Programme et résumés des présentations

Pour de nombreuses disciplines les données du Web représentent une source inédite et pléthorique. En sciences sociales aussi l'accès à ces données pose de multiples questions pour leur utilisation complémentaire à d'autres données plus traditionnelles, issues d'enquêtes ou de fichiers administratifs.

Au-delà de l'adéquation aux questions de recherche l'utilisation de données du Web soulève d'autres défis : technique par exemple pour le choix d'un outil adapté à l'automatisation d'une collecte dédiée, ou juridique quant à leurs bons usages pour les analyses.

Les différents exposés de cette session seront à la fois des témoignages sur l'utilisation technique de données du Web en sciences sociales, mais aussi sur les apports et contraintes dans la mise en œuvre des analyses.

13h45	**Arnaud Bringé (Ined) et Bénédicte Garnier (Ined)** **Accueil et informations**
14h00	**Etienne Ollion (EO, CREST)** **Données du web: l'abondance et ses revers** Depuis une quinzaine d'années, l'usage des données numériques s'est répandu dans les sciences sociales, suscitant autant d'enthousiasme que de critiques. Cette présentation présente les principaux éléments de ces discussions, qui éclairent chacun d'un jour particulier la particularité de ces données et les transformations qu'elles portent pour la recherche. Elle montre aussi que cette abondance de données numériques est source de reconfigurations à l'œuvre actuellement. Quatre aspects sont plus particulièrement explorés : les réorganisations disciplinaires, les transformations des méthodes quantitatives, l'accès et la gestion des données, les objets des sciences sociales et leur rapport à la théorie.
14h35	**Marie Bergström (Ined)** **(Re)faire la sociologie du couple avec des données massives** L'étude de la formation des couples repose traditionnellement sur l'exploitation de grandes enquêtes sur la conjugalité, ou sur des données administratives. Ces sources sont utiles pour connaître la nature des unions, mais elles captent mal le processus d'appariement des partenaires. Pour pallier à ces points aveugles, les données numériques s'avèrent utiles. La communication montre comment des données issues de sites et d'applications de rencontres peuvent enrichir la sociologie du couple, en permettant une observation quantifiée des préférences et des interactions.

15h10	**Corentin Roquebert (ENS, Centre Max Weber)** **Europresse, Youtube et Genius : retour sur trois expériences de scraping** Cette communication se propose de revenir sur les liens entre problématiques de recherche et résultats empiriques avec des données du web. Au-delà de considérations techniques, les enjeux principaux sont, d'une part, de maintenir un usage des données contrôlé par des questions de recherche plutôt que l'inverse et, d'autre part, de prendre en considération les enjeux juridiques spécifiques de ces matériaux. Tout d'abord, à travers l'exemple d'un tutoriel pour mettre en forme des articles de la base de données Europresse, je reviendrai sur les enjeux de publicisation de ces techniques et les problèmes que cela engendre. Puis, à travers l'exemple d'un scraping des données du site Genius et de Youtube pour construire une analyse de réseaux du rap français, j'essaierai d'analyser les apports techniques et scientifiques de ce type de données pour la recherche.
15h50	**Julien Boelaert (CERAPS, Université de Lille)** **Extraction automatique et harmonisation de données : retour sur une étude bibliométrique à partir de CAIRN** La présentation reviendra en détail sur la méthode de récolte, structuration et curation des données bibliométriques qui ont servi de base à l'analyse publiée dans « Les aléas de l'interdisciplinarité » (avec N. Mariot, É. Ollion et J. Pagis, *Genèses* 2015/3 n° 100-101). La base recense tous les articles publiés par plus de 15 000 auteurs dans 40 revues françaises de sciences sociales (1990-2014), compilés à partir des tables de matières de revues accessibles en ligne (sur Cairn, Persee et Jstor), et permet d'étudier l'évolution de l'interdisciplinarité des sciences sociales en France sur 25 ans.
16h25	**Frederic Vergnaud (Centre de Sociologie de l'Innovation - Mines-ParisTech - PSL - CNRS UMR 9217 i3)** **Le scraping de données conversationnelles avec Extractify** Si en théorie la manière de structurer en HTML et CSS des données sur le web est plutôt bien définie par tout un ensemble de normes et de standards énoncés par différentes instances promouvant la compatibilité des technologies web, en pratique on se rend compte assez vite de la grande hétérogénéité qui prévaut dans ce domaine, rendant la plupart des méthodes et logiciels inopérants s'ils reposent sur l'identification des structures classiques pour en extraire l'information voulue. Scraper des données en ligne nécessitera par conséquent d'acquérir en amont quelques connaissances sur la structure d'une page web afin d'en repérer et sélectionner les éléments balisant les contenus à extraire. Cette présentation se propose de participer à cette acquisition en présentant les fondamentaux du HTML et du CSS. Dans un second temps, nous illustrerons notre propos en présentant un scraping réalisé à l'aide du logiciel Extractify sur des données conversationnelles issues d'un forum de discussion.
17h10	**Discussion et conclusion**

Retrouvez les présentations et toutes les informations sur le séminaire sur statapp.site.ined.fr/