PROPOSITION SUJETS DE THÈSE
CONTRATS DOCTORAUX - 2020-2023

X **Contrat doctoral fléché FR Agorantic**

Directeur de thèse : MARREL Guillaume LBNC
Mail et Téléphone : guillaume.marrel@univ-avignon.fr - 06 72 64 89 35

Co-directeur : RENAUD Lise CNE
Co-encadrant : HAMZAOUI Ouassim LBNC

Titre en français : **Les sciences humaines et sociales face aux "data"**
Recompositions des pratiques de recherche en SHS entre ouverture et protection des données

Titre en anglais : *Human and social sciences facing "data": recomposition of research practices in HSS between data openness and protection*

Résumé en 7 lignes : Le projet de thèse vise à explorer et caractériser la manière dont les SHS sont aujourd'hui affectées structurellement et dans leurs pratiques de recherche par l'exploitation de données numériques. Laissant de côté les effets des Big data sur la démarche scientifique, la recherche s'intéresse plus spécifiquement aux conséquences, en termes de réajustements organisationnels, méthodologiques et professionnels du champ des SHS, des deux processus que sont la préparation et la mise en oeuvre du Règlement Général pour la Protection des Données à caractère personnel (RGPD) d'une part, et le mouvement d'incitation à l'ouverture des résultats et des corpus de données de recherche d'autre part (Open science).

Mots clés: Open science - RGPD - pratiques de recherche - Sciences humaines et sociales - données numériques

1 - Présentation du sujet

Le projet de thèse interroge les usages vraisemblablement croissants de données numérisées ou nativement numériques en sciences humaines et sociales (SHS) et les recompositions des conditions de production des connaissances qu'ils engendrent. Cette question est discutée depuis plus d'une décennie, mais souvent sous le seul angle de la massivité prétendument inédite de ces données numériques (Plantin & Russo, 2016). La focalisation sur les implications ontologiques (Anderson, 2008), identitaires (Mounier, 2010) et analytiques (Boullier, 2015) des *Big Data* dans la production scientifique est le fruit d'un effet de mode qu'il convient de relativiser (Ollion & Bollaert, 2015). Les enjeux cruciaux que ces études soulèvent font encore l'objet de controverses sur les avantages comparés de ce nouvel état des données en termes de possibilités analytiques en sciences sociales (Boyadjian, 2017). Surtout, ils ne doivent pas occulter les autres transformations par

lesquelles l'exploitation de données numériques affecte, de façon structurelle et matérielle, la recherche scientifique.

Dans le cadre de ce projet, nous identifions principalement deux mutations en cours : le processus de préparation et de mise en oeuvre du Règlement Général pour la Protection des Données à caractère personnel (RGPD) d'une part, et le mouvement d'incitation à l'ouverture des résultats et des corpus de données de recherche d'autre part (Open science). Ces deux phénomènes participent effectivement d'une dynamique de "datafication" des SHS, en ce qu'ils tendent à donner une place centrale, ou du moins nodale, au travail et aux activités de mise en forme juridique et technique des données traitées dans la chaîne de production scientifique (Marres, 2012).

Mêlant des approches relevant aussi bien de la science politique que des sciences de l'information et de la communication, ainsi que de la sociologie des sciences et des techniques (*Science and Technology Studies*), la recherche vise à explorer et caractériser la série de réajustements organisationnels, méthodologiques et professionnels par lesquels se manifestent ce "souci de la donnée numérique" en grande partie renouvelé par les "data" en SHS. L'intérêt d'une approche pluridisciplinaire de science politique et sciences de l'information et de la communication est d'offrir un positionnement multifocale qui prenne en considération non seulement les stratégies d'acteurs, les logiques et injonctions institutionnelles, mais au-delà le processus socio-symbolique même de production des données lorsqu'il s'incarne dans des dispositifs et des instruments.

Comprendre ce procès de production de la recherche en SHS est un enjeu épistémologique pluridisciplinaire dans la mesure où les données numériques, comme construits sociaux, induisent et construisent une certaine vision du monde, que leur utilisation et leur mobilisation contribuent à rendre légitime (Bullich & Clavier, 2018).

Contexte et utilité sociétale

La question des données de recherche en sciences sociales devient un problème dans l'espace public en 2018. Après le scandale mondiale *Facebook-Cambridge Analytica* peu de temps avant l'entrée en vigueur du RGPD, nous observons un rare et exceptionnel moment où la légitimité d'une recherche scientifique « à caractère public » se retrouve violemment prise à partie dans le champ politique, dans le cadre de l'Affaire Benalla, qui éclabousse la Présidence de la République en France, en juillet et août 2018, et le "*Desinfogate*" qui lui succède. Par sa couverture médiatique, les stratégies de politisation croisées auxquelles il a donné lieu, ainsi que le volume des plaintes qu'il a suscitées, le *Desinfogate* a mis en évidence de façon inédite le domaine de la recherche en sciences humaines et sociales dans l'action de régulation de la CNIL, jusqu'ici limitée à la recherche médicale et à la statistique publique. Cet épisode politico-judiciaire pose des questions relatives à l'application pratique du RGPD dans le cadre d'activités de recherches en SHS, celle du statut public des données personnelles issues du microblogging, celle de la conformation au principe de "minimisation", et enfin celle relative aux possibilités pratiques d'open research. Le *Desinfogate* apparaît donc comme une séquence accélératrice d'un processus de mise sur agenda et de mise en politique du traitement des données personnelles à des fins de recherche. Ce processus est au coeur du projet de thèse. Il se

développe dès le début des années 2010 à la faveurs de cinq principaux éléments de contexte à prendre en compte :

1) la préparation dès 2012, l'adoption le 14 avril 2016 au niveau européen, la déclinaison nationale le 25 mai 2018 et la mise en oeuvre depuis cette date du RGPD (*General Data Protection Regulation - GDPR*), notamment par la nomination des Délégués à la Protection des Données (Data Protection Officers - DPO) dans les Universités et les organismes de recherche, en remplacement des Correspondants Informatiques et Liberté (CIL) liés à la CNIL, mais aussi par la publication de guides de bonnes pratiques dans le traitement des données à caractère personnel (DCP), comme celui de l'INSHS mis en ligne sur l'intranet du CNRS le 3 juin 2019 ;

2) le mouvement de réflexion et de codification international de l'Open science, pensé comme un nouveau paradigme, résumé par les principes du FAIR (*Findable, Accessible, Interoperable and Resable*[1]), appliqués aux données du programme de recherche H2020 de l'Union européenne, et ses déclinaisons Open source et Open edition ;

3) le développement d'un véritable écosystème d'infrastructures de la recherche publique impliquées en France dans l'hébergement, la gestion, l'archivage et la protection des données, qui forment un écosystème de ressources et de solutions mutualisées : les TGIR Huma-Num (2013) et Progedo (2008), le CASD (Centre d'Accès Sécurisé aux Données) et le CINES (Centre Informatique National de l'Enseignement Supérieur) ;

4) la mobilisation d'acteurs réformateurs et d'entrepreneurs de morale, impliquant des personnels du champ de la documentation, des ingénieurs d'étude et de recherche, des juristes et des universitaires en sciences humaines et sociales, en droit et en informatique, notamment, autour du réseau Éthique et Droit ou encore du blog Silex de Calimaq, tous impliqués dans l'animation d'arènes scientifiques régulières consacrées à la "datafication" des SHS, tout au long de la décennie 2010 ;

5) le développement de champs de recherche hybrides donnant naissance à de nouveaux labels et à un ensemble de discours d'escorte visant à encourager et légitimer des formes d'usages des données numériques en SHS : "humanités numériques", "social data science", "data-sociologie", "opinion mining" et "sentiment analysis", "computational social science" et "digital methods".

Ce sujet de thèse est une invitation à relier ces éléments contextuels et ces mutations, et à en explorer d'autres, afin de rendre compte des implications profondes induites par l'usage apparemment croissant des données numériques dans la production des connaissances des sciences humaines et sociales sur la société. La thèse a donc vocation à contribuer à l'examen des conditions pratiques de la production du savoir sur la société. Elle interroge la légitimité même des SHS, actuellement, dans le rapport entre "savoir" et "pouvoir" ou entre "connaissance" et "gouvernement", décrit par Michel Foucault. En inscrivant les "data" dans l'histoire de la quantification des comportements et des pratiques sociales, la thèse doit questionner les promesses

[1] Trouvables, accessibles, interopérables et réutilisables.

de connaissances comportementales inédites, le retour du mythe de la prédiction algorithmique, mais aussi une certaine déstabilisation du monopole de l'Etat en termes de connaissance des populations et des publics gouvernés, au profit d'une appropriation privée des données comportementales et personnelles par les GAFAM.

Enjeux théoriques et hypothèses académiques

La thèse s'inscrit dans une sociologie des pratiques de la recherche et dans un travail épistémologique autour de la notion de "donnée". Dans ce cadre, cinq hypothèses peuvent guider la recherche :

La première hypothèse concerne les effets de la "datafication" des SHS sur les logiques de mobilisation de certaines catégories d'acteurs dans la division sociale du travail scientifique, impliquant des recompositions des relations du travail dans les laboratoires, une redistribution des rôles, mais aussi l'émergence de nouvelles expertises et la revendication de compétences spécifiques liées au travail des données et à la mise en conformité des traitements qui leur sont réservés.

La deuxième hypothèse envisage la "datafication" des SHS comme un processus non simplement technique mais résultant plus globalement d'un faisceau d'injonctions qui prescrivent et légitiment une conception de la science, des données et de la société. Elle considère le caractère agissant et structurant des discours d'escorte (Jeanneret, 2014), des instruments et des dispositifs de régulation et d'incitation dans la production de la connaissance. La construction de normes peut en ce sens être envisagée comme un processus communicationnel qui configure des pratiques mais aussi qui circule et s'incarne dans un ensemble de supports et de dispositifs matériels.

La troisième hypothèse aborde la question du rôle des principes et règlements juridiques comme instrument de "normalisation" de la production scientifique, en ce qu'ils participent et accélèrent en premier lieu le mouvement d'homogénéisation internationale des formes légitimes des traitements de données, parce qu'ils s'instituent ensuite en points de passages obligés du processus éminemment social de construction scientifique de la "donnée" (Latour et Woolgar, 1988) et enfin, car cette focalisation du champ de la recherche en SHS sur les enjeux réglementaires relatifs à la protection des données personnelles, occulte d'autres dynamiques comme l'appropriation privée des moyens de produire la connaissance sur la société.

La quatrième hypothèse porte sur les potentialités managériales des logiques d'Open science et Open access. Associées à l'essor du ranking scientifique et de la scientométrie, dans l'évaluation de la qualité des publications et ses effets de réputation et de carrière, elles ne feraient que prolonger les dynamiques sociales caractéristiques du benchmarking (Bruno et Didier 2013), par-delà la reconfiguration du modèle économique et social de l'édition scientifique et l'institutionnalisation discursive d'une légitimité démocratique.

La thèse pourra enfin explorer une cinquième hypothèse relative à la refondation de la légitimité des SHS face à la puissance du Big data (Boullier 2015), entre d'une part, *opinion mining* et

sentiment analysis exploitant les traces massives de comportements au profit d'un capitalisme de surveillance (Zuboff, 2019) et, d'autre part, la prophétie provocatrice de Chris Anderson relative à la disparition de la théorie (Anderson 2008), et partant de l'utilité sociétale des sciences sociales face aux capacités corrélatives et prédictives du machine learning.

Cinq cadres théoriques pourront être mobilisés dans l'examen des hypothèses proposées :

- La sociologie des sciences et des techniques et l'analyse des technologies cognitives nourries de l'anthropologie des artefacts cognitifs (Goody 1979, Hutchins 1984), pour lesquelles il n'existe pas de savoir sans les supports matériels d'élaboration, de mémorisation, de discussion et de circulation (Akrich, Callon, Latour, 2006).
- Une sociologie des usages du numérique, appuyée sur la sociologie pragmatique qui étudie les situations d'activités équipées (Thévenot, Conein & Dodier, 1993) et l'approche ethnométhodologique des appropriations des médiations techniques, qui rendent compte des ajustements, des habitudes et des familiarités dans la naturalisation du couplage homme-machine.
- La sociologie de l'activité scientifique qui insiste sur « l'encastrement organisationnel des pratiques scientifiques » (Owen-Smith, 2001) et tout particulièrement sur la façon dont les processus de procéduralisation travaillent l'organisation des laboratoires, transforment les stratégies collectives et individuelles, modifient les tâches quotidiennes des chercheurs, et reconfigurent les hiérarchies et spécialisations professionnelles. Cette démarche invite ainsi notamment à déconstruire la distinction institutionnelle entre activités de recherche et activités dites "de support à la recherche", et donc à reconnaître l'impact des activités d'autres catégories de travailleurs dans l'élaboration des "chaînes de médiation" (Latour, 2011) nécessaires à la "production scientifique".
- La socio-sémiotique de la médiation scientifique qui montre combien les enjeux de reconnaissance professionnelle jouent dans les logiques, principes et formes de la médiation scientifique (Jeanneret, 1994 ; Jacobi, 1999).
- La sociologie politique des problèmes publics (Neveu, 2015) et la sociologie de l'action publique (Lascoumes & Le Galès 2007) qui permettent d'objectiver la façon dont la politique de recherche et d'innovation se définit comme un vaste espace de négociation entre une multitude d'acteurs et d'organisations qui, en fonction des contextes institutionnels, de leurs logiques internes et des modalités de rationalisation embarquées dans les instruments qu'ils/elles utilisent, produisent des cadrages qui instituent des faits et pratiques sociales en "problèmes publics", nécessitant une reconfiguration des logiques et des dispositifs de régulation.

Adéquation entre le projet et les axes scientifiques de la FR

Ce sujet de thèse procède des pistes de recherche ouvertes dans le cadre du projet GOOW (GOuvernance des cOrpus de recherche issus du Web) financé par la FR Agorantic en 2018. Il porte un regard à la fois rétrospectif et prospectif sur les transformations des conditions pratiques de la recherche en SHS face à la massification de données numériques plus ou moins disponibles et

exploitables, les incitations et opportunités d'approches pluridisciplinaires pertinentes, les politiques publiques d'ouverture des données de la science et les mesures de protection et de sécurités qui les accompagnent dans le cadre du RGPD. Autant d'enjeux qui sont aujourd'hui au coeur des missions d'une Fédération promouvant la recherche interdisciplinaire sur les "sociétés numériques".

Terrains et méthodologies

La recherche combinera des méthodologies sociologiques, ethnographiques et d'analyse de discours. La thèse explorera tout ou partie des terrains suivants :

- **Sociologie nationale des DPD** : qui sont les Délégués à la Protection des Données (DPD) dans l'espace scientifique aujourd'hui ? Quels sont leurs profils, leurs missions, leurs environnement de travail et leurs pratiques ? L'enquête concernera la population des DPD dans les 59 universités françaises, ainsi que dans les organismes de recherche en SHS (CNRS, INSHS, INRAE...). Elle pourra procéder d'un questionnaire auto-administré avec relance, permettant de nourrir une base de données sociographique, destinée à une analyse quantitative, que des recherches documentaires pourront enrichir. Des entretiens semi-directifs, avec des DPD et leurs interlocuteurs, destinés à l'analyse plus précise des lettres de mission et des pratiques contextualisées, pourront être conduits sur quelques sites identifiés pour leur représentativité ou leur singularité. En mobilisant certains apports de l'analyse de discours, un travail plus spécifique sera mené autour du lexique mobilisé par les DPO et sur la rhétorique qu'ils déploient afin de justifier leur mission d'un point de vue scientifique et social.
- **Observations *in situ* dans des organismes de recherche** : L'enquête de terrain par immersion, observations et entretiens pourra être conduite dans plusieurs organismes de recherche, dont une Très Grande Infrastructure de Recherche (TGIR), comme Humanum ou Progedo. Il s'agira notamment d'analyser l'émergence de ce type d'infrastructures dans le paysage scientifique français, les missions qui lui sont assignées, son organisation, les moyens, supports communicationnels et les effectifs affectés, l'activité et les résultats. L'enquête pourra également concerner une ou plusieurs Unités Mixtes de Recherche, Équipes d'Accueil ou Fédérations de recherche au sein d'une université, où il s'agirait davantage ici d'observer et analyser les transformations en cours des relations entre chercheurs, ingénieurs et DPO et la mise en place des Plans de gestion de données sur les projets de recherche (ANR...).
- **Terrain comparatif** : Le doctorant pourra envisager une dimension internationale et éventuellement comparative à la recherche, à l'occasion d'un stage de recherche de six mois dans un laboratoire étranger. Il s'agit notamment d'examiner la circulation des modèles et des standards, et leurs déclinaisons nationales, en particulier au sein de l'Union européenne où le "FAIR" est désormais une norme pour les pratiques de recherche.

2 - Profil du candidat :

La personne candidate devra disposer d'une solide formation de sciences sociales, en sociologie politique, science de l'information et de la communication ou sociologie des sciences et des techniques. Elle fera preuve d'une grande capacité de distanciation réflexive afin d'investir le champ thématique de l'analyse des pratiques de quantification dans les sciences humaines et sociales. Si une formation aux méthodes quantitatives constitue un atout, celle-ci n'est pas un prérequis tant que la personne fait montre d'un intérêt pour l'évolution des outils numériques de mesure et d'objectivation statistique ainsi que de dispositions au travail pluridisciplinaire. Elle sera invitée à participer activement aux travaux de la fédération de recherche Agorantic (FR 3621).

3 - Références bibliographiques :

- **Anderson C., 2008**. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Wired Magazine, https://www.wired.com/2008/06/pb-theory/.
- **Boullier D., 2015**. « Les sciences sociales face aux traces du big data. Société, opinion ou vibrations ? » *Revue française de science politique* 65 (5‑6): 805‑28.
- **Boyadjian J., 2017**. « Les conditions de scientificité des Big Data en science politique ». *Revue française de science politique* 67 (5): 919‑29.
- **Broudoux É., Chartron G., (dir.) 2015**. *Open data, big data : quelles valeurs, quels enjeux ?*, Bruxelles, De Boeck.
- **Bullich V., Clavier V., (dir) 2018**. « Production des données, "production de la société". Les Big Data et algorithmes au regard des sciences de l'information et la communication », *Les Enjeux de l'information et la communication* 19 (2).
- **Chartron G., Schöpfel J., (dir.) 2017**. « Open access et Open science en débat », *Revue française des sciences de l'information et de la communication* 11, http://journals.openedition.org/rfsic/3331.
- **Ibekwe-Sanjuan F., 2014**. « Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ? », XIXème Congrès de la Sfsic. *Penser les techniques et les technologies : Apports des Sciences de l'Information et de la Communication et perspectives de recherches*, Toulon, France: 1-10, hal-01066202.
- **Lascoumes P., Le Galès P.. 2007**. *Sociologie de l'action publique : Domaines et approches*, Paris, Armand Colin.
- **Latour B., 2011**. *Pasteur : guerre et paix des microbes. Suivi de Irréductions,* Paris, La Découverte.
- **Latour B., Woolgar S., 1988**. *La Vie de laboratoire. La production des faits scientifiques*, Paris, La Découverte.
- **Marres N., 2012**. « The redistribution of methods: on intervention in digital social research, broadly conceived », *The Sociological Review* 60: 139-165.
- **Mounier P., 2010**. « Manifeste des Digital Humanities », *Journal des anthropologues* 122-123 : 447-452.
- **Neveu É., 2015**. *Sociologie politique des problèmes publics*, Paris, Armand Colin.
- **Ollion E., Boelaert J., 2015**. « Au-delà des Big Data. Les sciences sociales et la multiplication des données numériques », *Sociologie* 6(3): 295-310.
- **Owen-Smith J., 2001**. « Managing laboratory work through skepticism : processes of evaluation and control ». *American Sociological Review* 66 (2): 427-452.
- **Paquienséguy F., (dir.) 2016**. *Open data*, Paris, Archives contemporaines.
- **Plantin J.-R., Mabi C., Monnoyer-Smith L., (dir.) 2017**. *Ouvrir, partager, réutiliser : Regards critiques sur les données numériques*, Paris, Éditions de la Maison des sciences de l'homme, http://books.openedition.org/editionsmsh/9026.
- **Plantin J.-R., Russo F., 2016**. « D'abord les données, ensuite la méthode ?. Big data et déterminisme en sciences sociales ». *Socio. La nouvelle revue des sciences sociales* 6 (mai): 97‑115.

4 - Opportunités de mobilité à l'international du doctorant :

Le.la candidat.e sera invité.e à effetuer un séjour de recherche de 3 à 6 mois au Centre for Science and Technology Studies de l'Université de Leiden (Pays-Bas), et plus spécifiquement dans le cadre du "Research Hub" dirigé par Thed van Leeuwen et André Brasil. Ce dispositif est consacré à l'Open science et propose d'analyser l'impulsion politique à l'ouverture et ses conséquences à différents niveaux du système scientifique, mais aussi de collecter des données sur les pratiques de science ouverte (publication en libre accès, référentiels de données ouverts, l'innovation ouverte, laboratoires ouverts et science citoyenne) et d'élaborer des indicateurs de développement des infrastructures et des pratiques scientifiques ouvertes. Il entreprend également une réflexion sur les effets de l'ouverture sur la culture et la pratique de l'évaluation de la recherche, sa relation avec les systèmes de récompenses académiques et les applications des formes ouvertes d'évaluation.