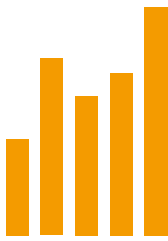


Du 7 au 11 décembre 2020

SEM AINE



DATA/SHS

Maison des Sciences de l'Homme de Dijon

POUR PLUS D'INFORMATIONS

➔ CLIQUEZ ICI

Programme

LUNDI 07/12	MARDI 08/12	MERCREDI 09/12	JEUDI 10/12	VENDREDI 11/12
<p>9h30 Accueil</p> <p>10h00 Ouverture - Jean VIGREUX, Directeur MSH Dijon - Jimmy LOPEZ, Responsable scientifique PUDD</p> <p>Invité : Pascal Neige, Vice Président de l'université de Bourgogne</p>	<p>SESSION 2 10h00-11h00</p> <p>Présentation 3 : L. Finez (Psy-DREPI) « Les enquêtes quantitatives : de la corrélation à la causalité. »</p>	<p>09h00-12h00</p> <p>• CONFÉRENCE 2 : Textométrie et application S. Heiden (ENS Lyon & IHRIM) & V. Lethier (MCF en Sciences du langage, ELLIADD, Université de Franche-Comté)</p>	<p>SESSION 5 9h30-10h30</p> <p>Présentation 10 : D. Andolfatto (CREDSEPO) & D. Labbé (Pacte CNRS, UGA) « Covid-19 : ce que nous apprennent les statistiques hospitalières »</p>	<p>9h00-11h00</p> <p>• CONFÉRENCE 3 : SMART CITY Comment FAIR.E métropole dans une ville Smart : D. Hameau, N. Martiny, Y. Richard, A. Jégou, J. Pergaud, C. Cruz, B. Pohl, T. Thevenin</p>
<p>10h00-12h00</p> <p>• CONFÉRENCE 1 : S. Lecorgne (CDSP, Sciences Po) & E. Zolotoukhine (ADISP, TGIR PROGEDO) « Les statistiques publiques : données ESS, ADISP »</p>	<p>11h00-12h00</p> <p>Présentation 4 : H. Marchal (LIR3S) & Louis Olagnon (co-fondateur de Cartodébat) « Retour critique sur une démarche de participation citoyenne »</p>		<p>10h30-11h30</p> <p>Présentation 11 : L. de Mesnard (CREGO) « Les méthodes biproportionnelles, un outil de reconstitution ou de comparaison de tableaux de contingence »</p>	<p>SESSION 7 11h15-12h15</p> <p>Présentation 14 : B. Quost, L. Granjon, L. Pillot, L. Saligny, E. Fovet, E. Chevigny, J. Vidal, D. Goguey, Y. Pautrat, J. Benard, M. Delcamp (GéoBFC) « Un LiDAR pour le Parc national de forêts : Présentation de la démarche et focus sur la problématique archéologique »</p>
<i>Déjeuner</i>				
<p>13h30-15h00</p> <p>• CONFÉRENCE 1 (suite) : J. Lopez (LEDi) « Données CASD »</p>	<p>SESSION 3 13h30-14h30</p> <p>Présentation 5 : A. Millereux (MSH) & T. Thevenin (MSH, ThéMA) « Acquisition de données géohistoriques : entre automatisation et intervention humaine »</p>	<p>SESSION 4 13h30-14h30</p> <p>Présentation 8 : V. Liegeois (TIL) « Bulletins météo télévisés et vulgarisation scientifique - Analyse critique des données (non-) disponibles »</p>	<p>SESSION 6 13h30-14h30</p> <p>Présentation 12 : C. Peyron & A. Péliissier (LEDi) « L'apport et la mise en œuvre de la méthode des choix discrets pour analyser les préférences individuelle »</p>	<p>12h15-13h00</p> <p>Présentation 15 : M. Bach (TIL ; MSH Dijon), A. Da Costa (MSH Dijon) « Twitter et la linguistique située : réflexions méthodologiques à partir de l'exemple de tweets sur la métropole de Dijon »</p>
<p>SESSION 1 Présentation 1 : O. Vernerey & J. Lopez (LEDi) « Product Market Regulations Impact on the Labor Market »</p>	<p>14h30-15h30</p> <p>Présentation 6 : M. Bunel (LEDi) « Impact du temps domicile travail sur les choix professionnels et familiaux : une analyse longitudinale à l'aide de données individuelles »</p>	<p>14h30-15h30</p> <p>Présentation 9 : A. Drouhin (ARTEHIS) « L'Analyse du fond documentaire du prieuré de Jully-les-Nonnains grâce aux outils numériques, avantages et problématiques »</p>	<p>14h30-15h30</p> <p>Présentation 13 : G. Poujade & J-F Raze (LEDi) « Des Nuées Dynamiques aux Essaims Particulaires : Plaidoyer pour un rapprochement »</p>	<p>13h00</p> <p>Clôture de la Semaine DATA/SHS</p>
<i>Pause</i>				
<p>15h15-16h15</p> <p>Présentation 2 : M. Mbome (LEDi) & Cyrinus Elégbédé (PUDD) « Accidents du travail dans l'UE-15 et méthodes d'apprentissage automatique »</p>	<p>15h40-16h45</p> <p>Présentation 7 : R. Konate (IREDU) « Déterminants sociaux, scolaires et contextuels de la réussite académique dans le premier cycle universitaire : enfants issus de l'immigration »</p>			

**CLIQUEZ SUR LA CASE DES CONFÉRENCES
POUR VOUS Y CONNECTER**

Lundi

9h30
10h

Accueil

Ouverture

- Jean VIGREUX, *Directeur MSH Dijon*

- Jimmy LOPEZ, *Responsable scientifique PUDD*

Invité : Pascal Neige, Vice Président
de l'université de Bourgogne

10h - 12h **Conférence 1**

« Comparaisons internationales. Utiliser les données de l'Enquête sociale européenne (ESS) »

Simon Le Corgne (CDSP, SciencesPo)

La présentation proposera de découvrir l'Enquête sociale européenne à travers ses résultats. Par l'exploration en ligne des données rendue possible grâce à l'intégration de l'outil Nesstar, les participants pourront découvrir les données disponibles dans l'enquête ESS. Ils seront amenés à produire des tables de fréquences, des tableaux croisés ainsi que des analyses de corrélation et de régression. La présentation s'intéressera également aux possibilités de mise en forme des résultats, en particulier la production de graphiques. Enfin, l'export d'objets depuis Nesstar vers les outils bureautiques classiques sera évoqué.

- **Source des données utilisées :**

- Enquête sociale européenne (ESS), édition 9

- **Type d'enquêtes :** quantitatives

- **Méthodologie d'analyse :** Production de tris à plat, de tris croisés, de graphiques et d'analyses multifactorielles

« Le portail de données de Quetelet-Progedo Diffusion »

Erik Zolotoukhine (ADISP, PROGEDO)

En préambule, on présentera le dispositif français de diffusion de données, ses partenaires et le périmètre de QPD.

On fera ensuite un tour d'horizon du (des) catalogue(s) des données disponibles pour finir par un rappel des outils d'informations et de commande proposés aux utilisateurs de données.

- **Source des données utilisées :**

- Réseau Quetelet-PROGEDO diffusion

- **Type d'enquêtes :** quantitatives

SESSION 1

13h30 - 15h00 *Présentation 1*

« *Product Market Regulations Impact on the Labor Market* »

Océane Vernerey & Jimmy Lopez (LEDi).

Depuis la crise financière de 2008, les gouvernements européens ont mis en place de nombreuses réformes des réglementations anticoncurrentielles des marchés des biens et services, soit des réformes des lois et règlements dans les secteurs des services et des réseaux pouvant restreindre la concurrence et les choix des entreprises. Ces réformes sont soutenues par des organismes internationaux, comme l'OCDE et le FMI, et par une littérature abondante concluant à des effets négatifs de ces réglementations sur l'innovation et la productivité des entreprises. Cependant, leurs effets sur le marché du travail sont assez méconnus.

Ce papier étudie les effets de ces réglementations sur les marchés du travail des régions Européennes grâce aux indicateurs OCDE sur les réglementations anticoncurrentielles des marchés des biens et services et aux données individuelles de l'enquête emploi européenne fournies par Eurostat. La base ainsi constituée couvre la plupart des régions européennes sur la période 1985-2016. Notre approche en double-différence nous permet de tester si l'effet d'une régulation anticoncurrentielle sur un marché du travail régional dépend de la composition sectorielle de celui-ci. Plus exactement, nous testons si l'effet des réglementations d'un secteur sur un marché du travail local dépend : (i) de la proportion de ses travailleurs employés dans le secteur régulé ; et (ii) de la proportion de ses travailleurs employés dans des secteurs utilisant les biens intermédiaires produits par le secteur régulé.

- **Source des données utilisées :**

- Labour Force Survey, Eurostat
- Product Market Regulation indicators, OECD

- **Type d'enquêtes :** quantitatives

- **Méthodologie d'analyse :** Économétrie des données de panel, approche en double-différence

15h15 - 16h15 *Présentation 2*

« *Accidents du travail dans l'UE-15 et méthodes d'apprentissage automatique* »

Marie Germaine Mbome (LEDi) et Cyrinus Elegbede (PUDD, MSH).

Les méthodes d'apprentissage automatique ont été utilisées comme outil de prédiction dans de nombreux domaines, mais leur utilisation en santé et sécurité au travail est relativement nouvelle. C'est la raison pour laquelle, il serait intéressant d'utiliser ces méthodes dans la prévision des accidents du travail, sur les données d'enquêtes européennes sur les conditions de travail. L'objectif de ce travail est de tester les performances des techniques d'apprentissage automatique dans la modélisation et la prédiction d'accidents du travail. A cette fin, nous utiliserons trois modèles : (les forêts aléatoires) (RF), Support Vector Machine (SVM), modèle logistique. Nous observons que, la performance des modèles dépend des critères d'évaluation choisis.

- **Source des données utilisées :**

- Enquêtes Européennes sur les Conditions de Travail (EECT)

- **Type d'enquêtes :** qualitatives

- **Méthodologie d'analyse :** Intelligence artificielle et méthodes d'apprentissage automatique

SESSION 2

10h - 11h

Présentation 3

« *Les enquêtes quantitatives : de la corrélation à la causalité.*

Méthodologies utilisées pour investiguer le besoin d'acceptation sociale et l'état de pleine conscience (papiers crayons, plateforme Qualtrics, etc.) »

Lucie Finez, Raphael Laurin, Anthony Clain, Anaïs Genion, & Anne Masselin-Dubois (Psy-DREPI).

Nous présenterons des études que nous avons mises en place pour étudier la sensibilité à l'acceptation sociale et l'état de pleine conscience (*mindfulness*). Des études ont d'abord été mises en place pour tester l'existence de liens de corrélation entre les variables ciblées : les variables A et B sont-elles liées ? Des études ont ensuite été menées pour tester des liens de causalité : la variable A, a-t-elle un effet sur la variable B ? Des supports variés ont été utilisés pour mesurer et/ou manipuler nos variables : questionnaires papiers, questionnaires en ligne (sur la plateforme Qualtrics), mises en situations variées. Nous présenterons la façon dont nos outils ont été sélectionnés et adaptés selon l'objectif ciblé (ex. mesurer une réponse en contexte). Chaque étude présentée est indépendante et a été menée sur un échantillon de plus de 100 participants. Au travers des différentes études présentées, nous verrons qu'un même objet d'étude peut être investigué selon des angles variés.

- **Source des données utilisées :**

- Données recueillies via des questionnaires papier-crayons, questionnaires en ligne ou mises en situation

- **Type d'enquêtes :** quantitatives

- **Méthodologie d'analyse :** Régressions multiples.

11h - 12h

Présentation 4

« *Retour critique sur une démarche de débat citoyen en ligne* »

Hervé Marchal (LIR3S), Louis Olagnon (co-fondateur de Cartodébat)

La présentation reviendra sur une démarche de débat citoyen organisé à partir d'une plateforme numérique pensée collectivement durant la crise sanitaire liée au Covid-19. Il s'agit d'opérer un retour critique au sens où nous interrogerons les tenants et les aboutissants d'une telle démarche, entendons les limites et les points forts, les doutes et les certitudes.

Notre présentation s'articulera autour de trois questions :

Qu'est-ce qu'une « ingénierie du sensible » ?

Qu'est-ce que veut dire « représentatif » ?

Qu'est-ce que veut dire « scientifique » ?

Parallèlement, nous reviendrons sur la nécessité que nous avons ressentie de tenir un journal de bord analytique qui, précisément, fait suite à ces questionnements et entend assumer des questions trop souvent mises en suspens : quels sont la philosophie et les implicites qui président au traitement des données ? Quelles attentes politiques accompagnent le projet tout en étant à son origine ?

De même, des pièges, à commencer par celui de la fétichisation des données, ont été discutés et prévenus dans la mesure du possible.

Ce retour critique doit donc être compris, *in fine*, comme une démarche réflexive qui interroge sans détour les conditions réelles de production de données numériques et les conditions mêmes de leur intérêt scientifique, politique et éthique.

- **Source des données utilisées :**

- Plateforme numérique « métamorphose » mise en place par la société Cartodébat durant la crise sanitaire liée au Covid-19 en lien avec de nombreux Conseils de développement durable au niveau national et avec le soutien de diverses institutions.

- **Type d'enquêtes :** qualitatives

- **Méthodologie d'analyse :** Travail collectif de définition des questionnements qui président à la mise en place de la plateforme ici concernée ainsi que des grandes thématiques proposées.

Construction collective du cadre d'analyse.

Travail collectif, manuel (humain) du « tagage » des données numériques et mise en évidence collective des tensions, des points saillants et des métamorphoses sociétales. .

Mise en place d'une « ingénierie du sensible » pour diversifier les sources et les données.

13h30 - 14h30 Présentation 5

« *Acquisition de données géohistoriques : entre automatisation et intervention humaine.* »

Arnaud Millereux (MSH) et Thomas Thévenin (ThéMA).

L'acquisition de données géohistoriques est régulièrement le fait d'un ou plusieurs chercheurs motivés pour répondre à une problématique précise. Ce travail est souvent long et aussi coûteux que fastidieux. Ainsi, les promesses de la numérisation automatique ont fait rêver plus d'un chercheur. Or, l'approche totalement informatisée implique un temps non négligeable pour détecter les erreurs, trouver la source et effectuer les modifications. Ainsi, le choix entre instrumentation complète et saisie manuelle se pose systématiquement au début de chaque collecte de données. Cette intervention propose de faire un retour d'expérience sur l'acquisition de données géohistoriques sur les voies de communications françaises de 1750 jusqu'à nos jours. Un focus sur la vectorisation des données fluviales sera tout d'abord effectué avant de proposer la construction d'un système d'information géohistorique d'un réseau de transport multimodal combinant les voies fluviales, les chemins de fer et les routes.

- **Source des données utilisées :**

- Données IGN (BD TOPO, BD Carthage)
- Cartes historiques : Cassini, Etat-Major, Michelin
- Site web, ouvrage et archives

- **Type d'enquêtes :** quantitatives

- **Méthodologie d'analyse :**

- WebScraping
- Acquisition collaborative : Mapathon, Arpenteur topographe
- Système d'information géographique
- Graphe et réseau

14h30 - 15h30 Présentation 6

« *Impact du temps domicile travail sur les choix professionnels et familiaux : une analyse longitudinale à l'aide de données individuelles* »

Mathieu Bunel (LEDi), Dominique Meurs (Université Nanterre- Economix-Ined) et Elisabeth Tovar (Université Nanterre-Economix) .

L'objectif de ce travail est de mobiliser des données exhaustives portant sur l'ensemble des salariés (relevant des entreprises) issus des entreprises du secteur de l'énergie afin de caractériser les interactions entre temps de transport-domicile-travail et carrière professionnelle. Il s'agira de préciser l'impact du facteur « temps de transport » sur le temps de travail, les évolutions de carrière et les choix familiaux (se marier et avoir un ou plusieurs enfants). La causalité inverse est également considérée. Une analyse particulière permettra d'identifier les différences hommes-femmes.

Dans les travaux standards sur l'offre de travail, le temps hors travail (incluant le temps de transport) est appréhendé comme du temps de loisir et les préférences individuelles pour ces deux temps « libre » sont supposées identiques. Or il paraît peu réaliste de considérer que les choix de temps de déplacements domicile-travail soit de même nature que les choix de temps de loisir. Il est donc utile de les différencier.

L'originalité et la qualité des données disponibles permettent de relâcher cette hypothèse. L'échantillon constitué offre la possibilité de mesurer précisément au niveau individuel et sur une période longue de près de 15 ans : d'une part le temps de transport domicile-travail et le temps de travail dont les heures supplémentaires et les absences pour raisons personnelles, d'autre part de connaître l'évolution précise des rémunérations (salaire de base, heures supplémentaires et primes) et des promotions et enfin de suivre les différents changements familiaux (vie en couple, avoir des enfants).

Ces informations permettent d'analyser finement à court et à moyen terme les interactions entre le temps de transport domicile-travail et ces autres dimensions tout en contrôlant la partie fixe de l'hétérogénéité inobservée associée aux individus et aux établissements/divisions de l'entreprise. La taille de notre échantillon offre la possibilité également de tester l'existence éventuelle d'hétérogénéité dans ces relations selon les caractéristiques de l'entreprise, celles de l'individu (genre, l'âge et la CSP) et la zone géographique considérée.

- **Source des données utilisées :**

Données exhaustives sur les salariés issus de groupe/ d'entreprises/ d'établissements/ de divisions du secteur énergétique sur la période 2005 à 2019. Panel de plus de 100 000 salariés sur 15 ans.

Données ODOMATRIX permettant de calculer des distances routières et des temps de trajet de centroïde à centroïde pour l'ensemble des communes de France (matrice 35 000 par 35 000)

- **Type d'enquêtes :** quantitatives

- **Méthodologie d'analyse :** - Analyse spatiale

- Econométrie des données de panel

15h40 - 16h40 **Présentation 7**

« Déterminants sociaux, scolaires et contextuels de la réussite académique dans le premier cycle universitaire : enfants issus de l'immigration »

Ramadan Konaté (IREDU)

L'étude vise à analyser les conditions de vie et d'études des étudiants dans le premier cycle universitaire et les variables qui déterminent la réussite académique. Il en ressort des différentes formes de réussite en fonction de l'origine migratoire qui est une dimension de la variable origine sociale. Les étudiants ayant une ascendance immigrée réussissent moins bien que leurs homologues autochtones. L'étude a également permis de mettre en exergue l'importance des caractéristiques sociodémographiques et scolaires dans la réussite académique. La catégorie socioprofessionnelle des parents, le passé scolaire et le degré de satisfaction que l'étudiant déclare avoir de sa formation sont des facteurs déterminants.

Mise en perspective : étude comparative des systèmes d'enseignement supérieur franco-allemands

• **Source des données utilisées :**

- Base de données Conditions de Vie des Etudiants (CVE) 2016 de l'Observatoire nationale de la Vie Etudiante (OVE)

• **Type d'enquêtes :** quantitatives

Mercredi

09h00-12h00 Conférence 2 : textométrie et application

Programme

9h00 - 10h30 Serge Heiden (ENS Lyon & IHRIM)

Introduction historique et méthodologique à la textométrie et au logiciel TXM (3/4h)

Démonstration des logiciels «TXM pour poste» et «portail TXM web» (1/2h)

10h45 - 11h45 Virginie Lethier (MCF en Sciences du langage. Université de Franche-Comté)

- Application de la textométrie : Textométrie et analyse du discours institutionnel (3/4h)

- Présentation des structures et des personnes ressources en ADT des MSH Dijon et Besançon (1/4h)

11h45 - 12h00 : Questions et Conclusion (1/4h)

Introduction historique et méthodologique à la textométrie et au logiciel TXM » (Serge Heiden)

Développée dans la continuité de la lexicométrie et de l'analyse de données « à la française », la textométrie est une méthode d'analyse de corpus textuels s'appuyant sur des observables textuels construits à partir de l'enrichissement linguistique automatique de textes par les outils de TAL (catégories grammaticales et lemmes de mots...) et sur la structuration des textes encodée en XML selon les recommandations du consortium TEI (www.tei-c.org).

La méthode est implémentée dans le logiciel TXM - open-source et gratuit - depuis 2009 par les partenaires du projet ANR Textométrie ; Il est depuis co-développé par l'UMR IHRIM de l'ENS de Lyon et par l'EA ELLIAD de l'UFC.

Il permet de construire des corpus textuels pour les explorer d'une part avec des outils quantitatifs d'analyse « en masse » comme :

- l'analyse factorielle des correspondances
- La classification hiérarchique ascendante
- Le calcul de mots spécifiques
- Ou encore de cooccurrents statistiques

et d'autre part avec des outils qualitatifs complémentaires, pour affiner l'analyse, comme :

- Les concordanciers
- Les listes de mots avec leur fréquence
- Les visualisations d'apparitions longitudinales

- Ou encore la lecture hypertextuelle d'éditions de textes.

Il peut traiter quatre paradigmes différents de corpus textuels :

- Les textes écrits (dans divers formats source TXT, DOCX, ODT, XML, TEI...)
- Les transcriptions d'enregistrements synchronisées (aux formats TXT tabulé, DOC tabulé, ODT tabulé ou XML Transcriber, et accompagnées de fichiers vidéo ou audio source)
- Les corpus alignés multilingues ou multiversions (au format XML TMX)
- Les corpus en tableau, comme les réponses à des questionnaires d'enquête ou des échanges de tweets (aux formats : Excel, ODS, CSV).

Dans TXM, un texte peut être structuré à n'importe quel niveau (livre, chapitre, section, paragraphe...), délimité (discours direct, prise de parole...), équipé d'apparat critique (notes de bas de page) et encodé jusqu'au niveau des mots si nécessaire. Les textes peuvent être regroupés en sous-corpus ou en partitions à partir de leurs métadonnées (auteur, date, genre...). Tous leurs mots sont annotés automatiquement à l'import par une catégorie grammaticale et un lemme qui seront mobilisables dans les analyses au même titre que la forme graphique des mots. Langues disponibles notamment : français contemporain, français parlé, français ancien, anglais, allemand, italien, espagnol, portugais, latin, grec, grec ancien, russe, chinois. Le logiciel est régulièrement utilisé dans de nombreux domaines scientifiques : Linguistique, Langue, Littérature, Civilisation, Géographie, Documentation, Sciences de l'information et de la communication, Didactique, Sciences de l'éducation, Histoire, Sociologie, Informatique, Mathématiques, Statistiques, Philosophie, Psychologie, Sciences politiques, Urbanisme, Transports, Économie, Administration, Sciences de gestion, Sciences biologiques, Anthropologie, Sciences de l'art.

Il est disponible pour poste de travail Windows, Mac ou Linux – téléchargement ici : <http://www.textometrie.org>

La plateforme logicielle TXM permet également l'accès aux corpus en ligne avec sa version portail TXM pour serveur web – exemples de portails publics ici : https://groupes.renater.fr/wiki/txm-users/public/references_portails.

Pour en savoir plus sur TXM et sa communauté d'utilisateurs, manuel, tutoriaux et documents de référence, n'hésitez pas à visiter le site suivant : <http://www.textometrie.org>.

- **Source des données utilisées :**

- Corpus 1- Discours (« speeches ») en langue anglaise de la Banque Centrale Européenne (1998-2020). Données accessibles ici : <https://www.ecb.europa.eu/press/key/html/index.en.html>
- Corpus 2- Plans de développement urbain (PDU) - conçu par Thomas Buhler. Pour une description détaillée de l'échantillon, se reporter à : <https://www.cairn.info/revue-flux-2018-3-page-93.htm>.

- **Type d'enquêtes :** qualitatives - quantitatives

- **Méthodologie d'analyse :** Analyse du discours ; textométrie.

« Textométrie et analyse du discours institutionnel » (Virginie Lethier) :

En complément des méthodes classiques qu'elles mobilisent pour interroger leurs objets traditionnels, de nombreuses disciplines des sciences sociales s'ouvrent depuis la fin des années 1990 aux pratiques numériques d'études de discours. Cette dynamique, largement encouragée par l'ère des « Humanités Numériques » (Dacos, Mounier, 2014), favorise notamment l'essor de nombreuses analyses de discours institutionnels (Krieg-Planque, 2012). Pour les développer, les chercheur.e.s de sciences sociales sont invité.e.s à se familiariser avec les méthodes de l'analyse de données textuelles, au sein de laquelle cohabitent différentes traditions (Jenny, 1997).

En vue de permettre une meilleure compréhension des singularités de la textométrie et des possibilités offertes par cette pratique d'analyse de données textuelles faisant suite à la lexicométrie (Tournier, 1980 ; Lebart et Salem, 1994), notre contribution proposera un retour d'expériences sur deux cas d'étude de vastes corpus de discours institutionnels menées grâce à l'outil de référence TXM (Heiden et al., 2010). Elaborées dans le cadre de coopérations interdisciplinaires, les analyses que nous proposerons trouvent leur point d'ancrage dans des champs disciplinaires aux prises avec la dimension langagière d'institutions qui agissent par le discours (Krieg-Planque, 2012) : d'une part, l'analyse des politiques monétaires ; d'autre part, l'analyse des politiques urbaines.

Notre communication s'organisera en deux principaux temps.

Dans un premier mouvement, nous insisterons sur une première singularité de la démarche d'analyse textométrique : celle-ci permet en effet la construction de parcours de lecture nourris par l'articulation de prises de vue globales (issues de différents calculs) et de prises de vue locales sur la matérialité textuelle (Pincemin, 2010). Afin d'illustrer concrètement les apports de cette démarche mettant à mal l'opposition traditionnellement admise entre analyse quantitative et analyse qualitative (Duchastel & Laberge, 2014; Mayaffre, 2014), nous montrerons comment la textométrie permet de mettre au jour des contrastes difficilement décelables par une lecture humaine dans un corpus de documents de planification urbaine, réputés être aussi neutres que

lisses et « techniques » (Offner, 2006) et des pistes de recherche ouvertes par de telles analyses différentielles et respectueuses de la matérialité (Buhler et al., 2018 ; Buhler & Lethier, 2019).

Dans un deuxième mouvement, nous mettrons la focale sur la façon dont la textométrie renouvelle le programme de la lexicométrie en étendant ses unités de segmentation et d'indexation. Afin d'illustrer les potentialités de l'exploration des dimensions plurielles de la textualité, nous rendrons compte de classes d'analyse menées sur les discours (« speeches ») délivrés par les différents représentants de la Banque Centrale Européenne (BCE) lors d'événements publics entre 1998 et 2020. Alors que des analyses thématiques pointent une permanence du stock lexical de la BCE en dépit des crises que traverse l'Europe à partir de 2007 (Lebaron, 2015), nous montrerons comment l'analyse de la consistance morphosyntaxique de ces discours livre, pour sa part, un faisceau d'indices attestant d'un tournant effectif des stratégies communicationnelles de la BCE dès 2009 (Lethier, 2019).

Bibliographie

Buhler T., Lethier V. (2019) « *Analysing urban policy discourses using textometry: An application to French urban transport plans (2000-2015)* ». Urban Studies, SAGE Publications, (10.1177/0042098019873824).

Buhler T., Bendinelli M., Lethier V. et al. (2018) « *La textométrie pour révéler évolutions et oppositions dans les discours des politiques urbaines. Le cas des PDU français (2000-2015)* ». Flux, 2018/3 (N° 113), p. 93-110. DOI : 10.3917/flux1.113.0093. URL : <https://www.cairn.info/revue-flux-2018-3-page-93.htm>

Dacos M., Mounier P. (2014). Humanités numériques. « *État des lieux et positionnement de la recherche française dans le contexte international* ». Paris : Institut Français, <http://www.institutfrancais.com/sites/default/files/if_humanites-numeriques.pdf>.

Duchastel J., Laberge D. (2014) « *Au-delà de l'opposition quantitatif/qualitatif. Convergence des opérations de la recherche en analyse du discours* ». Corela [En ligne], HS-15 | 2014. URL : <http://journals.openedition.org/corela/3524> ; DOI : <https://doi.org/10.4000/corela.3524>

Guilbert T. (2014) « *Introduction : articuler les approches qualitatives et quantitatives dans l'analyse de discours* ». Corela [En ligne], HS-15 | 2014. URL : <http://journals.openedition.org/corela/3545> ; DOI : 10.4000/corela.3545

Heiden S., Magué J.-P., Pincemin B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », in I. C. Bolasco S. (éd.), Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010, 2, Rome, Edizioni Universitarie di Lettere Economia Diritto, pp.1021-1032.

13h30 - 14h30 *Présentation 8***« Bulletins météo télévisés et vulgarisation scientifique – Analyse critique des données (non-) disponibles. »**

Vincent Liegeois (TIL).

Les bulletins météo sont, du point de vue linguistique et discursif, des textes très intéressants car ils transmettent une explication scientifique relativement compliquée à la population et représentent donc une forme de vulgarisation scientifique ou de communication scientifique externe.

Plusieurs questions apparaissent alors : Quelles différences apparaissent entre les bulletins météo destinés aux scientifiques et ceux destinés au grand public ? D'où proviennent les données scientifiques – ou les structures linguistiques scientifiques – remplacées par des modes d'expression plus expressifs ou plus subjectifs ?... Le projet de recherche « Analyse textuelle et modélisation syntactico-sémantique de bulletins météorologiques à partir de schémas lexico-grammaticaux » s'intéresse à ces questions, avec une attention particulière pour l'évolution linguistique de ces textes pendant la seconde moitié du siècle dernier.

Néanmoins, il est très difficile de collecter ces bulletins télévisés. Pendant le siècle dernier, beaucoup de bulletins météo télévisés ont été perdus parce que les archives n'étaient pas encore préparées au recollement des données vidéographiques et, ensuite, parce que la conservation de ces données ne semblait pas si importante que la conservation du journal télévisé ou de données et graphiques météorologiques.

Par conséquent, les archives ne renferment souvent qu'une collection fragmentaire de ce type de textes et parfois leur qualité est aussi discutable. Un bon exemple se trouve à l'Inathèque, où l'on peut trouver l'enregistrement d'un présentateur se faisant canarder par de nombreux produits alimentaires (Inathèque, URL : « <https://www.ina.fr/video/CPC03008653/meteo-video.html> ») Il n'y a aucun doute que des bulletins météorologiques de ce type posent des problèmes pour l'analyse de l'évolution linguistique des textes : ils présentent des structures non-prototypiques et ils diminuent la qualité du corpus. D'une part, nous aborderons cette question avec une certaine « méthodologie de recollement », qui concerne la recherche d'archives et les métadonnées disponibles pour tous les textes. D'autre part, nous approcherons ce problème avec une « méthodologie linguistique », qui observe l'analyse linguistique finale et l'importance de ces métadonnées et la variation textuelle.

• Source des données utilisées :

- Archives de télévision:
- français : Inathèque (France, <http://inatheque.ina.fr/>), RTS archives (Suisse, <https://www.rts.ch/>), RTBF archives (Belgique, <https://www.rtbf.be/>)
- néerlandais : VRT archives (Belgique, <https://www.vrt.be/nl/>), Nederlands Instituut voor Beeld en Geluid (Pays-Bas, <https://www.beeldengeluid.nl/>)
- italien: RAI archives (Italie, <http://www.rai.it/>), RSI archives (Suisse, <https://www.rsi.ch/>)
- allemand : NDR archives (Allemagne, <https://www.ndr.de/index.html>), ZDF archives (Allemagne, <https://www.zdf.de/>), ORF archives (Autriche, <https://orf.at/>)
- Autres archives
- français: Archives Nationales (France, <http://www.archives-nationales.culture.gouv.fr/>), Bibliothèque nationale de France (France, <https://www.bnf.fr/fr>)
- néerlandais: Archives de l'État (Belgique, <https://search.arch.be/fr/>), Nationaal Archief (Pays-Bas, <https://www.nationaalarchief.nl/>)

• Type d'enquêtes : qualitatives et quantitatives**• Méthodologie d'analyse :** - Méthodologie pour la collection des textes : critique historique et sociolinguistique historique

- Méthodologie pour l'analyse des textes : linguistique de corpus, sémantique des prototypes, sémantiques des cadres, linguistique textuelle

14h30 - 15h30 **Présentation 9**

« L'Analyse du fond documentaire du prieuré de Jully-les-Nonnains grâce aux outils numériques, avantages et problématiques. »

Aurore Drouhin (ARTEHIS).

Cette présentation s'invite à détailler l'ensemble des outils numériques appliqués aux recherches réalisées dans le cadre d'une thèse de doctorat en histoire médiévale concernant le prieuré de Jully-les-Nonnains.

Plusieurs outils numériques sont employés. Tout d'abord l'automatisation robotisée du classement des données est réalisé grâce à UiPath. De plus, l'exploitation et l'analyse des données collectées ont été facilitées par l'emploi de Gephi et de R. Ces logiciels permettent d'une part de modéliser les relations entre les personnes et les lieux, et d'autre part, de lemmatiser ces données pour effectuer des recherches spécifiques sur le vocabulaire notamment. L'utilisation de ces logiciels pour un sujet d'histoire devait répondre à des impératifs de praticité et d'efficacité tant pour la collecte des données que pour leur traitement. Il s'agira donc de définir en quoi l'utilisation des outils numériques se révèle pertinente pour cette étude, et quelles sont les limites et les problèmes liés à cet usage.

• **Source des données utilisées :**

- Fond documentaire du prieuré de Jully-les-Nonnains, Archives départementales de la Côte d'Or et Archives départementales de l'Yonne

• **Type d'enquêtes :** quantitatives

• **Méthodologie d'analyse :** La méthodologie d'analyse utilisée dans le cadre de cette thèse couple plusieurs méthodes. Elle associe la microhistoire et l'étude sur le temps long. Ce à quoi s'ajoute, l'analyse des réseaux de personnes et de lieux, ainsi que le processus de l'enquête quantitative qui permet de mettre en lumière la récurrence dans la documentation des noms de personnes, de lieux et de certains mots grâce à la lemmatisation.

SESSION 5

9h30 - 10h30 **Présentation 10**

« Covid-19 : ce que nous apprennent les statistiques hospitalières »

Dominique Andolfatto (CREDESPO &

Dominique Labbé (Pacte-Cnrs, UGA)

Bilan de l'épidémie de covid-19 en France à l'aide des statistiques hospitalières départementales mises en ligne par Santé publique France.

Ces statistiques montrent une grande hétérogénéité des situations selon les départements, spécialement en ce qui concerne la mortalité. Comment interpréter cette situation ? Une analyse à chaud a montré que ces différences ne sont pas le fait du hasard et ne sont pas liées à l'intensité de l'épidémie (mesurée par le nombre des hospitalisés). Elles pourraient tenir à l'organisation et à la nature des soins apportés aux malades. Avec le recul, il s'agira donc de vérifier cette première analyse.

Plus au fond, outre une analyse classique de données sociales – celles de Santé publique France ayant été finalement assez peu exploitées, en particulier de façon longitudinale et systématique – et leur remise en perspective, historique et comparée, il s'agira d'interroger les choix de politiques sanitaires face à l'épidémie de Covid-19.

• **Source des données utilisées :**

- Données Santé publique France
- Articles publiés dans la Revue politique et parlementaire (2020) : *D. Andolfatto, D. Labbé, « Covid-19 et surmortalité », Revue politique et parlementaire, 10 avril 2020.* »
En ligne : <https://www.revuepolitique.fr/covid-19-et-surmortalite-en-france/>
- « *Destin du covid-19* », *RPP, 22 avril 2020.* »
En ligne : <https://www.revuepolitique.fr/destin-du-covid-19/>
- « *Covid-19 : ce que nous apprennent les statistiques hospitalières* », *RPP, 20 mai 2020.* »
En ligne : <https://www.revuepolitique.fr/covid-19-et-surmortalite-en-france/>
- « *Covid-19 : premier bilan de l'épidémie* », *RPP, 5 juin 2020.* »
En ligne : <https://www.revuepolitique.fr/covid-19-premier-bilan-de-lepidemie/>
- « *Covid-19 : une défaite française* », *RPP, 22 juin 2020.* »
En ligne : <https://www.revuepolitique.fr/covid-19-une-defaite-francaise/>

Une version abrégée de ce dernier texte a été publiée par les PUG, dans la collection « Le virus de la recherche », sous le titre : Les malades privés de leur médecin de famille.

• **Type d'enquêtes :** qualitative

• **Méthodologie d'analyse :** Analyse quantitative (données sanitaires) et analyse de politique publique.

10h30 - 11h30 **Présentation 11**

« Les méthodes biproportionnelles, un outil de reconstitution ou de comparaison de tableaux de contingence »

Louis de Mesnard (CREGO)

Un tableau de nombres dont on connaît le total de ses colonnes et lignes (ou « marges ») est appelé « tableau de contingence ». Un tel tableau se rencontre quand on étudie les échanges entre pays ou régions, les flux migratoires ou financiers, etc. Le premier problème qu'on étudie est la reconstitution du tableau quand on ne connaît que ses marges, mais pas les données du tableau elles-mêmes, sachant que celles-ci sont néanmoins proches de celles d'un autre tableau. C'est le cas par exemple quand on veut prévoir l'origine des échanges entre pays en connaissant seulement le total de ce qu'ils vont exporter et importer. Le deuxième problème étudié est l'évaluation de ce qui a changé entre deux tableaux, par exemple l'évolution d'un tableau des migrations dans le temps. L'outil pour traiter ce genre de problèmes est non-stochastique : il s'agit des méthodes biproportionnelles. La communication présentera les deux problèmes, puis comment on peut les traiter par projection biproportionnelle, et enfin on donnera (d'une manière simple) les propriétés intéressantes de cet outil et aussi ses particularités calculatoires. Je me baserai sur les très nombreuses publications sur le sujet.

• **Source des données utilisées :**

- Pas de données : présentation d'une méthode originale

• **Type d'enquêtes :** Pas d'enquêtes : présentation d'une méthode originale

• **Méthodologie d'analyse :** Approche théorique, méthode non stochastique.

SESSION 6

13h30 - 14h30 *Présentation 12*

« *L'apport et la mise en œuvre de la méthode des choix discrets pour analyser les préférences individuelles* »

Christine Peyron & Aurore Pélissier (LEDi).

En prenant appui sur deux recherches ayant utilisé la méthode des choix discrets, nous présenterons l'objectif, la logique générale et l'apport de cette méthode. Nous préciserons les différentes étapes nécessaires pour mettre en place de telles enquêtes ainsi que les méthodes économétriques possibles pour exploiter les données collectées et estimer les préférences des répondants.

• **Source des données utilisées :**

- La présentation s'appuiera sur deux enquêtes *ad hoc*, la première porte sur un échantillon de 513 parents d'enfants atteints de la maladies rares et d'anomalies du développement, la seconde sur un échantillon représentatif de la population française (2501 répondants)

• **Type d'enquêtes :** Quantitative

• **Méthodologie d'analyse :** Méthode des choix discrets, économétrie des variables qualitatives.

14h30 - 15h30 *Présentation 13*

« *Des Nuées Dynamiques aux Essaims Particulaires : Plaidoyer pour un rapprochement* »

Guillaume Pujade et Jean-François Raze (LEDi.)

La 1^{ère} Partie de notre présentation a pour objectif de montrer l'existence d'une « analogie technique » entre d'une part le principe de fonctionnement de la méthode des Nuées Dynamiques en Classification Automatique (Diday, 1973) et d'autre part certaines méthodes de résolution "Heuristiques" utilisées en Localisation Optimale (Hansen, Peeters, Thisse, 1981). La 2^{ème} Partie de notre présentation est consacrée à l'exploitation technique de cette correspondance entre les deux champs disciplinaires, que sont la Statistique d'une part et l'Optimisation d'autre part. Plus précisément, nous montrons comment les avancées récentes dans le champ de la Localisation Optimale tant du point de vue de la formalisation que de celui de la résolution (Laporte, Nickel, Saldanha, 2017), peuvent être utilisées à des fins classificatrices. Enfin, la 3^{ème} Partie est une mise en application de notre démarche. Développé en langage Python, nous proposons un nouvel algorithme, reposant sur une formalisation de type Districting-Location Problem (Kalcsics, 2015) et une résolution de type Particle Swarm Optimization (Clerc, 2006). Testé sur un grand nombre d'échantillons, les résultats font l'objet d'une présentation et d'une comparaison avec ceux obtenus à partir de logiciels d'Analyse des Données (SPSS-IBM, R).

• **Source des données utilisées :**

- Échantillons générés aléatoirement pour simulations numériques

• **Type d'enquêtes :** Quantitative

• **Méthodologie d'analyse :** Aborder les "traditionnelles" méthodes de Classification Automatique à partir de la formalisation de problèmes et de techniques de résolution propres à l'Optimisation Combinatoire

Vendredi

9h - 11h Conférence 3

« Comment FAIR.E métropole dans une ville Smart »

Denis Hameau, Nadège Martiny, Yves Richard, Anne Jégou, Julien Pergaud, Christophe Cruz, Benjamin Pohl, Thomas Thevenin.

Programme :

- 1/ **Ouverture de la matinée Denis Hameau** : de la transition numérique à la transition socio-écologique Ou De l'infrastructure Smart aux actions citoyennes
- 2/ **Intervention Yves Richard - Nadège Martiny - Anne Jégou** : De Mustard à Qaémléo : la construction d'un observatoire environnemental participatif
- 3/ **Intervention Julien Pergaud - Christophe Cruz** : Pour une approche FAIR de l'observation environnementale
- 4/ **Table Ronde** : Fair.e métropole avec les chercheurs pour les citoyens

Depuis 2017, Dijon Métropole développe un projet de ville intelligente pour améliorer la gestion urbaine au quotidien. Ce nouveau récit métropolitain s'inscrit également dans une approche originale pour garantir un accès sécurisé aux données produites par la ville intelligente. Parallèlement à cette action, les chercheurs du campus dijonnais se sont engagés, dans le cadre du programme POPSU, à constituer un observatoire intelligent pour la qualité environnementale selon les principes FAIR : Facile à trouver, Accessible, Interopérable, Réutilisable. Se pose alors la question de comment faire converger ces infrastructures smart pour FAIR.E métropole ?

• Source des données utilisées :

- Mesures de température et qualité de l'air géolocalisées
- Base de données météo France
- Base de données IGN (BD TOPO)
- Image satellites

• Type d'enquêtes : quantitatives

• Méthodologie d'analyse :

- Modèle de climat
- Modèle d'estimation de la qualité de l'air
- Modèle de croissance urbaine
- Typologie des ambiances thermiques (LCZ)

• Mots clefs :

- Démarche FAIR
- Gouvernance de la donnée
- Open data
- Stratégie métropolitaine sur le numérique et recherche académique

SESSION 7

11h - 12h « Un LiDAR pour le Parc national de forêts : Présentation de la démarche et focus sur la problématique archéologique. »

B. Quost, L. Granjon, L. Pillot, L. Saligny, E. Fovet, E. Chevigny, J. Vidal, D. Goguey, Y. Pautrat, J. Benard, M. Delcamp (GéoBFC).

En 2011, le GIP de préfiguration du futur Parc national (aujourd'hui Parc national de forêts) se penche sur la mise en place d'une acquisition LiDAR sur la partie Côte-d'Or de son territoire, représentant une surface d'environ 400 km². A cet effet, le GIP fait appelle à la Plateforme GeoBFC de la Maison des Sciences de l'Homme de Dijon pour l'accompagner dans le suivi et l'expertise de cette acquisition.

Cette présentation propose de s'intéresser dans un premier temps aux objectifs initiaux du LiDAR pour le Parc national, d'établir un tour d'horizon des résultats obtenus dans différentes thématiques (archéologiques, forestières, etc.), et enfin de présenter l'apport de cette technologie pour un Parc national.

Dans un second temps, un focus sera fait sur les travaux engagés dans le cadre du partenariat entre le GIP et la MSH de Dijon. Nous présenterons la démarche mise en place sur le Châtillonnais, dans le cadre d'une problématique archéologique : du suivi de l'acquisition LiDAR, en passant par la mise en place d'un protocole de détection de structures, jusqu'à la restitution graphique des résultats et leurs valorisations.

• Source des données utilisées :

- LiDAR, Modèle Numérique de Terrain (MNT)

• Type d'enquêtes : quantitatives

• **Méthodologie d'analyse** : Protocole de vectorisation, Télédétection, Vérification, Calcul de densité, Cartographie.

12h15 - 13h « Twitter et la linguistique située : réflexions méthodologiques à partir de l'exemple de tweets sur la métropole de Dijon »

Matthieu Bach (TIL ; MSH Dijon), Arnaud Da Costa (MSH Dijon).

Cette communication s'inscrit dans le cadre du projet de recherche interdisciplinaire POPSU visant à entre autres à analyser l'identité métropolitaine et identifier les intérêts des citoyens. Pour alimenter la discussion proposée, nous exploiterons un corpus original de 300 000 tweets collectés en temps réel depuis juillet 2019 dans le cadre du projet interdisciplinaire POPSU Dijon, hébergé à la MSH Dijon et dont le but est d'explorer les interactions des utilisateurs de Twitter lorsqu'ils parlent de Dijon et de la métropole.

Nous commencerons par discuter les aspects liés à la collecte de données Twitter :

- juridiques : protections des données personnelles, Open Data ;
- relatifs à la constitution du jeu de données : où placer le curseur entre l'exhaustivité des données recueillies et le risque de recueillir des données trop éloignées de l'objet d'étude (au niveau temporel, géographique ou thématique) ;
- techniques : liés à l'API gratuite de Twitter (différentes méthodes de collectes et leur exhaustivité) ;
- relatifs aux jeux de données dont les traitements ne nécessitent pas d'infrastructure Big Data.

Les questionnements informatiques sont consubstantiels aux interrogations linguistiques qui ont émaillé l'ensemble de ce projet à l'interface entre informatique, linguistique et géographie. Les données collectées nous ont amenés à évaluer leur intégrité et leur mise en corpus ; cette phase de réflexion est fondamentale dans le cadre de l'analyse de discours, qui était le but initial du projet de recherche. Or, si l'intégrité du corpus n'est pas garantie, une analyse de discours ne peut être envisagée en tant que telle (cf. la définition séminale de Busse/Teubert 1994). Cette mise en défaut n'est toutefois pas rédhibitoire dans la mesure où les données Twitter peuvent alimenter d'autres réflexions linguistiques et amener à considérer d'autres nœuds entre matérialisations linguistiques et structures épistémiques.

En tant que source de données hypersynchroniques, Twitter permet une analyse de phénomènes linguistiques contemporains et authentiques (cf. la réflexion menée dans Bach 2020) ; toutefois, Twitter n'est pas un instrument magique pour la linguistique située, et un certain nombre de desiderata sont à formuler en ce sens pour une démarche scientifique rigoureuse (c'est-à-dire

falsifiable et en accord avec les innovations issues des autres champs de la cognition). Ce projet a en effet mis en évidence deux types de limitations : techniques et méthodologiques. Une grande partie de ces limitations ont été contournées, et il nous est possible d'ajuster la méthodologie pour les futurs projets. Notre expérience montre que, sur le long terme, une collaboration étroite et régulière avec une implication continue et forte de chacun des acteurs est nécessaire.

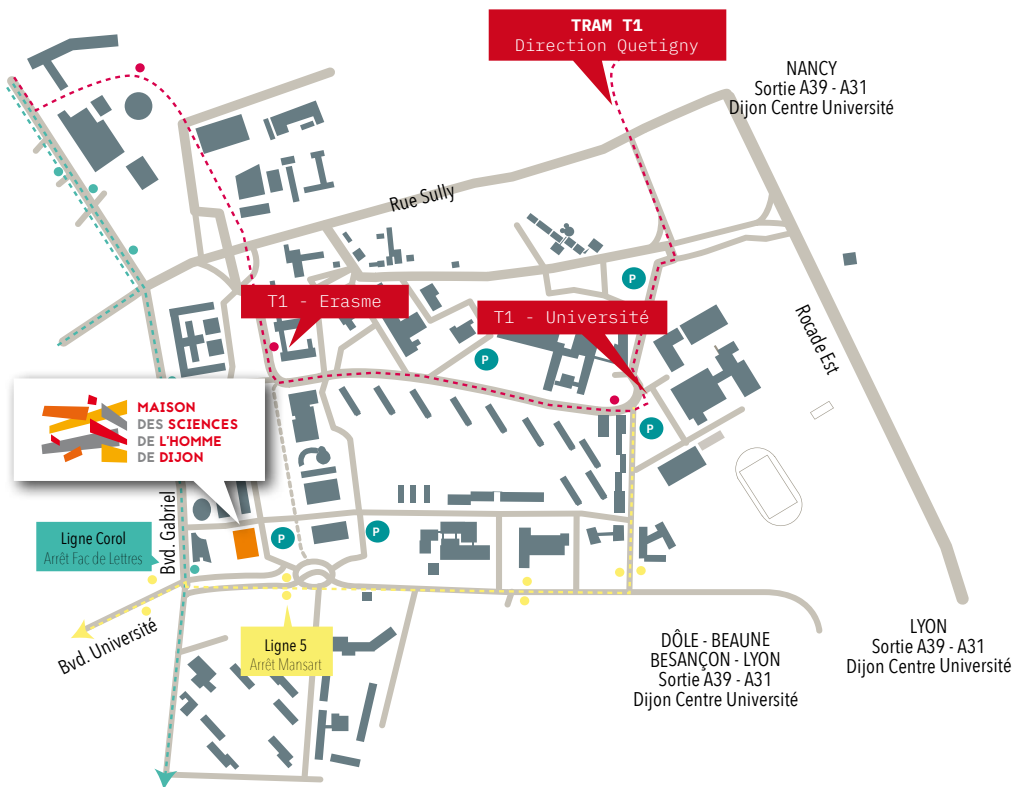
- **Source des données utilisées :**

- Corpus POPSU – Twitter

- **Type d'enquêtes :** qualitatives

- **Méthodologie d'analyse :** Linguistique textuelle.

La semaine DATA SHS est un événement composé de conférences et présentations quotidiennes. Après les succès de l'an dernier, l'édition de cette année, malgré la crise sanitaire, fait intervenir des chercheurs en SHS du grand campus dijonnais ainsi que des partenaires professionnels (la métropole de Dijon, la plateforme cartoDEBAT) autour de travaux de recherche mobilisant des données et des méthodes quantitatives.



Les activités de la PUDD de la MSH de Dijon ainsi que les événements organisés le sont dans le cadre de la TGIR PROGEDO (Très Grande Infrastructure de Recherche PROduction et GEstion de DONnées). Elle apporte le financement nécessaire au fonctionnement de la PUDD ainsi qu'au montage d'événements spécifiques.