De l'analyse informatique de données de la société à l'analyse sociale de l'informatique. Un cheminement guidé par les études de genre vers un décloisonnement disciplinaire et une posture réflexive.

Cécile Favre

MÉMOIRE

en vue de l'obtention de l'

Habilitation à diriger des recherches

délivrée par

l'Université Lumière Lyon 2

Discipline INFORMATIQUE

présentée par

CÉCILE FAVRE

Unité de recherche : ERIC

De l'analyse informatique de données de la société à l'analyse sociale de l'informatique.

Un cheminement guidé par les études de genre vers un décloisonnement disciplinaire et une posture réflexive.

---

soutenue le jeudi 7 mars 2024

JURY

Christine LARGERON	Professeure en informatique, Univ. Jean Monnet Saint-Étienne	*présidente*
Patrice BELLOT	Professeur en informatique, Univ. Aix-Marseille	*rapporteur*
Bénédicte LE GRAND	Professeure en informatique, Univ. Paris 1 Panthéon Sorbonne	*rapportrice*
Patrick MARCEL	Professeur en informatique, Univ. d'Orléans	*rapporteur*
Isabelle COLLET	Professeure en sciences de l'éducation, Univ. de Genève (Suisse)	*examinatrice*
Anne LAURENT	Professeure en informatique, Univ. de Montpellier	*examinatrice*
Phyllis RIPPEY	Professeure en sociologie, Univ. d'Ottawa (Canada)	*examinatrice*
Jérôme DARMONT	Professeur en informatique, Univ. Lumière Lyon 2	*garant*

Mémoire composé avec LaTeX 2ε.
Version du 25 janvier 2024.

Cécile FAVRE

De l'analyse informatique de données de la société à l'analyse sociale de l'informatique. Un cheminement guidé par les études de genre vers un décloisonnement disciplinaire et une posture réflexive.

Résumé

Suite à un double cursus de formation portant à la fois sur l'informatique décisionnelle et sur la fouille de données, les recherches menées après la prise de poste se sont poursuivies sur ces deux volets, amenant à une analyse de données de la société s'appuyant sur des contributions en informatique (modélisation d'entrepôts de données et de métadonnées pour les lacs de données, analyse de la diffusion dans les médias sociaux et détection de fausses informations par des approches multimodales).

La rencontre avec les études de genre a amené également une ouverture vers de nouveaux axes de recherche, avec un regard en prise avec des enjeux sociétaux. Ainsi, une perspective scientométrique a permis d'analyser l'empreinte carbone liée aux déplacements en conférences et la place des femmes au sein de la discipline informatique. Par ailleurs, une démarche sociologique exploratoire a permis d'entrevoir l'analyse des politiques de quotas dans les comités de sélection en informatique. Ces axes ont ainsi amené à une analyse sociale de l'informatique. Le parcours réalisé en recherche et en enseignement, en étant basée dans une UFR de Sciences Humaines et Sociales, a finalement conduit à un décloisonnement disciplinaire en direction de l'interdisciplinarité, mais aussi vers une posture réflexive, débouchant sur des questionnements épistémologiques. Ceci conduit à définir un programme de recherche post-HDR ambitieux, reflet de cette identité scientifique, qui fasse sens et science pour la société.

Mots-clés

réseaux sociaux • *Graph OLAP* • modélisation d'entrepôt de données • lac de données • diffusion de l'information • détection de fausses informations • scientométrie • bibliométrie • écologie • études de genre • carrières scientifiques • politiques de quotas • interdisciplinarité • réflexivité • épistémologie

ERIC – UR 3083 Université Lumière Lyon 2

Cécile FAVRE

From the computer analysis of society's data to the social analysis of computer science. A path guided by gender studies towards the decompartmentalization of disciplines and reflexivity.

Abstract

Following a dual training course covering both business intelligence and data mining, the research carried out after taking up the position of associate professor continued in these two areas, leading to an analysis of society's data based on contributions in computer science (modelling of data warehouses and metadata for data lakes, analysis of social media distribution and detection of fake news using multimodal approaches).

The encounter with gender studies has opened the door to new research areas, with a focus on societal issues. For example, a scientometric perspective has made it possible to analyze the carbon footprint associated with conference travelling, and the place of women within the computer science. In addition, an exploratory sociological approach has provided insights on quota policies in selection committees in the field of computer science. These perspectives have led to a social analysis of the discipline.

This research and teaching path anchored in a humanities and social sciences unit has led to decompartmentalizing disciplines and towards interdisciplinarity, but also towards a reflexive posture, resulting in epistemological questions. This has prompted the definition of an ambitious post-HDR research program reflecting this scientific identity, and which makes sense and science for society.

Keywords

Social Networks • Graph OLAP • Data Warehouse Modeling • Data Lake • Information Diffusion • Fake News Detection • Scientometrics • Bibliometrics • Ecology • Gender Studies • Scientific Careers • Quotas policies • Interdisciplinarity • Reflexivity • Epistemology

ERIC – UR 3083 Université Lumière Lyon 2

Au cheminement individuel dans, avec, grâce et au service du collectif…

Remerciements

La gratitude va de pair avec l'humilité comme la santé avec l'équilibre.
Citation traduite attribuée à Elizabeth Goudge (1900-1984), romancière britannique, « L'Arche dans la tempête » (1934)

Les remerciements dans un mémoire constitue l'espace de rédaction sans doute le plus personnel, même si ce mémoire d'HDR sera fortement coloré de qui je suis. Si cette partie peut être jugée de circonstance et sa rédaction vécue comme un « passage obligé », elle peut aussi être un espace d'expression authentique, en lien avec son origine étymologique de l'ancien français « mercier » : exprimer la gratitude.

Dans la pratique rédactionnelle d'une thèse de doctorat ou d'une habilitation à diriger des recherches, la section des remerciements est souvent une section rédigée à la fin si je me réfère à quelques échanges sur le sujet.

Dans cette habilitation à diriger des recherches, le premier jet de cette section a été rédigé avant même le démarrage de la rédaction de la partie scientifique elle-même. Il n'y a aucun doute pour moi que ce document d'HDR est le fruit d'un cheminement scientifique, aussi plus personnel, mais en aucun cas solitaire. Ce besoin de remercier avant de commencer la rédaction est apparu comme une nécessité, étape précédant l'écriture sur des contributions souvent collectives et sur des réflexions personnelles nourries d'échanges.

La vie est faite de rencontres, souvent inspirantes, qu'elles soient éphémères ou qu'elles marquent le démarrage d'une relation plus nourrie dans le temps. Je tiens ici à honorer ces rencontres qui m'ont forgée bien au-delà de ce que peuvent imaginer les personnes rencontrées. Je saisis alors cette occasion unique qui m'est donnée de le mettre par écrit, sans me limiter dans les mots, donc sans contrainte de longueur.

Lister des remerciements, c'est prendre le risque d'oublier, de ne pas parvenir à être exhaustive. C'est aussi se risquer à un processus de catégorisation, qui ne rend pas forcé-

ment compte de la réalité de la porosité des catégories ainsi nommées.

La démarche d'énumération prend le risque d'être excluante, et ce n'est pas mon intention ici, bien évidemment. Je m'excuse d'avance pour les personnes qui ne se reconnaîtraient pas dans ces remerciements, car si elles ont croisé mon chemin elles ont sans aucun doute d'une façon ou d'une autre contribué à qui je suis aujourd'hui. Alors Merci.

Je tiens à m'excuser par avance également auprès des personnes pour qui la catégorie mise en avant ne rendrait pas assez compte de la réalité de notre lien ou de notre vécu partagé.

Mettre en mots ces remerciements du cœur, c'est prendre conscience de l'étendue de mon réseau de connaissances, et il est tellement vaste. Quel plaisir alors de retracer mon cheminement avec le prisme des relations humaines qui l'ont accompagné.

Je précise ici qu'il ne s'agit pas ici d'accorder trop d'importance à l'ordre dans ces remerciements, mais de mettre en lumière des personnes qui ont permis que ce chemin soit lumineux... et elles sont nombreuses.

Je commence ces remerciements, de manière inhabituelle et non conventionnelle, par mentionner amicalement Guillaume Cleuziou dont un de nos échanges d'il y a plusieurs années était marqué par le conseil d'attendre le moment juste pour entreprendre la démarche d'habilitation à diriger des recherches, se sentir prêt·es plutôt que de céder à une forme de pression, ou d'injonction, parce que les éléments factuels pour l'entreprendre sont là. Je suis heureuse d'avoir suivi ce conseil, ce qui a donné un vrai sens à ma démarche.

Je remercie très chaleureusement Jérôme Darmont qui m'a accompagnée dans cette démarche, de la meilleure façon qui soit par rapport à mon projet, tout au long de mon processus. Cette HDR, ce fut pendant longtemps l'Arlésienne, une chose dont on parlait beaucoup mais qui ne se montrait pas...Alors merci pour la patience, et d'avoir aussi accuilli les changements de rythme. Ce temps était nécessaire pour aboutir à ce résultat, à ce que je souhaitais porter..., à cet aboutissement qui me ressemble. C'était précieux d'avoir pu suivre le rythme juste. Merci d'avoir toujours soutenu cette démarche en lien avec les études de genre, y compris dans ton rôle de directeur de laboratoire. Je t'ai connu comme enseignant à l'IUP ISEA, que de souvenirs de cours enthousiasmants de bases de données et de programmation auxquels j'adorais assister. Que de chemin parcouru depuis. Regarder en arrière avec cette perspective des personnes avec qui on a cheminé, c'est aussi l'occasion de projeter notre parcours dans une évolution qui va bien au-delà de ce qui a été fait : celle de qui nous sommes devenu·es. Merci pour cela.

Merci à Christine Largeron d'avoir accepté de présider ce jury. Merci à Patrice Bellot, Bénédicte Le Grand et Patrick Marcel pour le travail de rapport qu'elle et ils ont accepté de réaliser. Merci à Anne Laurent, Isabelle Collet et Phyllis Rippey d'avoir accepté de siéger dans ce jury en qualité d'examinatrices, pour alimenter les échanges grâce à leur expérience. Le choix que j'ai opéré pour ce jury me permet de témoigner du profond respect que j'ai pour votre travail, pour vos contributions à la science, mais aussi pour la manière

dont vous incarnez ces figures scientifiques inspirantes. Ce sont aussi nos échanges qui m'ont amenée à vous donner ce rôle dans cette étape de ma vie, car j'étais convaincue que cela amènerait à une session de discussion de soutenance enrichissante et passionnante.

Merci à Laurence Tain, à cette rencontre improbable, à ton parcours si inspirant par ta navigation disciplinaire, à nos échanges et nos voyages si enrichissants. Si j'ai fait ce chemin vers les études de genre, c'est avant tout grâce à toi. Merci aussi pour la Maison L.O.R.C.A, ce Lieu Ouvert à la Recherche et à la Création Artistique.

Merci à Nicky Lefeuvre pour ses conseils éclairés me permettant d'aller poser un peu mes pas sur le chemin d'une démarche et d'un outillage sociologique.

Le début de l'histoire de cette habilitation, c'était à l'issue d'une session du Conseil National des Universités (CNU), section Informatique, pour l'attribution des PEDR (Prime d'encadrement doctoral et de recherche) dans laquelle je siégeais. Ces sessions de CNU représentent beaucoup de travail, mais aussi une formidable occasion de rencontrer, d'échanger avec des collègues. Alors je remercie ici vivement les collègues que j'ai pu côtoyer, et notamment Véronique Benzaken, pour les échanges que j'ai eus et qui ont sans aucun doute contribué à ma décision de me lancer dans cette aventure de l'habilitation à diriger des recherches. Je remercie également les membres du bureau de la Section, car cette entrée au CNU n'a pas été anodine pour moi, et les échanges que nous avons eus, leur accueil, leur partage ont été précieux. Je remercie vivement sa présidente Annick Montanvert. Merci à Fabien Torre, Dorian Petit, Frédéric Saubion, Philippe Renevier Gonin et Régine Laleau.

J'exprime également un chaleureux remerciement à mon binôme de CNU, Guillaume Cabanac, qui au-delà du binôme que nous avons formé, a été et est toujours une personne si inspirante à bien des égards. Merci pour nos partages si riches, merci pour ton amitié. Merci aussi pour ce style LaTeX! Et maintenant que cette HDR est finalisée, nous allons pouvoir replonger dans la collaboration... Merci aussi pour les opportunités de belles rencontre que notre lien a permis. C'est le cas de ma rencontre avec Chérifa Boukacem. Je la remercie ici car mon élan d'entreprendre cette habilitation s'est confirmé en partageant avec elle, lors d'une rencontre un peu impromptue, sur toute la symbolique de cette étape.

Je tiens à remercier toutes les personnes avec qui j'ai eu l'occasion de collaborer sur le plan scientifique en informatique, dont les travaux sont la base de ce manuscrit et aux personnes avec qui je collabore actuellement. Merci aux stagiaires, aux doctorant·es, aux co-encadrant·es. Merci notamment à : Adrien Guille, Wararat Jakawat, Abderrazek Azri, Étienne Scholly, Yoann Pitarch, Irina Proskurina, Djamel Zighed, Sabine Loudcher, Jérôme Darmont, Ricco Rakotomalala et Nouria Harbi. Ces collaborations dans le cadre de l'accompagnement à la recherche ont été tellement formatrices pour moi. Ma conviction profonde est que la recherche est avant tout une affaire de rencontres humaines. Merci pour tout ce que vous m'avez apporté bien au-delà des échanges scientifiques. Merci pour le caractère très amical dans la manière de vivre ces co-encadrements, et qui amène une dimension de soutien plus personnel par moment. Merci plus particulièrement à toi Sa-

bine pour cela.

Par rapport à mon enquête exploratoire, je remercie l'ensemble des collègues qui m'ont accordé de leur temps précieux pour des entretiens riches d'apprentissages.

L'environnement professionnel dans l'enseignement supérieur et la recherche est multiple. Le mien comprend des personnes avec qui j'ai la chance de beaucoup échanger. Que les personnes se reconnaissent individuellement dans ces remerciements, même quand elles ne sont pas nommées explicitement. Merci ainsi aux structures et aux collectifs dans lesquels j'ai eu l'occasion d'évoluer, ainsi qu'à toutes les personnes qui font partie de ces collectifs ou y sont passées : enseignantes-chercheuses et enseignants-chercheurs, personnel administratif, personnel des services techniques ou d'entretien, étudiant·es, personnes non affiliées au milieu académique.

Je remercie pour sa patience le personnel de sécurité qui a eu l'occasion de me mettre plusieurs fois à la porte de mon bureau, à des horaires tardifs !

Merci à Shahi Derky pour la gestion administrative de ce processus d'HDR au niveau de la Direction de la Recherche, ainsi qu'à Isabelle von B.

Merci à l'ensemble du personnel support de l'université. C'est en prenant conscience du nombre de personnes que je connais à présent dans l'institution, dans des services support si multiples, que je prends conscience du temps qui est passé ! N'est-ce pas Alain Giacomazzo et Erick Haro ?

Merci à l'ensemble de l'équipe « SID » du laboratoire ERIC, notamment aux permanent·es : Sabine Loudcher sa responsable, Fadila Bentayeb, Omar Boussaïd, Jérôme Darmont, Gérald Gavin, Nouria Harbi, Nadia Kabachi, Mohamed-Lamine Messai, et Juba Agoun nouvellement arrivé. Merci également à Olga Cherednichenko pour accéder à un autre regard. Quelle étape que de soutenir cette HDR dans la même équipe où j'ai débuté en recherche ! J'ai depuis su prendre mon envol, je pense, sous vos yeux Fadila et Omar. Une pensée à Michel Rougié. J'ai aussi une pensée émue pour Nicolas Nicoloyannis.

Merci plus globalement aux membres du laboratoire ERIC, du passé et du présent. Merci notamment à celles et ceux avec qui les échanges viennent nourrir les réflexions : merci à Jairo Cugliari, à Antoine Rolland (en particulier pour ta recommandation de l'ouvrage *Statactivisme*). Merci à Guillaume Metzler et Julien Velcin, notamment pour l'opportunité de travailler avec Irina Proskurina. Merci aux gestionnaires administratives et financières qui se sont succédées Valérie Gabrièle et Habiba Osman ; un merci également à Julien Crevel.

Un grand merci à l'équipe « Dynamiques sociales et politiques de la vie privée » du Centre Max Weber où je suis chercheuse associée depuis quelques années à présent.

J'adresse des remerciements chaleureux à l'ensemble des membres de l'équipe SIGMA du Laboratoire d'Informatique de Grenoble, et à son responsable Cyril Labbé, qui m'ont accueillie dans cette belle équipe, à bureau ouvert (merci Christine Verdier et Agnès Front). Cet accueil fut l'occasion d'encouragements répétés sur le site grenoblois pour

cette HDR : merci aussi à Marlène Villanova et Jérôme Gensel. Cet accueil fut aussi l'occasion de démarrer EDUC'ACTION : sa réorientation au travers des échanges avec Sihem Amer-Yahia qui a débouché sur la dimension éthique du projet a été un virage pour moi, merci Sihem, merci Emilie Hoareau, merci à l'ensemble des membres de l'action !

Merci également aux membres (du passé et du présent) de la Maison des Sciences de l'Homme, et anciennement à l'Institut des Sciences de l'Homme. Bien que je regrette l'usage de ce terme d'Homme, les équipes que j'y ai et que je peux y côtoyer sont profondément humaines. Merci à Valérie Bernardo, Emma Bessieres, Sofiane Bouzid, Jean-Luc Caroujat, Jennifer Chanteloup, Christian Dury, Céline Faure, Amélie Hugot, Didier Leblanc, Mylène Pardoen et aux autres membres.

Merci aux collectifs de recherche pour les collaborations et les discussions enrichissantes : BI4people, EDUC'ACTION, SO-COEQUAL, HF, GDR MaDICS, GDS EcoInfo, AISLF…

Une pensée pour feu le « Cluster Web Intelligence » en Rhône-Alpes, qui a marqué mon démarrage de carrière avec des rencontres scientifico-amicales, et le développement de mon expérience d'organisation d'évènements ! Une pensée spéciale pour Catherine Garbay, Mohand-Said Hacid, Jean-Paul Jamont, Nicolas Lumineau, Laurent Vercouter, et les autres membres (notamment les personnes citées ailleurs dans ces remerciements).

Je remercie également les personnes avec qui j'ai collaboré et collabore encore dans le cadre de mon implication dans des sociétés savantes, notamment l'association INFORSID et l'association EGC. Je salue également les membres de la communauté EDA. Merci aux membres « passé et présent » du bureau d'INFORSID de leur soutien et leur confiance : Rebecca Deneckere, Agnès Front, Régine Laleau, Franck Ravat et Christian Sallaberry. Cette communauté si familiale m'a vue grandir ! Merci à la communauté EGC qui m'est si chère également, et merci à feu le Groupe de Travail Fouille de Données Complexes et ses co-animateurs ! J'ai également une pensée émue pour Arnaud Martin parti si tôt.

Un grand merci à l'équipe pédagogique de la licence MIASHS (Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales) et notamment à Jacques Viallaneix, Loic Bonneval et Monique Dalud-Vincent, ainsi qu'aux gestionnaires de scolarité. Merci également à Isabelle Robert de Saint Victor pour ce binôme de choc ! Cette filière a sans aucun doute, notamment au travers des co-encadrements de mémoire bidisciplinaire, façonné ma manière d'envisager la question de la pluridisciplinarité et de l'interdisciplinarité.

Un grand merci à l'ensemble de l'équipe pédagogique du Master en Etudes sur le Genre de Lyon. Un merci particulier à Laurence Tain qui a initié cette belle dynamique et qui m'a donné l'opportunité de faire partie de cette équipe pédagogique dès la naissance du Master EGALES. L'intégration dans cette équipe constitue un point d'ancrage particulier dans les Sciences Humaines et Sociales, en particulier sur la dimension de genre de cette habilitation. Un merci particulier aux différent·es responsables pédagogiques ainsi

qu'aux gestionnaires de scolarité qui se sont succédées. Merci à Estelle Bonnet, Yannick Chevalier, Aurélie Épron, Marie-Pierre Harder, Marie-Clémence Le Pape, Manuela Martini, Virginie Nicaise (avec une pensée spéciale pour le lycée Boissy d'Anglas d'Annonay!), Cécile Ottogalli, Corinne Rostaing, Muriel Salle, Marianne Thivend, Marie-Jeanne Zenetti et aux ATER qui se sont succédées : Laurine Thizy, Laure Sizaire, Lucia Valdivia, Daria Sobocinska, Aurore Turbiau.

Je remercie à nouveau et plus spécifiquement Yannick Chevalier et Marie-Clémence Le Pape qui, en me permettant d'assister à certains de leurs cours, m'ont donné la chance de continuer à apprendre de manière si pertinente vis-à-vis de la réalisation de mon habilitation. Merci Muriel Salle pour ces moments d'échange si stimulants, si précieux.

Merci au consortium du réseau EGALES, la richesse de ce réseau international a été l'occasion de si belles rencontres, de si beaux échanges, j'en remercie toutes et tous les membres. Le développement de la dimension internationale du Master EGALES constitue pour moi également un sacré challenge, et je remercie ici les différentes personnes du service des relations internationales pour leur travail et l'agréabilité de nos échanges : Merci à Meriem Benmessaoud, Dorothée Orjol-Sousa, Jim Walker et aux membres du personnel qui se sont succédé·es.

Pour une autre aventure, très enrichissante (mais qui commence à dater un peu!), merci au consortium du Master Erasmus Mundus DMKM : à Djamel Zighed pour sa confiance et aux collègues pour la dimension pédagogique, ainsi qu'à Delphine Ferrand…

Merci aux étudiant·es avec qui j'ai tellement appris, quelle que soit leur filière. Nombre de personnes étudiantes sont dans mes pensées.

Merci en particulier aux étudiant·es de formation continue des Masters EGALITES et OPSIE. Ce public est sans aucun doute source d'une belle occasion de cheminement pour moi. Votre implication dans la reprise d'études est inspirante et force le respect! Merci à Nouria Harbi pour sa confiance par rapport à mon implication dans le Master OPSIE.

Merci à la *Team Burger Roi* 2023 d'OPSIE! Je n'oublierai pas!!! J'ai une pensée particulière pour les personnes ayant suivi le cursus EGALITES-FC, qui témoigne de mon attachement à la reprise d'études, à la formation tout au long de la vie, et qui plus est quand il s'agit d'enjeux d'égalité. Vos parcours sont inspirants.

Un merci plus particulier aussi à Andreia Franca pour nos discussions et ton mémoire de master EGALES sur la pédagogie queer soutenu en septembre 2023, ce fut inspirant de finesse d'analyse, de positionnement et d'affirmation.

Le développement de la Formation Continue dans le cursus d'études sur le genre a donc également marqué mon parcours. Je remercie ici chaleureusement Hervé Rozier qui a toujours été d'un grand soutien pendant toute la période où il travaillait dans le service de formation continue. Merci au service de formation continue. Merci également à la petite équipe pédagogique que j'ai constituée par rapport aux cours spécifiques de formation continue, nous donnant l'opportunité de monter une formation si pertinente à mon

sens au fur et à mesure du temps. Un immense merci à Marie-Clémence Le Pape qui, en acceptant la co-responsablité de ce parcours, nous donne l'opportunité de travailler en confiance, avec beaucoup de plaisir, de sérénité et d'efficacité.

Merci aux enseignant·es chercheuses et chercheurs de l'UFR Anthropologie, Sociologie et Science Politique, à son personnel administratif et aux directions qui se sont succédées : merci à David Garibay, Stéphane Chrétien, Anne-Joëlle Bottemer et Sandrine David. Merci à Nathalie Dompnier de son soutien aux questions de genre alors qu'elle était encore directrice de l'UFR et qui se poursuit en tant que présidente de l'Université. Merci d'être si inspirante dans tes mots. Un merci également à Hélène Turlan, pour son soutien et ces moments partagés nourrissants à tout point de vue ! Merci Caroline Frau : je veux bien avoir encore de nouvelles recommandations de lecture et musique de travail !

Je remercie l'ensemble du personne administratif avec qui nous avons œuvré et œuvrons pour que les formations puissent « fonctionner » au mieux possible. Un merci plus particulier à Fatima Haddane, Fadila Taïbi, Zohra Salhi et Anne de Crescenzo, dont le travail a été un appui majeur, sur la fin de cette écriture en particulier. Merci également à Frédéric Bisinger, Olivier Damour, Perrine Guenault, Reza Hadjikhani, Delphine Labadie, Charlotte Lesueur, Aline Maitrias, Adeline Perardelle, Cécile Wolff, et l'ensemble des gestionnaires. J'ai une pensée émue pour Sonia De Nardi.

Merci à l'IREF (Institut de Recherches et d'Etudes Féministes) et à ses membres pour leur accueil chaleureux et ces deux séjours si fructueux pendant mon CRCT en 2019 et mon séjour de 2023. Merci à ses directrices successives : Rachel Chagnon et Thérèse St Gelais qui ont rendu ces séjours possibles. Je ne peux énumérer toutes les personnes... Mais merci à Rébecca Beauvais et Rosemarie Fournier-Guillemette pour m'avoir fait une place dans leur cours et donner la chance d'enseigner à l'UQAM (alors même que j'aurais tellement voulu y faire une partie de mes études). Merci Mélanie Millette pour nos échanges et les opportunités de rencontre grâce au séminaire du LabCMO ! Merci à Chiara Piazzesi pour son accueil italiano-québécois ! Merci aux personnes qui ont partagé des bouts de mon quotidien à l'IREF durant mes séjours. Et en particulier, Marie-Noëlle Bourdiec, Bronja Hildgen. Et *last but not least...* un immense merci à Alice Van der Klei pour TOUT. Les mots me manquent... mais tu sais...

Merci à Pascale Kuntz de m'avoir suivie et soutenue dans l'idée d'organisation de l'atelier PRISME-G (Penser la Recherche en Informatique comme pouvant être Située, Multidisciplinaire et Genrée) à EGC 2018, qui fut un tournant sur ma route ; ce titre d'atelier que j'avais proposé résume tellement ma perception plus que jamais d'actualité.

Quelques mercis également à des personnes avec qui les échanges ont été tellement inspirants et/ou soutenants (désolée pour les oublis) - par ordre inversement alphabétique pour un peu d'originalité : Marie-Jeanne Zenetti, Lise Wagner, Elise Vinet, Jean-Marc Vincent, Marie Vialaret, Genoveva Vargas-Solar, Karine Vanthuyne, Alice Van der Klei, Nathalie Valles-Parlangeau, Ronan Tournier, Sylvie Tomolillo, Stéphanie Tralongo, Emilie Stora, Thérèse St-Gelais, Chantal Soulé-Dupuy, Harriet Silius, Florence Sèdes, Mu-

riel Salle, Virginie Rozée, Claudia Roncancio, Thibaut Rioufreyt, Charlotte Rimaud (un clin d'œil à l'école Malleval d'Annonay après ces retrouvailles québécoises!), Sophie Quinton, Patrice Quinton, Véronique Prudhomme, Joyce Portilla, Chiara Piazzesi, Cécile Ottogalli, Aurélie Olivesi, Aurélie Navarro, Chantal Morley, Claire Morandeau, Mélanie Millette, Alain Mille, Marion Maudet (mention spéciale pour une co-bureau toujours là au bon moment!), Claudia Marinica, Giorgia Magni, Philippe Lopistéguy, Mathieu Lizotte, Anne-Laure Ligozat, Thérèse Libourel, Marie-Clémence Le Pape, Nicky Lefeuvre, Nathalie Lapeyre, Josée Lafond, Nicolas Labroche, Fabien Labarthe, Pascale Kuntz, Stéphanie Kunert, Brigitte Kervella, Zhanna Karimova, Julie Jarty, Emilie Hoareau, Mounira Harzallah, Marie-Pascale Halary, Gabriele Griffin, Adeline Gilbert, Albane Geslin, Isabelle Garcin-Marrou, Simon Gadras, Caroline Frau, Florence Françon, Rosa Figueiredo, Marion Fabre, Olivier Ferret, Montserrat Emperador-Badimon, Violaine Dutrop, Jean Dupuy, Anca Dohotariu, Marie Després-Lonnet, Cyril de Runz, Caroline Dayer, Jérôme David, Michèle Cros, Bruno Crémilleux, Céline Coutrix, Juliette Cleuziou, Yannick Chevalier, Rachel Chagnon, Amandine Catala, Pilar Carrasquer, Gaëlle Calvary, Guillaume Cabanac, Pierrick Bruneau, Maxine Brun, Raphaëlle Bour, Julia Bonaccorsi, Soline Blanchard, Nathalie Bertrand, Ionela Băluță, Mathilde Azzouz, Mathieu Azcué, Clothilde Arnaud, Pierre-Emmanuel Arduin, Sihem Amer Yahia. Merci à vous pour tous ces échanges si enrichissants!

Merci à toutes les féministes qui m'ont inspirée par leurs écrits ou par nos échanges. J'ai une pensée spéciale pour Miriam Grossi et Pinar Selek.

Merci aux collègues informaticiennes mobilisées sur les enjeux d'égalité, merci pour nos échanges qui font se sentir moins seules.

Je poursuis avec des remerciements familiaux. Je tiens d'abord à honorer ici la mémoire de mon grand-père Robert Favre, qui fut Professeur de Littérature du 18ème siècle, dans la même université où j'exerce mon métier aujourd'hui. Je me souviens de sa joie à l'annonce de ma décision de rejoindre l'Université Lyon 2 comme enseignante-chercheuse en 2009, avec la perspective d'une sorte de filiation qui s'opérait. Et mes liens aujourd'hui avec des personnes de l'UFR LESLA me réjouissent aussi de ce point de vue là.

J'ai une pensée emplie d'affection pour ma grand-mère maternelle Corinne Cortesi. Je garde en mémoire le souvenir de mon pot de thèse où tu avais tant œuvré avec Maman pour en faire un moment de dégustation de la cuisine familiale que j'affectionne tant. Le pot d'HDR ne sera pas le même sans ta touche... Tu m'as tant apporté! Ce début d'études supérieures où tu m'as accueillie chez toi, partageant ce quotidien durant deux années, fut un cadeau si précieux dans ma vie.

Mes remerciements vont vers l'ensemble des membres de ma famille.

Merci à mes parents en particulier, Sylvie et Etienne, pour leur soutien indéfectible, et aux personnes qui les accompagnent à présent dans leur vie, Christian et Véronique. Une pensée en direction de Londres pour mon frère Laurent, Lucie, Gustave et Edgar. Une

pensée pour ma sœur Valérie et Clément. Une pensée pour Lilian, Estelle, Manu, Maeva et Emmie.

Un clin d'oeil spéciale à toi ma soeur, Valérie, qui chemine également sur la voie académique, j'apprécie de plus en plus les partages sur nos vécus et questionnements.

Merci à ma cousine Delphine. Nous ne partageons que trop peu de moments mais si je poursuis ces orientations aujourd'hui sur l'inclusion, ça résonne par rapport à toi.

Merci à mon parrain Daniel et à Patricia pour nos discussions si riches et enrichissantes même si peu nombreuses.

Une pensée pour ma marraine Isabelle. Assister à ta soutenance de thèse a été très inspirant.

Merci aux membres de ma « belle »-famille, vous m'apportez tant… Merci à Simone, Myriam et Loic, Inyan et Donoma. Merci à Jacques, « beau-papa », d'être encore présent à ta façon.

Merci à ma famille québécoise ! 30 ans déjà…

Un merci pour les liens de sororité au-delà des liens du sang : Emma, Gaëlle, Myriam, Wendy, Yannick… Très chère Wendy, elle est enfin écrite ! Merci pour ce soutien toutes ces années durant et cet encouragement sans cesse renouvelé à déployer mes ailes…

Merci aux personnes qui ont été ou sont sur mon chemin musical amical, dans ces différents collectifs ! Cela fait partie de mon équilibre de vie ! Une pensée spéciale pour les Coquelicots et la famille Delord qui est en or : Christophe et Gaëlle, Margot et Mathias. Merci à Mariette Wilson qui a su semer des graines de fleur en moi ;-)

Merci aux personnes avec qui j'ai cheminé ici ou là, à telle ou telle place, elles sont si nombreuses et se reconnaîtront j'espère. Merci plus particulièrement à Thierry Wambli pour tout ce chemin partagé qui m'a fait grandir. Une pensée spéciale pour ce clan du stage 3 de novembre 2023 et pour Guillaume. Merci à Anpo Wi. Un immense Merci à Hehaka Win de m'avoir aidée à suivre le fil. Merci aussi à Natacha Traber. Merci pour les apprentissages au cœur des traditions autochtones au Québec : Dominique Rankin, Marie Josée Tardif, Dolorès Contré. Merci Katia Vuichet, Merryl Dellea, Tania Bonnel, Marie-Claire Prieur. Merci à la *dream team* : Denis Perrichon et Marie Guillomot. Merci à Alex Cros, Sylvie Savoie, Shanaz Moussa, Suzanne Blouin et Huguette Lucas de l'autre côté de l'Atlantique, ainsi qu'à Estelle Santerre qui a fait le pont entre deux de mes mondes !

Merci au groupe des bons virages : Blandine Belghit, Lise Wagner et Emilie Stora.

Merci à l'ensemble des Ami·es que je m'excuse de ne pas toutes et tous lister ! Même si le temps passé ensemble est parfois limité, vous êtes dans mon cœur.

Merci à l'ensemble de mes enseignant·es depuis la maternelle du Champ de Mars à Annonay jusqu'aux universités Lyon 1 et Lyon 2, en passant par l'école primaire Malleval, le collège des Perrières et le lycée Boissy d'Anglas, également à Annonay dans mon Ardèche de jeunesse, qui m'ont donné l'envie de transmettre.

Merci à Cédric Wemmert, j'espère que l'histoire que tu m'as invitée à raconter ne fera pas dormir ! Merci pour la relecture attentive et rigoureuse de la ponctuation ;-)

Merci à l'ensemble des personnes qui ont pris de leur temps pour des relectures et des retours précieux : Abderrazek Azri, Patricia Boiron, Yannick Chevalier, Jérôme Darmont, Valérie Favre, Agnès Front, Nouria Harbi, Stéphanie Tralongo, Sébastien Valat, Cédric Wemmert, Marie-Jeanne Zenetti.

Merci Marie-Pascale Halary et Marie-Jeanne Zenetti pour tout le soutien apporté grâce à vos précieux conseils de « chercheuses HDRisées ». Merci Marie-Pascale pour les précisions étymologiques. Merci Marie-Jeanne pour ton expertise en épistémologie et pour le reste... Merci aussi aux échanges avec les collègues qui vivent ce processus d'HDR. Merci notamment à Cécile Ottogalli, Nancy Venel, Veronika Peralta et Lylia Abrouk. Merci à Emilie Hoareau que j'encourage chaleureusement dans son beau projet d'HDR, j'ai hâte de te lire.

Je remercie toutes et tous les collègues des différents laboratoires de France avec leur relance sur l'HDR lors des congrès ou autres occasions qui ont fait que cette perspective d'écriture est restée bien vivante dans la durée ! Une mention spéciale dans cette catégorie à Max Chevalier, Olivier Teste et Franck Ravat !

Je décerne une mention spéciale à Patrice Bellot, me présentant dans une soutenance de thèse qu'il présidait, par anticipation, comme maîtresse de conférences habilitée à diriger des recherches !

Merci Marie pour ton accompagnement kinésiologique si précieux et si soutenant dans la durée ! Ca y est... enfin !

Gratitude aux peuples autochtones et à leurs savoirs qui me portent depuis tant d'années.

Je remercie du fond du coeur mon compagnon de vie, Sébastien Valat. Les mots ne sauraient exprimer toute la gratitude que j'ai pour tout le précieux de ce que nous partageons, nos échanges si riches, pour tout ce soutien sous des formes multiples. Merci pour les relectures et pour la patience vis-à-vis de ce projet si personnel et collectif à la fois !

Merci finalement aux lieux qui m'ont accueillie pour écrire ! Ce manuscrit est le résultat d'une écriture elle-même caractérisée par une forme importante de nomadisme : de Montréal au gîte « Shamb-Dolma » à La Biolle, en passant par celui de La Biolette (merci à la gardienne du lieu Danièle), par des bibliothèques et d'autres ailleurs !

Ce processus d'écriture fut un véritable processus de transformation intérieure profonde, je célèbre l'envol du papillon !

Merci, enfin, à la belle énergie inspirée et créative de Tawana qui a permis de suivre ce fil, allant de découverte en découverte, d'émerveillement en émerveillement !

Mitakuye Oyasin...

Migwetch ! Merci !

Table des matières

Préambule : charte rédactionnelle

> Si les efforts réalisés depuis le XIIIe siècle afin de renforcer la domination masculine font partie de l'histoire de notre langue, rien n'oblige à les transmettre en héritage aux générations futures, ni à les laisser parasiter nos imaginaires et retarder l'avènement de l'égalité.
>
> *Éliane Viennot, linguiste française, « Le langage inclusif : Pourquoi ? Comment ? » (2018)*

Ce mémoire constituant un travail de rédaction particulier sous différents aspects, il m'apparaît nécessaire d'avoir une forme de réflexivité sur la manière d'écrire, et de présenter en prélude les choix d'usage sous forme d'une charte rédactionnelle. Il s'agit de présenter non seulement les choix faits, mais de prendre le temps de les discuter. Cette charte aborde deux aspects importants, qui nécessitent selon moi d'être précisément explicités, voire même motivés : 1) le premier concerne le choix du pronom personnel ; 2) le second concerne le choix d'un protocole rédactionnel non discriminant.

Du « je » et du « nous »

En matière d'écriture d'une Habilitation à Diriger des Recherches (HDR), il m'apparaissait incontournable de se poser la question du choix de l'usage du pronom personnel « je » ou « nous » ? Le « je » pourrait être préféré parce que ce mémoire vise à démontrer ma capacité à diriger des recherches. Le « nous » pourrait être choisi, retranscrivant ainsi la dimension collective des travaux présentés, dans le cadre d'encadrement de travaux de recherche ou de collaborations. Mais aussi parce qu'en matière d'écriture scientifique, le « nous » est souvent d'usage en France.

S'il apparaît classiquement de faire un usage homogène du pronom pendant tout un document, je m'autoriserai dans ce mémoire à l'usage des deux. Il s'agit d'une part de

pouvoir rendre compte du travail collectif réalisé (notamment dans ce mémoire qui met en avant des travaux fruits pour la plupart d'un encadrement scientifique auquel j'ai participé). Il s'agit d'autre part de rendre compte de raisonnements, de points de vue qui m'appartiennent, notamment en lien avec une réflexivité sur les travaux menés et retranscrivant ma propre construction d'identité scientifique, en ayant donc également recours au « je ».

Protocole rédactionnel non discriminant

En 2009, j'étais recrutée sur un poste de Maître de Conférences. Quelques années plus tard, je commençais à me présenter comme Maîtresse de Conférences. Une évolution linguistique qui n'était pas anodine mais assez spontanée à l'époque !

Lors de mon recrutement, quand Laurence Tain m'invitait à intégrer l'équipe pédagogique du Master ÉGALES (Études Genre et Actions Liées à l'Egalité dans la Société) qu'elle était en train de créer, je dois reconnaître que c'est avec une certaine naïveté que j'acceptais. Naïveté politique… mais un enthousiasme certain de la perspective attractive d'un travail collectif dans une équipe pluridisciplinaire !

Ma rencontre avec Yannick Chevalier dans cette équipe fut déterminante dans mon usage de la langue française. Maître de Conférences en stylistique française que je dénommais à l'époque « grammairien », je découvrais cette manière d'écrire avec des « ·e », que d'ailleurs je ne nommais pas forcément ! Sans doute dans une démarche d'intégration collective au sein du Master, j'adoptais quelques nouveaux principes d'écriture, avec une application certaine, notamment dans mes mails aux étudiant·es.

Ce n'est qu'un peu plus tard, fin 2017, lorsque le *buzz* sur l'écriture inclusive battait son plein, que j'étais finalement amenée, en pleine médiatisation des controverses autour du langage, à prendre conscience de la symbolique politique si forte, alors même que c'était entré dans ma pratique quotidienne depuis plusieurs années. Je me demandais alors même ce qu'aurait pensé mon cher grand-père, Robert Favre, Maître de Conférences en lettres à l'Université Lyon 2 au début des années 80, spécialiste du 18ème siècle, de cette évolution de la langue française, imaginant les discussions que nous aurions pu avoir sur le sujet !

Manière d'écrire utilisée en toute simplicité depuis l'ouverture du Master Genre en 2011, je ne voyais pas cela comme un combat en tant que tel que j'aurais mené en tant que féministe. Et si j'avais adopté quelques principes d'écriture sans me poser trop de questions, la lecture des chartes rédactionnelles, que nous demandions à nos étudiant·es de Master Genre en préambule de leur mémoire après avoir suivi un cours de Yannick Chevalier sur le sujet, me permettait de peu à peu conscientiser l'enjeu de cet usage.

Voilà plus d'une dizaine d'année que j'évolue dans un environnement où les questions d'égalité femmes-hommes sont abordées dans un cadre d'apprentissage que constitue un

Master en Études sur le genre. M'impliquer dans cette équipe pédagogique, ce fut pour moi la prise de conscience de l'importance de contributions disciplinaires multiples traitant d'un même objet. C'est notamment au travers de ces questions d'égalité femmes-hommes que j'ai pu appréhender l'enjeu de l'interdisciplinarité. Concevoir que les mots sont importants m'apparaît aisé, mais toucher à ce que tout ce que le langage véhicule, y compris d'un point de vue de la construction sociale, c'est amener la réflexion sur sa propre manière d'exprimer les idées, c'est s'attacher à la forme pour à la fois mieux exprimer le fond, et le faire en accord avec des valeurs. L'égalité est une des valeurs qui a donné l'élan de ce travail de rédaction. C'est dans ce cadre que l'adoption d'un protocole rédactionnel non discriminant relève pour moi d'une nécessité, d'un automatisme, même s'il pourrait être critiqué pour sa lourdeur visuelle (argument auquel j'aurai tendance à rétorquer qu'il s'agit d'avantage d'une question d'habitude).

Début 2019, alors que la perspective de la rédaction de cette HDR approchait, je profitais avec joie de l'organisation d'un cours autour des questions de langue de Yannick Chevalier auprès des étudiant·es du Master EGALITES en Formation Continue dont j'avais la responsabilité, pour enfin assister à ce cours avec son accord.

Et si cette charte n'a pas vocation à convaincre sur l'adoption d'un tel protocole, mais plutôt à préciser celui-ci et ma motivation à le faire, je ne peux qu'inviter à découvrir les six arguments avancés par Yannick Chevalier sur ce point, écrit au début de son mandat de vice-président en charge de l'égalité et de la vie citoyenne à l'Université Lyon 2 [1].

En 2018, lorsque nous co-organisions l'évènement « L'informatique, pourquoi pas moi? » [2] dans le cadre du Webcome Lyon [3], j'étais alors sidérée d'entendre des collègiennes dire qu'elles avaient entendu de la bouche de leur conseillère d'orientation que l'informatique n'était pas pour les filles! Si un des objectifs que je porte est la liberté du choix de l'orientation professionnelle, notamment indépendamment du sexe, je n'ai pu que constater à l'occasion de différents travaux réalisés ou lus, l'importance des représentations dans les choix d'orientation.

Et si les représentations concernant le fait que l'informatique serait plutôt pour les garçons tiennent notamment à des stéréotypes (dont la figure du geek), la langue a aussi sa responsabilité dans la construction des représentations, avec l'absence ou non de l'usage d'une déclinaison au féminin des métiers. « Informaticien » doit pouvoir se décliner au féminin dans une forme visible. Et si l'invisibilisation des informaticiennes dans l'histoire de l'informatique est une réalité, celle des femmes de manière générale par l'usage de la langue française en est aussi une.

Dans sa séance du jeudi 28 février 2019, l'Académie française, institution dont la fonction historique est de normaliser et de perfectionner la langue française, adoptait finalement à une large majorité le rapport sur la féminisation des noms de métiers et de

1. https://www.rue89lyon.fr/2017/03/07/6-arguments-pour-inclure-les-femmes-dans-votre-langage/
2. http://institut-gaston-berger.insa-lyon.fr/content/linformatique-pourquoi-pas-moi-jeunes-ont-pris-parole
3. https://webcomelyon.fr/

fonctions [4]. Une adoption qui est tardive si l'on se réfère aux multiples guides qui ont eu le temps d'émerger autour de ces questions dans différentes sphères. Mentionnons par exemple en 1999 le document « Femme, j'écris ton nom... Guide d'aide à la féminisation des noms de métiers, titres, grades et fonctions », construit au CNRS par l'Institut National de la Langue Française, préfacé de Lionel Jospin, premier ministre à l'époque. La même année, en 1999, un guide en anglais et en français [5], dont la version française a pour titre « Pour l'égalité des sexes dans le langage » produit par l'UNESCO [6] qui précise que c'est la troisième édition, la 1ère intitulée « Pour un langage non sexiste » ayant été proposée en 1987, soit trente ans avant le *buzz* de 2017 en France autour de la controverse sur le langage dit inclusif. Plus récemment, en 2015, l'institution du Haut Conseil à l'Égalité entre les femmes et les hommes a édité un « Guide pour une communication publique sans stéréotype de sexe » [7].

Bien d'autres initiatives autour de la production de guides de ce type ont vu le jour depuis plusieurs années, alors que l'Académie française adoptait donc ce rapport « La féminisation des noms de métiers et de fonctions » en février 2019. Une étape pour l'Académie française qui apparaît néanmoins comme un virage historique quand on a en mémoire qu'en 2017, celle-ci avait adopté à l'unanimité sa position contre l'écriture inclusive, considérant qu'elle était un « péril mortel » pour la langue française et si l'on s'en réfère également à la perspective historique proposée par Viennot *et al.* dans l'ouvrage intitulé « L'Académie contre la langue française. Le dossier féminisation » (Viennot et al., 2016) qui retrace la lutte de l'Académie contre la « féminisation » de la langue française.

Je porte ici l'idée que la langue ne doit pas invisibiliser les femmes au détour d'un usage d'un masculin englobant, générique ou neutre qui serait pensé comme universel.

Si cet enjeu d'un protocole rédactionnel non discriminant a pour but de ne pas laisser transparaître de sexisme, il s'agit ici également de rendre visible dans la langue la place qu'occupent les femmes dans la société, et plus particulièrement dans l'enseignement supérieur et la recherche en informatique, thème abordé dans ce mémoire. En effet, la place des femmes dans la lanque française a été invisibilisée, que ce soit par la disparition de certains noms (de métiers par exemple) ou par la règle de grammaire qui veut que « le masculin l'emporte sur le féminin » car comme le justifiait alors l'abbé Bouhours en 1765 : « lorsque les deux genres se rencontrent, il faut que le plus noble l'emporte » et Nicolas Beauzée en 1767 de compléter : « le genre masculin est réputé plus noble que le féminin à cause de la supériorité du mâle sur la femelle ».

Et si depuis 2019, période où j'avais initié l'écriture de ce protocole pour mon habili-

4. http://www.academie-francaise.fr/sites/academie-francaise.fr/files/rapport_feminisation_noms_\de_metier_et_de_fonction.pdf
5. http://s3.amazonaws.com/inee-assets/resources/UNESCO_Guidelines_Gender_neutral_language_\1999_ENG_FRE.pdf
6. *United Nations Educational, Scientific and Cultural Organization*, institution spécialisée de l'ONU
7. https://www.haut-conseil-egalite.gouv.fr/IMG/pdf/guide_pour_une_communication_publique_sans_stereotype_de_sexe_vf_2016_11_02.compressed.pdf

tation, ces questions de langue ont encore beaucoup fait débat avec différentes controverses qui ont émergé autour notamment de la facilité de lecture et de précédents juridiques également, j'ai fait le choix de conserver ce protocole rédactionnel en l'état de sa version de 2019.

Ainsi, me saisissant de la responsabilité que chacun·e porte dans sa manière de s'exprimer, je présente dans ce qui suit mes choix en matière de rédaction, dans l'optique de préciser le protocole rédactionnel non discriminant adopté dans ce mémoire [8]

— Dans les pages qui suivent, le masculin et le féminin sont utilisés pour désigner les hommes et les femmes respectivement. Le masculin, qui n'est ni un neutre, ni un générique, ne sert donc pas à désigner des collectifs mixtes.
— Lorsque j'évoquerai un groupe mixte, je veillerai à utiliser des graphies appropriées afin de montrer que les personnes dont je parle sont des femmes ET des hommes. J'utiliserai le point médian lorsqu'il n'y a pas de variation syllabique dans le mot (par exemple : les enseignant·es-chercheur·es) et la barre oblique lorsqu'il existe une variation syllabique (par exemple : « évaluateur/trices »), et ce contrairement, par exemple, aux parenthèses qui ont pour effet de hiérarchiser en plaçant le féminin au second plan. Par ailleurs, j'accorderai les adjectifs rattachés à ces noms de la même manière. Les déterminants, les pronoms personnels et les pronoms démonstratifs seront adaptés de même (ex : son/sa, tous/toutes, celui/celle, ...). Je pourrai, par exemple, parler des « maître·sses de conférence rencontré·es » ou de « ceux/celles-ci ». La marque du pluriel sera accrochée à la dernière syllabe.
— J'opterai également pour des doublets pour accentuer une explicitation. Par exemple, j'utiliserai « les lecteurs et les lectrices ».
— Enfin, lorsque cela sera possible et pertinent, j'utiliserai de préférence des termes épicènes c'est-à-dire désignant indifféremment l'un ou l'autre sexe. Nous pourrons, par exemple, parler des « personnes candidatant à des postes dans le milieu académique » ou des « membres du comité ».
— Les noms de métiers, fonctions, grades et titres seront employés au féminin quand cela concerne une personne s'identifiant comme femme.

Une remarque concerne l'usage des termes chercheuse et chercheure. Ayant moi-même alterné l'usage de ces deux termes dans ma manière de me présenter, je m'autorise dans ce document à des variations. Ainsi, je parlerai de chercheuse quand ce terme n'est pas décliné au masculin et, dans le cas contraire, je pourrai avoir recours à chercheur·e, comme par exemple pour « enseignant·es-chercheur·es », mais aussi à des doublets comme « chercheuses et chercheurs ».

Il est à noter que ce procole rédactionnel non discriminant concernera les passages que j'ai moi-même rédigés et non les citations d'ouvrages/d'articles, qui garderont leur forme d'origine. Les citations issues des entretiens seront conservées également dans leur

8. Les erreurs qui pourraient se glisser ici ou là dans ce manuscrit seront à mettre sur le compte d'erreurs d'inattention non détectées dans la relecture ou de la difficulté à sortir d'anciens automatismes d'écriture.

forme d'origine afin de préserver la manière de parler des interviewées.

Introduction générale : saisir l'opportunité d'un contexte qui façonnera une identité scientifique

Introduction générale

On ne naît pas femme, on le devient.
Simone de Beauvoir (1908-1986), philosophe française, « Le deuxième sexe, tome 1 : Les faits et les mythes » (1949)

L'INTRODUCTION générale de ce mémoire vise à resituer la démarche de cette Habilitation à Diriger des Recherches (HDR) grâce à des éléments contextuels, à préciser les contours thématiques et le contexte de réalisation de mes travaux de recherche, ainsi qu'à présenter l'organisation de ce manuscrit.

Eléments contextuels

LA démarche d'aller vers une rédaction et une soutenance pour obtenir l'HDR constitue une étape particulière dans la carrière d'un·e chercheur·e. De mon point de vue, différentes motivations peuvent amener à aller vers l'obtention du plus haut diplôme de l'enseignement supérieur français. Ces motivations variables peuvent recouvrir des dimensions très différentes, notamment liées à des pratiques disciplinaires différentes. Il est à noter par exemple que l'implication dans des co-encadrements avant même de soutenir une HDR est une pratique courante (voire recommandée/nécessaire) en informatique, ce qui ne l'est pas dans d'autres disciplines où l'obtension de l'HDR marque réellement le fait de commencer à (co-)encadrer des thèses.

Cette habilitation constitue pour moi, de manière assez classique, à la fois une étape de bilan des recherches effectuées depuis ma prise de poste en tant que maîtresse de conférences au sein de l'Université Lumière Lyon 2 en septembre 2009, notamment en matière d'encadrement scientifique, et une projection de ce sur quoi portera ma recherche dans les années à venir.

Le démarrage de l'écriture s'est fait en 2019, une dizaine d'années après ma prise de poste. Il s'agissait pour moi de ressentir profondément l'élan de remettre en perspective mes contributions et mon positionnement scientifiques à la lumière de qui j'étais/je suis. Le processus d'écriture fut long, un temps de maturation qui était nécessaire, un cheminement en soi, qui allait au-delà d'un processus assez classique d'évolution de carrière. Il correspond en effet, pour moi, sans aucun doute à une porte à franchir, un rite de passage.

Cette HDR est l'occasion pour moi de manifester sur un plan concret le résultat de mon cheminement scientifique et professionnel. Le terme de « manifeste » est venu à moi un matin au réveil par rapport au sens que je voulais donner à cette HDR, sans forcément avoir en tête précisément le sens de ce mot. Alors, je suis allée à la recherche de ce qu'un dictionnaire avait à m'apprendre sur ce terme !

Manifeste[9] : nom masculin ; italien manifesto, déclaration ; du latin manifestus, un adjectif dont le sens commun rendait compte d'une forme d'évidence[10].

Je retiens ici les deux premières définitions données par le Larousse pour le mot *Manifeste* :

— « Écrit public par lequel un chef d'État, un gouvernement, un parti, etc., rend compte de son mandat ou expose son programme, son point de vue sur un problème politique. »
— « Exposé théorique par lequel des écrivains, des artistes lancent un nouveau mouvement. »

Pour moi, il s'agit de combiner différents éléments de ces deux définitions.

La terminologie de « problème politique » est forte, mais je retiens ici l'idée que dans ce document, il s'agit pour moi d'**exposer mon point de vue**, au travers des contributions scientifiques, en abordant deux aspects présents dans le titre que j'ai choisi pour ce mémoire[11] : le décloisonnement disciplinaire et le genre, ces deux aspects revêtant indéniablement une dimension éminemment **politique**.

Cette dimension politique peut apparaître contraire à la dimension scientifique, j'en ai bien conscience. Mais ce document permettra justement de faire une place à la discussion à ce sujet, afin de mettre en perspective ce que les Sciences Humaines et Sociales m'ont apporté sur cet aspect.

Cela n'empêchera pas que cela soit un **exposé théorique**, dans le sens de la scientificité qu'il revêt, et des contributions scientifiques qui seront présentées. C'est en effet, peut-être avant tout, un document permettant d'accéder à un diplôme reconnu de l'Université, avec une évaluation par les « pairs », qui constitue un **écrit public**.

Ce texte comprend également mon **programme** de recherche, sans doute avec l'envie d'emmener avec moi sur ce chemin des collègues qui souhaiteraient s'inscrire dans

9. https://www.larousse.fr/dictionnaires/francais/manifeste/49163
10. https://www.dicolatin.com/Latin/Lemme/0/manifestus
11. De l'analyse informatique de données de la société à l'analyse sociale de l'informatique. Un cheminement guidé par les études de genre vers un décloisonnement disciplinaire et une posture réflexive.

cette démarche de décloisonnement disciplinaire, vers une forme d'interdisciplinarité (je reviendrai ultérieurement sur ces appelations) avec une perspective de genre, sans forcément avoir la prétention de lancer un **mouvement**. Pour autant, ce texte constitue pour moi l'occasion de manifester trois idées fortes qui seront discutées plus globalement dans le mémoire en dressant le bilan de mes contributions :

1. la recherche en informatique a à gagner en incluant une réelle interdisciplinarité qui peut être amenée par des personnes qui n'ont pas seulement un goût pour l'ouverture disciplinaire, mais qui prennent le temps de la construire ;
2. il est important de pouvoir intégrer une perspective de genre à la fois dans la manière de faire certaines recherches, mais également vis-à-vis de comment se construit la discipline et comment nous y prenons part en tant qu'enseignant·es-chercheur·es du domaine ;
3. la question de la posture, du positionnement, de la réflexivité de chercheur·e est sans doute un impensé pour beaucoup d'entre nous dans le domaine de l'informatique ; je défends l'idée qu'il y a un enjeu fort sur l'épistémologie, en termes de l'analyse de la construction des connaissances en informatique, là où ces questions sont beaucoup plus présentes pour d'autres sciences (y compris en termes de contenu de la formation sur le domaine).

Au-delà des contributions scientifiques et du projet de recherche, cette HDR rend compte de l'histoire de la construction de mon identité scientifique, mais plus encore et de façon plus générale de mon identité professionnelle. Cette identité s'est construite au travers de mon cheminement quotidien, partant de mon statut d'enseignante en informatique rattachée à une Unité de Formation et de Recherche en Anthropologie, Sociologie et Science Politique et de membre d'un laboratoire d'informatique (ERIC) dont la tutelle était initialement l'Université Lumière Lyon 2 (depuis ERIC a également l'Université Lyon 1 Claude Bernard comme tutelle), université tournée vers le domaine des Lettres, des Sciences Humaines et Sociales.

L'université est un lieu de « production » (ou plutôt construction) des savoirs via la recherche et un lieu de circulation de ceux-ci puisque les enseignant·es-chercheur·es sont amené·es transmettre ces savoirs dans leur activité d'enseignement pour former les étudiant·es. Il est alors important pour moi de préciser que la dimension pédagogique de mon métier a transformé profondément la dimension recherche de celui-ci, mettant en lumière le mouvement de la dimension pédagogique vers la dimension recherche, sans doute inverse au mouvement pensé classiquement.

En 2009, Laurence Tain faisait partie du jury de recrutement sur le poste auquel je candidatais. Ce poste était axé plus particulièrement sur la filière MIASHS (Mathématiques Informatique Appliquées aux Sciences Humaines et Sociales) , ou plus précisément MISASHS à l'époque (Mathématiques Informatique Statistiques Appliquées aux Sciences Humaines et Sociales). Très rapidement, elle me proposait d'être intégrée dans la nouvelle équipe pédagogique qu'elle était en train de former dans la perspective de l'ouver-

ture d'un nouveau Master en Études sur le Genre. L'enjeu de monter alors des cours de statistiques descriptives, en développant des capacités professionnalisantes d'usage d'un tableur au prisme du genre, a constitué tout un défi qui m'a amenée à plonger dans le domaine des études de genre. C'est cette perspective pédagogique qui m'a finalement amenée à penser la recherche différemment de ce que j'avais pu le faire avant, pendant ma thèse notamment.

Évoluer dans un environnement empreint de Sciences Humaines et Sociales a confirmé mon goût sur les questionnements autour de la construction de la société, des causes des inégalités, m'a poussée à me familiariser aux notions de stéréotype et de représentation, avec une réflexion certaine autour de la déconstruction au sens de la posture sociologique que j'ai pu explorer. Cela m'a amenée à questionner mes propres représentations.

Cette HDR a finalement nécessité la déconstruction de certaines représentations au fur et à mesure de leur mise en lumière, côtoyer des personnes en Lettres, Sciences Humaines et Sociales a été déterminant pour ce faire. Déconstruire la représentation du modèle type de ce qu'est un·e enseignant·e chercheur·e en informatique, déconstruire aussi l'image de « professeur des universités » reflétée par mon grand-père paternel, intellectuel aux cheveux blancs avec une certaine carrure, déconstruire mes représentations sur la posture de recherche.

Déconstruire n'était pas une fin en soi, pas un choix conscient. C'est le regard porté aujourd'hui qui permet de nommer ce processus.

Déconstruire, oui, mais pour aboutir à quoi ?

Déconstruire pour construire. Construire une identité professionnelle, une identité scientifique, une posture empreinte de réflexivité, une posture pacifiée avec ma discipline d'origine.

Et la construction d'une identité prend du temps même si ce n'était pas un objectif en soi. C'est un processus qui peut être long, qui aura pris son temps, en tous cas en ce qui me concerne, et qui se poursuit.

Cette HDR, c'est également une étape où je suis donc prête à assumer un ressenti d'un décalage plutôt que de le subir silencieusement. Ce décalage ressenti est lié aux éléments saillants suivants qui caractérisent mon positionnement professionnel : cela n'a rien de « classique » d'être Maîtresse de Conférences en informatique dans une UFR d'Anthropologie, Sociologie et Science Politique, avec des responsabilités importantes dans une mention de Master Études sur le Genre. Cela a participé à la construction d'une posture féministe, notamment au sens de promouvoir l'égalité entre les femmes et les hommes. L'afficher, l'assumer au sein de nos environnements de travail est tout sauf anodin. Car il s'agit bien d'environnements au pluriel, et notamment l'environnement en tant que discipline. Car ce sentiment de décalage peut se vivre aussi bien côté informatique que sciences sociales. Est-ce le propre des personnes qui construisent une interdisciplinarité ?

Cela met en exergue une dimension atypique, qui s'exprime aujourd'hui dans l'écri-

ture de ce mémoire. En effet, sans doute que le format de cette HDR pourra être déstabilisant pour les lecteur/trices, en cela aussi qu'il revêt un côté personnel. Et sans doute que l'exercice d'égo-histoire [12] demandé à nos collègues d'histoire dans le cadre de l'HDR est inspirant. Si le format revêt une forme d'originalité, il est à l'image d'une identité revendiquée, pour rendre compte notamment de mes travaux de recherche en informatique, nourrie de la richesse de rencontrer et côtoyer ces collègues et ami·es inspirant·es de disciplines multiples.

Après tout, l'arrêté du 23 novembre 1988 relatif à l'habilitation à diriger des recherches stipule dans son Article 1 : « L'habilitation à diriger des recherches sanctionne la reconnaissance du haut niveau scientifique du candidat, du caractère original de sa démarche dans un domaine de la science, de son aptitude à maîtriser une stratégie de recherche dans un domaine scientifique ou technologique suffisamment large et de sa capacité à encadrer de jeunes chercheurs. ». Ainsi, je vais m'approprier le caractère original de ma démarche qui répond finalement aux attentes de la démarche d'une HDR.

Le choix d'un titre : explicitation

Sans vouloir divulgâcher [13] le contenu de ce document dès son introduction, il apparaît opportun de revenir sur le choix de titre de ce manuscrit, qui correspond à l'identité affichée de ce travail. En effet, si ce travail d'HDR rend bien compte de la construction de mon identité scientifique, le choix des mots mérite une explicitation.

Ce travail d'explicitation des termes, de définition, doit être à mon avis davantage systématisé, dans une dimension de rigueur scientifique. Je mesure à quel point la précision des mots est quelque chose d'important pour moi. Nul doute sur le fait qu'à la lecture, certain·es pourraient penser que je pourrais aller plus rapidemenet à l'essentiel, mais les longueurs verbales permettent précisément de contextualiser, d'amener la nuance.

« De l'analyse informatique de données de la société à l'analyse sociale de l'informatique. Un cheminement guidé par les études de genre vers un décloisonnement disciplinaire et une posture réflexive. »

« analyse informatique » : il s'agit d'un raccourci qui vise à inclure le fait que pour procéder à l'analyse de données, nous ayons procédé à des recherches relevant de la discipline de l'informatique. Il est à noter que cela n'inclut pas uniquement l'analyse des données elle-même, mais regroupe également les enjeux de modélisation/structuration de données qui permettent d'aboutir à une analyse effective des

12. D'après ce qui a été résumé dans Wikipedia https://fr.wikipedia.org/wiki/%C3%89go-histoire : « L'égo-histoire (parfois orthographiée sans accent) définit une forme d'approche historiographique et de courant d'écriture historique à travers laquelle l'historien est censé analyser son propre parcours et ses méthodes de manière réflexive et distanciée »

13. Terme utilisé par nos ami·es québecois·es pour ne pas utiliser l'anglicisme de « spoiler »

données (entrepôts de données / lacs de données). Les thématiques seront abordées dans la section suivante.

« données de la société » : il s'agit ici d'avoir une approche assez large de ce que peuvent recouvrir ces données, à la fois en terme de thématiques (données de la science, données plus ou moins personnelles, etc.), au sens où elles sont produites dans la société dans laquelle nous vivons

« analyse sociale de l'informatique » : ces années d'enseignante-chercheuse m'ont amenée à voir l'informatique au-delà de ma discipline, mais aussi comme pouvant être un objet de recherche, de questionnements, méritant analyse, et plus particulièrement en ayant un regard sur l'informatique et ce qui se produit d'un point de vue social, vis à vis des dynamiques présentes (incluant à la fois la dimension institutionnelle mais également les individus)

« cheminement guidé par les études de genre » : la construction de mon identité scientifique est indéniablement marquée par les études de genre dans lesquelles j'ai plongé dès ma prise de poste de MCF en intégrant l'équipe pédagogique des Masters EGALES et EGALITES en 2011, c'est un véritable cheminement, pas à pas qui s'est opéré

« décloisonnement disciplinaire » : ce terme fait ici référence à cette capacité à aller vers d'autres disciplines pour s'imprégner des méthodes, des résultats de ces autres disciplines, avec l'idée sous-jacente que la discipline d'origine puisse en bénéficier (cette question du bénéfice pourrait faire l'objet d'une discussion en soi, mais cette HDR constitue elle-même peut-être la démonstration de ce bénéfice)

« posture réflexive » : sans entrer ici dans l'immédiat sur une définition formelle, il s'agit de rendre compte pour moi d'une posture de recherche qui amène à des questionnements et réflexions, notamment du point de vue méthodologique.

Ce titre vise à rendre compte de ce processus du cheminement que j'ai réalisé, au-delà même des contributions au sens premier du terme, un processus qui se poursuivra au-delà de ce manuscrit, puisqu'il s'agit en soi d'une manière de faire de la recherche. La dernière partie de ce mémoire reviendra sur ce titre et discutera de différents aspects qui méritent qu'on s'y arrête un peu plus que sur un paragraphe.

Contrairement aux thèses de doctorat, en parcourant quelques mémoires d'HDR, il apparaît que sur la forme il n'existe pas de réel standard. Cette absence de « norme » constitue selon moi une opportunité, voire une autorisation, d'être dans une forme de créativité. Ce mémoire contient bon nombre de réflexions issues de cette prise de recul par rapport à presque quinze années dans cette fonction de maîtresse de conférences. Pour autant ce mémoire reprend classiquement (mais succinctement) les apports scientifiques en informatique auxquels j'ai contribué, qui concernent différentes thématiques que je décris à présent.

Thématiques abordées

Ce mémoire retrace notamment les différentes contributions scientifiques dans le domaine de l'informatique auxquelles j'ai participé, qui se trouvent s'ancrer dans ce que je pourrais qualifier de deux thématiques de l'informatique, bien que leurs frontières soient poreuses. Pour les caractériser, je me réfère ici à la *Nomenclature thématique*[14] utilisée dans la section Informatique (section 27) de l'instance nationale du Conseil National des Universités (CNU), encore en cours en 2023. Cette nomenclature permet de rendre compte de la diversité des thématiques qu'englobe la section informatique et qui sont organisées en deux niveaux de granularité. Ainsi, lors des dépôts de dossiers divers qui sont amenés à être pris en charge par la section, il est ainsi attendu de pouvoir spécifier où nous nous situons, ce qui permet à la fois une meilleure gestion de l'évaluation par les pairs, mais aussi la réalisation de potentiels suivis quantitatifs.

Pour ma part, il s'agit d'une part de la thématique des *Systèmes d'Information* qui recouvre les aspects liés à l'entreposage de données et aux lacs de données, et la thématique de l'*Intelligence Artificielle* qui recouvre notamment l'apprentissage machine et la science des données [15].

Ces deux thématiques s'inscrivent dans la continuité de mon double cursus académique en cinquième année d'études supérieures qui comprenait à la fois l'obtention d'un DESS en informatique décisionnelle notamment en lien avec les entrepôts de données et d'un DEA en fouille de données.

J'évoque ici quelques éléments sur les thématiques couvertes par nos contributions, sans pour autant en donner un état de l'art contextualisant. Cette contextualisation sera plutôt évoquée succinctement en discussion, après chaque présentation de contribution. Il est à noter que ces thématiques ne sont pas à percevoir comme disjointes puisque les contributions présentées s'articulent parfois sur deux d'entre elles simultanément.

Entrepôts et analyse *OLAP*

« Les entrepôts de données sont nés au sein des entreprises pour répondre à des besoins d'analyse pour l'aide à la décision. Un entrepôt de données peut être vu comme une grosse base de données modélisée pour accueillir, après nettoyage et homogénéisation, les informations en provenance de différents systèmes de production de l'entreprise. L'un des points-clés de la réussite de l'entreposage réside alors dans la conception du schéma de l'entrepôt qui doit permettre de répondre aux besoins d'analyse pour l'aide à la décision. » Ce sont les mots que j'écrivais dans ma thèse en 2007 pour ancrer ma problématique sur l'enjeu d'intégration de la connaissance des utilisateur/trices dans l'entrepôt de

14. https://cnu27.univ-lille.fr/nomenclature.html

15. Il est à noter que certaines thématiques détaillées se retrouvent rattachées à deux thématiques principales comme c'est le cas de la science des données, ce qui pourrait faire l'objet d'une discussion sur la catégorisation à part entière dans laquelle je n'entrerai pas ici)

données pour la personnalisation des analyses.

L'enjeu de la place des utilisateur/trices dans le processus d'analyse dans un entrepôt de données (*data warehouse*) au travers de ce qui est appelé l'*OLAP* pour *OnLine Analytical Processing* a été un point clé du succès de cette technologie. L'*OLAP* a été conçu pour permettre une navigation dans les données, par exemple en permettant des passages entre différents niveaux de granularité/détail. Travailler sur une forme de personnalisation des analyses nécessitant d'apporter une forme de flexibilité dans le schéma de données initial qui conditionne ces analyses possibles a été un des premiers travaux d'après-thèse, inspiré des résultats de celle-ci mais avec une toute autre dimension. Cette contribution a été initialement pensée dans le cadre de données médicales mais a été conçue de façon générique. Ainsi, l'application à des données issues de médias sociaux serait tout à fait possible.

En effet, les données de la société dont il est question dans ce mémoire proviennent également de ce qui est produit par la société elle-même, notamment au travers des applications de médias sociaux et leur expansion massive. Une des sources privilégiées qui a été considérée dans nos travaux est Twitter, le réseau social de microblogage, dont une des particularités qui constituait un avantage était l'accès gratuit à des données, même si les données distribuées répondaient à certains critères vis-à-vis du plafonnement de l'API (Application Programming Interface) de Twitter.

Face à l'explosion de la masse des données issues des médias sociaux tels que Twitter, permettre l'analyse de celles-ci via une approche d'analyse en ligne constitue un réel enjeu. Les défis liés à la modélisation des entrepôts de données concernent alors la prise en compte de données des médias sociaux. C'est ainsi que, dans ce contexte, nous avons proposé des pistes pour ce type d'analyse.

Une des caractéristiques de ces médias sociaux est précisément la dimension de réseau. Ainsi, dans l'évolution des approches d'entreposage des données, nous avons considéré l'enjeu de combiner l'aspect réseau représenté sous forme de graphe et la dynamique d'analyse en ligne dans une perspective de *Graph OLAP*, qui passe notamment par des enjeux de modélisation. Ceci a été développé sur des données bibliographiques, qui constituent une partie des données de la recherche, qui sont à la base de la scientométrie qui vise à la mesure et l'analyse de la science.

Ainsi, le travail réalisé dans la thématique des entrepôts de données a permis de se pencher sur différents types de données, en particulier sur des enjeux de modélisation, pour permettre des analyses qui prennent en compte les besoins d'analyse.

Lacs de données

Cette multiplicité des types de données est aussi à considérer dans leur dimension de complexité, et donc de leur variété.

C'est ainsi que la thématique des lacs de données (*data lake*) a émergé plus récem-

ment, avec l'idée d'un stockage sans pré-traitement et sans connaître précisément l'usage futur qui pourrait en être fait. Cela diffère du processus d'entreposage de données évoqué plus tôt, qui nécessite un important travail de modélisation et d'alimentation en vue des analyses qui sont envisagées en amont.

Pour ce faire, l'enjeu de reporter la modélisation des données vers les métadonnées est crucial. C'est dans cette perspective que nous avons travaillé à la modélisation et la métamodélisation des métadonnées dans un contexte de lac de données.

Ce travail, qui a un caractère générique, a porté sur des données en lien avec les habitats sociaux. Si les contributions se sont focalisées sur les aspects de modélisation des métadonnées pour permettre un stockage adéquat, il est entendu que les données ont ensuite vocation à être interrogées, analysées, notamment grâce à des processus de fouille de données, même si ce n'était pas la portée de ce travail directement.

Fouille de données

Les travaux menés en fouille de données ont quant à eux plutôt porté sur l'analyse de médias sociaux que nous avons évoqués précédemment.

Sur Twitter, il s'agissait de permettre des traitements automatiques, non réalisables manuellement du fait du volume de données, en permettant l'analyse de ce qui est dit sur Twitter, notamment avec la détection des évènements qui ont fait le *buzz*, leur diffusion, ainsi que les personnes qui influencent cette diffusion. La détection des rumeurs avec le développement d'approches multimodales fut également un point important qui a un fort impact sociétal, dans un contexte de profusion de fausses informations.

Ainsi l'apprentissage de données a porté à la fois sur des données « classiques », en réseaux, et des images.

Scientométrie

Le terme de scientométrie désigne l'étude quantitative des sciences et de l'innovation. L'analyse bibliométrique se focalise sur les publications, et peut être considérée comme une sous-partie de la scientométrie d'un certain point de vue.

C'est en considérant les publications comme données de travail pour l'approche mixant graphe et analyse *OLAP* que s'est faite l'entrée dans l'analyse des données de la science.

Cette plongée en scientométrie s'est poursuivie en se focalisant sur les enjeux de la place des femmes en informatique et les enjeux écologiques de mobilité liés aux déplacements des chercheuses et chercheurs pour les conférences.

La scientométrie est largement articulée avec la sociologie des sciences. Et la proximité pédagogique avec des sociologues fut l'occasion d'aller un peu plus loin dans cette

considération de la discipline comme objet de recherche, en allant vers les méthodes qualitatives chères à la sociologie.

Vers un travail sociologique sur l'informatique

Au-delà de ces thématiques de l'informatique, l'imprégnation des Sciences Humaines et Sociales m'a amenée à une prise de recul et, après avoir travaillé sur les données bibliographiques, à m'intéresser aux personnes qui produisent des connaissances en informatique. Avec l'analyse de l'informatique dans une perspective de genre, je me suis alors penchée sur la question des carrières dans l'enseignement supérieur et la recherche en informatique, sous l'angle du questionnement des politiques de quotas dans les comités de sélection. Aller vers une méthodologie issue de la sociologie permet d'aller vers une interdisciplinarité qui m'apparaît être riche en terme de réflexions. Ceci est en accord avec mon cursus initial post-baccalauréat, où j'avais entrepris un DEUG MASS (Mathématiques Appliquées et Sciences Sociales), compte-tenu de mon appétence pour les Sciences Humaines et Sociales « malgré » un cursus au lycée en filière scientifique.

L'ensemble des contributions décrites dans le présent document se base sur des travaux à la fois collectifs et d'encadrement de thèses, dans un environnement particulier. Le contexte de travail de développement de ces travaux est ainsi décrit dans ce qui suit.

Environnement(s) de travail

Depuis mon recrutement en 2009, j'ai évolué dans un environnement que je qualifierais d'hétéroclite, je pourrais même plutôt parler d'environnements de travail divers, qui ont nécessité le développement de nombreuses capacités. Cette habilitation rend compte des travaux réalisés au travers d'un cheminement dans ce métier en étant rattachée, comme indiqué précédemment, à une Unité de Formation et de Recherche en Anthropologie, Sociologie et Science Politique, en tant qu'informaticienne, enseignante et chercheuse, ayant des responsabilités pédagogiques, et en tant que femme dans une discipline majoritairement occupée par des hommes.

Ce positionnement dans l'UFR, où nous sommes deux titulaires en informatique, a été l'occasion de côtoyer quotidiennement des sociologues, des anthropologues, des politistes.

J'ai eu des co-responsabilités ou responsabilités pédagogiques dès ma prise de poste. Tout d'abord dans le DESS IIDEE en informatique décisionnelle (cursus que j'avais suivi en tant qu'étudiante) comme co-responsable, découvrant ainsi le public de formation continue.

Puis, j'ai eu des responsabilités croissantes dans les Masters en Etudes sur le genre de l'Université Lyon 2, jusqu'à porter la construction d'un parcours pour la formation conti-

nue, pour des étudiant·es en reprise d'études aux parcours professionnels si diversifiés et parfois incroyables. Intégrer cette équipe pédagogique fut une ouverture sur des disciplines multiples abordant le genre au-delà même de l'anthropologie, de la sociologie et de la science politique, avec l'histoire, la littérature, les sciences de l'information et de la communication, etc.

Ce fut aussi une ouverture sur la découverte d'une autre manière de faire de la recherche. La participation chaque année à une dizaine de jurys de soutenance de mémoires de master en études de genre a construit une partie de ma propre formation.

J'ai également été très impliquée (et le suis toujours) dans la filière MIASHS, qui, à Lyon 2, correspond à une formation où les Mathématiques et l'Informatique sont associées aujourd'hui à 6 Sciences Humaines et Sociales possibles (géographie, histoire, psychologie, sciences cognitives, sciences du langage, sociologie). Cela m'a notamment amenée à côtoyer dans le cadre d'encadrement de mémoires d'initiation à la recherche bidisciplinaires des collègues de ces différentes disciplines.

Par ailleurs, je ne peux passer à côté de la description de la dimension internationale de mon environnement de travail. C'était une belle invitation à me « décentrer » grâce au réseau international du master en études de genre EGALES, dont j'ai pris la responsabilité depuis plusieurs années. Ce fut l'opportunité de discussions marquantes et enrichissantes avec des collègues européennes représentantes de pays comme la Roumanie, l'Espagne, l'Italie, la Suisse, ...Mais aussi grâce au Québec (Canada), où mes séjours à l'Institut de Recherches et d'Etudes Féministes (IREF) ont été très structurants dans la construction de mon identité.

Ces environnements ont alimenté diverses réflexions personnelles dont je rendrai compte dans ce document, sans omettre que la base de ce manuscrit correspond aux différents travaux que j'ai développés en informatique depuis bientôt une quinzaine d'années et qui ont été réalisés au travers de collaborations multiples, donnant lieu au développement d'un encadrement scientifique, qui constitue en soi une réelle aventure humaine.

La figure 1 illustre justement le déroulement temporel, depuis mon recrutement en 2009, de mon activité liée à l'encadrement scientifique (en informatique), puisque c'est ce qui est au cœur des compétences liées à cette habilitation à diriger des recherches. Pour autant, il est à noter que ce mémoire ne se limitera pas aux travaux encadrés et comprendra également des contributions développées dans le cadre de collaborations, que j'ai initiées ou non, car il m'apparaît important de ne pas perdre de vue la dimension de chercheuse au-delà de l'encadrement.

Cette activité d'encadrement se décline ainsi en trois types :

- **Collaboration doctorale :** en 2009, j'ai été sollicitée pour une collaboration par Yoann Pitarch qui réalisait sa thèse à Montpellier [16], sous l'encadrement d'Anne Laurent et

16. Yoann Pitarch. Résumé de Flots de Données : motifs, Cubes et Hiérarchies. Thèse en informatique

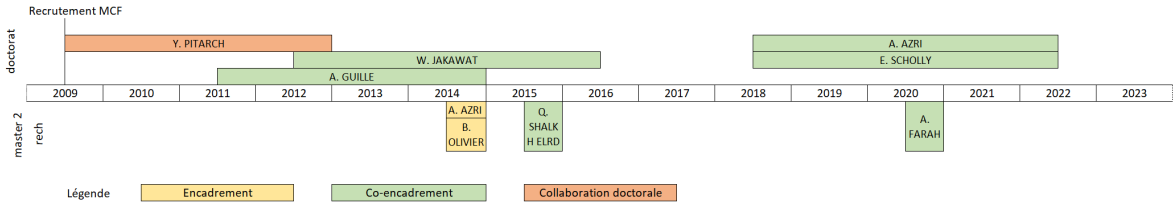


Figure 1 – Synthèse de mon activité liée à l'encadrement scientifique en informatique.

de Pascal Poncelet.

- **Co-encadrements :** cela correspond à la fois à des co-encadrements de master et de thèse de doctorat.
  - — Deux étudiants en Master : Qutiba Shalkh Elrd [17], co-encadré avec Sabine Loudcher et Abderrahmane Farah [18], co-encadré avec Abderrazek Azri.
  - — Quatre étudiant·es en thèse de doctorat : les quatre thèses ont été soutenues. Les deux premières thèses soutenues ont été réalisées par Adrien Guille [19] d'une part, co-dirigé avec Djamel Zighed et dont la soutenance s'est déroulée en novembre 2014 ; et par Wararat Jakawat [20], co-dirigée avec Sabine Loudcher et dont la soutenance s'est déroulée en septembre 2016. Les deux dernières thèses soutenues sont les travaux développés par Etienne Scholly [21] d'une part, thèse co-encadrée avec Sabine Loudcher (et Eric Ferey du côté de l'entreprise Bial-X puisqu'il s'agit d'une thèse en convention CIFRE) et soutenue en mai 2022 et Abderrazek Azri [22] d'autre part, thèse co-encadrée avec Nouria Harbi et Jérôme Darmont et soutenue en juillet 2022.
- **Encadrements :** j'ai eu l'occasion d'encadrer deux étudiants de Master : Brice Olivier [23] et Abderrazek Azri [24]. Abderrazek Azri a ensuite poursuivi un travail de thèse que j'ai co-encadré, comme indiqué plus haut.

soutenue le 10 mai 2011

17. Qutiba Shalkh Elrd. OLAP Analysis on tweets and blogs. Mémoire de Master en informatique soutenu en juillet 2015

18. Abderrahmane Farah. Constitution d'un jeu de données test pour la conception d'un outil de veille pour la détection de rumeurs dans les réseaux sociaux. Mémoire de Master en informatique soutenu en novembre 2020

19. Adrien Guille. Diffusion de l'information dans les médias sociaux : modélisation et analyse. Thèse en informatique soutenue le 25 novembre 2014

20. Wararat Jakawat. Graphs enriched by Cubes (GreC) : a new approach for OLAP on information networks. Thèse en informatique soutenue le 27 septembre 2016

21. Etienne Scholly. De la modélisation des métadonnées à la conception d'un lac de données : Application à l'habitat social. Thèse en informatique soutenue le 23 mai 2022

22. Abderrazek Azri. Approches multimodales d'apprentissage automatique pour la détection des rumeurs dans les microblogs. Thèse en informatique soutenue le 7 juillet 2022

23. Brice Olivier. Entrepôts de données et OLAP pour des données textuelles. Mémoire de Master en informatique soutenu en septembre 2014

24. Abderrazek Azri. Vers une navigation en mode OLAP dans les images satellitaires et leurs segmentations. Mémoire de Master soutenu en septembre 2014

Les travaux réalisés seront présentés selon les choix opérés avec un niveau de détail plus ou moins. L'organisation du mémoire est précisée ci-après.

Organisation du mémoire

Les deux premières parties présentent une synthèse de nos contributions. Une troisième partie empreinte de réflexivité est proposée, avant la présentation d'une conclusion générale.

Globalement, les deux premières parties sont construites avec un chapitre d'introduction, trois chapitres présentant une synthèse des contributions, et un chapitre de conclusion.

Cette synthèse aura pour objectifs de préciser 1) les données concernées et les enjeux, 2) les contributions, 3) une discussion qui permettra à la fois de resituer celles-ci par rapport aux travaux existants du moment de façon synthétique et d'aborder le cas échéant quelques remarques réflexives sur le travail, avec le regard que je porte aujourd'hui sur le travail.

La lecture de cette présentation succincte peut être complétée par celle des contributions qui sous-tendent le chapitre si davantage de détails sont souhaités. Chaque chapitre débutera donc par la liste des contributions (publications et/ou mémoires) sur lesquelles porte le chapitre.

La première partie (partie I) vise à décrire différents travaux permettant de montrer comment l'informatique rend possible une démarche d'analyse de données de la société. Cette partie présente des travaux qui s'appuient sur une démarche ancrée en informatique décisionnelle et/ou sur une démarche ancrée en fouille de données, correspondant à mes deux axes de formation initiale de Bac+5 comme précisé précédemment.

Cette première partie comprend donc trois axes de contributions principales se déclinant en trois chapitres.

1. Chapitre I.2 : la modélisation est au cœur des possibilités d'analyse de données et sera au cœur de ce chapitre. Deux axes seront développés : 1) la modélisation dans les entrepôts de données pour permettre des analyses personnalisées, en développant le concept de satellite (ce travail a été réalisé dans le cadre de la collaboration doctorale avec Yoann Pitarch) et 2) la modélisation des métadonnées nécessaire dans le contexte des lacs de données (ce travail a été réalisé dans le cadre de la thèse d'Etienne Scholly que j'ai co-encadrée avec Sabine Loudcher ainsi que dans le cadre de la petite équipe de « datalakers » au sein du laboratoire).
2. Chapitre I.3 : l'analyse de la diffusion de l'information sur Twitter grâce à l'apprentissage de données avec d'une part la détection d'évènements, leur diffusion et les personnes influençant cette diffusion (ce travail a été l'aboutissement de la thèse d'Adrien Guille que j'ai co-encadré avec Djamel Zighed) et d'autre part la détection

de rumeurs à partir d'approches multimodales s'appuyant notamment sur l'analyse d'images (proposées dans la thèse d'Abderrazek Azri que j'ai co-encadrée avec Jérôme Darmont et Nouria Harbi).

3. Chapitre I.4 : une approche de navigation pour des données organisées en graphe enrichie par des cubes de données « façon *OLAP* » (ce travail a été réalisé dans le cadre de la thèse de Wararat Jakawat que j'ai co-encadrée avec Sabine Loudcher).

Après avoir montré au travers de ces contributions comment l'informatique peut contribuer à l'analyse de données de la société, la deuxième partie permet de montrer le passage vers une analyse sociale de l'informatique.

Ainsi, la deuxième partie (partie II) présente donc des travaux qui s'appuient à la fois sur des contributions en informatique dans le domaine de la scientométrie, dans des perspectives écologique et de genre par rapport aux personnes produisant des contributions en informatique, au travers de l'analyse d'une communauté d'enseignant·es-chercheur·es. Il s'agit également d'aller vers une démarche sociologique qui a concerné une politique d'égalité qui est la mise en place de quotas dans les comités de sélection pour les recrutements d'enseignant·es-chercheur·es. Ainsi, cette partie amène vers une analyse plus sociale de l'informatique en tant que discipline de recherche.

Cette deuxième partie comprend trois contributions principales qui s'organisent en trois chapitres.

1. Chapitre II.2 : une analyse chiffrée selon l'axe de l'écologie des données bibliographiques, fruit d'un travail collaboratif initié avec Sébastien Valat sur les données bibliographiques de la conférence Extraction et Gestion des Connaissances (EGC), qui permet également d'apporter une discussion sur la mobilité des chercheuses et chercheurs en conférences à la lumière du vécu de la pandémie et du passage en distanciel du déroulement des conférences.
2. Chapitre II.3 : une analyse chiffrée avec une perspective de genre, illustrée sur les données bibliographiques d'EGC, qui débouche sur une analyse plus globale de la communauté des enseignant·es-chercheur·es en informatique ; ce travail est issu en partie d'un travail de collaboration collective autour des données d'EGC, où j'avais pris en charge la partie sur l'analyse sexuée des données.
3. Chapitre II.4 : une contribution dans une démarche sociologique, en chaussant les lunettes du genre autour de la question des carrières, et plus précisément le questionnement de la mise en place d'une politique d'égalité telle que les quotas dans les comités de sélection ; ce travail a été mené en étant guidée par Laurence Tain.

Après deux parties présentant des contributions, la dernière partie (partie III) comprendra trois volets en amenant au cœur d'une posture de recherche réflexive.

1. Chapitre III.2 : ce chapitre présente un bilan réflexif comprenant des points transversaux sur les contributions proposées, en abordant trois aspects : 1) le positionnement atypique qui vient nourrir ma recherche, 2) une discussion autour de la construction

des catégories d'analyse et de leur visibilisation, et 3) des éléments autour de la non neutralité des analyses et la responsabilité qui en découle.

2. Chapitre III.3 : dans ce chapitre, j'aborde des réflexions que je qualifie d'épistémologiques (je reviendrai sur ce terme le moment venu) par rapport à la recherche en informatique, en organisant mon propos en trois points : 1) l'enjeu de situer les savoirs, 2) le décloisonnement disciplinaire et ses déclinaisons, 3) un focus sur « Science, sens, société et engagement ».
3. Chapitre III.4 : les deux premiers chapitres de cette partie me permettent d'aborder un chapitre de perspectives nourri de cette posture de recherche réflexive, qui se déclinera en fonction de différentes facettes de mon « identité ». En effet, les perspectives seront articulées comme suit : 1) une recherche en tant que chercheuse en informatique, 2) une recherche en tant que chercheuse en études sur le genre, 3) une recherche pluri/interdisciplinaire. Un point sur la mise en œuvre de ce programme sera fait.

Une conclusion générale viendra clore ce mémoire. Elle comprendra à la fois un bilan sur les contributions et les perspectives présentées, mais aussi un bilan plus personnel.

Première partie

L'informatique : une science pour l'analyse des données de la société

1 Introduction

> La science amène les gens à atteindre de manière désintéressée la vérité et l'objectivité ; elle apprend aux gens à accepter la réalité, avec émerveillement et admiration, sans parler de la profonde crainte et de la joie que l'ordre naturel des choses apporte au vrai scientifique .
>
> *Lise Meitner (1878-1968), physicienne et chimiste autrichienne naturalisée suédoise*

LA CITATION venant illustrer ce chapitre introductif retrace sans doute la représentation que j'ai pu avoir initialement de la science. Cette brillante scientifique qu'est Lise Meitner est renommée pour ses travaux sur la radioactivité et la physique nucléaire, en particulier par rapport à son rôle dans la découverte de la fission nucléaire. Il est à noter que Lise Meitner est souvent mentionnée comme l'un des cas les plus flagrants de victime de l'effet Matilda [1] injustement ignorée par le comité attribuant le prix Nobel.

Lise Meitner rend compte au travers de ses mots de ses champs disciplinaires que sont la physique et la chimie, et de son propre rapport à ceux-ci, voire même de sa posture de recherche (« manière désintéressée », « la vérité et l'objectivité » par exemple).

Cela constitue pour moi une manière d'introduire un rapport initial aux sciences, avant que les Sciences Humaines et Sociales m'amènent à développer un positionnement construit qui questionne certains mots utilisés ici selon la discipline scientifique dans laquelle on s'inscrit, j'y reviendrai dans la troisième partie du manuscrit.

Cette « quête » de vérité et objectivité peut guider le besoin d'analyse de données présent aujourd'hui dans notre société, avec des enjeux plus ou moins importants.

1. Définition du Larousse de l'effet Matilda : phénomène consistant à minimiser, voire à nier, la contribution des femmes à la recherche scientifique, au profit d'une postérité essentiellement masculine. (Le phénomène a été théorisé sous ce nom dans les années 1980 par l'historienne des sciences américaine Margaret W. Rossiter [née en 1944], en référence à la militante féministe américaine Matilda Joslyn Gage [1826-1898], qui, la première, remarqua l'invisibilisation des femmes dans l'histoire des sciences.)

À l'ère des « données ouvertes », et de la multiplication des données dans des contextes plus fermés (comme par exemple dans des entreprises, des associations, etc.), l'enjeu majeur réside précisément dans la capacité à analyser ces données.

Si la chaîne de traitement pour l'analyse de données a connu diverses variantes imagées selon l'angle adopté, je choisis ici de ne pas nécessairement m'y référer pour mettre en lien des travaux qui relèvent de contextes différents, même s'il est vrai que l'on peut toujours percevoir les entrepôts de données ou les lacs de données comme des étapes faisant partie de cette chaîne de traitement avant analyse.

Les chapitres de contributions présentés dans cette partie ont porté sur des données s'inscrivant dans des thématiques différentes (santé et habitat social, données des réseaux sociaux, données bibliographiques), avec des enjeux spécifiques dans leur prise en compte. Et nous nous intéresserons ici à 3 volets pouvant être effectivement combinés dans une chaîne de traitement mais qui ont été abordés séparément dans les travaux pour lesquels j'ai participé à l'encadrement.

Il sera d'abord question de traiter des enjeux de modélisation, laquelle détermine inévitablement la manière dont peuvent être analysées par la suite les données.

Ainsi, nous verrons dans le chapitre I.2 deux axes de contributions en lien avec les enjeux de modélisation.

Le premier axe répondra au questionnement suivant : comment, dans les entrepôts de données, la modélisation des hiérarchies de dimension peut gagner en flexibilité d'analyse en remettant l'utilisateur/trice au cœur du processus d'analyse ? Cette contribution s'appuie sur un contexte de données médicales.

Le second axe s'inscrit dans l'émergence des lacs de données qui visent à stocker les données dans leur format d'origine, évitant ainsi le travail de prétraitement nécessaire à la construction d'un entrepôt de données, et ouvrant la possibilité à des données « moins » structurées d'être considérées. Il s'agit alors, dans ce cadre, de montrer à quel point la modélisation constitue un enjeu crucial pour des possibles interrogations de données. Ainsi nous aborderons la modélisation des métadonnées pour les lacs de données. Ce travail s'inscrit dans le contexte de l'habitat social.

Par la suite, dans le chapitre I.3, nous aborderons les enjeux de l'analyse des médias sociaux, et plus particulièrement dans le cadre de Twitter. Il s'agira de présenter les travaux sur l'analyse de la diffusion de l'information grâce à l'apprentissage de données. Ceci inclut la détection d'évènements et leur diffusion, l'analyse des personnes influençant cette diffusion, ainsi que la détection de rumeurs, en s'appuyant dans ce dernier axe particulièrement sur des approches multimodales intégrant les images.

Puis, le chapitre I.4 présentera notre vision et approche de *Graph OLAP* s'appuyant sur les données bibliographiques (données de la science). Nous nous attacherons à montrer comment les structures de graphes combinées à une approche *OLAP* permettent d'accéder à une nouvelle manière d'observer les données.

Dans cette partie, chaque contribution sera articulée autour de trois points de présentation : les données considérées en explicitant les enjeux, la contribution elle-même et enfin une discussion exprimant d'une part le positionnement succinct de la contribution dans le paysage des travaux existants du moment et d'autre part une discussion plus large. Des réflexions conclusives viendront compléter chaque chapitre.

Enfin, une conclusion viendra clore cette partie dans le chapitre I.5 pour faire une synthèse des contributions et apporter quelques éléments de discussion transversaux.

2 La modélisation au service de l'analyse de données

Ne laissez personne vous voler votre imagination, votre créativité ou votre curiosité. C'est ta place dans le monde, c'est ta vie. Continuez et faites tout ce que vous pouvez avec elle, et faites-en la vie que vous voulez vivre.

Citation attribuée à Mae Carol Jemison (née en 1956), première femme afro-américaine astronaute dans l'espace... notamment!

Contributions sur lesquelles se base ce chapitre

Axe modélisation pour la personnalisation des analyses dans les entrepôts de données

- > Publications dans des conférences internationales
  - Y. Pitarch, **C. Favre**, A. Laurent & P. Poncelet, Enhancing Flexibility and Expressivity of Contextual Hierarchies, FUZZ-IEEE 2012, Brisbane, Australia, 2012, 809-816. (Pitarch et al., 2012b)
  - Y. Pitarch, **C. Favre**, A. Laurent & P. Poncelet, Context-aware generalization for cube measures, DOLAP 2010, Toronto, Ontario, Canada, 2010, 99-104. (Pitarch et al., 2010b)
- > Article dans une revue nationale
  - Y. Pitarch, **C. Favre**, A. Laurent & P. Poncelet, Généralisation contextuelle de mesures dans les entrepôts de données. Application aux entrepôts de données médicales, Revue ISI, Vol.16, N°6, 2011, 67-90. (Pitarch et al., 2011)
- > Publications dans des conférences nationales
  - **C. Favre**, A. Laurent, Y. Pitarch & P. Poncelet, Représentation graphique des hiérarchies contextuelles : modèle avec satellites, EDA 2011, Clermont-Ferrand, 23-37. (Favre et al., 2011)
  - Y. Pitarch, **C. Favre**, A. Laurent & P. Poncelet, Analyse flexible dans les entrepôts de données : quand les contextes s'en mêlent, EDA 2010, Djerba, Tunisie, 60-74. (Pitarch et al., 2010a)
- > Publication dans un atelier national
  - Y. Pitarch, **C. Favre**, A. Laurent, P. Poncelet, Vers davantage de flexibilité et d'expressivité dans les hiérarchies contextuelles des entrepôts de données, Atelier FDC 2012 @EGC2012, Bordeaux, 55-66. (Pitarch et al., 2012a)

Axe modélisation des métadonnées pour les lacs de données

- > Mémoire de thèse d'Etienne Scholly
  - E. Scholly, De la modélisation des métadonnées à la conception d'un lac de données : Application à l'habitat social, Université Lumière Lyon 2. Thèse en informatique soutenue le 23 mai 2022. (Scholly, 2022)
- > Publications dans des conférences internationales
  - E. Scholly, P. N. Sawadogo, P. Liu, J. A. Espinosa Oviedo, **C. Favre**, S. Loudcher, J. Darmont & C. Noûs, Coining goldMEDAL : A New Contribution to Data Lake Generic Metadata Modeling, DOLAP 2021, Nicosia, Cyprus, 2021, 31-40. (Scholly, Sawadogo, Liu, Espinosa-Oviedo et al., 2021)
  - E. Scholly, **C. Favre**, E. Ferey, S. Loudcher, HOUDAL : A Data Lake Implemented for Public Housing, ICEIS 2021, Online Streaming, 2021, 39-50. (Scholly, Favre et al., 2021)
  - J. Darmont, **C. Favre**, S. Loudcher & C. Noûs, Data Lakes for Digital Humanities, DTUC 2020, Hammamet, Tunisia - Online Streaming, 38-41. (Darmont et al., 2020)
- > Publications dans des conférences nationales
  - E. Scholly, P. N. Sawadogo, P. Liu, J. A. Espinosa Oviedo, **C. Favre** and S. Loudcher, J. Darmont & C. Noûs, goldMEDAL : une nouvelle contribution à la modélisation générique des métadonnées des lacs de données, EDA 2021, Toulouse en distanciel, 55-58. (Scholly, Sawadogo, Liu, Espinosa-Oviedo et al., 2021a)
  - E. Scholly, P. N. Sawadogo, P. Liu, J. A. Espinosa Oviedo, **C. Favre** and S. Loudcher, J. Darmont & C. Noûs, goldMEDAL : A Data Lake Generic Metadata Model (résumé), BDA 2021, Paris en distanciel, 19-20. (Scholly, Sawadogo, Liu, Espinosa-Oviedo et al., 2021b)
  - E. Scholly, P. N. Sawadogo, **C. Favre**, E. Ferey, S. Loudcher & J. Darmont, Systèmes de métadonnées dans les lacs de données : modélisation et fonctionnalités, EDA 2019, Montpellier, 77-92. (Scholly et al., 2019)
- > Publication dans un atelier international
  - P. N. Sawadogo, E. Scholly, **C. Favre**, E. Ferey, S. Loudcher & J. Darmont, Metadata Systems for Data Lakes : Models and Features, Workshop BBIGAP 2019 @ADBIS2019, Bled, Slovenia, 440-451. (Sawadogo et al., 2019)

LA MODÉLISATION de données est une étape fondamentale qui conditionne les analyses possibles de ces données. Pour aller au-delà de ce que permettent les bases de données « classiques », on a vu émerger dans les années 90 le *data warehouse* (entrepôt de données), et dans les années 2010, le *data lake* (lac de données).

Ce chapitre retrace deux axes de contributions qui ont pour point commun d'amener de la flexibilité pour permettre des analyses répondant mieux aux besoins des utilisateur/trices.

Ainsi, le chapitre est organisé de la manière suivante. La section I.2.1 permet d'introduire les deux domaines considérés dans ce chapitre que sont les entrepôts de données et les lacs de données. Puis, nous présenterons dans la section I.2.2 notre contribution à la modélisation de hiérarchies de dimension flexibles grâce au concept de satellite dans les entrepôts de données. Ensuite, nous aborderons notre apport à la modélisation de métadonnées dans les lacs de données avec *MEDAL* et *goldMEDAL* dans la section I.2.3. Nous terminerons ce chapitre par quelques réflexions conclusives dans la section I.2.4.

2.1 Préambule

Les entrepôts de données (Inmon, 1996) permettent de consolider, stocker et organiser des données à des fins d'analyse. Des faits peuvent alors être analysés à travers des indicateurs (les mesures) selon différents axes d'analyse (les dimensions). En s'appuyant sur des mécanismes d'agrégation, les outils *OLAP* (*On Line Analytical Processing*) (Agrawal et al., 1997 ; M. Chen et al., 1996 ; Han, 1997) permettent de naviguer aisément le long des hiérarchies des dimensions. La puissance de ces outils a placé les entrepôts au centre des systèmes d'information décisionnels (Mallach, 2000). Toutefois, la manière d'agréger les données étant déterminée par la modélisation des hiérarchies de dimension, celle-ci peut souffrir d'un manque de flexibilité. C'est donc ce qui fait l'objet d'une de nos contributions.

Le travail de préparation, de structuration des données a pu constituer une limite dans le recours aux entrepôts de données, notamment face à l'ampleur des données à considérer (contexte du *big data*). Les lacs de données (Dixon, 2010) ont émergé comme une alternative avec deux propriétés caractérisantes : la variété des données et le fait de garder les données dans leur format d'origine, amenant une flexibilité. La contrepartie pour permettre cette flexibilité réside dans le fait de disposer d'un système de métadonnées (données décrivant les données elles-mêmes) efficace, qui constitue le cœur de la contribution présentée ici.

2.2 Modélisation de hiérarchies de dimension dans les entrepôts de données

La recherche sur la modélisation des entrepôts de données a évolué depuis son émergence dans les années 90, pour prendre au fur et à mesure en compte certains besoins d'analyse, jusqu'à répondre à des besoins d'analyse nécessitant une certaine personnalisation. Nous avions pu faire un état des lieux des approches permettant cette personnalisation dans le cadre d'une collaboration avec Franck Ravat et Olivier Teste de l'Université de Toulouse en 2009 dans un article intitulé « Personnalisation dans les entrepôts de données : bilan et perspectives » (Bentayeb et al., 2009). Cet article avait été rédigé à l'issue de ma thèse qui portait précisément sur la personnalisation des analyses en ligne de type *OLAP* à partir de l'expression de règles par des utilisateur/trices, (Favre, 2007).

Au moment de mon recrutement, en 2009, j'ai eu l'occasion de collaborer avec Yoann Pitarch sur ce sujet qui recoupait certaines de ses préoccupations scientifiques, dans le cadre de sa thèse [1] qu'il réalisait à l'Université Montpellier 2, sous la co-direction d'Anne Laurent et de Pascal Poncelet, thèse intitulée « Résumé de Flots de Données : Motifs, Cubes et Hiérarchies » (Pitarch, 2011).

2.2.1 Données considérées et explicitation des enjeux

Ce travail a été réalisé dans le contexte de données médicales. Un des enjeux forts réside alors dans le fait de pouvoir apporter un soutien dans l'aide à la décision, en rendant possible des analyses prenant en compte les données des patient·es mais aussi les savoirs médicaux.

Pour illustrer la problématique et la solution apportée, considérons le cas d'un hôpital qui souhaiterait mettre en œuvre un entrepôt de données enregistrant pour chaque patient·e de son service de réanimation sa tension ainsi que les médicaments prescrits au fil du temps. Ces valeurs seraient mesurées par des capteurs et alimenteraient directement l'entrepôt.

Afin de réaliser un suivi efficace des patient·es, un·e médecin souhaiterait par exemple connaître les personnes qui ont eu une tension artérielle basse au cours de la nuit. Pouvoir formuler ce type de requête suppose l'existence d'une hiérarchie sur la tension artérielle dont le premier niveau d'agrégation serait une catégorisation de la tension artérielle (*e.g.*, basse, normale, élevée). Toutefois, cette catégorisation est délicate car elle dépend à la fois de la tension artérielle mesurée mais aussi de certaines caractéristiques physiologiques de la personne (âge, tabagie, ...). Dès lors, une même tension peut être généralisée différemment selon le contexte d'analyse considéré. Par exemple, une valeur de 13 pour la tension artérielle est **élevée** chez un **nourrisson** alors qu'il s'agit d'une tension **normale**

1. Ce travail a été réalisé dans le cadre du projet ANR MIDAS (ANR-07-MDCO-008)

chez une personne **adulte**.

L'entrepôt permettrait d'observer deux faits : la tension et la posologie. Les dimensions considérées sont les suivantes. La dimension temps (partagée par les deux faits), la dimension médicament rattachée au fait posologie et la dimension patient (partagée également par les deux faits). Le schéma correspondant, en adoptant le formalisme graphique introduit par Golfarelli et al., 1998b puis étendu par Ravat et al., 2007b, est présenté dans la figure I.2.1. Nous nous attardons plus spécifiquement sur la dernière dimension. Chaque patient·e possède un identifiant unique et est décrit·e par son nom, son âge, son sexe, la ville d'habitation et par un attribut « Fumeur » qui indique si la personne fume ou non. L'âge peut être considéré selon trois niveaux de détail différents : Age, SubCatAge et CatAge.

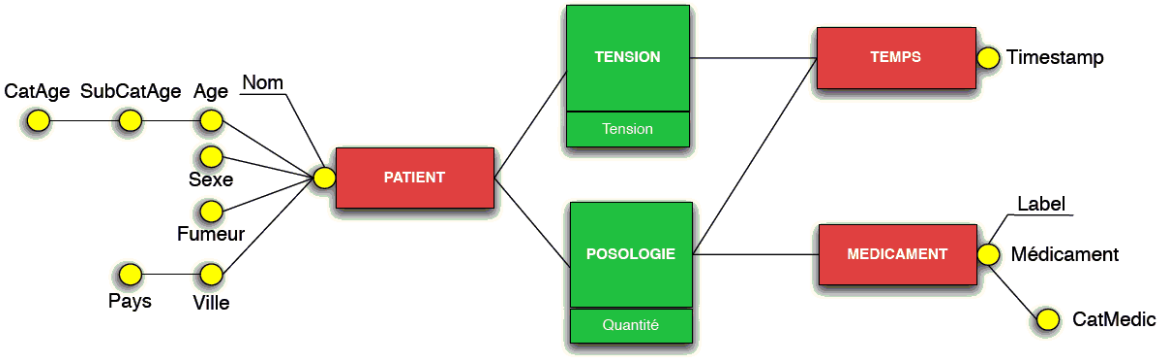


Figure I.2.1 – Schéma classique de l'entrepôt de données pour l'analyse de la tension et de la posologie.

Afin d'assurer un suivi efficace des patient·es du service, il est souhaitable de pouvoir formuler des requêtes telles que « Quels/quelles sont les patient·es dont la tension artérielle a été élevée pendant la nuit ? » ou « Quels/quelles sont les patient·es qui se sont vu·es prescrire une quantité trop importante de médicament X ? », etc.

Malheureusement, les modèles traditionnels d'entrepôts (celui de la figure I.2.1 par exemple) ne permettent pas de répondre à de telles requêtes pour deux raisons. Premièrement, la notion de tension (resp. posologie) élevée peut être considérée comme une généralisation de la tension mesurée (resp. de la quantité prescrite). Dans la mesure où les modèles classiques ne permettent pas d'établir une hiérarchie sur les mesures, ces requêtes ne peuvent être formulées.

De plus, même si l'on suppose que de telles requêtes sont formulables, la généralisation correcte de valeurs numériques est bien souvent contextuelle. Dans ce cas d'étude, nous considérons que les notions de tension élevée, de posologie élevée sont directement liées à certaines caractéristiques des patient·es et/ou des médicaments prescrits. Par exemple, un bébé ne doit pas recevoir la même quantité d'un médicament qu'une personne adulte. Ainsi, une même posologie pourra être considérée comme faible, normale ou élevée selon l'âge considéré. Une connaissance experte est alors nécessaire pour (1) définir quels sont les attributs qui impactent la généralisation d'un attribut (attributs

contextualisant) et (2) décrire cette généralisation en fonction des valeurs prises par ces attributs contextualisant.

C'est ainsi que notre contribution a visé à proposer une modélisation et sa mise en œuvre pour rendre plus flexibles les analyses issues de l'entrepôt de données.

2.2.2 Contributions : une modélisation à base de « satellite »

Les contributions se sont articulées autour des enjeux de modélisation, avec notamment la proposition d'un modèle formel, mais également la mise en œuvre de cette modélisation. Une extension de cette proposition a aussi été faite pour permettre la prise en compte de la logique floue dans la manière de considérer les connaissances.

La contribution majeure ici étant l'apport du concept de satellite dans la modélisation, j'ai choisi de me focaliser sur la présentation du modèle graphique qui illustre le modèle formel proposé disponible dans les publications associées.

L'apport est de pouvoir représenter des hiérarchies contextuelles aussi bien au niveau des dimensions que des mesures. Rappelons en préambule que la notion de contexte ne correspond ni au concept traditionnel dans les entrepôts de données de contexte d'analyse, ni au contexte au sens où l'on peut l'entendre lorsqu'un processus de personnalisation est mis en œuvre. Il s'agit bien ici d'exprimer le fait que pour une hiérarchie, au moins un des liens de généralisation entre deux niveaux ne peut être simplement déterminé grâce à un lien un à plusieurs prédéfini dans la mesure où d'autres informations (relevant d'autres dimensions par exemple) sont nécessaires.

Pour supporter le processus qui vise à la prise en compte de contextes par rapport à la détermination de la valeur de certains attributs généralisant les mesures ou des attributs de dimension, il est alors crucial de disposer d'un modèle d'entrepôt qui retrace cette contextualisation, par conséquent un modèle plus flexible. Le principe était de partir d'un modèle classique en constellation [2] que nous avons étendu par le concept de contexte.

Pour illustrer notre contribution, nous nous focalisons sur la catégorisation d'une tension et celle de l'âge. Le tableau I.2.1 présente quelques exemples de connaissances sur la catégorisation d'une tension en fonction des attributs SubCatAge et Fumeur d'une personne hospitalisée. Par exemple, une tension à 13 est normale chez une personne adulte qui fume mais est élevée chez un nourrisson [3].

La généralisation d'un âge (qui est un attribut de dimension et non une mesure) au niveau SubCatAge peut, elle aussi, être contextuelle. Nous supposons dans ce cas d'étude

2. Traditionnellement pour les entrepôts de données (et de manière très succincte dans la présentation), il est question de modèle en étoile, en flocon de neige et en constellation. Le modèle en étoile retranscrit le fait à analyser et ses dimensions d'analyse. Le modèle en flocon de neige présente le fait avec ses dimensions qui sont organisées en hiérarchies. Le modèle en constellation présente plusieurs faits qui sont analysés, avec des hiérarchies de dimension qui peuvent être partagées par plusieurs faits.

3. Les connaissances utilisées ici dans les exemples n'ont qu'une portée illustrative et ne doivent pas être considérées comme exactes.

SubCatAge	Fumeur	Tension	CatTension
Nourrisson	Oui ou Non	>12	Elevée
Adulte	Oui	>14	Elevée
$3^{ème}$ âge	Oui ou Non	> 16	Elevée
Nourrisson	Oui ou Non	Entre 10 (inclus) et 12 (inclus)	Normale
Adulte	Oui	Entre 12 (inclus) et 14 (inclus)	Normale
…	…	…	…

TABLEAU I.2.1 – Exemple de règles expertes décrivant la catégorie d'une tension (CatTension) en fonction de la tension mesurée, de la classe d'âge (SubCatAge) et de l'attribut Fumeur.

qu'elle dépend par exemple à la fois de l'âge mais aussi de la valeur associée à l'attribut Pays. Ceci permet de signifier que l'espérance de vie n'est pas la même dans les différents pays et que, de ce fait, la catégorie d'âge d'une personne peut varier en fonction de cette caractéristique. Le tableau I.2.2 présente d'ailleurs un extrait des règles permettant une généralisation correcte. Dès lors, généraliser correctement les tensions mesurées implique au préalable d'avoir correctement généralisé l'âge de la personne au niveau SubCatAge.

Age	Pays	SubCatAge
Plus de 75 ans	France	$3^{ème}$ âge
Plus de 50 ans	Swaziland	$3^{ème}$ âge
…	…	…

TABLEAU I.2.2 – Exemple de règles expertes décrivant la généralisation au niveau SubCatAge en fonction de l'âge de la personne et de son pays.

Pour permettre la contextualisation dans les généralisations d'attributs, nous introduisons le concept de satellite. Si nous appliquons la représentation graphique proposée à notre cas d'étude, nous obtenons le modèle de la figure I.2.2.

Nous pouvons relever la présence de trois hiérarchies contextuelles. Une d'entre elles, (IdPatient ; Age ; SubCatAge ; CatAge ; ALLPatient), est une hiérarchie contextuelle dimension alors que les deux autres, i.e., (Tension ; CatTension ; NormaliteTension) et (Quantite ; CatQuantite ; NormaliteQuantite), sont des hiérarchies contextuelles de mesures.

Un tel formalisme graphique rend assez immédiate la compréhension de l'entrepôt modélisé, et la possibilité de le discuter avec des expert·es prenant part aux décisions. En effet, le formalisme reste très proche des modèles classiques, en présentant une extension aisément interprétable grâce à la représentation du concept de satellite, qui demeure dans la métaphore classique céleste (étoile, flocon, constellation et galaxie de Ravat et al., 2007a pour une évolution plus récente). Ainsi, l'objectif de favoriser les interactions entre le niveau conception de l'entrepôt et le niveau expertise/décision peut être atteint

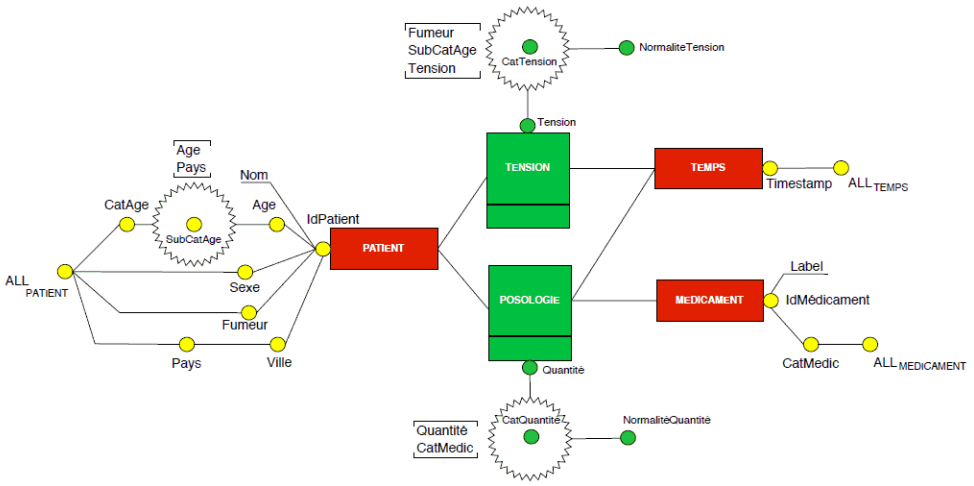


Figure I.2.2 – Représentation graphique de l'entrepôt de données médicales avec satellites.

(Moody & Shanks, 1994).

Bien évidemment, ce modèle graphique se situe au niveau de la représentation structurelle. Cela permet de structurer les connaissances à exprimer et de les confronter avec toutes les parties prenantes de la modélisation.

Pour l'explicitation des connaissances elles-mêmes, il s'agit de s'intéresser ensuite aux instances. Une fois la modélisation et le stockage des connaissances assurés, il s'agit alors de se focaliser sur le processus d'exploitation des données, et entre autres le processus d'agrégation (opération *Roll Up* par exemple), prenant en compte les hiérarchies contextuelles.

Cette proposition a été mise en œuvre et notamment présentée dans (Pitarch et al., 2010b). Les connaissances expertes (structures et instances de contexte) ont été stockées via une base de données relationnelles externe afin de garantir la généricité du stockage des différents contextes présents dans l'entrepôt (application web basée sur le SGBD PostgreSql).

En exploitant efficacement la connaissance stockée et en permettant une gestion facile de la base de données externe, nous avons prouvé que l'approche proposée présente un réel intérêt dans le contexte des entrepôts de données médicales.

Cette contribution a permis un apport de flexibilité dans le contexte des entrepôts de données, avec la prise en compte de l'utilisateur/trice qui est placé·e au cœur du processus de l'analyse en ligne, avec l'expression de ses connaissances.

L'approche que nous proposons permet un gain d'expressivité important dans les entrepôts de données médicales en permettant la modélisation, le stockage et l'exploitation de connaissances expertes dans le processus de généralisation de mesures. Néanmoins, un autre aspect essentiel doit être considéré ici : gérer, maintenir et mettre à jour cette connaissance. En effet, en fonction des personnes accueillies dans le service ou des pro-

grès de la médecine, certains contextes et instances de contexte peuvent être ajoutés, modifiés ou bien supprimés.

2.2.3 Discussion

2.2.3.1 Conception de schéma d'entrepôt de données et modélisation de hiérarchies de dimension

En 2009, une étude présentant les méthodologies de modélisation multidimensionnelle a été proposée par Romero et Abelló, 2009. Au-delà d'une description des travaux, une étude comparative est menée selon différents critères. Parmi ces critères figure ce que les auteurs ont appelé « paradigme ». Du point de vue de la conception du schéma de l'entrepôt, nous distinguons dans la littérature trois grandes approches par rapport à ce paradigme : celle guidée par les données, qualifiée également d'ascendante ; celle guidée par les besoins d'analyse, dénommée également descendante et l'approche mixte qui combine les deux précédentes (Soussi et al., 2005).

L'approche orientée données ignore les besoins d'analyse a priori. Elle concerne en particulier les travaux sur l'automatisation de la conception de schéma. En effet, cette approche consiste à construire le schéma de l'entrepôt à partir de ceux des sources de données et suppose que le schéma qui sera construit pourra répondre à tous les besoins d'analyse. Par exemple, Golfarelli et al., 1998a proposent une méthodologie semi-automatique pour construire un schéma d'entrepôt de données à partir des schémas entité-relation qui représentent les bases de données sources. Peralta et al., 2003 représentent les connaissances sur la construction de l'entrepôt de données sous forme de règles. Un algorithme gère alors l'ordre d'exécution des règles qui permettent une succession de transformations sur le schéma source pour obtenir le schéma logique final de l'entrepôt.

Les approches orientées besoins d'analyse, quant à elles, proposent de définir le schéma de l'entrepôt en fonction des besoins d'analyse et supposent que les données disponibles permettront la mise en œuvre d'un tel schéma, ou tout du moins que la confrontation avec les données réelles se fera dans un second temps (Prat et al., 2006).

Enfin, l'approche mixte considère à la fois les besoins d'analyse et les données pour la construction du schéma. L'idée générale est de construire des schémas candidats à partir des données (démarche ascendante) et de les confronter aux schémas définis selon les besoins (démarche descendante) (Bonifati et al., 2001 ; Phipps & Davis, 2002 ; Soussi et al., 2005). Quant à Romero et Abelló, 2010, ils proposent une méthode de dérivation des schémas multidimensionnels en fonction des besoins d'analyse, méthode incrémentale guidée par les exemples. Ainsi, le schéma construit constitue une réponse aux besoins réels d'analyse et il est également possible de le mettre en œuvre avec les sources de données. Il apparaît donc important dans le cadre de cette démarche de pouvoir discuter des schémas compréhensibles avec les utilisateur/trices (ayant l'expertise du domaine mé-

tier).

Dans le travail que nous avons réalisé, l'enjeu se situait plutôt sur l'expression des connaissances, à partir d'un schéma d'entrepôt existant, et des données qu'il contient, afin de modéliser des hiérarchies de dimension et de mesure contextualisées pertinentes. Il apparaît alors nécessaire de revenir sur l'existant en matière de modélisation de hiérarchies.

Différents travaux se sont intéressés à la modélisation des hiérarchies de dimension, la notion de hiérarchie de mesure n'étant pas traditionnelle pour la modélisation des entrepôts de données. Deux sont mentionnés ici principalement. Malinowski et Zimányi, 2004 ont proposé une classification des différents types de hiérarchies, en se basant sur des situations réelles. Ce travail met en avant des notations graphiques basées sur les modèles Entités/Associations, en exploitant entre autres les cardinalités. Différents types de hiérarchies sont représentés : symétriques/asymétriques, strictes/non strictes, multiples, parallèles, etc.

Un autre travail a été développé par Ghozzi et al., 2003 autour des bases de données multidimensionnelles contraintes en exprimant différents types de contraintes sur les hiérarchies de dimension (contrainte inter-dimensions et intra-dimension) avec une proposition de représentation graphique au niveau des opérations sur les hiérarchies, permettant d'assurer la consistance des données de l'entrepôt.

Ces travaux mettent en avant l'intérêt d'une modélisation graphique pour une meilleure compréhension du modèle. Toutefois, la modélisation des hiérarchies contextuelles telles que nous en avions besoin pour représenter la réalité de nos données médicales n'était pas prise en charge par ces formalismes. Ainsi, la modélisation graphique de ces hiérarchies contextuelles constituait le caractère original de notre proposition. Cette proposition a également été étendue pour intégrer la logique floue, pour encore davantage de flexibilité.

2.2.3.2 Place des utilisateur/trices

Au travers de ce travail, nous avons mis en avant l'importance d'impliquer les utilisateur/trices dans la phase de conception du modèle et dans l'expression des savoirs pour permettre la définition contextualisée des hiérarchies de dimension.

D'un point de vue méthode, il s'agit de pouvoir échanger sur un modèle facilement compréhensible qui représente bien la réalité des données et la manière de les analyser. Nous constatons alors l'émergence de deux caractéristiques importantes qui peuvent apparaître contradictoires d'un premier abord, à savoir : l'expressivité d'un modèle (à quel point l'on peut exprimer au travers du modèle des situations complexes posées par la réalité des données) et la simplicité du modèle (pour permettre la discussion entre les personnes qui réalisent cette conception et les utilisateur/trices de l'entrepôt de données qui connaissent le domaine métier). Une présentation graphique du modèle permet sans

aucun doute d'accéder à cette simplicité (Moody & Shanks, 1994).

Cette flexibilité amenée sous la forme d'une prise en compte des connaissances pose alors la question du partage de ces connaissances et de savoir à quel point une démarche collaborative est souhaitable dans l'analyse. Est-ce que cela doit rester au niveau d'une expression individualisée de ces connaissances ou s'agit-il de proposer un fonctionnement basé sur des connaissances partagées? Si cela se base sur des connaissances partagées, cela mérite alors sans doute de pouvoir être discuté, et ce grâce à une formalisation qui permet de le faire facilement pour que cela soit efficace.

Rappelons qu'historiquement, les premiers modèles d'entrepôt de données (étoile, flocon, constellation proposés par Kimball, 1996) ont émergé et connu un vif succès au sein des entreprises. Il est assez naturel de faire l'hypothèse que la simplicité de lecture/d'interprétation de ces modèles graphiques a sans doute participé à leur succès. D'un point de vue support d'échange entre les personnes se situant sur les plans de la conception et de l'utilisation (niveau décision / expertise du domaine métier), il paraît assez naturel de faire un parallèle entre le modèle entité/association dans le contexte des bases de données et les modèles en étoile/flocon/constellation dans le contexte des entrepôts de données.

Le travail proposé permet en tous cas d'intégrer une forme de flexibilité aux modèles plus traditionnels, en intégrant l'humain au cœur du système d'information décisionnel, alors même que c'est une des caractéristiques initiales de ce qui était mis en avant pour ces systèmes décisionnels.

Une autre manière d'amener une forme de flexibilité dans l'analyse est de considérer le lac de données comme source pour analyser les données, ce qui nécessite un système de modélisation des métadonnées. C'est la contribution que nous abordons dans la section suivante.

2.3 Modélisation de métadonnées pour les lacs de données

La thèse d'Etienne Scholly, co-encadrée avec Sabine Loudcher, s'est déroulée dans le cadre d'un dispositif CIFRE [4](Convention Industrielle de Formation par la REcherche), amenant ainsi une collaboration avec l'entreprise BIAL-X, qui travaille sur la gestion et la transformation de données. Ainsi Eric Ferey a également participé à l'encadrement du côté de l'entreprise. Elle a été soutenue en 2022 (Scholly, 2022), avec pour titre « De la modélisation des métadonnées à la conception d'un lac de données. Application à l'habitat social. » et rend compte du travail notamment réalisé sur la modélisation des métadonnées dans le cadre des lacs de données, avec un focus sur une mise en œuvre dans la thématique de l'habitat social.

Au sein du laboratoire ERIC, compte-tenu du fait qu'il y avait différentes thèses et

4. https://www.anrt.asso.fr/fr/le-dispositif-cifre-7844

collaborations portant sur les lacs de données, dont celle d'Etienne Scholly évoquée à l'instant, une petite équipe s'est constituée à compter de 2019 autour de ce thème de recherche. L'idée était de pouvoir travailler ensemble autour de la modélisation des métadonnées, et de tirer parti des différents cas d'études spécifiques en termes de choix d'implémentation. Cette équipe des *Data Lakers* s'est organisée autour de trois personnes enseignantes-chercheuses permanentes (Jérôme Darmont, Sabine Loudcher et moi-même), deux doctorants (Etienne Scholly et Pegdwendé Nicolas Sawadogo), rejoints par Pengfei Liu et Javier A. Espinosa Oviedo dans le cadre de leur post-doctorat. Ceci nous a amené à travailler sur la modélisation de métadonnées à partir de trois cas d'usages différents.

Avant d'aborder la modélisation des métadonnées, il est nécessaire de revenir plus précisément sur la définition de lac de données.

La définition la plus complète des lacs de données est celle de Madera et Laurent (2016), qui en plus de la variété des données et de l'approche *schema-on-read* (le schéma des données est défini seulement à l'interrogation), définit des caractéristiques supplémentaires.

En ajustant cette définition par rapport à quelques points, la proposition de définition sur laquelle nous avons basé la suite des travaux est la suivante :

Un lac de données est un système évolutif (en termes de passage à l'échelle) de stockage et d'analyse de données de tous types, dans leur format natif, utilisé principalement par des spécialistes des données (statisticien·nes, *data scientists*, *data analysts*) pour l'extraction de connaissances. Les caractéristiques d'un lac de données incluent :

1. un catalogue de métadonnées qui assure la qualité des données ;
2. une politique et des outils de gouvernance des données ;
3. l'ouverture à tous types d'utilisateur/trices ;
4. l'intégration de données de tous types ;
5. une organisation logique et physique ;
6. le passage à l'échelle.

Les entrepôts de données (qui ont été abordés dans la contribution précédente) et les lacs de données présentent des différences notoires qui sont synthétisées dans le tableau I.2.3

Il me paraît judicieux de préciser que les lacs de données ne sont pas forcément l'avenir des entrepôts de données. En effet, il s'agit bien de préciser les besoins décisionnels en amont, quelles sont les données à considérer, pour décider quel type d'architecture est pertinent vis-à-vis de la situation.

Nous abordons à présent les données de la société qui ont été considérées pour ce travail et ses enjeux.

TABLEAU I.2.3 – Différences principales entre entrepôt de données et lac de données.

Critère / Système	Entrepôt de données	Lac de données
Type de données stockées	Données structurées (ou de nature homogène)	Données de tous types
Intégration des données	ETL (Extract - Transform - Load)	ELT (Extract - Load - Transform)
Définition du schéma	À l'insertion (schema-on-write)	À l'interrogation (schema-on-read)
Flexibilité	Rigide, peu évolutif	Flexible, mais besoin d'un système de métadonnées
Type d'analyses	Analyses prédéfinies	Tous types d'analyses

2.3.1 Données considérées et explicitation des enjeux

L'entreprise BIAL-X est historiquement un cabinet d'expertise en informatique décisionnelle (*business intelligence*), où les consultant·es de la société accompagnent leur clientèle dans des projets pour la mise en place et la maintenance d'un système d'information décisionnel.

BIAL-X travaille notamment en lien avec des bailleurs sociaux. Un bailleur social est un organisme qui loue des logements sociaux à des ménages pour un loyer modéré, sous condition de ressources.

Il y a un fort enjeu pour les bailleurs sociaux à disposer d'un système décisionnel permettant d'optimiser la gestion des biens immobiliers, par exemple pour diminuer le temps d'inoccupation lié à des besoins de travaux entre deux locations, en étant capable de prévoir ces besoins de travaux.

Prenons un exemple concret. Les bailleurs disposent généralement d'un grand nombre d'informations sur leurs logements (superficie, nombre de pièces, type de chauffage, date de construction, etc.), mais ont en revanche peu voire pas d'informations sur l'environnement dudit logement (transports, services publics, sécurité, emploi, etc.). Ces dernières informations sont néanmoins de plus en plus mises à disposition sur le web en tant que données ouvertes. Croiser ces données externes concernant l'environnement d'un logement avec les données internes du logement permettrait de répondre à des questions auxquelles il est difficile de trouver des réponses uniquement avec les données des bailleurs. Par exemple, le fait qu'un logement refait à neuf, spacieux et lumineux mais restant vacant peut difficilement s'expliquer, sauf à avoir la connaissance qu'il est situé dans un quartier isolé, avec peu de transports ou une forte insécurité.

Ainsi l'habitat social constituait le cas d'étude pour le travail d'Etienne Scholly, auquel se sont ajoutés deux autres cas d'étude au niveau de l'équipe des *Data Lakers*. La thèse de Pegdwendé N. Sawadogo était financée par la Région Auvergne-Rhône-Alpes à travers le projet AURA-PMI, et avait pour objectif d'analyser l'avancée de la servicisation et la

digitalisation dans les Petites et Moyennes Industries (PMI) de la Région Rhône-Alpes-Auvergne. Par ailleurs, le projet HyperThesau portait sur la mise en œuvre d'un lac de données pour l'exploitation de données archéologiques dans lequel ont collaboré Pengfei Liu et Javier A. Espinosa Oviedo.

Ces divers cas ont été discutés lors des réunions portant sur la modélisation des métadonnées qui constitue un enjeu majeur des lacs de données. L'intérêt du lac de données peut être réel en permettant à la fois une variété des données qu'il est capable de prendre en charge, et une approche dans laquelle le schéma des données n'est défini qu'à leur interrogation (*schema-on-read*). Ces deux caractéristiques font qu'un lac de données est un système souple et adaptatif, mais il nécessite en contrepartie de disposer d'un système de métadonnées efficace. En l'absence d'un schéma fixe de données, les métadonnées sont alors indispensables pour empêcher que le lac se transforme en marécage de données (*data swamp*), un lac inutilisable!

Ainsi un des enjeux forts ici était de pouvoir s'intéresser à la modélisation des métadonnées. Notre cas d'étude se basait sur l'habitat social. Mais l'intérêt d'avoir travaillé en équipe, alors que plusieurs cas d'étude étaient couverts, était incontestablement de pouvoir réfléchir au-delà de celui-ci, allant plus facilement vers une généricité de la modélisation en la mettant à l'épreuve des différents cas.

2.3.2 Contributions : modèle et métamodèle de métadonnées pour les lacs de données

Le manque de généricité dans les approches proposées dans la littérature pour la modélisation des métadonnées et l'impossibilité de prendre en compte tous types de données, nous ont amené·es à proposer un modèle de métadonnées baptisé *MEDAL*, que nous avons par la suite, à la lumière de travaux plus récents, fait évoluer en un métamodèle de métadonnées nommé *goldMEDAL*.

Dans l'optique de cadrer notre travail de modélisation des métadonnées dans le contexte des lacs de données, il s'agit de préciser au préalable les informations servant à constituer les métadonnées, afin qu'elles soient les plus complètes possible. Pour ce faire, nous partons sur une typologie de métadonnées.

Pour l'introduire, nous avons considéré un concept générique représentant tout ensemble homogène de données que le modèle doit traiter. Certains travaux sur les lacs de données ont proposé les concepts d'unité de données, d'entité, de jeu de données (*dataset*) ou d'objet. Nous avons adopté la notion d'objet, qui nous semble plus appropriée pour représenter de façon abstraite un ensemble de données. Plus concrètement, un objet peut se matérialiser par une table relationnelle ou un fichier physique (document de tableur, XML ou JSON, document textuel, collection de tweets, image, vidéo, etc.).

Notre typologie organise les métadonnées en trois catégories : intra (les métadonnées contenues au sein d'un objet), inter (les métadonnées permettant d'interconnecter les

objets) et globales.

Les métadonnées intra regroupent :

— les **propriétés** qui fournissent une description générale de l'objet ;
— les **résumés et prévisualisations** qui ont pour rôle de donner un aperçu du contenu ou de la structure d'un objet ;
— les **versions** qui permettent de rendre compte de mises à jour et les **représentations** qui prennent en compte des opérations de formatage des données brutes ;
— les **métadonnées sémantiques** qui sont des annotations qui permettent de comprendre le sens des données.

Les métadonnées inter regroupent :

— les **regroupements d'objets** qui consistent à organiser les objets du lac en collections, chaque objet pouvant appartenir simultanément à plusieurs collections ;
— les **liaisons de similarité** qui traduisent la force de la ressemblance entre deux objets ;
— les **relations de parenté** qui traduisent le fait qu'un objet peut être issu de la jointure de plusieurs autres.

Les métadonnées globales comprennent :

— les **ressources sémantiques** qui sont essentiellement des bases de connaissances (ontologies, taxonomies, thésaurus, dictionnaires) utilisées à la fois pour générer d'autres métadonnées et améliorer les analyses ;
— les **index et index inversés** qui sont des structures de données permettant de retrouver rapidement un objet sur la base de caractéristiques précises ;
— les **journaux d'évènements** qui permettent de tracer les interactions entre les utilisateur/trices et le lac de données.

Le diagramme de classes de la figure I.2.3 présente de manière visuelle les concepts du modèle *MEDAL*. Les différentes classes et classes d'association comportent différents attributs qui ne sont pas représentés ici par souci de clarté. D'un point de vue logique, *MEDAL* adopte une représentation à base de graphes.

Dans un deuxième temps, nous avons fait évoluer *MEDAL* vers plus d'abstraction pour couvrir la possibilité d'une variété de cas qui aurait pu échapper à *MEDAL*, en proposant un métamodèle de métadonnées, baptisé *goldMEDAL*.

En nous inspirant des différentes notions introduites dans *MEDAL*, nous basons le modèle conceptuel de *goldMEDAL* sur quatre concepts principaux : entité de données, groupement, lien et processus (figure I.2.4).

Entité de données Les entités de données sont les unités de base du modèle de métadonnées. Elles sont flexibles en termes de granularité des données. Par exemple, une entité de données peut représenter un fichier de tableur, un document textuel ou semi-structuré, une image, une table de base de données, un tuple ou une base de données entière. L'introduction de tout nouvel élément dans le lac de données en-

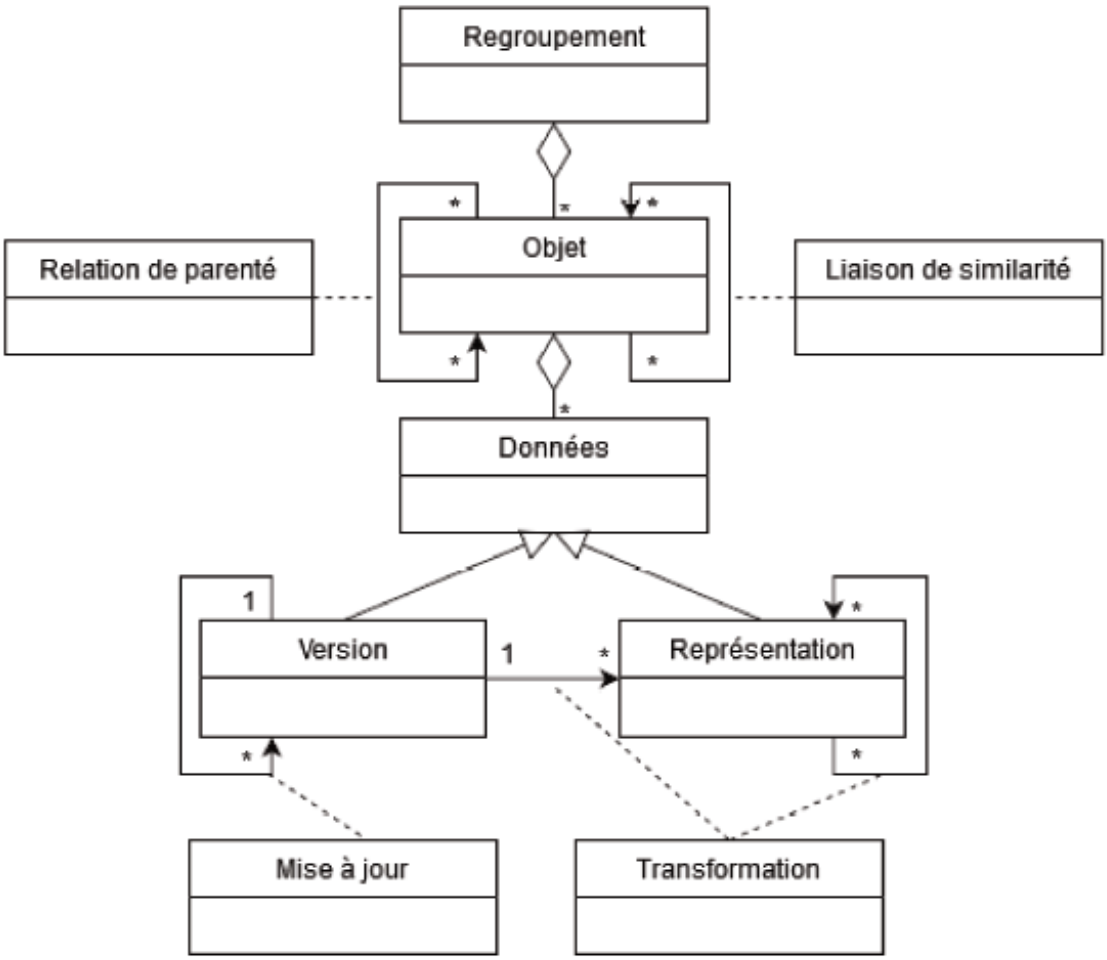


Figure I.2.3 – Diagramme de classes UML des concepts de *MEDAL*.

traîne la création d'une nouvelle entité de données.

Groupement Un groupement est un ensemble de groupes ; un groupe rassemble des entités de données sur la base de propriétés communes. Par exemple, des zones de données brutes et prétraitées d'un lac forment les groupes d'un groupement de zones. Un autre exemple est un regroupement de documents textuels en fonction de la langue d'écriture.

Lien Les liens sont utilisés pour associer soit des entités de données entre elles, soit des groupes d'entités de données entre eux. Ils peuvent être orientés ou non. Ils permettent d'exprimer, par exemple, de simples liens de similarité entre entités de données ou des hiérarchies entre groupes. Par exemple, une hiérarchie temporelle mois → trimestre aurait les mois de janvier, février et mars liés au premier trimestre d'une année donnée.

Processus Un processus désigne toute transformation appliquée à un ensemble d'entités de données qui produit un nouvel ensemble d'entités de données.

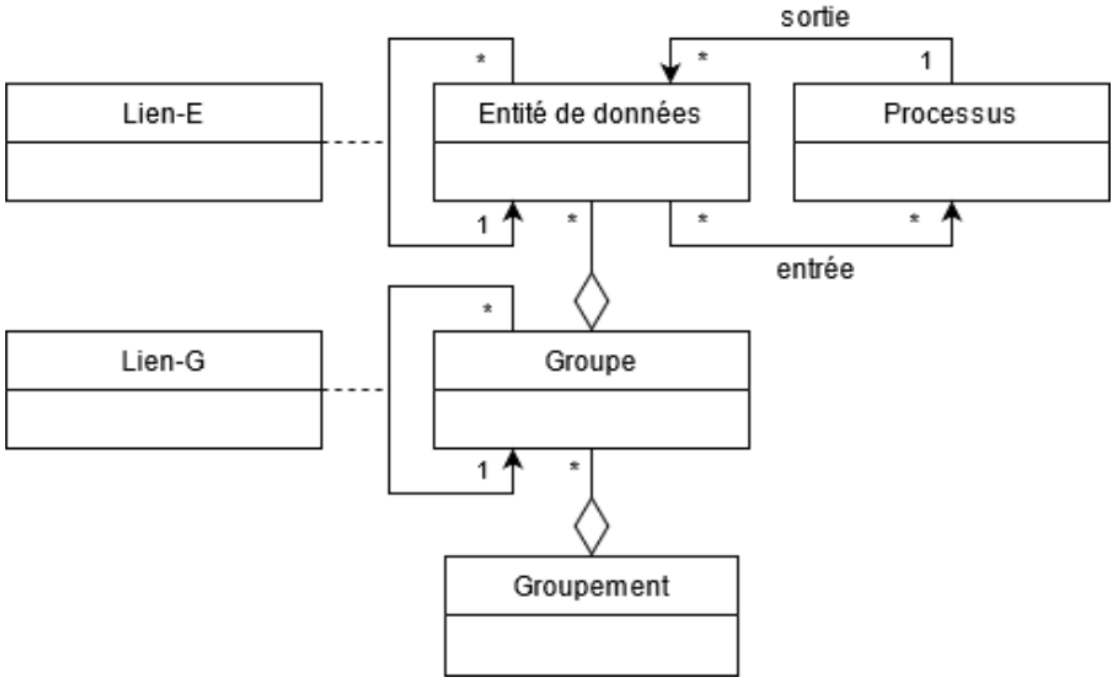


Figure I.2.4 – Diagramme de classes UML des concepts de *goldMEDAL*.

Au niveau logique, les concepts de *goldMEDAL* sont représentés à travers un graphe. Ainsi, les entités de données sont représentées par des nœuds. Les liens deviennent des arêtes. Enfin, les groupes et processus sont traduits en hyper-arêtes. Au niveau physique, *goldMEDAL* a été implémenté dans trois cas d'usage différents.

Pour le cas d'étude sur l'habitat social qui nous intéresse ici plus particulièrement, la mise en œuvre effective de cette proposition s'est traduite par HOUDAL qui constitue une implémentation de lac de données dédié à l'habitat social en s'appuyant sur le système de gestion de base de données graphes Neo4J [5] pour fournir un service de stockage et d'analyse de données principalement structurées pour notre cas d'étude.

2.3.3 Discussion

L'étude des modèles de métadonnées des lacs de données est un sujet de recherche assez actif et prolifique, ayant donné lieu à plusieurs propositions.

Parmi celles-ci, le modèle *MEDAL*, que nous avons initialement proposé, représente les données à travers trois concepts principaux : les objets qui correspondent à un ensemble de données homogènes, les représentations qui résultent de transformations de l'objet associé et les versions qui représentent les mises à jour d'un objet. Cependant, le modèle *MEDAL* ne peut pas représenter simultanément différents niveaux de granularité des données.

5. https://neo4j.com

Ravat et Zhao (2019) proposent un modèle qui met en avant la notion de métadonnées de zone, qui spécifie la zone où se trouvent les données (par exemple, zone de données brutes, zone de données traitées). Toutefois, ce modèle ne prend pas non plus en charge les niveaux de granularité multiples des données.

Eichler et al. (2020) introduisent le modèle *HANDLE*, qui utilise le concept générique d'entité de données pour représenter à la fois des fichiers de données et des parties de fichiers de données. Ainsi, n'importe quel niveau de granularité peut être pris en charge. Chaque entité de données est associée à des étiquettes qui représentent des zones, des niveaux de granularité ou des catégorisations. Ceci étant, *HANDLE* ne prend pas en compte le versionnement des données.

En termes de modélisation de métadonnées, *goldMEDAL* correspond à la contribution qui est venue parfaire la généricité qui était nécessaire. Ainsi, à travers les trois modèles physiques implémentés avec *goldMEDAL*, nous avons pu démontrer la faisabilité ainsi que la flexibilité de notre modèle de métadonnées.

Les concepts de *goldMEDAL* sont venus généraliser ceux des modèles les plus récents : notre proposition précédente de *MEDAL*, celle de Ravat et Zhao (2019) et *HANDLE* (Eichler et al., 2020). Cela fait de *goldMEDAL* le modèle le plus générique pour la modélisation des métadonnées de lacs de données à ce jour.

Bien entendu, cette métamodélisation des métadonnées n'est qu'une étape pour rendre compte ensuite de la réalité des métadonnées dans le cas à traiter pour permettre la phase d'analyse des données du lac.

2.4 Réflexions conclusives : de la modélisation de données au questionnement sur la construction de catégories d'analyse, un enjeu pour l'informatique décisionnelle

Cette section de réflexions conclusives n'aura pas pour objectif de dresser des perspectives directes à ces travaux, puisque ceux-ci font moins partie des travaux que je tends à développer selon le même axe de travail. Mais je tiens à apporter ici quelques éléments de réflexions sur ces travaux à l'aune de mon parcours.

La dimension modélisation a été un volet important des travaux dans lesquels je me suis impliquée depuis ma prise de poste, en témoignent les sections précédentes présentant des contributions en la matière. En effet, les deux sections précédentes ont montré les apports respectifs en matière de modélisation pour les entrepôts d'une part et pour les lacs de données d'autre part.

Par ailleurs, il apparaît important de souligner que les entrepôts comme les lacs de données se placent dans le contexte de l'aide à la décision. Ainsi, il est important de resituer cette dimension de décision, qui est primordiale, notamment en fonction du do-

maine d'application, et donc des enjeux de modélisation qui conditionnent finalement les éléments sur lesquels vont s'appuyer les décisions. Le cas du domaine médical illustre parfaitement cet enjeu vital, dans les deux sens du terme !

Ceci est vrai que ce soit dans la modélisation des entrepôts de données, et des spécificités qui peuvent être prises en compte, par exemple dans le cas des hiérarchies contextuelles comme nous l'avons proposé, mais aussi dans le contexte des lacs de données où ce qui va être pris en compte ou non dans la modélisation des métadonnées va impacter les analyses qui pourront être faites.

Ainsi, l'impact de la modélisation sur les analyses possibles est fort et implique que ces choix de modélisation ne sont pas neutres.

Lorsque les analyses vont s'appuyer sur des mécanismes d'agrégation (qui sont à la base des approches des systèmes décisionnels, que ce soit pour les entrepôts de données ou les lacs de données), cet impact est renforcé par la manière dont sont définis les chemins d'agrégation, et ce, que ce soit au niveau de la structure, ou des instances.

Ceci est directement lié à un processus de catégorisation. La définition des catégories a un impact sur les résultats obtenus. Prenons l'exemple des âges. Selon la manière de les catégoriser, nous obtenons des résultats différents au moment de l'agrégation, nous amenant à regarder les données différemment selon les catégories qui sont mises en œuvre, sans forcément être en mesure de comparer différentes manières d'agréger les données.

Cette question de catégorisation se retrouve aussi au cœur du contexte de l'apprentissage de données, thématique abordée dans le chapitre suivant.

3 Analyse de la diffusion de l'information dans les médias sociaux grâce à l'apprentissage de données

Quoi qu'il en soit, quand un sujet se prête à de nombreuses controverses [...] on ne peut espérer dire la vérité et on doit se contenter d'indiquer le chemin suivi pour parvenir à l'opinion qu'on soutient. .

Citation traduite de Virginia Woolf (1882-1941), femme de lettres anglaise, « A Room of One's Own » (1929)

Contributions sur lesquelles se base ce chapitre

Axe détection d'évènements et diffusion de l'information

- > Mémoire de thèse d'Adrien Guille
  - A. Guille, Diffusion de l'information dans les médias sociaux : modélisation et analyse, Université Lumière Lyon 2. Thèse en informatique soutenue le 25 novembre 2014. (Guille, 2014)
- > Articles dans des revues internationales
  - A. Guille & **C. Favre**, Event detection, tracking, and visualization in Twitter : a mention-anomaly-based approach, Social Network Analysis and Mining, Vol.5, N°1, 2015, 1-18. (Guille & Favre, 2015)
  - A. Guille, H. Hacid, **C. Favre** & D. A. Zighed, Information Diffusion in Online Social Networks : A Survey, SIGMOD Record, Vol.42, N°2, 2013, 17-28. (Guille, Hacid, Favre & Zighed, 2013)
- > Publications dans des conférences internationales
  - A. Guille & **C. Favre**, Mention-anomaly-based Event Detection and tracking in Twitter, ASONAM 2014, Beijing, China, 375-382. (Guille & Favre, 2014a)
  - A. Guille, **C. Favre**, H. Hacid & D. A. Zighed, SONDY : an open source platform for social dynamics mining and analysis, SIGMOD 2013, New York, USA, 1005-1008. Papier démo. (Guille, Favre, Hacid et al., 2013)
- > Publications dans des conférences nationales
  - A. Guille & **C. Favre**, Une méthode pour la détection de thématiques populaires sur Twitter, EGC 2014, Rennes, 83-88. Papier court. (Guille & Favre, 2014c)
  - A. Guille & **C. Favre**, Un système de détection de thématiques populaires sur Twitter, EGC 2014, Rennes, 605-608. Papier démo. (Guille & Favre, 2014b)
  - A. Guille, **C. Favre** & D. A. Zighed, SONDY : une plateforme open-source d'analyse et de fouille pour les réseaux sociaux en ligne, EGC 2013, Toulouse. Papier démo. (Guille, Favre & Zighed, 2013)
  - A. Guille, H. Hacid & **C. Favre**, Une approche multidimensionnelle basée sur les comportements individuels pour la prédiction de la diffusion de l'information sur Twitter, EGC 2012, Bordeaux, 405-410. (Guille et al., 2012)
- > Communication (sélection sur résumé)
  - A. Guille & **C. Favre**, Analyse et fouille pour les réseaux sociaux en ligne : la plateforme SONDY, édition 2013 des journées Big Data Mining and Visualization d'EGC, Paris. (Guille & Favre, 2013)

Axe détection de rumeurs

- > Mémoire de thèse d'Abderrazek Azri
  - A. Azri, Approches multimodales d'apprentissage automatique pour la détection des rumeurs dans les microblogs, Université Lumière Lyon 2. Thèse en informatique soutenue le 7 juillet 2022. (Azri, 2022)
- > Article dans une revue internationale
  - A. Azri, **C. Favre**, N. Harbi, J. Darmont & C. Noûs, Rumor Classification through a Multimodal Fusion Framework and Ensemble Learning, Information Systems Frontiers, Vol.25, N°5, 2023, 1795-1810. (Azri et al., 2023)
- > Publications dans des conférences internationales
  - A. Azri, **C. Favre**, N. Harbi, J. Darmont & C. Noûs, Calling to CNN-LSTM for Rumor Detection : A Deep Multi-channel Model for Message Veracity Classification in Microblogs, ECML-PKDD 2021, Bilbao, Spain, 497-513. (Azri et al., 2021a)
  - A. Azri, **C. Favre**, N. Harbi, J. Darmont & C. Noûs, MONITOR : A Multimodal Fusion Framework to Assess Message Veracity in Social Networks, ADBIS 2021, Tartu, Estonia, 73–87. (Azri et al., 2021b)
  - A. Azri, **C. Favre**, N. Harbi, J. Darmont, Including images into message veracity assessment in social media, INTIS 2019, Tangier, Morocco, 7 p. (Azri et al., 2019a)
- > Publications dans des conférences nationales
  - A. Azri & **C. Favre**, N. Harbi & J. Darmont, Vers une analyse des rumeurs dans les réseaux sociaux basée sur la véracité des images : état de l'art, EDA 2019, Montpellier, 125-142. (Azri et al., 2019b)

LES MÉDIAS SOCIAUX constituent des vecteurs d'informations qui ont pris beaucoup d'ampleur ces dernières années. Ainsi, un des enjeux était d'aider à la compréhension du phénomène de diffusion de l'information dans les médias sociaux, en fournissant des moyens d'analyse et de modélisation, ainsi que de pouvoir détecter les rumeurs.

L'ensemble des travaux présentés dans ce chapitre ont été menés sur Twitter. Il est à noter que ces travaux ont été menés avant le rachat de Twitter par Elon Musk en octobre 2022, rachat qui a induit de nombreux rebondissements médiatiques par rapport à des décisions prises impactant l'usage de ce média social.

La section I.3.1 présentera quelques éléments en préambule des contributions. Puis la section I.3.2 donnera à voir une synthèse des travaux menés sur la détection d'évènements et leur diffusion sur Twitter. Ensuite la section I.3.3 abordera les travaux sur la détection de rumeurs dans Twitter. Des réflexions conclusives viendront clore ce chapitre dans la section I.3.4

3.1 Préambule

L'émergence du Web 2.0, appelé quelques fois Web participatif, a donné lieu à la multiplication de plateformes diverses où les utilisateur/trices ont été en mesure de contribuer au Web. C'est notamment le cas des plateformes de *microblogging* qui permettent de partager en temps réel du contenu assez court.

Twitter a été un des services de *microblogging* les plus connus et plébiscités en Occident. Les utilisateur/trices des médias sociaux étant à la fois dans la production et la consommation d'information, l'augmentation continue de leur nombre s'accompagne d'une augmentation continue du volume de messages publiés.

Notons que le mécanisme d'abonnement aux comptes et la possibilité de repartager à sa propre audience des messages (principe de *retweet*) a induit des mécanismes de diffusion de l'information conséquents. Finalement, le principal facteur limitant la diffusion de l'information aujourd'hui n'est plus la disponibilité de l'information, mais la disponibilité de l'attention des personnes réceptrices de l'information. Face à cette situation de surcharge informationnelle, parfois appelée infobésité, le développement d'outils peut constituer une solution.

Ainsi, de nombreux travaux, notamment en informatique, se sont alors tournés sur l'analyse de ces sources de données, qui donnent à voir la société sous un certain prisme. D'autres disciplines se sont également intéressées à l'analyse de ces espaces « virtuels ». Mais ici, notre propos se focalisera sur les travaux menés en informatique pour mener à bien ces analyses, même si je suis convaincue de l'intérêt de croisements disciplinaires pour aborder tels types de données.

La volonté de développer des outils génériques ne doit pas masquer que le fait de considérer des données issues de Twitter, peut amener à des adaptations nécessaires en

considérant d'autres types de *microblog*.

En effet, certaines approches peuvent se baser sur des spécificités liées aux plateformes. Citons, par exemple, le fait que l'« abonnement » n'est pas soumis à validation sur Twitter – à savoir la possibilité de suivre les comptes voulus sans accord de la personne qui est suivie (non réciprocité pour la mise en lien) –, ce qui donne lieu à une diffusion différente de celle qu'il peut y avoir sur d'autres plateformes que Twitter.

Ces mécanismes de diffusion d'informations ont suscité des intérêts d'analyse qui couvrent à la fois des enjeux de détection d'évènements importants (au sens qu'ils suscitent l'intérêt des utilisateur/trices), des enjeux de prédiction de la diffusion de l'information, des enjeux de détection de l'influence dans la diffusion de l'information, mais aussi la détection de fausses informations (rumeurs). Ces points sont donc développés dans les deux sections qui suivent.

3.2 Détection d'évènements et diffusion

Cette section est basée sur les travaux d'Adrien Guille, menés dans le cadre de sa thèse de doctorat, que j'ai co-encadrée avec Djamel Zighed, à partir de septembre 2011 et qu'il a soutenue le 25 novembre 2014, avec pour titre « Diffusion de l'information dans les médias sociaux. Modélisation et analyse » (Guille, 2014).

3.2.1 Données considérées et explicitation des enjeux

Un des enjeux forts pour pouvoir évaluer des approches d'analyse de médias sociaux réside dans la disponibilité de jeux de données.

Il y a une dizaine d'années, il n'y avait pas forcément beaucoup de corpus disponibles à la communauté pour l'analyse, quand bien même des auteur/trices proposaient des méthodes d'analyse.

Afin d'éprouver la méthode pour la détection d'évènements qui était le point de départ des travaux, deux corpus ont été utilisés, l'un en anglais, l'autre en français. Le premier corpus – noté $\mathscr{C}_{en}$ – contient 1 437 126 tweets rédigés en anglais, collectés avec une stratégie centrée utilisateur/trice par Yang et Leskovec (2011). Ils correspondent à l'intégralité des tweets publiés durant le mois de novembre 2009 par 52 494 utilisateur/trices américain·es de Twitter. Ce corpus contient beaucoup de bruit. Selon l'étude menée par PearAnalytics (2009), la proportion de tweets sans rapport avec aucun évènement pourrait atteindre 50%. Ce corpus a également été utilisé pour le travail sur la modélisation et la prédiction de la diffusion d'information en le décomposant en jeux de données d'apprentissage et jeux de données de test.

Le second corpus – noté $\mathscr{C}_{fr}$ – contient 2 086 136 de tweets rédigés en français collectés avec une stratégie centrée-mots-clés en mars 2012, durant la campagne pour l'élec-

TABLEAU I.3.1 – Statistiques sur les corpus (@ : proportion de *tweets* qui contiennent des mentions, *RT* : proportion de *retweets*).

Corpus	# tweets	# auteurs	@	*RT*
$\mathscr{C}_{en}$	1 437 126	52 494	0,54	0,17
$\mathscr{C}_{fr}$	2 086 136	150 209	0,68	0,43

tion présidentielle en France. Adrien Guille avait obtenu ces tweets via l'*API streaming* de Twitter, en utilisant les noms des principaux candidats comme mots-clés. Ce corpus cible donc les thématiques politiques en rapport avec la France. Les mots triviaux sont retirés des messages à l'aide de listes de mots vides francophones et anglophones. La table I.3.1 donne des détails supplémentaires à propos de chaque corpus.

3.2.2 Contributions : détection d'évènements de *microblog* et suivi de la diffusion de l'information

À partir du phénomène de diffusion de l'information dans les médias sociaux, trois problématiques peuvent en découler : (i) détecter les évènements importants qui suscitent l'intérêt des utilisateur/trices, (ii) modéliser et prévoir la diffusion de l'information, et (iii) identifier des utilisateur/trices influençant la diffusion de l'information. Les travaux réalisés dans le cadre de la thèse d'Adrien Guille ont ainsi couvert ces trois problématiques en y apportant trois contributions décrites succinctement dans ce document.

— *MABED* (*Mention-Anomaly-Based Event Detection*) : une méthode statistique pour la détection et le suivi des évènements dans les médias sociaux.
— *T-BASIC* (*Time-Based ASynchronous Independent Cascades*) : un modèle probabiliste pour prévoir la diffusion de l'information dans les médias sociaux.
— *SONDY* (*SOcial Network DYnamics*) : un logiciel implémentant des méthodes de la littérature pour la détection d'évènements et l'identification d'utilisateur/trices influent·es.

3.2.2.1 Détecter les évènements avec MABED

Problématique. Les utilisateur/trices des médias sociaux partagent, discutent et retransmettent de l'information à propos d'évènements divers – allant d'évènements personnels et/ou banals à des évènements importants et/ou globaux – en temps réel. Le volume sans-cesse croissant amène un phénomène de surcharge informationnelle et il est de plus en plus difficile d'identifier les évènements importants dans ces médias. Par « évènement important » il est ici entendu un évènement réel et susceptible d'être couvert par les médias traditionnels. Cela avait amené à formuler la question suivante : comment exploiter les données des médias sociaux pour détecter automatiquement les évènements importants ? Répondre à cette question avait pour objectif de permettre l'analyse

des évènements, ou des types d'évènements, qui suscitaient le plus d'intérêt chez les utilisateur/trices des médias sociaux – ce qui pouvait être intéressant dans une perspective de veille d'informations, du journalisme de données, etc.

Contribution. Détecter automatiquement les évènements importants à partir des médias sociaux est une tâche complexe, puisque les messages se rapportant à ces évènements sont noyés dans un grand volume de messages sans rapport (*i.e.* du bruit). Nous avons proposé *MABED*, une méthode statistique pour détecter automatiquement les évènements importants à partir du flux de messages publiés, dont l'originalité est d'exploiter la fréquence des interactions sociales entre utilisateur/trices, en plus du contenu textuel des messages. La méthode *MABED* différait par ailleurs des méthodes qui existaient parce qu'elle estimait dynamiquement la durée de chaque évènement, plutôt que de supposer une durée commune et fixée à l'avance pour l'ensemble des évènements. Les expérimentations menées montraient la pertinence de la méthode proposée, notamment en comparant les performances de *MABED* avec des méthodes de la littérature, et démontraient que la prise en compte des interactions sociales entre utilisateur/trices conduisait à une détection plus précise des évènements importants, avec une robustesse accrue en présence de contenu bruité. Par ailleurs *MABED* facilitait l'interprétation des évènements détectés en fournissant des descriptions claires et précises, tant sur le plan sémantique que temporel.

Les expérimentations menées à l'aide de divers jeux de données collectés sur le média social Twitter avaient démontré la pertinence des propositions et mettent en lumière des propriétés qui nous aident à mieux comprendre les mécanismes régissant la diffusion de l'information.

3.2.2.2 Modéliser et prévoir la diffusion de l'information avec T-BASIC

Problématique. Au-delà de la détection *a posteriori* des évènements ayant suscité l'intérêt des utilisateur/trices d'un média social, il est également utile, dans certains cas, de pouvoir anticiper la réaction des utilisateur/trices à un évènement spécifique, *i.e.* anticiper la diffusion de l'information liée à cet évènement. La prédiction du phénomène de diffusion de l'information dans les médias sociaux est une tâche qui a suscité un fort intérêt de la part de la communauté scientifique en fouille de données. Cependant, la manière dont cette tâche était abordée faisait que nous en savions encore peu à propos des facteurs qui sous-tendaient le processus de diffusion. Cela nous a amené·es à formuler les questions suivantes. D'une part, *quels facteurs influent sur la diffusion de l'information dans les médias sociaux ?* D'autre part, *comment prévoir la diffusion de l'information à partir de ces facteurs ?* Répondre à ces questions nous permettait de mieux comprendre le phénomène de diffusion dans les médias sociaux et nous permettait de mieux l'anticiper – ce qui serait utile dans le cadre de la communication de crise par exemple.

Contribution. *T-BASIC* est un modèle probabiliste basé sur la structure de réseau sous-jacente aux médias sociaux pour prévoir la diffusion de l'information, plus précisé-

ment l'évolution du volume d'utilisateur/trices relayant une information donnée au fil du temps. Contrairement aux modèles basés sur la structure du réseau proposés précédemment, la probabilité qu'une information donnée se diffuse entre deux utilisateurs connectés n'est pas constante mais dépendante du temps. Les expérimentations menées ont montré la validité de la procédure d'estimation des paramètres, ainsi que l'intérêt d'avoir des probabilités dépendantes du temps (non constantes), ce qui permet de prendre en compte dans *T-BASIC* la fluctuation du niveau de réceptivité des utilisateurs des médias sociaux au fil du temps. Par ailleurs, nous avons montré comment, et dans quelle mesure, les caractéristiques sociales, thématiques et temporelles des utilisateur/trices affectent la diffusion de l'information.

3.2.2.3 Identifier les utilisateur/trices influent·es avec SONDY

Problématique. Dans la société, différents types de structure (*e.g.* les entreprises, les services gouvernementaux, les journalistes) cherchent à exploiter et analyser les médias sociaux à des fins diverses (*e.g.* analyser la réaction des consommateur/trices à propos de certains produits et les promouvoir, détecter des informations et utilisateur/trices à caractère dangereux, détecter des évènements importants et interroger les utilisateur/trices). Généralement, la démarche qui était mise en œuvre consistait à détecter les évènements animant les discussions des utilisateur/trices, puis à identifier les personnes influentes par rapport à ces évènements, afin de prendre des décisions et éventuellement agir. Pour que cette démarche soit efficace, elle doit reposer sur des méthodes de détection d'évènements et d'analyse de l'influence adaptées au contexte des médias sociaux. Néanmoins, nous avions constaté que, dans le domaine de la recherche, les personnes qui développaient de telles méthodes ne partageaient pas systématiquement leurs implémentations. Ceci amène à une impossibilité de reproductibilité des résultats (ce qui rend difficile la comparaison des approches).

Cela nous amène à formuler les deux questions suivantes. D'une part : comment permettre à des personnes non-expertes d'analyser efficacement des données collectées sur les médias sociaux ? D'autre part : comment favoriser le partage et la réutilisation des implémentations des méthodes nécessaires à cette analyse ? Les réponses à ces questions bénéficieraient autant aux personnes non-expertes ayant besoin d'analyser les données dont elles disposent, qu'au milieu de la recherche, en permettant d'une part de partager les nouvelles méthodes proposées et d'autre part en permettant de réutiliser les méthodes existantes.

Contribution. Adrien Guille a développé *SONDY*, un logiciel libre et extensible qui implémentait des méthodes tirées de la littérature pour la fouille et l'analyse des données issues des médias sociaux. Le logiciel traite deux types de données : les messages publiés par les utilisateur/trices, et la structure du réseau social interconnectant ces personnes. Contrairement aux logiciels académiques existants qui se concentraient soit sur l'analyse des messages, soit sur l'analyse du réseau, *SONDY* permettait d'analyser ces deux types

de données conjointement en permettant l'analyse de l'influence par rapport aux évènements détectés. Utilisé comme logiciel autonome, *SONDY* offrait une interface d'utilisation avancée, accessible aux personnes non-expertes, et des visualisations adaptées. Utilisé comme bibliothèque, il permet d'intégrer facilement les méthodes implémentées dans d'autres programmes, par exemple pour automatiser la comparaison de leurs performances.

3.2.3 Discussion

La figure I.3.1 permet de présenter une synthèse des travaux présentés précédemment.

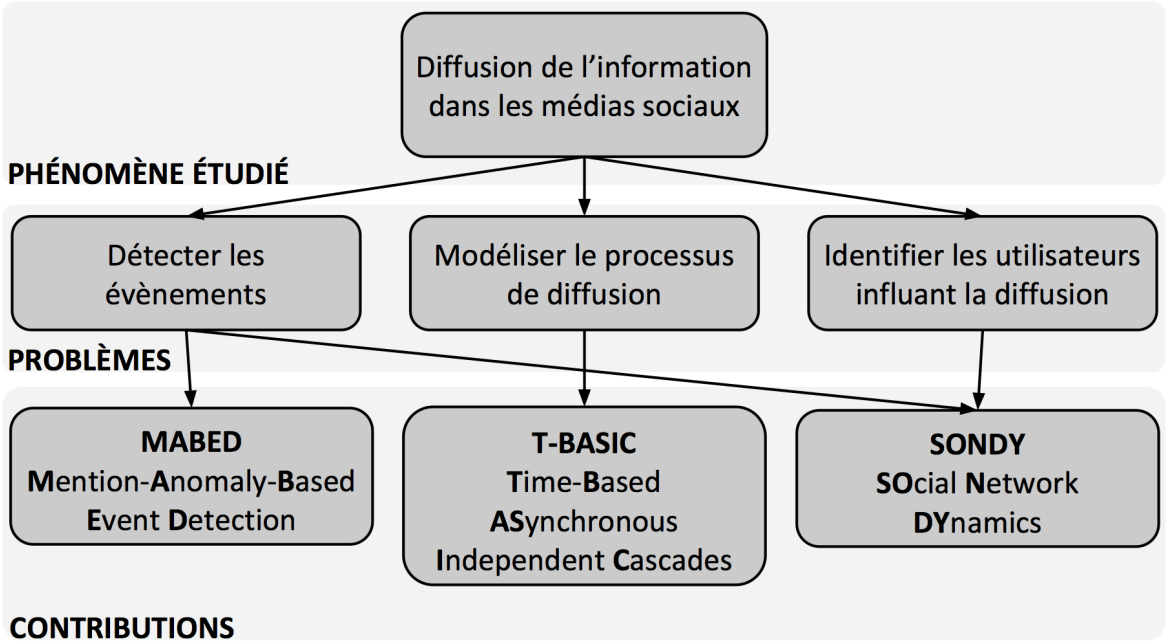


Figure I.3.1 – Structuration visuelle réalisée par Adrien Guille pour la présentation de ses travaux de thèse sur la diffusion de l'information dans les *microblogs*. De haut en bas : le phénomène étudié, les problématiques de recherche et les contributions apportées.

L'état de l'art dans le domaine commençait à être foisonnant, avec une évolution rapide. Nous avons tout de même réalisé un article *survey* (Guille, Hacid, Favre & Zighed, 2013) pour synthétiser l'état de l'art sur ces différents aspects. Il est à noter que malgré une évolution rapide des travaux sur ces sujets, cet article de synthèse a été particulièrement cité (1130 citations directes à ce jour – novembre 2023 – d'après Google Scholar par des papiers qui ont eux-même été particulièrement cités).

Il est à noter que ces travaux, qui ont été menés dans une perspective purement informatique et qui portaient sur des données issues d'un média social, auraient pu bénéficier d'un regard disciplinaire complémentaire. Par exemple une perspective sociologique ou d'information et communication, notamment en lien avec les usages des médias sociaux,

aurait pu être pertinente pour mieux comprendre les facteurs influençant la diffusion de l'information.

En effet, modéliser et prévoir le phénomène de diffusion de l'information à travers les médias sociaux est une tâche ardue en raison de l'intrication entre les dynamiques humaines et les structures sociales de grande ampleur.

Cette diffusion à grande échelle des informations pose alors nécessairement le problème de la diffusion à grande échelle de fausses informations. En effet, cette diffusion de fausses informations peut avoir des impacts importants et graves selon les cas. C'est ainsi que la détection de rumeurs a été traitée dans le cadre de la thèse d'Abderrazek Azri, sujet que nous présentons donc dans la suite de ce chapitre.

3.3 Détection de rumeurs

Permettre la détection de rumeurs dans les *microblogs* constitue un enjeu sociétal important.

Après qu'Abderrazek Azri a effectué un stage de master avec mon encadrement sur des problématiques ayant trait à l'analyse d'images, il fut assez naturel de lui proposer un sujet de thèse, quelques années plus tard lors de son retour en France, qui mobiliserait les connaissances qu'il avait acquises.

Dès lors, aborder la détection de rumeurs en exploitant la dimension visuelle était assez naturelle, en proposant des approches multimodales qui ne se contentent pas bien sûr d'utiliser que le volet image.

Ainsi, ce qui est présenté dans cette section s'appuie sur les travaux de thèse d'Abderrazek Azri, que j'ai co-encadrés avec Jérôme Darmont et Nouria Harbi à partir d'octobre 2018 et qui ont été défendus le 7 juillet 2022 lors de la soutenance, et dont le mémoire de thèse s'intitulait « Approches multimodales d'apprentissage automatique pour la détection des rumeurs dans les *microblogs* » (Azri, 2022).

3.3.1 Données considérées et explicitation des enjeux

Le problème des fausses informations en ligne a suscité une attention croissante de la part des chercheuses et des chercheurs, mais aussi des professionnel·les de l'information et de la communication (comme les journalistes).

Dans son manuscrit de thèse, Abderrazek Azri explicitait, au-delà des aspects contextuels et techniques des plateformes qui favorisaient une prolifération des fausses informations, les motivations financières et idéologiques pouvant être à l'origine de cette diffusion.

Il détaillait également l'ensemble de facteurs relevant d'aspects psychologiques, cognitifs et sociaux pouvant expliquer la difficulté à discerner la qualité des informations et le

pouvoir d'influence des fausses informations et donc la vulnérabilité des utilisateur/trices de médias sociaux.

Pour contrer ces diffusions de fausses informations, initialement, des approches basées sur la détection manuelle ont été proposées par les professionnel·les. Ces approches se sont traduites, notamment, par l'apparition de plusieurs sites de vérification de faits (*fact checking*) ou de plateformes de la vérification participative (*crowdsourced fact checking*).

Mais avec une grande quantité d'informations générée par les réseaux sociaux, la vérification manuelle devient une tâche laborieuse et coûteuse, alors même que la diffusion de fausses informations peut affecter sérieusement la crédibilité des plateformes et engendrer des conséquences désastreuses dans la vie réelle.

Ainsi, au cours du temps, et au fur et à mesure que des travaux étaient développés sur les *microblogs*, différents jeux de données ont été créés pour permettre la validation d'approches d'apprentissage automatique développées pour aider à la détection de fausses informations.

Travailler avec des données de médias sociaux amène à ne pas considérer que le contenu même des messages, en incluant par exemple la dimension sociale.

Par ailleurs, le contenu visuel des publications joue un rôle important dans la diffusion de l'information. À titre d'exemple, dans le cas de Twitter, les statistiques montrent qu'un *tweet* incluant une image obtient 150 % de *retweets*, 89 % de *likes* et 18 % de clics de plus qu'un *tweet* ne comportant aucune image [1]. Par conséquent, une stratégie pour analyser la véracité du message est d'analyser la véracité de l'image jointe au message.

Le fait de vouloir travailler avec une perspective multimodale mobilisant les images a induit des difficultés, par rapport à des jeux de données existant en nombre limité.

Ainsi, nous avons initialement testé nos travaux sur deux jeux de données portant sur la détection de rumeurs et comportant des images. Les statistiques détaillées de ces deux jeux de données sont listées dans le tableau I.3.2.

Tableau I.3.2 – Statistiques des jeux de données MediaEval et FakeNewsNet.

Dataset	Set	Tweets		Images
		Real	Fake	
MediaEval	Ensemble d'entraînement	5 008	6 841	361
	Ensemble de test	1 217	717	50
FakeNewsNet	Ensemble d'entraînement	25 673	19 422	47 870
	Ensemble de test	6 466	4 808	11 968

1. https://www.blogdumoderateur.com/chiffres-twitter/
2. https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2015

MediaEval[2] (Boididou et al., 2015) est collecté à partir de Twitter et comprend les trois caractéristiques : texte, contexte social et images. Il est conçu pour la vérification de la véracité des messages. Le jeu de données comprend deux parties : un jeu de développement contenant environ 9 000 faux *tweets* et 6 000 vrais *tweets* provenant de 17 évènements ; un jeu de test contenant environ 2 000 *tweets* provenant d'un autre lot de 35 évènements liés aux rumeurs. Nous avons supprimé les *tweets* sans texte ni image, obtenant ainsi un jeu de données final comprenant 411 images distinctes associées à 6 225 vrais et 7 558 faux *tweets*, respectivement.

FakeNewsNet[3] (Shu et al., 2020) est l'un des principaux référentiels de détection des fausses nouvelles. Les articles de fausses et de vraies nouvelles sont collectés sur les sites web de vérification des faits PolitiFact[4] et GossipCop[5]. Comme nous étions particulièrement intéressés par les images, nous avons procédé à l'extraction et à l'exploitation des informations relatives aux images de tous les *tweets*. Étant donné que la plupart des algorithmes de classification de l'apprentissage automatique sont conçus en partant de l'hypothèse d'une distribution équilibrée des classes, nous avons choisi aléatoirement 2566 vraies nouvelles et 2587 fausses. Après avoir supprimé les *tweets* sans image, nous avons obtenu 56 369 *tweets* et 59 838 images.

Ainsi, pour répondre au problème du peu de jeux de données disponibles pour la détection multimodale des rumeurs en exploitant notamment les images (ce n'était pas l'objectif premier des deux jeux de données présentés précédemment), nous avons été amené·es à penser la construction d'un jeu de données, dénommé DAT@Z21, qui a fait l'objet d'une publication à la conférence DAWAK 2023 : un jeu de données multimédia volumineux issues de Twitter, avec un étiquetage qui s'appuie sur une vérité terrain collectée à partir d'un site de *fact checking*. Il inclut toutes les caractéristiques nécessaires, à savoir les données textuelles et linguistiques, visuelles, spatio-temporelles, ainsi que des données relatives à l'engagement social et au comportement des utilisateur/trices. Le jeu de données a été rendu accessible à la communauté en respectant la politique de mise à disposition des données Twitter avec les identifiants des *tweets*[6].

Ainsi, l'enjeu était de permettre de détecter des fausses informations avec une perspective multimodale, incluant notamment les images. C'est ainsi que ces différents jeux de données ont été utilisés pour la validation de la pertinence et l'efficacité des approches de détection de rumeurs qui sont succinctement présentées dans la section suivante.

3. https://github.com/KaiDMML/FakeNewsNet
4. https://www.politifact.com/
5. https://www.gossipcop.com/
6. https://git.msh-lse.fr/eric/dataz21

3.3.2 Contributions : des approches multimodales utilisant les images pour la détection de rumeurs dans les *microblogs*

Compte-tenu du succès grandissant des techniques d'apprentissage automatique dans plusieurs domaines et, d'autre part, par les données riches en informations offertes par les sites des réseaux sociaux comme le texte, le contexte social de la diffusion, l'information visuelle, etc., nous avons exploré plusieurs paradigmes d'apprentissage automatique et profond pour prendre en compte ces modalités. Nous avons réservé une attention particulière au contenu visuel, notamment les images, dont le potentiel pour la détection des rumeurs demeurait insuffisamment exploité par les travaux de recherche.

Les travaux réalisés dans le cadre de la thèse d'Abderrzek Azri ont amené la proposition de trois approches pour détecter les fausses informations.

— *MONITOR* : une plateforme de fusion multimodale qui utilise des caractéristiques extraites du contenu textuel du message, du contexte social, ainsi que des caractéristiques des images.
— *Extension ensembliste de MONITOR* : cette extension explore l'utilisation de plusieurs modèles de *metalearning* de l'apprentissage ensembliste.
— *deepMONITOR* : un modèle d'apprentissage profond en utilisant les trois caractéristiques des messages, en l'occurence le contenu textuel et visuel des messages, ainsi que les signaux sentimentaux.

3.3.2.1 *MONITOR*

Nous avons proposé le *framework MONITOR*, qui exploite et fusionne tous les types de caractéristiques des messages (c'est-à-dire le texte, le contexte social et les caractéristiques des images) par des algorithmes supervisés d'apprentissage automatique.

Son organisation générale est présentée dans la figure I.3.2.

Ainsi, nous définissons un message comme un n-uplet de texte, de contexte social et de contenu d'image. *MONITOR* prend les caractéristiques de ces modalités et vise à apprendre un vecteur de caractéristiques de fusion multimodale comme une agrégation de ces aspects du message.

L'originalité de notre approche réside dans le fait que pour les caractéristiques visuelles, nous proposons un ensemble de métriques d'image inspirées du domaine de l'évaluation de la qualité des images (*Image Quality Assessment* – IQA). Nous avons montré que ces métriques ont contribué très efficacement à la vérification de la véracité des messages. Ces métriques estiment le taux de bruit et quantifient la quantité de dégradation visuelle de tout type de modification dans une image. Il est prouvé qu'elles sont de bons indicateurs pour la détection de fausses images, même pour celles générées par des techniques avancées.

Notre approche comporte donc deux étapes principales.

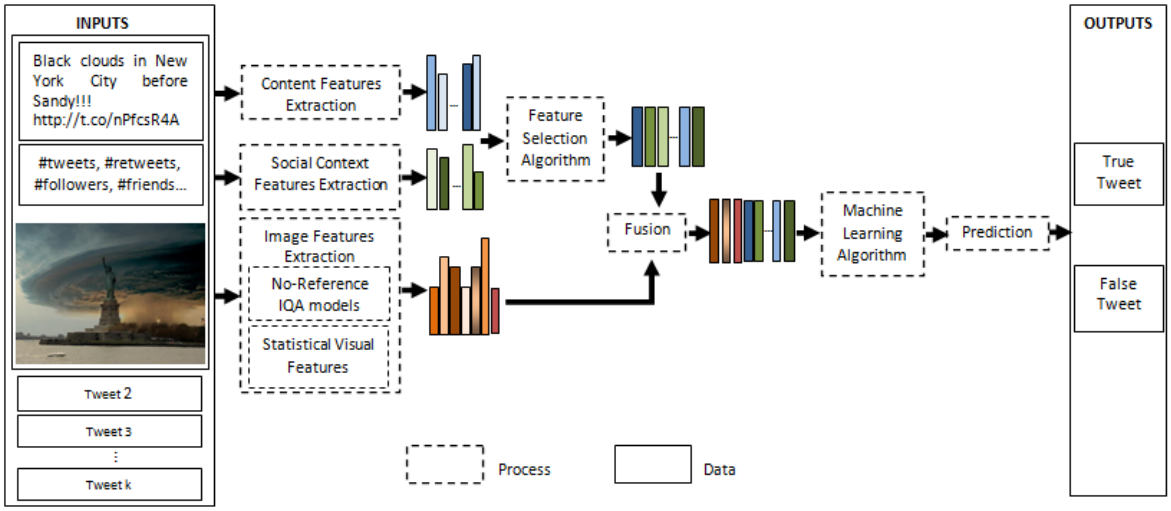


Figure I.3.2 – Aperçu général du *framework MONITOR*.

1. **Extraction et sélection des caractéristiques.** Nous procédons à l'extraction de plusieurs caractéristiques utiles à partir du texte du message et du contexte social, puis nous exécutons un algorithme de *feature selection* pour identifier les caractéristiques pertinentes, qui forment un premier ensemble de caractéristiques textuelles. Ensuite, à partir de l'image jointe, nous élaborons des statistiques et des caractéristiques visuelles efficaces inspirées du domaine de l'IQA, qui forment un deuxième ensemble de caractéristiques d'images.
2. **Apprentissage du modèle.** Les deux ensembles de caractéristiques textuelles et images sont ensuite concaténés, standardisés et normalisés pour former le vecteur de fusion comme étant la représentation multimodale finale du message. Ensuite, plusieurs classifieurs d'apprentissage automatique sont utilisés pour apprendre à partir du vecteur de la fusion afin de distinguer la véracité du message (c'est-à-dire vrai ou faux).

Concernant les caractéristiques visuelles, comme nous ne disposions pas d'une éventuelle version originale de l'image postée par l'utilisateur/trice sur le *microblog*, il s'agissait d'utiliser des métriques d'évaluation sans référence à une image originale. À cette fin, nous avons utilisé trois algorithmes, dont l'efficacité a été démontrée dans des travaux de recherche de l'IQA qui vise à mesurer un indicateur de qualité de l'image.

— Le *Blind/Referenceless Image Spatial Quality Evaluator* (BRISQUE) (Mittal et al., 2011) est entraîné sur une base de données d'images avec des distorsions connues, et est limité à l'évaluation de la qualité des images avec le même type de distorsion. BRISQUE est *opinion-aware*, ce qui signifie que des scores de qualité subjectifs donnés par des expert·es sont associés aux images d'entraînement.
— Le *Naturalness Image Quality Evaluator* (NIQE) (Mittal et al., 2012) est entraîné sur une base de données d'images sans aucune distorsion. Il peut mesurer la qualité d'images présentant une distorsion arbitraire. NIQE est une métrique *opinion-*

unaware, c'est-à-dire qu'elle ne tient pas compte des scores de qualité subjectifs des expert·es.

— Le *Perception based Image Quality Evaluator* (PIQE) (Venkatanath et al., 2015) est un algorithme non-supervisé (ne nécessitant donc pas d'entraînement sur des données étiquetées) et ne tient pas compte des scores de qualité subjectifs de l'expertise humaine. PIQE peut mesurer la qualité d'images présentant une distorsion arbitraire.

Les expérimentations menées sur les jeux de données MediaEval et FakeNewsNet ont permis de montrer l'intérêt de ces caractéristiques visuelles. En effet, sur les deux jeux de données utilisés, les caractéristiques visuelles figuraient parmi les cinq premières caractéristiques importantes pour l'apprentissage. Les autres caractéristiques sont un mélange de caractéristiques de contenu textuel et de contexte social.

3.3.2.2 Extension ensembliste de *MONITOR*

Dans un deuxième temps, nous avons utilisé les approches d'apprentissage ensembliste pour améliorer les performances de *MONITOR*. Pour ce faire, nous avons adapté cinq algorithmes de *metalearning* : *soft voting*, *weighted average voting*, *stacking*, *blending* et un *super learner ensemble*. Ces modèles ensemblistes utilisent les quatre algorithmes d'apprentissage automatique utilisés par *MONITOR* comme modèles de base.

Le terme *metalearning* fait ici référence aux algorithmes d'apprentissage qui apprennent à partir d'autres algorithmes d'apprentissage. Le plus souvent, cela signifie l'utilisation d'algorithmes d'apprentissage automatique qui apprennent à combiner au mieux les prédictions d'autres algorithmes, dans le domaine de l'apprentissage ensembliste.

Les expériences menées sur les deux jeux de données MediaEval et FakeNewsNet ont montré l'utilité de l'apprentissage ensembliste pour la tâche de classification de rumeurs en comparant les performances des cinq algorithmes de *metalearning* avec le meilleur modèle individuel obtenu avec *MONITOR*. Les résultats obtenus montrent clairement que l'approche ensembliste peut améliorer notablement les performances de *MONITOR*.

3.3.2.3 *deepMONITOR*

Nous avons proposé *deepMONITOR*, qui est basé sur des réseaux de neurones profonds pour apprendre les représentations et fusionner les contenus textuels, les sentiments et les images des messages. Un aperçu est donné dans la figure I.3.3.

L'originalité de notre approche réside particulièrement dans l'intégration, en plus du texte, de l'apprentissage multimodal de la dimension visuelle et l'analyse des signes sentimentaux dans un contexte de détection de rumeurs. Ainsi, *deepMONITOR* peut exploiter des informations provenant de différentes modalités et capturer les dépendances sous-jacentes entre le contexte, les émotions et les informations visuelles d'une rumeur.

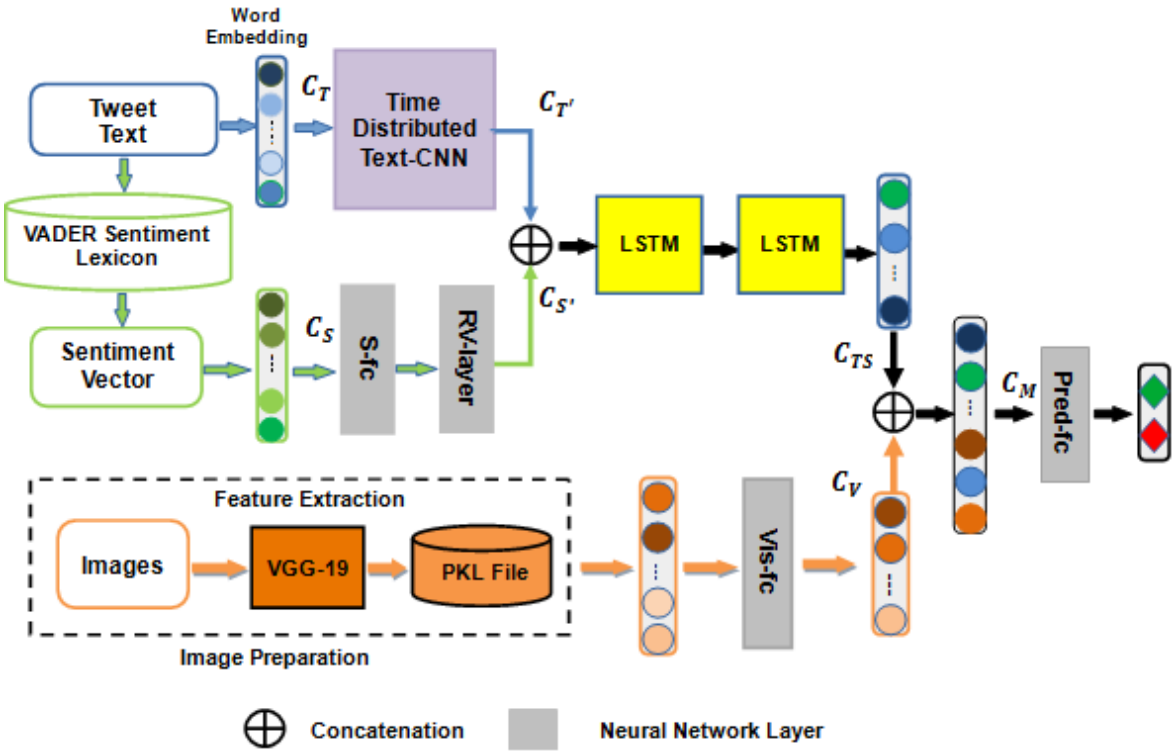


Figure I.3.3 – Aperçu du fonctionnement de *deepMONITOR*.

Concrètement, *deepMONITOR* est un modèle profond multicanal dans lequel nous employons d'abord un réseau de type *Long-term Recurrent Convolutional Network* (LRCN) pour capturer et représenter la sémantique des textes et les sentiments à travers des lexiques des sentiments. Cette architecture combine les avantages du CNN pour l'extraction des caractéristiques locales et la capacité de mémoire des réseaux *Long Short-Term Memory* (LSTM) pour bien connecter les caractéristiques extraites. Ensuite, nous utilisons le modèle pré-entraîné VGG-19 (Simonyan & Zisserman, 2015), fréquemment utilisé en vision par ordinateur pour extraire les caractéristiques visuelles saillantes des images attachées aux messages. Les caractéristiques des images sont ensuite fusionnées avec les représentations conjointes du texte et du sentiment pour classer les messages.

Nous avons montré expérimentalement que *deepMONITOR* surclasse les modèles de l'état de l'art de détection des rumeurs sur deux grands jeux de données multimédia collectés depuis Twitter, à savoir FakeNewsNet et DAT@Z21.

3.3.3 Discussion

Les travaux d'Abderrazek Azri prennent leur place dans un état de l'art varié d'approches de détection des rumeurs qui sont diverses et qui peuvent être catégorisées selon 3 types repris dans la figure I.3.4.

Les approches basées sur la propagation exploitent le cheminement du message, qui

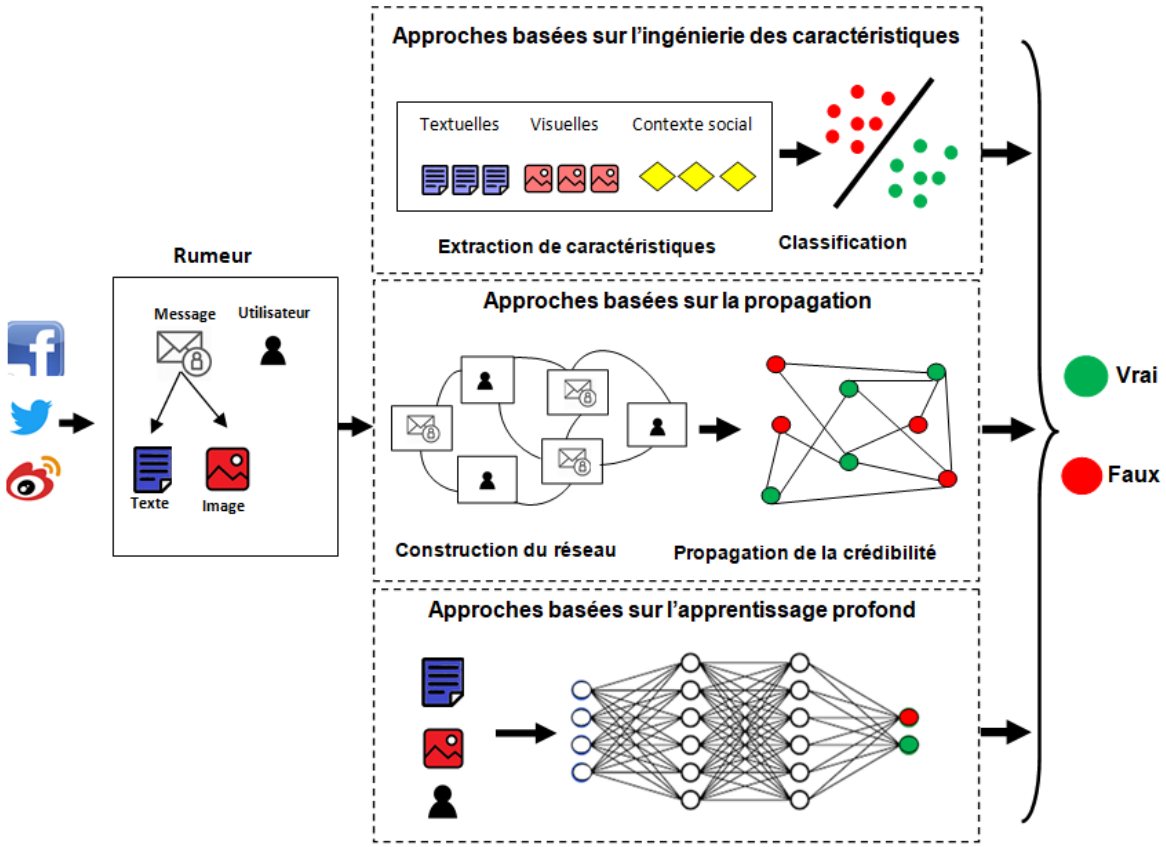


Figure I.3.4 – Familles d'approches pour prédire la véracité des rumeurs.

correspond à son historique, soit toutes les personnes ayant partagé ce message depuis son auteur/trice original·e jusqu'à l'utilisateur/trice associé·e au message à traiter.

La majorité des travaux basée sur l'ingénierie des caractéristiques suivent une approche générale de fouille de données pour la détection des rumeurs qui comprend deux phases : (1) extraction de caractéristiques et (2) construction du modèle. La phase d'extraction de caractéristiques vise à représenter le contenu du message et les informations connexes dans une structure mathématique formelle, tandis que la phase de construction du modèle permet de construire des modèles d'apprentissage supervisé pour mieux différencier les fausses et vraies informations en se basant sur les représentations des caractéristiques. Les caractéristiques peuvent être extraites principalement à partir du contenu des messages et du contexte social. C'est l'approche dans laquelle s'inscrit *MONITOR*.

Contrairement au grand nombre de travaux qui exploitent le contenu textuel pour la classification des rumeurs, le contenu visuel, notamment les images, est peu exploré. Seuls quelques travaux très récents avaient tenté d'extraire des caractéristiques du contenu visuel des messages contenant des rumeurs, dont nous décrivons ici les principaux.

A. Gupta et al., 2013 font un premier effort pour comprendre la diffusion de fausses

images sur Twitter pendant l'ouragan Sandy. Ils proposent un modèle de classification pour identifier les fausses images. Ils affirment que 86 % des *tweets* diffusant des fausses images sont des *retweets*. Cependant, leur travail ignore le contenu des images. Ce travail se base principalement sur des caractéristiques textuelles extraites du contenu des *tweets* et des profils des utilisateur/trices associés. M. Gupta et al., 2012 définissent une caractéristique permettant de noter si l'utilisateur/trice a une image de profil afin d'évaluer la crédibilité des personnes. Wu et al., 2015 utilisent une caractéristique *has multimedia* pour indiquer la présence d'un contenu multimédia attaché à un *tweet* (si le *tweet* est accompagné d'une image, d'une vidéo ou d'un fichier audio). Lors d'une investigation relative à la perception des utilisateur/trices concernant la crédibilité du contenu sur Twitter, Morris et al., 2012 ont découvert que les indicateurs importants sur lesquels les utilisateur/trices jugent la crédibilité sont les informations visibles au premier regard, notamment celles relatives au profil de la personne (son nom et son image), qui ont un grand impact sur la crédibilité des messages publiés par la personne. Ainsi l'extraction d'indice de qualité des images a été un réel apport pour *MONITOR* en comparaison de l'état de l'art.

Contrairement aux modèles basés sur l'apprentissage classique, qui dépendent de caractéristiques élaborées manuellement, les modèles profonds sont capables d'extraire automatiquement des caractéristiques et des représentations cachées dans le texte et les images. Les travaux de ces approches utilisent essentiellement deux structures de réseaux de neurones, les réseaux de neurones récurrents (*Recurrent Neural Networks* – RNN) qui modélisent les données de la rumeur comme des données séquentielles, et les réseaux de neurones convolutifs (*Convolutional Neural Network* – CNN) pour capturer des caractéristiques locales et globales.

Ainsi, *deepMONITOR*, modèle multimodal basé sur les techniques d'apprentissage profond, est composé essentiellement de réseaux CNN et RNN qui permettent d'extraire et de fusionner conjointement les informations de plusieurs modalités des messages.

Les méthodes basées sur la propagation de la crédibilité utilisent la structure hétérogène du réseau social. Les messages et les utilisateur/trices sont reliés dans un réseau entier par des méthodes d'optimisation basées sur les graphes et leurs crédibilités sont évaluées dans leur ensemble. Cependant, il est évident que ces travaux ignorent une composante importante de la rumeur, à savoir l'information textuelle contenue dans les messages.

Les méthodes basées sur l'ingénierie des caractéristiques utilisent ces caractéristiques pour décrire la distribution des rumeurs dans un espace à haute dimension. L'hyperplan de séparation entre les classes est appris par des classifieurs d'apprentissage automatique traditionnel. La phase d'ingénierie des caractéristiques est une étape cruciale dans le processus d'analyse des rumeurs. Elle permet d'intégrer divers types d'informations dans le processus d'apprentissage. Un autre avantage de ces approches est qu'elles sont en mesure de procurer certains éléments d'explicabilité et d'interprétabilité quant aux décisions prises. Nous pensons que de telles explications sont nécessaires, particulièrement

dans le contexte des rumeurs où la vie privée des personnes est en jeu. Cependant, dans le cas où la rumeur manque de certaines caractéristiques discriminantes, ces méthodes conduisent souvent à des résultats instables.

Les approches basées sur l'apprentissage profond exploitent la capacité des différentes structures de réseaux de neurones pour extraire automatiquement et apprendre des caractéristiques complexes des rumeurs. Comparées aux deux familles d'approches précédentes, celles-ci améliorent considérablement les performances de la prédiction. Cependant, l'inconvénient majeur de ces techniques est le manque d'éléments d'explicabilité et d'interprétabilité dans les résultats de classification.

Les travaux de thèse d'Abderrazek nous ont conduit, comme dans tout travail de thèse, vers une évaluation par les pairs des approches proposées. Nous nous sommes beaucoup heurté·es initialement à des rejets de notre approche *MONITOR*, pour le motif que nos premières propositions, basées sur de l'apprentissage « classique », n'étaient pas comparées avec des approches d'apprentissage profond.

Il est important de préciser ici que *MONITOR*, approche basée sur l'ingénierie des caractéristiques, dont l'originalité était de mobiliser des indices de qualité d'image, a manqué à convaincre au départ par l'absence de comparatif à l'apprentissage profond. Ceci peut interpeller par rapport à ce que peuvent apporter les approches plus classiques en termes d'interprétation des résultats. S'agit-il de rendre l'apprentissage profond explicable ou de considérer que l'apprentissage plus traditionnel demeure pertinent ? Pour répondre à cette question, un autre angle d'approche peut également résider dans la considération de la dimension écologique, compte-tenu de la consommation énergétique (Heguerte et al., 2023). Les enjeux d'apprentissage frugal sont réels et les approches en plein développement.

3.4 Réflexions conclusives : quelles contributions sur les médias sociaux avec quelles données ?

En 2015, dans son manuscrit de thèse, Adrien Guille précisait la motivation de travailler avec Twitter en ces termes :
« Deux raisons principales motivent ce choix. Premièrement, l'engouement pour Twitter est un phénomène global qui pousse chaque jour des internautes du monde entier à s'inscrire puis prendre part aux discussions. Par conséquent, Twitter occupe une place sans cesse plus importante dans notre environnement médiatique. [...] Deuxièmement, Twitter – contrairement à la majorité des médias sociaux – permet d'accéder gratuitement à une part importante de ses données, ce qui pousse beaucoup de chercheurs à l'étudier. »

L'intérêt de travailler avec les données de Twitter au-delà de sa gratuité dans le contexte académique était de pouvoir avoir accès à un *microblog* largement utilisé, notamment par rapport à la diffusion d'informations qui circulaient largement via ce média.

Depuis son rachat par Elon Musk en octobre 2022, Twitter, qu'il a renommé X, différentes personnalités, structures - telles que des universités notamment - se sont retirées de ce média. Ainsi, qu'est-ce que cela veut dire de prendre comme source X aujourd'hui dans nos analyses informatiques, compte-tenu des évolutions induites après ce rachat ? Ainsi il apparaît important de se poser la question de l'évolution des travaux qui s'appuyaient sur ce média.

Historiquement, une API pour la recherche académique était disponible. Pour y avoir accès, il s'agissait de compléter un formulaire qui présentait le projet de recherche en créant un compte dédié. C'est ce que nous avions fait dans le cadre du travail sur la détection de rumeurs. Il se trouve que suite au rachat et une nouvelle politique d'accès aux données mise en place, Twitter a supprimé l'API de recherche académique (notre compte dédié a été tout simplement supprimé). Des milliers de projets de recherche dans le monde entier ont donc été mis en péril à cette occasion.

En effet, dans un message du 30 mars 2023, depuis le compte des développeurs, l'annonce des nouvelles API disponibles était faite. Une mention était à destination du monde académique : « For Academia, we are looking at new ways to continue serving this community. Stay tuned to @TwitterDev to learn more. In the meantime, Free, Basic and Enterprise are available for academics. »

Le 19 juillet, un autre message venait repréciser les messages d'erreur qui pouvaient apparaître en raison du changement de droit d'accès.

Il est important de noter que cette API à destination du monde académique permettait une récolte des données sans échantillonnage incontrôlable. Ainsi, cela pose question sur l'avenir des travaux académiques qui prenaient comme source les données de Twitter.

En effet, cela pose dans un premier temps la question de l'usage des jeux de données dans une perspective comparative des travaux. Précédemment, la politique était déjà de ne pas diffuser les données elles-mêmes, mais seulement les identifiants des *tweets*. L'impact est ici alors de ne plus nécessairement être en capacité de reconstituer complètement les jeux de données, à moins de s'inscrire dans une perspective d'achat des données, ce qui amène des discussions sur différents aspects. Cela pose ensuite des questions en termes de poursuite même des travaux.

Peut-être que ces discussions vont émerger ou ont peut-être déjà commencé à émerger au sein même des communautés de recherche, induisant une analyse fine sur les API encore accessibles et ce qu'elles permettent.

Cela témoigne en tous cas de la non neutralité de décisions économiques/politiques sur des sources de données qui prennent place dans la société sur le plan numérique, et que nous traitons d'un point de vue informatique.

4 Apport de l'*OLAP* pour la navigation dans des données de type graphe

Toutes sortes de choses peuvent se produire lorsque vous êtes ouvert aux nouvelles idées et que vous jouez avec les choses.

Citation attribuée à Stephanie Kwolek (1923-2014), chimiste américaine

Contributions sur lesquelles se base ce chapitre

- > Mémoire de thèse de Wararat Jakawat
  - W. Jakawat, Graphs enriched by Cubes (GreC) : a new approach for OLAP on information networks, Université Lumière Lyon 2. Thèse en informatique soutenue le 27 septembre 2016. (Jakawat, 2016)
- > Articles dans des revues internationales
  - W. Jakawat, **C. Favre** & S. Loudcher, Graphs enriched by cubes for OLAP on bibliographic networks, IJBIDM, Vol.11, N°1, 2016, 85-107. (Jakawat et al., 2016a)
  - S. Loudcher, W. Jakawat, E.P. Soriano-Morales, **C. Favre**, Combining OLAP and information networks for bibliographic data analysis : a survey, Scientometrics, Vol.103, N°2, 2015, 471-487. (Loudcher, Jakawat, Soriano-Morales et al., 2015)
- > Publication dans une conférence internationale
  - W. Jakawat, **C. Favre**, S. Loudcher, OLAP Cube-based Graph Approach for Bibliographic Data, SOFSEM 2016, Harrachov – Czech Republic, 87-99. Track Student Research Forum Papers. (Jakawat et al., 2016b)
- > Publications dans des conférences nationales
  - **C. Favre**, W. Jakawat, S. Loudcher, Graphes enrichis par des Cubes (GreC) : une approche innovante pour l'OLAP sur des réseaux d'information, INFORSID 2017, Toulouse, 293-308. (Favre, Jakawat et al., 2017)

    S. Loudcher, **C. Favre**, W. Jakawat, Que peut apporter l'OLAP à l'analyse de réseaux d'informations bibliographiques?, MARAMI 2013, Saint-Etienne. (Loudcher et al., 2013)
- > Publication dans un atelier international
  - W. Jakawat, **C. Favre**, S. Loudcher, OLAP on Information Networks : A New Framework for Dealing with Bibliographic Data, Workshop SoBI 2013 @ADBIS2013, Genoa, Italy, 361-370. (Jakawat et al., 2013)

L'ANALYSE EN LIGNE *OLAP* (*Online Analytical Processing*) développée dans le cadre des entrepôts de données propose une navigation dans les cubes de données en s'appuyant notamment sur des opérateurs comme l'agrégation.

Parallèlement, les données sont parfois organisées en graphes et nécessitent des opérateurs spécifiques lorsque les données sont volumineuses pour permettre une visualisation adéquate du graphe.

Ce chapitre se focalise alors sur la combinaison des données de type graphe avec l'analyse *OLAP* qui ouvre à de nouvelles perspectives d'analyse. Un préambule sera présenté dans la section I.4.1. Nous présenterons ensuite dans la section I.4.2 notre contribution d'approche de *Graph OLAP* : *GreC* pour Graphes enrichis par des Cubes. Nous terminerons ce chapitre par des réflexions conclusives dans la section I.4.3.

4.1 Préambule

Historiquement, l'analyse *OLAP* a été développée et utilisée dans un contexte de données assez classiques, structurées dans des entrepôts de données. L'émergence de nouveaux types de données à considérer, comme par exemple le texte, les images, les réseaux d'information, a soulevé de nouveaux défis à relever pour permettre une extension de cette technologie, entre autres en revisitant les concepts, en recherchant comment transposer ce qui existait à de nouveaux types de données, en développant de nouvelles approches prenant en compte ces nouveaux types de données, et ce pour tirer parti de la richesse de leurs spécificités.

Dans le paysage des données complexes, les réseaux d'information constituent un type de données particulièrement riche compte-tenu non seulement de la multiplicité des données, mais aussi de leurs liens. La modélisation sous forme de graphes avec des nœuds et des arêtes peut prendre différentes formes selon les besoins de représentation : graphe valué ou non pour la pondération des arcs, graphe homogène (un seul type de nœud) ou hétérogène, etc.

Ainsi, dans ce chapitre, nous proposons une approche de *Graph OLAP* combinant les données de type graphe et l'analyse *OLAP*.

4.2 Approche de navigation dans les graphes enrichis de cubes de données

Les travaux synthétisés dans ce chapitre sont basés sur les contributions de la thèse de Wararat Jakawat que j'ai co-dirigée avec Sabine Loudcher, thèse intitulée « Graphs enriched by Cubes (*GreC*) : a new approach for OLAP on information networks » soutenue en 2016, et qui s'est plus particulièrement focalisée sur les données bibliographiques que

nous présentons dans ce qui suit.

4.2.1 Données considérées et explicitation des enjeux

Les travaux de thèse de Wararat Jakawat ont constitué une première manière d'aborder les données de la science sous l'angle de la bibliométrie, données représentées par des réseaux d'informations.

En effet, les données bibliographiques se prêtent particulièrement bien à la représentation sous forme de graphes. Ces données ont d'ailleurs fait l'objet des premières approches qui ont tenté de combiner les graphes et l'approche *OLAP*.

Il apparaît qu'une des caractéristiques importantes des données bibliographiques réside dans le fait que, de par leur nature, elles sont liées entre elles et peuvent donner lieu à une représentation sous forme de graphe. Par exemple, le fait que deux auteur/trices aient publié ensemble induit le fait que sur un graphe d'auteur/trices, si nous nous intéressons à la co-publication, l'arête reliant ces deux personnes pourra être valuée par le nombre de papiers qu'elles ont écrits ensemble. Ainsi, dans la figure I.4.1, J. Han et Y. Sun ont collaboré dans 3 publications.

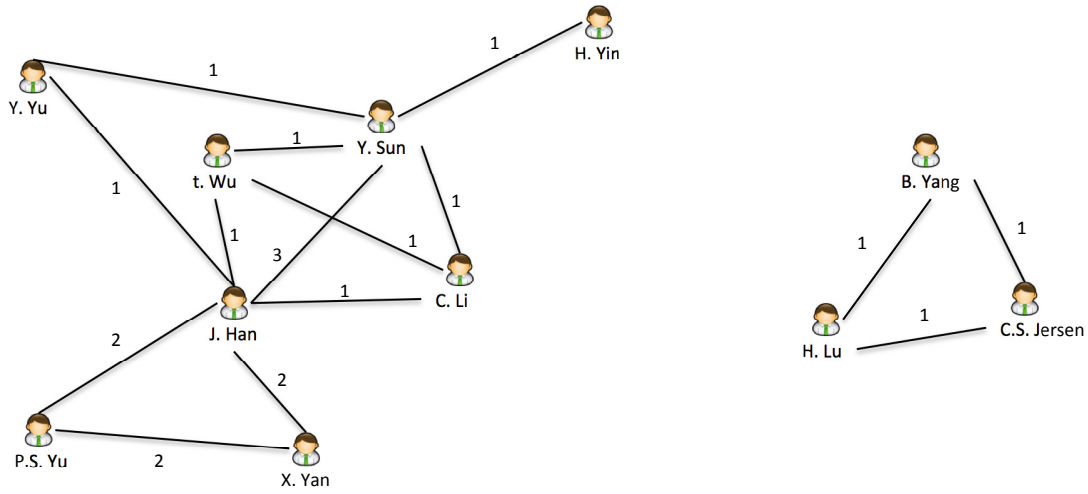


Figure I.4.1 – Graphe d'auteur/trices représentant les co-publications à un instant t.

Nombre travaux ont été développés autour de la visualisation de graphes comme ceux de Heymann et Grand (2013). Néanmoins, dans cet exemple, on peut constater que le pouvoir informatif de ce graphe reste assez faible. En effet, cette représentation ne prend pas en compte la dynamique des données (c'est une photo à un instant t de l'état des co-publications) ; par ailleurs, cette représentation ne permet pas de rendre compte de différentes informations caractérisant les publications dénombrées, telles que l'année, le lieu de publication, la thématique, etc.

Une autre alternative de visualisation pour rendre compte de ces informations correspond à ce qui est proposé par l'analyse *OLAP* avec une représentation multidimensionnelle sous forme de cube. Par exemple, dans la figure I.4.2, il est possible d'analyser le FAIT (objet d'analyse) « production scientifique », au travers d'une MESURE (indicateur) qui est le « nombre de publications », en fonction de différentes DIMENSIONS (axes d'analyse) qui sont ici au nombre de trois : « auteur/trice », « temps » et « lieu » (*venue*).

Ces dimensions peuvent être organisées sous forme de hiérarchies, organisées en différents niveaux de granularité. Par exemple, la dimension « lieu » a une hiérarchie en deux niveaux : un niveau avec le nom du lieu de publication (*venue name*) et un niveau domaine (*area*). La présence d'une dimension hiérarchisée est un élément important de la navigation dans les données. En effet, l'*OLAP* traditionnel propose différentes opérations de navigation dans les données. Parmi les plus utilisées, il y a les opérations qui permettent de naviguer à travers les niveaux de détail des données selon les dimensions hiérarchisées, avec un processus d'agrégation : le *Roll Up* (forage ascendant) qui permet d'obtenir les données à un niveau agrégé et le *Drill Down* (forage descendant) qui fait l'inverse. Il est à noter que dans cette représentation multidimensionnelle qui comporte davantage d'informations, le fait que les auteur/trices soient en lien au travers de ces publications (co-publications) n'apparaît pas du tout.

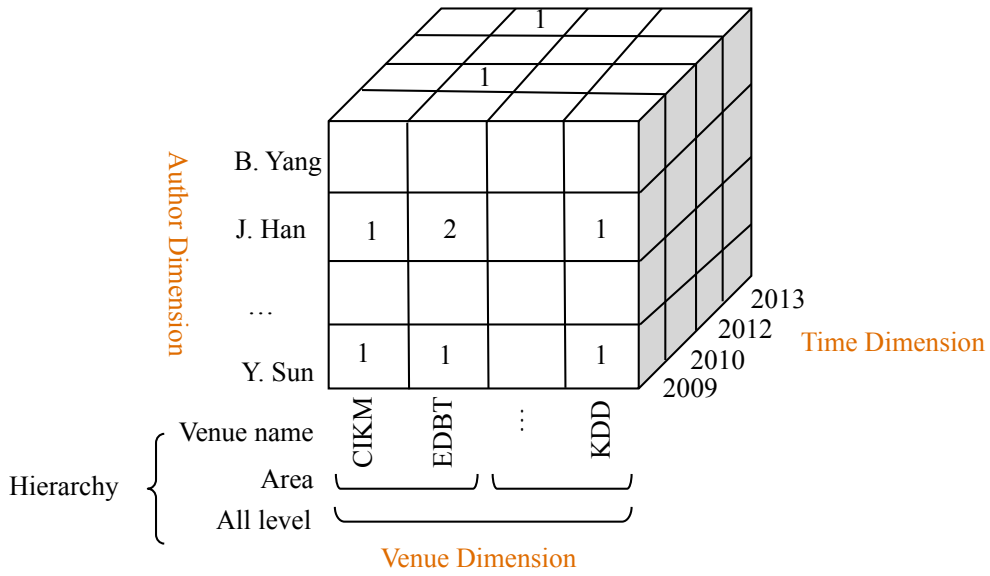


Figure I.4.2 – Structure d'un cube de données dans le contexte *OLAP* pour des données bibliographiques.

L'enjeu est alors de combiner le pouvoir informationnel du graphe, qui met en lumière les réseaux de collaboration, et le détail des informations pouvant être portées par des cubes (comprenant possiblement aussi une dimension temporelle permettant de retracer une dynamique). Ainsi, l'utilisateur/trice doit pouvoir naviguer au sein d'un graphe

enrichi de cubes selon différents niveaux d'analyse, avec des opérateurs dédiés.

Pour permettre la mise en œuvre de l'approche *GreC*, les données de base doivent être modélisées. La figure I.4.3 montre le modèle considéré pour les données bibliographiques. Ce modèle couvre l'ensemble des données prises en compte et intègre les liens entre les données. Ces données de base correspondent à un graphe hétérogène.

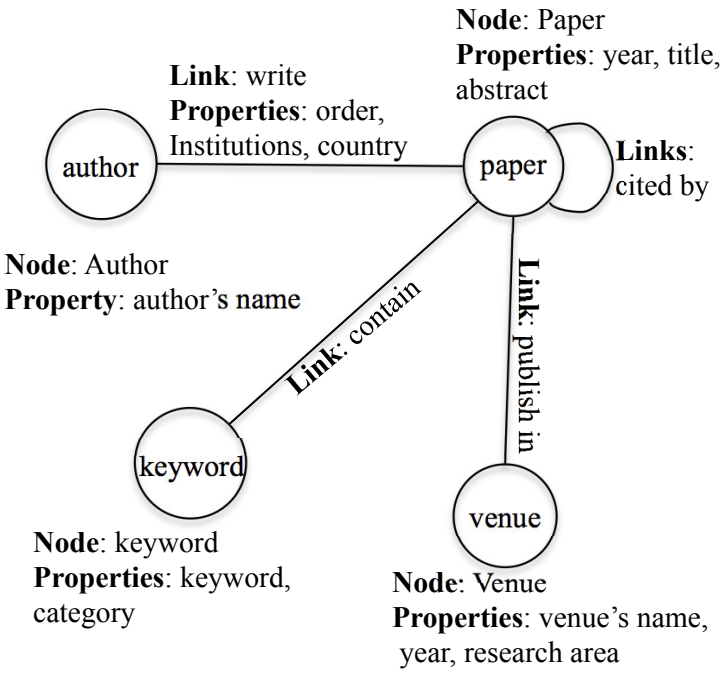


Figure I.4.3 – Modèle du graphe des données bibliographiques de base.

L'implémentation de *GreC* a été réalisée en combinant les données bibliographiques de DBLP, ACM et Microsoft Research Area. Ceci est notamment justifié par la complémentarité des données, en termes de récupération des informations sur les affiliations des auteur/trices des papiers entre autres.

Dans le cadre de l'approche proposée, des jeux de données ont été constitués pour l'expérimentation. Ainsi, quatre jeux de données de tailles différentes ont été constitués avec les caractéristiques décrites dans le tableau I.4.1, tableau présenté ici pour donner un aperçu en termes de volume de données.

4.2.2 Contributions : *GreC* comme nouvelle approche de *Graph OLAP*

Nous avons proposé une nouvelle façon de considérer la combinaison de l'*OLAP* et des graphes en construisant un graphe qui réponde aux besoins de l'utilisateur/trice avec l'enrichissement par des cubes de données pour valuer les nœuds et/ou les arêtes selon les besoins d'analyse. De plus, la présence d'une dimension temporelle dans les cubes

Jeux de Données	Nb de Publications	Réseau de co-auteur/trices		Réseau d'institutions	
		Nb de nœuds	Nb d'arêtes	Nb de nœuds	Nb d'arêtes
D1	1 000	2 216	4 322	696	959
D2	2 000	3 790	8 094	1 157	1 820
D3	3 000	5 335	12 150	1 573	2 711
D4	4 000	7 038	16 107	2 051	3 575

TABLEAU I.4.1 – Jeux de données pour l'expérimentation de *GreC*.

qui valuent les nœuds et/ou les arêtes va notamment permettre de rendre compte de la dynamique du graphe. Par ailleurs, pour enrichir l'analyse, notre attention s'est focalisée sur deux apports : d'une part les types de mesures possibles ; d'autre part les opérateurs de navigation proposés. Notre approche s'intitule donc *GreC* pour Graphes enrichis par des Cubes.

4.2.2.1 Cadre général de l'approche *GreC*

L'approche *GreC* a constitué une nouvelle façon de considérer la combinaison de l'*OLAP* et des graphes pour l'analyse de réseaux d'information. Elle permet de construire un graphe qui réponde aux besoins d'analyse de l'utilisateur/trice et de l'enrichir avec des cubes de données qui vont décrire et valuer les nœuds et/ou les arêtes selon les besoins d'analyse.

L'utilisateur/trice peut ainsi avoir une vue globale d'une partie du réseau avec des informations multidimensionnelles et faire des analyses intéressantes en naviguant au sein du graphe enrichi avec des opérateurs dédiés. *GreC* considère la structure du réseau pour permettre des opérations *OLAP* topologiques, et pas seulement des opérations *OLAP* classiques et informationnelles. La présence d'une dimension temporelle dans les cubes qui valuent les nœuds et/ou les arêtes permet de rendre compte d'une certaine façon de la dynamique du graphe.

La figure I.4.4 illustre le fonctionnement global de l'approche *GreC* appliquée aux données bibliographiques. Le point de départ est la phase préparatoire (couche A), qui correspond au pré-traitement des données. Diverses bases de données bibliographiques sont fusionnées et intégrées dans des fichiers XML qui alimentent une base de données orientée graphe (*PRE-PROCESSING*).

Il s'agit ici de considérer un graphe hétérogène comportant l'ensemble de toutes les données. Ensuite, à partir du graphe hétérogène des données de base, les graphes enrichis par des cubes sont construits (*GRAPHS with CUBES PROCESSING*). Notons que l'ensemble des graphes est calculé en amont pour assurer de bonnes performances pour l'utilisateur/trice.

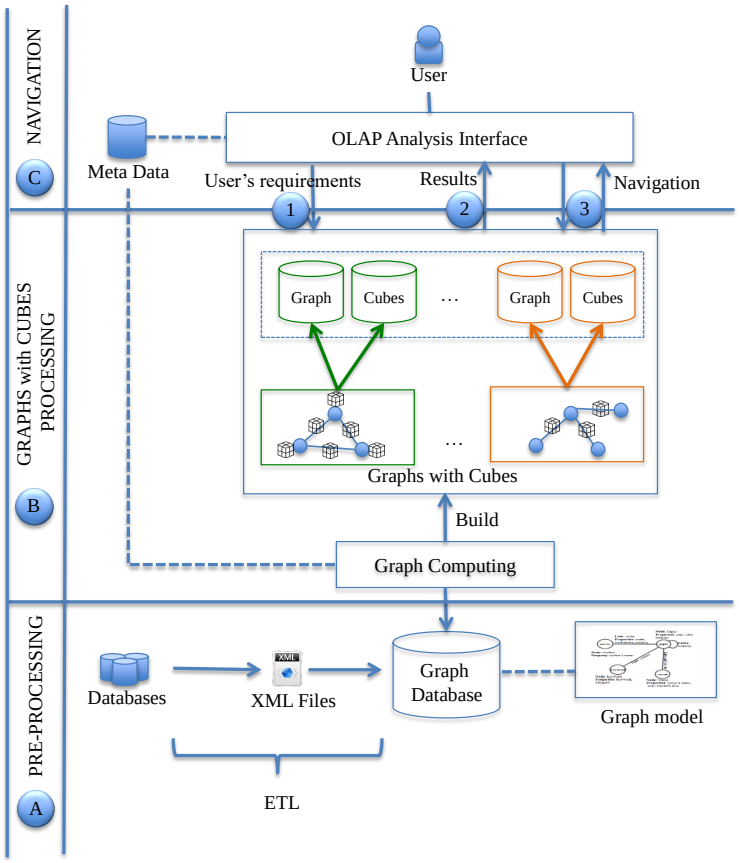


Figure I.4.4 – Processus de *GreC*.

Cette construction se décompose en deux étapes : construction du graphe lui-même, puis celle des cubes de données qui valuent les nœuds et/ou les arêtes. Ces deux étapes se répètent pour construire l'ensemble des graphes enrichis par les cubes.

Puis, grâce à une interface de navigation (*NAVIGATION*), l'utilisateur/trice exprime ses besoins d'analyse, ce qui permet de sélectionner le graphe adéquat. Une fois le graphe adéquat obtenu, l'utilisateur/trice peut naviguer grâce à des opérateurs adaptés, à la fois par rapport au graphe, mais aussi par rapport aux cubes qui lui sont associés.

Pour permettre l'analyse en ligne de graphes enrichis par des cubes ainsi décrite, nous avons été amenées à redéfinir et à étendre les concepts de l'*OLAP* dans le contexte de *GreC*.

Tout comme dans l'approche classique d'analyse en ligne, dans *GreC*, il s'agit d'analyser un fait. Par exemple, dans le cadre des données bibliographiques, il peut s'agir d'analyser la production scientifique ou la co-publication. En revanche, le fait n'est pas directement analysé au travers d'une mesure, mais au travers d'un graphe. En fonction du fait, et des besoins d'analyse, des métadonnées permettent de déterminer si des cubes de données valuent des nœuds et/ou des arêtes.

La notion de cube dans *GreC* correspond à un cube classique, qui contient dans chacune de ses cellules la valeur d'une ou plusieurs mesures numériques ; ces mesures peuvent être « simples » (additives) comme le nombre de publications ou elles peuvent être basées sur des graphes comme par exemple une mesure de degré de centralité.

Nous retrouvons deux types de dimension (comme dans l'approche *Graph OLAP* initiale) : dimension informationnelle et dimension topologique. Les dimensions informationnelles correspondent aux dimensions définissant les cubes de données attenants aux nœuds ou aux arêtes. Les dimensions topologiques correspondent aux dimensions par rapports aux éléments représentés au niveau du graphe, avec dans les deux cas, la possibilité d'une hiérarchisation.

Par exemple, la dimension topologique *auteur/trice* est hiérarchisée avec un niveau *institution*. Ceci permettra de passer du graphe des auteurs au graphe des institutions par exemple. De plus, nous parlons, non pas d'opérateurs *OLAP*, mais d'opérateurs *OLAP* informationnels ou topologiques, déterminant ainsi si l'opération (que ce soit un *Roll Up*, *Drill Down*, etc.) est appliquée par rapport au graphe en question selon une dimension topologique, ou aux cubes de ce graphe selon une dimension informationnelle.

Pour rendre opérationnelle l'approche *Grec* sur l'ensemble du processus, nous avons eu besoin de définir des métadonnées (en plus du modèle de données spécifique à chaque réseau d'informations) et de développer de nouveaux algorithmes pour construire les graphes et les cubes, calculer les mesures et adapter les concepts *OLAP*.

4.2.2.2 Opérationnalisation de *GreC*

Rendre opérationnelle l'approche *GreC* pour les données bibliographiques a nécessité la définition de métadonnées.

À partir du graphe hétérogène complet, pour extraire et construire les graphes enrichis par des cubes ainsi que pour assurer le fonctionnement de l'interface de navigation, nous introduisons des métadonnées, consignées dans le métamodèle simplifié de la figure I.4.5.

En adoptant le formalisme du modèle conceptuel entité-association (EA), les principales entités sont les « faits », « mesures », « dimensions », « hiérarchies », « niveaux » et « graphes ». Ceci a pour but de représenter les contextes d'analyse possibles pour l'utilisateur/trice, en partant des faits à analyser et de comment le faire : avec quel graphe, quels cubes, et ce, grâce à l'instanciation de ces métadonnées.

Les principales associations permettent alors de déterminer quels graphes sont possibles par rapport à un fait à analyser (association entre *FACTS* et *GRAPHS*) ; quels indicateurs sont possibles (association entre *FACTS* et *MEASURES*). L'entité *DIMENSIONS* est associée à la fois à l'entité *GRAPHS* et *FACTS*, ce qui permet de préciser les dimensions topologiques et informationnelles respectivement. L'entité *DIMENSIONS* est également associée à l'entité *HIERARCHY*, associée elle-même à l'entité *LEVELS*, ce qui permet de

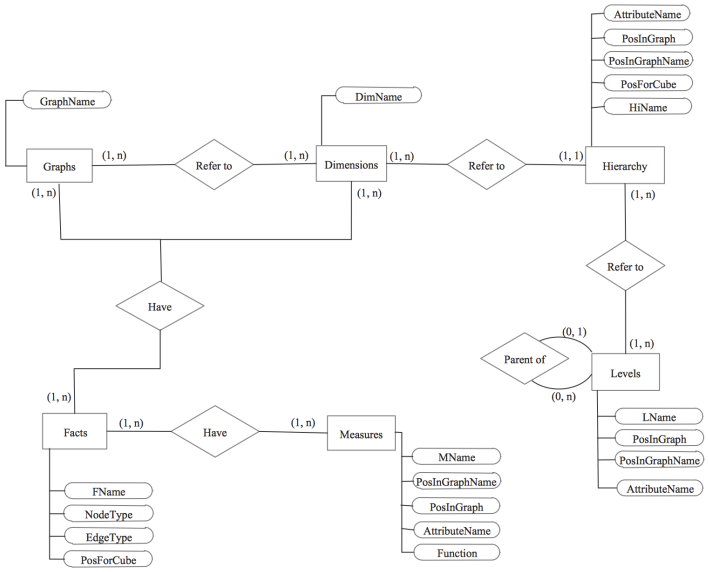


Figure I.4.5 – Métamodèle simplifié de *GreC*.

déterminer les niveaux de navigation. La définition des mesures et des dimensions informationnelles déterminent la structure du cube.

Chaque entité est bien sûr décrite par différents attributs. Le métamodèle ici présenté reprend seulement les attributs ayant un rôle particulier.

Cette modélisation peut s'illustrer sur l'exemple des données bibliographiques en instanciant ces métadonnées :

— Si l'utilisateur/trice veut analyser la co-publication, l'entité *FACTS* permet de tracer le fait que le réseau des co-publications est un graphe où les nœuds sont les auteur/trices et les arêtes entre ces nœuds indiquent que les auteur/trices ont co-écrit ensemble. Cela permet de préciser aussi que ce graphe a des cubes seulement sur les arêtes, puisque l'analyse est centrée ici sur la co-publication et non sur la publication. Si c'est la publication qui était analysée, des cubes seraient à la fois sur les arêtes et sur les nœuds, spécifiant que des publications peuvent être écrites par un·e unique auteur/trice sans collaboration.

— Si le fait est la co-publication, la mesure peut être le nombre de papiers, cela peut aussi être une mesure basée sur le graphe comme le degré de centralité. Dans le premier cas, nous avons donc des cubes au niveau des arêtes, mais pour le degré de centralité qui permet d'estimer l'activité d'un nœud, le cube contenant ce type de mesure se trouverait au niveau des nœuds. Ainsi, dans ce dernier cas, il s'agit pour chaque auteur/trice du graphe d'avoir un cube qui caractérise le nombre total de liens (avec des co-auteur/trices différent·es) en fonction des dimensions du cube choisies, permettant d'analyser les auteur/trices présent·es dans des collaborations variées.

— Si le fait est la co-publication, plusieurs dimensions comme *time* et *venue* peuvent

être utilisées au niveau des cubes qui seront définis. Ainsi, si la mesure est le nombre de papiers pour ce fait, cela induit qu'au niveau des arêtes, il y a des cubes qui déterminent le nombre de papiers co-publiés par année et par conférence.

— Une dimension peut être structurée selon une hiérarchie. Par exemple la dimension institution a une hiérarchie du style : *author name* / *institution name* / *country*, *country* étant un niveau plus élevé de *institution name*. Dans le cadre de l'analyse de co-publication, cette hiérarchie de dimension topologique permettra de faire des opérations pour analyser le phénomène de co-publication à l'échelle des auteur/trices, mais également des institutions, ou même des pays pour analyser l'internationalisation des collaborations scientifiques.

Le métamodèle permet ainsi de faire le lien entre les faits à analyser, les graphes à construire et l'emplacement des cubes (au niveau des nœuds et/ou des arêtes). Il permet également de décrire les concepts *OLAP* (faits, mesures, dimensions, etc.) et de stocker leur instanciation. Les métadonnées sont utilisées par les différents algorithmes qui construisent les graphes et les cubes, qui calculent les mesures et qui réalisent les opérations *OLAP* redéfinies. Elles conditionnent également l'interfaçage de l'application.

Différents algorithmes ont donc été implémentés pour mettre en œuvre *GreC*, en se basant sur l'usage des métadonnées, le graphe initial hétérogène et les besoins d'analyse.

Notons que pour éviter certains problèmes d'additivité, le retour aux données sources est souvent nécessaire pour différents calculs lors de la phase de manipulation des cubes ou du graphe, au travers d'une représentation à base de chemins.

Ce retour aux données sources est important d'un point de vue calculatoire pour deux raisons principales. La première est que cela permet de prendre en compte le fait que lorsqu'un *Roll Up* topologique est fait, les résultats demeurent cohérents. Par exemple, prenons le cas de l'analyse des co-publications avec le nombre de papiers comme mesure. Supposons qu'un papier a été écrit par deux auteur/trices du même établissement, ce papier sera comptabilisé pour les co-publications entre auteur/trices ; si nous passons au niveau des établissements, ce papier ne sera pas comptabilisé car il ne s'agit pas d'une collaboration inter-établissements. La deuxième raison réside dans le fait que nous prenons en compte l'évolution des données, notamment le fait qu'un·e auteur/trice peut changer d'établissement dans le temps. Ainsi, nous nous ramenons toujours à la donnée de base qui est la publication. Il s'agit de fait de pouvoir récupérer l'affiliation indiquée pour l'auteur/trice dans le papier en question. Ainsi, l'affiliation d'un·e auteur/trice qui évolue dans le temps est bien prise en compte dans les différents calculs de graphes et dans les opérations *OLAP* appliquées.

Concernant l'implémentation, les données bibliographiques de base ont été centralisées dans le système NoSQL Neo4j. Les différents graphes correspondant aux différents faits et leurs cubes associés sont générés à partir des données de base du réseau hétérogène et des métadonnées. Les cubes sont ensuite stockés également dans Neo4j. Les métadonnées sont stockées dans le système relationnel Oracle. Les interfaces pour l'utilisateur/trice ont été développées en Java.

4.2.3 Discussion

Afin de tirer parti de l'intérêt des visualisations de graphe et de cube, un nouveau champ de recherche est apparu : *Graph OLAP* (C. Chen et al., 2008).

Le *Graph OLAP* a fait l'objet de plusieurs publications proposant des améliorations et des extensions (Jin et al., 2010; Qu et al., 2011; Zhao et al., 2011). L'idée sur laquelle repose initialement le *Graph OLAP* consiste à construire un cube de graphes dans lequel il est possible de naviguer, grâce à différents opérateurs *OLAP* qui ont été redéfinis pour prendre en compte ce nouveau cadre d'analyse.

Dans ces approches de *Graph OLAP*, il s'agit de considérer des cubes définis selon des dimensions dites informationnelles, et les mesures contenues dans les cellules correspondent, non pas à des indicateurs numériques comme traditionnellement, mais à des graphes ou plus exactement à des sous-graphes.

Par exemple, dans la figure I.4.6, par rapport aux données considérées ici, les dimensions informationnelles sont le temps, le lieu et les mots-clés.

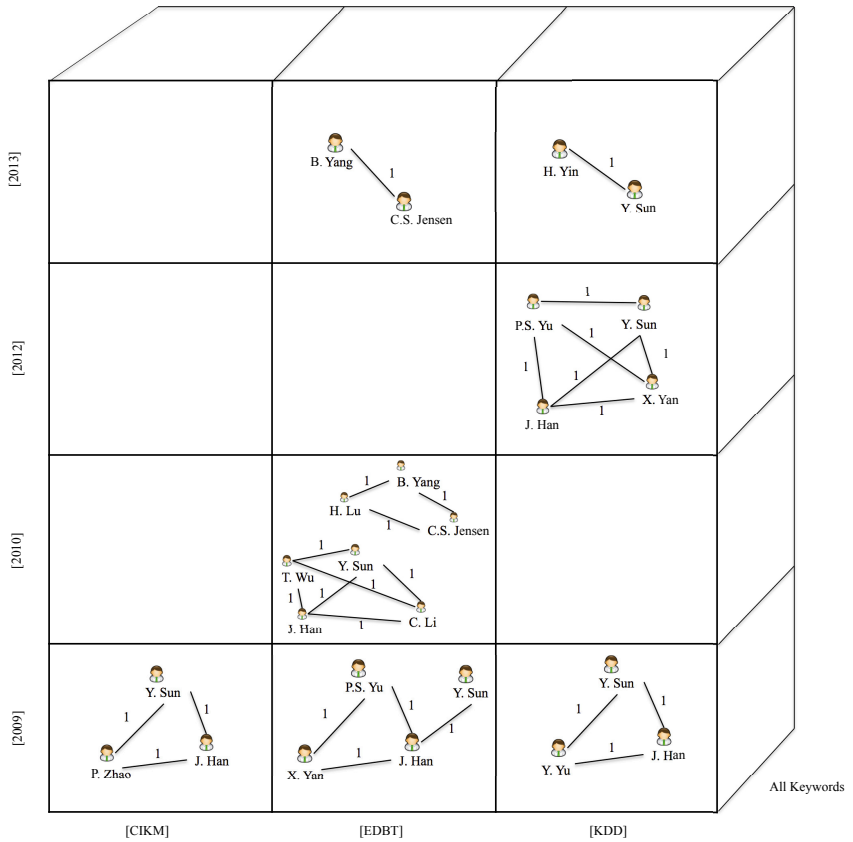


Figure I.4.6 – Cube de graphes sur des données bibliographiques pour analyser les liens de co-publication.

Dans les approches initiales de *Graph OLAP*, au niveau de la modélisation, deux types de dimensions ont été définis : les dimensions informationnelles et les dimensions to-

pologiques. Les dimensions informationnelles vont donc conditionner les manipulations du cube.

Ainsi, lors d'opérations informationnelles sur le cube, les graphes à l'intérieur des cellules vont être recalculés. Les dimensions topologiques, quant à elles, se rapportent à la modélisation des réseaux eux-mêmes dans les cellules. Les opérations topologiques sont caractérisées par un changement du type de nœuds dans les graphes.

Par exemple, à partir d'un réseau d'auteur/trices présent dans une cellule, nous passons à un réseau d'institutions, si nous effectuons un *Roll Up* topologique selon la dimension auteur/trice, dont la hiérarchie comprend un niveau institution.

Initialement, la combinaison de l'*OLAP* et des graphes s'est donc faite au travers de plusieurs approches basées sur des cubes de graphes avec dans leurs cellules, un graphe comme mesure (Beheshti et al., 2012 ; Morfonios & Koutrika, 2008 ; Tian et al., 2008 ; Yin et al., 2012). Ces approches permettent de visualiser des « instantanés » de graphes en fonction des dimensions d'analyse choisies (c'est à dire l'état d'un graphe à un instant donné), grâce à différents opérateurs qui ont été proposés pour naviguer dans le cube de graphes.

Nous avions proposé un état de l'art et une étude comparative de ces différentes approches (Loudcher, Jakawat, Morales et al., 2015). Un des enseignements tirés de ce travail est que dans cette combinaison de l'*OLAP* et des graphes basée sur des cubes de graphes, la visualisation plus globale du graphe est perdue, alors même que celle-ci peut être intéressante d'un point de vue analytique.

Parallèlement, la dynamique des données est importante pour l'analyse du graphe, et ceci n'est pas toujours bien perceptible dans la visualisation des parties de graphe. En effet, malgré la présence d'une dimension temporelle, en croisant celle-ci avec une ou plusieurs autres dimensions, il est difficile de se rendre compte de la dynamique même d'un graphe (évolution des arêtes ou des nœuds).

GreC constitue alors une approche originale et complémentaire, avec un paradigme différent des approches basées sur une construction d'un cube de graphes, puisqu'il s'agit de naviguer dans un graphe qui est enrichi de cubes de données qui décrivent les nœuds et/ou les arêtes du réseau.

4.3 Réflexions conclusives : travailler avec les données de la science, pourquoi et comment ?

Considérer des données de la recherche dans un travail d'analyse, ici dans une perspective de bibliométrie, c'est entrouvrir la porte du questionnement sur la recherche et ses normes de publications.

Quelles données sont considérées ou non ? Qu'est-ce que nous voyons ou pas de cette analyse (éventuels angles morts) ? etc. Autant de questions qui amène finalement à une première étape de réflexivité sur sa propre communauté (en considérant ici des données sur les publications en informatique par exemple), entrouvrir en quelque sorte la perspective de se prendre comme objet de recherche en soi.

Il est important ici de préciser que pousser la porte de l'analyse des données de la science n'induit pas un positionnement en faveur d'une évaluation de la recherche par le quantitatif de manière absolue. La démarche même d'ailleurs ne s'articule pas nécessairement avec une perspective d'évaluation. Pour ma part, il s'agit plutôt d'une démarche compréhensive, et non évaluative.

Ainsi, l'analyse des données bibliographiques dans une perspective de *Graph OLAP* permet d'aller par exemple explorer des dynamiques de collaboration dans le temps.

Cet accompagnement de thèse qui a porté sur des données de la science a ainsi amorcé la volonté de travailler davantage sur ce type de données, ce qui permettait, au-delà de la contribution en informatique, de venir nourrir des réflexions sur notre fonctionnement académique et ses dynamiques.

C'est ainsi que la volonté de travailler sur les données de la science, posant un premier pas sur le chemin de la scientométrie, allait se poursuivre. En effet, j'y ai vu une manière de pouvoir faire appel à l'apport des sciences sociales pour alimenter ces réflexions, et d'envisager non seulement l'informatique comme pouvant aider à l'analyse, mais aussi comme objet de recherche lui-même en tant que discipline étudiée.

L'intérêt de se questionner m'apparaît essentiel, ma conviction étant que nous faisons partie intégrante d'un système académique qui a ses instances d'évaluation, mais considérant que nous faisons partie de ce système pleinement, nous avons aussi des marges d'action pour impulser des manières d'envisager les choses. C'est d'autant plus le cas dans un système où l'évaluation par les pairs est structurante. Par ailleurs, analyser cela au regard des pratiques diverses dans les différentes disciplines est source d'un apprentissage certain.

Il s'agit en tous cas d'avoir une posture claire, selon moi, vis-à-vis de la dérive qui pourrait être induite sur l'évaluation quantitative de la recherche (h-index et autres indicateurs). L'intention est alors clairement posée sur une volonté de questionner, pour mieux comprendre et peut-être développer des moyens d'action sur ce qui poserait problème.

5 Conclusion

> Ce qu'il faut cultiver, ce n'est pas l'idée absurde qu'on a toujours pensé ce qu'on pense aujourd'hui, c'est l'art de se mettre à distance de soi-même, de se regarder comme « de loin », et de comprendre pourquoi on change, et surtout pourquoi on devra changer encore pour être à la hauteur des responsabilités qui seront les nôtres dans le monde de demain.
>
> *Irène Théry (née Noizet, en 1952), sociologue française, entretien réalisé par l'équipe de La Cité doc[a] au sujet du documentaire d'Étienne Chaillou et Mathias Théry « La Sociologue et l'Ourson » (2015)*

a. https://www.lacitedoc.com/entretien-irene-thery

CETTE PREMIÈRE PARTIE a permis de montrer différents aspects liés à l'analyse de données de la société. Ici le terme de données de la société est bien sûr à considérer dans un sens large, émanant de la société, et recouvrant diverses thématiques puisque les travaux portaient à la fois sur des données médicales hospitalières, des données en lien avec l'habitat social, des données issues des médias sociaux et des données de la science.

Les contributions présentées dans cette partie, principalement issues de mes participations à du co-encadrement doctoral, ont permis de témoigner d'une recherche « en largeur », diversifiée, car touchant à différents champs tels que la modélisation en entrepôts et lacs de données, l'analyse *OLAP*, la fouille de données. Cela témoigne de la poursuite d'une recherche marquée par mon cursus de formation double en informatique décisionnelle et en fouille de données.

Ces travaux s'inscrivent finalement en science des données, dans son acception large, si l'on considère que ce terme recouvre tous les aspects du traitement des données, de la collecte, en passant par la modélisation, jusqu'aux connaissances que l'on peut en tirer.

Il est intéressant alors de revenir sur cette expression qui met les données au cœur de l'approche envisagée, ce qui amène finalement une diversité d'approches pour envisager

la manière de faire de la recherche en science des données.

En effet, cette diversité des données amène à des contributions qui peuvent être génériques ou plus spécifiques aux données considérées.

Et il est important pour moi de spécifier que la quête de généricité ne doit pas être absolue, elle peut avoir du sens dans certains contextes, mais ne doit en aucun cas être le critère de mesure de la valeur scientifique du travail. Autrement dit : le caractère générique d'une contribution n'est pas directement lié à la valeur scientifique de cette contribution.

Une pluralité des approches qui articulent données et généricité existe dans les travaux présentés, cette pluralité est à considérer en tant que telle, dans sa nécessité.

Parfois, des contributions vont émerger justement par rapport à une situation spécifique qui émane des données à traiter. Cela peut déboucher sur une contribution qui pourrait être déployée dans d'autres contextes. Ce fut le cas du travail sur la modélisation de hiérarchies de mesure et de dimension contextualisées avec la modélisation à base de satellite. En effet, ce besoin de modélisation a émergé dans le contexte spécifique des données de la santé, mais la solution apportée pourrait se déployer dans un autre champ d'études.

C'est aussi le cas de la proposition de *Graph OLAP* sur les données bibliographiques. En effet, cette approche pourrait par exemple être appliquée dans le cadre de l'analyse de messages issus de Twitter en se focalisant sur le graphe des *followers* enrichi par des cubes informant de l'activité propre en nombre de *tweets* selon le temps, la thématique, etc. ; des cubes sur les arêtes se focalisant sur les mentions entre deux *twittos* renseignant sur la production de *tweets* selon les mêmes axes se restreignant aux mentions des comptes dans les *tweets* produits par exemple.

Il se peut qu'une volonté initiale soit de proposer une contribution générique. Dans ce cas, les données vont permettre d'illustrer un cas d'usage, voire même, la multiplication des cas d'usage peut aider à penser la généricité de la solution. C'est le cas des travaux sur la modélisation de métadonnées pour les lacs de données, avec l'équipe des *Data Lakers*.

Concernant les travaux sur les données de médias sociaux, il est souvent mis en avant le type de données (*microblog*). Pour autant, il est important de considérer les spécificités que peuvent présenter les sources de données utilisées, car des subtilités sont à prendre en compte dans l'analyse. Ainsi, dans les travaux qui ont été développés sur la diffusion d'informations et la détection de fausses informations, ces travaux ont été réalisés sur les données Twitter. Il est alors important de prendre en compte les spécificités de ce *microblog*, par exemple sur le mode de fonctionnement des abonnements qui peut induire des spécificités sur les cascades informationnelles. Certains principes peuvent être généralisés, mais des spécificités sont à prendre en compte, notamment sur l'implémentation, mais aussi sur la manière d'analyser, d'interpréter, les résultats obtenus.

L'ensemble de ces données « de la société » a donc permis l'élaboration de différentes contributions scientifiques en informatique relevant de la science des données, que ce soit en termes de modélisation, d'apprentissage ou de méthodes pour leur exploration.

Dans la partie suivante, nous allons aborder comment l'informatique est une science qui peut ouvrir sur une analyse des dynamiques sociales, notamment en considérant l'informatique comme discipline permettant d'analyser ces dynamiques sociales, mais aussi comme discipline dont on peut analyser les modes de fonctionnement.

Deuxième partie

L'informatique : une science ouvrant sur une analyse des dynamiques sociales

1 Introduction

> Sans le vouloir, j'étais passé de l'ignorance de mon ignorance à la conscience de mon état de conscience. Et le pire aspect de cet état de conscience, c'est que j'ignorais ce dont j'étais consciente.
>
> *Citation attribuée à Maya Angelou, née Marguerite Annie Johnson (1928-2014), écrivaine, poétesse, essayiste, actrice, scénariste, productrice, documentariste (entre autres) américaine.*

Le travail de thèse de Wararat Jakawat présenté précédemment sur le *Graph OLAP* a été l'occasion de faire un pas vers la scientométrie et l'analyse de données de la science.

Un tournant s'est opéré par rapport à l'analyse des données, en allant au-delà de méthodes et d'un outillage innovants, avec une démarche prenant en compte également une dimension marquée par les Sciences Humaines et Sociales.

Ainsi, dans cette partie, je vais présenter la poursuite du travail sur les données de la science, avec pour particularité d'ancrer ce travail dans une forme d'ouverture disciplinaire, notamment sociologique.

L'association internationale francophone EGC (Extraction et Gestion des Connaissances) est un association loi 1901 qui peut être qualifiée de société savante et qui rassemble une communauté de scientifiques travaillant en « science des données », pouvant relever à la fois du domaine de l'informatique et des statistiques.

La conférence EGC a été créée en 2001. Elle a su fédérer autour de son existence une communauté scientifique qui rassemble des personnes dont les intérêts de recherche la définissent. Cette communauté est constituée de personnes qui contribuent scientifiquement à la conférence.

En effet, une conférence est organisée chaque année. Depuis 2016, des données sont mises à disposition par l'association sous forme de défi à relever. Les données sont mises

à disposition plusieurs mois avant la date de soumission des articles pour la conférence. En effet, à partir des résultats obtenus pour les défis, un article est écrit et soumis à la conférence pour évaluation.

En 2016, le défi lancé était intitulé « Communauté EGC : quelle histoire et quel avenir ? » et les données mises à disposition concernaient les articles publiés à la conférence EGC depuis 2004. L'objectif était formulé de la manière suivante : « Les objectifs de ce premier défi sont volontairement ouverts : surprenez-nous ! Le principe est d'appliquer des techniques d'extraction et de gestion de connaissances afin d'expliquer la structure et l'évolution de l'ensemble des données au fil des éditions (thématiques, communautés, atypiques, ...) ».

Dans le cadre d'une collaboration pour ce défi qui a donné lieu à un article (Cabanac et al., 2016), j'avais initié un travail pour permettre une analyse sexuée de la communauté EGC, afin d'apporter une coloration « études sur le genre » pour analyser ces données. Pour ce faire, j'avais procédé à un étiquetage « manuel » des données pour la sexuation, à la fois des auteur/trices, mais en complétant également les données par les membres des comités de programme.

En 2020, à l'occasion de la 20ème édition, l'association a remis en place un défi à partir de ses propres données pour un défi intitulé « Défi EGC 2020 : 20 ans d'histoire pour quel avenir ? ».

À cette occasion, l'association a mis à disposition diverses données :
— les articles publiés à la conférence EGC depuis 2004 ;
— les données issues de DBLP [1] ;
— les messages de la liste de diffusion EGC entre 2006 et 2018 ;
— les données de Twitter et Facebook sur EGC.

Pour ce nouveau défi, dans le cadre d'une nouvelle collaboration, l'envie était d'aller au-delà en matière d'analyse de genre ; et au vu du matériel mis à disposition, d'autres idées ont émergé, avec le fil conducteur d'enjeux sociétaux d'actualité. Les résultats ont donné lieu à un article intitulé « Regards d'actualité au prisme des enjeux sociétaux sur les données historisées d'EGC » (Valat & Favre, 2020) qui a remporté le prix du défi EGC 2020. La manière dont nous avons considéré ces données pour en extraire de l'information présentait une certaine originalité.

Dans l'ouvrage « Qu'est-ce que le travail scientifique des données ? Big data, little data, no data » (plus précisément dans sa traduction de l'ouvrage anglais correspondant), Borgman (2020, p. 82) indique : « Les pratiques en matière de données sont tout aussi variables que les autres aspects du travail scientifique. Mener des recherches s'apprend tout au long d'une carrière et requiert une expertise et une expérience profondes. La théorie, les questions, les méthodes et les ressources influent sur le choix des données pour résoudre un problème. À l'inverse, s'apercevoir que quelque chose pourrait servir de donnée influence la manière de poser les questions et d'appliquer les méthodes. Ces décisions reposent

1. https://dblp.org/

souvent sur un savoir tacite et sur des suppositions qu'il est difficile d'expliciter. ».

Au-delà du contenu scientifique produit dans le cadre d'EGC, EGC désigne aussi une communauté scientifique composée de citoyen·nes de la société, s'insérant dans un environnement social traversé par différents enjeux sociétaux pouvant nous interpeller. Partant de cette réalité et de la demande du défi « 20 ans d'histoire pour quel avenir? », il a paru pertinent de nous focaliser sur comment EGC pouvait se saisir de certains de ces enjeux sociétaux.

En effet, nous ne concentrons pas l'analyse sur le contenu scientifique produit (par exemple en termes d'analyse de l'évolution de thématiques, ou de la dimension collaborative), mais nous faisons le choix d'axer notre analyse au regard de trois enjeux sociétaux qui nous paraissent fondamentaux : le rapport au travail (articulation des temps de vie), l'égalité femmes-hommes et l'écologie, invitant ainsi à une réflexion plus large sur nos postures professionnelles.

L'originalité de ce travail résidait alors dans la manière de se saisir des matériaux et de ce qu'ils renfermaient comme potentiels questionnements, en y ajoutant d'autres matériaux également, permettant donc notamment d'aborder des thématiques sur l'écologie au travers des déplacements des scientifiques, le temps de travail que nous avions abordé dans le défi de manière générale (cette thématique sera amenée dans l'HDR sous l'angle de l'analyse genrée), et enfin l'égalité via une analyse genrée de la communauté scientifique au-delà des auteur/trices et composition des comités de programme.

Parallèlement aux analyses quantitatives sur la sous-représentation des femmes dans l'informatique, j'étais sollicitée depuis plusieurs années à participer à de nombreux (à mon sens) comités de sélection pour des recrutements de maître·sse de conférences.

Ainsi, chaque année, le mois de mai est mis à mal dans son déroulement avec une phase d'évaluation de dossiers et de planification de déplacements pour les auditions, plus ou moins longue selon le nombre de comités assurés. Je précise que la participation à ces comités présente des intérêts certains, à des niveaux différents. Le problème réside plutôt éventuellement dans une sur-sollicitation et ce qu'elle produit, alors que le vivier de femmes est statistiquement ce qu'il est.

La question qui se pose alors est : à quel point une politique qui se veut d'égalité contribuerait finalement à un ralentissement de l'activité scientifique valorisée pendant cette période, produisant ainsi de l'inégalité?

Alors que les questions de genre devenaient de plus en plus présentes dans mes préoccupations, j'ai finalement entrepris une étude exploratoire sur le sujet, dans une une démarche sociologique avec l'aide de Laurence Tain, dont je rendrai compte ici succinctement.

Cette partie est alors structurée de la façon suivante. Le chapitre II.2 reviendra sur la dimension écologique de nos travaux, en abordant la ciculation des savoirs scientifiques, au travers de la mobilité des chercheuses et chercheurs par rapport au déplacement à la conférence de la communauté EGC. Puis, je reviendrai sur l'analyse de la place

des femmes au sein de cette communauté dans le chapitre II.3, soulignant certains aspects ayant trait aux enjeux d'égalité et d'inégalité. Puis seront questionnées les politiques de quotas comme politiques d'égalité, dans une perspective sociologique, dans le chapitre II.4. Enfin,nous conclurons cette partie dans le chapitre II.5.

2 Circulations des savoirs scientifiques et enjeux écologiques

Plus nous concentrons notre attention sur les merveilles et les réalités de l'univers qui nous entourent, moins nous aurons de goût pour la destruction.

Rachel Carson, biologiste marine, écologiste, zoologiste, essayiste américaine (1907-1964)

Contributions sur lesquelles se base ce chapitre

- > Publication dans une conférence nationale
  - S. Valat & **C. Favre**, Regards d'actualité au prisme des enjeux sociétaux sur les données historisées d'EGC, EGC 2020, Bruxelles, 205-216. *A remporté le Prix du défi EGC 2020 « 20 ans d'histoire pour quel avenir ? »*. (Valat & Favre, 2020)
- > Communication (sélection sur résumé)
  - **C. Favre** & S. Valat, Quelles normes scientifiques ? Etude de la production d'une conférence en informatique et statistiques, Colloque de l'Association Internationale des Sociologues en Langues Française (AISLF 2021), Comité de recherche « Science et innovation technologique », Tunis (en distanciel compte-tenu de la situation sanitaire). (Favre & Valat, 2021)

AU FIL DES ANNÉES qui passent, et dans un état d'urgence climatique avéré, l'environnement prend une place de plus en plus importante dans les débats questionnant l'organisation de nos sociétés. Parmi les préoccupations actuelles, on dénote sur un plan politique les émissions de gaz à effets de serre, principalement le CO2.

Dans le domaine de la recherche, une part des déplacements à forte émission est en général attribué aux conférences avec les déplacements nécessaires vers une zone géographique habituellement éloignée de son lieu de travail commun.

Différentes initiatives ou journées de réflexion [1] émergent sur ce sujet et se structurent petit à petit pour questionner la manière d'exercer une profession mais également la dimension écologique de l'informatique plus largement. Par exemple, EcoInfo [2] est un groupement de services du CNRS, composé notamment d'informaticien·nes qui a pour objectif général d'évaluer puis de réduire les impacts au niveau environnemental de l'informatique, notamment au sein de l'enseignement supérieur et de la recherche. Ce groupe a analysé ses propres émissions montrant la prépondérance des déplacements dans leurs émissions en 2019 (EcoInfo, 2019).

Ainsi, après quelques éléments de préambule dans la section II.2.1, je présenterai dans la section II.2.2 le travail mené sur le calcul de l'empreinte carbone de la conférence EGC au fil des années (travail réalisé avec Sébastien Valat qui a pris en charge l'ensemble de la partie technique). Enfin, quelques réflexions conclusives en lien avec le questionnement autour des mobilités des chercheuses et chercheurs viendront clore ce chapitre dans la section II.2.3.

2.1 Préambule

La dimension écologique constitue aujourd'hui un enjeu fondamental, notamment du point de vue du réchauffement climatique. Alors même que la réflexion globale en matière d'écologie numérique est croissante, il s'agit d'amorcer des réflexions vis-à-vis des émissions de carbone, et le transport constitue un des domaines où des actions sont et doivent être entreprises. L'ADEME (Agence De l'Environnement et de la Maîtrise de l'Energie), établissement public sous la tutelle conjointe du ministère de la Transition écologique et de la Cohésion des territoires, du ministère de la Transition énergétique et du ministère de l'Enseignement supérieur et de la Recherche, précise qu'en France il est considéré que les transports sont responsables d'1/3 des émissions de gaz à effet de serre.

Dans le cadre du défi EGC 2020, juste avant la pandémie que nous ne pouvions imaginer, nous avions choisi de traiter des enjeux écologiques sous l'angle des déplacements aux conférences.

1. http://cedd-pes.com/nos-activites-2019/les-chercheurs-prennent-trop-lavion/
2. https://ecoinfo.cnrs.fr/

En effet, nous souhaitons pouvoir disposer d'une estimation en termes de distance et d'émission de carbone liées au déplacement pour la conférence. N'ayant pas accès aux participant·es de chaque édition, il s'agit pour nous d'aller vers le calcul d'un minimum en s'appuyant sur l'hypothèse que le déplacement a été effectué par la personne en première position sur le papier. Cette hypothèse de travail est nécessaire, même si elle peut être forte dans certaines situations où le papier est la production d'une collaboration entre des personnes d'organismes distants géographiquement. L'objectif visé n'est pas ici une estimation exacte, notamment car hormis les auteur/trices des papiers, d'autres personnes participent à la conférence, mais bien d'avoir une première idée des kilomètres parcourus et de ce que cela peut générer en termes d'émissions de carbone « au moins », en lien avec l'historique d'EGC.

C'était bien avant que s'opère la bascule vers des conférences avec un déroulement en distanciel. Ce moment de la pandémie, historique à bien des égards, viendra alimenter les réflexions conclusives de cette partie sur la mobilité des chercheuses et chercheurs.

.

2.2 Calcul de l'empreinte carbone de la conférence EGC au fil des années

2.2.1 Données considérées et explicitation des enjeux

En 2016, lors d'un précédent défi d'EGC sur les données de la conférence elle-même, la publication de Kergosien et al., 2016 mentionnait une distribution des distances parcourues par les participant·es en ayant manuellement rempli leur base de donnée d'auteur/trices et ville.

De notre côté, pour le défi EGC 2020, nous nous proposions non seulement d'automatiser ce processus afin d'obtenir les distances parcourues par les personnes en première position sur les papiers pour l'ensemble du corpus de publications EGC qui était disponible, et d'aller par ailleurs vers une estimation totale minimale d'émissions de carbone.

Pour ce faire, un travail important de préparation a été réalisé en développant un processus le plus automatisé possible. Il s'agissait de partir des documents PDF eux-mêmes pour accéder à l'affiliation des auteur/trices. Ainsi, compte-tenu de la disponibilité des données sur le site de l'éditeur RNTI, cette analyse démarre à partir de 2006 et va jusqu'en 2019.

L'enjeu était d'aller vers une estimation totale minimale d'émissions de carbone en se basant sur des estimations kilométriques par type de transport (exprimées dans l'unité gCO2e/passager.km : grammes CO2 équivalent par passager et par kilomètre). En effet, selon la distance de l'auteur/trice au lieu de la conférence, il est intéressant de prendre en compte que le moyen de transport ne sera pas le même. Par exemple, si l'on consi-

dère le déroulement de l'édition 2010 de la conférence à Tunis, il est assez clair que les participant·es de France ne vont pas y aller en TGV.

Il est entendu ici qu'il s'agit d'une estimation minimale puisque la conférence ne rassemble pas seulement une personne par papier présenté.

2.2.2 Contributions : vers une automatisation du traitement des données bibliographiques pour un calcul d'empreinte carbone

2.2.2.1 Méthode d'extraction de données

L'analyse des données débute avec comme jeu d'entrée les informations DBLP sur les publications des éditions d'EGC et les PDF associés extraits de RNTI.

Afin de calculer la distance de la première personne au lieu de la conférence, il s'agit d'obtenir l'établissement de rattachement. L'extraction du texte a permis d'obtenir un résultat valide pour 700 publications sur les 1390 analysées.

Pour compenser ce manque, nous avons appliqué une méthode additionnelle sur les publications problématiques. La méthode la plus fiable que nous ayons trouvée consiste à générer une image de la première page de la publication et à effectuer une reconnaissance de caractère pour l'astérisque d'affiliation.

Quelques corrections manuelles (300 articles) ont été nécessaires lorsqu'il y avait un mauvais formatage de la part des auteur/trices (nous sous-estimons peut-être parfois l'importance du respect des modèles d'articles !).

2.2.2.2 Gestion des distances

Après avoir obtenu le laboratoire des auteur/trices de la personne en première position sur l'article dont nous considérerons qu'elle est la seule à aller à la conférence, une géolocalisation du laboratoire a été réalisée grâce aux services de Google de manière automatique.

En effet, à ce moment-là nous n'avions pas encore connaissance des travaux de Maisonobe et al. (2018) qui ont permis de développer la méthode de géocodage NETSCITY qui a été conçue pour précisément traiter les affiliations universitaires (Maisonobe et al., 2019).

Le traitement a donné une extraction viable de 916 publications sur les 1390. Nous avons travaillé sur cette base pour construire une estimation minimale.

Les erreurs proviennent soit d'une mauvaise extraction des laboratoires depuis les papiers, soit d'un échec d'extraction de localisation reporté par Google. Ceci représente un total d'une perte de 34% qui est relativement importante mais peut, si on le souhaite, être amélioré en ajoutant des *fix* automatiques sur quelques erreurs communes et récurrentes

Année	Nb papiers traités	Nb papiers publiés	% traitement automatique
2006	84	102	82%
2007	77	92	84%
2008	83	93	89%
2009	74	82	90%
2010	88	115	77%
2011	84	100	84%
2012	44	59	75%
2013	46	56	82%
2014	67	87	77%
2015	46	67	69%
2016	61	84	73%
2017	60	73	82%
2018	53	71	75%
2019	49	69	71%

TABLEAU II.2.1 – Traitement des données pour le calcul de distances sur la période 2006-2019.

dans la méthode d'extraction des laboratoires. Les taux de succès de géolocalisation sont présentés dans le tableau II.2.1.

On peut dans un premier temps observer la distance sommée associée aux publications de chaque année (en considérant la personne en première position sur le papier) et les comparer à la circonférence terrestre qui est d'environ 40 075 km (Moureau & Brace, 2000). Les résultats sont reportés dans le graphique II.2.1.

Toutes les distances supérieures à 800 km ont été vérifiées à la main, on obtient donc un minimum considérant les publications non traitées ainsi que la considération unique de la personne en première position sur le papier.

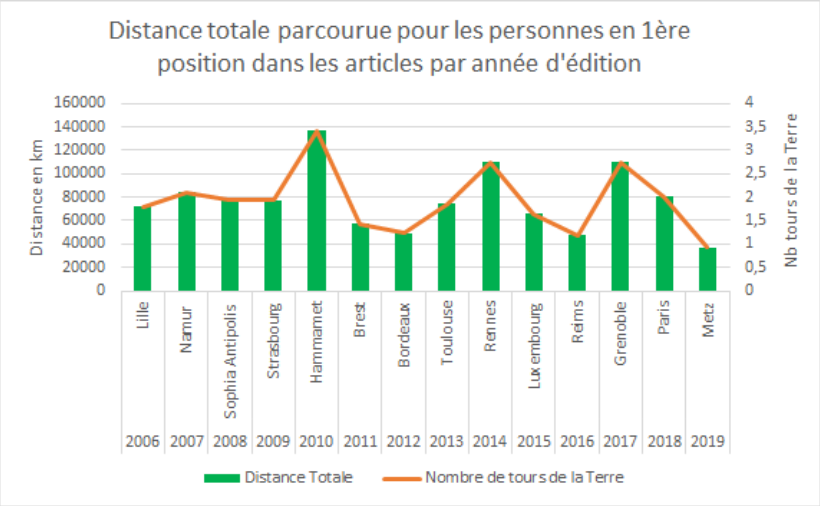


Figure II.2.1 – Distance globale parcourue par édition pour les personnes en première position sur les papiers.

Ainsi globalement, pour l'ensemble des années, nous arrivons à une estimation basse de 1 079 931 km, soit 27 tours de la Terre [3] !

3. Circonférence terrestre ≈ 40 075 km)

2.2.2.3 Estimation de l'empreinte carbone

Pour trouver une estimation minimale des émissions liées aux éditions EGC de 2006 à 2019, il s'agissait de considérer les différents types de transport, qui ont chacun des émissions différentes.

On peut considérer 3 types de transport : le train que l'on a considéré de type TGV (3,69 gCO2e/passager.km, source eco-info/ADEME) pour obtenir une fourchette minimale. Les TER consommant plus (8.91 gCO2e/passager.km) seront ignorés en cherchant une fourchette basse des émissions. Concernant les avions, nous prendrons les bases de l'aviation civile [4] en considérant 2 types de trajets : ceux de plus de 1000km et ceux de moins de 1000km, les consommations à prendre en compte étant différentes selon la catégorie. On prendra donc pour ces deux catégories une moyenne de type Paris/Marseille (117 gCO2e/passager.km) et Paris/New York (82 gCO2e/passager.km) respectivement.

Une estimation plus juste peut être obtenue en utilisant les API publiques et données ouvertes de la SNCF, par exemple la liste des gares et les API d'obtention des horaires fournies par la SNCF, avec les émissions associées. Pour l'instant, cette méthode n'a pu être appliquée que sur 500 des publications et demandent un travail supplémentaire, on se limitera donc à l'estimation minimaliste avec les moyennes par moyen de transport. On considérera le train si une gare est trouvée à moins de 50 km du lieu de travail et du lieu de la conférence ; sinon, le moyen de transport considéré sera l'avion.

Les résultats donnent une estimation basse de 105 tonnes de CO2 équivalent émises pour le total des trajets considérés. Le maximum à Tunis donnant 15 tonnes émises et les années les deux plus basses donnant respectivement 2 tonnes en 2004 (Clermont Ferrand) et 3 tonnes en 2019 à Metz. On notera qu'une prise en compte des trajets réels effectués en train pour les extractions faites avec succès ne change pas significativement ces résultats, les nombres étant dominés par les trajets en avion.

Si l'on considère l'émission moyenne en CO2 en France à environ 9 tonnes par an et par personne (avant la pandémie au vu de la temporalité de notre étude), l'émission en carbone de la conférence correspondait donc à un peu plus de 11 années de cette émission moyenne.

Pour l'édition anniversaire qui s'est déroulée à Bruxelles en 2020, en Belgique, le même calcul a donné un résultat de 53 136 km avec 2,7 tonnes eq. CO2 émises. A noter qu'il avait été envisagé que l'édition se déroule en Guadeloupe, à Pointe-à-Pitre, auquel cas, les résultats auraient été de 465 330 km (9 fois plus que par rapport à Bruxelles) pour une estimation de 38,1 tonnes eq. CO2 (14 fois plus que par rapport à Bruxelles).

4. https://eco-calculateur.dta.aviation-civile.gouv.fr/

2.2.3 Discussion

Nous avons travaillé sur le déplacement en conférence mais ce n'est bien évidemment pas le seul motif de déplacement géographique des scientifiques.

La question des déplacements des scientifiques n'est pas si simple, dans un contexte où l'internationalisation de la recherche sous-tend et valorise un modèle de collaborations internationales et de déplacements associés à ces collaborations et de diffusion des connaissances à cette échelle internationale, notamment par le biais des conférences.

Ce travail est sous bien des angles perfectible mais il a eu le mérite de donner un éclairage sur la mesure minimale en terme d'empreinte carbone d'une conférence francophone, à taille modeste si l'on considère le déroulé de conférences internationales à large échelle sur des sujets similaires à la communauté EGC.

La poursuite de ce travail s'est faite en collaboration avec Guillaume Cabanac dans une proposition de quantification de l'empreinte carbone d'une conférence à partir des données bibliographiques, en envisageant différents scénarios sur « qui va voyager et présenter le travail en conférence ? », afin de ne pas se limiter à l'hypothèse de la personne en première position de l'article qui voyage, et d'évaluer l'impact de cette hypothèse. Ce travail, qui n'est pas à l'heure actuelle encore publié, s'appuie notamment sur une modélisation d'une base permettant de stocker les données produites sur les différents scénarios.

Il est à noter que ce travail s'est poursuivi dans un contexte de pandémie, qui a amené une bascule du déroulement des conférences en distanciel, en particulier dans la discipline informatique. Ceci a ouvert encore plus largement les réflexions sur la mobilité des chercheuses et chercheurs.

Ainsi, j'apporte dans la section suivante différents éléments qui s'appuient sur ce travail encore non publié dans lequel nous avons réfléchi à des points de discussion quant aux enjeux de la mobilité des chercheuses et chercheurs pour la communication de leurs résultats au regard des enjeux écologiques de limitation de ces déplacements.

2.3 Réflexions conclusives : les mobilités des chercheuses et chercheurs

Les réflexions évoquées ici n'ont pas la prétention d'être exhaustives mais permettent de mettre en lumière la complexité des questions soulevées.

Il s'agit de préciser en amont que ces réflexions ont été largement nourries par une réalité qui nous a rattrapé·es durant la pandémie. Et que si ces réflexions étaient d'ores et déjà présentes pour certaines chercheuses et chercheurs, vivre le passage en distanciel des conférences a induit des réflexions qui se sont incarnées dans des réalités que nous ne pouvions imaginer ! En effet, il s'agit aussi d'avoir expérimenté au-delà de réflexions basées sur un récit fictionnel, notamment en termes d'implications sociales !

Il apparaît aussi important de souligner comment le niveau individuel de sensibilisation à ces sujets vient s'entrecroiser avec un niveau plus collectif, voire systémique, au travers de l'institution, et des modes de fonctionnement au sein de la recherche académique.

La pandémie ayant été finalement l'occasion d'expérimentations à grande échelle d'autres modes de fonctionnement, et les enjeux écologiques pouvant s'entrecroiser avec des aspects économiques, des politiques éco(nomiques/logiques) commencent à voir le jour dans certains établissements, dans une perspective par exemple de diminution des mobilités.

La « Responsabilité Sociétale des Entreprises » (RSE) est définie par la commission européenne comme « l'intégration volontaire par les entreprises de préoccupations sociales et environnementales à leurs activités commerciales et leurs relations avec les parties prenantes ». Ainsi on peut s'interroger sur sa déclinaison dans les institutions académiques, ce qu'ont fait Traon et al. (2015) ou encore Barbot et Juban (2018) en revenant sur les concepts de « Responsabilité sociétale d'une Organisation ou d'un Organisme » (RSO), et de « Responsabilité sociétale des Universités » (RSU) pour leurs déclinaisons au-delà du contexte des entreprises, et notamment dans la sphère publique.

Ainsi, les enjeux écologiques ne sont pas à considérer indépendamment des enjeux sociaux. Si nous prenons par exemple la conférence en recherche d'information ECIR, le bilan de ce passage en distanciel évoquait notamment le fait d'avoir beaucoup appris avec un nouveau dispositif, plus distant, mais aussi plus inclusif[5]. Ce sont donc les deux dimensions écologique et sociale que nous abordons dans ce qui suit.

2.3.1 Dimension écologique

Le travail présenté dans ce chapitre a permis de montrer à l'échelle d'une conférence de taille relativement modeste, l'impact des déplacements des chercheuses et chercheurs au niveau environnemental.

L'alternative à cet impact peut sembler évidente grâce à un passage des conférences en distanciel pour éviter ces déplacements géographiques.

Notons qu'il s'agit de recontextualiser par rapport à la discipline, à savoir l'informatique, pour laquelle les conférences internationales présentent parfois une sélectivité sur la base de contributions sous forme de papiers détaillant les contributions (ce qui n'est pas le cas pour toutes les disciplines) plus importante et valorisante que certaines revues internationales.

Bien évidemment, il s'agit d'avoir également en tête que les pratiques disciplinaires étant variées (y compris sur le mode de fonctionnement des conférences et de leur coût d'entrée par exemple), il n'est pas anodin de réfléchir, soit d'un point de vue de la disci-

5. https://irsg.bcs.org/informer/2020/05/42nd-ecir-2020-the-1st-online-ir-conference-an-overview/

pline dans laquelle nous nous inscrivons, soit de manière comparative, avec sans doute des bilans à dresser.

Il est important de pouvoir évaluer non seulement l'économie au niveau carbone, mais également les effets rebonds [6]. En effet, la solution d'un passage en distanciel grâce à la mobilisation du numérique amène des effets qui peuvent limiter les bénéfices au niveau écologique.

Ainsi, l'évaluation de la consommation énergétique liée aux ordinateurs, aux systèmes de visio-conférence, à la connexion, et tout un ensemble d'éléments liés à la mise en œuvre d'une solution en distanciel est nécessaire.

L'initiative Labos 1point5 [7] en France, qui rassemble un collectif de membres du monde académique, de toutes disciplines et sur tout le territoire, s'inscrit d'ailleurs dans cette évaluation, afin de mieux comprendre et réduire l'impact des activités de recherche scientifique sur l'environnement. Les initiatives mondiales pour un monde universitaire plus vert s'efforcent de réduire l'empreinte écologique des activités scientifiques (Quinton, 2020).

Pour atteindre l'objectif sur le plan écologique, considérant qu'il faut réduire les déplacements, notamment les conférences en présentiel, le passage sur un format de conférence en ligne mène à plusieurs questions, mais en particulier peut-être : comment faire à la fois plus distant et plus inclusif, dans une perspective plus sociale ?

2.3.2 Dimension sociale

Il est à noter que certain·es n'ont pas attendu la pandémie pour réfléchir à la question du passage en virtuel des conférences scientifiques ! Par exemple, Malik (2011) présente différentes réflexions.

Mentionnons des travaux visant à comparer les conférences virtuelles et en présentiel (Sá et al., 2019) et qui concluent effectivement sur la réduction des inégalités, notamment de genre, de race, de classe : « the virtual conference dimension may facilitate the academics' participation, reducing the inequalities that take place in the global scientific world resulting from factors such as gender, race/ethnicity or social class ».

Il est en tout cas facile d'avoir en considération que par rapport au coût que représente une publication de papier en conférence (au travers du coût d'inscription à celle-ci), du déplacement et des frais d'hébergement que cela représente, qu'un passage en distanciel peut ouvrir des perspectives à un plus grand nombre quand il est question de déplacements internationaux, avec des contextes économiques pour la recherche qui sont très variables d'un pays à un autre.

Et si l'organisation en distanciel parvient à offrir des possibilités de rencontres, de

6. https://ecoinfo.cnrs.fr/2015/12/23/les-effets-rebond-du-numerique/
7. https://labos1point5.org/

socialisation, au-delà de la « simple » présentation des papiers, comme c'est le cas des conférences en présentiel, alors c'est une dimension importante qui peut-être prise en compte au-delà de la dimension écologique, mais également dans une perspective plus équitable d'envisager la recherche. Car, en effet, il ne s'agit pas de sous-estimer ces espaces de rencontres réelles, d'espaces d'échanges informels (comme la fameuse pause thé/café) qui permettent par exemple des discussions scientifiques poussées, des prises de contact pour des postdocs, pour des montages de projets, etc.

Dans cette perspective, il est intéressant de voir comment des outils ont émergé pour favoriser cette socialisation malgré la distanciation. Ce fut le cas par exemple pour la conférence CAiSE [8] (Schroeder, 2020).

Par ailleurs, concernant la circulation des savoirs, il est intéressant de voir comment dans l'organisation de conférences en distanciel, des choix d'enregistrement de la présentation ont pu être faits pour pallier d'éventuels problèmes techniques. Il s'agit peut-être ici de considérer aussi la persistance de ces présentations, au-delà de la conférence, pouvant avoir un impact positif en terme de visibilité des travaux, si tant est qu'elles soient accessibles au plus grand nombre bien évidemment, et ce dans la durée.

Il s'agit aussi de voir comment, dans une perspective de collaboration dans les papiers, plusieurs personnes puissent plus facilement participer à la conférence. Ce qui peut être particulièrement intéressant, notamment pour les jeunes chercheuses et chercheurs durant leur doctorat.

Si le passage en distanciel peut être vu comme une ouverture du point de vue économique, il s'agit bien de ne pas nourrir une vision manichéenne de la question. Tout comme les solutions sont à discuter également. Puisque, par exemple, le passage en distanciel ne doit pas faire oublier des problèmes de mauvaise connexion amenant à une expérience dégradée, voire à pas d'expérience du tout, de la conférence.

Au vu des inégalités de genre qu'il y a dans la société, et que l'on retrouve dans le domaine académique, on pourrait imaginer que cela facilite la situation pour des femmes qui se déplaceraient moins en raison de contextes familiaux par exemple. Mais là-aussi, la pandémie a été l'occasion de vérifier ces inégalités [9] accrues.

Bien évidemment, cela pose la question d'un point de vue très pratique également des aspects de décalage horaire.

Vardi (2020) invite l'ACM [10] (*Association for Computing Machinery*) à reconnaître l'urgence climatique à laquelle nous sommes confronté·es et à faire sa part pour réduire son empreinte environnementale, puisqu'ACM est impliquée dans nombre de conférences. Point de vue qui amène des discussions, des points de vue différents (Wasserman et al., 2020).

8. https://play.google.com/store/apps/details?id=com.whova.event&hl=fr
9. https://www.natureindex.com/news-blog/decline-women-scientist-research-publishing-production-coronavirus-pandemic
10. https://www.acm.org/

Des positionnements forts existent déjà vis-à-vis de ces enjeux, par exemple sous forme de manifeste « Manifesto : Pledge for sustainable research in theoretical computer science » [11]. Cela émane de chercheuses et chercheurs travaillant en informatique théorique, posant l'objectif de s'engager à réduire les émissions d'au moins 50 % avant 2030 par rapport aux niveaux d'avant 2020. Il est alors intéressant de voir dans cette formulation, comment les choses sont envisagées, en tant que chercheuses et chercheurs (*individual researchers*), en tant qu'organisateur/trices de conférences et d'ateliers (*organizers of conferences and workshops*) et en tant que groupes de chercheuses et chercheurs (*research groups*).

Ces différentes places, depuis lesquelles il est intéressant de placer la réflexion, ne doivent pas faire oublier le contexte plus systémique. Par exemple, la « course » aux publications contribue à accentuer la problématique d'empreinte carbone liée aux déplacements en conférence.

Il est important de voir comment la section CNU 27 (informatique) a récemment (juin 2023) rappelé son évaluation sur la qualité plutôt que sur la quantité des productions scientifiques [12]. Cette quête de qualité rejoint notamment une manière de faire de la recherche qui prendrait sans doute davantage de temps, à l'image de l'appel à un mouvement *Slow Science* (Candau, 2022a) initié en 2012 par un chercheur français en anthropologie, qui en a fait le bilan 10 ans après (Candau, 2022b).

De là à mentionner que réfléchir sur ces questions écologiques peut constituer également un lien avec les enjeux de science ouverte et accessible à toutes et tous, il n'y a qu'un pas que je franchis !

La circulation des connaissances, qui constitue une des perspectives du déroulement des conférences et inhérente à l'internationalisation de la recherche, est sans doute à repenser plus largement !

En définitive, ce travail réalisé sous l'angle écologique, avec l'estimation de l'empreinte carbone d'une conférence et les réflexions qui en découlent, donne à voir certaines dynamiques sociales à l'œuvre dans la recherche en informatique.

La pandémie, moment marquant par rapport aux enjeux écologiques de déplacements en conférence, a aussi été révélatrice d'inégalité de genre. Et la science n'a pas échappé à ce constat. Ainsi différents travaux ont mis en lumière cela comme Myers et al., 2020, à la fois d'un point de vue de travaux de recherche en scientométrie, mais également sur un plan marqué plus politiquement, comme par exemple un rapport émis au niveau de la commission européenne (Commission et al., 2023).

Au-delà de ce contexte, nous abordons justement dans ce qui suit la place des femmes en informatique, toujours en se basant sur le travail mené pour la communauté EGC, nous permettant d'aborder l'analyse de l'informatique du point de vue d'une dynamique du genre.

11. Engagement pour une recherche durable en informatique théorique : https://tcs4f.org
12. https://cnu27.univ-lille.fr/documents/publication-note.pdf

3 Place des femmes au sein d'une communauté scientifique en informatique : où et quand ?

« Qu'est-ce qui leur prend, soudain, aux femmes ? Voilà qu'elles se mettent toutes à écrire des livres. Qu'ont-elles donc à dire de si important ? » demandait récemment un hebdomadaire qui ne s'était jamais posé la question de savoir pourquoi les hommes écrivaient, eux, depuis deux mille ans et ce qui leur restait encore à dire !

Benoîte Groult, journaliste et romancière française (1920-2016), « Ainsi soit-elle » (2006)

Contributions sur lesquelles se base ce chapitre

- > Publication dans une revue internationale
  - S. Amer Yahia, A. Bonifati, **C. Favre**, E. Fromont, N. Labroche, G. Melançon, F. Sèdes, A. Soulet & A. Termier, Diversity and Inclusion Activities in EGC - A 2022 Report, SIGKDD Explor 24(1), 52-56. (Amer-Yahia et al., 2022)
- > Publications dans une conférence nationale
  - S. Valat & **C. Favre**, Regards d'actualité au prisme des enjeux sociétaux sur les données historisées d'EGC, EGC 2020, Bruxelles, 205-216. *A remporté le Prix du défi EGC 2020 « 20 ans d'histoire pour quel avenir ? »*. (Valat & Favre, 2020)
  - G. Cabanac, G. Hubert, H. D. Tran, **C. Favre** & C. Labbé, Un regard lexico-scientométrique sur le défi EGC 2016, EGC 2016, Reims, 419-424. Papier court. (Cabanac et al., 2016)
- > Publication dans un atelier international
  - G. Vargas-Solar, T. Cerquitelli, A. Montorsi, S. Salvai, J. Darmont, **C. Favre**, Promoting equity, diversity and inclusion : policies, strategies and future directions in higher education, research communities and business, 2nd International Workshop on Data science for equality, inclusion and well-being challenges (DS4EIW@BigData 2022), Osaka, Japan, 4710-4718. (Vargas-Solar et al., 2022)
- > Publication dans un atelier national
  - **C. Favre**, Les données de la recherche vues au travers des lunettes du genre : quand l'informatique rencontre les sciences humaines et sociales pour rendre visible le non visible, Atelier VADOR 2017 @INFORSID2017, Toulouse, 58-66. (Favre, 2017a)
- > Communications (sélection sur résumé)
  - **C. Favre**, G. Cabanac, G. Hubert & C Labbé, Du bon usage de l'interdisciplinarité pour l'analyse de données sexuées : le cas des données bibliographiques de la conférence EGC, édition 2017 des journées Big Data Mining and Visualization d'EGC - Regards croisés sur les data, Lille. (Favre, Cabanac et al., 2017)
  - **C. Favre**, Les statistiques sexuées comme miroirs des inégalités de genre : apports et limites de leur production en démocratie, colloque en sciences politiques SCOPE 2017, Bucarest, Roumanie. (Favre, 2017b)
  - **C. Favre**, Femmes et recherche en informatique : d'une analyse sexuée d'une communauté scientifique aux questions de genre, colloque en sociologie AISLF 2016, Montréal, Canada. (Favre, 2016)

LA COMMUNAUTÉ EGC (Extraction et Gestion des Connaissances), reliée à la conférence nationale du même nom, lançait donc en 2016, comme évoqué plus tôt, son premier défi [1], sous la responsabilité de Christine Largeron. Ce premier défi consistait en la mise à disposition des données concernant la communauté avec pour objectif de répondre à la question suivante : « Communauté EGC, quelle histoire et quel avenir ? ». Les données étaient rendues disponibles en amont, et les résultats devaient faire l'objet d'un article soumis à la conférence.

Nous avons « monté une équipe », composée de membres rassemblant plusieurs laboratoires de recherche, à savoir le LIG (Grenoble), l'IRIT (Toulouse) et ERIC (Lyon), pour tenter de relever le défi, en proposant « Un regard lexico-scientométrique sur le défi EGC 2016 ». Ce défi fut pour moi l'occasion de proposer une première perspective de genre pour analyser ces données.

Pour l'édition 2020, avec Sébastien Valat, nous lancions un travail pour répondre au défi organisé pour l'édition anniversaire des 20 ans de la conférence. À cette occasion, le défi revenait avec des données concernant la communauté (après avoir porté durant quelques années sur des données diverses) et s'intitulait « Défi EGC 2020 : 20 ans d'histoire pour quel avenir ? » [2], organisé sous la responsabilité d'Arnaud Martin.

Compte-tenu de l'angle adopté pour ce défi, nous intitulions notre réponse : « Regards d'actualité au prisme des enjeux sociétaux sur les données historisées d'EGC ». La présentation de ce chapitre se focalisera sur les travaux de 2020, puisqu'ils sont une extension à la fois des données considérées (4 années de plus) et des axes d'analyse développés par rapport au défi de 2016.

Le chapitre est organisé de la façon suivante. Un préambule est présenté dans la section II.3.1. Puis j'évoquerai les données considérées pour ce travail dans la section II.3.2. Ensuite, dans la section II.3.3, je présenterai les résultats de notre travail scientométrique qui va au-delà d'un travail bibliométrique d'analyse des publications (qui sont un des matériaux du défi). Des éléments seront ensuite discutés dans la section II.3.4. Finalement, des réflexions conclusives seront apportées dans la section II.3.5 afin d'aborder ce qu'il en est au-delà des constats de la place des femmes dans la communauté.

3.1 Préambule

Dans un contexte de généralisation de l'accessibilité des données, avec des initiatives au niveau de la recherche elle-même comme par exemple sur le plan européen avec OpenAIRE [3], l'informatique s'inscrit dans la dynamique de contribuer à la scientométrie, étude quantitative de la science et de l'innovation (Leydesdorff & Milojević, 2015). Il s'agit

1. http://www.egc.asso.fr/manifestations/defi-egc/defi-egc-2016.html
2. https://www.egc.asso.fr/manifestations/defi-egc/defi-egc-2020-20-ans-dhistoire-pour-quel-avenir.html
3. https://www.europeandataportal.eu/sites/default/files/2014_open_research_in_europe.pdf

notamment de faire face à l'enjeu du traitement de données de la recherche pouvant être à la fois volumineuses et complexes.

De nombreux travaux ont émergé en scientométrie, notamment sur l'exploitation de la production scientifique (caractérisant le champ de la bibliométrie), constituant différents miroirs de la science, et ce dans de nombreuses disciplines telles que la sociologie des sciences, les sciences de l'information, l'histoire des sciences, etc.

Parallèlement, les études de genre constituent elles aussi un domaine de recherche étudié par des disciplines multiples. Le terme de « genre », qui a été largement controversé au travers d'un emploi inapproprié du concept, renvoie notamment à la notion de rapports sociaux de sexe, induisant sémantiquement le fait qu'il y a un rapport hiérarchique entre les sexes qui s'est construit socialement.

Les études de genre impliquent l'usage d'un certain nombre de concepts. Quand bien même il existe différents courants de pensée, et différentes manières d'appréhender ce concept, il nous paraît surtout important de pouvoir revenir, dans le cadre de ce travail, sur la différence entre le concept de sexe et celui de genre [4]. Bereni *et al.* précisent que « dans un premier temps, le « genre » a été distingué de la notion commune de « sexe » pour désigner les différences sociales entre hommes et femmes qui n'étaient pas directement liées à la biologie [...] Cette « dénaturalisation » est un enjeu politique majeur : si l'invocation de la « nature » sert souvent à justifier les inégalités, la mise en avant de l'« histoire » contribue au contraire à rendre ces inégalités plus arbitraires aux yeux de ceux qui les subissent, et facilite ainsi leur remise en cause » (Bereni et al., 2008).

Utiliser les lunettes du genre, ce n'est, en aucun cas, nier les différences biologiques entre femmes et hommes. Utiliser les lunettes du genre, pour l'analyse de données sexuées, c'est remettre en perspective qu'au-delà d'utiliser la variable sexe, il y a une construction sociale. Comme a pu l'énoncer Simone de Beauvoir dans le Deuxième sexe, « On ne naît pas femme, on le devient » (de Beauvoir, 1949). Faire des analyses sexuées induit ici (et plus largement) une bi-catégorisation femmes-hommes, elle-même sujette à discussion voire controverse. Ceci est notamment dû au fait qu'un pourcentage non négligeable d'enfants naisse avec une indétermination au niveau du sexe : les personnes dites hermaphrodites. Fausto-Sterling et al. (2013) précise que le pourcentage des sujets dits hermaphrodites est estimé à 4% de l'humanité.

Mentionner ces éléments nous permet d'indiquer qu'une analyse sexuée avec une bi-catégorisation femmes-hommes est utilisée ici par rapport à la mobilisation des données accessibles, ce qui n'a pas pour but de remettre en cause une perception plus globale de cette question. En effet, adopter une démarche de quantification sexuée peut renforcer cette bi-catégorisation sexuée. Mais elle est utilisée ici, dans une perspective de genre, pour l'observation d'inégalités femmes-hommes. Ceci rejoint la perspective des études de genre qui amène une dimension politique, et ce de façon historique, notamment d'un point de vue féministe.

4. http://www.ecoledugenre.com

Avant tout, pour recontextualiser l'analyse qui va suivre, il s'agit de préciser que la communauté EGC concerne des personnes qui seraient plutôt catégorisées, selon la nomenclature CNU, en sections 27 (informatique) et 26 (mathématiques appliquées et applications des mathématiques). D'après les fiches démographiques disponibles par rapport à notre étude menée en 2019, fournies par le ministère [5], la section 27 comprenait 24% de femmes et la section 26 comprenait 27% de femmes (Maître·sse de Conférences - MCF - et Professeur·e des Universités - PR - confondu·es). Ceci peut servir de base de comparaison sur la composition en matière de représentation sexuée, positionnant ainsi les résultats en perspective du « vivier » d'enseignant·es-chercheur·es, même si les statuts de chercheur·es (CR et DR) ne sont pas considérés dans ces données du ministère.

3.2 Données considérées et explicitation des enjeux

Parmi les données que l'assocation EGC a mises à disposition pour le défi, nous avons retenu les suivantes pour les analyses présentées dans ce chapitre qui interroge la place des femmes dans la communauté :

— les articles publiés à la conférence EGC depuis 2004 ;
— les données issues de DBLP [6] (Ley, 2002) ;
— les messages de la liste de diffusion EGC entre 2006 et 2018 ;

Notre objectif était de pouvoir exploiter le maximum de données pour traiter les axes que nous avions choisis, allant au-delà donc des données fournies. La période considérée pour chacun des axe n'est pas systématiquement la même en fonction de la disponibilité des données.

Sur la dimension de la place des femmes, nous avons œuvré manuellement, pour compléter les données fournies, avec des données du site *Web* de l'association EGC et des sites *Web* des éditions. Nous avons récolté ainsi les personnes ayant présenté les conférences invitées, la composition des comités de programme et les différentes présidences : les personnes invitées à la présidence d'honneur et celles qui ont présidé les comités de programme et d'organisation, nous permettant de faire une analyse de 2001 à 2019 inclus.

Un important travail a été réalisé pour procéder à un étiquetage manuel de la variable sexe. L'objectif était d'avoir des données précises, là où les algorithmes d'étiquetage automatique dédiés ne produisent pas un résultat à l'unité près. Ainsi, ce travail s'est appuyé sur un temps de recherche important sur les pages *Web* professionnelles des personnes quand elles existaient, mais également - et en dernier recours - des dictionnaires de prénoms. Ceci a trait au questionnement dans d'autres disciplines de comment « s'affiche » le sexe sur le *Web* (exemple d'un site référençant les professionnel·les et utilisant des photos types lorsque celle-ci n'est pas fournie par la personne elle-même avec un col en V pour

5. https://www.enseignementsup-recherche.gouv.fr/cid85019/fiches-demographiques-des-sections-du-cnu.html
6. http://dblp.uni-trier.de/xml

les femmes et une cravate pour les hommes).

Pour la dimension analyse de l'articulation des temps de vie avec l'analyse des heures d'envoi des mails, ceux fournis allaient du lundi 10 juillet 2006 au dimanche 9 septembre 2018, ils sont au nombre de 7 075 provenant de 811 adresses mails différentes.

3.3 Contributions : un regard pas seulement bibliométrique!

3.3.1 Auteur/trices

Sur 19 éditions, parmi les co-auteur/trices des papiers (toutes catégories confondues qui sont présentes dans les actes), nous trouvons 27% de femmes en moyenne.

Dans la figure II.3.1, nous pouvons observer l'évolution de cette répartition sexuée en nombre et en pourcentage. Le pourcentage a oscillé entre 21 et 31%. Ainsi, nous pouvons considérer que la communauté EGC est dans la moyenne de ce qui s'observe en termes de répartition sexuée des sections CNU concernées.

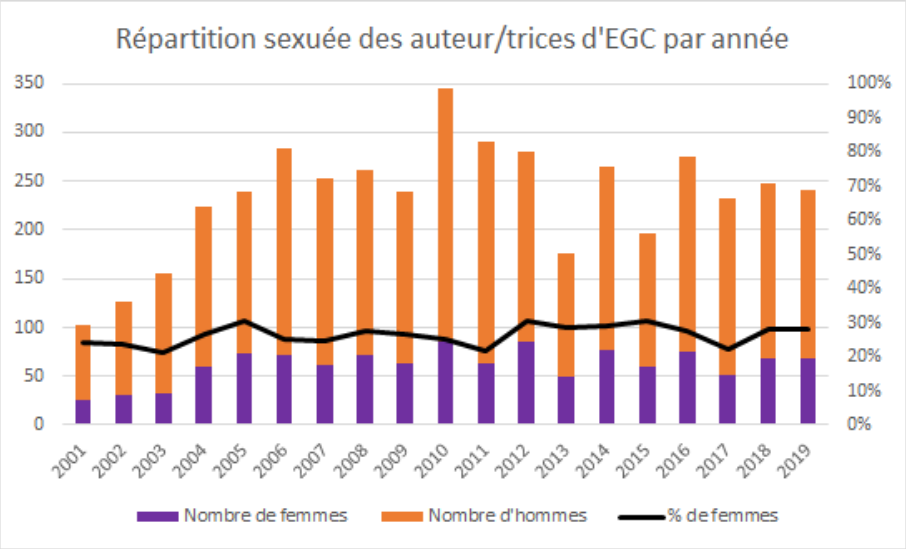


Figure II.3.1 – Représentation sexuée des auteur/trices sur la période 2001-2019.

Considérons à présent les personnes qui ont le plus contribué à la conférence sur les 19 ans, en nombre de papiers (quelle que soit la taille des papiers et moyennant les choix d'édition de certaines années [7]). Dans la figure II.3.2, nous pouvons observer la représentation sexuée de ces personnes ayant le plus contribué à EGC. Si l'on prend en compte les personnes ayant au moins 10 papiers sur les 19 éditions, elles sont 48, dont 12 femmes

7. Par exemple, lors de certaines éditions, les actes n'ont pas comporté les papiers « démonstration de logiciel »

(soit 25%). Nous restons donc sur les mêmes ordres de grandeur, ce qui montre que les femmes font bien partie aussi des personnes contribuant de façon importante et régulière à EGC.

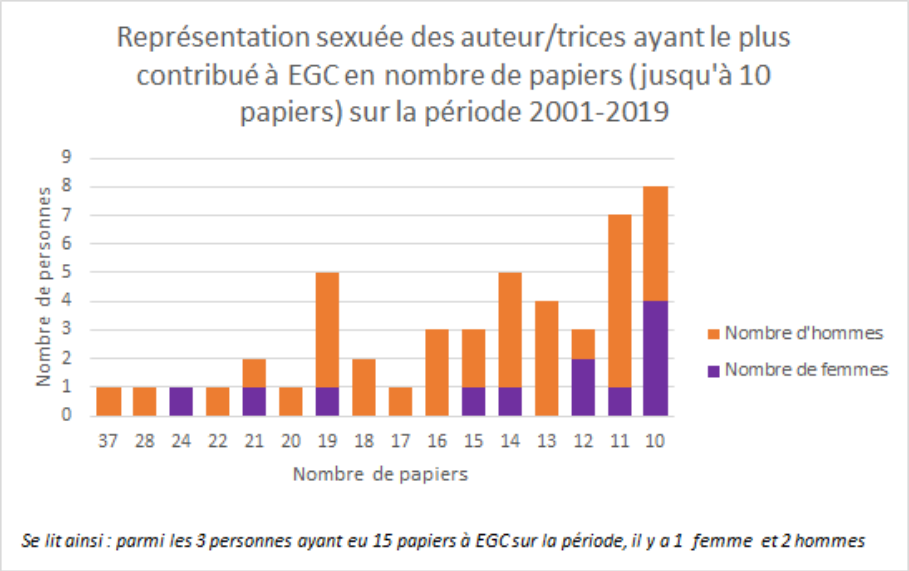


Figure II.3.2 – Représentation sexuée des auteur/trices ayant le plus contribué sur 2001-2019.

3.3.2 Membres du comité de programme

Nous nous sommes également penché·es sur la présence des femmes dans le comité de programme. En étude sur le genre, un des indicateurs quantitatifs utilisés est le rapport de masculinité. Il est exprimé en nombre d'hommes pour 100 femmes (à la naissance, il est classiquement de 105 garçons pour 100 filles). Dans la figure II.3.3, nous constatons que, depuis 2015, ce rapport de masculinité a diminué jusqu'à 200 hommes pour 100 femmes, tendant vers un meilleur équilibre sexué sur la place des femmes dans le comité de programme. En 2019, cela représente 34% de femmes dans le CP, c'est une manière de reconnaître leur compétence.

3.3.3 Présidences de la conférence

Un autre aspect intéressant est le travail et la visibilité qui découlent du rôle dans une présidence. Nous avons recensé les présidences d'honneur, de comité de programme (CP) et d'organisation (CO). Dans la figure II.3.4, nous pouvons observer le nombre de femmes et d'hommes impliqué·es. Ainsi, le pourcentage de femmes est de 17% pour la présidence d'honneur, 24% pour la présidence de comité de programme et 33% pour la présidence de comité d'organisation. Il est à noter que, parfois, des binômes de personnes ont pu être constitués, soit pour la gestion du comité de programme (surtout dans les premières édi-

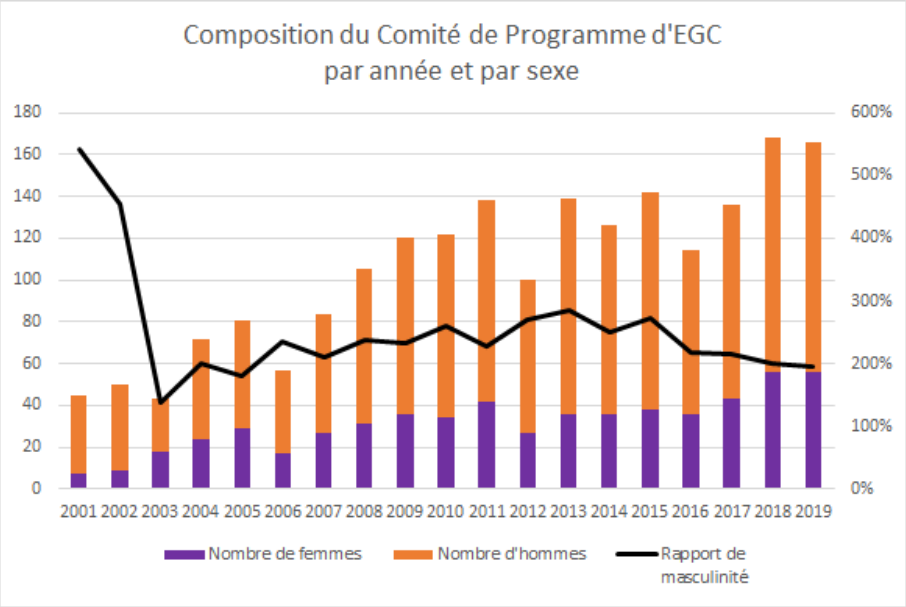


Figure II.3.3 – Composition sexuée du Comité de Programme par année et rapport de masculinité.

tions), soit pour la gestion de l'organisation. Si l'on considère le « prestige » pouvant être associé aux différents types de présidence, la dimension scientifique pouvant être globalement plus valorisée, il est intéressant de noter que l'on retrouve davantage de femmes sur l'aspect organisationnel des conférences, qui pourrait être mis en perspective d'analyses sociologiques sur la question des tâches au prisme du genre.

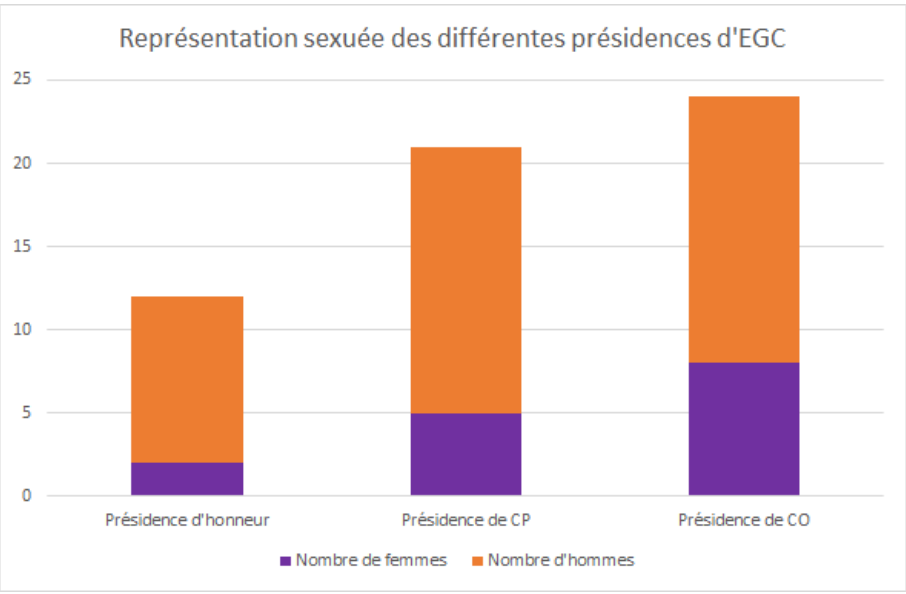


Figure II.3.4 – Représentation sexuée pour la présidence d'honneur, la présidence du Comité de Programme et la présidence du Comité d'Organisation.

3.3.4 Conférences invitées

Dans les carrières académiques, la visibilité et le rayonnement scientifique sont des éléments importants, notamment du point de vue de l'avancement dans la carrière, ou l'accès à des primes telles que anciennement la PEDR (Prime d'Encadrement Doctoral et de Recherche) ou à présent la RIPEC (Régime Indemnitaire des Personnels Enseignants et Chercheurs) dans sa composante individuelle.

Au delà d'une diffusion des connaissances, il apparaît que l'invitation pour présenter une conférence invitée dans le cadre d'un colloque peut contribuer à cet enjeu de visibilité. Ainsi, bien que cet aspect ne figure pas dans les données initiales fournies pour le défi, nous avons choisi de les collecter manuellement à la fois via le site *Web* EGC qui récapitule partiellement ces éléments dans le menu dédié [8], et nous avons complété les données qui s'avéraient lacunaires par une recherche dans les publications, compte-tenu du fait que les conférences invitées font l'objet d'un résumé d'une page maximum dans les actes de la conférence. Pour l'année 2001 et 2002, nous n'avons pu être en mesure de déterminer l'éventuelle présence de personnes donnant une conférence invitée, supposant que sur les deux premières éditions, il était possible qu'il n'y en ait pas eu.

Ainsi, entre 2003 et 2019, 65 personnes se sont succédées en session plénière pour présenter une conférence invitée, selon la répartition par année présentée dans la figure II.3.5. Parmi ces personnes, nous notons la présence de 16 femmes et 49 hommes, soit 25% de femmes. Compte-tenu du pourcentage de femmes dans la communauté, il y a une certaine forme de représentativité. Mais nous pourrions regretter que, dans certains cas, il y ait eu entre 3 et 5 personnes en conférence invitée, sans aucune femme (4 éditions concernées).

La présence en conférence invitée peut répondre à différentes logiques (y compris à celle du refus des personnes invitées) mais il apparaît nécessaire d'avoir une vigilance sur ce point, d'autant plus que la conférence EGC ne se limite pas à une seule invitation par édition.

3.3.5 Articulation des temps de vie

Le droit à la déconnexion est une disposition issue de la loi El Khomri de 2016, dite « loi Travail ». Elle est entrée en vigueur le 1er janvier 2017. Un des objectifs de cette loi est d'adapter le droit du travail à l'ère du numérique. Concernant en particulier le droit à la déconnexion, l'objectif est de permettre aux salarié·es de concilier vie personnelle et vie professionnelle, tout en luttant contre les risques de burn out. Pour cela, ils/elles doivent avoir la possibilité de ne pas se connecter aux outils numériques et de ne pas être contacté·es par leur employeur en dehors de leur temps de travail (congés payés, jours de RTT, *week-end*, soirées...).

8. https://www.egc.asso.fr/category/publications/conferences-invitees

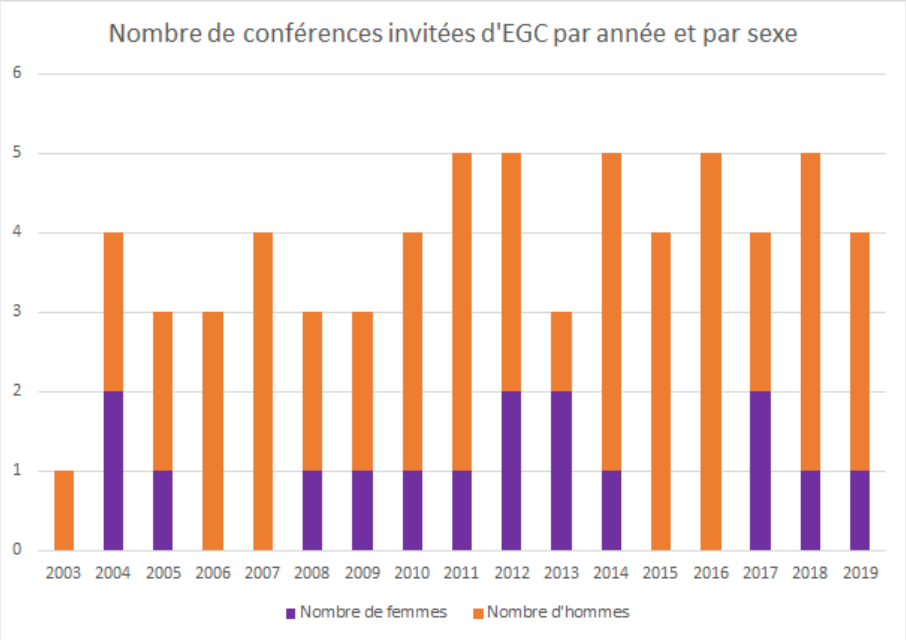


Figure II.3.5 – Historique des conférences invitées : représentation sexuée.

Ce droit à la déconnexion concerne toutes les personnes salariées. Cependant, il est à noter que cette loi ne s'appliquait pas dans la fonction publique de façon réglementaire au moment de notre étude (le droit à la déconnexion dans la fonction publique est consacré dans l'accord télétravail du 13 juillet 2021, en dépassant toutefois le champ du télétravail). Des initiatives pouvaient néanmoins être prises dans les institutions. La question se pose alors sur la capacité à respecter ce temps de déconnexion, dans des environnements dont le temps de travail peut être une variable assez « élastique », et c'est le cas pour beaucoup d'enseignant·es-chercheur·es.

Les travaux scientifiques en Sciences Humaines et Sociales se multiplient autour de l'épuisement au travail, de l'articulation des temps de vie, sans que le métier d'enseignant·e-chercheur·e ne soit épargné par ces problématiques, et ce dans différents pays (Hechiche-Salah et al., 2018 ; Sousa, 2015). L'usage du mail apparaît être comme une des composantes pertinentes à analyser. C'est ce que font d'ailleurs Chaulet et Datchary, 2014 dans un article intitulé *Moduler sa connexion : les enseignants-chercheurs aux prises avec leur courriel*, analysant sur la base d'entretiens l'usage du mail. Au vu des données fournies pour le défi EGC sur les mails de la liste de diffusion, il nous a paru pertinent de faire un focus sur l'analyse des heures d'envoi de ces mails, pour traiter de cette question du temps de travail.

Rappelons que le jeu de données de la liste de diffusion EGC comporte 7 075 mails écrits entre le lundi 10 juillet 2006 et le dimanche 9 septembre 2018, provenant de 811 adresses mails différentes.

Pour travailler sur les horaires, nous avons utilisé le champ *date* des emails consi-

dérant qu'il reflète normalement l'heure locale[9] pour la personne qui envoie le mail, contrairement au champ *received* donnant la date du côté serveur SMTP potentiellement distant.

Nous avons choisi de considérer l'envoi de mails dans le cadre professionnel de 8h à 20h comme « classique ». Nous avons pu dénombrer 1 116 mails écrits en dehors de cette plage horaire, soit 16% des mails. Ceci est une trace du travail non négligeable effectué en dehors d'horaires classiques de travail, car si les mails aux listes de diffusion peuvent être considérés marginaux dans l'activité professionnelle, ces derniers sont la trace de cette activité.

Dans la figure II.3.6, nous pouvons constater que l'augmentation de l'activité sur la liste de diffusion va de paire avec une augmentation du pourcentage de mails envoyés hors horaires classiques, qui oscille à présent autour de 15%.

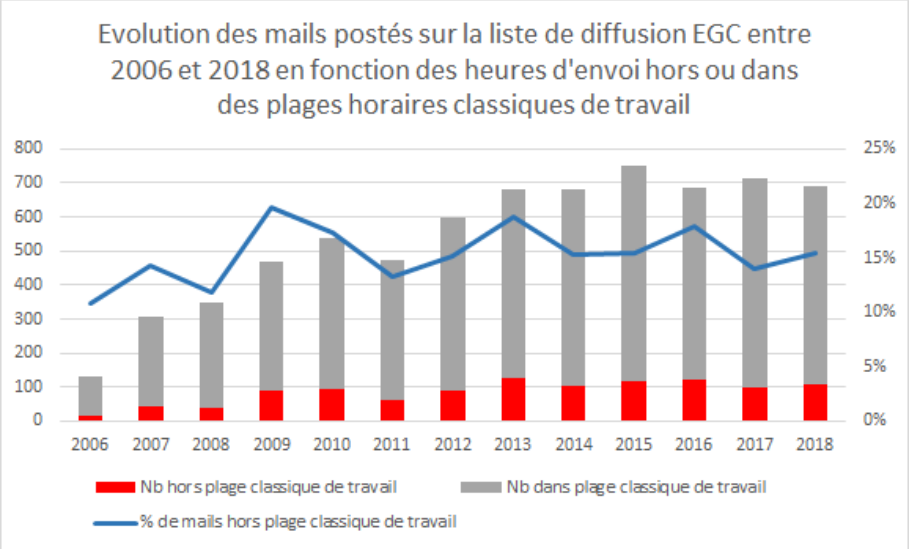


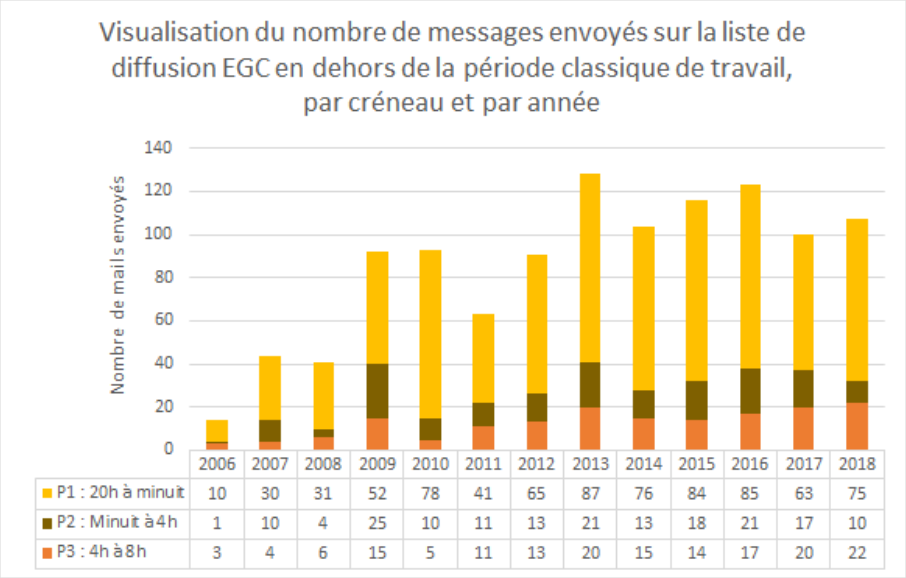
Figure II.3.6 – Évolution des mails postés sur la liste de diffusion EGC en fonction des plages horaires dédiées ou non classiquement au travail.

Pour aller plus en détails, nous avons découpé la période non classique en 3 créneaux : de 20h à minuit, de minuit à 4h et de 4h à 8h. Les résultats sont fournis dans la figure II.3.7. Nous pouvons y observer que la tranche 20h-minuit est celle qui est la plus utilisée, correspondant à une activité de travail en soirée. Parmi les 1116 mails hors créneaux classiques, 70% ont été écrits de 20h à minuit; 15% de minuit à 4h et 15% de 4h à 8h. Ce travail en dehors d'horaires classiques doit nous interpeller sur le rapport au travail et notre organisation du temps de travail.

En reprenant cette étude, en associant chaque adresse mail à la chercheuse ou au chercheur correspondant·e quand cela était possible (l'usage du mail professionnel a fa-

9. en tenant compte du fuseau horaire local selon la norme RFC 4021, sous réserve que les mails envoyés ne l'aient pas été durant un déplacement à l'étranger avec décalage horaire

cilité cet appariement, nous avons exclu de notre étude les adresses génériques), il s'avère qu'il n'y a pas de différence femmes-hommes quant à l'envoi tardif de messages.



	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P1 : 20h à minuit	10	30	31	52	78	41	65	87	76	84	85	63	75
P2 : Minuit à 4h	1	10	4	25	10	11	13	21	13	18	21	17	10
P3 : 4h à 8h	3	4	6	15	5	11	13	20	15	14	17	20	22

Figure II.3.7 – Évolution des mails postés sur la liste de diffusion EGC en fonction des plages horaires et des années pour les mails postés en dehors des plages classiques de travail.

3.4 Discussion

Gaudel et Rozoy, 2017 proposent dans un bulletin de la Société Informatique de France un état des lieux dans l'enseignement supérieur et la recherche pour les femmes et l'informatique, rassemblant un certain nombre de données chiffrées.

Différentes études relevant de la scientométrie se sont intéressées à une analyse sexuée des données (Demarest et al., 2014 ; Hartley & Cabanac, 2014 ; Larivière et al., 2013 ; Paul-Hus et al., 2015). Ces études publiées en anglais utilisent le terme de « gender ». Il s'agit pour nous de bien alors distinguer l'analyse sexuée de l'analyse genrée. La première étant le fait de construire une analyse considérant le sexe comme un paramètre de l'étude, la seconde cherchant à aller au-delà des résultats obtenus, pour les remettre en perspective à l'aune du concept de genre, fondé sur l'idée de rapports sociaux de sexe. Pour ce faire, utiliser les lunettes du genre nous pousse à avoir recours à des travaux en études de genre mobilisant différentes disciplines, ce qui constitue un point clé de la démarche, là où l'analyse sexuée des données ne constitue alors qu'une étape préalable.

Par ailleurs, la démarche de travailler sur des gros volumes de données, avec un étiquetage automatique (en se basant sur des dictionnaires de prénoms) diffère d'une démarche d'analyse d'une communauté, où il s'agissait de pouvoir déterminer cette étiquetage « plus finement », manuellement, en multipliant les sources d'information (sans contacter les personnes).

Pour ma part, ce travail de réflexion sur l'analyse genrée de données de la recherche a démarré à l'occasion du défi de la conférence EGC 2016, où j'avais amené l'idée d'une analyse sexuée des données de la Conférence EGC, focalisant mon attention sur trois aspects : les publications, les membres de comité de programme et les présidences de comité de programme. Ces résultats ont été consignés, avec les autres éléments de notre proposition groupée (Cabanac et al., 2016), pour être accessibles sous forme d'une anthologie des congrès, en ligne [10]. Le défi 2020 a permis d'aller au-delà en termes d'axes d'analyse (conférences invitées, différents types de présidence, articulation des temps de vie).

Mentionnons que ces analyses constituent autant de miroirs de cette communauté, pouvant être déformants lorsque l'on se focalise sur un axe à la fois. Il s'agit alors d'aller au-delà de ce que les informations chiffrées montrent, soit pour les comprendre, soit pour voir ce qu'elles ne disent pas, aller au-delà de ces constats de représentativité.

Afin d'avoir une analyse plus fine, différentes perspectives sont à envisager, certaines sont réalisables, d'autres son plus délicates compte-tenu de l'absence de données.

Parmi les perspectives que je n'ai pas encore pris le temps de poursuivre, il y aurait la question du statut MCF/PR dans les participations au comité de programme ou d'organisation. Cela nécessite de compléter par les statuts nos données, mais aussi d'historiser ce statut par rapport aux années de passage du statut de MCF à PR.

Concernant l'articulation des temps de vie, s'il n'y a pas eu de différence notable femmes-hommes, il y aurait grand intérêt à voir ce qu'il en est par rapport à une configuration familiale, à savoir la présence ou non d'enfants. Différents travaux sociologiques ont montré comment l'arrivée du premier enfant marque un tournant dans un fonctionnement moins égalitaire (sur la répartition des tâches domestiques notamment).

Et il serait alors judicieux, d'explorer à la fois une approche plus fine des découpages de créneaux analysés (en isolant notamment le créneau 7h-8h ou 20h-21h). Pour affiner cette analyse de l'articulation des temps de vie, il pourrait être judicieux de prendre en compte les *week end* et jours féries, voire même les vacances (même si ce point peut être plus délicat au niveau de la construction des données compte-tenu de l'existence des zones et du choix de la semaine quand il n'y a qu'une semaine de vacances universitaires sur les 2 semaines de vacances scolaires).

Par ailleurs, j'avais entrepris de réitérer ce travail sur la communauté INFORSID. Certains aspects sont à poursuivre pour parvenir à une comparaison sur les différents points. Ainsi, le travail sur l'archivage des données d'une communauté peut trouver de la valeur pour ce type d'exploration.

L'intérêt de la comparaison avec la communauté INFORSID, qui présente le fait que je la connaisse bien comme EGC, est qu'elle est marquée par une présence globalement plus importante des femmes dans ce domaine que sont les systèmes d'informations. Ainsi,

10. https://dbrech.irit.fr/rechpub/cabanac_egc.accueil

construire les différents indicateurs est intéressant.

Il ne s'agit pas pour autant de vouloir s'intéresser à l'ensemble des communautés françaises sur les différentes thématiques. Non pas que ça n'ait pas d'intérêt. Mais plutôt parce que je réfléchis également en termes de transformation par la communication de ces résultats. Mon hypothèse est qu'en étant au cœur de la communauté, contribuer à des changements est plus envisageable, dans ces cas précis. C'est ce que j'ai pu constater dans mon expérience.

En outre, le recueil d'autres données sexuées de la recherche permettrait de rendre compte d'autres aspects, comme par exemple les questions autour des projets de recherche, de leurs financements, de la participation aux recrutements dans le cadre des comités de sélection, des PEDR et à présent RIPEC, etc. Deux points nous semblent intéressants à développer ici par rapport à cette analyse : 1) le caractère genré de la discipline ; 2) la notion de carrière académique impactée directement en termes de genre.

L'informatique est une discipline que nous pouvons qualifier de genrée, dans le sens ici où il y a un déséquilibre de représentativité femmes-hommes (dans l'enseignement supérieur et la recherche mais plus généralement également dans les métiers liés à l'informatique). Cette sous-représentation des femmes n'est pas due à la « nature », parce que les femmes seraient moins aptes à faire de l'informatique… Mais elle est le produit d'une construction socio-historique, notamment accentuée par des stéréotypes de genre qui aujourd'hui sont très présents (notamment sur les représentations liées à ce qu'est un « informaticien »).

Ce travail avait pour objectif de poser les bases d'une discussion sur l'analyse genrée des données de la recherche. Il en ressort deux aspects majeurs. Le premier porte sur les enjeux de la pluridisciplinarité, voire de l'interdisciplinarité, pour aller au-delà de la production de données chiffrées, avec un besoin de recontextualisation pour la compréhension et l'analyse de ces données grâce aux lunettes du genre. Le second a trait à l'enjeu de cette production de données chiffrées, pouvant rendre visibles certaines problématiques liées notamment aux questions d'égalité/d'inégalité. Ce second aspect renvoie à la dimension politisée des études de genre de manière générale.

3.5 Réflexions conclusives : constat de la place des femmes et après?

La place des femmes dans l'enseignement supérieur et la recherche [11] reste une thématique d'actualité. Le travail réalisé sur la communauté EGC avec une perspective d'analyse en termes de genre a permis de faire un état des lieux sur la place des femmes dans cette communauté, au-delà des contributions en termes de publications.

11. Rédaction en 2014 d'un Livre Blanc « Le genre dans l'enseignement supérieur et la recherche » par des chercheur·es de l'ANEF (Association Nationale des Études Féministes)

Finalement, la question principale qui émerge est : après avoir procédé à des analyses fournissant différents résultats sur la place des femmes, que faire face à certains constats qui demeurent problématiques?

Historiquement, ce déséquilibre en informatique n'a pas toujours été aussi marqué, voire même, l'informatique était initialement un métier féminin avant la généralisation du micro-ordinateur. Dans (Collet, 2006), Isabelle Collet retrace la dimension historique de l'évolution de cette représentativité. En 2011, Isabelle Collet écrivait déjà qu'« en l'espace de vingt ans, la part des femmes en informatique a été divisée par deux » (Collet, 2011).

Précisons qu'il ne s'agit pas de la question « faut-il absolument avoir 50% de femmes et 50% d'hommes dans ces métiers? », mais il s'agit plutôt de la question de la liberté d'orientation professionnelle, notamment vis-à-vis des stéréotypes de genre qui touchent directement les métiers de l'informatique. Dans ce cadre, la mobilisation de travaux en science de l'éducation est essentielle (Collet, 2004).

Morley et Collet (2017) sont revenues sur deux expériences positives d'inclusion : la Norwegian University of Science and Technology (NTNU) et la Carnegie Mellon University (CMU). L'absence des femmes dans le monde digital ne serait pas une fatalité. C'est le message que porte Isabelle Collet dans son dernier ouvrage, dans lequel elle livre une analyse de la place des femmes dans le numérique, de façon socio-historisée notamment (Collet, 2019).

Au-delà des questions d'effectifs, il s'agit également de s'intéresser aux formes d'inégalités professionnelles, au-delà même d'une grille de salaire qui par essence n'est pas genrée.

Cela doit passer donc notamment par les réflexions et actions menées au sein des institutions académiques, en termes de politiques publiques, de politiques d'établissement, mais aussi comment, au sein même des institutions, les personnes se saisissent de ces questions.

La mise en place d'un atelier « Egalité en actions! » avec Claudia Roncancio, dans le cadre de la conférence Inforsid 2023, fut l'occasion de mettre en avant l'intérêt de partager sur ces questions, en tant que chercheuses et chercheurs, mais aussi en tant qu'enseignant·es, et responsables pédagogiques.

Au-delà de ces aspects au sein des institutions vis-à-vis de notre quotidien académique, il y a l'éternelle question du vivier. Ayant participé, à ma mesure, à des actions notamment de sensibilisation auprès des jeunes, comme par exemple dans le cadre d'une journée « Filles, Maths, Informatique : une équation lumineuse[12] », j'ai pu constater à quel point les stéréotypes sont bien ancrés. Comme l'énonce Isabelle Collet, la sous-représentation des femmes dans l'informatique relève plutôt d'un problème de censure sociale que d'auto-censure.

12. https://filles-et-maths.fr/

Il est vrai qu'après de nombreuses années d'actions multiples, le découragement pourrait se faire sentir. Il n'est pas toujours facile et/ou possible de mesurer l'impact des actions menées. Mais beaucoup œuvrons à un travail de fond pour changer la situation, avec l'endurance du castor [13] qui rëconstruit inlassablement les barrages, quoi qu'il arrive.

Ainsi, à l'occasion des premières Assises de la féminisation des métiers et filières du numérique [14], la fondation Femmes@Numérique, initiatrice et porteuse de ce projet, a rassemblé de nombreux partenaires et a communiqué en février 2023 un plaidoyer [15] pour la féminisation des métiers du numérique qui comprend différentes propositions d'action.

Il est à noter également que le rapport détaillé du Haut Conseil à l'Égalité intitulé « La Femme Invisible dans le numérique. Le cercle vicieux du sexisme [16] », sorti en novembre 2023, présente des constats alarmants et des recommandations.

Bien sûr, beaucoup d'autres initiatives existent. Mais je ne discuterai pas ici plus en avant des actions. L'enjeu était plutôt de pointer l'articulation entre constats et nécessité de mise en place de ces actions.

13. *Amik* en *anicinapek*
14. https://www.assises-feminisation-metiers-numerique.fr
15. https://www.assises-feminisation-metiers-numerique.fr/wp-content/uploads/2023/02/Plaidoyer_Assises-nationales23_VF.pdf
16. https://www.haut-conseil-egalite.gouv.fr/parite/actualites/article/rapport-la-femme-invisible-dans-le-numerique-le-cercle-vicieux-du-sexisme

4 Les politiques de quotas dans l'informatique en question

Parce qu'il y a eu toutes sortes d'acquis légaux - ô combien fragiles ! -, [des jeunes femmes] s'imaginent qu'il n'y a plus rien à faire. Or il reste l'essentiel : changer les mentalités.

Françoise Héritier (1933-2017), anthropologue française

Contributions sur lesquelles se base ce chapitre

- > Article dans une revue facultaire
  - **C. Favre** & L. Tain, Les quotas : levier ou frein au déroulement des carrières des femmes ? Analyse suite à une enquête préliminaire dans le cas de l'enseignement supérieur et la recherche en France dans le domaine de l'informatique, Annals of the University of Bucharest / Political science series, ANNUL XX, 2018, N°2, 37-54. Publication suite à la communication dans SCOPE 2018 dans les Annales de l'Université de Bucarest, Série Sciences Politiques. (Favre & Tain, 2018b)
- > Communications (sélection sur résumé)
  - **C. Favre** & L. Tain, Femme et enseignante chercheure en informatique : des contextes pluriels, une pluralité des vécus, SCOPE 2018, Bucarest, Roumanie. (Favre & Tain, 2018a)
  - **C. Favre**, Quotas in higher education and research : leverage or brake for gender equality ? Study in the field of computer science in France, Fazendo Gênero 2017, Florianopolis, Brazil. (Favre, 2017c)

M'INITIER à la démarche d'enquête sociologique traduit l'opportunité que j'ai saisie au sein de mon environnement de travail, un apprentissage au-delà des enseignements sociologiques théoriques que j'avais reçus durant mes études, et surtout un véritable *challenge*.

Comme cela est souvent le cas en Sciences Humaines et Sociales, choisir un sujet n'a rien de neutre. Le désir de traiter des politiques de quotas au sein des comités de sélection est né de mon propre expérience par rapport à ce sujet, avec la volonté d'explorer des vécus divers. Ainsi ce chapitre tend à rendre compte de notre analyse exploratoire qui a été menée en 2017-2018.

Le chapitre est organisé selon les sections suivantes. Quelques éléments seront tout d'abord présentés en préambule dans la section II.4.1. Puis le matériau d'enquête qui a permis l'analyse sera succinctement évoqué dans la section II.4.2. Dans la section II.4.3, sera restituée les résultats de notre analyse exploratoire sur les politiques de quotas. Une discussion suivra dans la section II.4.4. La section II.4.5 viendra clore ce chapitre avec des réflexions conclusives sur l'évaluation des politiques d'égalité.

4.1 Préambule

Si dans certains contextes, la mise en place de politiques des quotas sur un plan légal a permis des avancées certaines quant à la place que prennent les femmes aujourd'hui dans la société (en politique, au niveau des conseils d'administration en entreprise par exemple), sa généralisation dans le monde académique a pu induire des questionnements sur les résultats obtenus.

Ainsi, il s'agit de questionner celle-ci, notamment quand il s'agit de quotas au niveau de comités de sélection, dans un domaine tel que l'informatique caractérisé par une sous-représentation des femmes. En 2017, moment de l'enquête, les femmes représentent moins du quart de la discipline informatique dans l'ESR et moins de 20% du corps des professeur·es plus spécifiquement.

Le métier d'enseignant·e-chercheur·e revêt de nombreuses facettes en termes de tâches. Quels objectifs sont visés par de telles politiques et qu'est-ce qu'elles produisent, sur les plans individuels et collectifs ? Il s'agit de pouvoir éclairer les enjeux de la tension entre égalité professionnelle et politique des quotas.

L'hypothèse que nous pouvons formuler est que, dans le domaine de l'informatique, la mise en place de ces quotas pourrait induire une forme de sur-sollicitation des femmes contre-productive par rapport aux objectifs initiaux, notamment en termes d'égalité dans le déroulement de la carrière académique (ce qui pourrait être qualifié d'« effet rebond »).

Il est ici à rappeler en préambule le schéma type d'une carrière académique en France des enseignant·es-chercheur·es dans les universités (hors postes de chercheur·es), en se focalisant sur les deux étapes clés que sont l'obtention d'un poste de Maître·sse de Confé-

rences (MCF) et de Professeur·e des Universités (PR), sans détailler ici les différents grades qui composent eux-mêmes chacun de ces corps en matière d'avancement dans la carrière. Une fois la thèse de doctorat soutenue, une candidature à la section CNU (Conseil National Universitaire) de la discipline concernée doit être déposée pour obtenir de cette instance nationale la « Qualification Aux Fonctions de Maître·sse de Conférences ». Une fois cette qualification obtenue, il est alors possible de candidater sur les postes à pourvoir de Maître·sse de Conférences proposés dans les universités. Pour de nombreuses disciplines, dont l'informatique, au-delà de l'ancienneté, l'obtention d'un poste de PR nécessite l'obtention du diplôme d'Habilitation à Diriger les Recherches, diplôme correspondant au niveau le plus élevé qu'il soit possible d'obtenir à l'université. Ce diplôme nécessite la rédaction d'un mémoire plus ou moins conséquent selon les pratiques, avec une soutenance devant un jury composé de « pairs ». Il n'y a à présent plus de « Qualification Aux Fonctions de Professeur·e des Universités » à obtenir auprès de l'instance du CNU, pour les personnes qui étaient déjà en poste en tant que MCF. Les candidatures se font ensuite sur des postes PR. Ce n'est pas en général une transformation de son propre poste de MCF en PR (sauf exception d'une procédure de repyramidage mis en place récemment), même si le passage entre MCF est PR est en général qualifié de promotion dans le langage commun.

4.2 Données considérées et explicitation des enjeux

Les données considérées pour cette analyse diffèrent de l'ensemble des données traitées dans le reste du manuscrit, puisque celles-ci correspondent à du « matériau » issu des entretiens semi-directifs, une approche qualitative bien connue des sciences sociales.

Il s'agissait d'une enquête exploratoire menée avec un petit nombre d'entretiens (5) auprès de femmes enseignantes-chercheuses en informatique.

Il est toujours plus aisé de comprendre certains mécanismes sociaux via une approche comparative. Ce qui est présenté ici de notre enquête exploratoire s'appuie sur le recueil de paroles de femmes. Le fait ici de n'avoir enquêté qu'auprès de femmes peut amener un risque de « sur-spécialisation » des expériences féminines.

Nous aurions souhaité poursuivre cette enquête plus largement dans une perspective comparative mais finalement cela n'a pas été le cas. Par ailleurs, un problème informatique amenant à une disparition de données (c'est un comble !) nous a amenées à aménager la restitution, et son niveau de détails.

Malgré ce contexte, il paraissait opportun de pouvoir rendre compte de ces éléments autour de l'ambivalence qui émane des politiques de quotas, dans un cadre spécifique qu'est celui de l'ESR.

L'enjeu est donc de tenter d'éclairer les liens entre l'égalité professionnelle femmes-hommes et la politique des quotas à partir de l'expérience de sa mise en œuvre dans

le domaine spécifique de l'ESR en France dans la discipline de l'informatique, dans le cadre plus particulier des comités de sélection (CoS), instance de recrutement des postes d'enseignant·e-chercheur·e.

Précisons que nous ferons référence par la suite aux données de l'Institut National d'Etudes Démographiques (INED [1]), quant à l'âge moyen par rapport à la maternité. Cette précision nous a paru utile au regard des déroulés de carrière et des propos présents dans certains entretiens, sur les questions d'articulation des temps de vie.

L'idée sera ici de rendre compte de portraits emblématiques des situations que nous avons rencontrées lors des entretiens.

4.3 Contributions : une enquête exploratoire pour aboutir sur une hypothèse d'un effet de génération

En guise de contributions, nous présentons ici la restitution de l'analyse opérée à partir des données de notre enquête exploratoire.

Le déroulé reprendra donc les 3 périodes distinguées : les générations nées dans les années 60-70, les générations nées dans les années 80 et les générations nées dans les années 90. Pour chaque période, nous commencerons par présenter le contexte du déroulé de carrière académique en informatique avec la spécificité de ses temporalités, puis nous appuierons l'analyse sur l'évocation d'une trajectoire emblématique d'une informaticienne.

Il est à noter que nous avons recours ici au terme de génération pour distinguer les périodes de naissance des interviewées, sans considérer le sens démographique exact de ce terme.

4.3.1 Les générations nées en 1960-70 : ancrage d'un quota levier

4.3.1.1 Le contexte des carrières des générations nées en 1960-70

Les personnes nées dans les années 60-70 vivent une enfance et une adolescence essentiellement hors informatique. En effet, cette période correspond à l'émergence de la micro-informatique au-delà d'un usage spécialisé : le premier micro-ordinateur vendu tout assemblé n'apparaît qu'en mai 1973. Néanmoins, l'usage n'est pas encore généralisé au sein des familles françaises. Par conséquent, les personnes nées dans les années 60-70 ont grandi dans des familles sans matériel informatique à disposition à la maison.

Un autre élément de contexte concerne le déroulement de la carrière au regard de l'émergence de la discipline. L'accès aux études supérieures se fait dans les années 80-90

1. https://www.ined.fr/

au moment même où se développe la discipline informatique à l'université. C'est au démarrage de l'informatique qu'on observe une féminisation la plus importante au cours du temps, notamment dans les écoles d'ingénieures avec un % de femmes de 20% en 1983, même si les étudiant.es suivant ce cursus ne sont pas dans une répartition sexuée équilibrée, comme le précise Isabelle Collet (2007) dans Le Monde Diplomatique « L'informatique a-t-elle un sexe ? », qui reprend des éléments de son ouvrage (Collet 2006). L'ouverture de postes sur une discipline émergente est favorable à un recrutement et à des promotions rapides entre MCF et PR.

Un autre élément de contexte a trait aux repères temporels liés d'une part à l'introduction de quotas et d'autre part à l'âge moyen à la maternité de ces générations. En effet, pour les femmes nées dans les années 60-70, la mise en œuvre des politiques de quotas dans l'Enseignement Supérieur et la Recherche arrive en 2012-2013, c'est à dire au moment où elles ont entre 40 et 50 ans. Par ailleurs, cette introduction des quotas intervient à un âge postérieur au calendrier de la maternité. En effet pour la génération née en 1960 l'âge moyen à la maternité est de 27,7 ans selon les données de l'INED .

Ce constat de maternité « précoce », ou dans la continuité tout du moins des précédentes générations par rapport à celles qui suivront, est confirmé par un autre indicateur, l'indicateur transversal de l'âge moyen des mères à la naissance des enfants. L'âge moyen à la maternité est en effet de 26,5 ans en 1977, année qui correspond au minimum de l'âge à la maternité observé en France depuis la deuxième guerre mondiale (Toulemon & Mazuy 2001).

4.3.1.2 Un portrait emblématique d'informaticienne des générations 1960-70

Le portrait de Sandrine N. s'avère emblématique des trajectoires d'informaticiennes nées dans les années 60-70. Sandrine N., née en 1970, nous raconte que son enfance et son adolescence se sont déroulées en dehors d'un univers informatique. Certes, il y a eu un minitel dans sa famille mais c'était le seul outil numérique à disposition. Elle a pu avoir un usage d'un micro-ordinateur de façon quotidienne seulement au moment où elle faisait ses études dans le supérieur dans le domaine. Elle met au monde son premier enfant alors qu'elle termine ses deux années de poste d'attachée temporaire d'enseignement et de recherche. Elle obtient son poste de MCF en 1998 dans la foulée. Elle passera son habilitation à diriger les recherches en 2011, pour obtenir finalement son poste de PR en 2015.

En 2012-2013 lorsque les décrets d'application sont publiés cette politique lui est plutôt favorable comme elle l'explique dans l'extrait d'entretien ci-dessous. La sollicitation intensive pour des comités à l'échelle nationale lui permet en effet de renforcer des réseaux professionnels, d'accroître sa visibilité et donc de valoriser ses acquis antérieurs facilitant ainsi sa promotion en 2015 au sein du corps des professeurs, à l'âge de 45 ans.

« parce que ça moi qui ai beaucoup vécu dans l'ombre, [...] et du coup pour moi c'est intéressant d'aller faire des COS dans d'autres universités, à la fois bah pour découvrir d'autres pratiques, d'autres façons de faire, voilà, rencontrer des collègues qui sont dans des thématiques quand même proches des miennes et constituer enfin dans le bon sens du terme un réseau professionnel, donc si-si pour moi ça ça a un effet vrai – enfin.. positif pour moi, à titre personnel oui. » Sandrine N, professeure en informatique, née en 1970

4.3.1.3 Une politique de quotas ressource de légitimité pour les informaticiennes

Cette étude exploratoire nous suggère l'esquisse d'une première expérience de la politique des quotas liée aux générations nées dans les années 1960-1970. Ces femmes ont pu bénéficier de l'ouverture de la discipline à l'université. L'expérience des quotas survenant au cours de la carrière est vécue de façon plutôt favorable : elle n'est pas un obstacle à la publication scientifique et correspond plutôt à une période de valorisation impliquant la consolidation de réseaux professionnels, favorisant l'évolution de carrière pour des personnes dont celle-ci a pu être ralentie notamment de part une implication familiale forte. L'incitation à participer à des comités de sélection dans d'autres universités de France contribue à bâtir la visibilité et donc la légitimité institutionnelle.

4.3.2 Génération 80 : diffusion d'un quota ambivalent

4.3.2.1 Le contexte des carrières des générations nées dans les années 1980

Pour les générations nées dans les années 1980, l'outillage numérique fait peu à peu partie de l'environnement quotidien. C'est d'abord le minitel qui apparaît puis les micro-ordinateurs qui servent à l'ensemble de la famille. Il y a aussi un usage ludique particulièrement mobilisé par les garçons.

Ces générations voient le grand essor de la discipline informatique à l'université avec la multiplication de filières spécialisées. Simultanément, cet essor de la discipline s'accompagne peu à peu d'un recul de la féminisation du public étudiant dans cette discipline. Ainsi, par exemple, en 2000, Isabelle Collet (2007) note que la proportion des femmes dans les écoles d'ingénieur.es pour la spécialité informatique est de 11%. Le pourcentage est donc presque revenu à son niveau inital qui était de 9% dans les années 70. La proportion de femmes en informatique rejoint celle des disciplines de mécanique et de défense qui sont traditionnellement des bastions masculins. Le développement de la discipline qui reste une jeune discipline universitaire est encore favorable à la création de postes et au recrutement d'informaticien·nes.

Les temporalités des trajectoires de vie de ces générations sont bien différentes des temporalités des générations des années 1960-1970. Le recul de l'âge moyen à la maternité est caractéristique de ces générations, recul qui a commencé déjà un peu avant. En effet, l'âge moyen à la maternité pour la génération née en 1980 est de 30,1 selon les don-

nées de l'INED.

Par ailleurs l'introduction de la politique des quotas arrive à un autre moment de leur carrière. Elles sont au début de leur carrière académique et subissent donc une double pression : d'une part enchaîner la participation aux comités de sélection et d'autre part accumuler les publications. C'est en effet également le moment où l'on peut voir apparaître une transformation du fonctionnement de la recherche avec une accélération de l'injonction à la publication, avec l'usage généralisé du « PUBLISH OR PERISH ».

4.3.2.2 Un portrait emblématique des générations nées dans les années 1980

Céline G., MCF en informatique, est née en 1980. Pendant son enfance, à la maison, il y a un ordinateur qui est l'ordinateur pour les jeux et qui est largement utilisée par son frère. Elle n'est pas du tout familière de l'outil, et naïvement, elle tentera même un jour de l'utiliser de façon inadéquate comme traitement de texte. Elle découvrira l'informatique a proprement parler via la programmation et son goût pour la logique algorithmique seulement en 1ère année d'université à l'occasion de la préparation de sa première année de diplôme de Mathématiques Appliquées aux Sciences Sociales. Elle a toujours aimé les mots et les Sciences Humaines et Sociales, mais comme elle avait le niveau pour suivre un baccalauréat scientifique (au sens de l'appellation utilisée pour désigner à l'époque la filière avec notamment les mathématiques poussées), elle a suivi cette filière-là. Elle poursuivra finalement un cursus d'informatique jusqu'au doctorat, sans forcément avoir l'impression d'évoluer dans une discipline masculine. Lorsqu'elle candidate aux postes de MCF, à son grand étonnement, elle a le choix des postes. Malgré son manque de confiance en elle, dont elle parle volontiers, elle est classée sur plusieurs postes auxquels elle a candidaté. Et si à l'époque il y a déjà une pression en termes d'accès aux postes, que l'ensemble des candidat·es n'obtiennent pas forcément de poste, elle se sentira chanceuse d'avoir un certain choix. C'est à ce moment-là que pour la première fois, elle prendra conscience de la dimension genrée de la discipline et elle sera « ramenée » à son sexe, comme le montre l'anecdote dont elle fait part.

« Je me souviens bien de la première fois où j'ai su que j'aurai un poste de MCF. J'avais fait une audition devant un comité où y'avait que des hommes. Honnêtement, je m'étais sentie un peu malmenée dans les questions. Non pas qu'elles étaient pas pertinentes, mais plutôt sur une manière un peu agressive de les poser. J'arrivais d'un poste d'ATER où je faisais de l'informatique pour des étu qui n'étaient pas informaticiens purs et durs et qui avaient aussi une formation en SHS. Et là, à cette audition, on me demande un peu de façon bousculante, euh bousculante car ça m'avait rappelé un mauvais souvenir d'entretien RH pour une entreprise quelques temps avant, où je m'étais faite descendre en gros car j'avais une thèse et que j'étais pas opérationnelle. Bref, là on me demande si je me sens vraiment capable d'enseigner à des informaticiens en IUT. Avec le recul, le problème c'était + la forme de la question que le fond même si c'est quand même un peu gonflé, car j'ai quand même un doctorat d'informatique. Je l'ai pas eu dans une pochette surprise ! J'ai finalement été classée première pour ce poste, et quand je l'ai su, j'espérais vraiment que y'aurait d'autres postes possibles car c'était mon premier résultat et ça donnait pas très envie d'aller travailler là-bas. Un an après, quand j'ai recroisé en conf un des mecs de ce comité qui travaillait là-bas, j'en ai profité pour lui faire part de mon étonnement sur ce classement de 1er alors que franchement j'avais pas bien vécu l'audition. Et là il m'a dit qu'il voulait féminiser l'équipe. À dire vrai, sur le coup, je l'ai pas super bien pris au fond de moi. Je pense que j'ai un peu oublié que peut-être y'avait un bout de phrase du type à dossier équivalent on voudrait féminiser l'équipe. Ou tout du moins que c'était pas seulement parce que j'étais une fille ! À ce niveau-là, on recrute quand même pas juste par rapport au sexe ! »
Céline G., maîtresse de conférences en informatique, née en 1980

Céline G. est une personne très impliquée dans son travail, notamment dans la dimension collective, pédagogique. C'est une personne qui se dit très volontaire à rendre service. À partir de 2013, elle est sollicitée régulièrement pour faire partie de comité de sélection. Elle n'identifie pas forcément que c'est dû au départ à des questions de parité. Mais peu à peu, elle prend conscience de la charge de travail que cela représente et a la sensation que c'est une tâche qui revient régulièrement chaque année, peut-être un peu moins que pour ses collègues hommes. Elle réalise peu à peu que les règles de constitution des comités de sélection font qu'elle est peut-être davantage sollicitée et elle se pose des questions sur ce que cela induit, comme le montre l'extrait d'entretien présenté ci-après.

« au départ, quand on me sollicitait pour les comités de sélection, y'avait une part de moi toute contente, la sensation d'être visible, comme une forme de reconnaissance. Mais en fait, c'était des fois plutôt parce qu'il y avait besoin de femmes pour le comité. Alors à un moment, c'est pas très agréable, ça met des doutes sur pourquoi j'étais choisie. [...] Et puis j'me suis rendue compte qu'au bout du compte ça prend quand même un sacré temps, l'évaluation des dossiers, se déplacer pour aller en comité. Contente de voir du monde et en même temps, pendant ce temps-là, ben je fais pas autre chose, quoi. Pas de la recherche. C'est pas ce qu'il y a de plus valorisant pour la carrière, c'est pas du travail rémunéré. Alors quand on en fait beaucoup, ben c'est pas anodin. Puis moi je suis hyper impliquée dans mon travail, donc du coup, ce que je f'sais pas pendant la journée, ben je le faisais le soir et le week end. À 38 ans, la vie a fait qu'en fait ben j'avais toujours pas de famille à moi, je veux dire pas de compagnon stable et pas d'enfant, quoi. Donc sans doute que c'est plus facile pour moi de bouger beaucoup pour ce type d'invitation. »
Céline G., maîtresse de conférences en informatique, née en 1980

Comme le montre l'extrait présenté juste avant, Céline G. a une implication dans son travail qui se caractérise parallèlement par l'absence de développement d'une vie familiale. À 38 ans, elle n'a pas encore d'enfants même si elle l'envisage sérieusement après avoir rencontré un compagnon. Elle incarne ce recul de l'âge à la maternité.

4.3.2.3 Les effets ambivalents de la politique des quotas

Pour cette génération, la mise en œuvre de la politique des quotas se révèle beaucoup plus ambivalente. Certes, d'une part elle accroit la visibilité des jeunes informaticiennes en raison de leur participation à de multiples comités au niveau national. Cette reconnaissance peut être un levier pour leur promotion dans le grade de professeur. Néanmoins, simultanément, compte-tenu d'une répartition non paritaire dans la discipline, les jeunes informaticiennes peuvent se retrouver très mobilisées pour participer à différentes instances. Cette sur-sollicitation peut s'avérer contre-productive pour le développement de leur carrière individuelle. Il est en effet difficile de cumuler des publications de haut niveau et des implications fortes dans les tâches institutionnelles liées à la participation à différents comités. De plus, ce recours à des femmes informaticiennes parce qu'elles sont femmes induit au niveau subjectif un sentiment de non-légitimité (renforçant possiblement un éventuel « syndrôme de l'imposteur ») qui est évoqué par les interviewées. Ce sentiment peut susciter une baisse de la confiance dans leurs compétences scientifiques, conformément aux attributions stéréotypées du système de genre. Cette expérience est parfois renforcée par les réactions de collègues masculins opposés à la politique des quotas.

Pour les MCF, c'est donc le passage au professorat qui peut poser problème et être rendu difficile. Pour les professeures, c'est l'avancement en grade qui peut être délicat. Ceci est explicité dans le schéma qui suit de la figure II.4.1, en évoquant le mécanisme que peut induire une politique de quotas en terme d'avancement de carrière pour les

femmes.

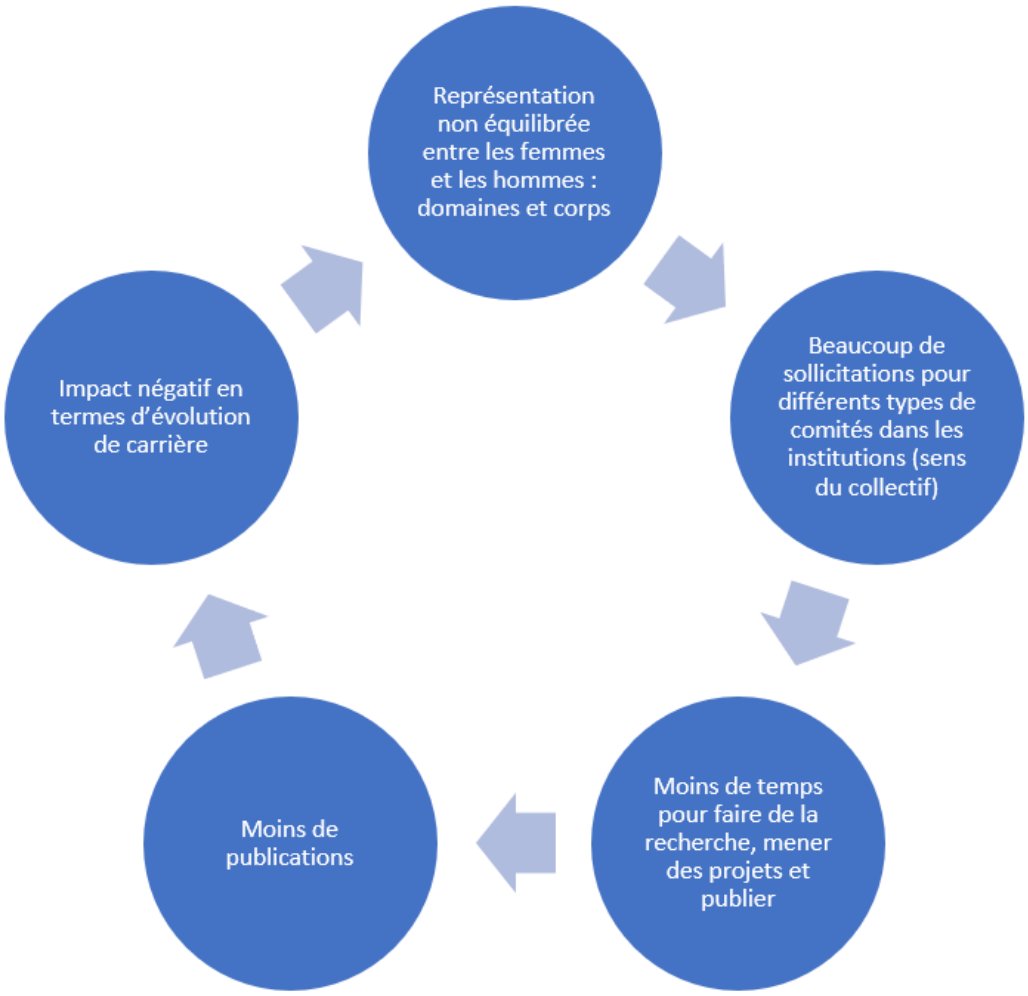


Figure II.4.1 – Etapes du cycle des politiques de quotas sur les CoS qui peuvent mener aux inégalités dans le déroulement de carrière dans le contexte de l'Enseignement Supérieur et la Recherche dans un domaine fortement déséquilibré au niveau sexué.

4.3.3 Génération 90 : démarrage/mise en route d'un quota frein

4.3.3.1 Le contexte des carrières des générations nées dans les années 1990

Un contexte bien différent apparaît avec les générations nées en 1990. En effet la discipline informatique a changé de statut et irrigue l'ensemble de la société. Pour autant, la création de postes dans la discipline reste limitée compte-tenu de la raréfaction de la création des postes dans l'Enseignement Supérieur et la Recherche de manière globale (étant entendu que l'informatique a pu être particulièrement bien dotée comparativement à d'autres disciplines).

Dès l'enfance les personnes nées dans les années 90 s'initient à l'environnement numérique. Cette initiation prend plusieurs formes les jeux, les réseaux sociaux... À l'école aussi les enfants acquièrent des compétences en micro-informatique. Simultanément cette familiarité avec les outils numériques s'avère très genrée : la figure du geek qui apparaît relayée par les médias est une figure masculine.

La période des études universitaires confirme ce recul de la présence des femmes dans la discipline. Il y a à la fois moins de femmes se spécialisant en informatique et moins de postes.

Les évolutions de calendrier constatées pour les générations des années 1980 sont désormais ancrées dans l'expérience quotidienne. C'est le cas pour la politique des quotas qui intervient dès leur entrée à l'université. C'est aussi le cas pour le recul de l'âge à la maternité pour les femmes ayant fait des études supérieures. Ce recul est acté. Il est même intégré dans les pratiques de médicalisation des maternités dites tardives. Il a été banalisé par des entreprises comme Google qui annonce en 2014 que l'entreprise propose déjà la congélation d'ovocytes à des femmes pour qu'elles puissent concilier leur carrière et leur maternité.

4.3.3.2 Un portrait emblématique des générations nées dans les années 1990

L'enfance d'Elodie D. se déroule dans une famille où le numérique se développe pleinement. À la maison, il y a un ordinateur familial, avec lequel elle joue beaucoup avec ses copains. Elle est la seule fille de son groupe d'amis passionnés d'ordinateur et de jeux. C'est au collège qu'Elodie se familiarise avec la pratique de l'informatique. Etant bonne élève, elle s'investit et réussit dans cette discipline, ce qui la conduit à s'engager dans une filière informatique à l'université. Elle se retrouve alors dans une section informatique où les filles sont très minoritaires, ce qu'elle évoque dans l'extrait d'entretien ci-après.

> « J'étais toujours un peu la seule fille. J'étais considérée comme un garçon manqué par ma famille car mes copains de jeux informatiques étaient tous des garçons. Ça a continué à l'université où nous n'étions pas très nombreuses. »
> Elodie D., jeune MCF en informatique, née en 1992

Elle réussit brillamment à l'université et la possibilité d'obtenir un financement de doctorat lui permet de poursuivre son cursus en thèse. Malgré le faible nombre de postes, elle réussit à être recrutée comme maîtresse de conférences un an après sa thèse.

4.3.3.3 Une politique de quotas qui risque d'être un frein à la carrière des informaticiennes

Cette génération inscrit sa carrière académique d'emblée dans un contexte institutionnel où s'appliquent les décrets concernant les quotas. Cette situation devrait être favorable à la promotion des femmes. Néanmoins l'enquête exploratoire amène un résultat plus contrasté car d'autres forces sociales interviennent.

Un premier obstacle à la féminisation de la discipline provient de l'évolution de la composition sexuée de l'informatique. Les femmes pionnières informaticiennes sont devenues invisibles et la discipline informatique est devenue un bastion masculin qui résiste à l'injonction de quotas : « pourquoi donc vouloir qu'il y ait 50% de femmes en informatique ? »

Un deuxième obstacle tient à la diffusion de l'image du Geek dans la société et à ses relais institutionnels. C'est une figure masculine qui induit l'idée que l'informatique n'est pas faite pour les femmes. La famille, l'école, les structures d'orientation répandent ce stéréotype qui freine l'accès des femmes en informatique. Ainsi, les femmes qui réussissent à surmonter ces deux obstacles sont actuellement dans une transgression de norme de genre.

4.4 Discussion

La mise en place de politiques de quotas s'est généralisée, et constitue un des instruments des politiques d'égalité. Les liens entre instauration de l'égalité et politiques des quotas ont toutefois suivi un itinéraire complexe en France durant ces dernières années (Laufer & Paoletti, 2015).

C'est ce que développe Lépinard (2007) dans son ouvrage « L'égalité introuvable. La parité, les féministes et la République ». Comme le résume Marques-Pereira (2011), après avoir esquissé une « généalogie internationale », Éléonore Lépinard retrace le long cheminement polémique qui a été nécessaire pour aboutir à une « exception française » compatible avec l'idéal Républicain.

Si le terme de démocratie paritaire est apparu au Conseil de l'Europe dès 1989, il a fallu repenser les liens entre féminisme et institutions de la démocratie pour prendre en compte la volonté d'une partie des mouvements féministes français de susciter des mesures législatives contraignantes visant à impulser une dynamique de féminisation des élites.

Le débat théorique sous-jacent initial était cristallisé autour d'une opposition entre universalisme républicain et différence des sexes. L'enjeu pourrait s'exprimer avec la question suivante : dans quelle mesure l'instauration de quotas risquerait-elle d'ancrer une vision différentialiste des catégories de sexe et donc de perpétuer des représentations sexuées pour les hommes et les femmes respectivement ?

Bien qu'aucun consensus n'ait été possible au sein des mouvements féministes français, une alliance réformiste originale s'est nouée, permettant l'instauration d'un cadre légal diversifié à différents niveaux et dans différents domaines. Selon la synthèse produite par le Secrétariat d'État à l'égalité femmes-hommes, le principe fondateur s'enracine dans la modification en 1999 des articles 3 et 4 de la Constitution, puis en 2008 de l'article 1 introduisant ainsi l'égal accès des femmes et des hommes aux mandats électo-

raux et aux fonctions électives. Ces modifications avaient été précédées par la loi Roudy sur l'égalité professionnelle en 1983.

À la suite de ces lois cadre, des dispositions plus spécifiques ont été votées concernant l'égalité professionnelle (loi Génisson de 2001) favorisant l'égal accès des femmes et des hommes aux mandats politiques (lois de 2007 et 2008), concernant la représentation équilibrée des femmes et des hommes au sein des conseils d'administration avec un minimum de 40% de femmes (loi de 2011).

Dans le domaine de la fonction publique, c'est la loi dite Sauvadet de mars 2012 qui va mettre en place des objectifs chiffrés à hauteur de 15% de femmes en 2012, 20% en 2013, 30% en 2015 et 40% en 2018 pour améliorer l'égalité entre les sexes. De même, la loi du 22 juillet 2013 va instaurer des dispositifs chiffrés contraignants pour favoriser la parité dans l'Enseignement Supérieur et la Recherche.

Ensuite, la discipline informatique, comme l'ensemble des autres disciplines universitaires, est soumise à l'encadrement réglementaire concernant la mise en œuvre des quotas à l'université. Suite aux décrets d'application, la loi du 22 juillet 2013 introduit ainsi la parité sexuée au niveau national dans les différents conseils et au niveau de chaque université dans la constitution des listes de candidature aux élections universitaires, dans la nomination de personnalités extérieures, etc. Selon les instances d'application de ces quotas, (conseils centraux de l'établissement, comité de sélection pour le recrutement d'enseignant.es chercheur·es, etc.), les enjeux et impacts peuvent être variés. Mais ce qui est notable, c'est la multiplication de l'application de quotas dans des instances multiples, plus ou moins importantes. Il ne s'agit pas forcément d'assurer l'accès à un lieu de pouvoir comme peut le constituer un conseil d'administration dans une grande entreprise.

Le domaine de l'informatique, compte-tenu de sa composition sexuée présentée précédemment, est particulièrement concerné par les résistances et les ambiguïtés de la mise en œuvre de cette politique de quotas. Ce contexte en demi-teinte à l'université a été analysé de façon globale par Lemercier (2015). « La parité dans les conseils académiques restreints fait ainsi l'objet d'un recours de la Conférence des présidents d'université auprès du Conseil constitutionnel. Si le principe de l'amélioration de la place des femmes à l'université est rarement remis en question publiquement, les quotas de femmes font débats dans les coulisses des instances universitaires, souvent sous forme de commentaires ironiques ou d'agacements. Les arguments opposés insistent sur l'insuffisance d'un « vivier de femmes », la complexification – parfois indiscutable – d'une double voire triple parité et les résistances s'intensifient pour les espaces de pouvoir les plus importants (conseils académiques restreints et comités de sélection pour les recrutements de professeur-e-s). ».

Nous avons fait part de notre analyse à partir des résultats d'une première enquête exploratoire sur l'ambivalence des politiques de quotas dans un contexte particulier qu'est celui de l'Enseignement Supérieur et de la Recherche dans le domaine très masculinisé de l'informatique. Il s'agit de pouvoir bien identifier quelles sont ces spécificités pour

pouvoir généraliser notre réflexion sur finalement les contextes où les politiques de quotas peuvent avoir des effets contre-productifs, au-delà même du vécu individuel qu'elles peuvent amener, en ramenant sans cesse la personne, et dans ce cas précis, les femmes à leur sexe. Ainsi, au-delà de la controverse théorique et des débats suscités sur ces politiques, il s'agit de pouvoir observer, quantifier ces effets contre-productifs à l'égalité professionnelle, notamment au sens d'une accentuation du retard d'avancement dans la carrière, au-delà de ce que peut d'ores et déjà induire la maternité. Ainsi, cette étude montre que la politique des quotas peut s'avérer un frein ou un levier selon les générations d'informaticiennes et selon le moment où cette politique de quotas s'est appliquée.

Le premier résultat met en évidence que la politique de quotas peut s'avérer un levier. C'est le cas pour les informaticiennes nées dans les années 60-70 grâce à une mise en visibilité qui contrebalance les difficultés de carrière liée à des maternités qui surviennent au début de carrière.

Néanmoins, cette enquête exploratoire montre aussi que cette politique de quotas peut amener des freins au déroulé de carrière. D'une part, au niveau individuel, la politique de quotas renforce les arguments sexistes d'une promotion des femmes parce qu'elles sont femmes et non pas grâce à leurs compétences. Il s'en suit un vécu de manque de confiance et de légitimité au niveau individuel qui amène à une auto-censure en termes d'évolution de carrières. D'autre part, la sur-sollicitation des informaticiennes, notamment pour des comités de sélection, au moment même où elles sont soumises à une injonction de publications internationales pourrait freiner leur évolution de carrière.

4.5 Réflexions conclusives : des politiques d'égalité à évaluer

La question de la place des femmes est bien au-devant de la scène, induisant parfois une certaine exaspération face à certaines politiques d'égalité. C'est le cas des politiques de constitution des comités de sélection pour le recrutement des enseignant·es-chercheur·es.

Différentes actions ont été mises en place face aux divers problèmes soulevés. Nous mentionnons ici quelques initiatives sans prétention d'exhaustivité.

Tout d'abord, nous pouvons mentionner l'initiative de certaines universités à sensibiliser sur les biais de recrutement, dont deux sont mentionnées ici. Une simple recherche via un moteur permet d'accéder de façon ouverte par exemple à ce que l'Université d'Angers propose [2].

L'université de Toulouse, quant à elle, a rassemblé différents pointeurs :

— s'auto-tester [3], avec notamment les tests « Gender and Carreer » et « Gender and

2. https://moodle.univ-angers.fr/course/view.php?id=22716
3. https://implicit.harvard.edu/implicit/takeatest.html

Science » ;
— prendre connaissance de la fiche [4] proposée par le ministère sur les comités de sélection ;
— lire et s'inspirer du document [5] réalisé à l'IRIF (Institut de Recherche en Informatique Fondamentale - Université Paris Cité) [6]
— regarder les vidéos : vidéo [7] sur la première réunion - sélection des dossiers - de 4min élaborée par l'Université de Lausanne et vidéo [8] sur la seconde réunion - auditions - de 7min proposée par Sorbonne Université.

Au-delà de cette sensibilisation par rapport au déroulé de ces comités de sélection, l'enjeu initial réside dans la constitution du comité, avec ses multiples critères.

Les quotas peuvent amener une sur-sollicitation de certaines personnes. Mais il est important de noter que, finalement, cette sur-sollicitation peut ne pas concerner l'ensemble des femmes (dans le cas de l'informatique), en se concentrant sur certaines d'entre elles, particulièrement visibles, avec un effet d'entrainement. Plus elles participent à des comités de sélection, plus leur visibilité par rapport à cette tâche augmente, créant une accélération du phénomène.

Face au nombre de sollicitations (sur-sollicitations) pour participer à des comités de sélection et un positionnement de refus qui peut ne pas être évident, l'article [9] s'appuyant sur un travail quantitatif, paru en juin 2022 et rédigé par Anne-Cécile Orgerie et Camille Maumet avec la commission égalité femmes hommes de l'IRISA/Inria Rennes est très instructif. Cet article, dont l'intitulé est très parlant : « Comités de sélection : combien d'invitations accepter ? », peut constituer un repère pour faciliter le positionnement (d'un refus éventuel).

Les sollicitations à ces comités sont très dépendantes d'une forme de visibilité - visibilité qui se base plutôt sur la dimension recherche de la carrière - alors même que les composantes enseignement et recherche doivent toutes deux être représentées pour le recrutement. Il pourrait alors être pertinent de pouvoir disposer d'un annuaire qui recense les enseignantes-chercheuses et leurs domaines d'expertise (en recherche et en enseignement), un peu à la manière du site des Expertes [10]. Ce site émane d'une initiative qui visait à pouvoir fournir aux médias un annuaire gratuit, dans le contexte où, en 2020, « seulement 41% des expert·ess invité·es dans les médias français étaient des femmes ».

Mais finalement, l'un des enjeux forts réside dans l'analyse des politiques mises en place, en termes de résultats, sans partir du principe que l'intention initiale positive va garantir sa réussite. Que produit finalement la mise en place des politiques de quotas

4. https://cache.media.education.gouv.fr/file/27/15/8/ensup504_1303158.pdf
5. https://www.irif.fr/_media/postes/recommandations-parite.pdf
6. https://www.irif.fr/egalite-fh
7. https://youtu.be/TQG7zySAyaE
8. https://youtu.be/l4rCUxIBZnw
9. https://egalite-fh.irisa.fr/presentation-recommandations/comites-de-selection-combien-dinvitations-accepter/
10. https://expertes.fr/

dans les comités de sélection ?

5 Conclusion

> Certaines personnes - des hommes, bien sûr – m'ont découragée en disant que la science n'était pas une bonne carrière pour les femmes, ce qui m'a poussé encore plus à persévérer.
>
> *Citation attribuée à Françoise Barré-Sinoussi (née en 1947), chercheuse française en virologie*

Les études de genre m'ont amenée à explorer les données d'une autre manière, à envisager leur analyse sous un autre angle, ayant pour ancrage une volonté de recontextualisation, souvent socio-historique.

Les travaux développés dans cette partie sont plutôt de l'ordre de la collaboration scientifique que de l'encadrement de la recherche. Pour autant, il était important de faire une place à ces travaux, car pour être dans un rôle d'encadrement de la recherche, il est nécessaire de porter une vision de ce qui est à accomplir pour la suite (peut-être notamment pour proposer des sujets de recherche), et poursuivre ses propres travaux est une alternative qui peut avoir du sens.

Si les projets de thèse en Sciences Humaines et Sociales trouvent souvent leur origine chez les doctorant·es qui vont porter leur sujet, il est plus fréquent en informatique que les sujets de thèse soient initiés par les chercheuses et chercheurs qui vont encadrer le travail.

Ainsi, la capacité à mener des travaux, sans être nécessairement dans une forme d'accompagnement à la recherche, me paraît particulièrement pertinente.

Par ailleurs, aller explorer la mise en œuvre d'une démarche sociologique, autour des politiques de quotas dans les comités de sélection, sujet qui me tenait particulièrement à cœur, fut source d'une richesse d'apprentissages.

Cette deuxième partie du manuscrit a ainsi démontré l'intérêt de porter un regard

nourri d'interdisciplinarité sur les données. C'est l'originalité de ce regard qui a été source de résultats intéressants, du point de vue des connaissances qui ont pu être extraites à partir des données, et ce, sans forcément être dans une originalité des méthodes.

J'ai par ailleurs la conviction que cette manière d'envisager ces contributions scientifiques trouve sa source dans ce cheminement fait de rencontres avec des collègues de multiples disciplines (sociologie, histoire, littérature, science politique, science de l'information et de la communication) et bien d'autres.

C'est ainsi qu'il a été possible d'analyser l'informatique sous l'angle de dynamiques sociales, ouvrant des pistes pertinentes et originales sur la manière de se saisir d'un matériau.

Il apparaît alors intéressant de faire un pas de côté pour porter un regard sur ce qui a été fait, comment, etc. C'est ainsi que j'ai souhaité que cette HDR soit nourrie de ces réflexions issues de ce « retour en arrière ». J'ai la certitude que ce travail de prise de recul peut constituer un apport en soi.

Dès lors, après la présentation des contributions dans les deux premières parties, j'ai souhaité développer une troisième partie qui inclura un focus sur la démarche réflexive et sur l'épistémologie, avant de laisser une large place à la présentation des perspectives de recherche post-HDR.

Troisième partie

Réflexivité, épistémologie et perspectives de recherche

1 Introduction

> Nous avons tous besoin d'une permission pour faire de la science, mais, pour des raisons profondément ancrées dans notre histoire, cette permission est bien plus souvent donnée aux hommes qu'aux femmes.
>
> *Citation de Vera Rubin, née Cooper (1928-2016), astronome américaine.*

Si mon parcours académique a été marqué par mon cursus en informatique dans lequel je n'ai pas forcément eu conscience d'un besoin de permission pour faire de la science, il est sûr que la mise en œuvre de cette habilitation à diriger des recherches en a nécessité plus d'une. Sans doute parce que cette permission est intrinsèquement liée à la notion de légitimité.

La volonté de retranscrire, dans ce mémoire, la démarche de réflexion qui est la mienne, et ce bien au-delà des résultats scientifiques obtenus durant ces années, est sans doute à l'origine de ce besoin de permission.

Au-delà de tous les encouragements professionnels reçus pour aller au bout de cette mise en œuvre, qui constituent en soi autant de permissions, l'enjeu est indéniablement de se donner cette permission à soi-même. En l'occurrence, une permission de questionner, de penser ce cheminement scientifique.

Les deux premières parties de ce mémoire ont été l'occasion de présenter divers résultats (fruits du travail collaboratif, mené ou non dans le cadre d'un travail d'encadrement de la recherche). Avant de pouvoir présenter les perspectives de recherche qui s'ouvrent pour l'avenir, il m'apparaissait important de rendre compte de ma démarche réflexive et de réflexions épistémologiques. Je prends ce dernier terme dans son sens large, relatif à la production de connaissances scientifiques.

Plusieurs disciplines se sont intéressées à considérer la science comme objet de recherche. Parmi elles, nous pouvons penser naturellement à la philosophie des sciences, l'histoire des sciences, la sociologie des sciences.

Je ne suis ni philosophe, ni historienne, et pas même sociologue (même si je chemine vers une interdisciplinarité de manière que je qualifierais d'empirique en mobilisant une démarche sociologique dans certains travaux). Mais il m'apparaissait pertinent d'apporter un regard depuis ce positionnement d'informaticienne en contact étroit avec nombre de disciplines où ces réflexions, notamment sur la posture de recherche, sont présentes, voire même structurantes pour celle-ci.

Ainsi, je souhaite amener une dimension réflexive, et ce, sous l'angle expérientiel, à la lumière de mon propre cheminement intellectuel, en présentant des points saillants qui ouvriront sur des réflexions autour de la question épistémologique dans ma pratique scientifique.

En toute transparence, c'est la transmission de cela qui fait partie intégrante de l'envie de mener ce travail de rédaction d'HDR.

Réflexivité et épistémologie, ce sont deux concepts que je n'avais jamais abordés de manière explicite dans mon cursus de formation initiale. Et je ne crois pas trop m'avancer en indiquant qu'il s'agit souvent de concepts qui ne sont pas forcément beaucoup abordés dans la recherche en informatique, au vu des échanges que j'ai pu avoir avec mes collègues, depuis que mon projet d'HDR se dessine en intégrant ces aspects. Peut-être que cela est « normal » et intrinsèque à la discipline ? Ou peut-être pas ?

C'est ainsi que cette troisième partie commencera par deux chapitres qui illustreront comment ces deux concepts se sont invités (imposés) à moi, en tentant de démontrer leur pertinence quant à la recherche en informatique. Ainsi, le chapitre III.2 est un bilan réflexif qui me permet de présenter ce que je retiens comme éléments importants des recherches menées, éléments qui sont inter-reliés à la question de posture de recherche. Ensuite, le chapitre III.3 permet de retranscrire quelques éléments de réflexion en matière d'épistémologie, en tant que chercheuse en informatique qui a évolué dans un environnement de Sciences Humaines et Sociales. Ces deux chapitres m'apparaissaient nécessaires, avant de pouvoir présenter mes perspectives de recherche dans le chapitre III.4. Une conclusion viendra clore cette partie dans le chapitre III.5.

2 Bilan réflexif

> À mesure qu'on acquiert des connaissances, on apprend à douter de celles qu'on croyait certaines.
>
> *Marie-Geneviève-Charlotte Thiroux d'Arconville, née Darlus (1720-1805), femme de lettres et de sciences française, « Les pensées et réflexions morales » (1760)*

LA CITATION choisie pour ce chapitre de bilan réflexif illustre ce double mouvement que j'expérimente dans ma carrière, à savoir une avancée dans mes connaissances au fur et à mesure du temps passé à l'université et de mon expérience, et en même temps cet apprentissage de remettre en question ce qui est acquis, inhérent à la science.

Si le fait de se poser des questions est sans doute un processus de base pour toute personne « scientifique » (au sens qui veut contribuer à la science), je reconnais volontiers que tendre vers des questionnements réflexifs, en me posant des questions sur mes propres biais, sur ma manière de faire de la recherche est arrivé assez tardivement, mais peut-être pas si tardivement, relativement aux pratiques de la discipline qu'est l'informatique.

Ces questionnements, que des personnes pourraient qualifier de « métaphysiques » - avec une connotation négative - sont multiples, foisonnants. Et il n'est pas question ici de les aborder tous dans une forme de recensement, ou de les discuter, n'étant pas plus philosophe, historienne que sociologue des sciences.

Par contre, je choisis ici d'en rendre compte avec un regard d'informaticienne qui est allée explorer ce que d'autres disciplines peuvent nous appporter, un peu à la manière d'un papillon qui va de fleur en fleur, mais en prenant tout de même le temps de s'arrêter.

Ainsi, mon intention était d'apporter un bilan par rapport à mes années de recherche, bilan ne s'appuyant pas nécessairement sur des références théoriques vis-à-vis des questionnements ainsi soulevés.

Dans la section III.2.1, j'introduirai ce qui m'a amenée à la construction de ce chapitre dans un préambule. Je présenterai ensuite ces éléments de réflexivité selon trois axes. Je donnerai tout d'abord quelques éléments d'analyse de mon positionnement dans les recherches menées/encadrées dans la section III.2.2. Je développerai ensuite dans la section III.2.3 des éléments autour de la construction de catégories d'analyse, thème qui se retrouve être transversal dans différents travaux présentés précédemment. Puis, j'aborderai la dimension de la restitution des analyses en croisant cela avec la notion de responsabilité dans la section III.2.4, qui constitue un point central de ce que j'ai eu ensuite à cœur de transmettre dans certains de mes enseignements qui s'y prêtaient. Enfin, j'apporterai des réflexions conclusives permettant d'amener à la notion de posture de recherche réflexive dans la section III.2.5, qui servira de point d'appui à la discussion qui suivra autour de la notion d'épistémologie dans le chapitre suivant.

2.1 Préambule

Après avoir d'ores et déjà utilisé à de multiple reprise ces termes de réflexif/réflexivité dans ce manuscrit, il est opportun de revenir un peu sur cette terminologie avant de poursuivre.

Dans le dictionnaire des concepts de la professionnalisation, le concept est alors introduit de la manière suivante (Carnus & Mias, 2013) : « Dans le paysage que composent les concepts fondamentaux ayant servi, tout au long de l'histoire de la philosophie, à expliquer la nature de l'esprit, le concept de réflexivité se détache tout particulièrement. La réflexivité, c'est la réflexion spontanée se prenant elle-même pour objet en élaborant des critères épistémologiques d'ordre rationnel. Elle renvoie donc à un retour sur soi, une mise en relation de soi avec soi-même, caractéristique également de la définition qu'en font les mathématiques. Dans le domaine de la professionnalité, cette métaréflexion renvoie aux mécanismes que mettent en place les professionnels pour tirer parti de leur expérience et ainsi poursuivre une démarche de formation autonome ».

Au-delà de cette définition, il faut savoir que la littérature est assez foisonnante sur cette démarche, en particulier en Sciences Humaines et Sociales. L'article intitulé « Place de la réflexivité dans les Sciences Humaines et Sociales : quelques jalons » (Bertucci, 2009) a retenu mon attention. L'interrogation suivante est posée : « Quel est l'intérêt de la réflexivité pour la recherche en sciences humaines et plus largement pour tout chercheur en relation avec un terrain ? Autrement dit comment concilier l'objectivation inhérente à la recherche avec la part de subjectivité propre à chaque chercheur et que dire de soi chercheur au bout du compte en évitant l'anecdotique et le particulier ? ». Les prémices d'une réponse apportées à la suite sont : « Ceci suppose de considérer que la recherche est une expérience humaine qui se constitue en tant que recherche par le processus réflexif.[...] On insistera d'emblée sur le fait que la réflexivité n'est pas « une introspection psychologisante et autocentrée du chercheur » mais qu'elle est constitutive de la posture

de recherche car elle suppose un travail constant du chercheur sur ses positionnements, ses angles d'attaque et une réactivité permanente [...]. »

Le questionnement réflexif est un processus qui s'est invité dans mon cheminement professionnel, petit à petit. J'en attribue l'explication notamment à mon environnement au sein de l'UFR [1] Anthropologie, Sociologie et Science Politique à laquelle je suis rattachée pour l'enseignement depuis ma prise de poste en 2009, à la démarche sociologique dont je me suis rapprochée et aux études de genre qui m'ont amenée à redéfinir bon nombre de mes connaissances et ma propre posture de recherche, en les questionnant. La participation, chaque année, à un nombre de jury de mémoires de master en études de genre assez conséquent (une dizaine environ par an) y a contribué également. Ces mémoires, qui relatent un travail de recherche dans une centaine de pages, ont illustré la démarche des étudiant·es à se positionner sur un sujet de recherche de leur choix, en lien ou non avec leur terrain de stage, explicitant notamment leur rapport à leur objet de recherche, des questionnements réflexifs dans leur démarche de recherche.

Ainsi, j'attribue volontiers aux Sciences Humaines et Sociales une habitude de réflexivité, que je me suis appropriée, et qui contribue indéniablement à une prise de recul sur les travaux menés, ce qui caractérise notamment l'étape d'HDR.

Il s'agissait donc pour moi de rendre compte de manière explicite de ces réflexions, convaincue que ce processus d'explicitation participerait à la construction d'une manière de mener des travaux de recherche pour la suite, et au-delà même, d'une manière d'être enseignante-chercheuse.

2.2 Un positionnement atypique qui vient nourrir ma recherche : à la croisée des chemins

Lors de mon cursus de formation universitaire centré sur l'informatique et les mathématiques, j'avais eu l'occasion d'être initiée un peu à la sociologie.

L'occasion de prendre un poste de maîtresse de conférences en informatique au sein de l'UFR d'Anthropologie, de Sociologie et de Science Politique fut l'occasion de côtoyer plus précisément des collègues anthropologues, sociologues et politistes.

L'engagement dans la mention de master en études sur le genre depuis ma prise de poste fut aussi l'occasion d'être en lien avec des collègues en Sciences Humaines et Sociales, mais aussi en histoire, en lettres, en linguistique, en information et communication, en psychologie, et encore bien d'autres, compte-tenu de la perspective pluridisciplinaire des études de genre suivie à l'Université Lyon 2.

Ce fut un réel apprentissage qui n'est, de mon point de vue, pas neutre dans la manière d'aborder la recherche. Ainsi, si je me réfère au défi EGC 2020, le fait de traiter celui-ci sous

1. Unité de Formation et de Recherche

l'angle des enjeux sociétaux d'actualité qu'étaient la place des femmes - aussi abordée dans le défi 2016 - et l'écologie, mon bagage en études de genre et mon imprégnation en Sciences Humaines et Sociales ont été déterminants dans cette perspective d'aborder ces défis.

Compréhension sociale et dimension politique, c'est peut-être une manière de résumer l'essence de ce qui vient alimenter ma réflexion dans la manière d'aborder l'informatique et les contributions.

Cela vient alors questionner comment l'originalité d'« un regard sur » peut être valorisé, au delà d'une originalité dans la méthode, du point de vue de la science des données.

Le prix du défi EGC 2020 nous a été remis grâce à un mélange de vote du public (majoritairement composé de chercheuses et chercheurs) suite à la présentation des résultats en session plénière et d'avis du comité scientifique du défi.

Les réactions entendues à l'annonce du prix ont suscité chez moi nombre d'interrogations. En effet, il y eut des félicitations de vive voix. Mais aussi, quelques commentaires indirects, moins agréables bien entendu, qui soulignaient l'étonnement de cette remise de prix. Cet étonnement soulevait en particulier la simplicité de la méthode (statistiques descriptives ne présentant pas une contribution en termes de méthodes, position clairement affichée de notre part), et donc l'absence d'une contribution « scientifique » caractérisée par la nouveauté.

Ceci ouvre la question sur ce qui fait science au sein d'une communauté scientifique par rapport à une discipline. Ainsi, finalement, qu'est-ce qu'une contribution scientifique en science des données ? Qu'est-ce qu'une contribution scientifique en informatique ?

Je trouve que se poser ces questions n'a rien d'inintéressant et ce, notamment dans un contexte structurel de course à la publication. En effet, ce contexte de course met en scène comment chaque publication présenterait une contribution, qui souvent peut faire l'objet d'un découpage artificiel pour multiplier les publications sur un même objet de recherche.

Toujours est-il que si les réactions à ce prix ont suscité bon nombre de questionnements, il est vrai que ces deux défis de 2016 et 2020 ont aussi été l'occasion d'observer également les autres propositions de réponse et la manière d'y répondre.

Cela a contribué à accentuer ma conviction selon laquelle mon cheminement marqué par une curiosité, un intérêt et une ouverture vers les Sciences Humaines et Sociales, et vers les études de genre en particulier, nourrissent ma manière de faire de la recherche.

Je le constate notamment dans la manière de venir questionner à un niveau plus méta certains aspects, avec une perspective parfois différente. C'est le cas par exemple sur le sujet de la catégorisation abordée dans la section suivante.

2.3 Catégories d'analyse : de la construction à la visibilisation

La construction de catégories est au coeur de l'analyse de données. Dans les travaux présentés précédemment dans ce manuscrit, cette construction des catégories est présente dans de multiples contextes (et ce, de manière plus ou moins explicite) :

— elle est au cœur même des travaux sur l'expression de connaissances des utilisateur/trices pour créer des niveaux d'analyse dans le cadre des entrepôts de données ;
— elle est présente dans la modélisation adoptée des lacs de données dans certains contextes ;
— la définition de catégories pour l'apprentissage supervisé de données (comme dans le cadre de certains des travaux menés sur des données issues de *Twitter*) constitue le point de départ du travail ;
— les multiples catégories d'analyse sont incontournables pour la scientométrie, comme c'est le cas de l'analyse de la place des femmes dans la communauté EGC.

Ainsi il m'apparaît important de revenir sur cette notion de catégorie, et plus particulièrement sur le processus de catégorisation. Car il s'agit bien là d'un processus.

Il est sans doute important de préciser que cette construction ne va pas forcément de soi, ou en tous cas que cela ne devrait pas forcément être le cas. La catégorisation retenue a en effet un impact direct sur la manière d'observer/d'analyser les données, et de rendre compte ou non de certaines « réalités ».

Le questionnement sur les catégories d'analyse est une réflexion qui s'est invitée au long cours pour moi, et en particulier la notion de construction sociale des catégories, qui amène à une production sociale des statistiques.

Cette réflexion a pris un essor particulier en avril 2021, lorsque j'ai eu l'occasion de présenter un séminaire en distanciel pour l'Institut de Recherches et d'Etudes Féministes de l'Université du Québec A Montréal. J'avais choisi d'intituler ce séminaire : « Faire parler les données chiffrées avec une perspective de genre : à quel prix? ».

J'y ai évoqué notamment les travaux de Statistique Canada[2] qui correspond dans la statistique publique nationale à l'équivalent de l'INSEE[3] en France.

Statistique Canada a retravaillé les catégories d'identité de genre et d'orientation sexuelle en préparation du recensement réalisé en 2021 et dont les résultats ont été publiés en mai 2022.

Il avait été alors proposé de participer à la « consultation sur les normes de métadonnées statistiques sur la diversité de genre et la diversité sexuelle » du 2 février 2021 au 12 mars 2021. Cette consultation était publique, sans nécessairement avoir été très publici-

2. https://www.statcan.gc.ca
3. Institut National de la Statistique et des Etudes Economique https://www.insee.fr

sée (si je me réfère aux réponses que j'ai eues au sujet de cette communication auprès d'étudiant·es lors de cours que j'ai pu faire à Montréal).

Elle témoigne à la fois du travail de ce processus de catégorisation qui a fait l'objet d'une évolution, démontrant que les catégories peuvent être temporellement définies, qu'elles peuvent être également géographiques (socio-culturelles).

En effet, il est par exemple à noter la modalité de bi-spiritualité [4] qui rend compte d'une réalité catégorielle propre aux autochtones des « premiers peuples », avec lesquels un processus important de réconciliation [5] est en cours au Canada, après une colonisation marquée notamment par des violences, par la mise en place de pensionnats visant à l'assimilation culturelle.

Cela remet en perspective la dimension de construction des catégories, quand bien même l'on souhaite que ces catégories retranscrivent une « réalité », une « vérité », etc.

Ce recensement a été l'occasion d'une première récolte de données [6] donnant à voir une quantification des personnes relativement à ces catégories et leurs modalités.

Quantifier devient alors un moyen de visibilisation, notamment pour les « minorités ». Ainsi, cette mise en visibilité est inhérente à la construction des catégories et à la définition de leurs modalités.

Ces catégories, qui ont servi au recensement et donc à la quantification, ont aussi vocation à être utilisées comme variables d'analyse dans le futur, permettant notamment de mettre en lumière d'éventuelles situations d'inégalité.

Ces éléments m'ont donné à voir les impacts de ces processus de catégorisation, qui peuvent par exemple se jouer à des niveaux politiques (Lizotte, 2021), quand bien même il y aurait une volonté d'« objectivité ».

Finalement, une réflexion autour de la catégorisation s'est matérialisée dans un article dans le cadre d'une collaboration avec Marie Vialaret, statisticienne et féministe engagée, pour la troisième édition de l'atelier international qui se déroule dans le cadre de la conférence *Big Data* « Data science for equality, inclusion and well-being challenges » (Favre & Vialaret, 2023).

J'ai fait le choix ici de ne synthétiser que certains aspects de la réflexion menée, mais l'article comprend également bon nombre de références bibliographiques sur le sujet, issues de disciplines diverses également.

Le point clé ici présenté demeure la non neutralité de ce processus. Ceci caractérise, notamment en Sciences Sociales, le regard porté sur un objet de recherche avec une dimension d'analyse quantitative. Par exemple, Bacot et al. (2012) évoquent dans leur article intitulé « Le langage des chiffres en politique » le fait que l'une des questions « qui se pose concerne cette transformation de l'objet mis en chiffres, qui se trouve ainsi discré-

4. https://cihr-irsc.gc.ca/f/52214.html
5. https://www.rcaanc-cirnac.gc.ca/fra/1400782178444/1529183710887
6. https://www150.statcan.gc.ca/n1/daily-quotidien/220914/dq220914c-fra.htm

tisé, catégorisé, dans un processus de naturalisation, de réification, d'objectivation tendant à faire oublier les « conditions d'équivalence qui ont présidé à leur construction » ».

Les processus de catégorisation ont un impact direct sur l'analyse qu'il est fait des données, ce qui amène finalement une non neutralité de cette analyse. Au-delà de l'impact de ces catégories vis-à-vis de cette non neutralité, d'autres aspects viennent s'ajouter à cela, c'est ce qui est abordé dans la section suivante.

2.4 De la non neutralité des visualisations, une responsabilité dans l'analyse

Le traitement de données par l'informatique décisionnelle vise à outiller les utilisateur/trices grâce à la production de visualisations (au travers du *reporting* plus particulièrement). Le recours à l'apprentissage induit également souvent la construction de visualisations, fournis par exemple par les logiciels développés.

Il apparaissait ainsi pertinent de se questionner sur la portée de ces visualisations qui permettent de fournir des informations, et donc sur leur production dont la conception émane aussi des informaticien·nes.

Le terme d'« informer » est présent dès l'ancien français sous la forme d'« enformer ». Il vient du latin « informare » qui signifie « façonner, former » et au sens figuré « représenter idéalement, former dans l'esprit ». Ainsi, en faisant le lien entre « informer » et « mettre en forme »

Un des enjeux pour moi, réside alors dans la prise de conscience de la non-neutralité de ces visualisations, en ce sens que l'analyse qui en découle rend compte d'un point de vue, éloignant peut-être de la fameuse '« objectivité », posée souvent en étendard de scientificité, notion sur laquelle je reviendrai dans le chapitre suivant.

C'est ainsi que dans les enseignements au sein de la mention de master Études sur le genre, j'illustre depuis quelques années ce processus grâce à une image trouvée sur internet [7] représentée dans la figure III.2.1. Ainsi, considérons le fait analysé, l'objet de recherche, sa vérité, comme étant le cylindre. Le processus d'analyse (et notamment la réalisation d'un graphique) peut être vu comme la projection de ce cylindre à la lumière d'un projecteur sur un mur, qui donnerait lieu soit à un rectangle, soit à un cercle, selon le point de vue utilisé, le choix d'un graphique pouvant constituer un point de vue. Ceci est notamment le cas lorsque le graphique choisi permet de mettre en avant une manière de voir l'objet, tout en masquant d'autres facettes de cet objet.

Sur ce point, Bacot et al. (2012) exprimaient également qu'« au-delà de la mise en chiffres, il y a la mise en images, en schémas, en graphiques, en courbes, en tableaux, en camemberts et autres modélisations de toutes sortes. Il s'agit d'une seconde transfor-

7. https://www.jeuxvideo.com/forums/42-68-70371367-1-0-1-0-dans-cette-image-ou-se-situe-la-verite.htm

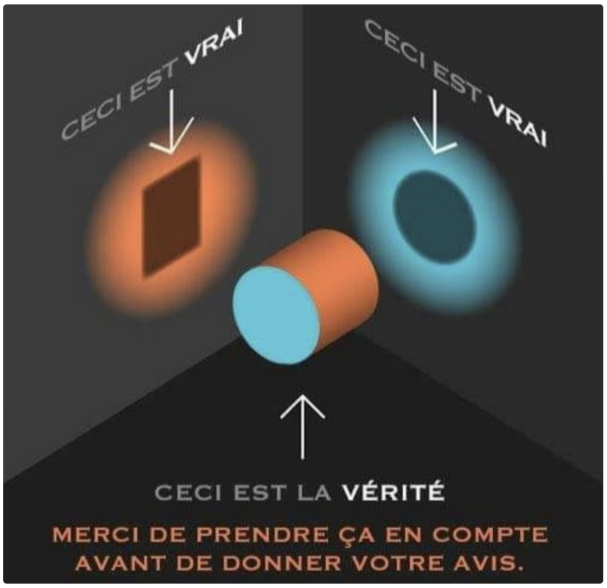


Figure III.2.1 – Représentation de points de vue.

mation sémiotique, car les chiffres eux-mêmes sont alors iconisés. Ils sont perçus comme de belles images, souvent mises en valeur par une profusion de couleurs. Dans une certaine mesure, on peut dire que cette formalisation constitue un niveau supplémentaire de réduction de la réalité, car ces modélisations ne rendent compte qu'imparfaitement des résultats chiffrés pris en eux-mêmes ».

Ceci rejoint, sur un autre aspect, la réflexion de Patrice Belot autour de la nécessité d'une pluralité des résultats présentés aux utilisateur/trices dans le domaine de l'apprentissage de données, amenant aussi la notion de responsabilité (Bellot, 2020) : « Présenter aux utilisateurs les résultats d'un seul modèle omnipotent est un choix qui doit être remis en question. Un tel modèle ne peut être suffisant car, au-delà des biais d'apprentissage et de l'attraction majoritaire centrale pour assurer la convergence de l'apprentissage, toute approche algorithmique reflète des points de vue sur la nature des données et sur la notion de pertinence. Il ne s'agit pas de problèmes propres aux traitements automatiques ou à l'intelligence artificielle : chaque humain a lui aussi ses a priori, ses opinions et sa méthode, ses algorithmes conscients et inconscients. Ça n'est pas tant la mainmise des algorithmes sur notre quotidien qui présente un danger que l'unicité des algorithmes et des modèles, associée à la non perception des orientations idéologiques et morales souvent involontaires induites par les différents biais, les algorithmes et les choix sur les données à observer, conserver ou supprimer. »

Ainsi, cette mise en images revêt un enjeu important, d'autant plus si nous considérons l'importance de ces visualisations dans les prises de décision (je pense notamment ici à l'informatique décisionnelle). Ainsi, ses préoccupations peuvent rejoindre les travaux qui se développent actuellement autour de la narration de données, et sa personnalisation pour les utilisateur/trices (Amer-Yahia et al., 2023 ; Chanson et al., 2022).

2.5 Réflexions conclusives : un pas de côté profitable pour construire une posture réflexive

Il était important pour moi de rendre compte de cette démarche réflexive acquise dans mon cheminement, au contact des Sciences Humaines et Sociales (grâce aux discussions avec les collègues mais également lectures des mémoires des étudiant·es, curiosité du questionnement nourrie par des lectures diverses, etc.).

J'ai fait le choix d'aborder ce chapitre centré sur la réflexivité plutôt par rapport à des réflexions issues de l'ensemble des contributions proposées précédemment. Il s'agissait alors d'illustrer ce qui pouvait en ressortir de pertinent, pour nourrir la réflexion en termes de posture professionnelle.

Je suis convaincue que cela apporte beaucoup et qu'une fois cette dimension intégrée, aucun retour en arrière n'est possible pour fonctionner autrement.

Finalement, si l'on considère le processus collaboratif d'alimentation de Wikipédia pour observer la synthèse vulgarisée qui a été faite de ce concept de « réflexivité », la page ainsi constituée précise entre parenthèses le rattachement aux Sciences Sociales, en étant vue comme une démarche méthodologique. Il est précisé ensuite que « plus généralement, une démarche réflexive en science consiste en une prise de conscience et en un examen approfondi de sa propre démarche scientifique. Le chercheur doit réaliser qu'il s'inscrit lui-même dans des traditions culturelles, dans des cadres sociaux, etc. Il s'agit de sortir des « mécanismes d'explications » qui donnent l'illusion de comprendre son objet d'analyse de façon transparente. Pour une bonne « réflexivité », le chercheur doit comprendre son habitus et ses schèmes sociaux afin d'objectiver sa relation à l'objet (pourquoi ça l'intéresse, etc.). Cette réflexion doit être réalisée à deux reprises : lors de l'élaboration du terrain et lors de l'interprétation des données. Elle rappelle fortement l'idée de neutralité axiologique. »

Ce terme de « neutralité axiologique » nous amène à la notion de posture de recherche. Ainsi, cette réflexivité qui amène à questionner différents points sur la manière de faire de la recherche notamment, m'a amenée au concept d'épistémologie, qui constitue le point central du chapitre suivant.

3 Réflexions épistémologiques

> Si vous savez que vous êtes sur la bonne voie, si vous avez cette connaissance intérieure, alors personne ne peut vous décourager... peu importe ce qu'ils disent.
>
> *Barbara McClintock (1902-1992), cytogénéticienne américaine*

ÉPISTÉMOLOGIQUE est un adjectif relatif à l'épistémologie qui peut elle-même être définie de bien des manières. S'il y avait une évidence à présenter ces réflexions, il a été aussi question de pouvoir construire cette démarche par des lectures. En voulant étayer mes propos par des lectures, face à ces multitudes de références pouvant être identifiées comme pertinentes pour aborder ce sujet, je qualifierais de vertigineux l'effet de prendre conscience du travail que cela représentait.

Je prends le terme d'épistémologie dans son sens large relatif à la production de connaissances scientifiques. Plusieurs disciplines se sont intéressées à considérer la science comme objet de recherche. Parmi elles, nous pouvons penser naturellement à l'histoire des sciences, la philosophie des sciences, la sociologie des sciences.

Ces disciplines, associées à ce champ qu'est l'analyse de la science, sont très inspirantes et ont beaucoup à nous apprendre. Je n'en suis pas spécialiste, mais le cheminement que je fais vers une interdisciplinarité de manière que je qualifierais d'empirique m'amène à des réflexions que je voulais partager, et ce depuis le début de ce projet d'HDR.

J'ai souhaité construire dans ce chapitre un bilan expérientiel de ma perception de la question épistémologique dans ma pratique scientifique car c'est une question qui n'avait jamais été abordée de manière explicite dans mon cursus, et je ne crois pas trop m'avancer en indiquant qu'il s'agit souvent (et non toujours) d'une question impensée dans la recherche en informatique.

Et si le contenu de mon HDR a souvent été discuté lors de rencontres avec de multiples

collègues, avec parfois le conseil de rester dans un format plus classique pour arriver plus rapidement au terme de sa rédaction, il y aura eu mille occasions de renoncer à ce chapitre. Mais comme l'indique la citation choisie et présentée plus haut, rien ne pouvait aller à l'encontre de ce qui m'avait donner envie d'aller vers ce projet d'HDR (faire part de mes réflexions sur la manière d'envisager la recherche en informatique) avec cette profonde conviction que j'étais « sur la bonne voie ».

Réfléchir sur la manière d'envisager la recherche en informatique pourrait faire l'objet d'une vaste discussion en soi, que je trouverais passionnante (sans que cette opinion fasse l'unanimité au sein de la communauté scientifique, j'en ai bien conscience). J'ai choisi ici de retenir trois axes qui me tiennent à cœur et qui ont émergé de mon cheminement.

Ces trois axes seront présentés après la section III.3.1, où j'introduirai quelques éléments en préambule de ce chapitre. J'aborderai donc ensuite, dans la section III.3.2, la dimension de savoirs situés. Puis je questionnerai alors les aspects de disciplines dans la section III.3.3, avant de porter un regard sur les questions d'engagement en science dans la section III.3.4. Enfin, je tenterai d'apporter une conclusion sur ces réflexions dans la section III.3.5, qui servira de point d'appui au développement de mes perspectives de recherche qui seront développées dans le chapitre suivant.

3.1 Préambule

Il est délicat de parler d'épistémologie sans revenir sur toute une contextualisation historique de sa définition, sur des manières d'appréhender ce concept, qui peuvent être inhérentes à des disciplines notamment. Ainsi, en choisir une définition, qui serait sans doute la bienvenue pour éclairer la lecture de ce chapitre, n'est pas aisé non plus.

J'ai donc plutôt choisi ici de donner l'intention que j'avais en utilisant ce terme, à savoir discuter ce qui est relatif à une posture de recherche au sein de la discipline informatique, en la nourrissant de questionnements et réflexions qui sont plus habituels dans d'autres disciplines.

Par posture de recherche, je n'entends pas aller vers une discussion sur les méthodes scientifiques employées pour la validation, qui peuvent être une manière d'aborder l'épistémologie. Je souhaite plutôt aller vers ma propre compréhension de la manière dont j'envisage la recherche en informatique, en terme de positionnement.

En 2017, naissait mon idée d'organiser un atelier qui avait l'objectif de créer un espace de partages d'échanges autour d'une perception de l'informatique qui pour moi évoluait au fur et à mesure d'une réflexion nourrie à la fois par mon imprégnation des questions (et des travaux) en études de genre, et des multiples contacts (et collaborations) à la fois au niveau pédagogique et scientifique avec des personnes issues de disciplines différentes (sociologie, science politique, histoire, sciences de l'information et de la communication, lettres, etc.).

« Penser la Recherche en Informatique comme pouvant être Située, Multidisciplinaire Et Genrée », avec l'acronyme « PRISME-G [1] », c'est avec cette idée que je sollicitais Pascale Kuntz qui était alors chargée de mission égalité dans son établissement, pour une co-organisation dans le cadre de la conférence EGC 2018.

L'appel à communications était introduit de la manière suivante :

> Aujourd'hui, la question de l'égalité femmes-hommes est largement médiatisée et l'institutionnalisation des études de genre sur le plan à la fois de la recherche et de la formation propose un cadre de réflexion pour penser cette question sociale qui est transversale et qui touche à des aspects de politiques publiques, de carrières, d'éducation, de recherche, etc. Ces réflexions concernent pleinement l'informatique qui est un domaine fortement genré à la fois en tant que domaine économique et en tant que discipline académique. Les questions soulevées concernent la conférence EGC sous deux angles : naturellement par sa place dans le domaine de l'informatique mais aussi pour ses contributions méthodologiques puisque la pluridisciplinarité des questions soulevées nécessite le traitement de données multi-sources et multi-échelles.
> Cet atelier vise à proposer un espace d'échanges et de discussions autour de la question de l'égalité femmes-hommes qui concerne aussi bien les objets de recherche et les personnes qui font la recherche en informatique. Les thèmes seront abordés au travers d'exposés invités pour poser un cadre de réflexion et au travers de propositions de communications relatant des retours d'expériences ou proposant une réflexion sur les thèmes qui suivent.
> Il s'adresse à des enseignant·es chercheur·es en informatique qui s'intéressent aux aspects épistémologiques de leur discipline, des enseignant·es-chercheur·es ancré·es dans d'autres disciplines que l'informatique et s'intéressant à la discipline de l'informatique.

L'idée de cet atelier venait déjà mettre en mots une manière d'appréhender la recherche. Les exposés invités, les contributions de personnes en Sciences Humaines et Sociales dont l'informatique était l'objet de recherche, ainsi que la table ronde, sont venus nourrir la réflexion.

Les études sur le genre m'avaient déjà amenée à la notion de « savoirs situés ». Elles m'avaient aussi conduite à renforcer une manière d'appréhender les disciplines de façon décloisonnée. Finalement, les enjeux des études sur le genre, éminemment politiques, m'avaient aussi interpellée autour de la question du rôle des sciences dans la société.

3.2 Savoirs situés

En 1988, la biologiste et philosophe féministe Donna Haraway théorise le concept de « situated knowledges », qui peut être traduit par « connaissances ou savoirs situés » (Haraway, 1988).

1. https://eric.univ-lyon2.fr/prisme-g/

Ce concept a été développé à la fois en réaction à une conception dominante de l'objectivité scientifique, selon laquelle la ou le scientifique pourrait tout voir depuis n'importe où, et en réaction au relativisme, qui anéantit les objectifs de l'objectivité en mettant sur un pied d'égalité toutes les opinions.

L'émergence des *Cultural Studies* dans les années 1960 qui s'intéressent à de nouveaux objets et qui déplacent le regard des sciences sociales vers les marges et les cultures dites « minoritaires » amène une transformation de l'étude des sciences dites exactes. C'est ainsi que, par exemple, les travaux du chercheur Bruno Latour, connu pour ses contributions en sociologie des sciences, vont souligner l'importance des conditions matérielles et politiques qui influencent le contenu des énoncés scientifiques.

Ainsi, au moment où émerge une critique de la prétendue « neutralité » de l'objectivité scientifique, est théorisée la notion de savoirs situés.

Les savoirs situés impliquent le questionnement 1) de la position depuis laquelle la connaissance est produite d'une part, mais aussi 2) des limites qu'impliquent cette position sur sa vision, et encore 3) des relations de pouvoir qui peuvent exister.

Cette manière d'aborder des questionnements a notamment permis d'entrevoir, dans une perspective de genre, comment finalement l'identité de chercheuse ou chercheur, sa construction sociale, etc. a un impact sur la recherche.

S'il peut être assez aisé d'envisager ces points sur un travail en Sciences Humaines et Sociales, comme la sociologie par exemple, il est intéressant / utile / voire nécessaire de projeter cette idée sur un domaine comme l'informatique, en particulier à la science des données par exemple. La problématique autour de biais dans les données d'apprentissage est à un certain niveau intrinsèquement liée au fait de ne pas avoir intégré cette réalité des savoirs situés. Les travaux autour des questions de biais, et plus largement de *fairness* se développent dans le domaine de l'apprentissage de données, comme en témoigne par exemple l'article *survey* de Choudhary et al. (2022), avec un focus sur les données de type graphe. C'est l'objet même du projet Diké [2], financé par l'Agence Nationale de la Recherche (ANR) française, concernant plus spécifiquement les modèles de langue.

Il est également intéressant d'évoquer la notion de « strong objectivity », traduit par « objectivité forte », qui renvoie aux travaux de la philosophe féministe Sandra Harding (Harding, 1992), et qui est reliée à cette notion de savoirs situés.

Ainsi, la thèse défendue est qu'il y a un biais dans la recherche puisque les chercheurs masculins tentent d'être des chercheurs neutres, ce qui n'est pas possible a priori. Sandra Harding appelle à une réflexivité, une prise en compte de sa propre position et de la manière dont la recherche peut être affectée, qui donne lieu à une objectivité « plus forte » que dans la situation d'une prétendue neutralité.

Cette reconnaissance d'une forme de subjectivité, qui peut s'avérer dérangeante selon la réprésentation que l'on se fait de la science, est pourtant nécessaire pour poursuivre un

2. https://www.anr-dike.fr/

idéal d'une plus grande objectivité.

Le récent ouvrage de Courau et al., 2022 sur le genre des sciences souligne la nécessité de re-travailler les sciences dites dures (comme la médecine, la biologie, la pharmacologie, etc.) en prenant en compte les enjeux liés au genre. Ainsi les diverses disciplines peuvent être traversées par ces enjeux de situation des scientifiques qui y contribuent. L'informatique n'y échappe pas non plus.

Notons enfin que cette question des savoirs situés est beaucoup développée actuellement dans le contexte canadien, avec la reconnaissance des savoirs autochtones, notamment sur les enjeux écologiques, et une manière spécifique d'aborder le développement de ces savoirs (« construction » et transmission). Il est alors question de développer des espaces pour penser la co-construction des savoirs. C'est ce que j'ai pu constater en assistant à une table ronde « Langues et savoirs autochtones : voies vers le développement durable » organisée par l'UNESCO[3] dans le cadre de la 91ème édition du Congrès de l'Acfas (organisme à but non lucratif contribuant à l'avancement des sciences au Québec, dans la francophonie canadienne et sur la scène francophone internationale). C'est dans ce contexte, qu'un numéro des cahiers de Centre interuniversitaire d'études et de recherches autochtones (CIÉRA) a été diffusé en juin 2023[4], intitulé « La pertinence des épistémologies autochtones face à la crise climatique actuelle : Enjeux de protection et de préservation du territoire »(Ngono & Şükran, 2023).

Ainsi, ce principe de savoirs situés sous-tend différentes épistémologies qui ont en commun le fait de remettre en questions la possibilité d'une neutralité absolue du ou de la scientifique, qui permettrait de produire une théorie générale.

C'est ainsi qu'aborder des questions avec des perspectives disciplinaires différentes peut constituer un intérêt certain.

3.3 Du décloisonnement disciplinaire à l'interdisciplinarité en passant par le nomadisme disciplinaire

Lorsqu'il est question de décloisonnement disciplinaire, plusieurs concepts apparaissent et tentent de trouver leur place les uns par rapport aux autres : pluridisciplinarité, multidisciplinarité, interdisciplinarité, transdisciplinarité, etc. Leur définition peine à trouver un consensus, et ces termes ne sont pas toujours tous positionnés les uns par rapport aux autres dans la littérature, ce qui a même amené à la proposition d'une taxonomie (Kleinpeter, 2013).

Dans cette section, il m'importera plutôt de discuter mon regard sur mon chemine-

3. Organisation des Nations Unies pour l'Education, la Science et la Culture : institution spécialisée internationale de l'Organisation des Nations Unies

4. https://ciera-recherches.ca/publications/liste-des-numeros-parus/1899/?fbclid=IwAR2biFZP5tdTFT0JrNWfjCooeNuH1oC1CfSPKWbMYib4GMLNvxv0wgFropg

ment par rapport au décloisonnement disciplinaire que de rentrer dans le détail des définitions, tout en m'appuyant sur le besoin de quelques unes que j'emprunte au travail mené au Laboratoire interdisciplinaire littérature et mathématiques de l'Université de Sherbrooke au Canada [5].

Sur l'« interdisciplinarité », je retiendrai notamment : qu'elle consiste en « la mise en relation d'au moins deux disciplines, en vue d'élaborer une représentation originale d'une notion, d'une situation, d'une problématique » (Maingain et al., 2002). Par ailleurs, un travail interdisciplinaire peut être vu comme un « processus dans lequel on développe une capacité d'analyse et de synthèse à partir des perspectives de plusieurs disciplines ».

Ces dernières années ont vu émerger un discours prônant les collaborations pluridisciplinaires et l'interdisciplinarité, notamment dans les dotations d'appels à projet ou des créations de commissions interdisciplinaires dans certaines instances (comme au CNRS par exemple).

Pour autant, il est aisé de constater que le fonctionnement académique disciplinaire en silo prédomine (par exemple au travers des sections disciplinaires du CNU (Conseil National des Universités). Ceci n'est pas anodin non plus en termes d'évolution de carrière, comme l'indiquait le chercheur Julien Jourdan, professeur à HEC Paris Business School, dans son article « La pluridisciplinarité, un frein pour les chercheurs dans leur avancement de carrière [6] » dans *The Conversation*.

Ainsi, les enjeux de contributions scientifiques dans des contextes interdisciplinaires font notamment l'objet d'un besoin de discussions. Ce fut le cas lors de la séance en amont du $5^{ème}$ symposium du GDR MaDICS (Groupement de Recherche Masses de Données, Informations et Connaissances en Sciences) en mai 2023.

Il apparaît pertinent de pouvoir accéder à une réflexion sur le contexte actuel de l'interdisciplinarité comme l'ouvrage « Disciplines Academiques en Transformation. Entre Innovation et Resistance » (Paradeise et al., 2015) mais aussi sur la mise en œuvre de celle-ci, en analysant également les obstacles (Prud'homme & Gingras, 2015).

Ma proximité avec les Sciences Humaines et Sociales, les études de genre m'ont amenée à ce que je qualifierais de « nomadisme discplinaire ». Ce terme est venu de la déclinaison de ce que j'entendais de Laurence Tain [7] sur le fait d'avoir en recherche des sédentaires et des nomades.

Après avoir utilisé ce terme, une recherche bibliographique m'a amenée à constater son émergence dans des articles en sciences sociales. J'associe ce terme à la capacité que j'ai pu développer à aller m'imprégner des méthodes de la recherche d'autres dis-

5. https://www.usherbrooke.ca/litt-et-maths/fondements/inter-trans-multi-pluri-ou-intradisciplinarite/

6. https://theconversation.com/la-pluridisciplinarite-un-frein-pour-les-chercheurs-dans-leur-avancement-de-carriere-195984

7. Laurence Tain est historiquement agrégée de mathématiques et est devenue professeure de sociologie, avec une spécialisation en études de genre.

ciplines (en sciences sociales comme la sociologie en particulier, mais pas seulement), à lire des références qui permettent de croiser des points de vue sur des problématiques pour enrichir mon regard, etc. Plus largement, c'est l'occasion d'observer les différences de pratiques au sein des disciplines, par exemple en assistant aux soutenances de thèse et d'HDR de disciplines diverses (en sociologie, en information et communication, en littérature, etc.).

Ce nomadisme m'a permis de me « décentrer », pour également procéder à une forme d'analyse comparative implicite des pratiques scientifiques. Celle-ci n'est pas formalisée a proprement parler, mais elle a d'ores et déjà donné lieu à quelques réflexions.

Convaincue de la richesse que m'a apporté ce nomadisme disciplinaire, c'est une manière de construire petit à petit une approche interdisciplinaire, en m'appuyant sur un décloisonnement disciplinaire.

Ainsi, il s'agit de prendre ici un peu de hauteur sur l'organisation des disciplines, avoir cette vision d'en haut comme l'aigle qui est en même temps capable de regarder les détails grâce à son acuité visuelle.

Le mariage de l'informatique avec d'autres disciplines tend à se généraliser. C'est le cas notamment des humanités numériques, dans lesquelles je m'inscris par ailleurs. Ce fut aussi le cas de la biologie et de l'informatique dont l'alliance a donné lieu à un véritable champ disciplinaire à part entière reconnu.

Les travaux de Yann Renisio sur l'« origine sociale des disciplines » (Renisio, 2015) dans une perspective d'analyse quantitative mettent en lumière une hiérarchisation des disciplines, notamment véhiculée par des chercheuses et chercheurs.

Les politiques de financement de la recherche illustrent également cette hiérarchisation (à déconstruire), notamment des sciences dites « dures » versus des sciences dites « molles ». Ainsi, compte-tenu du fait que la recherche en informatique apparaît être plutôt bien dotée, il m'importe d'avoir une vigilance sur les guichets de financement dans le cadre de collaborations pluridisciplinaires, en ayant une posture que l'on peut caractériser d'alliée aux Sciences Humaines et Sociales.

Ceci questionne également la manière de considérer le croisement des disciplines pour construire des contributions, et pourquoi la valorisation doit s'opérer de manière cloisonnée du point de vue disciplinaire.

Ceci peut questionner, notamment si l'on se réfère historiquement au fait que des chercheuses et chercheurs endossaient plusieurs étiquettes disciplinaires, sans cette nécessité d'enjeu de valorisation disciplinaire revendiquée systématique.

Il est à noter que l'articulation de l'informatique avec d'autres disciplines est sans doute un argument de taille pour attirer davantage de filles en informatique, compte-tenu de leur socialisation qui les amène par exemple, en tendance, à un attrait pour le médical, ou les Sciences Humaines et Sociales.

Quand on évoque le nomadisme, il est important de préciser que pour moi il s'agit

néanmoins de préciser l'existence d'un point d'ancrage, dans la discipline qu'est l'informatique. Si je poursuis la métaphore du nomadisme, il est important de préciser que les peuples autochtones du Québec étaient des peuples nomades, ce qui impliquait qu'ils revenaient de manière cyclique à des endroits bien connus. Le nomadisme ne veut pas dire qu'il n'y a pas d'endroit de référence, en l'occurence pour moi la discipline de l'informatique.

Cette circulation dans les disciplines n'est pas toujours confortable, car elle amène notamment des remises en question, mais elle est toujours source de richesse. Elle permet de s'imprégner de connaissances, de les faire circuler, tout comme appréhender de nouveaux concepts. Nous avions évoqué dans la partie 2 du manuscrit la circulation géographique, ici il est question de la circulation entre les disciplines.

C'est ainsi qu'un des enjeux que je vois dans cette profusion des contributions abordées, tout en maintenant des recherches en profondeur, est un réel besoin de création de ponts, des jonctions entre les disciplines, et ce de manière solide.

Il s'agit donc d'accepter que la science ait un cheminement, et qu'il y a plusieurs manières de parcourir ce chemin. Aller à la rencontre des disciplines, c'est accepter de prendre des routes moins rapides, voire des chemins. Il faut le temps de l'exploration. Et c'est ainsi peut-être que l'on découvre parfois des trésors.

Il y a peut-être donc un enjeu à ne pas valoriser que la rapidité en nombre de kilomètres parcourus (avec la « production » de nombreux papiers), questionnant les critères d'évaluation de la recherche plus globalement.

Pour ma part, faire preuve de ce nomadisme disciplinaire peut être vertigineux, car sortir de son champ direct d'expertise nous amène à l'immensité de ce qe nous ne connaissons pas, à de l'humilité assez logiquement aussi.

Fabien Labarthe, maître de conférences en sciences de l'information et de la communication à l'Université Jean Monnet Saint-Etienne, dont un des champs d'expertise concerne les usages et les enjeux du numérique, m'avait soufflé le concept de « marginal sécant » à l'écoute de quelques éléments de mon parcours.

Dans son ouvrage intitulé « Pour une hybridation entre arts et sciences sociales » dans le chapitre « Le chercheur comme passeur et « marginal sécant » » (Grésillon, 2020), Boris Grésillon, professeur des universités en géographie, écrivait :

« À l'opposé du touche-à-tout, le passeur est celui qui approfondit en les pratiquant plusieurs genres (littéraires, scientifiques ou artistiques) – de même qu'il pratique également, souvent, plusieurs langues, ce qui lui permet d'être un passeur de langues, donc un passeur de mondes – et qui, par sa grande culture et sa plasticité intellectuelle, est capable de tisser des liens inédits entre ces genres et de les hybrider. Définie ainsi, la figure de passeur est voisine de celle du « marginal sécant » identifiée par les spécialistes de l'innovation dans les années 1970 et utilisée par Michel Crozier et Erhard Friedberg dans leur théorie dite de sociologie des organisations. Pour eux, le « marginal sécant » est « un acteur qui est partie prenante de plusieurs systèmes d'action en relation les uns avec les

autres » (Crozier & Friedberg, 1977). À la manière de « l'entrepreneur de méthode », il est capable d'entendre et de comprendre des discours différents et de les mettre en lien ; à la manière du passeur, il est capable de s'appuyer sur d'autres disciplines que la sienne pour mettre en lumière ce qu'elles ont en partage. La position de marginalité dans son domaine l'oblige d'autant plus à tisser des liens vers d'autres domaines, ce qui permet, à terme, de faire émerger une nouvelle situation de carrefour créateur. Ou quand la marge devient centralité.

Dans ce manifeste pour des sciences sociales « transversives », la figure du passeur, comme celle du « marginal sécant », est centrale ; et elle est inspirante. Même si elle est très exigeante intellectuellement et socialement, elle confère aux chercheurs contemporains un horizon, voire un idéal. Ma démarche, que cet essai illustre, s'inscrit clairement dans cette ligne qui est plus qu'une ligne scientifique : c'est une ligne de conduite intellectuelle. Ainsi, il s'agira de travailler non seulement sur les ponts [...] mais aussi et surtout sur les fils, plus ténus [...] ».

Le « marginal sécant » est ainsi un « acteur qui est partie prenante dans plusieurs systèmes d'action en relation les uns avec les autres et qui peut, de ce fait, jouer un rôle indispensable d'intermédiaire et d'interprète entre des logiques d'action différentes, voire contradictoires ».

Pour ma part, dans ce cheminement, j'y ai vu l'intérêt de collaborer, d'enrichissements mutuels, et c'est en tous cas l'occasion de choisir des sujets de recherche, avec une manière peut-être plus « holistique » de les aborder, ce qui répond au besoin de problèmes qui émanent directement de la société, ce qui donne un sens à notre travail de les aborder ainsi selon moi, ce point est détaillé dans la section suivante.

Il est à préciser que ce propos ne remet pas du tout en cause l'intérêt d'une recherche en informatique plus fondamentale ou d'une recherche en informatique plus cloisonnée en profondeur, mais prône une véritable complémentarité des manières d'envisager ces formes de recherche.

3.4 Science, sens, société et engagement

Initialement, le titre que je souhaitais choisir pour ce mémoire était : « De l'analyse informatique de données de la société à l'analyse sociale de l'informatique. Un cheminement guidé par les études de genre vers un décloisonnement disciplinaire et une posture réflexive en science pour la société. »

La longueur du titre m'a amenée à renoncer à la fin de celui-ci, sans pour autant renoncer au projet que cette expression de « science pour la société » sous-tend.

En mai 2021, la présidente de l'Université Lyon 2, Nathalie Dompnier, mettait en place une vice-présidence « en charge des relations Sciences et société », en démarrant la lettre

de mission pour Julia Bonaccorsi [8] de la façon suivante :

« En tant que Vice-présidente en charge des relations sciences et société, vos missions consisteront à mettre en œuvre la politique de l'établissement en matière de valorisation, de médiation scientifique, de diffusion des savoirs et d'ouverture de l'université sur le monde socio-économique, en étroite collaboration avec l'ensemble de l'équipe présidentielle.

Vous conduirez une politique visant à affirmer la « troisième mission » de l'université et développerez les liens Sciences - société pour renforcer la place de l'université au cœur de la société, non pas seulement par les disciplines qu'elle porte et qui interrogent les sociétés dans leur diversité et leur complexité, mais aussi par le rôle qu'elle entend jouer dans son environnement, auprès des acteurs socio-économiques, culturels et institutionnels. »

Dans la temporalité, cette création coïncide avec la présentation en avril 2021 par la Ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation de l'époque (Frédérique Vidal) des « mesures issues de la Loi de Programmation de la Recherche (LPR) visant à renforcer les relations entre science et société [9] ».

La Loi de Programmation de la Recherche a été l'objet de multiples réactions et mobilisations, dénonçant une « réforme en trompe-l'œil », avec des financements concentrés sur une recherche sélective, par projets, et une attaque du statut de fonctionnaire, en raison des nouvelles voies de recrutement des jeunes chercheuses et chercheurs, mobilisations finalement interrompues par la pandémie.

Sa promulgation (au 24 décembre 2020) vient finalement aussi asseoir la relation science-société en mettant en avant que celle-ci « doit désormais être reconnue comme une dimension à part entière de l'activité scientifique ».

Cette relation n'est bien évidemment pas nouvelle, y compris si l'on considère la partie financement de la recherche, par exemple au travers des financements de thèse CIFRE (Convention industrielle de formation par la recherche). Le dispositif naît en 1981 [10].

L'informatique est la discipline dans laquelle le dispositif est majoritaire si l'on se réfère au rapport de 2020 « Évaluation des effets du dispositif Cifre sur les entreprises et les doctorants [11] » établi par deux chercheurs de l'Institut des politiques publiques qui ont eu accès aux données du dispositif.

Ce dispositif concernait toute entreprise privée de droit français. Toutefois, depuis 2005, le ministère a décidé d'ouvrir la procédure Cifre à des structures non industrielles et peuvent ainsi désormais bénéficier du dispositif Cifre : Collectivités Territoriales, Asso-

8. https://www.univ-lyon2.fr/universite/gouvernance/julia-bonaccorsi-vp-sciences-et-societe
9. https://www.enseignementsup-recherche.gouv.fr/fr/science-avec-et-pour-la-societe-les-mesures-issues-de-la-lpr-49218
10. https://www.anrt.asso.fr/sites/default/files/anrt_cr_colloque_15_mars_40ans_cifre.pdf
11. https://www.entreprises.gouv.fr/files/files/en-pratique/etudes-et-statistiques/etudes/evaluation-dispositif-cifre-rapport-final-octobre-2020.pdf

ciations à vocation sociale, ONG, Etablissements consulaires : Chambres de Commerce et d'Industrie, d'Agriculture, des Métiers.

Afin de renforcer les liens entre recherche et action publique de l'État, les ministres de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI) et de la Transformation et de la Fonction publiques (MTFP) ont lancé, en mars 2022, une expérimentation portant sur le déploiement de conventions de formation par la recherche en administration (COFRA).

Ayant moi-même soutenu une thèse en financement CIFRE dans une banque, les années qui viennent de passer m'ont poussée à la réflexion.

Cette articulation science-société, avec différents acteurs, qui contribue sans aucun doute à pouvoir donner du sens à la recherche, ne doit pas faire oublier les enjeux de positionnement scientifique vis-à-vis d'un financement de la recherche, loin d'être neutre. Ceci peut finalement nous ramener aussi au sujet des savoirs situés, d'une dimension politique sur laquelle je vais revenir un peu plus loin.

Avec Erhard Friedberg, dans « L'Acteur et le Système », en 1977, Michel Crozier présente les éléments d'une théorie organisationnelle de l'action collective. Pour lui, la théorie sociologique n'est pas une fin en soi. Elle doit être utile, produire une connaissance pratique, une connaissance qui puisse être un outil du changement en permettant aux personnes de mieux comprendre la situation dans laquelle elles se trouvent et donc, d'être mieux à même de la changer. Très engagé, il a toujours cherché à faire coïncider son activité de recherche avec son engagement pour la réforme de la société et de l'État français.

Pour aller au-delà des propos de Michel Crozier, dans cette dimension d'articulation science-société, il s'agit, par ailleurs, de mentionner également le développement des formes plus concrètes de « recherche-action ». Ceci pourrait faire l'objet en soi d'une discussion importante tellement la littérature sur le sujet a été développée, avec des prismes différents, notamment disciplinaires. Je retiendrai ici, à travers cette notion, l'idée d'une recherche qui vise la transformation, et qui souvent part du principe d'une recherche « avec » plutôt qu'une recherche « sur ». La notion d'utilisateur/trice étant souvent très présente en informatique, cette forme de recherche mériterait sans doute d'être réfléchie dans la perspective disciplinaire de l'informatique.

Il s'agit aussi d'avoir en tête la spécificité peut-être de l'informatique en termes de « productions », car au-delà de connaissances, il s'agit parfois d'une production de logiciels. Ceci pose donc la question de ses destinataires, question qui peut relever d'une considération elle-même politique.

Considérons par exemple les enjeux de l'*open source*, c'est toute une « philosophie politique » qui est sous-jacente.

Ainsi la politique peut ne pas être très loin des considérations informatiques, même si cela peut être subtile, et parfois de manière plutôt implicite.

Comme le précisait Pestre (2006, p. 94) dans son « Introduction aux Sciences studies » :

« Malgré ce qu'énonce la philosophie spontanée des savants, il est difficile de laisser la science dans sa tour d'ivoire et de faire comme si c'était dans l'univers clos des laboratoires que se passait l'essentiel de ce qui la concerne. Parce que savoirs et pouvoirs ont beaucoup en commun, la question du rapport des sciences aux techniques, à l'économique et au politique (comme aux sociabilités et à la domination masculine) ne peut être évacuée ».

Cela amène bien évidemment la question de la recherche comme engagement (si l'on considère par exemple les études féministes de manière très explicite). Mais l'on peut considérer qu'adopter une perspective de genre dans une démarche scientométrique se positionne également sur une dimension politique.

Viennent alors les questions éthiques, en terme notamment de responsabilité dans notre posture de recherche. Les enjeux éthiques, notamment de formation, sont importants. Par exemple, l'école doctorale Infomaths de l'Université de Lyon précise, quant aux formations doctorales [12], l'obligation de suivre deux formations de 15h en ligne qui portent sur « Ethique de la Recherche [13] » porté par l'Université de Lyon et « Intégrité scientifique [14] » porté par l'Université de Bordeaux.

Cette question éthique comprend aussi le choix d'un objet de recherche. En 2018, sortait un article écrit par deux chercheurs de Stanford qui a porté sur la détection de l'orientation sexuelle à partir d'images de visages (Kosinski & Wang, 2018). Cette publication a eu un retentissement particulier dans certains médias en ligne. Par exemple, le magazine « Sciences et avenir » avait intitulé un article « Un algorithme plus fiable que l'humain pour deviner l'orientation sexuelle, vraiment ? [15] »

Les auteurs ont mis en place un document de réponse [16] qui a été ajusté dans le temps (dernière modification en 2022) pour repréciser leur intention et des points de vigilance.

Cette contribution pose en effet la question en termes de la relation entre recherche et société, et de la portée de la recherche dans la société. Ainsi, cette contribution est questionnable du point de vue de son intention et sa portée, en considérant, entre autres, que nombre de pays pénalisent encore aujourd'hui l'homosexualité par la peine de mort.

En effet, en 2023, les relations homosexuelles sont toujours illégales dans plus de 60 pays, et être homosexuel·le est passible de la peine de mort dans douze pays du monde [17] : l'Afghanistan, l'Arabie saoudite, Brunei, les Émirats arabes unis, l'Iran, la Mauritanie, le Nigeria (dans certaines régions seulement), l'Ouganda, le Pakistan, le Qatar, la Somalie et le Yémen.

12. https://edinfomaths.univ-lyon1.fr/these/formations-doctorales
13. https://www.fun-mooc.fr/fr/cours/ethique-de-la-recherche/
14. https://www.fun-mooc.fr/fr/cours/integrite-scientifique-dans-les-metiers-de-la-recherche/
15. https://www.sciencesetavenir.fr/high-tech/intelligence-artificielle/un-algorithme-plus-fiable-que-l-humain-pour-deviner-l-orientation-sexuelle-des-gens-vraiment_116423
16. https://docs.google.com/document/d/11oGZ1Ke3wK9E3BtOFfGfUQuuaSMR8AO2WfWH3aVke6U/edit
17. https://fr.statista.com/infographie/30849/pays-dans-lesquels-homosexualite-est-un-crime/

Des comités d'éthique sont parfois mis en place pour veiller à différents points de vigilance. Il est sans doute pertinent que la liberté académique, qui, en France, est encadrée par la loi, et exprimait notamment le fait de ne pas avoir de pressions économiques, politiques ou autres dans l'exercice de la recherche, soit pensée à l'aune du contexte actuel. Mentionnons d'ailleurs l'existence d'un indice de liberté académique qui indiquait en 2023, via l'éditeur de presse en ligne du secteur de l'éducation AEF info [18] que « 22 pays ont connu une baisse nette de la liberté académique entre 2012 et 2022. Ce recul concerne à la fois des régimes autoritaires et démocratiques. ».

Il s'agit donc aussi peut-être de pouvoir réfléchir à cette articulation entre liberté académique, financement de la recherche et science pour la société.

Il m'apparaît en tous cas essentiel, que la recherche que je mène ait du sens, en s'inscrivant dans un contexte qui n'est pas celui d'il y a quinze ans, ni par rapport à là où j'en étais dans mes réflexions.

Cette question du sens a émergé de manière assez importante ces dernières années, notamment vis-à-vis des enjeux écologiques. C'est ainsi que des initiatives voient le jour et permettent de proposer des espaces de réflexivité. C'est le cas par exemple de l'atelier SEnS [19] (Sciences, Environnements, Sociétés) dont voici les objectifs décrits sur leur site :

> L'atelier SEnS a pour objectifs de :
> — Fournir un espace et un cadre de discussion collective sur les conséquences de nos recherches, les valeurs qu'elles véhiculent et le rôle de la recherche scientifique dans l'Anthropocène.
> — Sensibiliser aux sciences humaines et sociales, et en particulier à la philosophie, l'histoire, et la sociologie des sciences.
> — Amorcer un travail collectif de construction d'une responsabilité sociale et environnementale de la recherche
>
> Il ne cherche pas à mettre tout le monde d'accord, mais plutôt à offrir des moyens de se positionner sur les enjeux environnementaux, au-delà des calculs d'empreinte environnementale.

Finalement, développer une science qui fait sens pour soi, et pour la société, constitue en soi un engagement éminemment politique, individuel avant tout, mais qui peut être l'occasion de fédérer du collectif : une opportunité de transformation (qui peut elle aussi être à la fois individuelle et collective). Ceci rejoint pleinement la dédicace de mon manuscrit : « Au cheminement individuel dans, avec, grâce et au service du collectif… »

18. https://www.aefinfo.fr/depeche/688745-plus-d-une-personne-sur-deux-vit-dans-un-pays-ou-la-liberte-academique-a-baisse-significativement-en-dix-ans-ifa

19. L'atelier SEnS a été conçu par Sophie Quinton (chercheuse et chargée de mission « Sciences, Environnements et Sociétés » à l'Inria Grenoble) et Eric Tannier (chercheur à l'Inria Lyon), dans le cadre du collectif SEnS-GRA (Sciences, Environnements et Sociétés à l'Inria Grenoble Rhône-Alpes) :
https://sens-gra.gitlabpages.inria.fr/atelier-impacts-recherche/

3.5 Réflexions conclusives : pour une épistémologie en informatique

Les différents points abordés dans ce chapitre ont permis de souligner des aspects relatifs à la manière dont j'envisage d'aborder la recherche en informatique. La réflexion épistémologique m'apparaît incontournable aujourd'hui, mon incursion en études de genre n'y est pas pour rien.

Le travail mené par Marie-Jeanne Zenetti, maîtresse de conférences en littérature, pour son HDR intitulée « Politiques des savoirs et des représentations : littératures documentaires (XXe-XXIe) » constitue justement un travail remarquable autour des savoirs situés et des épistémologies féministes, retraçant et analysant la littérature sur le sujet (Zenetti, 2023).

Cette notion d'épistémologie en informatique a émergé petit à petit, et nécessite sans doute de faire l'objet de discussions au sein des communautés scientifiques en informatique. En effet, aborder notamment cette non-neutralité dans l'informatique n'est pas courant, voire même reconnu. Dans son ouvrage « Politiques féministes et construction des savoirs. « Penser nous devons »! », de la Bellacasa (2013) précise que « [...] dans la culture académique occidentale l'exigence de faire preuve d'objectivité s'est aussi construite sur l'effort de ne pas éveiller le soupçon d'un quelconque manque de neutralité vis-à-vis des faits ».

Pour autant, ces questions épistémologiques ont fait émerger des travaux, y compris pour aborder le champ disciplinaire de l'informatique.

En janvier 2010, Gilles Dowek, connu comme informaticien et philosophe, faisait un exposé intitulé « Informatique et épistémologie [20] ».

Chantal Morley, professeure émérite à l'Institut Mines-Télécom - Business School (IMT-BS), avait présenté les bases d'une épistémologie féministe, en système d'information, et en informatique plus généralement lors de l'atelier PRISME-G [21]. Son parcours atypique a été marqué par l'obtention en 2008 du master Sociologie, mention Genre, Politique et Sexualité de l'EHESS (Ecole des Hautes Etudes en Sciences Sociales), alors qu'elle était titulaire d'un doctorat en sciences de gestion, avec une spécialité sur les systèmes d'information.

En 2018, Eglantine Schmitt soutenait une thèse intitulée « Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des big data » (Schmitt, 2018).

Le résumé de la thèse est formulé de la façon suivante : « L'enjeu de cette thèse de philosophie des sciences était de répondre à un problème pratique, qui s'est présenté à moi alors que je travaillais comme analyste sur des données massives chez un éditeur de logi-

20. https://edutice.hal.science/edutice-00560705/file/a1002d.htm
21. https://eric.univ-lyon2.fr/prisme-g/

ciel : comment produire des connaissances valides en manipulant de grandes masses de données que je n'ai pas constituées et qui ne sont pas le fruit d'une méthode scientifique reconnue ? En m'appuyant sur mon expérience de terrain, je propose un paradigme méthodologique pour la construction, l'exploration et l'interprétation des données massives. Par paradigme méthodologique, on entend un cadre théorique et pratique qui fournit autant de clés pour développer une méthode adaptée aux données et au projet épistémique envisagés. En faisant la part du mythe des big data et des pratiques effectives, je montre comment les données numériques, toujours déjà manipulables, sont construites techniquement et épistémologiquement à partir des traces laissées par les individus et leurs médiations sur les supports informatiques. Cette construction s'appuie sur une logique de constitution qui requiert un cadre interprétatif et une continuité épistémique entre la donnée et les connaissances que l'on cherche à produire. Les sciences de la culture fournissent ainsi un cadre nécessaire, mais pas suffisant, à assurer cette continuité. Le calcul, incarné par les sciences des données et l'intelligence artificielle, actualise et instrumente cette continuité au prix du renoncement à s'envisager comme fin en soi. Ni le cadre théorique, ni le calcul, ne suffisent toutefois à rendre intelligibles les connaissances ainsi produites : c'est la médiation de l'interprétation, du récit et de la conception logicielle (ou design) qui matérialise, donne à voir et contextualise les connaissances ainsi produites. Ces dernières, enfin, ne sont pas légitimées par leur pur caractère véridictionnel, mais par leur capacité à proposer ou faciliter des décisions d'action, bien souvent en entreprise, elles-mêmes productrices de nouvelles traces numériques manipulables. »

En 2021, Dubois et Brault (2021) proposaient un manuel d'épistémologie pour l'ingénieur.e.

Certaines disciplines avec lesquelles la distance peut paraître moins importante peut-être qu'avec les Sciences Sociales, comme c'est le cas des Sciences de gestion, sont peut-être à explorer du point de vue des propositions épistémologiques, comme par exemple les travaux de Avenier et Gavard-Perret (2012) et Avenier (2019).

L'informatique est finalement une discipline assez jeune qui a acquis ses traditions, ses normes de publications et d'évaluation. Elle a également acquis, parfois non sans difficulté, même encore aujourd'hui, sa reconnaissance en tant que science, etc.

Ainsi la dimension épistémologique de la discipline peut être parfois un impensé. Il est intéressant de regarder les questionnements épistémologiques sur les disciplines récentes, comme c'est le cas par exemple en science de l'information et de la communication. Ainsi, dans Fondin (2001), les questions posées sont les suivantes : « en tant que chercheur en Science de l'information : quel phénomène voulons-nous comprendre ? Quelle est la place et quel est l'objet de cette science ? Quelle est la posture épistémologique des chercheurs qui, implicitement ou explicitement, revendiquent leur appartenance à cette science ? ».

Ainsi, les ponts entre informatique et Sciences Sociales me paraissent plus que nécessaires à bâtir ; et ce, même s'ils ne sont pas forcément simples à bâtir. Car entre l'in-

formatique et les Sciences Sociales, c'est un entre-deux. En 1992, Serres (1992) retracait sa formation, son passage des sciences à la philosophie, expliquant sa méthode et situant sa pensée par rapport à la réflexion contemporaine, avec l'aide de Bruno Latour sous forme d'entretiens. Il précisait : « Ces espaces entre sont plus compliqués qu'on ne le pense[...]entre les sciences dures et les dites sciences humaines, le passage ressemble à un rivage dentelé, parsemé de glaces et variable[...] plutôt fractal que vraiment simple. Moins une jonction dominée qu'une aventure à courir. Voilà un espace étrangement dénué de chercheurs. ».

Je précise ici que, si mon propos est davantage tourné sur ce que peuvent apporter les Sciences Humaines et Sociales à l'informatique, cela ne signifie aucunement que l'inverse n'est pas vrai, bien au contraire, au regard par exemple de l'expansion des humanités numériques.

S'interroger d'un point de vue épistémologique sur notre posture et ce qui en découle m'apparaît incontournable. Quelle science voulons-nous ? Je crois à la discussion collective, et non à la pensée individuelle pour répondre à ces questions, qui pourraient en apparence peut-être paraître inutiles ou simplistes.

Il y a ainsi, de mon point de vue, un enjeu fort à ce que ces travaux qui amènent ces aspects d'épistémologie en informatique soient davantage diffusés et mieux connus au sein des communautés informatiques. Si l'informatique a gagné ses galons de reconnaissance en tant que science selon des chemins divers, le questionnement épistémologique est à mon sens un point important de cette reconnaissance.

Le chapitre précédent a amené un bilan réflexif sur le chemin de recherche qui a été parcouru, et le présent chapitre, lui, m'a permis de mettre en avant des réflexions en termes de posture de recherche, en les abordant sous l'angle de la notion d'épistémologie.

Formuler ces éléments constituait un véritable défi d'écriture, mais également une conviction forte sur la nécessité de l'expression de cette prise de recul. C'est finalement le terreau de la manière d'envisager la suite de mon parcours scientifique, à la fois en termes d'encadrement de la recherche, mais aussi d'objets de recherche. C'est ainsi que j'aborde dans le chapitre suivant mes perspectives.

4 Perspectives de recherche

> Ce travail est une grande aventure scientifique, mais c'est aussi une belle aventure humaine, l'humanité a fait des pas de géant, mais ce que nous savons est vraiment très, très petit par rapport à ce que nous devons encore savoir.
>
> *Citation attribuée à Fabiola Gianotti (née en 1960), physicienne italienne, directrice générale du CERN[a] depuis 2016*

a. CERN : Conseil Européen pour la Recherche Nucléaire : https://home.cern/fr

LES PERSPECTIVES DE RECHERCHE d'un mémoire d'habilitation à diriger des recherches constitue un espace privilégié de restitution de la réflexion sur la projection pour les années à venir, allant bien au-delà des perspectives directes développées en général à plus ou moins long terme dans les publications.

Il ne s'agit pas non plus de l'exercice de fin de thèse qui peut consister à tenter de structurer des perpectives, même à long terme, où l'on ne sait finalement pas si elles auront l'occasion réelle d'être développées ou non. À cette étape-là d'écriture lors de la thèse, cette incertitude tient au questionnement quant à une réelle poursuite de ce qui a été proposé qui est bien évidemment lié au contexte de travail à venir, en terme de poste après la thèse notamment, mais également aux opportunités, aux envies qui émergeront, dans une discipline où les objets de recherche peuvent être amenés à évoluer finalement assez rapidement, comparativement à d'autres disciplines. C'est en tous cas mon expérience si je regarde les travaux que j'ai développés après mon doctorat.

Je me saisis donc de cet espace pour livrer ma vision du travail de recherche que j'entends mener après cette étape dans les années à venir, les incertitudes quant aux possibilités de mener à bien ces recherches étant, pour de multiples raisons, sans doute plus relatives à cette étape de la carrière. Je garde en tête aussi cette idée soufflée de pouvoir imaginer ces perspectives sans limitation, en osant peut-être même le terme d'ambition,

au sens de la volonté de contribuer scientifiquement à des perspectives avec un impact sociétal positif non neutre.

4.1 Préambule

« Choisir, c'est renoncer » est une citation attribuée à André Gide. Se projeter sur l'avenir c'est faire des choix. Et pour une personne comme moi, qui voudrait explorer tellement de pistes, le problème n'est pas tant de choisir que de renoncer... Rattrapée par la réalité temporelle de journée de 24h, il s'agit alors de garder l'axe de ce qui fait sens pour moi, en ayant en tête ce proverbe autochtone entendu et réajusté : « fais en sorte que ce que tu fais aujourd'hui ait du sens pour toi, car tu vas y échanger une journée de ta vie ».

À quoi je choisis de consacrer mon temps de recherche ? À une science qui fait sens pour moi.

Avec qui ? Avec des personnes qui vont partager cette belle aventure humaine que peut constituer la recherche comme c'est énoncé dans la citation introductive de chapitre de Fabiola Gianotti.

C'est un luxe que de se poser la question de ces choix. J'en ai bien conscience. Alors il s'agit pour moi que ce luxe puisse permettre de contribuer à ma mesure, bien au-delà de raisonner en terme de carrière.

Ainsi, il ne s'agit pas de seulement aborder les perspectives en termes de contenu, mais aussi le « avec qui » d'où le fait que ces perspectives soient d'ores et déjà pensées parfois en termes de collaborations, car la science, c'est du collectif.

Il s'agit aussi de poursuivre ma manière de faire cette recherche, avec une forme de nomadisme, de continuer à faire des ponts, plutôt que de me restreindre à une seule voie vers laquelle j'irais en profondeur, multipliant par là-même les occasions de rencontres et d'enrichissements intellectuels et humains.

Le titre choisi pour ce manuscrit pourrait laisser croire à un cheminement linéaire qui amènerait à délaisser l'informatique pour une analyse sociale de celle-ci. Il n'en est rien. Mais il est important pour moi de garder cette articulation qui peut amener à réfléchir à la fois sur les pratiques individuelles de l'informatique, et à un niveau plus collectif, à l'échelle de la discipline également.

Je choisis ici de ne pas décliner ces perspectives de manière assez classique selon une temporalité (court/moyen/long terme), mais plutôt selon ce que je qualifierais de facettes constitutives de cette « identité scientifique » qui s'est construite pas après pas. Il est clair pour moi que segmenter ces facettes revient à une catégorisation un peu artificielle, où il faut bien voir une porosité des contours de cette catégorisation. J'évoquerai les perspectives en tant que chercheuse en informatique, puis celles en tant que chercheuse en études sur le genre avant d'aborder celles en tant que chercheuse interdisciplinaire. Bien entendu, ces perspectives, quelle que soit la catégorie de rattachement, sont envisagées à

l'aune de mon identité pleine et entière.

4.2 Une recherche en tant que chercheuse en informatique

Les deux axes que constitue l'informatique décisionnelle et la science des données dans lesquelles se sont inscrites les contributions en informatique présentées précédemment vont être poursuivis.

4.2.1 Une informatique décisionnelle inclusive accessible

Concernant l'informatique décisionnelle, il s'agit de poursuivre le travail sur la dimension collaborative, qui s'inscrit dans le prolongement des personnalisations d'analyse, pour y inclure une dimension collective.

Dans l'ouvrage « Statactivisme. Comment lutter avec des nombres » (Bruno et al., 2014), le résumé commence par « Les statistiques nous gouvernent. Argument d'autorité au service des managers, elles mettent en nombres le réel et maquillent des choix qui sont, en fait, politiques. Le parti pris de ce livre, qui rassemble les contributions de sociologues, d'artistes et de militants, procède du judo : prolonger le mouvement de l'adversaire afin de détourner sa force et la lui renvoyer en pleine face, faire de la statistique une arme critique. ». Cet ouvrage permet d'envisager les statistiques sous l'angle d'un regard critique, sans présenter un plaidoyer pour l'abolition de celles-ci.

Si les statistiques nous gouvernent, l'informatique décisionnelle n'est a priori pas en reste... Ainsi l'enjeu d'une informatique décisionnelle accessible au-delà des personnes sur des postes de décision est réel. Un accès pour toutes et tous. C'est l'ambition du projet BI4people porté par Jérôme Darmont, qui a le soutien de l'ANR. Le projet arrive à son terme quant au financement, mais je reste convaincue qu'il y a encore beaucoup à faire et que ce projet a été l'occasion d'ouvrir des prémisses.

L'objectif du projet était de pouvoir amener une accessibilité au-delà de la nécessité de maîtriser des compétences en décisionnel, l'enjeu d'aller davantage vers des utilisateur/trices aux profils variés, dans une démarche inclusive avec une interface qui l'est, est un autre challenge, qui s'avère de grande ampleur.

D'une part, considérons les utilisateur/trices dans le contexte des entreprises pour lesquelles l'informatique décisionnelle a été conçue préalablement. Si l'on considère l'inclusivité dans une acception large, il s'agit par exemple de travailler sur les différentes formes de handicap et des spécificités à développer, compte-tenu de l'accès de personnes en situation de handicap à ces postes en entreprise. Des enjeux se situent alors au niveau de la modélisation de ces spécificités elles-mêmes. Bien évidemment ce travail nécessite une articulation avec toute une littérature autour des questions du handicap. Je pense notamment au handicap visuel, mais d'autres angles pourraient être pertinents. J'y reviens

ci-dessous.

D'autre part, il s'agit de penser l'accessibilité au-delà des entreprises. En effet, les enjeux sont d'autant plus importants sur un autre plan, à l'ère des données ouvertes et de la possibilité de se saisir de données sur des sujets variés, notamment dans le contexte de la démocratie participative. Ceci s'inscrit directement dans la poursuite du projet, avec une accentuation d'un travail à réaliser avec la dimension des usages. Il s'agit de bien repenser la conception du système, en allant vers une conception de système d'information décisionnel que l'on pourrait qualifier de « responsable ». Ce concept de responsabilité peut regrouper un ensemble de critères à discuter. Ma participation à un atelier participatif sur les Systèmes d'Information Responsables et leur enseignement au sein d'INFORSID constitue une base de réflexions qui seront amenées à se développer. La dimension écologique fait partie intégrante de cette réflexion.

Au-delà de la dimension collaborative permettant de faire collectif, l'accessibilité est un enjeu fort pour que le collectif, dans le cadre de cette informatique décisionnelle, ne soit pas excluant.

L'inclusion est une valeur centrale du vivre ensemble dans la société. L'informatique peut simultanément être un facteur favorisant l'inclusivité et produire de l'exclusion, de la discrimination. Ceci fut l'objet de la table ronde « Science et société » animée par Sabine Loudcher dans le cadre des Entretiens Jacques Cartier 2023 sur « Tensions sociales et contexte numérique : équité, diversité et inclusion en France et au Canada », table ronde à laquelle j'ai eu la chance de pouvoir apporter mon regard.

Ainsi, par rapport aux dimensions très visuelles de l'informatique décisionnelle, il y a un défi à relever pour les personnes non voyantes, alors que de nombreuses avancées remarquables ont été faites en quelques années en termes d'accessibilité à l'information notamment, grâce au numérique, comme cela a été évoqué dans la table ronde.

L'application *Be My Eyes* est un outil important en soutien aux personnes non voyantes. Une intelligence artificielle permet de fournir un descriptif des images qui lui sont soumises.

Prenons par exemple la photo de la figure III.4.1 prise lors de cette table ronde. Le descriptif qui est fourni par l'application est le suivant :

« La photo montre une salle avec des murs de couleur orange et un haut plafond. Il y a un grand écran sur le mur du fond qui affiche des images de quatre personnes avec le texte « Science et société ». Il y a une grande table en bois au centre de la pièce avec trois femmes assises de l'autre côté. L'une d'elles porte des lunettes et un haut vert, une autre porte un haut coloré et la troisième porte un haut noir. Il y a des bouteilles d'eau et des tasses sur la table. Au premier plan, on voit l'arrière de la tête d'un homme et d'une femme qui semblent être assis en face des femmes à la table. Il y a aussi un grand blason sur le mur à côté de l'écran. »

Je dois reconnaître que j'ai été assez bluffée par les résultats de cette application, interpellée également du point de vue du genre quant aux reconnaissances faites. L'image fait

Figure III.4.1 – Table ronde « Science et société » dans le cadre des Entretiens Jacques Cartier 2023 sur « Tensions sociales et contexte numérique : équité, diversité et inclusion en France et au Canada »
Sur la photo de gauche à droite : Lise Wagner, Cécile Favre et Sabine Loudcher.
Crédit photo : Emilie Stora.

l'objet d'éléments descriptifs qui devraient aller bien au-delà si l'on veut envisager l'analyse de tableaux et de graphiques, éléments constitutifs de l'informatique décisionnelle.

Au-delà de la collaboration qui a eu lieu dans le cadre de l'ANR BI4people avec des volets visualisation de données et usages (notamment avec des personnes d'information et communication), il s'agit de pouvoir collaborer avec des personnes concernées que ce soit dans le contexte académique ou au-delà.

4.2.2 Science des données pour les données de la science

Les travaux en apprentissage auxquels j'ai participé ces dernières années ont porté sur les microblogs, et plus particulièrement Twitter pour la mise en œuvre. Comme évoqué précédemment, le rachat de Twitter et les changements de politique qui s'en sont suivis remettent sérieusement en cause la perspective de poursuite de travaux sur ces données de mon point de vue.

Lors de la soutenance d'Abderrazek Azri sur des approches multimodales pour aider à la détection de rumeurs dans les microblogs, Guillaume Cabanac a soulevé l'intérêt de

pouvoir explorer ces approches dans un contexte de détection, non pas de fausses informations, mais de faux articles scientifiques. Ce regard sur ce travail fait suite au travail mené par Guillaume Cabanac, Cyril Labbé et leurs collègues autour de la dépollution de la littérature scientifique (Cabanac et al., 2022), notamment sur la base de « phrases torturées » [1].

Les images font partie intégrante des articles scientifiques. Ainsi, explorer la dimension visuelle par rapport à la qualité des papiers constitue une piste qui pourrait être pertinente. Les disciplines scientifiques sont nombreuses, mais elles n'ont pas toutes recours à des images. Il s'agirait donc dans un premier temps de définir un domaine de travail. Cette perspective pourra être envisagée avec la collaboration d'Abderrazek Azri, qui poursuit une carrière d'enseignant-chercheur en Algérie, et Cédric Wemmert, professeur d'université à Strasbourg spécialisé notamment sur l'apprentissage de données et les images. Travailler sur les données de la science constitue une continuité par rapport aux travaux déjà entrepris, et cela relève d'une utilité certaine.

Un deuxième volet concernant les données de la science porte sur la poursuite de travaux en lien avec le prisme des enjeux sociétaux. Il s'agit notamment de poursuivre le travail sur l'aspect écologique avec Sébastien Valat et Guillaume Cabanac, ce qui constitue une question complexe dans notre contexte actuel. À plus court terme, cela concernera la finalisation et la publication des travaux menés.

Enfin, il s'agit de pouvoir explorer également comment les travaux sur les lacs de données peuvent être abordés dans une perspective scientométrique. En effet les travaux de Lin et al. (2023) qui proposent un lac de données ouvert rassemblant des données sur la science sont tout à fait intéressants et inspirants. Cette perspective serait intéressante à creuser puisqu'elle se situe également à la jonction de différents travaux déjà menés sur les lacs de données, par rapport aux données de la science, qui constituent une thématique de grand intérêt pour moi.

Si les travaux présentés comme perspectives portent sur les données de la recherche, d'autres perspectives peuvent également être envisagées à partir de ces données, avec l'enjeu d'un regard au prisme des enjeux sociétaux comme cela avait été initié dans les défis présentés précédemment. Il ne s'agirait pas d'analyser l'ensemble des communautés en informatique, mais plutôt d'aller vers une analyse d'autres disciplines pour mieux connaître celles-ci.

En effet, questionner notre propre discipline au regard de la connaissance d'autres apparaît tout à fait intéressant. Dans la continuité de la réflexion épistémologique, c'est un moyen de pouvoir remettre en perspective certains fonctionnements de la recherche. Par exemple, l'analyse des formes de valorisation dans les différentes disciplines amène au constat de l'importance de l'internationalisation dans notre discipline. Nous sommes nombreuses et nombreux à regretter le désinvestissement des communautés nationales, là où pour d'autres disciplines, les publications en français sont importantes.

1. https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener

Les disciplines peuvent être analysées sous plusieurs angles. Les pratiques au niveau du CNU peuvent constituer un de ces angles. Et si la section informatique du CNU a publié une note sur les publications en juin 2023 pour rappeler notamment un principe de qualité plutôt que de quantité, cette instance joue son rôle en contribuant à normer les pratiques de chercheuses et chercheurs compte-tenu de son rôle d'évaluation dans la carrière. Et si était imposé par exemple pour la qualification d'avoir publié au moins un article dans la communauté française de son domaine de spécialité ?

Je ne peux que constater dans mon parcours à quel point, sur la dimension Sciences Humaines et Sociales, les lectures de textes en français ont été très présentes (même si cela ne s'y limite pas). Je reconnais que c'est constitutif de ma construction interdisciplinaire, sans doute parce que cela représente pour moi aussi une finesse dans les mots exprimés, dans les formulations d'idées, dans la construction de la pensée, qui est plus délicate sur une langue étrangère. C'est ainsi que des travaux sur les données de la science autour de la citation de travaux de diverses langues, comme ceux développés par Bertin et Atanassova (2023), présentent un intérêt certain.

4.3 Une recherche en tant que chercheuse en études sur le genre

Les discussions autour de genre et informatique ont aujourd'hui un retentissement important à différents niveaux, et notamment au sein des lieux de formation, plutôt en raison de la masculinisation de ces formations d'ailleurs.

Isabelle Collet fut précurseuse sur ce sujet avec sa thèse en Sciences de l'Éducation intitulée « La masculinisation des études informatiques : savoir, pouvoir et genre »soutenue en 2005 sous la direction de Nicole Mosconi, à une époque où ce sujet n'intéressait pas vraiment, comme elle le rappelle régulièrement.

Au vu des enjeux qui se posent en apprentissage automatique de données autour des questions de genre, je souhaite pouvoir développer cette partie, afin de mettre à profit tous les apprentissages que j'ai faits autour du genre compte-tenu de mon parcours. Cela concerne à la fois la question de l'évaluation des modèles de langue vis à vis des stéréotypes de genre (travail entamé avec une doctorante du laboratoire Irina Proskurina qui travaille dans le projet Diké autour des biais dans les modèles compressés d'apprentissage profond de langue [2]).

Un autre volet concerne les traductions automatiques et comment se fait le genrage des individus lors des traductions. Une étude approfondie a démarré grâce au projet de Travail de fin d'Etudes et de Recherche (TER) que j'encadre avec deux étudiantes du diplôme M1 IDSM-Kharkiv en partenariat avec l'Ukraine, afin de challenger les systèmes de traduction et voir comment ces systèmes réagissent autour de la question du « genrage »

2. Projet ANR Diké : https://www.anr-dike.fr/

TABLEAU III.4.1 – Tests sur *Google Translate* par rapport au genrage de données, avec le hongrois pour lequel les pronoms sont neutres, report des traductions réalisé en novembre 2023.

Français	Français -> Hongrois	Hongrois -> Anglais	Hongrois -> Français
il fait la cuisine	ő főz	she/he cooks	elle cuisine
elle fait la cuisine	ő főz	she/he cooks	elle cuisine
il fait le ménage	ő intézi a házimunkát	she/he does the houseworks	elle fait le ménage
elle fait le ménage	ő intézi a házimunkát	she/he does the houseworks	elle fait le ménage
il peut s'occuper des enfants	ő tud vigyázni a gyerekekre	she/he can take care of the children	elle peut s'occuper des enfants
elle peut s'occuper des enfants	ő tud vigyázni a gyerekekre	she/he can take care of the children	elle peut s'occuper des enfants
il est passionné par l'informatique	szenvedélyesen rajong a számítógépekért	she/he is passionate about computers	il est passionné d'informatique
elle est passionnée par l'informatique	szenvedélyesen rajong a számítógépekért	she/he is passionate about computers	il est passionné d'informatique
l'informatique est une passion pour lui	számára az IT szenvedély	for him, IT is a passion	pour lui, l'informatique est une passion
l'informatique est une passion pour elle	számára az IT szenvedély	for him, IT is a passion	pour lui, l'informatique est une passion

des phrases, notamment quand le point de départ est une langue qui ne genre pas les personnes, comme c'est le cas avec le hongrois par exemple. Les petits tests que j'avais réalisés sur *Google Translate*, outil très utilisé, reportés dans le tableau III.4.1, sont assez éloquents. En effet, on y retrouve des stéréotypes pour la traduction en français (la cuisine est par exemple associée aux femmes et l'informatique aux hommes). Par ailleurs, malgré la prise en compte du genre en anglais, on peut également constater que la double traduction (féminin et masculin) ne fonctionne que pour les phrases où le pronom personnel est sujet. Ainsi, cette problématique de traduction est un sujet que je souhaite approfondir compte-tenu de l'apprentissage de données qui sous-tend ces traductions et de l'importance de ce qui est véhiculé dans les langues.

Ces travaux permettront de poursuivre la réflexion plus en profondeur autour des biais d'apprentissage, et reprendre les travaux portant sur les catégories qu'on cherche à apprendre (reflet de la manière de penser le monde, de l'organiser) en discutant plus préci-

sément de la dimension de genre, en s'inspirant notamment du travail de recatégorisation qui a été mené par Statistique Canada autour de l'identité de genre.

4.4 Une recherche pluri/interdisciplinaire favorisant les collaborations au-delà de l'informatique

Je souhaite axer mon projet autour des enjeux d'éducation, de transmission des savoirs, et plus particulièrement dans un contexte d'apprentissage à distance. La période COVID a accéléré les enjeux de ce type d'apprentissage. Cela a permis de repenser des méthodes pédagogiques. Et si le retour en présentiel apparaît pour moi nécessaire pour la transmission, cela pousse à réfléchir sur les plateformes dédiées à la transmission de connaissances. En effet, c'est une manière de faciliter la circulation des connaissances, et donc leur accessibilité de manière plus globale, peut-être au-delà de notre système d'apprentissage. Ainsi il est ici question notamment de didactique.

Il s'agit bien sûr d'alimenter cette recherche également du point de vue du genre, mais pas seulement.

Un premier volet concerne l'apprentissage de l'informatique. Les travaux menés en articulation avec la discipline « information-communication » avec une perspective de genre sur des plateformes où l'enseignement de programmation est diffusé constituent une base de travail, pour réfléchir en matière de pédagogie de l'enseignement, et notamment peut-être pour réfléchir aux enjeux de l'attrait de la discipline pour des jeunes filles. En effet, dans le cadre du projet SO COEQUAL[3], le stage réalisé par Noé Vaccari a donné à voir la dimension de genre dans l'apprentissage de l'informatique sur des MOOC (Massive Open Online Course). La poursuite de cette recherche pourrait permettre d'établir des recommandations, de soulever des points de vigilance.

Un deuxième volet concerne l'usage de l'intelligence artificielle pour l'éducation. J'ai participé au montage d'un atelier dans le cadre du GDR MaDICS[4] (Masses de Données, Informations et Connaissances en Sciences) intitulé EDUC'ACTION, piloté par Sihem Amer-Yahia, Emilie Hoareau et Philippe Dessus. Cette équipe s'appuie sur une multidisciplinarité où l'informatique, les Sciences de Gestion et les Sciences de l'Éducation sont représentées. Cet atelier porte sur l'éthique dans les Sciences de l'Éducation et de la Formation. Ces travaux s'inscrivent dans la perspective des travaux sur l'AIED (Artificial Intelligence for EDucation) et proposent d'aborder la question de la montée en compétences sous les angles informatique et Sciences Humaines et Sociales, fondée sur une réflexion éthique permettant de mettre en balance les bienfaits et les préjudices de l'intelligence artificielle. Ainsi l'enjeu est d'articuler une réflexion sur l'éthique pour permettre

3. SuppOrt COmputer EQUALity : vers une égalité femmes-hommes de l'apprentissage en informatique par une analyse genrée des IHM

4. https://www.madics.fr/

une conception de système, ou dans un premier temps des points de vigilances / recommandations.

L'intégration à la fédération de laboratoires SFR-RELYS (Recherche en éducation Lyon Saint-Etienne) qui regroupent des laboratoires de plusieurs disciplines va permettre de développer également des contacts pour mener à bien ce projet.

Un dernier volet a été inspiré par mon séjour au Québec, avec une imprégnation des cultures autochtones, où la tradition orale est un axe fort de la transmission, en s'appuyant notamment sur un ensemble de dimensions dont certaines correspondent à des aspects de Sciences de l'Éducation. La problématique est alors de réfléchir à comment inclure d'autres dimensions d'apprentissage. Un projet est en cours de montage avec une collègue de l'Université de Montréal en études autochtones (Dolorès Contré), autour de la pédagogie par symboles. Il s'agit d'un projet qui d'apparence se situe moins dans mon champ de compétences, mais il est vrai que les échanges que nous avons eus m'ont convaincue d'envisager ce projet très sérieusement, convaincue de la valeur de ce qui pourrait en découler.

4.5 Une recherche au-delà du prévisible

En 2007 je soutenais ma thèse de doctorat réalisée dans un contexte de financement CIFRE au sein d'une banque, autour de la personnalisation des analyses avec une perspective d'analyse de données commerciales, et de demandes marketing qui visaient à vendre davantage de produits bancaires.

À aucun moment je n'ai pu imaginer ce que serait mon parcours, ce cheminement dans l'informatique, guidée notamment par les Sciences Humaines et Sociales et les études de genre, quand bien même j'avais une appétence pour la sociologie.

La science est faite de rencontres, d'opportunités, toutes deux parfois inattendues.

Dans un contexte d'articulation des temps de vie, avec toute la difficulté de mettre des limites qu'implique un métier-passion, il s'agira de faire des choix parmi ce qui se présenterait au-delà de ce que j'ai prévu.

Mais grâce à ce cheminement qu'a permis la réalisation de cette habilitation, je sais aujourd'hui que je peux compter sur une grille d'analyse des opportunités qui s'est façonnée, empreinte de ce qui est important pour moi : une science qui fait sens pour moi. Dans cette notion de faire sens, j'inclus son utilité (pour la société), la possibilité d'explorations disciplinaires en articulation avec l'informatique, la dimension humaine pour la collaboration et avant tout l'enthousiasme, l'élan du cœur qu'elle provoque globalement.

4.6 Réflexions conclusives : une politique de recherche à mettre en œuvre

Les perspectives de recherche proposées précédemment amènent à se questionner sur sa mise en œuvre, et à défaut de préciser clairement le comment, il s'agit de préciser quelques points de vigilance.

Bien évidemment, il y a le volet enjeux de financement de ces recherches. En fonction des partenaires, les guichets de financement seront à définir (ANR française / Programme Samuel de Champlain avec le Canada / financements européens, internationaux). La collaboration sur base de fonds privés doit être réfléchie en fonction des sujets.

Par ailleurs, tout en travaillant de manière interdisciplinaire, il apparaît important d'avoir en tête que l'informatique est globalement mieux dotée en financement de projets par rapport aux Sciences Humaines et Sociales d'après mes observations de fonctionnement. Il m'apparaît ainsi important de pouvoir envisager que pour des projets mêlant l'informatique à d'autres disciplines, s'il est possible d'obtenir des financements par l'informatique, cela me paraît aligné avec le fait d'avoir conscience des inégalités, y compris au niveau de la science. Par exemple si l'on considère les fonds attribués à l'intelligence artificielle, il est a priori logique que soient financés sur ces fonds les aspects sociaux, éthique, et pas seulement la recherche pour d'avantage de contributions en intelligence artificielle elles-mêmes.

Il s'agit également de sujets de thèse à définir. Et pour ce faire la vigilance est de mise quant à la dimension pluri/interdisciplinaire. En effet, si cette coloration apparaît mise en valeur de manière affichée, sa valorisation peine encore parfois à s'ancrer au sein des espaces d'évaluation scientifiques parfois, dans une organisation selon des silots disciplinaires. Ainsi, il ne s'agit pas de pénaliser un·e doctorant·e avec un sujet de thèse qui serait d'emblée mal reconnu. Mais les échanges qui ont eu lieu fin mai 2023 à Troyes en amont du 5ème Symposium du GDR CNRS MaDICS ont permis de mettre en lumière cette dimension de l'évaluation autour de l'interdisciplinarité. Ainsi, il est possible de constater, d'une certaine mesure, un passage d'un affichage positif de l'interdisciplinarité (avec aussi des financements) à une réflexion et transformation sur sa valorisation au sein des communautés. Il a été discuté de la nécessité dans le démarrage parfois de passer par une étape où la dimension outil de l'informatique peut être nécessaire avant d'envisager des contributions scientifiques en informatique.

Il est important que ces réflexions aient lieu, car un cloisonnement disciplinaire ne peut favoriser des avancées scientifiques globales. D'autant qu'au delà des modalités de collaboration, on peut constater des pratiques disciplinaires différentes avec par exemple des enjeux de lieux de publications qui ne sont pas les mêmes, avec aussi des temporalités de la recherche qui peuvent être différentes. Il s'agit alors d'avoir davantage l'attention focalisée sur les objectifs qui peuvent être partagés.

Pour certaines des perspectives, arriver à avoir des personnes en thèse qui arrivent de parcours bidisciplinaires, ou ayant un goût pour les Sciences Humaines et Sociales apparaît comme quelque chose de positif, peut-être même nécessaire. Car il s'agit, de mon point de vue, de pouvoir notamment lire d'autres disciplines. Cette curiosité est riche d'enseignements. Les parcours de MIASHS ou d'Humanités Numériques sont tout à fait pertinents de ce point de vue.

Concernant le volet handicap, il s'agit d'envisager des collaborations non seulement avec le milieu académique, mais aussi des structures en dehors de l'université, entreprises, associations, etc., à la manière des recherches-actions en Sciences Sociales.

Pour conclure, j'ai utilisé dans le titre de section le terme de « politique de recherche ». Ainsi, il est temps de revenir sur la notion de manifeste que j'évoquais en introduction, en limitant la notion de politique. Mais il s'avère finalement que ce programme de perspectives de recherche revêt des enjeux éminemment politiques.

Et il est sans doute utile de faire référence ici à l'ouvrage de Winner (2020) : « La Baleine et le Réacteur : À la recherche de limites au temps de la haute technologie » dont la description indique « Initialement publié en 1986, ce célèbre examen, par Langdon Winner, des implications politiques, sociales et philosophiques de la technologie s'avère plus actuel que jamais. Il démontre que les choix technologiques, loin d'être neutres, déterminent le genre de monde dans lequel nous vivons et le type d'être humain qui y vit, qu'en adoptant une technologie, on adopte une politique – autrement dit, que les décisions techniques sont des décisions politiques, aux conséquences majeures en matière de pouvoir, de liberté et de justice. ».

C'est ainsi que les perspectives envisagées ne sont pas neutres et peuvent être qualifiées de politiques, elles s'inscrivent dans une volonté de penser l'informatique et son apport à la société. Les valeurs que l'informatique peut soutenir ou non sont à examiner, tout comme la dimension écologique en termes d'impact environnemental de celle-ci (informatique frugale [5]).

5. https://theconversation.com/informatique-frugale-a-quand-un-numerique-compatible-avec-les-limites-planetaires-204625

5 Conclusion

Sentir sous l'écorce
Captives mais invincibles
La montée des sèves
La pression des bourgeons
Semblables aux rêves tenaces
Qui fortifient nos vies

Andrée Chedid née Andrée Saab (1920-2011), poétesse française d'origine syro-libanaise, extrait du poème « Destination : arbre » dans « Tant de corps et tant d'âme » (1991)

CETTE TROISIÈME PARTIE avait pour objectif de prendre un pas de recul pour permettre de porter un regard sur la posture de chercheuse en informatique en amenant un peu de réflexivité, et d'ouvrir la réflexion sur la dimension épistémologique et ses enjeux, permettant de faire un bilan du chemin parcouru, pour permettre d'établir les grandes directions pour la suite en présentant les perspectives de recherche.

Cette citation d'Andrée Chedid retranscrit l'état d'esprit qui est le mien. Le séjour au Québec printemps-été 2023 a été l'occasion de prendre de la distance, non seulement physiquement mais intellectuellement, de mûrir cette partie et ce vers quoi je veux aller, mettre du temps et de l'énergie. La richesse culturelle m'a nourrie et les échanges intellectuels ont été déterminants.

Je mesure que je suis depuis longtemps à la croisée de différents carrefours. Mon poste à l'UFR d'Anthropologie, Sociologie et Science Politique aurait pu être lieu d'enseignement et que j'en reste là. Il s'avère que j'ai saisi l'opportunité qu'il constituait, parce que le choix de ce poste était aussi lié à cet attrait pour les Sciences Humaines et Sociales. J'ai choisi de grandir avec ce que ces disciplines pouvaient m'apporter, et avec les personnes qui y contribuent.

C'est ce qui amène une part d'originalité dans mon regard à différents égards. Et si c'est parfois déstabilisant, pouvant donner lieu à un sentiment d'isolement par le passé, c'est aussi ce qui m'a amenée ces dernières années à avoir des sollicitations qui amènent indéniablement du sens à cette posture un peu atypique, et la confirmation d'être sur le « bon chemin » pour moi.

C'est un chemin qui nécessite du temps. Mais cette démarche réflexive m'a permis de tendre vers la dimension intellectuelle [1] qu'incarnait mon grand-père paternel, professeur de littérature. Je m'autorise ici dans l'espace de cette habilitation à faire part de mes analyses et points de vue.

Par ailleurs, ce cheminement constitue pour moi l'expérience de ralentir pour mieux réfléchir, tendant aussi d'une certaine façon vers les aspirations du mouvement *slow science* pour une recherche de qualité.

Dans l'introduction, j'évoquais le terme de manifeste en ayant recours à des définitions et en en rapprochant des termes avec ce que représentait cette HDR pour moi. Je ne m'étonnerais pas que l'on puisse relever une dimension assez artificielle dans ces rapprochements initiaux, notamment sur le volet politique.

Pour autant, je pense que cette troisième partie illustrant la posture réflexive qui s'est construite et les perspectives proposées sont caractéristiques de la dimension politique finalement. Inclusion, égalité/équité, écologie, science de qualité en laquelle on peut avoir confiance, c'est finalement ce qui sous-tend les perspectives évoquées, et qui peuvent constituer un programme politique, et peut-être même une manière affichée, revendiquée de faire de la recherche. Cela rejoint finalement la notion de manifeste évoquée en introduction dans sa dimension politique.

Et si c'est le pédagogique qui m'a amenée aux études de genre et à côtoyer les Sciences Humaines et Sociales, le retour de tout ce que j'ai appris est fait pour être retranscrit dans la démarche pédagogique, au niveau des enseignements. C'est d'ores et déjà le cas avec le montage de quelques cours qui viennent à la croisée de l'informatique, du numérique et des études de genre notamment. Mais c'est un des points que je vais chercher à développer pour aller au-delà.

1. D'après Wikipedia : un intellectuel est une personne dont l'activité repose sur l'exercice de l'esprit, qui s'engage dans la sphère publique pour faire part de ses analyses, de ses points de vue sur les sujets les plus variés ou pour défendre des valeurs [...] (https://fr.wikipedia.org/wiki/Intellectuel)

Conclusion générale : quand la réflexivité s'invite jusque dans le bilan final pour pousser la réflexion

Conclusion générale

Choisissons pour nous-mêmes notre propre chemin de vie, et essayons de répandre des fleurs sur ce chemin.

Citation attribuée à Émilie du Châtelet (1706-1749), femme de lettres, mathématicienne et physicienne française du Siècle des Lumières

MERCI de m'avoir suivie dans ce périple ! Sur le chemin de mes réflexions au-delà de mes contributions !

La conclusion générale de ce mémoire aura pour objectif de boucler la boucle que constitue le cheminement de production de ce mémoire d'habilitation à diriger des recherches, cheminement qui aura nécessité quelques années entre l'émergence du projet et son aboutissement.

Boucler une boucle n'est pas une fin en soi, et constitue surtout la possibilité d'en ouvrir une nouvelle, dans un cycle qui se poursuit et qui, pour moi, se poursuivra avec une assise trouvée grâce à ce processus d'écriture qui met en lumière d'où je suis partie et où je suis aujourd'hui, sans que cela ne constitue en soi une arrivée.

Pour conclure ce manuscrit, j'apporterai tout d'abord un préambule, avant de dresser le bilan des contributions et celui des perspectives ; je reviendrai sur la dimension pédagogique dans son articulation avec la recherche, avant de dresser un bilan plus personnel.

Préambule

À ce stade de la dernière ligne droite, je ne peux m'empêcher de voir la démarche de recherche comme une manière de cheminer sur les voies de la connaissance. Plusieurs manières de le faire en fonction de qui nous sommes ou aussi du moment où nous abordons ce chemin. Parfois il s'agit d'emprunter des voies rapides, parfois nous avons besoin

d'un rythme plus lent, pour profiter du voyage, se laisser interpeller par tel ou tel détail, faire demi-tour pour revenir sur ses pas (dans une démarche réflexive). Le décloisonnement disciplinaire mis en avant dans ce travail, c'est une manière de permettre des jonctions entre les autoroutes et les routes, parfois c'est construire des ponts (et cela prend du temps, beaucoup de temps, si l'on veut qu'ils soient solides).

Il y a une co-existence de voies multiples, l'enjeu est que ce ne soit pas des routes parallèles qui le demeurent, parcourues à différentes vitesses. Mon idéal serait qu'il n'y ait pas de hiérarchisation des manières de s'engager sur la route, mais plutôt une co-existence, et ce dans le respect du chemin des autres, en acceptant en premier lieu que d'autres chemins sont possibles, et ce, sans les dénigrer.

Pour se rendre compte de ces ponts et de ces voies, rien de tel que de prendre de la hauteur, tel un aigle[2] qui s'envole haut dans le ciel, et qui a en même temps la faculté de voir les détails, avec précision, notamment pour détecter où se situent les points qui mériteraient d'être reliés.

Pour moi, ce parcours consistait à suivre le fil, celui-ci m'a emmenée en 2023 à être pendant près de trois mois au Québec, et avoir l'occasion de faire des rencontres incroyables, notament autour des sujets de l'équité, de la diversité et de l'inclusion, et des questions autochtones. Je ne doute pas que cela est venu nourrir mes réflexions.

Nous sommes les acteur/trices sur ces voies pour faire circuler les savoirs, les concepts, que ce soit au niveau des publications lues par d'autres experts, mais aussi via d'autres fomes, et notamment la médiation scientifique, avec l'enjeu pour moi d'œuvrer pour la société.

Tout le long de ce chemin d'enseignante-chercheuse, je n'ai pas eu de plan de carrière, comme si la destination n'était pas si importante. C'est le chemin qui a compté, fait de son lot de rencontres.

Le moment de l'HDR est peut-être plus que jamais le moment où on arrive à un carrefour. Après avoir marché dans le sillon des aîné·es, voici venu le moment de savoir si nous sommes prêt·es à tracer notre propre voie, en toute humilité bien évidemment.

Si mon grand-père paternel incarne une figure académique du professeur maîtrisant parfaitement les mots en tant que professeur de littérature (figure inspirante à qui je dois sans doute mon souci des mots utilisés et de l'expression des pensées), il s'agit pour moi de ne pas omettre l'autre branche généalogique avec ma grand-mère maternelle immigrée italienne de milieu ouvrier, qui apprit le français tardivement dans les livres, et qui m'a notamment transmis, parmi ses nombreuses qualités, sa simplicité, dans le meilleur sens (positif) que ce terme puisse signifier pour moi.

La rencontre de ces deux chemins en moi me permet de ne pas oublier l'importance que la Science ne doit pas être une « affaire de privilégié·es ».

Je souligne ici les points notables qui m'ont permis de suivre le chemin que j'ai choisi.

2. *Mikisi* en *anicinapek*

Je pense que le fait d'arriver dans une nouvelle discipline comme la sociologie dans une forme de curiosité intellectuelle, fut un atout indéniable, y compris par rapport au fait de ne pas avoir été formatée par des courants de pensées de ces disciplines, qui n'existent pas vraiment en informatique, en tous cas pas de manière explicite.

L'attrait plus général pour les Sciences Humaines et Sociales a permis de nourrir des réflexions importantes et ce nomadisme disciplinaire, sans remettre en cause un ancrage de départ, amène à explorer quelles sont les pratiques de ces diverses disciplines et notamment les manières d'évaluer : valeurs des espaces de publication, critères pour la qualification, contenu d'une HDR (alors même qu'un texte de loi régit cela). Ces pratiques sont ancrées dans un historique de construction et de perpétuation de la discipline.

Explorer ces pratiques invitent parfois à questionner les pratiques de sa propre discipline. Par exemple, j'ai toujours été étonnée que dans la dynamique des soutenances, la parole aux encadrant·es soit laissée après les membres du jury mais avant que le président ou la présidente prenne la parole, créant une sorte d'interruption des échanges scientifiques. J'ai pu observer dans nombre de disciplines (comme en sociologie, information et communication, littérature comparée, etc.) que les encadrant·es ont la parole après la personne qui soutient et qui a fait son propos liminaire, ce que je trouve beaucoup plus cohérent. Dernièrement, assister à des soutenances où le/la président·e de jury se donnait la parole pour venir clore la partie discussion scientifique, avant de donner la parole aux encadrant·es, m'a paru très pertinent.

J'ai donc exploré ces différents chemins qui s'offraient à moi toutes ces années, y compris dans le processus de cette HDR : revenir en arrière sur les contributions pour mieux penser l'avenir. Je reviens donc justement à présent sur une synthèse des contributions.

Bilan des contributions

Les contributions scientifiques relatées dans ce mémoire ont été organisées selon deux axes.

Le premier axe concerne les apports pour l'analyse de données de la société, comprenant, selon ces apports, des données médicales, des données sur l'habitat social, des données de médias sociaux, des données bibliographiques relatives à des articles scientifiques.

Les apports de ce premier axe s'articulent selon trois volets. Le premier volet dans le domaine de la *business intelligence* concerne la modélisation : 1) modélisation dans les entrepôts de données pour la prise en compte de connaissances utilisateur/trices pour l'analyse OLAP avec une application aux données médicales, 2) modélisation de métadonnées pour rendre possible l'analyse dans les lacs de données avec une application à l'habitat social. Le deuxième volet concerne l'apprentissage de données dans le contexte des médias sociaux, et plus particulièrement Twitter, avec la détection d'évènements, leur

diffusion et les personnes influençant cette diffusion, ainsi que la détection de rumeurs. Le troisième volet traite de l'analyse de données bibliographiques avec une approche de *Graph OLAP* basée sur l'apport de cubes pour valuer les sommets et les arêtes de graphes.

Le second axe concerne les apports qui ont permis de mettre en lumière des dynamiques sociales au sein de l'informatique, tendant ainsi à faire de la discipline un objet de recherche.

Les apports de ce second axe sont articulés en trois volets également, qui s'inscrivent dans une perspective d'enjeux sociétaux d'actualité relatifs à des aspects de politiques publiques que sont l'écologie et les enjeux d'égalité femmes-hommes notamment.

Le premier volet consiste en l'analyse d'une communauté scientifique dans une perspective écologique, allant vers une automatisation pour le calcul de l'empreinte carbone de la conférence liée à la communauté EGC, au fil du temps. Le second volet détaille une perspective de genre pour analyser la place des femmes dans cette communauté scientifique. Le troisième volet a permis d'adopter les principes d'une démarche sociologique dans le cadre d'une analyse exploratoire pour questionner les politiques de quotas en tant que politiques d'égalité dans le contexte de comité de sélection pour le recrutement d'enseignantes-chercheuses et d'enseignants-chercheurs en informatique.

Dans cette section de bilan des contributions, j'inclus également la dimension réflexive de ce travail, qui a abouti à des réflexions au niveau épistémologique par rapport à la démarche scientifique en informatique. Bien évidemment, ce travail épistémologique mériterait des lectures plus approfondies sur le sujet, sur l'histoire et la sociologie des sciences, sur les enjeux méthodologiques dans la construction des disciplines. Mais il s'agissait là pour moi de poser des bases de réflexions pour servir de socle aux perspectives de recherches, qui ont été développées et que je synthétise succinctement ci-après.

Bilan des perspectives de recherche

Les perspectives de recherche visaient à définir le plan de recherche à venir, à la fois en termes de pistes à poursuivre, mais également comment cela pouvait être envisagé.

Il est important d'avoir en tête que ce plan de route s'est construit au regard du cheminement parcouru ces dernières années, et ce, pas seulement dans un prolongement des perspectives directes des contributions déjà faites, sans doute un plan plus ambitieux finalement.

Ces perspectives se sont finalement articulées non pas de manière temporelle, mais plutôt en termes de positionnement vis-à-vis de différentes facettes identitaires : une chercheuse en informatique, une chercheuse en études sur le genre, une chercheuse nomade aimant aller vers d'autres disciplines. Ces facettes ne sont bien évidemment pas cloisonnées.

Sur le plan de la recherche en informatique, deux axes ont été développés autour d'une

informatique décisionnelle inclusive accessible et d'une science de données pour les données de la science. Ces deux axes sont motivés par l'envie de pouvoir contribuer d'un point de vue sociétal sur la capacité à se saisir de données et de les analyser d'une part, et d'autre part de contribuer à la qualité de nos modes de fonctionnement en recherche, car il en va de la confiance qui nous est attribuée par la société.

Sur le plan de la recherche en études sur le genre, les pistes sont bien évidemment reliées à l'informatique, se focalisant sur l'apprentissage de données au travers des enjeux de langue. En effet cette préoccupation autour de l'expression n'est pas récente même si jusque là, cela ne s'était pas concrétisé pleinement en termes de recherche. J'y vois là une trace de mon cheminement dans le souci des formulations exactes, les plus proches de nos intentions d'expression, qui peuvent subir un éloignement lorsque nous nous éloignons de notre langue d'origine.

Sur le plan du nomadisme disciplinaire, le projet dans lequel je souhaite m'impliquer a trait à l'éducation, qui porte à la fois sur des mises en œuvre, mais avant tout à des réflexions d'ordre éthique sur l'usage de l'intelligence artificielle dans de telles plateformes. Cela rejoint les points d'accessibilité aussi évoqués vis-à-vis de l'informatique décisionnelle. Cette volonté d'implication sur ce sujet part précisément de l'accès à la formation de manière inclusive.

Il s'agit d'un programme qui peut paraître assez large en termes de spectre, qui marque une continuité au niveau de la pluralité des thématiques par rapport à ce que j'ai réalisé ces dernières années, mais aussi poursuivant de manière plus large certains axes.

La réflexion autour de ces perspectives a nécessité de faire un pas de côté pour ne pas coller à des perspectives directes des contributions présentées, à l'image de ce qui est attendu de mon point de vue à cette étape de la carrière. C'est le fruit d'un réel processus de mâturation, qui a été largement nourri par la dimension pédagogique.

Du pédagogique à la recherche et vice-versa

Si le mémoire d'habilitation à diriger des recherches s'axe plutôt sur la dimension scientifique du travail, compte-tenu de l'impact fort qu'a eu la dimension enseignement sur mon parcours de scientifique, je ne peux pas ne pas faire une pause dans ce cheminement sans un peu de réflexion sur la dimension pédagogique.

C'est j'imagine aussi un point plus spécifique en tant qu'enseignante-chercheuse voulant assumer pleinement cette double casquette et en imaginant que ces deux volets du métier peuvent être interreliés et ne pas être vécus de manière séparée (sans mettre de côté le troisième volet souvent très chronophage des responsabilités, pédagogiques notamment).

Ma prise de poste à l'UFR d'Anthropologie, de Sociologie et de Science Politique a été

déterminante et m'a amenée vers un cheminement dont je ne pouvais imaginer l'issue, ou plutôt la destination à laquelle je suis présentement, en côtoyant quotidiennement des collègues de Sciences Humaines et Sociales. Les études de genre plus particulièrement ont contribué à l'évolution de mes objets de recherche, tout comme la manière d'envisager de faire ces recherches. Il n'est pas si simple de retranscrire en mots comment ce cheminement a façonné qui je suis aujourd'hui, en tant que chercheuse, mais aussi en tant qu'enseignante.

Mon parcours est marqué par une articulation croissante entre enseignement et recherche (elle n'est pas toujours effective selon ce que nous enseignons et les niveaux auxquels nous enseignons), avec un point saillant pour moi qui est le fait que c'est l'enseignement qui m'a amenée à transformer ma manière d'être chercheuse et à développer de nouveaux objets de recherche. Et ce qui a émergé au niveau recherche est aujourd'hui retransmis au niveau de l'enseignement (même si cela ne concerne bien évidemment pas l'ensemble de mes enseignements). Les cours donnés à l'Université du Québec A Montréal dans le cadre du partenariat avec l'Institut de Recherches et d'Etudes Féministes durant mon séjour de 2023 ont fini de me convaincre de l'intérêt de ma posture, de mes intérêts de recherche.

Aller vers l'interdisciplinarité ne peut être seulement une question de « lire sur ». Il y a un processus qui relève de l'appropriation, dans un sens positif, et pour ce faire, d'une forme d'« infusion ».

Il apparaît alors évident pour moi que cette conscientisation et cette mise en mots qu'a permis ce travail de rédaction d'habilitation à diriger des recherches impactent fortement la manière de concevoir la transmission d'une discipline qu'est l'informatique, que ce soit dans l'accompagnement à la recherche dans sa dimension scientifique, mais également dans son enseignement, qui a vocation à former des personnes qui auront ou non la vocation d'aller vers de la recherche.

Tout d'abord, la création de cours, qui permettent de présenter les résultats des travaux, notamment autour de la place des femmes dans l'informatique, a pu se concrétiser dans différents espaces de formation.

Par ailleurs, il ne s'agit pas d'oublier que dans les blocs de compétences définis au niveau national pour les mentions de master, la dimension numérique correspond au premier bloc qui apparaît sur la fiche RNCP (Répertoire national des certifications professionnelles[3]). Deux points y figurent :

1. Identifier les usages numériques et les impacts de leur évolution sur le ou les domaines concernés par la mention
2. Se servir de façon autonome des outils numériques avancés pour un ou plusieurs métiers ou secteurs de recherche du domaine

Ainsi, avoir une réflexion sur le croisement du numérique et des études genre est ce que nous tentons d'apporter aux étudiant·es de la mention études sur le genre avec dif-

3. https://www.francecompetences.fr/recherche/rncp/31494/

férents enseignements que j'ai proposés et pris en charge au cours des années. Alors, au-delà de la maîtrise d'outils comme un tableur et des statistiques descriptives dans la dimension d'analyse de données nécessaire aux questions de genre, un volet sur les questions d'usages et d'impacts en lien avec les études de genre a été développé. Je mentionne la séance de controverse autour de l'intelligence artificielle et du genre pour les étudiant·es de première année de master. Ainsi qu'une séance en deuxième année de master autour des questions de genre et numérique. Ces enjeux sont également présentés dans des enseignements transversaux à l'université Lyon 2 appelé MOTIF (MOdules Transversaux et Innovants de Formation) qui correspondent à des modules de formation pluridisciplinaires par et à la recherche destinés aux étudiant·es de Master sous forme d'options.

De plus, dans un monde où les applications numériques façonnent la société, il m'apparaît crucial de pouvoir apporter cette démarche de réflexivité de manière transversale aux étudiant·es en informatique. Il est question ici d'éthique, d'enjeux écologiques, d'enjeux sociaux, etc. Cela renvoie à la notion de RSE - Responsabilité Sociétale (ou Sociale) des Entreprises [4] évoquée dans le chapitre précédent. Ainsi, amener les étudiant·es à se questionner constitue un enjeu fort de mon point de vue pour alimenter aussi ce type de démarche présente en entreprise.

Forte des réflexions menées, il apparaît ainsi que la sensibilisation d'étudiant·es en informatique sur ces enjeux sociaux devient cruciale. Ceci rend compte de mon analyse personnelle concernant la dimension pédagogique, au regard de l'ensemble de mon cheminement. Je reviens à présent sur un bilan plus global et plus personnel dans la dernière section de ce chapitre de conclusion générale.

Bilan personnel

Compte-tenu du format relativement libre du mémoire d'habilitation à diriger des recherches, et compte-tenu de cette réflexivité développée tout au long de ces dernières années, je ne pouvais clore ce manuscrit sans un dernier retour réflexif, qui vise à un bilan que je qualifierais de plus personnel. En effet, poser des mots sur le vécu est sans doute un des outils les plus utiles pour moi afin de conscientiser le chemin parcouru. Et j'ai la conviction que cela constitue une étape importante pour pouvoir clore un cycle, avant d'en démarrer un autre, à l'image de la roue médicinale [5] dont la représentation iconographique est devenue un des symboles de certaines cultures autochtones inspirantes évoquées dans la troisième partie.

Derrière ce que peut représenter la soutenance d'une habilitation à diriger des recherches pour moi en termes de processus de légitimation d'une posture, elle offre avant tout l'occasion d'un dialogue de fond avec les pairs, qui se veut porter non seulement sur

4. RSE : https://www.economie.gouv.fr/entreprises/responsabilite-societale-entreprises-rse
5. https://fr.wikipedia.org/wiki/Roue_m%C3%A9decine

les contributions mais aussi sur les éléments présentés dans la partie 3, qui constitue une originalité en soi.

Le titre de ce mémoire « De l'analyse informatique de données de la société à l'analyse sociale de l'informatique : un cheminement guidé par les études de genre vers un décloisonnement disciplinaire et une posture réflexive. » rend justement compte de cette originalité. Le titre a eu son propre processus évolutif même s'il n'a qu'assez peu évolué finalement. Et il est vrai qu'au moment de fixer un titre initial, je ne mesurais pas à quel point le processus d'écriture constituerait en soi une partie de ce cheminement, et a fortiori que la mise en mots de ce cheminement constituerait une réelle étape. Sans doute est-ce dû au fait qu'en informatique, j'ai l'impression que les temps de rédaction sont habituellement assez courts (si je prends l'exemple de la thèse), en comparaison aux Sciences Humaines et Sociales notamment, dans lesquelles la rédaction elle-même tient un rôle particulier. Là où la rédaction correspond à une restitution des travaux menés pendant la durée de la thèse en informatique, pour d'autres disciplines, la rédaction fait partie intégrante d'un temps long sur la thèse. Ainsi, j'ai conscience que le présent mémoire présente une forme quelque peu originale par rapport à mon ancrage disciplinaire. Mais l'originalité d'une contribution n'est-elle pas un des critères importants qui comptent dans l'évaluation scientifique de notre discipline … ? Sans doute ai-je été tentée d'approcher des pratiques d'autres disciplines par ce moyen là. Je pense notamment à l'ego-histoire qui caractérise une forme d'approche historiographique et de courant d'écriture historique à travers laquelle l'historien·ne est censé·e analyser son propre parcours et ses méthodes de manière réflexive et distanciée, notamment dans le cadre d'une habilitation à diriger des recherches.

La place d'enseignante-chercheuse en informatique que j'occupe dans une composante de Sciences Humaines et Sociales (au sein de l'UFR d'Anthropologie, de Sociologie et de Science Politique), depuis 2009 en tant que maîtresse de conférences, a été pour moi un terreau très fertile de ma propre transformation et de la construction de ce que je pourrais appeler l'originalité de mon identité scientifique.

Il ne s'agit pas ici de discuter cette notion même d'identité qui pourrait faire l'objet d'échanges avec des perspectives multiples, en témoigne par exemple le travail de Baudry et Juchs, 2007. Mais, quoi qu'il en soit, ce manuscrit m'a permis de mettre en lumière différentes facettes de mon identité scientifique qui s'est construite peu à peu. Et si la question se posait de comment je me définis, je crois que je ne choisirais pas de case (catégorie) s'il n'y a pas besoin de comptage !

Je mets néanmoins ici en lumière ce que je peux qualifier comme étant le triptique que je choisirais qui fonde/guide celle-ci, et ce dans une articulation que j'explicite juste après : 1) ouverture disciplinaire, 2) réflexivité et éthique, et 3) importance du sens.

En côtoyant des collègues issu·es de disciplines diverses, et m'amenant à m'interroger, à intégrer le questionnement comme un processus à la base de la démarche scientifique (au-delà du comment : le pourquoi ?, pour qui ?, etc.), un intérêt pour l'ouverture disci-

plinaire a dépassé l'intérêt initial de mes études où j'avais pu découvrir notamment la sociologie, témoignant je pense de mon intégration réussie dans ma composante de rattachement de Sciences Humaines et Sociales.

Le cheminement vers l'interdisciplinarité est passionnant, riche, stimulant et surtout, il prend du temps. Pour aller vers une pratique de l'interdisciplinarité, il m'apparaît qu'un certain nombre de capacités sont nécessaires : la patience, la curiosité, l'écoute, la remise en question, une volonté collaborative ...

Je revendique ainsi une forme de nomadisme scientifique dans le sens qui a été détaillé précédemment dans sa dimension épistémologique. C'est une belle opportunité de tendre vers une maîtrise de connaissances multiples et de pouvoir faire des ponts entre les disciplines, dans un enrichissement mutuel.

Les riches questionnements qui émergent de ces échanges fructueux traduisent un besoin fort de donner du sens à ce qui est fait, au service de la société que l'on veut voir venir, ce qui nécessite de se questionner sans cesse.

Je reste convaincue que l'ouverture à d'autres disciplines permet une ouverture à ces questionnements qui sont aujourd'hui plus que nécessaires de mon point de vue. Il ne s'agit pas dans mon propos de demander à ce que chaque chercheuse ou chercheur se saisisse de ces questionnements de façon dogmatique, mais il est alors plus que jamais nécessaire d'accepter des profils divers relevant à la fois d'une sédentarité disciplinaire et d'un nomadisme disciplinaire comme évoqué précédemment, en essayant de déconstruire la hiérarchisation qui tend à s'opérer aujourd'hui.

Si l'informatique est une science très en lien avec la société, notamment si l'on considère le financement de celle-ci (financement de thèse par le dispositif CIFRE, financement de projets dans lesquels des entreprises sont impliquées, services de recherche et développement dans les entreprises, etc.), il est d'autant plus important que des questionnements émergent par rapport à la place de l'informatique dans la société, par exemple en lien avec les questions d'éthique. Ce n'est sans doute pas confortable individuellement, en termes de positionnement. Ce sont je crois les mêmes difficultés quand nous travaillons sur certains domaines de l'informatique en étant sensibles aux enjeux écologiques par exemple.

L'Arrêté du 26 août 2022 (modifiant l'Arrêté du 25 mai 2016) fixant le cadre national de la formation et les modalités conduisant à la délivrance du diplôme national de doctorat a fait notamment l'objet de l'ajout de l'article 19 bis pour qu'un serment soit prêté, dans lequel la réflexivité éthique est un concept central, cet article est rédigé de la façon suivante.

Article 19 bis
A l'issue de la soutenance et en cas d'admission, le docteur prête serment, individuellement en s'engageant à respecter les principes et exigences de l'intégrité scientifique dans la suite de sa carrière professionnelle, quel qu'en soit le secteur ou le domaine d'activité.
« Le serment des docteurs relatif à l'intégrité scientifique est le suivant :
En présence de mes pairs.
Parvenu(e) à l'issue de mon doctorat en [xxx], et ayant ainsi pratiqué, dans ma quête du savoir, l'exercice d'une recherche scientifique exigeante, en cultivant la rigueur intellectuelle, la **réflexivité éthique** et dans le respect des principes de l'intégrité scientifique, je m'engage, pour ce qui dépendra de moi, dans la suite de ma carrière professionnelle quel qu'en soit le secteur ou le domaine d'activité, à maintenir une conduite intègre dans mon rapport au savoir, mes méthodes et mes résultats. »

Une fiche pratique a été proposée [6].

Cette mise en œuvre, où l'accent est mis sur la rigueur intellectuelle, la réflexivité éthique et l'intégrité scientifique, a suscité diverses discussions et induit différents questionnements au-delà d'une initiative qui peut paraître positive.

Après avoir ponctué le démarrage de chaque chapitre avec des citations de femmes scientifiques et femmes de lettres qui illustraient des éléments importants quant à mes valeurs et mon cheminement, je souhaite faire la place ici à une citation d'homme bien connue qui fait le lien avec les enjeux d'éthique et de réflexivité en science.

« Science sans conscience n'est que ruine de l'âme. » est une citation de François Rabelais (Pantagruel, 1532), écrivain français humaniste de la Renaissance.

Cette citation me ramène à mes années lycées et les cours de philosophie que j'avais tant appréciés. Au-delà du souvenir de cette citation qui s'est rappelée à moi à l'écriture de ce manuscrit, elle symbolise également un rappel important pour chacune et chacun.

Rappel important au moment où il ne s'agit pas d'oublier la jeunesse de la discipline de l'informatique, qui a su défendre sa place en tant que science dans le paysage de la connaissance. Rappel important au moment où l'informatique n'a jamais été aussi présente dans la société. Rappel vital que l'informatique est articulée avec le politique, et qu'elle peut produire le meilleur, comme le pire...

Alors, il est utile de se rappeler à l'histoire des sciences qui ne saurait oublier l'histoire liée aux recherches sur le nucléaire et la bombe atomique.

Je rappelle ici les propos de Dominique Pestre, historien des sciences, déjà présentés précemment dans ce manuscrit, qui précisait dans l'ouvrage d'introduction au Science Studies en 2006 (Pestre, 2006) : « Malgré ce qu'énonce la philosophie spontanée des savants, il est difficile de laisser la science dans sa tour d'ivoire et de faire comme si c'était

6. https://www.hceres.fr/sites/default/files/media/files/fiche-serment-doctoral-integrite-scientifique-pdf1.pdf

dans l'univers clos des laboratoires que se passait l'essentiel de ce qui la concerne. [...], la question du rapport des sciences [...] à l'économique et au politique (comme aux sociabilités et à la domination masculine) ne peut être évacuée. »

Ce bilan personnel dans un mémoire à caractère académique peut paraître étonnant, en particulier dans notre discipline qu'est l'informatique, où l'on écrit finalement assez peu sur les difficultés rencontrées et comment elles sont surmontées, sur les « échecs » d'une recherche, sur la manière dont se déroule une recherche dans une perspective réflexive, ce qui, d'après moi, pourrait être source du point de vue scientifique d'une certaine richesse. J'ai sans doute la naïveté de croire que si nos pratiques changeaient sur ce point, cela permettrait d'amorcer un changement dans les critères d'excellence qui régissent notre système scientifique, des règles à un jeu que nous jouons en étant plus ou moins aveugles à la situation pour que cela reste supportable d'un point de vue personnel, des règles que nous décrions parfois dans certains espaces d'expression (en revendiquant la *Slow Science* (Noûs, 2020), en dénonçant le *publish or perish*), mais que nous continuons de jouer car nous sentant sans doute pris dans un engrenage que nous ne pouvons arrêter. J'ai bien conscience que ce « nous », qui se veut ici volontairement pluriel, exprime un positionnement personnel, pouvant tout aussi bien être partagé comme non partagé. Mais cela pose inévitablement la question de comment nous formons les jeunes chercheuses et chercheurs.

Je reviens donc ici sur la dimension de l'encadrement de la recherche. L'encadrement scientifique nécessite des compétences. C'est avant tout, d'après moi, une aventure humaine. Il m'apparaît évident après quatre co-directions de thèse qui ont pu aller jusqu'à la soutenance que ce fut en soi un apprentissage aussi pour moi.

Encadrer des recherches, cela constitue à la fois dans notre discipline une co-construction, mais aussi une transmission en étant en contact avec une génération généralement de personnes plus jeunes. Cela relève d'une responsabilité, à la fois de ce qui va être produit comme connaissances, de comment nous accompagnons ce processus, des valeurs que nous incarnons (éthique, etc.). Le temps de prêter serment à présent lors de la soutenance (mentionné plus tôt dans cette section à travers la présentation de l'arrêté de modification) est un beau rappel finalement de notre propre responsabilité, vis-à-vis de la chercheuse ou du chercheur que nous formons à la recherche. Bien évidemment, les formations suivies dans le cadre du cursus doctoral, participent et doivent participer à cela. Mais il ne s'agit pas à mon sens de se dé-responsabiliser par rapport au rôle que nous avons.

C'est l'expérience qui m'amène à formuler les choses de cette manière. Je n'ai peut-être pas toujours incarné l'encadrante idéale pour les personnes que j'ai accompagnées, bien que je l'ai toujours fait avec beaucoup d'implication et de volonté de faire au mieux. Ce rôle est définitivement un apprentissage à parfaire. Parfois, nous avons des modèles inspirants ou des contre-modèles qui permettent d'apprendre également sur des manières que nous ne voulons pas reproduire. Et si certains établissements proposent des formations à l'encadrement des doctorant·es, il s'avère que je n'ai pas eu l'occasion d'en

suivre dans ma carrière. Cela ramène en tous cas selon moi aux questionnements suivants, et de façon non exhaustive : Quelle encadrante je veux être ? Que recouvre ce rôle ? Qu'est-ce que je veux transmettre aux futures chercheuses et chercheurs et comment le transmettre ? Et en tout état de cause, il n'y a pas d'unique réponse à ces questions, car mon cheminement m'a montré qu'une des qualités que je considère première dans ma façon de concevoir ce rôle, est l'adaptabilité, car selon les personnes, le besoin d'accompagnement ne se situera pas sur les mêmes plans.

Et si les termes d'habilité et d'habileté, provenant de la même famille, ne se différencient dans l'écriture que par une lettre, leur sens est différent, et en même temps, il s'agit pour moi de développer une forme d'habileté à diriger des recherches qui va au-delà de l'habilitation retranscrivant le diplôme qui reconnaît l'habilité à le faire.

J'évoquais en introduction la notion de manifeste qui était présente dans la caractérisation du sens que je donnais à cette HDR au-delà de ce qu'elle pouvait signifier dans son sens premier politique. À l'issue de ce processus d'écriture, je suis moi-même convaincue de la pertinence du recours à ce terme, et j'espère que ce manuscrit aura également su convaincre de cette pertinence.

Je terminerai ce mémoire par quelques mots autour de la citation introductive de ce chapitre : « Choisissons pour nous-mêmes notre propre chemin de vie, et essayons de répandre des fleurs sur ce chemin. » d'Emilie du Chatelet.

Elle est inspirante à bien des égards. Tout d'abord parce qu'elle remet en perspective l'importance du choix pour nous-mêmes de notre propre chemin de vie, et ce message porté par une femme a une saveur toute particulière, notamment en tant que féministe. C'est une invitation d'une scientifique d'une autre époque qui, en amenant cette idée de répandre des fleurs, fait écho pour moi à comment cheminer dans un système de l'enseignement supérieur et de la recherche qui parfois nous amène plus que jamais à des limites difficiles à gérer. Derrière cette notion de choix, il y a des questions, beaucoup de questions, qui se posent à moi. Les principales sont peut-être : À quoi je veux contribuer ? Qu'est-ce que je veux transmettre et comment ? Quelle enseignante-chercheuse je veux être/incarner ? Et dans cette idée de répandre des fleurs, il y a pour moi cette intention de comment je veux marcher ce chemin non seulement en tant que chercheuse, mais aussi enseignante et responsable pédagogique, à partir du moment où j'ai choisi de le marcher, avec la volonté ferme de pouvoir contribuer tel un colibri qui fait sa part [7]. Ainsi, au-delà d'y répandre des fleurs, qui pour moi parle de faire en sorte que ce chemin soit le plus « beau » possible, je choisis surtout de pouvoir semer des graines. Et cette habilitation est porteuse de cette intention.

7. https://hannenorak.com/catalogue/le-vol-du-colibri/

Bibliographie

Quoi que vous cherchiez à savoir ou à ressentir, à comprendre ou à percevoir, à saisir ou à entrevoir, quelque part un livre répond à votre quête.

Christiane Taubira (née en 1952), femme politique française, « Baroque Sarabande » (2018)

Agrawal, R., Gupta, A. & Sarawagi, S. (1997). Modeling Multidimensional Databases. *Proceedings 13th International Conference on Data Engineering, ICDE*, 232-243. doi:10.1109/ICDE.1997.581777 (cité p. 36)

Amer-Yahia, S., Bonifati, A., **Favre, C.**, Fromont, É., Labroche, N., Melançon, G., Sèdes, F., Soulet, A. & Termier, A. (2022). Diversity and Inclusion Activities in EGC - A 2022 Report. *SIGKDD Explor.*, *24*(1), 52-56. doi:10.1145/3544903.3544911 (cité p. 110)

Amer-Yahia, S., Marcel, P. & Peralta, V. (2023). Data Narration for the People: Challenges and Opportunities. *Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023*, 855-858. doi:10.48786/EDBT.2023.82 (cité p. 156)

Avenier, M.-J. (2019). Les Sciences de l'artificiel : une conceptualisation révolutionnaire de sciences fondamentales à parachever. *Projectics / Proyéctica / Projectique*, *24*(3), 43-56. doi:10.3917/proj.024.0043 (cité p. 173)

Avenier, M.-J. & Gavard-Perret, M.-L. (2012). Inscrire son projet de recherche dans un cadre épistémologique. *Méthodologie de la recherche en sciences de gestion - Réussir son mémoire ou sa thèse en science de gestion, de M. L. Gavard-Perret, D. Gotteland, C. Haon and A. Jolibert* (p. 11-62). Pearson Education Universitaire. (cité p. 173).

Azri, A. (2022). *Approches multimodales d'apprentissage automatique pour la détection des rumeurs dans les microblogs* (thèse de doct.). Université Lumière Lyon 2. (cité p. 55, 62).

Azri, A., **Favre, C.**, Harbi, N. & Darmont, J. (2019a). Including images into message veracity assessment in social media. *8th International Conference on Innovation and New Trends in Information Technology (INTIS 2019), Tangier, Morocco*, 7 p (cité p. 55).

Azri, A., **Favre, C.**, Harbi, N. & Darmont, J. (2019b). Vers une analyse des rumeurs dans les réseaux sociaux basée sur la véracité des images : état de l'art. *Business Intelligence & Big Data, 15ème Edition de la conférence EDA, Montpellier, France, 3-4 octobre 2019, B-15*, 125-142 (cité p. 55).

Azri, A., **Favre, C.**, Harbi, N., Darmont, J. & Noûs, C. (2021a). Calling to CNN-LSTM for Rumor Detection: A Deep Multi-channel Model for Message Veracity Classification in Microblogs. *European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2021, Bilbao, Spain, Proceedings Part V, 12979*, 497-513. doi:10.1007/978-3-030-86517-7_31 (cité p. 55)

Azri, A., **Favre, C.**, Harbi, N., Darmont, J. & Noûs, C. (2021b). MONITOR: A Multimodal Fusion Framework to Assess Message Veracity in Social Networks. *25th European Conference on Advances in Databases and Information Systems (ADBIS 2021), Tartu, Estonia, August 24-26, 12843*, 73-87. doi:10.1007/978-3-030-82472-3_7 (cité p. 55)

Azri, A., **Favre, C.**, Harbi, N., Darmont, J. & Noûs, C. (2023). Rumor Classification through a Multimodal Fusion Framework and Ensemble Learning (Springer, Éd.). *Information Systems Frontiers, 25*(5), 1795-1810 (cité p. 55).

Bacot, P., Desmarchelier, D. & Rémi-Giraud, S. (2012). Le langage des chiffres en politique. *Mots. Les langages du politique, 100 - Numéro spécial : Chiffres et nombres dans l'argumentation politique*, 5-14 (cité p. 154, 155).

Barbot, G. & Juban, J.-Y. (2018). L'université peut-elle porter une conception renouvelée de la responsabilité sociale des entreprises (RSE)? *Management & Sciences Sociales, 2*(25), 14-27. doi:10.3917/mss.025.0014 (cité p. 105)

Baudry, R. & Juchs, J.-P. (2007). Définir l'identité. *Hypothèses, 10*(1), 155-167 (cité p. 198).

Beheshti, S.-M.-R., Benatallah, B., Motahari-Nezhad, H. R. & Allahbakhsh, M. (2012). A Framework and a Language for On-Line Analytical Processing on Graphs. *13th International Conference on Web Information Systems Engineering (WISE'12)*, 213-227 (cité p. 85).

Bellot, P. (2020). IA, recherche d'information et recommandation automatique : La diversification et la transparence des modèles comme rempart à l'uniformisation. *Implications philosophiques - Dossier « Philosophie et numérique »* (cité p. 156).

Bentayeb, F., Boussaid, O., **Favre, C.**, Ravat, F. & Teste, O. (2009). Personnalisation dans les entrepôts de données : bilan et perspectives. *Actes des 5èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne, EDA 2009, Montpellier, France, Juin 4-5, 2009*, 7-22 (cité p. 37).

Bereni, L., Chauvin, S., Jaunait, A. & Revillard, A. (2008). *Introduction aux Gender Studies. Manuel des études sur le genre*. De Boeck. (cité p. 112).

Bertin, M. & Atanassova, I. (2023). Citing Foreign Language Sources : an Analysis of the S2ORC Dataset. *13th International Workshop on Bibliometric-enhanced Information Retrieval, co-located with 45th European Conference on Information Retrieval, BIR@ECIR 2023, Dublin, Ireland* (cité p. 181).

Bertucci, M.-M. (2009). Place de la réflexivité dans les sciences humaines et sociales : quelques jalons. *Cahiers de sociolinguistique, 1*(14), 43-55 (cité p. 150).

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., Kompatsiaris, Y. et al. (2015). Verifying multimedia use at mediaeval 2015. *MediaEval, 3*(3), 7 (cité p. 64).

Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A. & Paraboschi, S. (2001). Designing Data Marts for Data Warehouses. *ACM Transactions on Software Engineering and Methodology, 10*(4), 452-483 (cité p. 42).

Borgman, C. L. (2020). *Qu'est-ce que le travail scientifique des données ? Big data, little data, no data*. OpenEdition Press. (cité p. 94).

Bruno, I., Didier, E. & Prévieux, J. (2014). *Statactivisme. Comment lutter avec des nombres*. Zones. (cité p. 177).

Cabanac, G., Hubert, G., Tran, H. D., **Favre, C.** & Labbé, C. (2016). Un regard lexico-scientométrique sur le défi EGC 2016. *16ème Journées Francophone Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, 419-424 (cité p. 94, 110, 121).

Cabanac, G., Labbé, C. & Magazinov, A. (2022). The 'Problematic Paper Screener' automatically selects suspect publications for post-publication (re)assessment. (cité p. 180).

Candau, J. (2022a). Pour un mouvement slow science. *Socio, 17*, 33-36 (cité p. 108).

Candau, J. (2022b). Slow science : l'appel de 2010 douze ans après. *Socio, 17*, 37-46 (cité p. 108).

Carnus, M.-F. & Mias, C. (2013). Dictionnaire des concepts de la professionnalisation. In A. Jorro (Éd.). De Boeck Supérieur. doi:10.3917/dbu.devel.2013.02.0269. (cité p. 150)

Chanson, A., Outa, F. E., Labroche, N., Marcel, P., Peralta, V., Verdeaux, W. & Jacquemart, L. (2022). Generating Personalized Data Narrations from EDA Notebooks. *Proceedings of the 24th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP) co-located with the 25th International Conference on Extending Database Technology and the 25th International Conference on Database Theory (EDBT/ICDT)2022, Edinburgh, UK, 2022*, *3130*, 91-95 (cité p. 156).

Chaulet, J. & Datchary, C. (2014). Moduler sa connexion : les enseignants-chercheurs aux prises avec leur courriel. *Réseaux*, *4/2014*(186), 105-140 (cité p. 118).

Chen, C., Yan, X., Zhu, F., Han, J. & Yu, P. S. (2008). Graph OLAP: Towards online analytical processing on graphs. *8th IEEE International Conference on Data Mining (ICDM'08)*, 103-112 (cité p. 84).

Chen, M., Han, J. & Yu, P. S. (1996). Data Mining: An Overview from a Database Perspective. *IEEE Trans. on Knowl. and Data Eng.*, *8*(6), 866-883 (cité p. 36).

Choudhary, M., Laclau, C. & Largeron, C. (2022). A Survey on Fairness for Machine Learning on Graphs. *ArXiv*, *abs/2205.05396*. https://api.semanticscholar.org/CorpusID:248693458 (cité p. 162)

Collet, I. (2004). La disparition des filles dans les études d'informatique : les conséquences d'un changement de représentation. *Carrefours de l'éducation*, *1*(17), 42-56 (cité p. 123).

Collet, I. (2006). *L'informatique a-t-elle un sexe ? Hackers, mythes et réalités*. Editions L'Harmattan. (cité p. 123).

Collet, I. (2011). Effet de genre : le paradoxe des études d'informatique. *TIC & Société*, *5*(1), En ligne (cité p. 123).

Collet, I. (2019). *Les oubliées du numérique. L'absence des femmes dans le monde digital n'est pas une fatalité*. Le Passeur. (cité p. 123).

Commission, E., for Research, D.-G. & Innovation. (2023). *COVID-19 impact on gender equality in research & innovation – Policy report*. Publications Office of the European Union. doi:doi/10.2777/171804. (cité p. 108)

Courau, T., Jarty, J. & Lapeyre, N. (2022). *Le genre des sciences. Approches épistémologiques et pluridisciplinaires*. Le Bord de l'eau. (cité p. 163).

Crozier, M. & Friedberg, E. (1977). *L'acteur et le système* (du Seuil, Éd.). (cité p. 167).

Darmont, J., **Favre, C.**, Loudcher, S. & Noûs, C. (2020). Data Lakes for Digital Humanities. *2nd International Digital Tools and Uses Congress (DTUC 2020), Hammamet, Tunisia*, 38-41 (cité p. 35).

de la Bellacasa, M. P. (2013). *Politiques féministes et construction des savoirs. « Penser nous devons » !* L'Harmattan. (cité p. 172).

de Beauvoir, S. (1949). *Le Deuxième Sexe, Tome II*. Gallimard. (cité p. 112).

Demarest, B., Freeman, G. & Sugimoto, C. R. (2014). The reviewer in the mirror: examining gendered and ethnicized notions of reciprocity in peer review. *Scientometrics*, *101*(1), 717-735 (cité p. 120).

Dixon, J. (2010). Pentaho, Hadoop, and Data Lakes. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. (cité p. 36)

Dubois, M. & Brault, N. (2021). Manuel d'épistémologie pour l'ingénieur.e. In É. Matériologiques (Éd.). (cité p. 173).

EcoInfo. (2019). Quelle est notre propre empreinte carbone ? https://ecoinfo.cnrs.fr/2019/07/02/quelle-est-notre-propre-empreinte-carbone/. (cité p. 99)

Eichler, R., Giebler, C., Gröger, C., Schwarz, H. & Mitschang, B. (2020). HANDLE - A Generic Metadata Model for Data Lakes. In M. Song, I. Song, G. Kotsis, A. M. Tjoa & I. Khalil (Éd.), *Big Data Analytics and Knowledge Discovery - 22nd International Conference, DaWaK 2020, Bratislava, Slovakia* (p. 73-88). Springer. doi:10.1007/978-3-030-59065-9_7. (cité p. 51)

Fausto-Sterling, A., Boterf, A. & Molinier, P. (2013). *Les cinq sexes: Pourquoi mâle et femelle ne sont pas suffisants*. Payot & Rivages. (cité p. 112).

Favre, C., Laurent, A., Pitarch, Y. & Poncelet, P. (2011). Représentation graphique des hiérarchies contextuelles : modèle avec satellites. *Actes des 7èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne, Clermont-Ferrand, France, EDA 2011, Juin 2011*, 23-37 (cité p. 34).

Favre, C. (2007). *Évolution de schémas dans les entrepôts de données : mise à jour de hiérarchies de dimension pour la personnalisation des analyses* (thèse de doct.). Université Lumière Lyon 2. (cité p. 37).

Favre, C. (2016). Femmes et recherche en informatique : d'une analyse sexuée d'une communauté scientifique aux questions de genre. *Communication au 20ème congrès international des sociologues en langue française (AISLF 2016), dans le cadre du comité de recherche « Sociologie des rapports sociaux de sexe », Montréal, Canada* (cité p. 110).

Favre, C. (2017a). Les données de la recherche vues au travers des lunettes du genre : quand l'informatique rencontre les sciences humaines et sociales pour rendre visible le non visible. *Actes du 1er atelier Valorisation et Analyse des Données de la Recherche (VADOR 2017) organisé durant la 35ème édition du congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2017), Toulouse, France*, 58-66 (cité p. 110).

Favre, C. (2017b). Les statistiques sexuées comme miroirs des inégalités de genre : apports et limites de leur production en démocratie. *Communication at the 5th International Interdisciplinary Conference of Political Research (SCOPE 2017), Bucarest, Roumanie* (cité p. 110).

Favre, C. (2017c). Quotas in higher education and research: leverage or brake for gender equality ? Study in the field of computer science in France. *Communication at the 13th Women's Worlds & 11th Fazendo Gênero, Florianopolis, Brazil* (cité p. 126).

Favre, C., Cabanac, G., Hubert, G. & Labbé, C. (2017). Du bon usage de l'interdisciplinarité pour l'analyse de données sexuées : le cas des données bibliographiques de la conférence EGC. *Communication lors de l'édition 2017 des journées Big Data Mining and Visualization de l'association EGC - Regards croisés sur les data, Lille* (cité p. 110).

Favre, C., Jakawat, W. & Loudcher, S. (2017). Graphes enrichis par des Cubes (GreC) : une approche innovante pour l'OLAP sur des réseaux d'information. *Actes du XXXVème Congrès INFORSID, Toulouse, France, May 30 - June 2, 2017*, 293-308 (cité p. 74).

Favre, C. & Tain, L. (2018a). Femme et enseignante chercheure en informatique : des contextes pluriels, une pluralité des vécus. *Communication at the 6th International Interdisciplinary Conference of Political Research (SCOPE 2018), Bucarest, Roumanie* (cité p. 126).

Favre, C. & Tain, L. (2018b). Les quotas : levier ou frein au déroulement des carrières des femmes ? Analyse suite à une enquête préliminaire dans le cas de l'enseignement supérieur et la recherche en France dans le domaine de l'informatique. *Annals of the University of Bucharest / Political science series, ANNUL XX*(2), 37-54 (cité p. 126).

Favre, C. & Valat, S. (2021). Quelles normes scientifiques ? Etude de la production d'une conférence en informatique et statistiques. *Communication au 21ème congrès international des sociologues en langue française (AISLF 2021), dans le cadre du comité de recherche « Science et innovation technologique », Tunis (en distanciel), Tunisie* (cité p. 98).

Favre, C. & Vialaret, M. (2023). A Perspective on Data Categorization with Regard to Equity, Diversity, and Inclusion in Data Science. *3rd International Workshop on Data science for equality, inclusion and well-being challenges in conjunction with IEEE International Conference on Big Data (DS4EIW 2023 @BigData2023), Sorrento, Italy* (cité p. 154).

Fondin, H. (2001). La science de l'information : posture épistémologique et spécificité disciplinaire. *Documentaliste-Sciences de l'Information, 38*(2), 112-122. doi:10.3917/docsi.382.0112 (cité p. 173)

Gaudel, M.-C. & Rozoy, B. (2017). Femmes et Informatique : Etat des lieux dans l'enseignement supérieur et la recherche. *1024 Bulletin de la société informatique de France*, (HS2), 71-82 (cité p. 120).

Ghozzi, F., Ravat, F., Teste, O. & Zurfluh, G. (2003). Constraints and Multidimensional Databases. *Vth International Conference on Enterprise Information Systems (ICEIS'03), Angers, France, 1*, 104-111 (cité p. 43).

Golfarelli, M., Maio, D. & Rizzi, S. (1998a). Conceptual Design of Data Warehouses from E/R Schemes. *XXXIst Annual Hawaii International Conference on System Sciences (HICSS'98), Big Island, Hawaii, USA, 7*, 334-343 (cité p. 42).

Golfarelli, M., Maio, D. & Rizzi, S. (1998b). The Dimensional Fact Model: A Conceptual Model for Data Warehouses. *International Journal of Cooperative Information Systems, 7*(2-3), 215-247 (cité p. 38).

Grésillon, B. (2020). Pour une hybridation entre arts et sciences sociales. In C. Éditions (Éd.). doi:10.4000/books.editionscnrs.32387. (cité p. 166)

Guille, A. (2014). *Diffusion de l'information dans les médias sociaux. Modélisation et analyse* (thèse de doct.). Université Lumière Lyon 2. (cité p. 54, 57).

Guille, A. & **Favre, C.** (2013). Analyse et fouille pour les réseaux sociaux en ligne : la plateforme SONDY. *Communication lors de l'édition 2013 des journées Big Data Mining and Visualization de l'association EGC, Paris* (cité p. 54).

Guille, A. & **Favre, C.** (2014a). Mention-anomaly-based Event Detection and tracking in Twitter. *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014*, 375-382. doi:10.1109/ASONAM.2014.6921613 (cité p. 54)

Guille, A. & **Favre, C.** (2014b). Un système de détection de thématiques populaires sur Twitter. *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014, Rennes, France, 28-32 Janvier, 2014*, 605-608 (cité p. 54).

Guille, A. & **Favre, C.** (2014c). Une méthode pour la détection de thématiques populaires sur Twitter. *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014, Rennes, France, 28-32 Janvier, 2014*, 83-88 (cité p. 54).

Guille, A. & **Favre, C.** (2015). Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach. *Social Netw. Analys. Mining, 5*(1), 18:1-18:18. doi:10.1007/s13278-015-0258-0 (cité p. 54)

Guille, A., **Favre, C.**, Hacid, H. & Zighed, D. A. (2013). SONDY: an open source platform for social dynamics mining and analysis. *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, 1005-1008. doi:10.1145/2463676.2463694 (cité p. 54)

Guille, A., **Favre, C.** & Zighed, D. A. (2013). SONDY : une plateforme open-source d'analyse et de fouille pour les rÃ©seaux sociaux en ligne. *13èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2013, Toulouse, France* (cité p. 54).

Guille, A., Hacid, H. & **Favre, C.** (2012). Une approche multidimensionnelle basée sur les comportements individuels pour la prédiction de la diffusion de l'information sur Twitter. *Extraction et gestion des connaissances (EGC'2012), Actes, janvier 31 - février 2012, Bordeaux, France*, 405-410 (cité p. 54).

Guille, A., Hacid, H., **Favre, C.** & Zighed, D. A. (2013). Information diffusion in online social networks: a survey. *SIGMOD Record, 42*(2), 17-28. doi:10.1145/2503792.2503797 (cité p. 54, 61)

Gupta, A., Lamba, H., Kumaraguru, P. & Joshi, A. (2013). Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. *Proceedings of the 22nd international conference on World Wide Web*, 729-736 (cité p. 69).

Gupta, M., Zhao, P. & Han, J. (2012). Evaluating event credibility on twitter. *Proceedings of the 2012 SIAM International Conference on Data Mining*, 153-164 (cité p. 70).

Han, J. (1997). OLAP Mining: An Integration of OLAP with Data Mining. *7th IFIP Working conference on Database Semantics*, 1-9 (cité p. 36).

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies, 14*(3), 575-599 (cité p. 161).

Harding, S. (1992). Rethinking Standpoint Epistemology: What is "Strong Objectivity?" *The Centennial Review, 36*(3), 437-470 (cité p. 162).

Hartley, J. & Cabanac, G. (2014). Do men and women differ in their use of tables and graphs in academic publications? *Scientometrics, 98*(2), 1161-1172 (cité p. 120).

Hechiche-Salah, L., Ouerdian, E. G.-B., Yahmadi, T. & Othman, S. B. (2018). Quand le stress professionnel dégénère en souffrance au travail : cas des enseignants-chercheurs tunisiens. *@GRH, 2/2018*(27), 57-82 (cité p. 118).

Heguerte, L. B., Bugeau, A. & Lannelongue, L. (2023). How to estimate carbon footprint when training deep learning models? A guide and review. *Environmental Research Communications*, *5*(11). doi:10.48550/ARXIV.2306.08323 (cité p. 71)

Heymann, S. & Grand, B. L. (2013). Visual Analysis of Complex Networks for Business Intelligence with Gephi. *2013 17th International Conference on Information Visualisation*, 307-312. doi:10.1109/IV.2013.39 (cité p. 76)

Inmon, W. H. (1996). *Building the Data Warehouse, 2nd Edition* (2e éd.). Wiley. (cité p. 36).

Jakawat, W. (2016). *Graphs enriched by Cubes (GreC) : a new approach for OLAP on information networks* (thèse de doct.). Université Lumière Lyon 2. (cité p. 74).

Jakawat, W., **Favre, C.** & Loudcher, S. (2013). OLAP on Information Networks: A New Framework for Dealing with Bibliographic Data. *New Trends in Databases and Information Systems, 1st International Workshop on Social Business Intelligence (SoBI 2013) in conjunction with the 17th East European Conference on Advances in Databases and Information Systems (ADBIS), Genoa, Italy, Proceedings II*, 361-370. doi:10.1007/978-3-319-01863-8_38 (cité p. 74)

Jakawat, W., **Favre, C.** & Loudcher, S. (2016a). Graphs enriched by cubes for OLAP on bibliographic networks. *International Journal of Business Intelligence and Data Mining (IJBIDM)*, *11*(1), 85-107. doi:10.1504/IJBIDM.2016.076435 (cité p. 74)

Jakawat, W., **Favre, C.** & Loudcher, S. (2016b). OLAP Cube-based Graph Approach for Bibliographic Data. *Proceedings of Student Research Forum Papers and Posters at SOFSEM 2016 co-located with 42nd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2016), Harrachov, Czech Republic, January 23-28, 2016.*, 87-99 (cité p. 74).

Jin, X., Han, J., Cao, L., Luo, J., Ding, B. & Lin, C. X. (2010). Visual Cube and On-Line Analytical Processing of Images. *19th ACM International Conference on Information and Knowledge Management (CIKM'10)* (cité p. 84).

Kergosien, E., Bessagnet, M.-N., Sallaberry, C., Le Parc-Lacayrelle, A. & Royer, A. (2016). Analyse géographique de séries de publications : application aux conférences EGC. *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, 371-382 (cité p. 100).

Kimball, R. (1996). *The Data Warehouse Toolkit*. John Wiley & Sons. (cité p. 44).

Kleinpeter, É. (2013). Taxinomie critique de l'interdisciplinarité (Hermès, Éd.). *La Revue*, *67*(3), 123-129 (cité p. 163).

Kosinski, M. & Wang, Y. (2018). Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images. *Journal of Personality and Social Psychology*, *114*(2), 246-257 (cité p. 170).

Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*, 211-213. doi:https://doi.org/10.1038/504211a (cité p. 120)

Laufer, J. & Paoletti, M. (2015). Quotas en tout genre ? *Travail, genre et sociétés*, *34*(2), 151-155 (cité p. 137).

Lemercier, É. (2015). À l'université : les dessous d'un consensus apparent. *Travail, genre et sociétés, 34(2): 175-180*, *34*(2), 175-180 (cité p. 138).

Lépinard, É. (2007). *L'égalité introuvable. La parité, les féministes et la République* (P. de Sciences Po, Éd.). Académique. (cité p. 137).

Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. *SPIRE'02 : Proceedings of the 9th international conference on String Processing and Information Retrieval*, *2476*, 1-10. doi:10.1007/3-540-45735-6_1 (cité p. 113)

Leydesdorff, L. & Milojević, S. (2015). Scientometrics. *International Encyclopedia of the Social & Behavioral Sciences, 2nd Edition*, 322-327 (cité p. 111).

Lin, Z., Yin, Y., Liu, L. & Wang, D. (2023). SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data* (cité p. 180).

Lizotte, M. (2021). If you do not deign to quantify, someone else will do it for you: In support of a balanced approach to the evaluation of science. *Social Science Information*, *60*(3), 363-371 (cité p. 154).

Loudcher, S., **Favre, C.** & Jakawat, W. (2013). *Que peut apporter l'OLAP à l'analyse de réseaux d'informations bibliographiques ?* (cité p. 74).

Loudcher, S., Jakawat, W., Morales, E. P. S. & **Favre, C.** (2015). Combining OLAP and information networks for bibliographic data analysis: a survey. *Scientometrics*, *103*(2), 471-487 (cité p. 85).

Loudcher, S., Jakawat, W., Soriano-Morales, E.-P. & **Favre, C.** (2015). Combining OLAP and information networks for bibliographic data analysis: a survey. *Scientometrics*, *103*(2), 471-487. doi:10.1007/s11192-015-1539-0 (cité p. 74)

Madera, C. & Laurent, A. (2016). The next information architecture evolution: the data lake wave. In R. Chbeir, R. Agrawal & I. Biskri (Éd.), *Proceedings of the 8th International Conference on Management of Digital EcoSystems, MEDES 2016, Biarritz, France* (p. 174-180). ACM. (cité p. 45).

Maingain, A., Dufour, B. & (direction), G. F. (2002). *Approches didactiques de l'interdisciplinarité* (D. B. Université, Éd.). (cité p. 164).

Maisonobe, M., Jégou, L. & Eckert, D. (2018). Delineating urban agglomerations across the world: A dataset for studying the spatial distribution of academic research at city level. *Cybergeo: European Journal of Geography*. doi:10.4000\/cybergeo.29637 (cité p. 101)

Maisonobe, M., Jégou, L., Yakimovich, N. & Cabanac, G. (2019). NETSCITY: A geospatial application to analyse and map world scale production and collaboration data between cities. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed & G. Ruocco (Éd.), *ISSI'19: 17th International Conference on Scientometrics and Informetrics* (p. 631-642). Edizioni Efesto. (cité p. 101).

Malik, K. (2011). Virtual Research Conferences: A Case Based Analysis. *International Journal of Virtual Communities and Social Networking*, *3*(4), 32-45. doi:10.4018/jvcsn.2011100103 (cité p. 106)

Malinowski, E. & Zimányi, E. (2004). OLAP Hierarchies: A Conceptual Perspective. *XVIth International Conference on Advanced Information Systems Engineering (CAiSE'04), Riga, Latvia*, *3084*, 477-491 (cité p. 43).

Mallach, E. G. (2000). *Decision Support and Data Warehouse Systems*. McGraw-Hill Higher Education. (cité p. 36).

Marques-Pereira, B. (2011). Éléonore Lépinard : L'égalité introuvable. La parité, les féministes et la République. - Résumé. *Nouvelles Questions Féministes*, *30*(1), 115-117 (cité p. 137).

Mittal, A., Moorthy, A. K. & Bovik, A. C. (2011). Blind/referenceless image spatial quality evaluator. *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, 723-727 (cité p. 66).

Mittal, A., Soundararajan, R. & Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, *20*(3), 209-212 (cité p. 66).

Moody, D. L. & Shanks, G. G. (1994). What Makes a Good Data Model? Evaluating the Quality of Entity Relationship Models. *13th International Conference on the Entity-Relationship Approach (ER'94), Manchester, U.K.*, *881*, 94-111 (cité p. 41, 44).

Morfonios, K. & Koutrika, G. (2008). OLAP Cubes for Social Searches: Standing on the Shoulders of Giants? *International Workshop on the Web and Databases (WebDB)* (cité p. 85).

Morley, C. & Collet, I. (2017). Femmes et métiers de l'informatique : un monde pour elles aussi. *Cahiers du Genre*, *62*(1), 183-202. doi:10.3917/cdge.062.0183 (cité p. 123)

Morris, M. R., Counts, S., Roseway, A., Hoff, A. & Schwarz, J. (2012). Tweeting is believing? Understanding microblog credibility perceptions. *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 441-450 (cité p. 70).

Moureau, M. & Brace, G. (2000). *Dictionnaire des sciences de la terre (anglais-français, français-anglais)*. Editions Technip. (cité p. 102).

Myers, K. R., Tham, W. Y., Yin, Y., Cohodes, N., Thursby, J. G., Thursby, M. C., Schiffer, P., Walsh, J. T., … Wang, D. (2020). Unequal effects of the COVID-19 pandemic on scientists. *Nature Human Behaviour*, *4* (cité p. 108).

Ngono, F. A. & Şükran, T. (2023). *La pertinence des épistémologies autochtones face à la crise climatique actuelle : Enjeux de protection et de préservation du territoire*. Cahiers du CIÉRA. (cité p. 163).

Noûs, C. (2020). Slow Science – la désexcellence. *Genèses*, *119*(2), 199-208 (cité p. 201).

Paradeise, C., Noël, M. & Goastellec, G. (2015). Disciplines académiques en transformation : entre innovation et résistances. Edition des Archives. (cité p. 164).

Paul-Hus, A., Sugimoto, C. R., Haustein, S. & Larivière, V. (2015). Is There a Gender Gap in Social Media Metrics? *15th International Society of Scientometrics and Informetrics Conference (ISSI), Istanbul, Turkey* (cité p. 120).

PearAnalytics. (2009). *Twitter study* (rapp. tech.). http://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf. (cité p. 57).

Peralta, V., Illarze, A. & Ruggia, R. (2003). On the Applicability of Rules to Automate Data Warehouse Logical Design. *Ist International Workshop on Decision Systems Engineering (DSE'03), in conjunction with the XVth International Conference on Advanced Information Systems Engineering (CAiSE'03), Klagenfurt/Velden, Austria*, *75* (cité p. 42).

Pestre, D. (2006). *Introduction aux Science Studies*. La Découverte, coll. « Repères ». doi:10.3917/dec.pestr.2006.01. (cité p. 169, 200)

Phipps, C. & Davis, K. C. (2002). Automating Data Warehouse Conceptual Schema Design and Evaluation. *IVth International Workshop on Design and Management of Data Warehouses (DMDW'02), Toronto, Canada*, *58*, 23-32 (cité p. 42).

Pitarch, Y., **Favre, C.**, Laurent, A. & Poncelet, P. (2012a). Vers davantage de flexibilité et d'expressivité dans les hiérarchies contextuelles des entrepôts de données. *9ème atelier Fouille de Données Complexes : complexité liée aux données multiples et massives (FDC 12), en conjonction avec la 12ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 12), Bordeaux*, 55-66 (cité p. 34).

Pitarch, Y., **Favre, C.**, Laurent, A. & Poncelet, P. (2010a). Analyse flexible dans les entrepôts de données : quand les contextes s'en mêlent. *Actes des 6èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne, EDA 2010, Djerba, Tunisie, Juin 2010*, 191-205 (cité p. 34).

Pitarch, Y., **Favre, C.**, Laurent, A. & Poncelet, P. (2010b). Context-aware generalization for cube measures. *ACM 13th International Workshop on Data Warehousing and OLAP (DOLAP 2010), Toronto, Ontario, Canada*, 99-104. doi:10.1145/1871940.1871961 (cité p. 34, 41)

Pitarch, Y., **Favre, C.**, Laurent, A. & Poncelet, P. (2011). Généralisation contextuelle de mesures dans les entrepôts de données. Application aux entrepôts de données médicales. *Ingénierie des Systèmes d'Information*, *16*(6), 67-90. doi:10.3166/isi.16.6.67-90 (cité p. 34)

Pitarch, Y., **Favre, C.**, Laurent, A. & Poncelet, P. (2012b). Enhancing flexibility and expressivity of contextual hierarchies. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2012), Brisbane, Australia*, 1-8. doi:10.1109/FUZZ-IEEE.2012.6251176 (cité p. 34)

Pitarch, Y. (2011). *Résumé de Flots de Données : Motifs, Cubes et Hiérarchies* (thèse de doct.). Université Montpellier 2. (cité p. 37).

Prat, N., Akoka, J. & Comyn-Wattiau, I. (2006). A UML-based data warehouse design method. *Decision Support System*, *42*, 1449-1473 (cité p. 42).

Prud'homme, J. & Gingras, Y. (2015). Les collaborations interdisciplinaires : raisons et obstacles. *Actes de la recherche en sciences sociales*, *210*(5), 40-49 (cité p. 164).

Qu, Q., Zhu, F., Yan, X., Han, J., Yu, P. & Li, H. (2011). Efficient Topological OLAP on Information Networks. *Proceedings of the 16th International Conference on DAtabase Systems For Advanced Applications (DASFAA'11)*, *1*, 389-403 (cité p. 84).

Quinton, J. N. (2020). Cutting the carbon cost of academic travel [News & Views]. *Nature Reviews Earth & Environment*, *1*(1), 13. doi:10.1038/s43017-019-0008-3 (cité p. 106)

Ravat, F., Teste, O., Tournier, R. & Zurfluh, G. (2007a). A Conceptual Model for Multidimensional Analysis of Documents. *International Conference on Conceptual Modeling (ER), Auckland, New Zealand*, (4801), 550-565 (cité p. 40).

Ravat, F., Teste, O., Tournier, R. & Zurfluh, G. (2007b). Graphical Querying of Multidimensional Databases. *11th East European Conference on Advances in Databases and Information Systems (ADBIS'07), Varna, Bulgaria*, *4690*, 298-313 (cité p. 38).

Ravat, F. & Zhao, Y. (2019). Metadata Management for Data Lakes. In T. Welzer, J. Eder, V. Podgorelec, R. Wrembel, M. Ivanovic, J. Gamper, M. Morzy, T. Tzouramanis, … A. K. Latific (Éd.), *New Trends in Databases and Information Systems, ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8-11, 2019, Proceedings* (p. 37-44). Springer. doi:10.1007/978-3-030-30278-8_5. (cité p. 51)

Renisio, Y. (2015). L'origine sociale des disciplines. *Actes de la recherche en sciences sociales*, *5*(210), 10-27. doi:10.3917/arss.210.0010 (cité p. 165)

Romero, O. & Abelló, A. (2009). A Survey of Multidimensional Modeling Methodologies. *International Journal of Data Warehousing and Mining: IJDWM*, *5*(2), 1-23 (cité p. 42).

Romero, O. & Abelló, A. (2010). Automatic validation of requirements to support multidimensional design. *Data Knowledge Engineering*, *69*, 917-942 (cité p. 42).

Sá, M. J., Ferreira, C. M. & Serpa, S. (2019). Virtual and Face-To-Face Academic Conferences: Comparison and Potentials. *Journal of Educational and Social Research*, *9*(2), 35-47. doi:10.2478/jesr-2019-0011 (cité p. 106)

Sawadogo, P. N., Scholly, É., **Favre, C.**, Ferey, É., Loudcher, S. & Darmont, J. (2019). Metadata Systems for Data Lakes: Models and Features. *1st International Workshop on BI and Big Data Applications in conjunction with the 23rd European Conference on Advances in Databases and Information Systems (BBIGAP@ADBIS2019), Bled, Slovenia, September 8-11, 2019*, *1064*, 440-451. doi:10.1007/978-3-030-30278-8_43 (cité p. 35)

Schmitt, E. (2018). *Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des big data.* (Histoire, Philosophie et Sociologie des sciences). Université de technologie de Compiègne. (cité p. 172).

Scholly, É. (2022). *De la modélisation des métadonnées à la conception d'un lac de données : Application à l'habitat social.* (thèse de doct.). Université Lumière Lyon 2, France. (cité p. 35, 44).

Scholly, É., **Favre, C.**, Ferey, É. & Loudcher, S. (2021). HOUDAL: A Data Lake Implemented for Public Housing. *23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*, 39-50. doi:10.5220/0010418200390050 (cité p. 35)

Scholly, É., Sawadogo, P. N., **Favre, C.**, Ferey, É., Loudcher, S. & Darmont, J. (2019). Systèmes de métadonnées dans les lacs de données : modélisation et fonctionnalités. *Business Intelligence & Big Data, 15ème Edition de la conférence EDA, Montpellier, France, 3-4 octobre 2019*, *B-15*, 77-92 (cité p. 35).

Scholly, É., Sawadogo, P. N., Liu, P., Espinosa-Oviedo, J. A., **Favre, C.**, Loudcher, S., Darmont, J. & Noûs, C. (2021a). goldMEDAL : une nouvelle contribution à la modélisation générique des métadonnées des lacs de données (résumé). *Business Intelligence & Big Data, Actes de la conférence EDA 2021, en distanciel, 01-02 juillet 2021*, *B-17*, 55-58 (cité p. 35).

Scholly, É., Sawadogo, P. N., Liu, P., Espinosa-Oviedo, J. A., **Favre, C.**, Loudcher, S., Darmont, J. & Noûs, C. (2021b). goldMEDAL: A Data Lake Generic Metadata Model (résumé). *37ème Conférence sur la Gestion de Données – Principles, Technologies et Applications (BDA 2021), Paris (en distanciel), Octobre 2021,* 19-20 (cité p. 35).

Scholly, É., Sawadogo, P. N., Liu, P., Espinosa-Oviedo, J., **Favre, C.**, Loudcher, S., Darmont, J. & Noûs, C. (2021). Coining goldMEDAL: A New Contribution to Data Lake Generic Metadata Modeling. *23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2021), Nicosia, Cyprus*, *2840*, 31-40 (cité p. 35).

Schroeder, T. (2020). Offer a User-Friendly Way To Launch New Chapters. *The Membership Management Report*, *16*(7), 4. doi:10.1002/mmr.31506 (cité p. 107)

Serres, M. (1992). *Eclaircissements - entretiens avec Bruno Latour* (Julliard, Éd.). (cité p. 174).

Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, *8*(3), 171-188 (cité p. 64).

Simonyan, K. & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Y. Bengio & Y. LeCun (Éd.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9*. (cité p. 68).

Sousa, A. P. D. (2015). Productivisme et souffrance chez les enseignants-chercheurs au Brésil. *Pensée plurielle*, *1/2015*(38), 45-66 (cité p. 118).

Soussi, A., Feki, J. & Gargouri, F. (2005). Approche semi-automatisée de conception de schémas multidimensionnels valides. *Ière journée sur les Entrepôts de Données et l'Analyse en ligne (EDA'05), Lyon, B-1*, 71-90 (cité p. 42).

Tian, Y., Hankins, R. & Patel, L. (2008). Efficient Aggregation for Graph Summarization. *ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*, 567-580 (cité p. 85).

Traon, C., Robledo, C. & Guérin, F. (2015). La RSE est-elle applicable dans les établissements publics comme dans n'importe quelle autre organisation ? Un exemple : les Universités. *11ème Congrès International Pluridisciplinaire en Qualité, Sûreté de Fonctionnement et Développement Durable (QUALITA'2015)* (cité p. 105).

Valat, S. & **Favre, C.** (2020). Regards d'actualité au prisme des enjeux sociétaux sur les données historisées d'EGC. *20ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2020, 27-31 Janvier 2020, Bruxelles, Belgique*, 205-216 (cité p. 94, 98, 110).

Vardi, M. Y. (2020). Publish *and* perish. *Communications of the ACM*, *63*(1), 7. doi:10.1145/3373386 (cité p. 107)

Vargas-Solar, G., Cerquitelli, T., Montorsi, A., Salvai, S., Sangineti, M. T., Darmont, J. & **Favre, C.** (2022). Promoting equity, diversity and inclusion: policies, strategies and future directions in higher education, research communities and business. *2nd International Workshop on Data science for equality, inclusion and well-being challenges in conjunction with IEEE International Conference on Big Data (DS4EIW 2022 @BigData2022), Osaka, Japan*, 4710-4718. doi:10.1109/BigData55660.2022.10020621 (cité p. 110)

Venkatanath, N., Praneeth, D., Bh, M. C., Channappayya, S. S. & Medasani, S. S. (2015). Blind image quality evaluation using perception based features. *2015 Twenty First National Conference on Communications (NCC)*, 1-6 (cité p. 67).

Viennot, É., Candea, M., Chevalier, Y., Duverger, S. & Houdebine, A.-M. (2016). *L'Académie contre la langue française. Le dossier féminisation*. iXe. (cité p. 4).

Wasserman, A. I., Vardi, M. Y., Chien, A. A., Kartoun, U., Ayres, M. & Meyer, B. (2020). Conferences and carbon impact [Letters to the editor]. *Communications of the ACM*, *63*(3), 6-7. doi:10.1145/3380448 (cité p. 107)

Winner, L. (2020). *La Baleine et le Réacteur : À la recherche de limites au temps de la haute technologie* (É. Libre, Éd. ; 2ème). (cité p. 186).

Wu, K., Yang, S. & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. *2015 IEEE 31st international conference on data engineering*, 651-662 (cité p. 70).

Yang, J. & Leskovec, J. (2011). Patterns of temporal variation in online media. *WSDM*, 177-186 (cité p. 57).

Yin, M., Wu, B. & Zeng, Z. (2012). HMGraph OLAP: a Novel Framework for Multi-dimensional Heterogeneous Network Analysis. *15th International Workshop on Data warehousing and OLAP (DOLAP'12)*, 137-144 (cité p. 85).

Zenetti, M.-J. (2023). *Politiques des savoirs et des représentations : littératures documentaires (XXe-XXIe)* [Habilitation à Diriger des Recherches. École des Hautes Etudes en Sciences Sociales]. (cité p. 172).

Zhao, P., Li, X., Xin, D. & Han, J. (2011). Graph Cube: On Warehousing and OLAP Multidimensional Networks. *ACM SIGMOD International Conference on Management of Data (SIGMOD'11)*, 853-864 (cité p. 84).

Liste des figures

Liste des tableaux