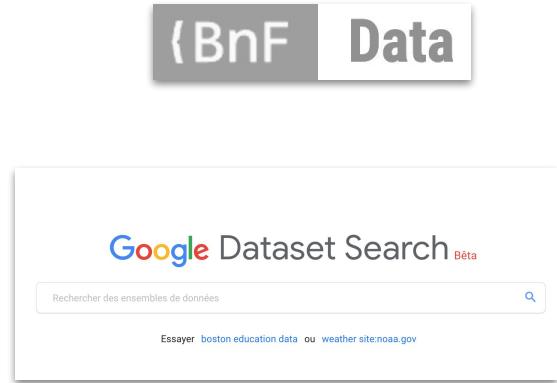
La variable "type de faits incriminés" comporte 13 valeurs correspondant au code suivant :

- 1 Violence physique ou verbale, agressions, menaces
- 2 Harcèlement
- 3 Diffamation, atteinte à l'image d'autrui ou de l'établissement (+ manquements à l'obligation de réserve)
- 4 Autres comportements irrespectueux ou inappropriés
- 5 Faits ou propos à caractère de haine
- 6 Faits ou propos à caractère sexistes et/ou sexuels
- 7 Atteinte aux biens d'autrui ou de l'établissement
- 8 Atteinte à l'intégrité scientifique et/ou à la propriété intellectuelle
- 9 Fraude à un examen
- 10 Autres fraudes et falsifications de document
- 11 Faits à caractère financier, corruption
- 12 Manquements aux obligations professionnelles
- 13 Autres atteintes à l'ordre et au bon fonctionnement de l'établissement (+ fautes administratives)

Ces catégories ne sont pas exclusives et sont séparées par des virgules lorsqu'une affaire correspond à plusieurs catégories. Par exemple, une affaire de harcèlement sexuel est répertoriée par les codes 2,6.

Trouver les données

Quelques grandes plateformes...



Quelques portails spécialisés

- **Insee** (Institut national de la statistique et des études économiques)
- **Drees** (Direction de la recherche, des études, de l'évaluation et des statistiques)
- **Mesri** (Ministère français de l'Enseignement supérieur, de la Recherche et de l'Innovation)
- **CGDD** (Commissariat général au développement durable)
- **Santé publique France**
- **SNCF, RATP...**



Dans le monde

Banque mondiale

Eurostat (Commission européenne)

Fao (Onu)

OCDE (Organisation de coopération et de développement économiques)

USGS (U.S. Geological Survey)

et bien d'autres...



Des données essentielles et démocratiques

La réserve parlementaire

Ensemble de subventions du budget de l'État qui permettait aux députés et sénateurs de financer des associations et des collectivités de leur circonscription (jusqu'en 2018).

http://www.senat.fr/dotation_daction_parlementaire/tableau.html

Haute Autorité pour la transparence de la vie publique

Publie les déclarations de situation patrimoniale et les déclarations d'intérêts de certains responsables publics (membres du Gouvernement, candidats à la présidentielle, députés...)

<https://www.hatvp.fr/>

La base de données publique Transparence - Santé

Liens d'intérêts que les entreprises entretiennent avec les acteurs du secteur de la santé (médecins...)

<https://www.transparence.sante.gouv.fr/flow/main?execution=e10s1>

Recherche par bénéficiaire

Critères de recherche

Noms (1) DARAI ;

Déclaration entre Semestre 1 - 2016 et Semestre 1 - 2021

Catégorie de bénéficiaires

Toutes



[Recherche avancée](#)

Cuez les cases de un ou plusieurs bénéficiaires puis cliquer sur le bouton « Valider » en bas de page

Cuez les cases de un ou plusieurs bénéficiaires puis cliquer sur le bouton « Valider » en bas de page pour afficher leurs déclarations.

12 Résultat(s) concernant les professionnels de santé

	RPPS (1)	Bénéficiaire	Code postal	Ville	Adresse	Profession
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	HOPITAL TENON (AP-HP) 4 RUE DE LA CHINE	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	MATERNITE GYNECOLOGIE	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	HOPITAL TENON	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	94400	VITRY-SUR-SEINE	58 voie lesueur	Médecin
<input type="checkbox"/>	10000543826	DARAI Emile	75020	PARIS	4 rue de la Chine	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS	4 RUE DE LA CHINE	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75012	PARIS	184 RUE DU FAUBOURG SAINT ANTOINE	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS	HOPITEL TENON	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	HOPITEL TENON	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS 20	HOPITEL TENON	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS	4, RUE DE LA CHINE	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	4 RUE DE LA CHINE	Médecin

12 Résultat(s) concernant les professionnels de santé

	RPPS (1)	Bénéficiaire	Code postal	Ville	Adresse	Profession
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	HOPITAL TENON (AP-HP) 4 RUE DE LA CHINE	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	MATERNITE GYNECOLOGIE	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	75970	PARIS CEDEX 20	HOPITAL TENON	Médecin
<input type="checkbox"/>	10000543826	DARAI EMILE	94400	VITRY-SUR-SEINE	58 voie lesueur	Médecin

Github



github.com

github

Incontournable pour les programmeurs, GitHub est une plateforme collaborative créée en 2008.

Les utilisateurs y déposent les codes qu'ils souhaitent partager. Chacun peut ensuite les consulter, les télécharger ou les enrichir.

Pour les médias, GitHub permet de mettre à disposition du public les données utilisées pour des productions, dans une démarche de transparence. Les Décodeurs ont ainsi un compte où figurent en libre accès le code source du Décodex et celui de leur tableau de bord de la présidentielle 2017.

Des questions ?

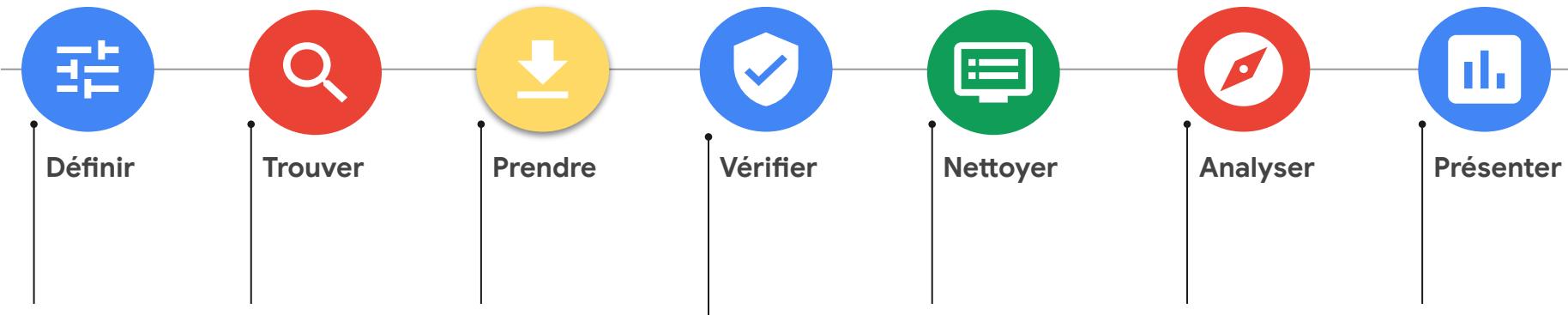
5 minutes de pause



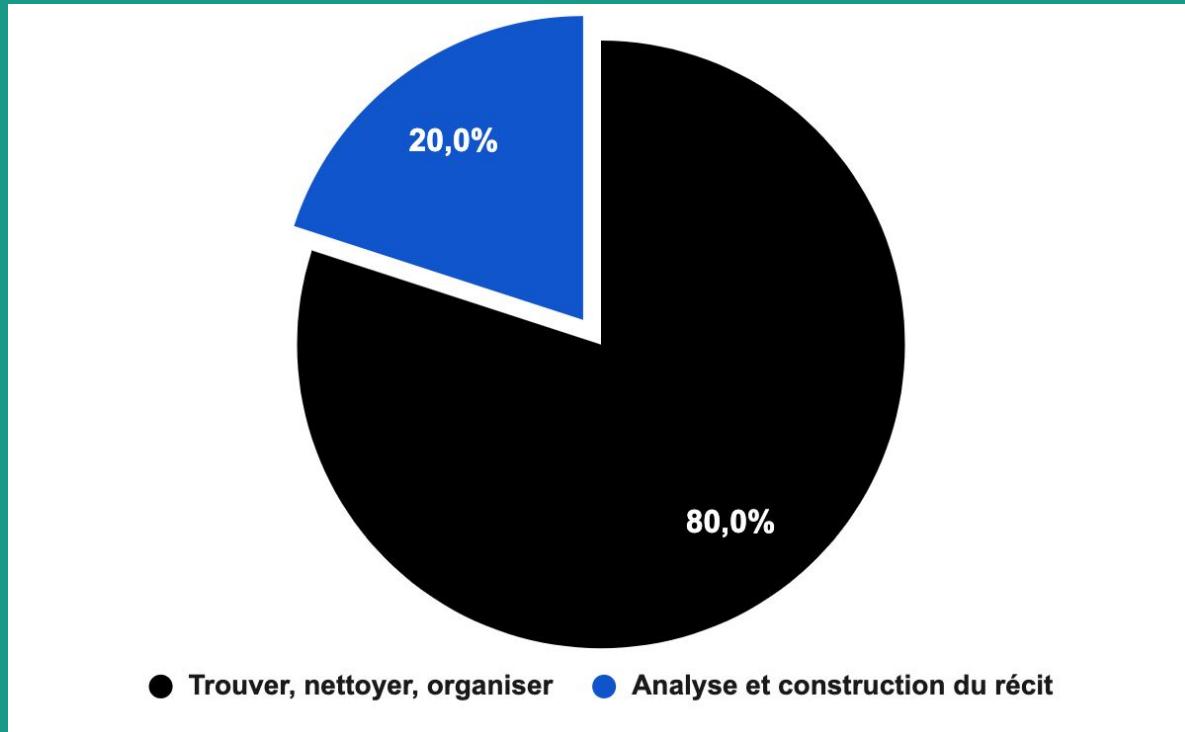
Le processus, les étapes clés

Flux de données

Les étapes essentielles pour élaborer des récits avec des données



Le travail du datajournaliste

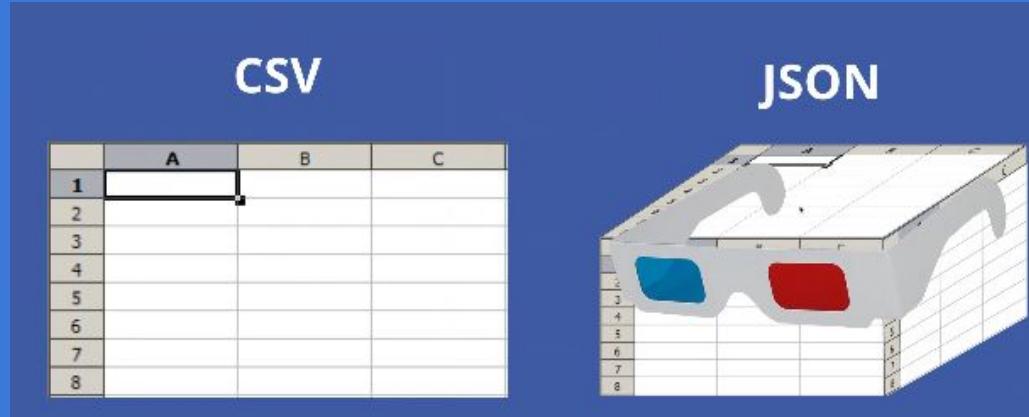


Datajournalisme : Quels sont les formats de stockage de données que vous connaissez?

Les formats de stockage des données

- tabulaires
 - .xlsx
 - .csv
- hiérarchiques
 - .json
 - .xml
- géographiques
 - .shp

Les formats



On peut convertir

```
<city>
  <cityname>Amsterdam</cityname>
  <country>nl</country>
  <population>850000</population>
</city>
```

```
city,country,population
Amsterdam,nl,850000
```

Plusieurs façons de json->csv

```
[  
  {  
    "city": "Amsterdam",  
    "country": "nl",  
    "population": [  
      {  
        "year": "2019",  
        "amount": "850000"  
      },  
      {  
        "year": "2014",  
        "amount": "822000"  
      }  
    ]  
  },  
  ...  
]
```

city	country	population_2019	population_2014
Amsterdam	nl	850000	822000

Tidy data

city	country	year	population
Amsterdam	nl	2014	822000
Amsterdam	nl	2019	850000

Qui a déjà expérimenté :

-Excel : *tapez 1*

-Spreadsheet (google) : *tapez 2*

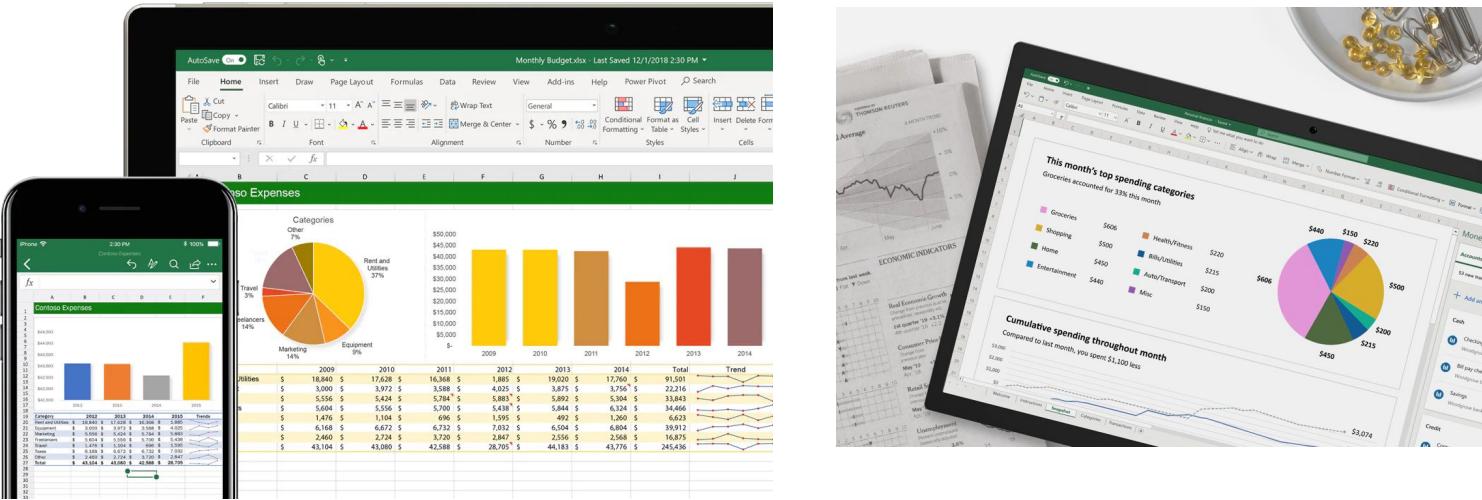
Les tableurs : Google Sheets & Excel

Excel

Excel



microsoft.com/fr-fr/microsoft-365/excel



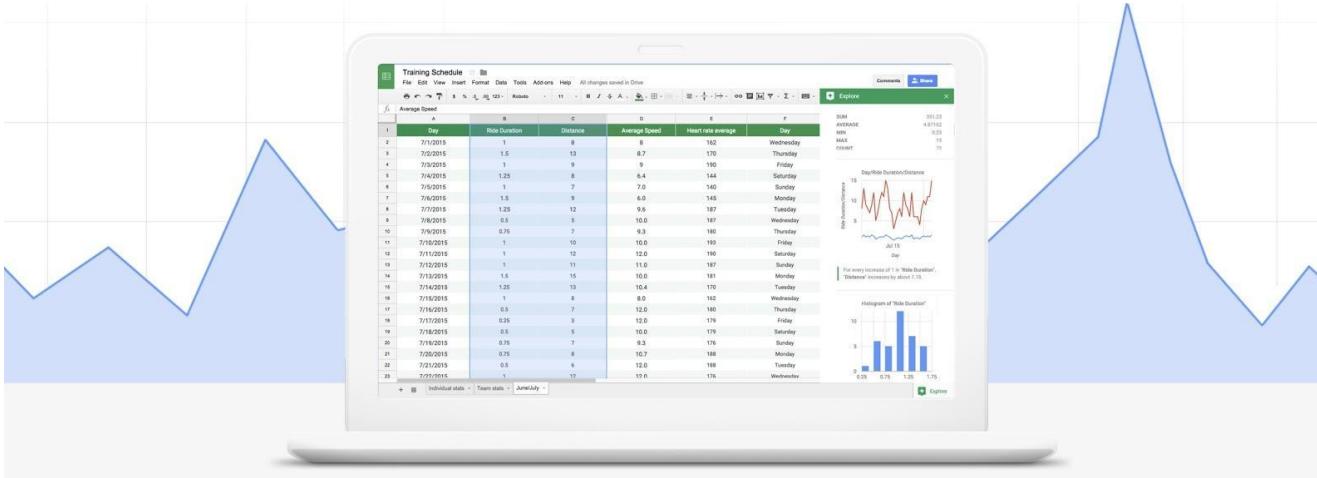
Payant, intégré à abonnement Microsoft 365 (parfois installé d'office sur les PC). Abonnements étudiants et packs entreprises possibles.

Google Sheets

Google Sheets



docs.google.com/spreadsheets



Gratuit, intégré à la suite Google Workspace, interface similaire à Excel, publication facile et partage contrôlé, interactif.

Sheets

- Ergonomique et facile d'utilisation
- "Cloud-based" : modifications automatiquement sauvegardées
- Modification visibles en temps-réel (+historique des révisions)
- Collaboratif (travail en équipe, chat..)
- Partage personnalisable (lecteur, éditeur..)
- Accès depuis n'importe quel appareil de n'importe où
- Gratuit

vs

Excel



Le match

- Puissance pour traiter des jeux de données volumineux (+ de 1000 lignes)
- Version "Cloud-based" possible mais payante
- Accès hors ligne performant (et synchro sur onedrive)
- Fonctions et formules + importantes et performantes (mises en forme conditionnelles..)
- Payant (au sein du pack office), mais largement utilisé dans les entreprises

Hands-on



Le jeu de données à copier dans une googlesheet à vous :

shorturl.at/ehryl

Source : open data du ministère de l'Enseignement supérieur et de la recherche

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION
L'enseignement
supérieur
et la recherche

OPEN DATA

Accueil | Explorez les jeux de données | Note

2 866 284
enregistrements

Aucun filtre actif

Filtres

Rechercher...

Année universitaire

Année universitaire	Nombre d'enregistrements
2020-21	427 812
2019-20	414 134
2016-17	409 809
2015-16	406 308
2017-18	406 306
2018-19	405 110

Effectifs d'étudiants inscrits dans les établissements publics sous tutelle du ministère en charge de l'Enseignement supérieur

Informations Tableau Tableau

Ce jeu de données est sous licence : Licence Ouverte

Formats de fichiers plats

CSV Jeu de données entier
Le CSV utilise le point-virgule (;) comme séparateur.

JSON Jeu de données entier

Excel Jeu de données entier

Effectifs d'étudiants inscrits dans les établissements publics sous tutelle du ministère en charge de l'Enseignement supérieur

shorturl.at/istyH

Deux mantras

- **On respire.** Même pas peur.
- **Google est ton ami** (comment ça, quelles données perso). Si on peut taper “comment faire un joint de salle de bain”, on peut aussi taper “comment chercher/remplacer dans google spreadsheet”.



Lire les données : une ligne, c'est quoi ?

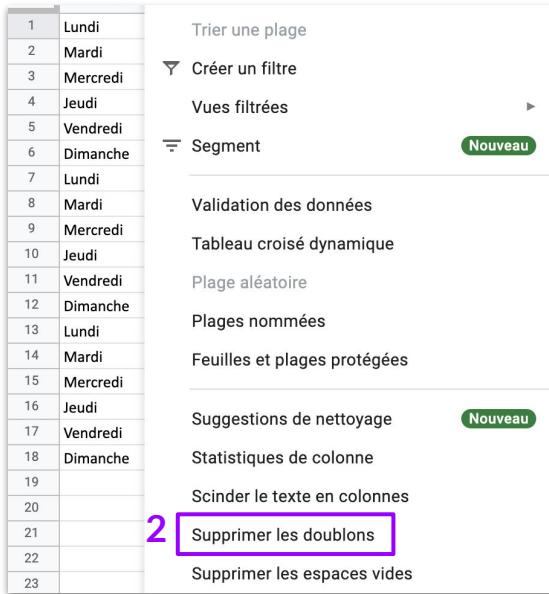
- Chaque ligne est une combinaison unique de toutes les variables (= colonnes). 3 000 lignes = 3 000 combinaisons possibles.
Pour la combinaison femmes, ayant eu un bac S, et poursuivant en DUT de gestion, etc., il y a 39 inscrits à Valenciennes par exemple.

Diplôme	Niveau dans le diplôme	GD_DISCIP LINE	Grande discipline	Discipline	Secteur disciplinaire	Spécialité de DUT	Nombre d'étudiants inscrits (inscriptions principales) hors étudiants inscrits en parallèle en CPGE	Nombre de nouveaux bacheliers inscrits hors étudiants inscrits en parallèle en CPGE
Autres formations	Sans précision	LLSH	Lettres, langue	Lettres, science Arts		Sans objet	3	0
Autres formations	Sans précision	LLSH	Lettres, langue	Lettres, science Arts		Sans objet	1	0
Autres formations	Sans précision	LLSH	Lettres, langue	Lettres, science Arts		Sans objet	1	0
Autres formations	Sans précision	LLSH	Lettres, langue	Lettres, science Arts		Sans objet	2	0
Diplôme universitaire 1ère année		DSA	Droit, sciences	Sciences écono	Sciences de ge	Techniques de	4	1

Tri et nettoyage : les bases

- Chaque colonne doit avoir UN seul type de données (1)
- La première ligne indique normalement le nom de la colonne (1)
- Enlever les lignes “doublons” (2)
- Attention aux fautes de frappes, à la typo..(3)
- Vous pouvez faire une copie du fichier source pour ne pas faire d'erreur dedans (4)

1	A	B	C	D	E	F	G
2	Civilité	Noms	Prénoms	Type d'instance	Commissions	Semestre	Date séance
	Madame ACS	Nathalie	Commission	C8 - Affaires fa...	1	19/4/2021	



Hauts-de-France
Normandy
Pyrenees



Apprivoiser les données : commencer par un filtre



- ça permet de voir, pour chaque variable (= colonne), les différentes valeurs (= contenus des cases) possibles

Sexe de l'étudiant	
F	
M	

- en sélectionnant “F”, on peut filtrer le jeu de données pour ne prendre que les lignes où les étudiants sont des femmes

Trier et filtrer les données : les calculs et manipulations de base

- Rechercher et remplacer (espaces superflus...)
- Supprimer les doublons
- Additionner une colonne ou une ligne de chiffres
- Déterminer la moyenne d'une colonne
- Calculer un pourcentage

Les symboles de calculs

- Dès qu'on veut faire une opération, on commence par écrire le signe égal (=) dans une case qui indique à l'outil que l'on va faire un calcul
- Le signe plus (+) pour ajouter un nombre à un autre
- Le signe moins (-) pour soustraire un nombre à un autre
- L'astérisque (*) pour multiplier un nombre à un autre
- Le backslash (/) pour diviser un nombre par un autre

...Suivi des lettres/nombres des cellules dans lesquelles vous voulez faire le calcul (A1, A2, B1, B2, etc.), séparés par le symbole du type de calcul.

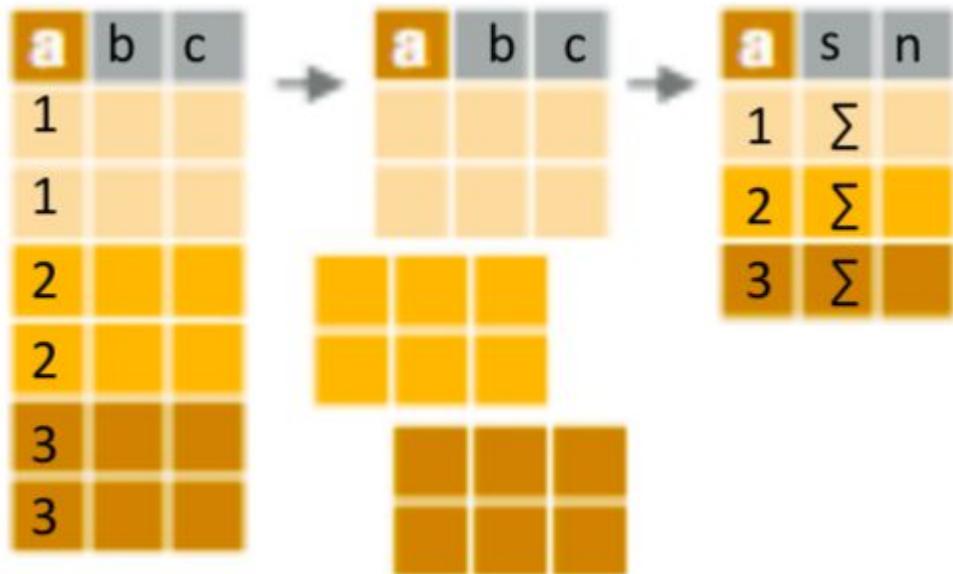
Le tableau croisé dynamique (pivot table en anglais)



L'arme nucléaire : le **tableau croisé dynamique** (ou “pivot table”)



- l'expression “group by” est assez parlante



L'arme nucléaire : le tableau croisé dynamique (ou “pivot table”)



Données Outils Modules complémentaires Aide

Trier la feuille à partir de la **colonne X**, A → Z
Trier la feuille à partir de la **colonne X**, Z → A

Trier une plage

Créer un filtre

Vues filtrées

Segment Nouveau

Validation des données

Tableau croisé dynamique

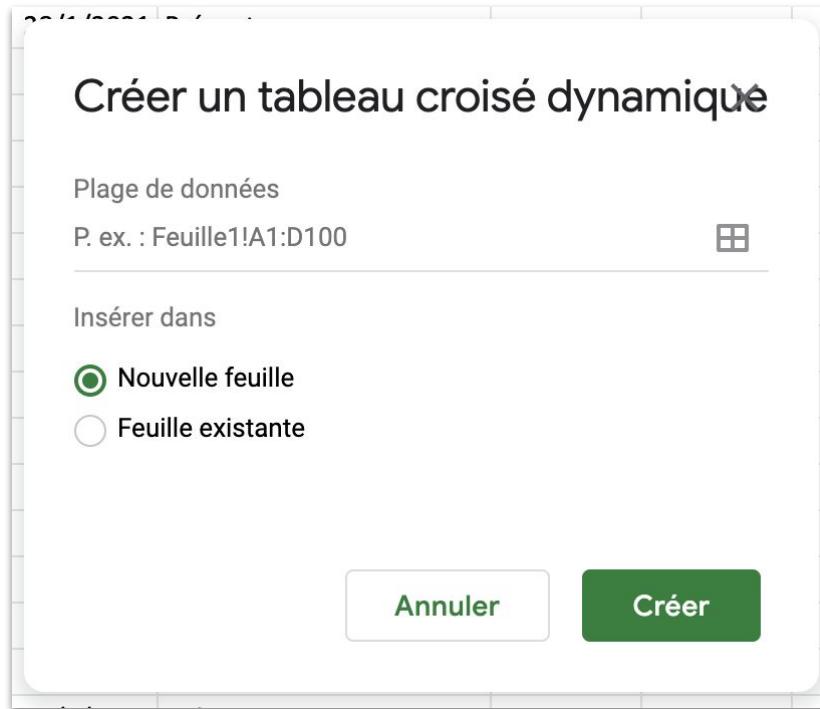
Plage aléatoire

Plages nommées

tableau croisé dynamique



L'arme nucléaire : le **tableau croisé dynamique** (ou “pivot table”)



L'arme nucléaire : le **tableau croisé dynamique** (ou “pivot table”)



quelle variable je veux
en catégories de lignes ?

en colonnes ?

et comme contenu dans
mes cases ?

Sheet1!A1:Y3536

Suggested

- Number of unique Regroupement de diplômes for each Discipline
- Number of unique Diplôme for each Spécialité de DUT
- Number of unique Secteur disciplinaire for each Série de baccalauréat

Rows Add

Columns Add

Values Add

Filters Add

Search

- Décomposition d...
- Sexe de l'étudiant
- Regroupement d...
- Série de baccala...
- Âge au baccalaur...
- Attractivité dépar...
- Attractivité régio...
- Attractivité intern...
- Mobilité internati...
- Type de diplôme
- CURSUS_LMD
- Regroupement d...
- Diplôme

L'arme nucléaire : le **tableau croisé dynamique** (ou “pivot table”)



Exerçons-nous :

- Combiens de femmes parmi les étudiants de Valenciennes ?
- Quelle est la part des DU ?
- Quelle série de bac ont fait les étudiants de Staps ? Droit ? Sciences ? SHS ?

L'arme nucléaire : le **tableau croisé dynamique** (ou “pivot table”)



A l'UPHF, 37% des étudiants en cursus santé en 2020-2021 ont fait un bac techno

Autres séries technologiques Baccalauréat professionnel Dispensés Série ES Série L
Série S Séries STMG (ex STG et STT)



Des questions ?

Quelques conseils et règles essentielles

**Organiser les données = les
rendre utiles pour nous**

Réfléchir en amont à la source de données

- Quel est le sujet /la problématique/les données **dont j'ai besoin?**
- Quel serait l'organisme/le ministère/l'association qui pourrait **disposer de ces données** (responsabilité, tutelle, fédération...)?
- Réfléchir au **financement** : cela peut indiquer un biais dans les données..
- Se “nourrir” sur son sujet de recherche (articles, émissions...) pour voir **quelles sources de données ont déjà été utilisées**, choisir d'y fouiller davantage ou en choisir d'autres...

Ne pas oublier de lire les notes/la méthodo!

45	
46	* Voir la méthodologie dans le paragraphe <i>Sources</i> .
47	Lecture : à long terme, lorsque le prix relatif des biscuits, biscuits et pâtisserie de conservation augmente (respectivement diminue), le volume consommé baisse fortement (respectivement augmente fortement).
48	Source : Insee, comptes nationaux base 2010, calcul des auteurs.
49	

11) Montants recouvrés : cf. tableau 102 à compter de 2010

12) Nombre de redevables ayant déposé et montants recouvrés : cf. tableaux 301n et 103 à compter de 2010

13) À partir de 2004, le montant de l'impôt sur le revenu comprend les titres courants, les titres antérieurs et la contribution sur les revenus locatifs.

À partir de 2004, le montant de la taxe foncière sur les propriétés bâties comprend en plus la taxe d'enlèvement des ordures ménagères

À partir de 2004, le montant de la taxe foncière sur les propriétés non bâties comprend en plus la taxe pour frais de chambre d'agriculture et la cotisation pour la caisse d'assurances des accidents agricoles

À partir de 2004, le montant de la taxe professionnelle comprend en plus la taxe pour frais de Chambre de Commerce et d'Industrie, la taxe pour frais de Chambre de Métiers, et la cotisation nationale de prééquation

À partir de 2004, le montant de l'impôt sur les sociétés comprend les versements spontanés y compris la contribution exceptionnelle et la contribution temporaire

14) Suite à réhaussement du seuil d'imposition de 0,8M€ à 1,3M€

15) Hors départements de la Charente-Maritime et de la Manche

Attention aux données :

- Provenant de **plusieurs pays**
- Aux **méthodes de mesure** des données (catégories de chômeurs, CVS, etc...). Des notes de synthèse et révision sont généralement disponibles.
- Aux **questions posées** et échantillons récoltés (lors de sondages, études..)
- A la **périodicité** des données : quand/à quelle fréquence les données sont collectées (mensuel, trimestriel...)?

Citez vos sources de données!

De la même façon que vous citerez une source dans un papier, il ne faut jamais oublier de **mentionner vos sources de données**.

Pour chaque élément de données que vous allez inclure dans votre récit, vous devez créer une “**biographie**” de celles-ci :

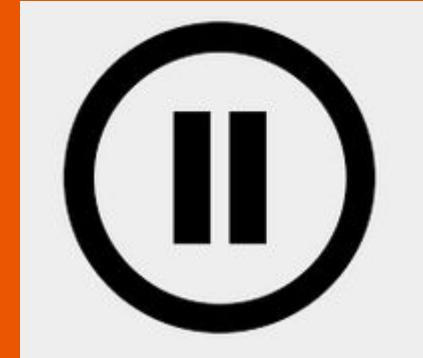
- le contexte, ou l'origine : cela vous aidera aussi à les comprendre
- la date de ces données

Bénéfices :

- pour pouvoir y revenir si besoin (utile lorsque l'on travaille sur plusieurs sujets en même temps)
- précision et transparence pour votre audience et vos confrères/consoeurs
- Check supplémentaires, possibilité de repérer une erreur dans vos données
- Vous devrez toujours pouvoir justifier de l'origine de vos informations à toutes les étapes de production de votre récit

Des questions ?

5 minutes de pause



Le projet

Projet

- **En équipe de 2 ou 3 personnes**
- Pas de conseils / limite sur le nombre de dataviz ou la longueur du texte
- **Focus sur le récit**, et dataviz qui vient en support ou complément (elle se justifie)
- **Publié en ligne** sur la plateforme que vous souhaitez (blog, etc...)
- Google est votre ami, les tutos aussi : **on encourage la débrouille !**
- Vous serez notés sur le contenu produit, mais aussi votre démarche, le jeu du collectif, à travers un **pitch à l'oral** le 15 octobre (chacun devra expliquer son rôle dans la construction du projet, les embûches rencontrées, la résolution du.es problème.s)

**Pour vous aider : différentes
méthodes d'approche**



Toute question est un **bon point de départ** pour explorer vos données!

- La plupart des journalistes ne fouillent pas les données par pur plaisir ou curiosité, car cela peut être très **chronophage** quand on ne sait pas quoi chercher.
- Souvent, ils ont une **histoire à raconter** ou un problème à résoudre. Encore plus souvent, ils ont une **question en tête**, telles que : “Ma ville natale/de résidence est-elle plus ensoleillée que la moyenne?”, “Comment l’équipe municipale dépense-t-elle l’argent public?”, “Le prix de l’essence connaît-il une hausse inédite en France?”
- Comprendre pour qui votre question est destinée vous aidera également à **définir votre public**, celui pour lequel vous allez travailler et celui qui vous **aidera à construire votre récit**.

Où sans question?

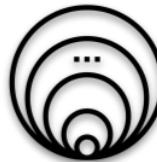
- Si pas de question = **exploration**
- Enquête commencée ou non : veuillez à toute **tendance inattendue**, résultat **inhabituel** ou tout ce qui vous surprendra.
- Si des données manquent, c'est aussi peut-être une information qui vaut la peine d'être racontée?
- Souvent, les histoires les **plus intéressantes ne sont pas celles que vous cherchez** à l'origine.

Trouver son angle

Voici les angles que les journalistes utilisent le plus souvent pour leurs récits de datajournalisme, selon Paul Bradshaw. Il a pris et analysé 100 publications de datajournalisme pour voir s'il est possible d'identifier la fréquence de recours à chacun de ces angles de narration.

7 common angles for data stories

Scale



Change



Ranking



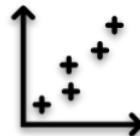
Variation



Explore



Relationships



Bad/open



(+ Leads)



7 common angles for data stories

1. Scale: 'This is how big an issue is'
2. Change/stasis: 'This is going up/down/not improving'
3. Outliers/ranking: 'The best/worst/where we rank'
4. Variation: "Postcode lotteries" and distributions
5. Exploration: Tools, simulators, analysis — and art
6. Relationships/debunking: 'Things are connected' — or not, networks and flows of power and money
7. Problems & solutions: 'Concerns over data', 'Missing data', 'Get the data'

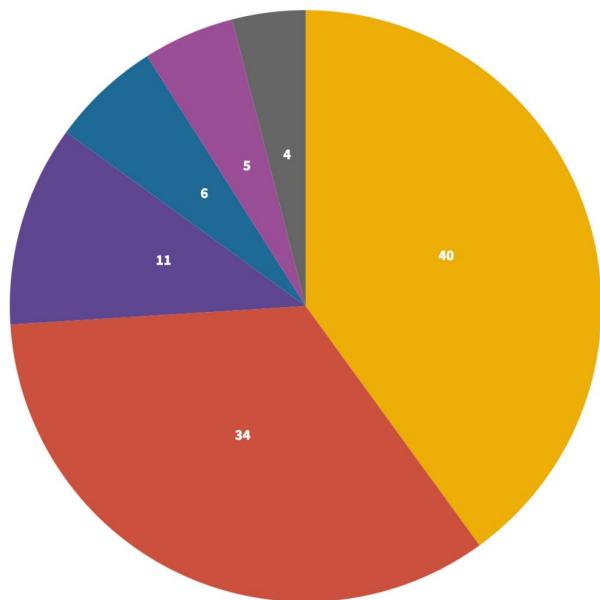
ONLINE JOURNALISM BLOG

Most common angles for 100 BBC Data Unit stories
By Paul on 27 May 2021

These are the most common angles adopted on data for 100 BBC Data Unit stories

Analysis of 100 stories by the BBC Data Unit and Shared Data Unit finds that scale and change stories are used much more often than any other type of angle.

■ Scale ■ Change (including lack of it) ■ Ranking/outliers ■ Variation ■ Open/bad data ■ Exploratory/tool

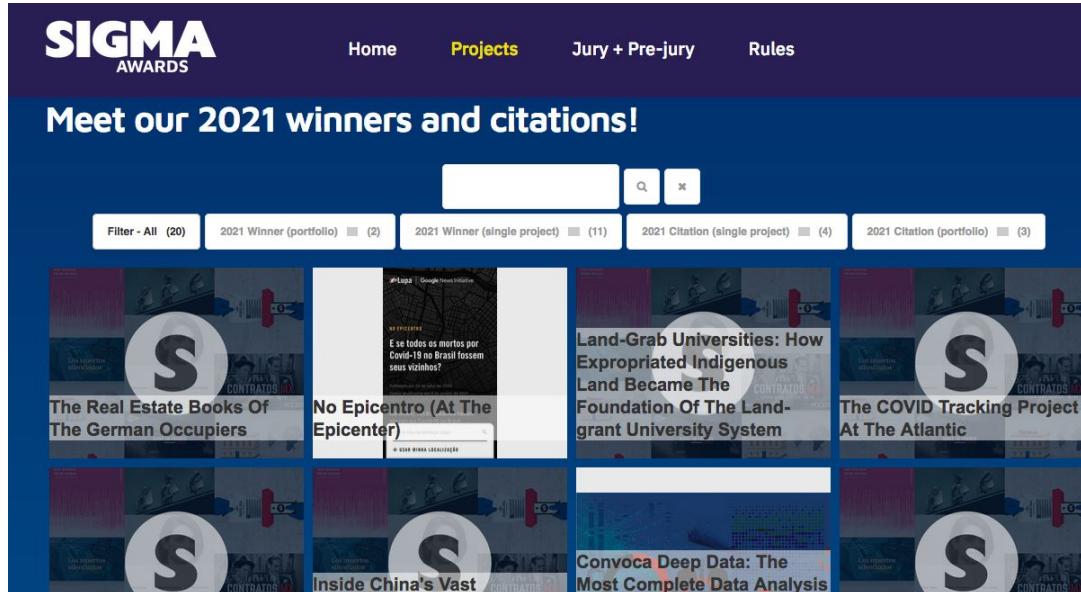


- Echelle :
- Changement/stabilité:
- Classement:
- Evolution:
- Exploration:
- Liens (ou non)
- Problèmes/solutions

Des ressources

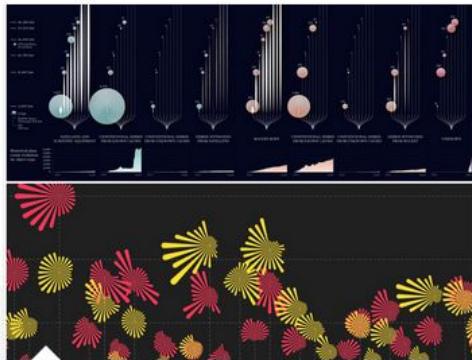
S'inspirer : les Sigma awards

Le **prix international de datajournalisme**, avec les coulisses du travail de chaque lauréat / candidat Sigma Awards (nouveau nom du concours depuis 2020) (<https://sigmaawards.org/>)



S'inspirer : information is beautiful

Information is beautiful (<https://www.informationisbeautifulawards.com/>)



OVER 1 YEAR AGO

Best of the Web: January 2020

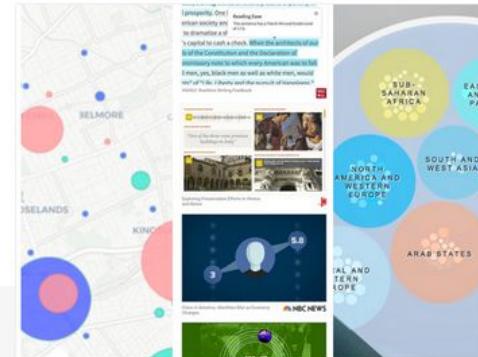
There's only one cure for the depths of January... bright, gorgeous, dazzling data. Scroll down to see which vizes got all the likes, shares, slow claps, and nods of approval in the past... →



OVER 1 YEAR AGO

Interview with 2019's Best Non-English-Language Winner

Our Best Non-English-Language category presents us with an opportunity to celebrate vizes featuring topics and commentary outside of the English-speaking world - i.e., most of the



OVER 1 YEAR AGO

6 Years of Outstanding Outfits

The Information is Beautiful Awards have been celebrating extraordinary outlets and astounding studios since 2014... that's 6 years of data visual goodness from some stunningly creative &... →

TEDX, talks, conférences...

de designers de dataviz, artistes, journalistes, scientifiques



Hans Rosling :
The best stats
you've ever seen

Manuel Lima : A visual history of human knowledge

Aaron Koblin Visualizing ourselves with crowd sourced data

David McCandless : The beauty of data visualization

Giorgia Lupi : How we can find
ourselves in data

Nate Silver : La question de la race
affecte-t-elle le vote?

Mona Chalabi :

-Designing Data for Maximum Impact &

-3 moyens de repérer une mauvaise
statistique

Anne Milgram : Pourquoi les statistiques
sont des éléments clefs pour combattre
le crime

Tutos en ligne

- **Data Journalism and Visualization with Free Tools (en anglais)**
<https://journalismcourses.org/course/dataviz/>
- [Doing Journalism with Data: First Steps, Skills and Tools](#)
- [Cleaning Data in Excel](#)
- [Charting Tools for the Newsroom](#)
- [Mapping for Journalists](#)
- [Python for journalists](#)
- [R for journalists](#)

Pour le prochain cours le 8 octobre :

- A partir des ressources, trouvez une dataviz / une enquête-récit / une vidéo qui vous a plu et expliquez-nous pourquoi en **2 minutes chrono**.

On commencera vendredi par un **tour de table** avec les coups de cœur, les remarques de chacun.nes

- Répartissez-vous en **groupes de 2-3 personnes** maxi et **commencez à réfléchir à votre projet final** : les groupes doivent être formés vendredi 8 octobre matin pour commencer à tester des outils de viz en petits groupes.

Au programme vendredi 8 octobre

- Discussion autour des ressources proposées (coups de cœur, etc...)
- Intro à la datavisualisation (si non vue le 4/10)
- Les erreurs de représentation visuelle + quiz
- Découverte de Datawrapper (Ana)
- Découverte de Flourish (Alice et Ana)
- Exercices/discussion

Des questions ?

Merci ! Good night and good
luck
