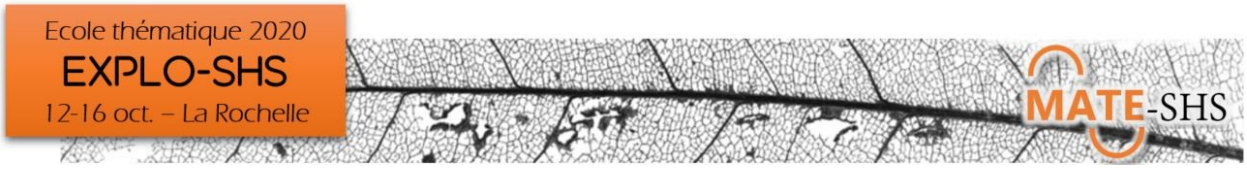


Ecole Thématique EXPLO-SHS 12-16 octobre 2020
Programme détaillé

Les dimensions épistémologiques de l'exploration

Jean-Daniel Fekete

Plénière d'ouverture, lundi 12 octobre 15h15–16h45

Présentation générale

Depuis le milieu du 20e siècle, les méthodes « exploratoires » ont pris un certain essor dans les sciences. Je présenterai plusieurs exemples de ces méthodes pour fixer les idées des participants, puis introduirai plus formellement les principes des méthodes exploratoires en les comparant avec le paradigme dominant« hypothético-déductif ».

Toujours à partir d'exemples, j'expliquerai comment suivre une démarche exploratoire, ses bénéfices et ses risques.

J'insisterai ensuite sur la validité de la démarche en décrivant les problèmes possibles, avec quelques solutions, et les critiques parfois mises en avant.

Objectifs

Introduire la notion d'exploration, sa validité épistémologique, donner des exemples, mettre en avant ses bénéfices et risques, montrer comment limiter les risques, et répondre à quelques faux procès.

Références bibliographiques

- Bertin ,Jacques, Sémiologie graphique, Paris, Mouton/Gauthier-Villars, 1967. Réédité aux Éditions de l'école des Hautes Études en Sciences Sociales (July 23, 2013)
- Bertin, Jacques, La graphique et le traitement graphique de l'information, Flammarion, 1977
- Tukey, John Wilder (1977). Exploratory Data Analysis. Addison-Wesley. ISBN 978-0-201-07616-5.
- Benzécri, Jean-Paul L'Analyse des données. Tome 1 et 2, Dunod, 1973
- Munzner, Tamara, Visualization Analysis and Design, AK Peters Visualization Series, 2014
- Tufte, Edward R. The Visual Display of Quantitative Information, Graphics Press, 2001

Principes de sémiologie graphique

Najla Touati

Focus, mardi 13 octobre 8h30–9h15

Présentation générale

Aujourd'hui, de simples logiciels permettent de générer algorithmiquement des millions d'images, sorte de "dataviz" ou tableau de bord, pour visualiser et explorer nos données. A ces images, il est souvent associé un graphisme spectaculaire mais où la sémiologie graphique -ensemble des techniques et méthodes visant à adapter un mode de représentation graphique à l'information représentée en fonction de codes et de conventions- est absente. Le résultat de cette absence : on ne voit rien et/ou on voit mal. La sémiologie graphique utilisée en phase d'exploration permet le traitement de l'information sous la forme d'une transcription graphique des données dont le but n'est pas de dessiner un graphique « une fois pour toute » mais de le re-construire et de le manipuler sans cesse jusqu'au moment où les relations perçues entre les données se dessinent.

Cette présentation s'articulera en trois temps. Une première partie sera consacrée à l'utilité d'explorer les données par la visualisation. Nous verrons en quoi les graphiques exploratoires sont des outils permettant au lecteur de trouver des relations non-attendues dans les données et en quoi ils facilitent les découvertes et l'analyse d'information pour celui qui manipule ces données.

La deuxième partie de la communication fera le point sur la notion de perception visuelle. La prise ne compte de la perception visuelle dans l'exploration des données joue un rôle important dans la visualisation par sa capacité à faire remonter rapidement et avec précisions les informations à notre cerveau. Enfin, nous verrons comment rendre l'exploration visuelle des données plus efficace par l'usage de la sémiologie graphique, en présentant les différents types de variables rétiennes et leurs applications en fonction du type de données (quantitatives ou qualitatives).

Comment explorer des corpus de textes

Bénédicte Garnier, Lucie Loubère, Gaëlle Delétraz

Atelier, mardi 13 octobre 9h30–12h15 et 13h45-16h30 (atelier répété 2 fois)

Présentation générale

L'exploration de textes avec un logiciel de statistique textuelle permet d'en déceler des structures, de produire des graphiques synthétiques ou d'extraire des verbatims.

Cette « fouille » de textes quasi automatique et très rapide nécessite de structurer les données disponibles en amont et de comprendre les méthodologies sous-jacentes issues de la linguistique (comme la lemmatisation) ou de la statistique multidimensionnelle (comme la classification automatique).

Lors de cet atelier nous illustrerons les méthodes et les appliquerons sur un corpus de réponses à une question ouverte dans une interface gratuite simple d'utilisation et qui fait référence : IraMuTeQ.

Environnement informatique

Utilisation du logiciel libre IRaMuTeQ sous Environnement PC ou MAC : http://www.iramuteq.org/

Il sera demandé aux stagiaires d'installer R et IRaMuTeQ avant l'ET. Un tutoriel d'installation et un fichier d'essai seront fournis en amont.

Type de données traitées - droit d'accès

Réponses aux questions ouvertes du questionnaire sur les attentes des participants de l'ET. Données anonymes et associées aux caractéristiques des répondants (sexe, statut, discipline, …).

Niveau requis

Aucun – Niveau initiation- pas besoin de savoir programmer.

Objectifs

- S'approprier le protocole scientifique d'exploration de données textuelles dans un outil dédié : IraMuTeQ
- Utiliser les méthodes et techniques permettant de détecter des structures dans de grands volumes de textes et les restituer
- Expliquer la méthodologie embarquée des procédures exploratoires intégrées dans l'outil : lemmatisation, classifications automatiques, extraction de verbatims (concordances)
- Interpréter et présenter les résultats.

Formule pédagogique

Atelier de 2h45 structuré en 3 temps :

- Exposé sur la méthodologie embarquée
- Présentation du logiciel et d'exemples de résultats
- Manipulation des participants sur un corpus fourni (en groupe de 2 ou 3 si les conditions le permettent) et retour d'expérience, trucs et astuce, …

Références bibliographiques

- Baril, Élodie, et France Guérin-Pace. 2016. Compétences à l'écrit des adultes et événements marquants de l'enfance : le traitement de l'enquête Information et vie quotidienne à l'aide des méthodes de la statistique textuelle. Économie et Statistique 490 (1): 17-36. https://doi.org/10.3406/estat.2016.10719.
- Baudelot, Christian, et Gollac, Michel. 1997. Faut-il travailler pour être heureux ? Insee Première, no 560. https://www.epsilon.insee.fr/jspui/bitstream/1/759/1/ip560.pdf.
- Benzécri, Jean-Paul. 1984. Description des textes et analyse documentaire. Cahiers de l'analyse des données Tome 9 (2): 205-11.
- Bonvalet, Catherine, Maison, Dominique, et Ortalda, Laurent. s. d. La place des univers familiaux, résidentiels et professionnels dans la structure du discours - Analyse textuelle des entretiens de “Proches et parents”. In La famille et ses proches, Ined-Puf diffusion. Travaux et documents 143. Ined.
- Demazière, Didier, Claire Brossaud, Patrick Trabal, et Karl Van Meter. s. d. Analyses textuelles en sociologie. Logiciels, méthodes, usages Presses universitaires de Rennes, Rennes (2006). 224 p.
- Garnier, Bénédicte, et France Guérin-Pace. 2010. Appliquer les méthodes de la statistique textuelle. Paris: CEPED. https://www.ceped.org/fr/publications-ressources/editions-du-ceped-1988-2011/les-clefs-pour/article/appliquer-les-methodes-de-la. Guérin-Pace, France. 1997. La statistique textuelle : un outil exploratoire en sciences sociales. In Population, revue de l'INED, 865-87. 4. https://www.persee.fr/doc/pop*0032-4663*1997*num*5*246*471.
- Guérin-Pace, France, Thérèse Saint Julien, et Anita W. Lau-Bignon. 2012. Une analyse lexicale des titres et mots-clés de 1972 à 2010. Espace géographique 41 (1): 4. https://doi.org/10.3917/eg.411.0004. http://lexicometrica.univ-paris3.fr/. s. d.
- Lebart, Ludovic, Bénédicte Pincemin, et Céline Poudat. 2019. Analyse des données textuelles. Mesure et évaluation 11. Québec: Presses de l'Université du Québec.
- Lebart, Ludovic, et André Salem. 1994. Statistique textuelle. Paris: Dunod. http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html.
- Reinert, Max. 1983. Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Cahiers de l'analyse des données 8 (2): 187-98.
- Reinert, Max. 1993. Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. Langage & société 66 (1): 5-39. https://doi.org/10.3406/lsoc.1993.2632.
- Reinert, Max. 2001. Alceste, une méthode statistique et sémiotique d'analyse de discours ; application aux ‘Rêveries du promeneur solitaire'. La revue française de psychiatrie et de psychologie médicale 49: 32-36.
- Reinert, Max. 2008. Mondes lexicaux stabilisés et analyse statistique de discours.

Comment explorer des données relationnelles ou de réseaux

Jean-Daniel Fekete, Pascal Cristofoli

Atelier, mardi 13 octobre 9h30–12h15 et 13h45-16h30 (atelier répété 2 fois)

Présentation générale

En SHS, il est fréquent de vouloir comprendre un phénomène en étudiant les relations qui associent ses artefacts. L'analyse des données relationnelles permet de mettre en œuvre cette démarche pour laquelle les données potentiellement mobilisables sont pléthoriques. En général, elles sont crées à partir de documents ou de dispositifs techniques qui référencent des personnes, des lieux et des organisations, ou encore à partir de protocoles d'enquêtes spécifiques. Elles peuvent concerner de nombreux sujets et s'intéresser à des types de relations très variés.

Nous allons rappeler brièvement l'historique de l'analyse de réseaux, en particulier sociaux, et montrer ensuite comment des réseaux peuvent être constitués et explorés, essentiellement visuellement. Nous mettrons l'accent sur la séparation entre l'exploration et la présentation à l'aide de visualisations.

Nous aborderons le cours en partant de plusieurs types de données classiques, stockées sous forme de tables de données, et nous montrerons comment les visualiser pour les explorer. Nous commencerons par la représentation à base de nœuds et liens, classique mais limitée, et montrerons les représentations alternatives qui permettent une meilleure exploration tout en étant souvent moins efficaces pour la présentation des résultats dans un article.

Nous montrerons comment visualiser et explorer des réseaux simples ou d'autres plus complexes: réseaux à un ou plusieurs modes, avec un nombre d'attributs et de relations plus ou moins grand, et/ou qui évoluent dans le temps (réseaux dynamiques ou longitudinaux).

L'atelier utilisera des logiciels faciles à utiliser et à s'approprier : en particulier les applications web The Vistorian et PAOVis, ainsi que le prototype PK-Clustering.

Environnement informatique

Les outils d'exploration des réseaux abordés dans l'atelier sont libres ou gratuits. Il s'agit d'applications web ne demandant aucune installation locale, mais nécessitant l'utilisation d'un navigateur web récent (nous conseillons d'installer la dernière version de Google Chrome : https://www.google.com/chrome/).

Pour interagir facilement avec les applications, il est aussi fortement conseillé de se munir d'une souris.

Le format de données en entrée de ces applications est le format tabulaire CSV. Les fichiers de ce type peuvent être crées dans un éditeur de texte ou un tableur (Libre office ou Open office de préférence, ou bien encore Excel).

Type de données traitées - droit d'accès

L'atelier se fonde sur l'exploitation de données relationnelles. Nous donnerons de nombreux exemples de types de réseaux et présenterons la façon de les collecter et de les coder dans un format tabulaire (type CSV). Les participants sont fortement encouragés à travailler avec leurs propres données. Ces données sont traitées par les logiciels que nous utiliserons sans être transmises à distance : elles restent sur le navigateur web de la machine et ne risquent pas de « fuiter » par erreur. L'application PK-Clustering les transfert à un serveur qui ne les garde pas après la session de travail.

Niveau requis

Aucune connaissance informatique n'est requise pour assister à l'atelier et utiliser les outils, il n'y aura pas de programmation ni d'utilisation de langage informatique. L'atelier suppose de comprendre le modèle relationnel et sa structuration en données tabulaires (ce qui sera rappelé en début de séance à l'aide d'exemples) ; le fait d'avoir préalablement constitué son propre corpus permettra aux auditeurs de mieux prendre en main les applications et s'approprier la démarche et les questions.

Objectifs

L'atelier vise à présenter les méthodes et outils dédiés à l'exploration de données relationnelles et de réseaux. Il propose d'utiliser quelques logiciels simples d'exploration interactive de réseaux, et de montrer comment les multiples représentations visuelles permettent d'explorer différents types de réseaux et, en diversifiant les points de vue, de faire progressivement connaissance avec un corpus de données.

Formule pédagogique

Les participants sont invités à utiliser leurs propres données pendant l'atelier. Le plus simple est de les saisir sous la forme de fichiers textes au format CSV ou dans un tableur en les exportant au format CSV (pour plus d'informations, consulter les pages de présentations des outils).

Références bibliographiques

- Edward Tufte, The Visual Display of Quantitative Information
- Alberto Cairo, The Functional Art
- Nathan Yau, Visualize This: The FlowingData Guide to Design, Visualization, and Statistics
- Tamara Munzner, Visualization Analysis and Design
- Wasserman, Stanley & Faust, Katherine (1994). Social Networks Analysis: Methods and Applications
- Di Battista, Giuseppe; Eades, Peter; Tamassia, Roberto; Tollis, Ioannis G. (1998), Graph Drawing: Algorithms for the Visualization of Graphs
- von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J., Fekete, J.-D. and Fellner, D. (2011), Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. Computer Graphics Forum, 30: 1719-1749. doi:10.1111/j.1467-8659.2011.01898.x

Liens

Les applications web abordées dans l'atelier sont présentées sur les pages suivantes :

- The Vistorian : https://vistorian.net/
- PAOHVis: https://www.aviz.fr/Research/Paohvis
- PK-Clustering : https://aviz.fr/pkclustering

Comment explorer des données à l'aide de graphiques statistiques

Maël Theulière (en visio), Hugues Pécout

Atelier, mardi 13 octobre 9h30-12h15 et 13h45-16h30 (atelier répété 2 fois)

Présentation générale
Cet atelier vise à initier les participants à la constructions de graphiques statistiques et de visualisations à partir de la logique de « Grammar of Graphics » initié par Leland Wilkinson dans les années 80.

Environnement informatique
logiciel R : ggplot2

Type de données traitées - droit d'accès
Données statistiques quantitatives et qualitatives

Niveau requis
Méthodologique : Connaissance de base en analyse univariée, bivariée

Technique: Savoir à minima manipuler des données sous R ou sous EXCEL

Objectifs
Parcourir les potentialités de représentation graphiques à partir de la logique Grammar of Graphics (https://www.r-graph-gallery.com/index.html , https://www.data-to-viz.com/)

Formule pédagogique
Atelier "learning by doing"!!

Références bibliographiques
https://www.routledge.com/R-Graphics-Third-Edition/Murrell/p/book/9781498789059
https://ggplot2-book.org/
https://clauswilke.com/dataviz/
https://mtes-mct.github.io/parcoursrmodule_datavisualisation/
https://socviz.co/

Liens
R (https://www.r-project.org/), Rstudio (https://rstudio.com/products/rstudio/download/), la liste des packages nécessaires au TP sera fournie en arrivant.

Bonnes pratiques et cadre réglementaire

Isabelle André-Poyaud

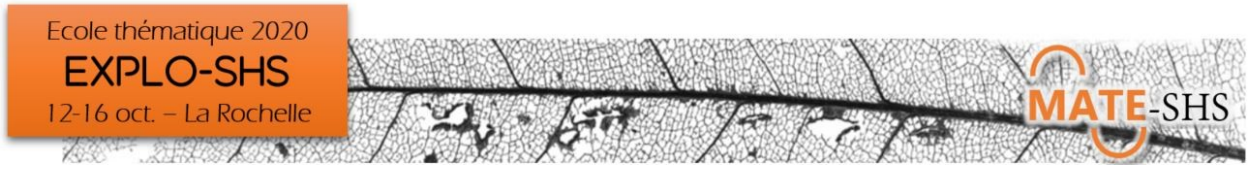
Focus, mercredi 14 octobre 8h30–9h15

Présentation générale

Dans un contexte d'interdisciplinarité des projets, de massification des données, de constitution de corpus multi-sources et d'open data, préparer des données et des métadonnées de qualité devient un enjeu majeur de nos pratiques de recherche. Cette qualité commence dès la phase de collecte en prenant en considération le cadre juridique associé aux données collectées et se poursuit à toutes les étapes d'un projet.

Dans cette séance, nous nous focaliserons particulièrement sur les données à caractère personnel si souvent mobilisées dans les recherches SHS. Depuis mai 2018, le RGDP encadre le traitement de ces données. Après une présentation des principes clés du RGPD, nous nous appuierons sur des exemples concrets pour savoir comment répondre au mieux à cette mise en conformité qui met l'accent sur le droit des personnes et la sécurité de leurs données.

Nous terminerons cette séance par une ouverture sur les plans de gestion des données. Nous verrons en quoi cet outil peut s'avérer utile dans le suivi des données que celles-ci soient des données à caractère personnel ou non en nous invitant à nous questionner sur nos corpus dès la conception du projet de recherche et ensuite à des étapes clés du projet : acquisition, traitements, partage, réutilisation potentielle... Cela permet soit d'aller vers l'ouverture des données (open data) lorsque cela est possible ou la restriction d'ouverture lorsque cela est nécessaire (ex : données non anonymisées ou sensibles).



Utiliser le nettoyage des données pour explorer, rendre compte d'un potentiel scientifique via le langage Python avec le module Pandas

Antoine Mazières (en visio), Julie Pierson

Atelier, mercredi 14 octobre 9h30-12h15

Présentation générale

Cet atelier vise à initier ses participants à la programmation avec le langage Python. Un intérêt particulier sera porté à l'analyse de données et la bibliothèque Pandas qui y est dédiée. Les débutants pourront apprendre quelques bases et découvrir ce à quoi ils peuvent aspirer, tandis que les participants plus expérimentés pourront expérimenter des fonctions plus avancées et faire un tour d'horizon pratique des possibilités offertes par cet ensemble d'outils.

Environnement informatique

Pour participer à cet atelier vous avez besoin d'installer au préalable Anaconda Python : https://www.anaconda.com/products/individual

Type de données traitées - droit d'accès

Utilisation de jeux de données ouverts, déjà formatés (XLS, CSV, JSON, etc.).

Niveau requis

Tous niveaux.

Objectifs

Initier à la programmation, au langage Python et à la bibliothèque Pandas qui permet de facilement traiter des données. Donner un aperçu de ce qu'il est possible de faire à terme avec ces outils, ce mais permettra aux utilisateurs plus avancés de pouvoir découvrir aussi des choses.

Liens

Pour en savoir plus sur l'intervenant : https://www.antonomase.fr/

Comment utiliser le nettoyage des données pour explorer, rendre compte d'un potentiel scientifique : le langage R

Maël Theulière (en visio), Hélène Mathian

Atelier, mercredi 14 octobre 9h30-12h15

Présentation générale

Nous souhaiterions aborder au cours de cet atelier un certain nombre de « procédures » qui permettent le nettoyage des données , mais qui nécessairement les interrogent et nous conduit à faire des choix de recodage, à construire des modèles de lecture des enregistrements.

En particulier les premières explorations s'attachent à explorer la cohérence des données. Ces tests peuvent être de simples tris à plat par exemple, ou être plus complexes.

Dans un premier temps on s'attachera à illustrer :

- Le recodage de données manquantes
- Le repérage d' « outliers »
- Le nettoyage et le recodage de chaînes de caractères

Dans une deuxième partie on abordera des tests de cohérences sur des dimensions plus spécifiques que sont le temps et l'espace.

Environnement informatique

R et Rstudio

Type de données traitées - droit d'accès

Données territoriales, données open data , données réseaux sociaux

Niveau requis

Être à l'aise avec la manipulation de données en général et avec le package dplyr sous R.

Objectifs

Illustrer quelques pratiques de "nettoyage" de données statistiques

Formule pédagogique

Fourniture des données préparées et des codes. Identification des étapes et discussion autour des choix de la mise en œuvre.

Liens

R (https://www.r-project.org/) , Rstudio (https://rstudio.com/products/rstudio/download/) la liste des packages nécessaires au TP sera fournie en arrivant.

Une méthode mixte à visée exploratoire : détermination de registres discursifs et recherche de liens avec des strates de population ou des variables

Gaëlle Delétraz, Bénédicte Garnier

Atelier, mercredi 14 octobre 9h30-12h15

Présentation générale

Cet atelier propose de découvrir certains des apports des logiciels de statistique textuelle dans une visée exploratoire, c'est-à-dire afin de naviguer dans un ou plusieurs corpus, de les découvrir, de formuler une première série d'hypothèses ou d'analyses au cours de leur lecture. Dans cette perspective, s'inspirer des fonctionnalités de codage et d'annotation offertes par une autre famille de logiciels (les CAQDAS) constitue un complément utile aux fonctionnalités automatiques au cœur des outils de textométrie. L'objectif est de repérer des registres discursifs plus finement que ne le permet une méthode automatique seule mais en permettant des choix et une orientation totalement ouverte aux objectifs poursuivis par le/la chercheur.e.

Cette détermination mixte de registres discursifs permet alors une exploration du corpus à la recherche de liens statistiques entre ces registres discursifs et des variables associées, des strates de la population étudiée, d'autres registres discursifs ou des paramètres temporels par exemple. Ces résultats exploratoires peuvent confirmer, réorienter les analyses fines ultérieures, voire ouvrir de nouvelles pistes.

Environnement informatique

ATTENTION : Windows uniquement. Nécessité d'un émulateur sous Mac.

Logiciel : SphinxIQ2 + module Quali : logiciel propriétaire. Ce logiciel est disponible pour les acteurs académiques de la recherche via la TGIR Huma-Num.

Installation du logiciel dans le cadre de l'école : Dans le cadre de l'école thématique, la société Le Sphinx met à disposition 20 clés de formation afin de faciliter la logistique logicielle de l'atelier. Il suffira de télécharger le logiciel sur le site et d'utiliser la clé temporaire qui sera transmise aux stagiaires. Les clés seront fonctionnelles jusqu'au 31/12/2020 pour permettre aux stagiaires de poursuivre leur découverte de l'outil après la formation + disponible via Huma-Num.

Type de données traitées - droit d'accès

Données textuelles de tous types : discours, entretien, focus groupes, extraction du web etc. associées ou non à des variables descriptives, sous forme tableur ou formatées selon un balisage assez simple sous Word ou autre format texte.

Niveau requis

Aucun pré-requis nécessaire.

Objectifs

- S'approprier un/des cheminements d'exploration de données qualitatives/textuelles avec le module « quali » de SphinxIQ2
- Étiqueter des unités textuelles selon des méthodes automatiques (thesaurus, regroupement par racine etc.)

Regrouper des mots ou expressions manuellement selon ses objectifs/hypothèses de recherche

- Créer des variables lexicales sur la base des regroupements de mots ou expressions • Croiser ces variables lexicales à toutes autres variables disponibles : variable associée, autre variable lexicale, variable temporelle, etc.
- Repérer les liaisons statistiques (khi-deux) entre variables lexicales créées
- Interpréter et présenter les pistes explorées

Formule pédagogique

Atelier de 2h45 structuré en 3 temps :

Présentation générale de la démarche et articulation quanti/quali

- Présentation du logiciel, de l'organisation de ses fonctionnalités
- Réalisation guidée du traitement par les participant.e.s sur un corpus fourni : 8000 réponses à une question ouverte portant sur le ressenti durant le confinement.

Références bibliographiques

- Bardin, Laurence. 2013. L'Analyse de contenu. Quadrige. Paris: Presses Universitaires de France.
- Lejeune, Christophe. 2019. Manuel d'analyse qualitative. Analyser sans compter ni classer. 2eme ed., Louvain-la-Neuve: De Boeck.
- Ganassali, Stéphane. 2014. Enquêtes et analyse de données avec Sphinx. Pearson Ed.
- Paillé, Pierre, et Alex Mucchielli. 2012. L'analyse qualitative en sciences humaines et sociales. U. Paris: Armand Colin.

Liens

https://www.lesphinx-developpement.fr/logiciels/enquete-analyse-sphinx-iq/

Explorer les données spatiales, géovisualiser

Hélène Mathian, Delphine Montagne

Atelier, mercredi 14 octobre 13h45-16h30

Présentation générale

L'atelier vise à sensibiliser à la notion d'exploration des organisations spatiales dans une démarche typique de "l'analyse spatiale", c'est à dire une démarche qui consiste à questionner comment un phénomène varie dans l'espace, identifier les niveaux d'organisation de ce phénomène. Si les SIG sont des outils majeurs de gestion et d'analyse de l'information géographique, ils nécessitent des compétences en géomatique pas forcément nécessaires à l'exploration statistiques de la dimension spatiale. Nous proposons de donner un aperçu de cette démarche en mobilisant des environnements existant (GéoDa, ExploratR).

Cet atelier visera aussi à réfléchir en atelier sur la conception de l'exploration de la dimension spatiale…

Environnement informatique

(liste prévisionnelle)

- GeoDa (logiciel à installer)
- ExploratR (environnement en ligne)
- Carto (plateforme avec inscription)

Type de données traitées - droit d'accès

- Données administratives sur Lyon
- Données de récit
- …

Niveau requis

Statistiques de base (univariées, bivariées, éventuellement multivariées) Les données seront fournies au bon format

Objectifs

Sensibiliser aux formes de la dimension spatiale que l'on peut explorer.

Formule pédagogique

- Intro méthodologique
- Manipulations
- Conception

Références bibliographiques

- Saint-Julien T., Pumain D. 1997 L'analyse spatiale: 1.localisation dans l'espace, Cursus, Armand Colin.
- Saint-Julien T., Pumain D. 2001 Les interactions spatiales, Cursus, Armand Colin.
- Feuillet Thierry - Cossart Etienne - Commenges Hadrien, 2019, Manuel de géographie quantitative - concepts, outils, méthodes

Liens

- GeoDa- au sujet de-> https://spatial.uchicago.edu/geoda pour télécharger le logiciel- http://geodacenter.github.io/download.html
- ExploratR - https://analytics.huma-num.fr/geographie-cites/ExploratR/
- CARTO - au sujet de-> https://carto.com/ pour se créer un compte d'essai gratuit: https://carto.com/signup/

Explorer des données historiques ou archéologiques

Pascal Cristofoli, Mélanie Lecouedic

Atelier, mercredi 14 octobre 13h45-16h30

Présentation générale

L'atelier proposé est envisagé comme une mise en perspective des potentialités offertes par les grands volumes de données désormais accessibles, sous forme numérique à l'aune des heuristiques et des méthodes élaborées et appliquées en histoire et en archéologie.

Comment peut-on construire une enquête scientifique sur une société passée, donc inaccessible par la voie d'une observation directe, par la passation de questionnaires ou encore la conduite d'entretiens ? De quelle manière est-il possible d'exploiter les « traces » des activités humaines, que celles-ci soient documentaires ou matérielles, pour répondre à une problématique de recherche ? Comment explorer un terrain ou des corpus de sources hétérogènes, comment « abstraire » des informations, soit formaliser et modéliser, pour « construire » des « données », et selon quelles conditions les analyser ?

L'atelier proposera une alternance de présentations et de petits exercices pour aborder quelques-unes des étapes saillantes d'une telle démarche et exposer leur mise en œuvre à l'aide d'outils simples visant à :

- L'exploitation des documents, la construction d'un échantillon, le croisement de sources hétérogènes et l'organisation des bases de données nominatives en histoire.
- La question du nettoyage/codage des données et celle de l'appariement des noms pour tenter de reconstituer des trajectoires d'individus (familiales, sociales, professionnelles, géographiques,…) ou de toponymes (parcelles, bâtiments, lieux)
- Les implications de la prise en compte de la dimension temporelle portée par les sources, tant dans l'élaboration des données que dans leur analyse.
- L'importance d'une réflexion sur les données manquantes, l'incertitude et/ou l'imprécision des informations recueillies ainsi que sur la manière de travailler en tenant compte de ce contexte.

Environnement informatique

- Tableurs
- **Openrefine** ; Application java multiplateforme qui fonctionne avec le navigateur par défaut de votre ordinateur. https://openrefine.org/download.html
- Précisions sur l'installation d'Openrefine : https://msaby.gitlab.io/formation-openrefine-Lyon20191122/installation-lancement-d%C3%A9sinstallation.html
- **Yed :** programme de création de schémas et de graphiques multiplateforme, permettant la réalisation de schémas et de graphiques UML (Unified Modeling Language)

Exploration des données à l'aide des arbres de décision

Grégoire Le Campion, Hugues Pécout

Atelier, mercredi 14 octobre 13h45-16h30

Présentation générale

Il existe en statistiques un certain nombre d'outils pour tenter de prédire, d'expliquer, classer des variables. Les méthodes les plus connues sont notamment l'analyse factorielle, ou encore les méthodes de régression.

Ces différentes méthodes bien qu'extrêmement intéressantes ont de nombreuses conditions qu'il n'est pas toujours aisé de remplir en SHS, et fournissent des résultats pas toujours simple à interpréter.

L'idée ici est de vous présenter une méthode alternative et tout-terrain : l'arbre de décision avec ses avantages et ses limites.

Environnement informatique

Tout environnement (Linux, Mac, Windows), en revanche nécessité d'avoir installer sur l'ordinateur un navigateur internet (firefox, chrome…). Dans l'optique d'un éventuel approfondissement du sujet (facultatif) R et Rstudio seront nécessaires.

Type de données traitées - droit d'accès

Données quantitatives et qualitatives, mais non textuelles, mise sous forme base de données.

Niveau requis

Aucun pré-requis nécessaire

Objectifs

Comprendre ce qu'est et permet un arbre de décision. Pouvoir utiliser cette méthode d'analyse à l'issue de l'atelier de manière autonome

Formule pédagogique

Atelier « bring your data »

Références bibliographiques

- T.Hastie, R.Tibshiraniet J.Friedman : The Elementsoff Statistical Learning : Data Mining, Inference, andPrediction.. Springer, 2nd ed., 2009
- T.Hastie, R.Tibshiraniet J.Friedman : The Elementsof Statistical Learning : Data Mining, Inference, andPrediction. Springer, 2nd ed., 2009

Explorer en codant – dialogue entre informatique et SHS

Guy Mélançon

Séance plénière, mercredi 14 octobre 17h-18h30

Présentation générale

La présentation qui précédera le « débat » reviendra sur l'expérience de l'orateur, informaticien et chercheur en « visual network analytics », à collaborer en contexte pluridisciplinaire avec les SHS. Les quelques leçons apprises au fil des collaborations, anecdotes, « guidelines » et bonnes pratiques pourront être mises au défi par l'auditoire. Deux affirmations pourront être utiles déjà pour situer l'intervention : « Les algorithmes ont peu de valeur sans données de qualité. Les données à elles seules ne permettent souvent pas de poser de bonnes questions. »

Niveau requis

Dans le but de donner un cadre au débat, d'alimenter les réflexions, je ferai une présentation d'une quarantaine de minutes en soulignant quelques idées empruntées à la communauté, dont certaines sont peut-être miennes. Je supposerai de l'auditoire qu'il s'est déjà frotté à la manipulation de données de recherche, voire s'est peut-être un peu frotté à la programmation.

Objectifs

Donner un éclairage peut-être nouveau sur la relation partenariale qui se tisse nécessairement dans toute collaboration entre informatique et disciplines SHS. Tenter de sortir d'une perspective « consumériste » de l'informatique par les SHS – l'informatique vu comme seul outil – tout comme d'une perspective faisant des SHS le prête-nom d'une pluridisciplinarité de surface.

Donner une lecture, depuis l'intérieur et en se référant à l'expérience de l'orateur, des enjeux du travail pluridisciplinaire ; proposer peut-être une (des) ligne(s) directrice(s) ou de bonnes pratiques du travail dans un tel contexte. Ce chemin nous amènera à décrire un contexte de travail où l'activité de codage vient épauler une démarche exploratoire où les questions sont à la fois affinées par ceux qui les amènent (les SHS) à mesure qu'elles sont challengées par les données, et par ricochet, par ceux qui les manipulent.

Formule pédagogique

Exposé « magistral » au démarrage du débat, puis questions/réponses avec la salle – éventuellement sollicitation de la salle au cous de la présentation.

Références bibliographiques

- Brehmer, M. and T. Munzner (2013). "A multi-level typology of abstract visualization tasks." Visualization and Computer Graphics, IEEE Transactions on 19(12): 2376-2385.
- Gray, J. (2007). Jim Gray on eScience: A Transformed Scientific Method. The Fourth Paradigm: Data-Intensive Scientific Discovery. (Transcript by T. Hey, S. Tansley and K. Tolle, Microsoft Research.
- Ollion, E. et J. Boelaert (2015). Au-delà des big data. Les sciences sociales et la multiplication des données numériques. Sociologie (3, vol. 6).
- Meyer, M., et al. (2012). The four-level nested model revisited: blocks and guidelines. Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization. Seattle, Washington, ACM: 1-6.
- Munzner, T. (2009). "A Nested Process Model for Visualization Design and Validation." IEEE Transactions on Visualization and Computer Graphics 15: 921-928.
- Pentland, Alex (2015). Social Physics – How Social Networks Can Make Us Smarter. Penguin Books, ISBN 9780143126331.

Massification des données, structures et langages du web : ce qu'il faut savoir avant de se lancer

Frédéric Vergnaud, Antoine Mazières (en visio) et Benjamin Ooghe Tabanou

Focus, jeudi 15 octobre 8h30-9h15

Présentation générale

Afin de donner aux stagiaires quelques repères fondamentaux sur le lexique et les technologies qui seront utilisés au cours des trois ateliers du jeudi matin, ce focus se propose dans un premier temps de jeter les bases du fonctionnement d'Internet, du web et de la structuration (HTML) et forme (CSS) d'une page web.

Dans un second temps, nous examinerons la manière de cibler et d'extraire des éléments de ces pages (scraping) grâce au modèle de page DOM et au langage de requête XPATH qui permet de s'y déplacer.

Enfin, dans une troisième partie, nous verrons en quoi le crawling se distingue du scraping et comment la structure hypertextuelle du web reposant sur les liens peut permettre de nouvelles formes d'analyse.

Fouiller son terrain en explorant le web avec Hyphe

Benjamin Ooghe Tabanou, Viviane Le Hay

Atelier, jeudi 15 octobre 9h30-12h15

Présentation générale
Le web nous oppose des défis à la fois méthodologiques et technologiques. Le médialab de Sciences Po a développé, et publié sous la forme d'un logiciel libre, Hyphe, un robot amasseur de données web aussi appelé «crawler», ajusté aux besoins de la recherche en sciences sociales. Il s'adresse aux sociologues qui veulent investiguer le web comme terrain d'enquête qualitative et en tirer des indicateurs quantitatifs. S'appuyant sur le modèle du web «en couches», il guide son utilisateur pour offrir aux chercheurs et étudiants un outil de création et nettoyage, itération après itération, un corpus de ressources et/ou d'acteurs.

La séance proposera tout d'abord une présentation générale de la méthode et de l'outil, ainsi qu'une rapide démonstration d'usage de Hyphe avant une mise en pratique par les participants.

Environnement informatique
Un ordinateur personnel avec un navigateur web connecté à Internet.

Type de données traitées - droit d'accès
Métadonnées sur des sites web et réseaux de liens entre les pages de ces sites. Libre accès.

Niveau requis
Tout public, simple connaissance des bases de fonctionnement du web préférable.

Objectifs
Découverte et utilisation exploratoire de l'outil Hyphe avec réalisation d'un premier corpus personnel sur la thématique de son choix.

Formule pédagogique
Travaux pratiques en petits groupes sur le logiciel après un exemple d'usage.

Références bibliographiques

- OOGHE-TABANOU, Benjamin, JACOMY, Mathieu, GIRARD, Paul & PLIQUE, Guillaume, "Hyperlink is not dead!", In Proceedings of the 2nd International Conference on Web Studies (WS.2 2018), Everardo Reyes, Mark Bernstein, Giancarlo Ruffo, and Imad Saleh (Eds.). ACM, New York, NY, USA, 12-18. DOI: https://doi.org/10.1145/3240431.3240434
- JACOMY, Mathieu, GIRARD, Paul, OOGHE-TABANOU, Benjamin, et al, "Hyphe, a curation-oriented approach to web crawling for the social sciences.", in International AAAI Conference on Web and Social Media. Association for the Advancement of Artificial Intelligence, 2016.

Liens
http://hyphe.medialab.sciences-po.fr/demo/
https://github.com/medialab/hyphe-browser/releases/tag/v1.1

Le scraping de données structurées web à l'aide d'Extractify. Focus sur les données conversationnelles

Frédéric Vergnaud, Pascal Cristofoli

Atelier, jeudi 15 octobre 9h30-12h15

Présentation générale

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.

L'atelier présente le logiciel libre Extractify, un plugin pour le navigateur Chrome, qui se propose de fournir à son utilisateur une interface simplifiée lui permettant de récolter n'importe quel type de données structurées en ligne. Après avoir décrit le logiciel, nous en étudierons les fonctions automatiques d'identification des structures HTML englobant les données recherchées. Dans un second temps, nous verrons qu'il est possible d'aller plus loin en utilisant les sélecteurs CSS. Enfin, dans le cadre d'un focus sur des données issues de forums de discussions, nous utiliserons le logiciel libre L@ME pour visualiser les données extraites et les exporter en vue de traitements statistiques ultérieurs.

Environnement informatique

Extractify : Liaison Internet & Navigateur Chrome L@ME : Java 7

Type de données traitées - droit d'accès

Tout type de données structurées et librement accessibles sur le web. Focus sur des forums de discussions publics.

Niveau requis

Tous niveaux. Un tour d'horizon rapide des sélecteurs CSS sera nécessaire pour les fonctions avancées.

Objectifs

Initiation au « scraping » de données en sciences sociales à l'aide d'une suite de logiciels libres développés pour des non-initiés.

Formule pédagogique

Après une présentation générale et un premier exemple rapide suivi par tous, chaque stagiaire pourra s'exercer à récolter son propre corpus.

Références bibliographiques

Frédéric Vergnaud. L@ME : un logiciel libre d'analyse et de traitement de messages électroniques. Tuto@MATE, 2017. ⟨hal-02393861⟩

Liens

https://github.com/fredericvergnaud/extractify
https://github.com/fredericvergnaud/lame

Collecter des données sur le Web avec Python

Antoine Mazières (en visio), Julie Pierson

Atelier, jeudi 15 octobre 9h30-12h15

Présentation générale
Cet atelier vise à initier ses participants à la construction d'une base de données à partir de source hétérogènes et non-structurées. Un intérêt particulier sera porté à l'acquisition de données depuis le Web et des APIs diverses, et aux bibliothèques LXML, Requests et au langage XPATH qui y sont dédiées. Les débutants pourront apprendre quelques bases et découvrir ce à quoi ils peuvent aspirer, tandis que les participants plus expérimentés pourront expérimenter des fonctions plus avancées et faire un tour d'horizon pratique des possibilités offertes par cet ensemble d'outils.

Environnement informatique
Pour participer à cet atelier vous avez besoin d'installer au préalable Anaconda Python : https://www.anaconda.com/products/individual

Type de données traitées - droit d'accès
Utilisation de sources de données publiques.

Niveau requis
Tous niveaux.

Objectifs
Avoir une idée de ce qu'il est possible de collecter ou pas sur le web, et pouvoir évaluer l'effort que cela implique. Présentation d'une suite d'outils pour ce faire (lxml, xpath, etc.). Réalisation de plusieurs exemples avec des degrés de difficulté graduels.

Liens
Pour en savoir plus sur l'intervenant : https://www.antonomase.fr/

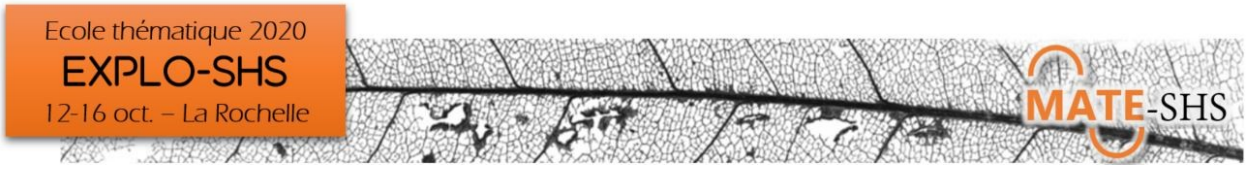
Explorer à l’aide du Web sémantique

Stéphane Pouyllau (en visio)

Séance plénière, jeudi 15 octobre 15h45-16h30

Présentation générale

Le Web sémantique est l'un des piliers des données et informations structurées sur le Web. Au cœur d’un grand nombre de réservoirs de données et de document, il permet une exploration structurée, massive et comparative des données. Associé aux méthodes et services utilisant l’apprentissage profond (Deep Learning) il offre un potentiel d’exploration comparative et cumulative qui dépasse le cadre strict de la science des données pour intéresser la recherche dans les humanités et les sciences sociales. L’intervention présentera ces potentiels tout en revenant sur le chemin parcouru depuis 10 ans dans ce secteur.



Comment faire des sciences sociales à partir des traces textuelles du web ?

Sylvain Parasie

Séance plénière, jeudi 15 octobre 17h-18h30

Présentation générale

L'essor du web et des réseaux sociaux offre aux chercheurs en sciences sociales un volume considérable de matériaux textuels qui sont le support d'un registre étendu d'activités sociales. Du fait de l'étendue des objets qu'ils permettent d'embrasser et de leur caractère non sollicité, ces matériaux intéressent un nombre croissant de chercheurs en sciences sociales.

Cet élargissement des sources de l'enquête en sciences sociales se heurte toutefois à plusieurs obstacles : (1) ces matériaux textuels sont produits par des plateformes numériques qui sollicitent, encadrent et mettent en forme les expressions et les échanges, mais contrôlent aussi la façon dont les chercheurs y ont accès. (2) l'essor de nouvelles techniques d'analyse textuelle issues des mondes informatiques ajoute un trouble supplémentaire lié à l'opacité des algorithmes et leur difficile appropriation par les chercheurs en sciences sociales. (3) l'ancrage social des personnes qui s'expriment sur le web et les réseaux sociaux demeure en grande partie inconnu. Font souvent défaut des informations aussi cruciales pour l'enquête que le niveau de revenus ou de diplôme, la catégorie sociale, et même l'âge ou le genre de ceux et celles qui prennent la parole en ligne.

Dans cette intervention, nous proposons un parcours critique d'un ensemble de recherches contemporains de sciences sociales, qui explorent des méthodes originales pour surmonter les obstacles associés au traitement quantitatif des matériaux textuels issus du web. Une attention particulière sera portée à la façon dont ces recherches parviennent à redonner une épaisseur sociale aux traces textuelles, et ainsi à mettre ces techniques au service d'un questionnement de sciences sociales.