Café des données - PUD Panels & corpus numérique

Alioscha Massein, Orline Poulat, Céline Faure, Alexandra Dugué, Hélène Kiefer

msh Lyon St-Étienne

Café des Données

PUD-PANELS & corpus numérique

Alioscha Massein, Céline Faure, Orline Poulat et Hélène Kieffer

07/02/2023

Le café des données c'est quoi ?

- Un mardi par mois, de 9h à 10h.
- Un aspect de l'utilisation et du traitement de données par café
- Orienté pour les SHS.
- Pensé pour la discussion avec les intervenant-es et les personnes présentes.

msh Lyon St-Étienne

PROGRAMME

Où trouver des données en SHS ?

Brice Lefèvre, L-VIS
7 mars 2023

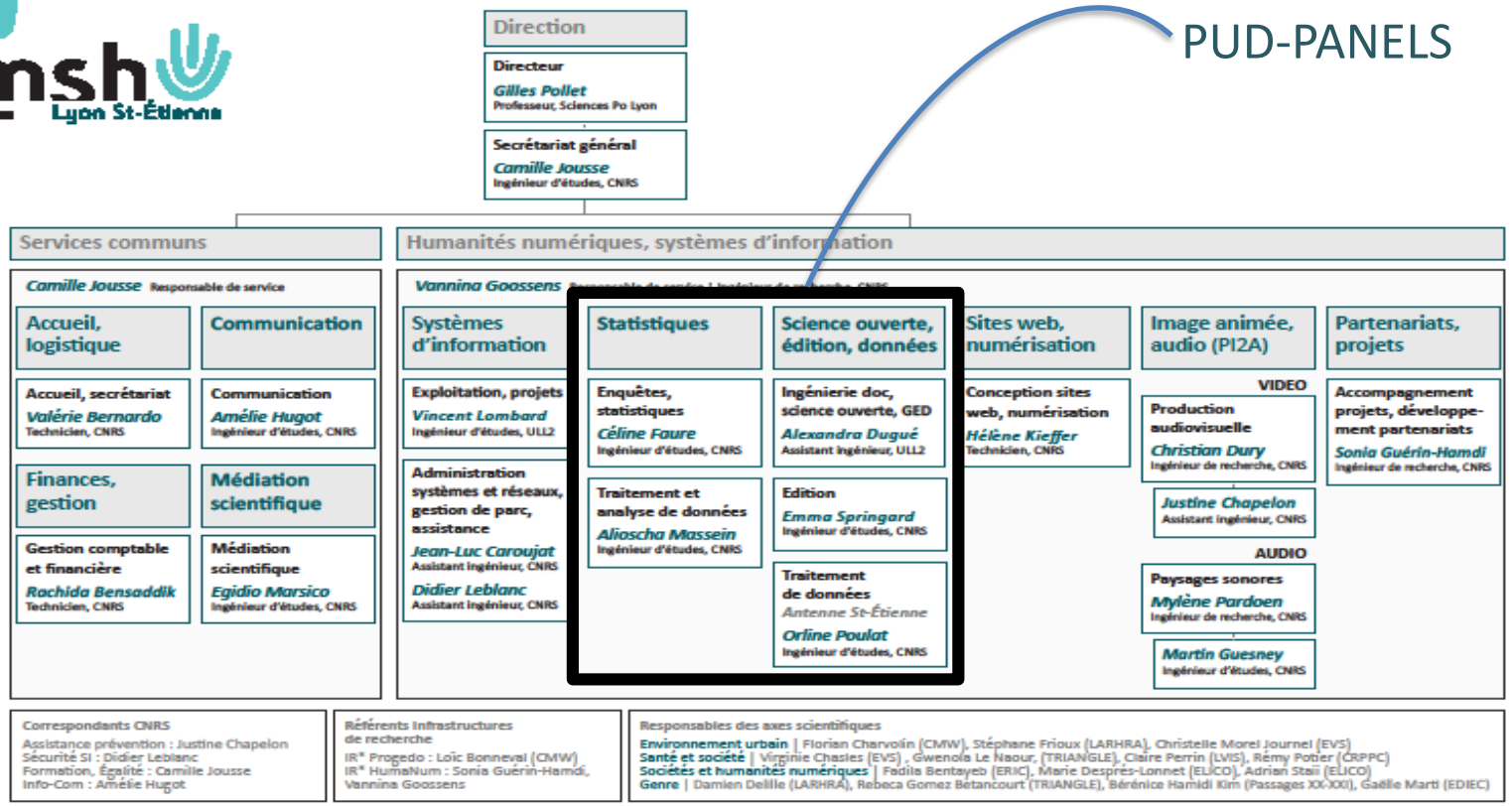
Données de cadrage et statistiques descriptives

Loïc Bonneval, CMW
18 avril 2023

Permanence libre : venez avec vos données

Intervenant-es en fonction des demandes.
9 mai 2023

Une Plateforme Universitaire de Données au sein de la MSH



Qu'est-ce qu'une PUD ?

Relais local de l' ***Infrastructure de Recherche * Progedo*** chargée d'impulser et structurer une politique publique des données pour la recherche en sciences sociales.

Quetelet PROGEDO Diffusion permet de rechercher et d'accéder à des données issues de la statistique publique nationale (grandes enquêtes, recensements, bases de données) et de grandes enquêtes provenant de la recherche française. (http://quetelet.progedo.fr/ords/)

Disciplines scientifiques couvertes : droit, économie, géographie, gestion, histoire, santé, sciences politiques et sociologie. L'ensemble des laboratoires de SHS sur le territoire de Lyon-Saint-Étienne sont couverts par PANELS et la MSH.

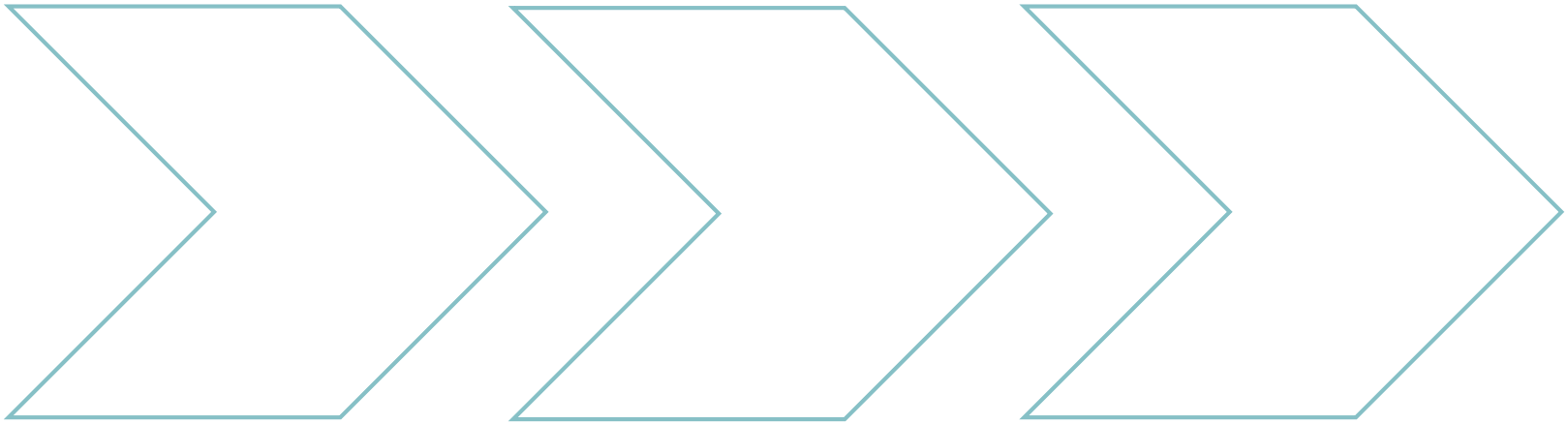
- Identifier les ***données disponibles*** correspondant aux besoins,

- Orienter les chercheurs vers ***les producteurs de données pertinents,***

- Sensibiliser former les chercheurs et les étudiants à la «culture de la donnée».

À quoi sert PANELS ?

Accompagnement des chercheur-ses sur demandes, dans toutes les étapes qui nécessitent la collecte, le traitement, l'analyse, le stockage de données : **le cycle des données.**



Création ou **réutilisation** de données :
- Questionnaires
- Entretiens
- Archives
- Corpus
- Jeux de données

Analyses et traitements :
- Statistiques
- Textométrie
- Expérimentation
- **Accompagnement** sur les logiciels
- Mise en conformité **RGPD**

Stockage des données
- Lien avec les IR* **Progedo** & **HumaNum**
- Conseils pour la mise en place de recherche **« science ouverte »** : FAIRisation des données.
- **Diffusion** pour réutilisation des données

Le pôle statistique

Le pôle statistique est rattaché au service HNSI et se compose de **2 ingénieurs d'études** spécialisés en collecte de données, traitements statistiques, montage d'enquête.

Disciplines scientifiques couvertes : droit, économie, géographie, gestion, histoire, sciences politiques, sociologie, linguistique…

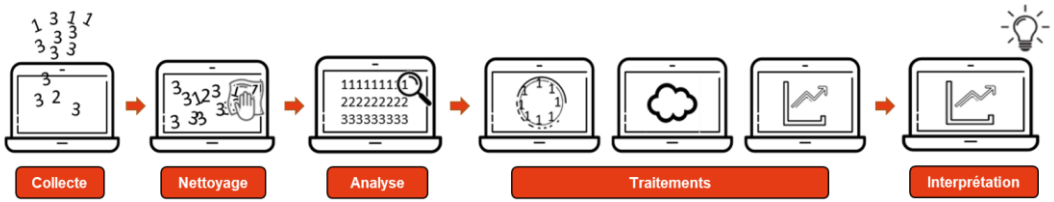
Leurs missions sont réalisées - **en partie** – dans le cadre de **PANELS :**

- Relais local de l'***Infrastructure de Recherche PROGEDO*** depuis 2009,
- Aide à l'utilisation des données mises à disposition par PROGEDO
- 3 missions principales : 1 - l'aide à la recherche et à la mise en œuvre des fichiers de données, 2 - la formation au traitement de ces données, et 3 - la formation aux outils informatiques et statistiques.

Compétences

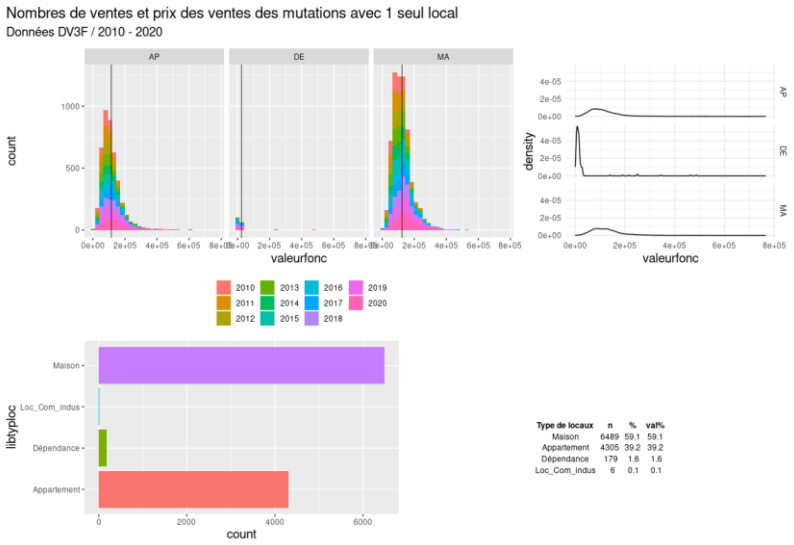
- **Informer et collecter les données en SHS :**
  - accès aux bases de données de la statistique publique nationale, internationales
  - Mise en place de dispositifs de recueil de données du Web (Web scraping, crawling).
  - Traitement statistique des données, analyse
  - Elaboration d'enquêtes
  - Accompagnement pour la mise en conformité des données utilisées (RGPD)

- **Traitement et analyse dans le cadre de projets :** analyse statistique, analyse de réseaux, cartographie, modélisation statistique, data-visualisation

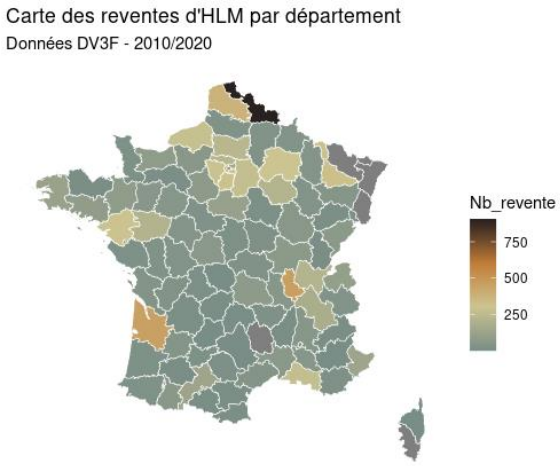


- **Organisation de formations :** Introduction à l'usage de logiciels, méthodes spécifiques (cartographie, échantillonnage, machine Learning, tests statistiques) et cycles de formation : semaine DATA-SHS, parcours cartographique, visualisation
  >> https://www.msh-lse.fr/services/formation/

1. Analyse des données issues d'une base foncière :



Type de locaux	n	%	val%
Maison	6489	59.1	59.1
Appartement	4305	39.2	39.2
Dépendance	179	1.6	1.6
Loc_Com_indus	6	0.1	0.1



2. Créer des enquêtes en ligne et se mettre en conformité sur les données personnelles traitées

Le pôle Science ouverte, édition et données

Au sein du service HNSI et du **pôle science ouverte, édition et données *(2 personnes)***, l'appui en matière de **gestion et traitement de données** est pris en charge par une ingénieure d'étude spécialisée sur un ensemble de méthodes et techniques directement lié aux Humanités numériques.

→ L'accompagnement des chercheurs s'effectue tant sur des traitements de données **qualitatives** que **quantitatives**, et propose un **service de proximité** sur le site de l'Université Jean Monnet à **Saint Etienne**.

Compétences - Service de gestion et traitement de données

- Constitution de **base de données** : modélisations et implémentation
- **Extraction automatisée** de données et **création de corpus**.
- **Analyses textuelles** : Iramuteq, R, Python -> lemmatisation, TF-IDF, moteurs de recherche…
- **Analyses statistiques** : statistique descriptive dont statistique multivariée.
- **Nettoyage, recodage et formatage** de jeux de données.
- **Edition numérique** : encodage en XML-TEI
- Au besoin, développement d'outils : Python, HTML/CSS/JavaScript, R, SQL/PHP…

- **Données bibliographiques et web de données** : conseil en matière de formats, standards et normes qui s'appliquent à l'indexation, aux thésaurus et aux métadonnées ;
- **Veille scientifique** : aide à la mise en place d'outils et conseils adaptés aux projets ou thématiques ;
- **Gestion des données** : formation au plan de gestion de données (projets, structures) ; conseils et méthodes sur le dépôt et la diffusion des données (choix des entrepôts), orientation pour les démarches d'archivage classique et numérique des laboratoires;
- Organisation de formation à la demande et sur des sujets spécifiques : Data Management Plan, Data Papers, RGPD, éthique et droit…

msh Lyon St-Étienne

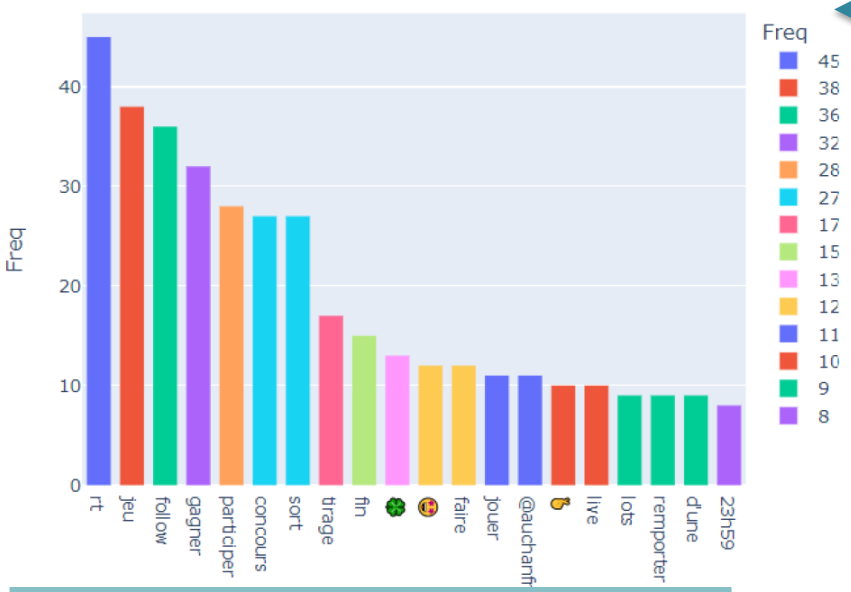
1. Analyser le vocabulaire des marques sur Twitter

Key	Type	Size	Value
post_1	dict	5	{'description':'C'est vendredi ! Et aujourd'hu
post_2	dict	5	{'description':'Ce midi c'est healthy avec la
post_3	dict	5	{'description':'Le dos de cabillaud est notre
post_4	dict	5	{'description':'Un bon burger ça vous dit ? No
post_5	dict	5	{'description'
post_6	dict	5	{'description'
post_7	dict	5	{'description'
post_8	dict	5	{'description'
post_9	dict	5	{'description'
post_10	dict	5	{'description'

→ *Webscrapping* et constitution de corpus

Key	Type	Size	Va
comment_nb	str	1	6
commentaires	list	6	['Avec du riz et du citron', 'En papi
description	str	132	Le dos de cabillaud est notre poisson
like	str	10	2107 likes
timeline	str	10	2 days ago

Les 20 mots les plus courants des tweets (>= 50 likes ; >= 50 rt)



→NLP / textométrie, moteur de recherche…

2. Formations

Données : bonnes pratiques, éthique

PROTECTION DES DONNÉES PERSONNELLES ET RECHERCHES SHS

Atelier-formation proposé par la MSH Lyon St-Etienne

1 journée par an. Le 27 septembre 2021, en distanciel | Programme & informations

PUBLICATIONS ET DONNÉES (EN) SHS SOUS L'ANGLE DE L'ÉTHIQUE ET DU DROIT

Conférences-ateliers, en collaboration avec l'Urfist de Lyon

4 séances. Du 18 octobre 2019 au 4 mai 2020, à Lyon | Programme & informations

Le pôle site web et numérisation

La numérisation est la **conversion** d'un signal (textes, images, sons) **sous forme de données** pouvant être traitées par un dispositif informatique.

En sciences humaines et sociales, la numérisation sert différents objectifs :

- pérenniser et archiver des documents à valeur patrimoniale ;
- créer des corpus (fonds) numériques constituant des objets de recherche ;
- rendre accessible ces fonds à la communauté scientifique, voire à la société.

La MSH mutualise du matériel et des outils de numérisation, mais apporte aussi des compétences spécifiques pour mieux **accompagner les chercheurs** en littérature, histoire, gestion, géographie, etc.

https://www.msh-lse.fr/services/numerisation/

Compétences

La MSH met à disposition des laboratoires du matériel adapté pour numériser différents supports.

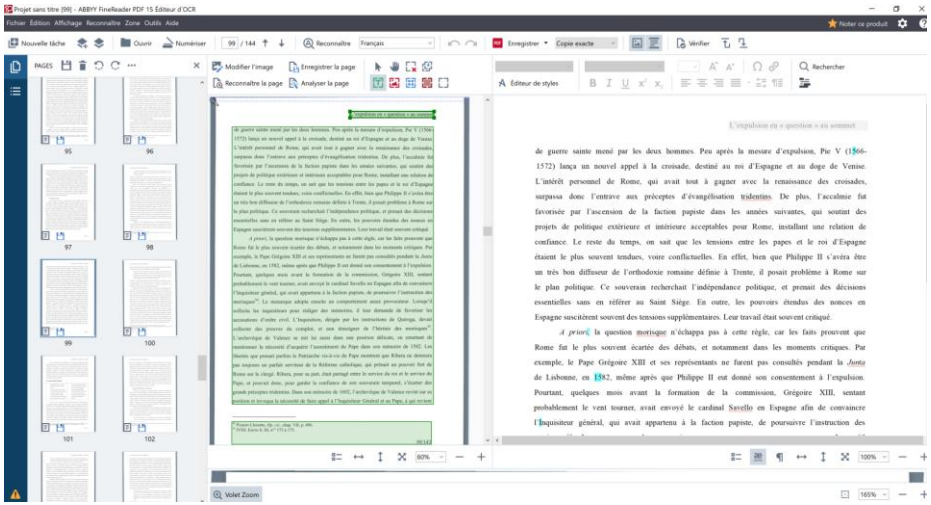
- Mise à disposition d'une **station de numérisation** pour numériser différents supports : supports papier, photographie, microfiches et microfilms…
- reconnaissance optique de caractères (**OCR**), mise en pages ;
- **Traitement d'images** (rééchantillonnage, recalibrage) ;

Ces travaux peuvent être effectués par la MSH pour remettre aux chercheurs des contenus directement exploitables (livraison du fac-similé d'un ouvrage récent, OCR de documents administratifs).

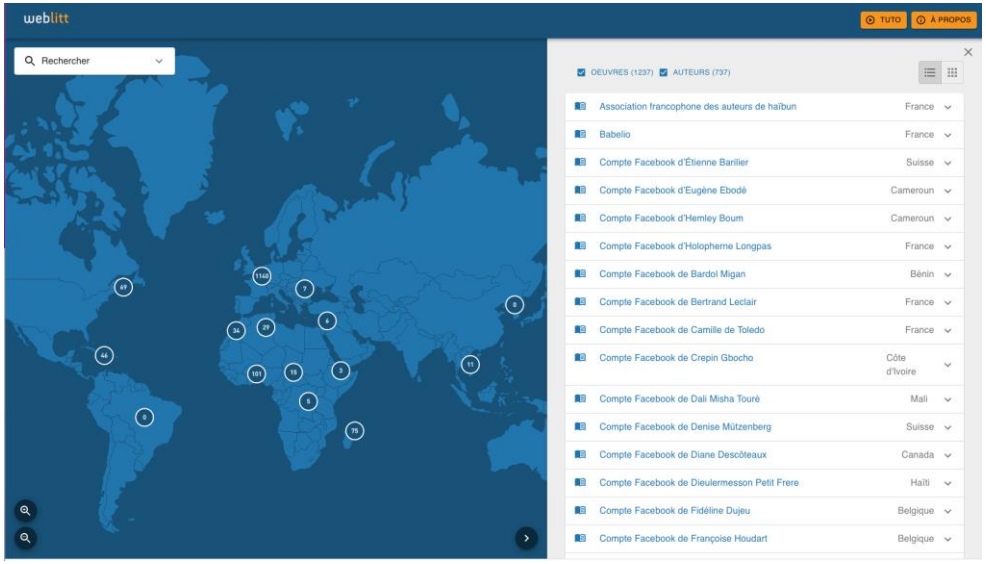
En plus de la numérisation d'autres services sont disponibles et mis en place :

- La **révision** des documents numérisés
- La mise en forme des textes
- Diffusion en ligne et mise en place de sites internet

1. Océrisation d'un corpus



2. Hébergement de sites web



msh Lyon St-Étienne

equipe.panels@msh-lse.fr

Statistiques :

https://www.msh-lse.fr/services/statistiques/

Céline Faure :
celine.faure@msh-lse.fr

- Enquêtes
- Traitements de données (outils, méthodes)
- Accompagnement sur la mise en conformité de données personnelles (LiL, RGPD)

Alioscha Massein
alioscha.massein@msh-lse.fr

- Traitement de données
- Statistiques descriptive, multivariée
- Informatique pour les SHS (R, Python, SQL)

Science ouverte – données :

https://www.msh-lse.fr/services/science-ouverte/

Alexandra Dugué :
alexandra.dugue@msh-lse.fr

- Gestion et diffusion des données de recherche
- Métadonnées, entrepôts de données
- « Fairisation »

Orline Poulat :
orline.poulat@msh-lse.fr

- Analyse textuelle
- Traitement de données
- Informatique pour les SHS (Python, javascript, SQL…)

Sites web et numérisation

Hélène Kieffer
helene.kieffer@msh-lse.fr

Edition numérique - Prairial

Emma Springard
emma.springard@msh-lse.fr

Un référent scientifique :

Loïc Bonneval
loic.bonneval@...fr