Programme et intervenants de la formation « Données du Web : collecte automatique, nettoyage et traitements »

Lundi 24 juin 2019 (14h) - Jeudi 27 juin 2019 (17h)

Séances par demi-journée	Intervenants	Modalités pédagogiques
Stratégies numériques pour les sciences sociales	Étienne Ollion (CR, CNRS)	Séance introductive, cours
Introduction à R	Julien Boelaert (MCF, UDL)	Séance de travail sur ordinateur
Comment s'écrit le web ? Comment le lire ?	Étienne Ollion (CR CNRS)	Cours suivi de travaux dirigés
Sélectionner des informations : le langage XPath	Julien Boelaert (MCF, UDL)	Cours suivi de travaux dirigés
Curation de données : expressions régulières, open refine.	Julien Boelaert, (MCF, UDL)	Cours suivi de travaux dirigés
Enjeux légaux et éthiques de la récolte de données	Thomas Soubiran (IE, CNRS)	Cours
Automatisation de la collecte	Julien Boelaert, (MCF, UDL)	Cours suivi de travaux dirigés

Les enjeux et objectifs

1) Présenter les enjeux scientifiques (épistémologiques, théoriques, méthodologiques) qui entourent la multiplication des données numériques dans nos pratiques de recherche. En effet, les personnels de recherche sont confrontés à une prolifération de termes (données de l'internet, *big data*, données numérique, web sémantique) qu'il s'agira de critiquer, de décoder, d'expliquer et de replacer dans les enjeux théoriques et méthodologiques des sciences humaines et sociales.

2) Identifier les diverses données numériques utilisables pour mener à bien un projet. Il s'agira notamment d'apprendre à les localiser et à en évaluer la qualité, la valeur et l'intérêt pour un projet de recherche ou un projet documentaire.

3) Maîtriser des techniques simples de collecte (ou extraction) de données de façon automatisée puis de nettoyage/curation.

4) Connaître la réglementation qui entoure ces données, leur collecte et leur exploitation. En effet, si les données numériques ouvrent des potentialités pour la recherche scientifique et la collecte documentaire, ce sont aussi des informations qui font l'objet d'un accès et d'un usage de plus en plus réglementé aussi bien au regard de la protection des données personnels que des droits de propriété.