Semaine Data SHS

Représentativité, collecte et nettoyage des données en sciences humaines et sociales

Du 8 au 12 décembre 2025

à la MSH Lyon St-Étienne, Lyon 7e / St-Étienne

CONFÉRENCES ET ATELIERS MÉTHODOLOGIQUES

Contexte et Objectifs

Dans le cadre de sa plateforme universitaire de données (PUD) PANELS, la MSH Lyon St-Etienne organise du 8 au 12 décembre 2025 sa semaine « Data SHS », qui a pour thématique générale « Représentativité, collecte et nettoyage des données en sciences humaines et sociales ». Cette semaine, inscrite dans le plan national de formations de PROGEDO, propose une série de présentations et d'ateliers pratiques autour des méthodes et enjeux du traitement des données.

Ces conférences et ateliers ont un double objectif : présenter en détails les fondements théoriques des méthodes de construction des jeux de données SHS et expliciter les moyens de leur mise en œuvre pratique. Le but est de permettre aux participants d'adapter ces méthodes à leurs propres travaux, de façon la plus autonome possible.

Public : ce programme s'adresse aux chercheurs, enseignants-chercheurs, ingénieurs et techniciens, ainsi qu'aux doctorants, des laboratoires SHS du site Lyon St-Étienne.

Thématiques

4 grandes thématiques ont été retenues pour cette semaine d'ateliers :

- Données de recherche : actualités Progedo, le Panel ELIPSS, et la représentativité des échantillons
- La collecte de données
- Le nettoyage des données
- Les données images

LUNDI 8 DÉCEMBRE 2025

Données de recherche : actualités Progedo, panel ELIPSS et la questions de la représentativité

Module 1 – Données de recherche : s'appuyer sur les infrastructures nationales | **Inscription**
Module 2 – L'échantillonnage en sciences humaines et sociales | **Inscriptions**

MARDI 9 DÉCEMBRE 2025

Collecter les données en SHS : Retour sur le processus de collecte et inititation au questionnaire avec LimeSurvey

Module 1 – Retour sur le processus de collecte | **Inscriptions**
Module 2 – Initiation à LimeSurvey | **Inscriptions**

MERCREDI 10 DÉCEMBRE 2025

Nettoyage et recodage des données : Retour sur le nettoyage et le recodage des données, initiation à OpenRefine

Module 1 – Recoder des données, entre impératif statistique et traduction du terrain | **Inscriptions**
Mosule 2 – Initiation à OpenRefine | **Inscriptions**

JEUDI 11 DÉCEMBRE 2025

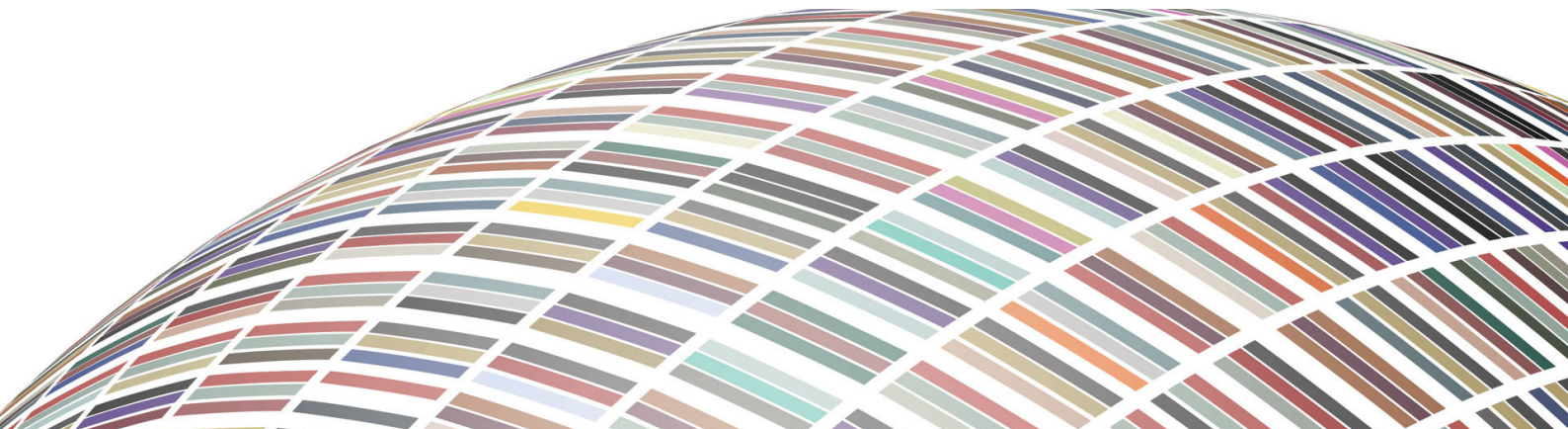
Collecter les données en SHS : Webscrapping

Module 1 – Initiation au webscrapping de site statistique avec Python et BeautifulSoup| **Inscriptions**

VENDREDI 12 DÉCEMBRE 2025

Manipuler, collecter et analyser des images

Module 1 – Gérer, traiter et diffuser ses images numériques | **Inscriptions**
Module 2 – Annotation d'images pour l'entraînement des modèles d'intelligence artificielle | **Inscriptions**



LUNDI 8 DÉCEMBRE 2025

Données de recherche : actualités Progedo, Panel ELIPSS et la questions de la représentativité

Module 1 – Données de recherche : s'appuyer sur les infrastructures nationales

Intervenants : Nicolas Sauger, Olivier Beaude, Alioscha Massein, Loïc Bonneval

10h – 10h45 : Présentation de la PUD-Panels

Les PUD sont des dispositifs nationaux implantés localement et qui accompagne le personnel de la recherche dans la collecte, le traitement et l'analyse des jeux de données, en particulier quantitatif. Sur le site de Lyon St-Étienne, c'est la PUD-Panels qui décline la politique de l'IR* PROGEDO sur le territoire.

La principale mission de PROGEDO est de développer la culture des données, impulser et structurer une politique des données d'enquêtes pour la recherche en sciences sociales. Son organisation est conçue pour les trois niveaux d'enjeux stratégiques où elle doit intervenir : européen, avec l'investissement dans les consortiums européens (ERIC) et le partage des grandes enquêtes ; national, avec les départements de la documentation et la diffusion des données par le département Quetelet-Progedo-Diffusion ; et régional, dans les MSH et les Universités avec les Plateformes Universitaires de Données (PUD).

Nous proposons cette séance pour vous faire découvrir l'accompagnement que nous proposons ainsi que l'actualité de l'IR* PROGEDO.

11h – 12h : Présentation de l'actualité de PROGEDO

Organisée par les Plateformes Universitaires de Données et le réseau des correspondants Huma-Num, la semaine dataSHS est coordonnée par les IR* Progedo et Huma-Num, avec l'IR RnMSH (les MSH et leur réseau). Cette semaine propose une série de conférences, de présentations, d'ateliers et de formations afin de sensibiliser la communauté scientifique à l'importance des données et des méthodes en sciences humaines et sociales.

Horaires : 10h-12h00
Lieu : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Inscription le 4 décembre 2025 > formulaire en ligne : Inscription

LUNDI 8 DÉCEMBRE 2025

Données de recherche : actualités Progedo, Panel ELIPSS et la questions de la représentativité

Module 2 – L'échantillonnage en sciences humaines et sociales

Intervenants : Alioscha Massein, Panel ELIPSS, Géraldine Charrance (INED) & Aurélie Santos (INED)

14h – 14h30 : Retour sur la représentativité dans les enquêtes

La question de la représentativité des données en SHS est un élément central de toute démarche de collecte et de traitement de données. Qu'il s'agisse de l'observation d'une population connue ou inédite, d'un petit ou d'un grand échantillon, d'une enquête ou de la sélection de document dans un corpus, la représentativité des informations nous permet d'inférer des résultats à une population plus large que celle étudiée sur le terrain. Nous reviendrons ici sur les principales notions pour appréhender au mieux les précautions à prendre dans la constitution d'un échantillon.

14h30 – 15h30 : Le Panel ELIPSS – un panel à disposition des projets de recherche en SHS

ELIPSS est un panel Internet, représentatif de la population française, constitué de plus de 2 200 personnes qui sont invitées à participer tous les mois à des recherches dans de nombreux domaines (santé, environnement, politique, sport et loisirs, etc.). Contrairement aux sondages d'opinion et aux études de marché, ces études sont élaborées par des chercheur·e·s sur des thèmes d'intérêt général et ont une finalité exclusivement scientifique. Cette présentation sera l'occasion de revenir sur l'origine du projet et d'exposer les caractéristiques actuelles du panel, mais aussi de mettre en avant son utilisation par la communauté scientifique. Ce sera l'occasion de rappeler les modalités de l'accès aux données, et de présenter des usages et travaux concrets réalisés à partir de ces données.

16h – 17h : L'enquête ChiPRe – Atteindre la représentativité sans base de sondage

Le projet ChIPRe, dans le cadre duquel la première enquête quantitative française spécifiquement dédiée aux immigrés chinois en Île-de-France a été réalisée, documente différentes dimensions de cette immigration. L'enquête a recouru à une méthode d'échantillonnage innovante (Network Sampling with Memory) qui ne fait pas appel à une base de sondage classiquement utilisée dans les méthodes d'échantillonnage. Nous écouterons Géraldine Charrance et Aurélie Santos nous détailler les mesures qu'elles ont mis en œuvre pour mettre en place cette méthode.

Horaires : 14h00-17h00
Lieu : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Inscription le 4 décembre 2025 > formulaire en ligne : Inscriptions

MARDI 9 DÉCEMBRE 2025

Collecter des données – 1ere journée

Module 1 – Retour sur le processus de collecte

Intervenants : Alioscha Massein, Céline Faure, Charlie Fabre, Julien Salanié

9h30 – 11h : Retour sur la collecte de données

La collecte de données est un processus nécessaire à tout travail de recherche. Une fois l'échantillon constitué, il s'agit de l'interroger : soit avec une série d'hypopthèse que l'on cherche à tester, soit de manière inductive. Dans les deux cas, la « captation » des données au moment de la collecte est fondamentale et pose de nombreux biais qu'il s'agit au mieux de prendre en compte, sinon d'éliminer. Nous reviendrons dans ce premier temps sur quelques aspects fondamentaux de la collecte.

11h15 – 11h45 : Collecter les données à partir d'un questionnaire, l'exemple de l'enquête identités transmasculines en France entre 2000 et 2022.

Cette présentation revient sur les questionnements pratiques et éthiques posées dans le cadre d'une enquête à destination de publics minorisés et en dehors des binarités de genre. Elle revient ainsi sur les différentes étapes de réflexion et de construction de l'enquête et la démarche de protection des données.

11h45 – 12h30 : Webscrapper des données, l'exemple du projet sur les algues vertes.

Ce projet étudie les déterminants du ramassage des algues vertes dans les communes du littoral atlantique touchées par les marées vertes. La médiatisation du phénomène nous intéresse particulièrement et nous utilisons un corpus d'articles de presse écrite pour créer des indicateurs locaux de pression médiatique à partir de la géolocalisation des articles et de la mesure de leur valence émotionnelle par analyse de sentiments.

Horaires : 9h30-12h30
Lieu : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Inscription le 4 décembre 2025 > formulaire en ligne : Inscriptions

MARDI 9 DÉCEMBRE 2025

Collecter des données – 1ere journée

Module 2 – Initiation à LimeSurvey

Intervenante : Céline Faure

14h-15h30 / 14h30 – 16h : Atelier LimeSurvey

Cette session est destinée à permettre une prise en main de l'outil Limesurvey, installé sur les serveurs de la MSH. L'idée est de transmettre les clés de réflexions préalables pour créer et réaliser leurs enquêtes en autonomie. Des enquêtes qualitatives ou quantitatives menées à l'aide d'un questionnaire (formulaire) en ligne permettent, notamment, de récolter des données dans le cadre de recherches en sociologie, économie, géographie, science politique, droit, gestion, langues, etc.

L'objectif de la séance est de présenter LimeSurvey, un outil d'enquête en ligne gratuit et simple d'utilisation, qui permet de créer de simples questionnaires ou des enquêtes élaborées, dans le cadre d'un projet de recherche. Vous seront expliqués le fonctionnement de l'outil, son interface, les paramétrages à apporter, les types de questions possibles, la gestion des réponses, etc. L'atelier sera l'occasion de rappeler les bonnes pratiques à avoir dans la conception et la mise en œuvre de ce type d'enquêtes, et abordera également les problématiques de protection des données personnelles.

Horaires : 14h00-16h00
Lieu : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Inscription le 4 décembre 2025 > formulaire en ligne : Inscriptions

MERCREDI 10 DÉCEMBRE 2025

Collecter des données - 2e journée

Module 1 – Recoder des données, entre impératif statistique et traduction du terrain

Intervenant : Alioscha Massein

9h30 – 11h : Retour sur le recodage des données

La collecte de données nous amène un ensemble de premières informations qui sont souvent peu exploitables dans un premier temps. Que ce soit pour des raisons éthiques, légales, avec l'anonymisation et la pseudonymisation des informations personnelles, la nécessité d'avoir une puissance statistique en recodant des informations qui sont trop peu nombreuses, ou encore pour éviter de réutiliser des catégories qui correspondent mal à notre population d'étude, pire, qui reproduisent des rapports de domination ou de légitimité dans les analyses, le nettoyage et le recodage de données est une étape incontournable du travail sur les données. Nous reviendrons sur les principaux concepts clés et des considérations pratiques du recodage.

11h15 – 12h : Retour d'expérience avec le projet d'étude sur l'expertise auprès de la Commission Européenne.

Depuis le milieu des années 2000, la Commission Européenne met à disposition un nombre croissant de jeux de données en libre accès pour mettre en œuvre sa politique de transparence. En s'intéressant aux mécanismes à l'œuvre dans la participation aux groupes d'experts qui interviennent auprès de la Commission, Cécile Robert s'est confrontée à plusieurs de ces jeux de données. Nous reviendrons sur le travail de nettoyage et de recodage des informations qui a été mis en œuvre non seulement pour permettre des traitements quantitatifs de qualité, mais aussi pour dépasser les limites des jeux de données initiaux, dont les origines sont multiples (contraintes pratiques et organisationnelles, évolution des modalités d'enregistrement, mais également luttes d'intérêts autour de la production mêmes de ces données.

Horaires : 9h30-12h00
Lieu : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Inscription le 4 décembre 2025 > formulaire en ligne : Inscriptions

MERCREDI 10 DÉCEMBRE 2025

Collecter des données - 2e journée

Module 2 – Initiation à OpenRefine

Intervenants : Sylvain Besson, Séverine Gedzelman
Pré-requis : un ordinateur

14h – 17h : Atelier OpenRefine

Une fois collectée, les données présentent un caractère globalement « brut », même si elles sont déjà circonscrites à une ou plusieurs problématiques de recherche. Le nettoyage est une étape indispensable dans un projet de travail comportant des données, qu'il est essentiel de prendre en compte. De même, le codage et le recodage des informations « brutes » peuvent s'avérer être une tâche pénible qui soulèvent de nombreuses questions que ce soit avec des données préalablement collectées ou avec des données que l'on réutilise.

Au cours de cet atelier, nous présenterons l'outil OpenRefine, boîte à outil indispensable pour le nettoyage des données, permettant également l'alignement et l'enrichissement des informations avec d'autres référentiels de données.

Horaires : 14h00-17h00
Lieu : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Inscription le 4 décembre 2025 > formulaire en ligne : Inscriptions

JEUDI 11 DÉCEMBRE 2025

Collecter les données en SHS : Webscrapping

Module 1 – Initiation au webscrapping de site statistique avec Python et BeautifulSoup

Intervenant : Alioscha Massein

Pré-requis : une installation de Python et des librairies nécessaires

9h30 – 12h30 : Webscrapper un site statique avec Pyhton et BeautifulSoup

Le webscrapping est une technique plutôt pratique pour récupérer des données depuis des sites internet. Internet étant fournie en page HTML, c'est-à-dire des données structurées, nous pouvons nous appuyer sur cette technologie pour récupérer automatiquement de l'information à partir des éléments propres à ce langage. En utilisant Python, les librairies requests et BeautifulSoup, nous verrons comment accéder à une page web, et extraire de l'information de celle-ci.

Cet atelier ne nécessite pas de pré-requis technique, mais une connaissance rudimentaire de python est un plus pour au mieux avancer sur cette matinée. Il faut également avoir installer quelques informations sur votre ordinateur. Vous pouvez prendre connaissance des éléments à cette **adresse : https://github.com/Geminy3/Webscrapping_formation**

Horaires : 9h30-12h30
Lieu : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Inscription avant le 4 décembre 2025 > formulaire en ligne : Inscriptions

VENDREDI 12 DÉCEMBRE 2025

Manipuler, collecter et analyser l'images numérique

Module 1 – Gérer, traiter et diffuser ses images numériques

Intervenants : Shannon Bruderer, Damien Petermann, Braz-ma Etheve

9h30 – 11h : Les images au coeur des projet de recherche

L'image est une donnée complexe qui exige une attention particulière portée à la structuration de ses métadonnées, à l'interprétation de son contenu, et à la mise en place de stratégies adaptées pour sa diffusion. Trois approches mettront en lumière la pluralité des usages de l'image : analyse environnementale, valorisation patrimoniale et gestion de corpus iconographiques avec Tropy.

Ces interventions mettront en lumière la richesse des pratiques mobilisant l'image et ouvriront la réflexion sur les enjeux techniques et méthodologiques que ce médium soulève dans les projets scientifiques contemporains.

11h45 – 12h30 : Wikimedia (Wikidata et Wikicommons)

Pré-requis : un ordinateur

Cet atelier propose une initiation à l'écosystème Wikimedia, en particulier à l'articulation entre Wikidata et Wikimedia Commons. Les participant·es découvriront comment ces plateformes structurent et relient les données et les images. Une introduction à l'utilisation de l'API Wikidata permettra de comprendre comment interagir automatiquement avec les données, avant de terminer par un exercice pratique de dépôt d'image illustrant le fonctionnement collaboratif de Wikimedia.

Horaires : 9h30-11h45
Lieu : Univ. Jean Monnet, Campus Tréfilerie, rue du 11 novembre, St-Étienne (salle K007)
Inscription le 4 décembre 2025 > formulaire en ligne : Inscriptions

VENDREDI 12 DÉCEMBRE 2025

Manipuler, collecter et analyser l'images numérique

Module 2 – Annotation d'images pour l'entraînement des modèles d'intelligence artificielle

Intervenants : Cédric Boscher, Shannon Bruderer

14h00 – 17h00 : Atelier Label Studio - Annoter un corpus d'images pour l'intelligence artificielle

Pré-requis : un ordinateur

Cet atelier propose une immersion dans la pratique de l'annotation d'images à des fins de recherche et d'apprentissage automatique. Après une présentation de la thèse doctorale de Cédric Boscher et une introduction à Label Studio (interface et fonctionnalités), la séance abordera les enjeux méthodologiques de l'annotation : vocabulaire contrôlé, protocoles d'annotation, et complexité de l'interprétation visuelle.

Horaires : 14h00-17h00
Lieu : Univ. Jean Monnet, Campus Tréfilerie, rue du 11 novembre, St-Étienne (salle K007)
Inscription le 4 décembre 2025 > formulaire en ligne : Inscriptions

Comité d'organisation

Alioscha Massein (CNRS, MSH Lyon St-Etienne)
Céline Faure (CNRS, MSH Lyon St-Etienne)
Loïc Bonneval (Univ. Lyon 2, CMW)
Shannon Bruderer (CNRS/Univ. St-Etienne, MSH Lyon St-Etienne)
Alexandra Dugué (Univ. Lyon 2, MSH Lyon St-Etienne)
Amélie Hugot (CNRS, MSH Lyon St-Etienne)
Syvain Besson (CNRS, MSH Lyon St-Etienne)

Informations pratiques

Les personnes intéressées pourront s'inscrire à un ou plusieurs modules (durée 1h à 3h).
La participation est gratuite, mais sur inscription obligatoire (nombre de places limité).

Localisation

Les séances se tiendront en présentiel ou en distanciel en fonction des séances :
Lundi, Mardi, Mercredi, Jeudi : MSH Lyon St-Etienne, 14 av. Berthelot, Lyon 7e (salle André Bollier, rdc).
Univ. Jean Monnet, Campus Tréfilerie, rue du 11 novembre, St-Étienne (salle K007).

Inscriptions

Merci de vous inscrire, au plus tard le jeudi 4 décembre 2025, à l'aide des formulaires en ligne spécifiés pour chacun des modules (cf programme ci-dessus et à télécharger en pdf).

Contacts

Alioscha Massein (MSH Lyon St-Etienne) - alioscha.massein@msh-lse.fr

Site web : **www.msh-lse.fr/semaine-data-shs-2025**