CREST - GENES
Cours doctoraux 2017 – 2018

Données numériques en sciences sociales
Collecte, traitement et analyse

Etienne OLLION
CNRS - Université de Strasbourg

Big data, données de l'internet données numériques, web sémantique, … ces termes ont fait une apparition remarquée dans les sciences sociales. Dans les discours d'abord, où ils sont régulièrement convoqués pour évoquer le futur de la recherche. Dans les pratiques ensuite, puisque les chercheurs sont régulièrement confrontés à des données de ce type, sans toujours pouvoir saisir les opportunités empiriques qu'elles offrent.
Qu'il s'agisse d'informations issues de l'internet, de bases de données ou d'informations stockées sur un disque dur, ou même de textes papiers scannés, un matériau parfois extrêmement riche est à portée de main, à condition de le repérer et savoir le traiter. Ces données intéressent les chercheurs en sciences sociales dans leur ensemble car tous peuvent avoir recours aux méthodes numériques pour collecter, stocker et traiter ces données dans le cadre d'un projet.
L'objectif de cette formation est de proposer une introduction à la collecte et à la curation de ces diverses données numériques. Il s'agira d'apprendre à les localiser, à mettre en place une stratégie pour les collecter, les nettoyer et les préparer en vue d'un traitement. Cet apprentissage ne sera pas dissocié d'une réflexion sur les enjeux que pose leur utilisation. Que peut-on apprendre avec ces diverses sources, et que nous masquent-elles? À quelles conditions l'abondance d'information est-elle bénéfique au savoir? Quel droit d'accès et d'usage pour les chercheurs face à ces données particulières ?
Afin de favoriser l'autonomie, le cours s'appuiera sur des exemples concrets (collecte et traitement de données de vote issues d'un site web, de données bibliométrique, croisement de bases de données, etc...).
Ce cours est conçu comme une introduction, et est destiné à des non-spécialistes. Les traitements seront intégralement réalisés sur R (une connaissance minimale du logiciel est bienvenue, mais pas indispensable). Un site compagnon offrira des matériaux supplémentaires (textes, scripts) pour prolonger l'analyse.

Plan du cours

1. Au-delà des *big data*
2. Comment on écrit le web ? (et comment le lire)
3. TD1 : Aspirer une page de site
4. Sélectionner des données (1) : Xpath
5. TD 2 : Principes de Xpath
6. TD 2 : Principes de Xpath
7. Droit, vie privée et sécurité
8. Automatisation et stockage
9. TD 3 : Boucles et stockage
10. Sélection des données (2) : Expressions régulières
11. TD 4 : Principes de Regex
12. Pistes de recherches

Cours	Lundis	13 Novembre 2017 20 Novembre 2017	De 14h à 17h00	Salle 1002
	Jeudis	16 Novembre 2017 27 Novembre 2017	De 14h à 17h00	Salle 1002

à l'ENSAE, - 5 Av. Henry Le Chatelier - Palaiseau (**REB B Massy Palaiseau & bus 9106C**)

Ces cours sont proposés aux étudiants de 3ème année de l'ENSAE, de l'ENSAI, ouverts aux étudiants de M2 ou inscrits en *thèse*. ***Une inscription préalable est demandée impérativement*** *pour tous les étudiants de l'ENSAE, de l'ENSAI, ou* extérieurs, à Lyza RACON : lyza.racon@ensae.fr ou par téléphone **au 0170266926 afin de pouvoir être admis dans les locaux de l'ENSAE et pouvoir être joints en cas de nécessité par les organisateurs du cours.**