Département d'évaluation de la recherche

DOCUMENT D'AUTOÉVALUATION DES UNITÉS DE RECHERCHE

CAMPAGNE D'ÉVALUATION 2025-2026
VAGUE A

Septembre 2024

1. INFORMATIONS GÉNÉRALES POUR LE CONTRAT EN COURS

1-1 Identification de l'unité

Nom de l'unité : ERIC

Acronyme : ERIC

Label et numéro :

Domaine scientifique principal :
ST : Sciences et Technologies

Panels scientifiques par ordre décroissant de pertinence :

Panel 1
ST6 : Sciences et technologies de l'information et de la communication - STIC

Panel 2
ST1 : Mathématiques

Équipe de direction :

- Julien Jacques, directeur
- Nadia Kabachi, directrice adjointe
- Sabine Loudcher, responsable équipe SID
- Guillaume Metzler, responsable équipe DMD

Liste des tutelles de l'unité de recherche : Université Lumière Lyon 2 et Université Claude Bernard Lyon 1

École(s) doctorale(s) de rattachement : ED 512 InfoMaths

1-2 Présentation de l'unité

Historique, localisation de l'unité

Fondée en 1995 à l'Université Lumière Lyon 2 comme jeune équipe, puis labélisée équipe d'accueil en 1999, ERIC a été rejointe par plusieurs collègues de l'Université Claude Bernard Lyon 1 en 2010. Ses tutelles sont les universités Lyon 2 et Lyon 1. Désormais unité de recherche, ERIC est localisée sur le campus Porte des Alpes de l'Université Lyon 2 (Bron) et dispose d'un demi-bureau sur le campus LyonTech-La Doua de l'Université Lyon 1 (Villeurbanne).

Organisation de l'unité

Le laboratoire ERIC est structuré en deux équipes de recherche rassemblant chacune des enseignants-chercheurs de Lyon 1 et Lyon 2 :
• Data Mining et Décision (DMD),
• Systèmes d'Information Décisionnels (SID).

Le laboratoire dispose d'un conseil de direction, composé du directeur, de son adjointe, de la responsable administrative ainsi que des deux responsables d'équipe.

Chaque équipe dispose de son propre budget pour le financement de ses activités de recherche (missions, recrutement de stagiaires…). L'équipement des personnels, les infrastructures de stockage et de calculs, les missions liées aux soutenances de thèse et d'HDR sont financés en central par le laboratoire.

Le premier lundi matin de chaque mois se réunit le conseil de laboratoire qui est constitué de l'ensemble des membres permanents du laboratoire ainsi que d'un représentant des doctorants et d'un représentant des membres associés. A l'issue de ce conseil, les équipes se retrouvent à leur réunion mensuelle, ouverte cette fois-ci à l'ensemble des personnels.

D'un point de vue scientifique, les équipes se retrouvent autour d'un séminaire commun qui a lieu approximativement une fois par mois, même s'il est compliqué de tenir un calendrier régulier.

Un certain nombre de moments de convivialité sont également organisés de sorte à consolider la cohésion entre les membres du laboratoire :

- un petit déjeuner ouvert à tous a lieu avant chaque conseil de laboratoire,
- un repas de fin d'année pris au sein du laboratoire juste avant les fêtes de fin d'année,
- une demi-journée fin juin destinée aux présentations de l'ensemble des travaux des stagiaires de l'année, suivi d'un pique-nique dans le parc de Parilly,
- une journée hors les murs, début juillet, mêlant présentation scientifique et activité récréative.

Nous pouvons également préciser que l'ensemble des activités en présentiel de l'unité se déroule au sein de ses locaux dans l'Université Lyon 2. Une visio-conférence est généralement mise en place pour les moments scientifiques, afin de faciliter la participation des membres Lyon 1 de l'unité.

Équipes, plateformes, services communs

Outre la répartition en deux équipes, le laboratoire s'est doté d'une cellule informatique, composée d'enseignants-chercheurs *bénévoles* du laboratoire, qui s'occupent de : la gestion du parc informatique des membres non permanents (stagiaires, doctorants…), la maintenance de notre site internet, la gestion et maintenance de nos serveurs de calculs.

Effectif de l'unité et de ses éventuelles équipes au 31/12/2024

Le laboratoire ERIC est composé de 24 membres permanents : 23 enseignants-chercheurs, dont 17 de Lyon 2 et 6 de Lyon 1, et une personnelle BIATSS (Lyon 2). Le laboratoire compte également un professeur émérite (Lyon 2), 2 ATER (Lyon 2) et 22 doctorants.
La répartition suivant les équipes est la suivante :
• DMD : 13 membres permanents (3 PR, 10 MCF dont 4 HDR), 1 professeur émérite, 1 ATER, 14 doctorants ;
• SID : 10 membres permanents (5 PR, 5 MCF), 1 ATER, 8 doctorants ;

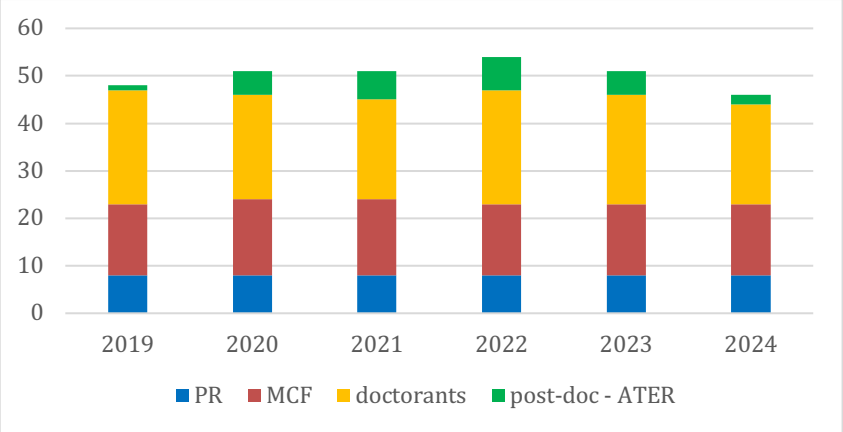

Figure 1 : Évolution des effectifs du laboratoire ERIC.

Les effectifs sont globalement stables sur la période 2019-2024, même si le décompte des membres permanents au 31/12/2024 masque en réalité une augmentation des effectifs (3 recrutements sont prévus en 2025, 1 PR et 2 MCF, sans départ prévu). En effet, si on prend en compte ces recrutements 2025, tous les départs auront été remplacés, et deux créations sont venues renforcer les effectifs (un PR26 en 2019 et un MCF 27 en 2023).
Concernant le renouvellement des membres du laboratoire, nous avons eu sur la période 4 départs, 5 recrutements, et 2 promotions de MCF à PR. Soit un renouvellement d'un sixième des membres du laboratoire. Précisons néanmoins, si l'on compte les 3 recrutements qui ont eu lieu en 2024 et les 3 qui auront lieu en 2025, que nous sommes actuellement dans une période charnière de l'évolution du laboratoire avec le renouvellement de près d'un quart des membres permanents sur 2024 et 2025. La figure 2 présente l'histogramme des âges des membres de l'unité au 31/12/2024, qui démontre une répartition assez équilibrée sur l'ensemble des classes d'âge.

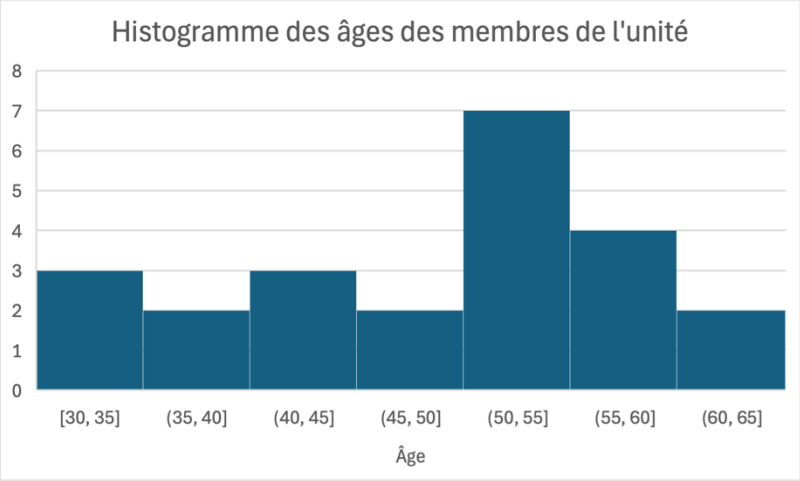

Figure 2 : Histogramme de l'âge des membres de l'unité.

Enfin, la figure 3 représente les connexions entre les membres permanents de l'unité, sur la base des co-publications sur la période évaluée. On y retrouve clairement les deux équipes (en haut à gauche l'équipe SID, en bas à droite l'équipe DMD), mais également une connexion importante entre les deux équipes, portée notamment par deux membres de l'unité.

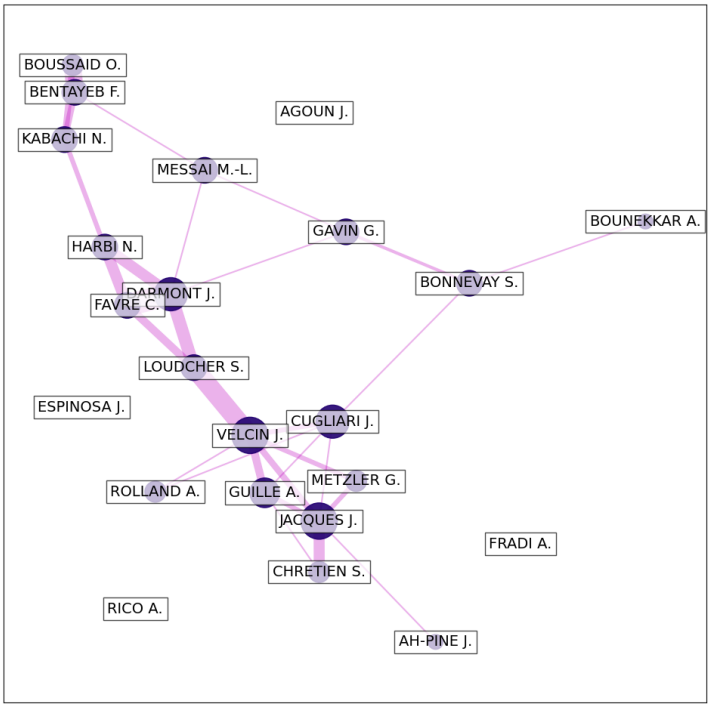

Figure 3 : Graphe de connexions entre les membres permanents du laboratoire. L'intensité du trait est proportionnelle au nombre de co-publications sur la période 2019-2024.

Thématiques scientifiques (par équipe le cas échéant)

Le laboratoire ERIC mène des recherches théoriques et appliquées dans les domaines de la science des données et de l'informatique décisionnelle. Elles visent à valoriser les grandes bases de données complexes, notamment dans les domaines des arts, lettres, langues, sciences humaines et sociales (ALLSHS) et se situent dans les domaines suivants :

- **Machine learning et décision** (équipe DMD) **:** développement de modèles et d'algorithmes d'apprentissage automatique (machine learning) pour les données complexes (de nature variée, notamment textuelles, séquentielles, catégorielles…) ; mise au point de techniques de prévision et d'agrégation multicritère pour l'aide à la décision.
- **Gestion et analyse de données massives** (équipe SID) **:** développement de modèles de gestion et d'analyse des données complexes, protection des données**.**
- **Humanités numériques** (équipes DMD et SID) **:** transdisciplinarité/interdisciplinarité avec les ALLSHS (méthodologies hybrides), la scientométrie, le genre, l'informatique verte...

La figure 4 représente de façon schématique nos thématiques de recherche.

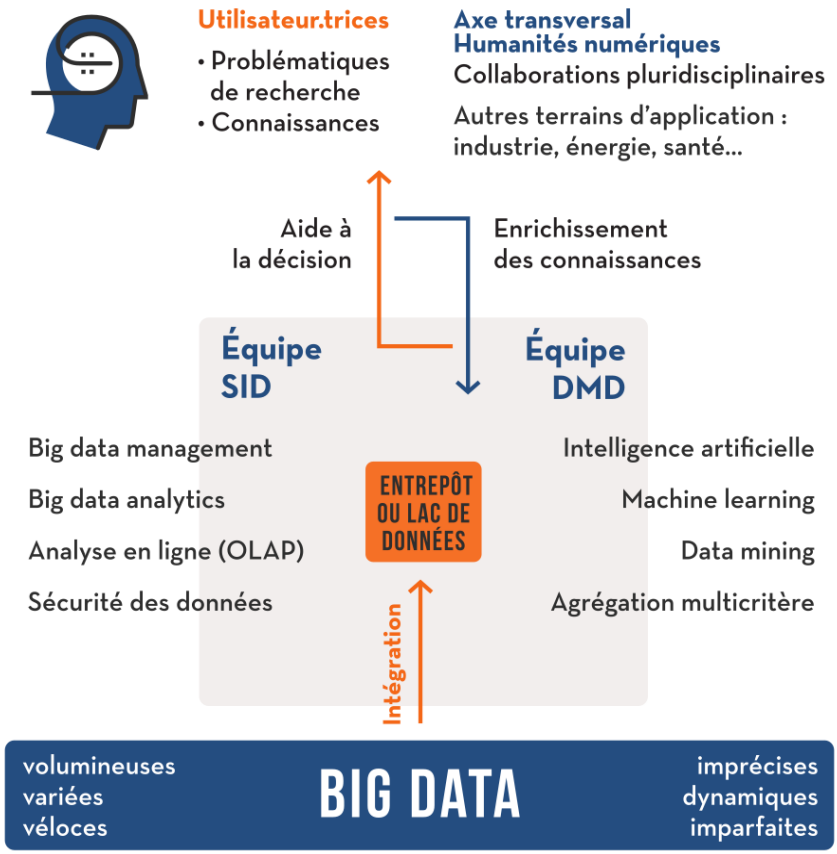

Figure 4 : Thématiques de recherche ERIC

1-3 Environnement de recherche

ERIC, un laboratoire de recherche en informatique et mathématiques appliquées

Le laboratoire ERIC a intégré la Fédération Informatique de Lyon (FIL) en 2021, même si l'officialisation de cette intégration ne sera réalisée qu'au moment de l'évaluation HCERES de la fédération, qui sera la bonne occasion pour déclarer son nouveau périmètre. En attendant, dès 2021, la direction du laboratoire a intégré le comité de direction de la FIL et des membres de l'unité ont également rejoint le conseil de la fédération. Le laboratoire ERIC rejoint ainsi 5 autres laboratoires d'informatique du site Lyon Saint-Etienne (CITI : télécom, CREATIS : santé, LabHC : physique et informatique, LIP : parallélisme, LIRIS : image, données, connaissances, services). Dans ce riche écosystème, le laboratoire ERIC se distingue à plusieurs niveaux : tout d'abord en se positionnant sur la thématique de l'informatique décisionnelle, même si des proximités thématiques existent avec des équipes des laboratoires LIRIS, DISP, LabHC, se traduisant par des collaborations régulières ; mais également en abordant la thématique de l'apprentissage automatique par une double vision informatique et mathématiques appliquées. En effet, si cette thématique de recherche est déjà présente au sein des laboratoires LabHC et LIRIS notamment, mais également au sein des laboratoires de mathématiques comme ICJ et UMPA qui ont récemment recruté dans ce domaine, la particularité d'ERIC est de regrouper dans une même équipe des chercheurs des deux disciplines.

ERIC fait également partie de l'initiative SciDoLySE, qui a vu le jour en 2019 dans le but de fédérer les chercheurs des laboratoires d'informatique et de mathématiques appliquées du site autour de la thématique du machine learning. Cette initiative a notamment débouché sur le dépôt du projet AILyS en réponse à l'appel à manifestation d'intérêt IA Cluster, dont deux membres du laboratoire ont fait partie du comité exécutif. Même si le projet AILyS n'a finalement pas été retenu, il n'est pas enterré et nous travaillons actuellement à son développement sous une autre forme.

Enfin, ERIC est également membre du groupe de recherche CNRS Masse de Données, Informations et Connaissances en Sciences (MaDICS, dont une action, ADOC, a été co-portée par un membre du laboratoire), partenaire de l'Institut du Genre (groupement d'intérêt scientifique) ainsi que co-fondateur du groupe de travail humanités numériques au sein de l'association EGC (DAHLIA) et co-fondateur du consortium ARIANE de l'infrastructure de recherche Huma-Num du CNRS ; toutes ces implications fortes renforcent le réseau interdisciplinaire et accroissent le rayonnement du laboratoire ERIC.

Le laboratoire ERIC est rattaché à l'école doctorale (ED) 512 InfoMaths. Le directeur de l'unité fait partie du conseil de l'ED, qui se réunit une à deux fois par an. Les attributions des contrats doctoraux des établissements aux unités de l'ED sont réalisées par une commission constituée des directions d'unités. Cette commission interclasse les dossiers candidat/sujet/direction, en respectant un équilibre entre les unités, avec un système de mémoire sur plusieurs années. Du fait de la pression importante du côté de l'Université Lyon 1 (peu de contrats et beaucoup de demandes), nous n'avons jamais obtenu de contrat doctoral de la part de cette tutelle. Cependant, l'Université Lyon 2 offrant un à deux contrats par an à cette ED, avec comme condition que la direction de thèse soit membre de cet établissement, cela nous place dans des conditions très favorables pour obtenir des contrats doctoraux de l'Université Lyon 2. Ainsi, sur la période évaluée, nous avons obtenu un contrat chaque année (soit au total 5 contrats sur la période, un ayant été rompu précocement).
Par ailleurs, le directeur de l'unité est également le représentant de l'école doctorale auprès de l'Université Lyon 2, et participe notamment aux événements à destination des doctorants ainsi qu'au jury d'attribution du prix de thèse.

ERIC, un acteur majeur dans le domaine des Humanités Numériques

Compte-tenu du positionnement d'ERIC dans une université de Arts, Lettres, Langues, Sciences Humaines et Sociales (ALLSHS), le développement de travaux en Humanités Numériques (HN) a permis de façonner un volet de son identité, à la fois sur un plan local, national et international. Ceci se traduit par des recherches développées de manière collaborative et qui trouvent également appui sur le plan structurel, ce qui amène un contexte favorable pour le développement de cet axe.

À la fois par son appartenance à des structures essentiellement de Sciences Humaines et Sociales (Maison des Sciences sociales et des Humanités : MSH Lyon St-Etienne, où une membre d'ERIC coanime l'axe scientifique Sociétés et Humanités Numériques ; Institut du Genre) et interdisciplinaires (labEx IMU, institut IXXI), par son implication forte dans le master Humanités Numériques co-accrédité par Lyon 2, Lyon 3, l'ENS Lyon et l'ENSSIB (dont la coordination de la mention), et par sa participation à de nombreux projets interdisciplinaires avec divers laboratoires de ALLSHS, ERIC est ainsi devenu un acteur reconnu des Humanités Numériques, non seulement sur le site de Lyon St-Étienne mais également en France. Cela nous a notamment valu d'être sollicités comme partenaire dans deux projets d'EUR déposés en 2019, mais non retenus : PASS – *Past Societies Studies* et GENDERING – *School of Gender Studies*.

Par ailleurs, notons que l'objectif d'intégrer les infrastructures nationales et européennes liées aux HN, soit à travers le master HN (consortiums européens CLARIN et DARIAH), soit à travers des projets de recherche où l'on a commencé à travailler, par exemple, avec l'infrastructure de recherche Huma-Num du CNRS, soit en adhérant aux sociétés savantes francophones et internationales comme Humanistica et ADHO a été atteint.

Sur la période 2020-2024, le laboratoire ERIC a eu des collaborations fortement liées aux HN avec des laboratoires du site et extérieurs à Lyon, dans des disciplines variées :

- Archéologie et Archéométrie (ARAR)
- Bibracte EPCC (archéologie), dans la Nièvre
- Centre Max Weber (CMW, sociologie)
- Centre de recherche en terminologie et traduction (CRTT), devenu en 2021 le CerLA : Centre de recherche en Linguistique Appliquée
- Coactis (Sciences de gestion)
- Environnements et sociétés de l'Orient ancien (Archéorient)
- Equipe de recherche de Lyon en sciences de l'information et de la communication (ELICO)
- Institut de Recherches Géographiques du laboratoire Environnement, Ville, Société (EVS)
- Interactions, corpus, apprentissages, représentations (ICAR)
- Institut des Textes et Manuscrits Modernes (ITEM), à Paris
- Institut de Recherches et d'Etudes Féministes (IREF) de l'UQAM (Université du Québec A Montréal)
- Laboratoire de recherche historique Rhône-Alpes (LARHRA)
- Site-musée d'Ullastret, Catalogne, Espagne (archéologie)
- MARGE (pluridisciplinaire littérature et information-communication)
- Théorie et histoire des arts et des littératures de la modernité (THALIM), à Paris
- Transversales (Droit)

En outre, une partie de ces collaborations a consisté en des codirections de thèses, ce qui est un témoin important de notre activité de recherche pluridisciplinaires. Ainsi, au cours de la période, 4 thèses ont été co-encadrées avec des collègues relevant de disciplines de ALLSHS (les deux dernières étant encore en cours) :

- Pegwende Nicolas Sawadogo. Des lacs de données à l'analyse assistée de documents textuels et tabulaires. Co-direction avec le laboratoire COACTIS (Gestion).
- Max Béligné. Remise en question d'une lecture kuhnienne de la géographie française : réflexions épistémologiques entre sciences sociales, humanités numériques et données massives. Co-direction avec le laboratoire EVS (Géographie).

- Francesco Amato. Mixed data temporal clustering for modelling longitudinal surveys. Co-direction avec le laboratoire COACTIS (Gestion).
- Alyssa Giraudo-Turgis. L'apport du numérique dans l'étude des objets et des sites du Haut-Empire romain. Co-direction avec le laboratoire ArAr (Archéologie).

ERIC, un laboratoire engagé dans ses établissements

Parmi les membres du laboratoire, sur la période évaluée, deux membres ont été élu au Conseil d'Administration de l'Université Lyon 2, un a été élu à la Commission Recherche de l'Université Lyon 2 et un autre à la Commission Recherche de l'Université Lyon 1. A noter qu'un membre de l'unité a également été Vice-présidente de la Section disciplinaire compétente à l'égard des usagers de l'Université Lyon 1 de 2021 à 2023.
Notons également que nous avons deux directeurs de composantes de l'Université Lyon 2 au sein de notre unité : le Directeur de l'Institut de la Communication (depuis 2024) et le Doyen de l'UFR d'Anthropologie, Sociologie et Science Politique (depuis 2021).

Les membres de l'unité assurent de nombreuses responsabilités pédagogiques dans les établissements tutelles. Coté Lyon 2, les membres de l'unité sont responsables des licences et masters Informatique, MIASHS et Humanités Numériques. A noter que le Master 2 Informatique comporte 6 parcours, dont 4 sont gérés par des membres de l'unité. Coté Lyon 1, les membres de l'unité assurent la responsabilité du Master MIAGE, d'un double diplôme avec l'Université Nationale de Ho Chi Minh (Vietnam).

L'Université Lyon 2 a pour ambition de développer une recherche dynamique et ouverte, en favorisant notamment « des questionnements transversaux et des approches transdisciplinaires susceptibles de répondre aux grands défis sociétaux »[1]. Notre politique de recherche s'insère totalement dans ce cadre, notamment à travers nos activités dans le domaine des Humanités Numériques. Par ailleurs, « Tout en s'assumant comme université de ALLSHS, Lyon 2 soutient les projets à l'interface des autres sciences »[1], ce qui constitue une politique particulièrement importante pour notre unité, qui nous permet d'avoir un soutien important de la part de l'Université Lyon 2, notamment d'un point de vue des ressources humaines (avec sur la période, deux créations de postes d'enseignants chercheurs, le renouvellement de tous les départs, et un contrat doctoral par année.

1-4 Prise en compte des recommandations du précédent rapport

Suite à la précédente évaluation du laboratoire, un certain nombre de recommandations nous ont été faites (résumées ci-dessous, sous trois aspects), pour lesquelles nous avons entrepris des actions permettant de nous améliorer sur ces différents aspects. Les actions et résultats sont discutés ci-dessous.

Produits et activités de la recherche

- Chercher à publier plus dans les conférences et revues les plus sélectives

Nous avons beaucoup communiqué auprès des membres du laboratoire, lors des différents conseils, sur le fait de viser les conférences et revues les plus sélectives (conférence CORE A ou A*, revues SJR Q1). Des listes des conférences avec les dates butoirs correspondantes ont été tenues à jour sur l'intranet du laboratoire (hébergé sur notre GitLab : https://git.msh-lse.fr/eric). Comme cela est visible sur la figure 8 (section 3.1), les nombres de publications de ce type ont en effet augmenté sur la période, ce dont on peut se féliciter. Ce résultat est néanmoins à nuancer si on prend en compte que le nombre total de publications du laboratoire a également augmenté (figure 7, section 3.1), rendant les proportions globalement stables (légère amélioration de la proportion de conférences A/A*, légère diminution de la proportion de revues Q1).

- Inciter les collègues MCF à passer une HDR

La charge importante en enseignement et responsabilités pédagogiques des collègues du laboratoire conduit souvent ces derniers à s'auto-censurer et à ne pas demander de CRCT ou de délégation, qui pourraient être mis à profit pour préparer des HDR. Néanmoins, même sans recourir à ces demandes, parmi les 9 collègues MCF qui étaient présents en 2019 et qui n'avaient pas encore soutenu d'HDR, 4 l'ont soutenue durant la période 2019-2024. Ce qui a notamment permis à deux d'entre eux d'obtenir un poste PR au sein de notre laboratoire (un recrutement et un avancement de grade).

- Réduire la durée moyenne des thèses

Durant la période précédente, la durée moyenne des thèses était de 4 ans. Nous avons beaucoup travaillé à cela, notamment grâce à la mise en place systématique des comités de suivi individuel. La durée médiane des thèses est désormais de 40 mois, avec un premier quartile à 38 mois et un troisième quartile à 51 mois. La durée des thèses est désormais, dans la grande majorité des cas, tout à fait convenable, même si pour quelques cas particuliers précis (thèse en cotutelle, thèse en SHS co-encadrée) la durée est beaucoup plus longue.

[1] Extrait de la note stratégique de l'Université Lyon 2 du 14 janvier 2025

- Montage de contrats internationaux

Outre quelques projets d'ampleur modérée pour des collaborations internationales avec l'Algérie et l'Uruguay, l'unité a encore des difficultés à monter ou intégrer des projets internationaux d'envergure. Néanmoins, elle a été récemment impliquée dans deux dépôts de projets européens.
Un premier, qui vient tout juste d'être lauréat et courra sur 2025-2026, dans le cadre de l'infrastructure européenne de recherche pour les arts et les humanités (DARIAH), a pour ambition de concevoir un outil « intelligent » de recommandation de métadonnées de documents littéraires, assistant ainsi la création de métadonnées pour les chercheurs en sciences humaines. L'outil utilisera des modèles et techniques d'intelligence artificielle pour proposer des recommandations de métadonnées basées sur l'analyse de données textuelles. Le projet rassemble des partenaires couvrant toutes les langues romanes d'Europe (français, espagnol, portugais, italien et roumain). Pour atteindre ses objectifs en seulement deux ans, le projet AMIS (Advanced Metadata Intelligent System) s'appuiera sur le recrutement d'un ingénieur full stack, d'un postdoctorant en machine learning et d'un ingénieur de recherche en Humanités numériques.
Un second projet européen, dans le cadre de l'appel MSCA-Doctoral Network, vise à développer un consortium européen de chercheurs académiques et industriels, organisés autour d'un réseau de doctorants, sur la thématique de l'analyse des données fonctionnelles (une des thématiques de l'équipe DMD).

Organisation de la vie de l'unité

- Intensifier le recrutement exogène

Lors de l'ensemble de nos recrutements, nous avons cherché à prioriser les recrutements externes. Notamment, en intensifiant notre politique de communication au sujet de nos recrutements, en démarchant de potentiels candidats. Nous avons eu sur la période 2 recrutements de PR, parmi lesquels 1 fut un recrutement externe (équipe DMD). Le recrutement interne sur le second poste PR étant principalement dû à la spécificité de la thématique de recherche (équipe SID), pour laquelle le nombre de candidats dans le profil est faible. Concernant les 4 recrutements MCF, 100% ont été des recrutements externes. Ce qui au total nous donne un taux de recrutement externe de 83%.

- Recruter un ingénieur de recherche

Malgré des demandes récurrentes à nos tutelles, parfois en commun avec d'autres laboratoires, nous n'avons toujours pas réussi à obtenir le recrutement d'un ingénieur de recherche, ne serait-ce qu'à temps partiel. C'est une perte indéniable pour le transfert de nos recherches, et nous le déplorons.

Projet et stratégies à 5 ans

- Pilotage de l'axe transversal Humanités Numériques et affichage des synergies entre équipes

Plusieurs chercheurs d'ERIC des deux équipes DMD et SID sont fortement impliqués dans des projets (ANR, financés par la MSH LSE ou par l'Université Lyon 2) avec des laboratoires de ALLSHS. Ces projets reposent non seulement sur des collaborations avec des chercheurs de disciplines relevant des ALLSHS mais ils impliquent également une collaboration forte des chercheurs d'ERIC entre les deux équipes et ils permettent de renforcer la synergie entre les équipes.
Une formalisation du pilotage de cet axe, à travers la nomination d'un responsable et l'affectation d'un budget dédié fait partie de notre projet pour le futur quinquennal.

- Bien identifier les questions de recherche en Humanités Numériques

Les recherches menées dans les projets relevant des humanités numériques ou dans les thèses co-encadrées par un chercheur d'ERIC et un chercheur en archéologie, en géographie, en gestion, ne sont pas seulement pluridisciplinaires avec l'application des modèles, méthodes, algorithmes informatiques à des domaines des ALLSHS. Elles sont réellement interdisciplinaires avec une synergie étroite entre les disciplines, un échange des idées et un mélange des approches pour élaborer une compréhension plus complète et des solutions innovantes. On peut ainsi citer quelques questions abordées : l'usage collaboratif pour tous et toutes des outils d'informatique décisionnelle, réflexions épistémologiques entre sciences sociales, humanités numériques et données massives, l'usage des outils de science des données pour des chercheurs non informaticiens, les dynamiques sociales au sein de l'informatique tendant ainsi à faire de la discipline un objet de recherche, etc.

- Intégrer la Fédération Informatique de Lyon

Notre demande d'intégration de la FIL a été acceptée au moment du changement de direction de la FIL, en 2021.

2. INTRODUCTION DU PORTFOLIO

Notre portfolio est constitué des portfolios des deux équipes, comprenant chacun 5 éléments introduits ci-dessous.

2- 1 Portfolio de l'équipe DMD

Nous avons choisi de mettre en avant pour l'équipe DMD un brevet, une activité de médiation scientifique, un ouvrage scientifique ainsi que deux publications dans des conférences. Outre les publications que l'on a choisies représentatives des travaux de l'équipe, les autres éléments ont été mis en avant car ils sont caractéristiques de notre activité.

1. Brevet Procédé et dispositif de pesée notamment pour la pesée de véhicules montés sur roues
2. Médiation Scientifique (nuit de la science) : Musée des moulages
3. Global Vectors for Node Representations, R. Briocher et al., WWW, 2019.
4. Statistical Machine Learning for Electricity Load Forecasting, J. Cugliari et al., Springer, 2024.
5. Serialized interacting Mixed membership Stochastic Block Model, G. Poux-Médard et al., ICDM 2022.

Nous avons en effet beaucoup de collaborations avec des entreprises, très souvent sous la forme de thèses CIFRE (figure 4, section 3-1), et la publication de ce premier brevet pour le laboratoire vient démontrer matériellement l'intérêt de nos travaux pour le monde socio-économique.

Nous avons également une activité de médiation scientifique très importante en regard de la taille de notre unité. C'est la raison pour laquelle nous avons choisi de mettre en avant la nuit de la science du 2 octobre 2020 qui a eu lieu au Musée des Moulages et qui était consacrée à l'utilisation de l'intelligence artificielle en Sciences Humaines et Sociales.

Nous avons une activité de recherche importante dans le domaine de l'analyse statistique des données fonctionnelles, dans le contexte de nombreuses tâches d'apprentissage (supervisé, non supervisé…). L'ouvrage choisi est centré sur une application pratique entrant dans le cadre de ces travaux, à savoir la prédiction de consommation électrique. Nous avons travaillé à de nombreuses reprises sur cette question, avec de multiples collaborateurs académiques mais également industriels (EDF, Enercoop, Enedis…). Cet ouvrage consacre l'ensemble de ces travaux.

La première publication que nous avons choisie (R. Briocher et al., WWW, 2019) propose une méthode originale d'apprentissage de représentation pour des données hétérogènes (graphes et données textuelles). Ces deux aspects ayant été beaucoup abordés au sein de l'équipe, nous pensons que cet article est relativement représentatif du type de travaux effectués dans l'équipe.

Le second article (G. Poux-Médard et al., ICDM 2022) propose un modèle de mélange probabiliste pour les données sous forme de graphes, généralisant plusieurs modèles de l'état de l'art, de type Stochastic Block Model, très utilisés par les systèmes de recommandation. Cet article a été sélectionné à la fois parce qu'il développe des approches par modèles probabilistes et également car il s'intéresse à une tâche de recommandation, deux aspects importants de nos recherches au sein de l'équipe.

2- 2 Portfolio de l'équipe SID

Nous avons choisi de mettre en avant pour l'équipe SID une organisation d'un évènement scientifique, un projet ANR, ainsi que trois publications scientifiques.

1. Organisation des trois conférences conjointes ADBIS-TPDL-EDA 2020
2. ANR BI4people
3. A New Physical Design for Distributed Big Data Warehouses in Hadoop, Y. Ramdane et al., International Conference on Conceptual Modelling (ER 2019)
4. Rumor Classification through a Multimodal Fusion Framework and Ensemble Learning, A. Azri et al., Information Systems Frontiers 2022.
5. A robust and efficient vector-based key management scheme for IoT networks, S. Bettayeb et al., Ad Hoc Networks, 2023.

Le choix de mettre en avant l'organisation des trois conférences conjointes *European Conference on Advances in Databases and Information Systems* (ADBIS), *International Conference on Theory and Practice of Digital* Libraries (TPDL) et les journées Business Intelligence & Big Data (EDA) a été motivé par le fait que l'équipe SID est à l'origine des journées EDA et que l'un des membres de l'équipe fait partie depuis de nombreuses années du comité de pilotage de la conférence ADBIS. Cette dernière est une conférence majeure dans le domaine des systèmes d'information et des bases de données. L'édition 2020 a été organisée par le laboratoire, présidée par un membre de l'équipe SID et a rassemblé plus de 500 personnes.

Le projet ANR BI4people a quant à lui été mis en avant car il est caractéristique des travaux développés par l'équipe SID sur l'informatique décisionnelle, et une grande partie de l'équipe a été impliquée dans le projet projet ; il est également caractéristique des collaborations interdisciplinaires.
Enfin, les deux publications ont été sélectionnées du fait qu'elles représentent les travaux de l'équipe et leur variété.

L'article (Y. Ramdane et al., 2019) présente « SkipSJoin », une nouvelle approche pour améliorer les performances des opérations de jointure en étoile dans les entrepôts de données distribués sur Hadoop. Bien que le partitionnement horizontal réduise le trafic réseau, les méthodes existantes nécessitent encore plusieurs cycles MapReduce. SkipSJoin combine des modèles orientés données et orientés charge de travail, permettant d'exécuter une jointure en étoile en une seule étape Spark tout en évitant le chargement de blocs HDFS inutiles. Les expérimentations montrent que cette approche surpasse les approches existantes en termes de temps d'exécution des requêtes.

L'article (A. Azri et al., 2022) apporte des contributions sur la détection de rumeurs dans les médias sociaux avec MONITOR (Multimodal Fusion Framework to Assess Message Veracity in Social Network). Ce travail s'inscrit dans le cadre des approches multimodales, avec l'originalité de traiter en particulier la véracité des images grâce à des indicateurs issus du domaine de l'IQA (Image Quality Assessment), en démontrant la pertinence d'inclure ces indicateurs pour cette tâche. Un accent est mis sur le recours à différentes approches ensemblistes dont les performances sont comparées, avec l'enjeu de proposer des approches dont les résultats sont explicables. Cet article de revue a été sélectionné notamment parce qu'il présente différentes contributions de détection de rumeurs dans les médias sociaux qui répondent à un fort enjeu sociétal, plus que jamais d'actualité, et ayant trait à l'intégrité des données, un aspect important abordé au sein de l'équipe.

L'article (S. Bettayeb et al., 2023) présente un nouveau schéma de gestion des clés, appelé EVKMS *(Efficient Vectors-based Key Management Scheme*), conçu pour sécuriser la collection de données dans les réseaux IoT (Internet des Objets). Ce schéma utilise des vecteurs pré-distribués pour masquer les clés, tout en divisant la zone de déploiement en sous-zones pour améliorer la flexibilité, l'évolutivité et la résilience aux attaques. EVKMS minimise les communications pour l'établissement de clés cryptographiques tout en offrant une sécurité renforcée contre les attaques de capture de nœuds et autres menaces connues. Les résultats de simulations montrent que ce schéma est plus efficace que les solutions existantes en termes de stockage, de communication et de consommation d'énergie, le rendant adapté aux réseaux IoT à ressources limitées.

3. AUTOÉVALUATION DU BILAN

L'argumentaire présenté dans cette section se base notamment sur les données fournies dans le Tableur « Données de Caractérisation et de Production », dénommé TDCP par la suite.

L'unité disposant de deux équipes semblables à de nombreux point de vue (organisation, résultats, rayonnement, …), nous avons choisi de décrire la majorité des éléments d'auto-évaluation au niveau de l'unité. Les auto-évaluations des équipes se concentrent sur l'aspect purement scientifique, en détaillant les références 1 des domaines 1 et 2, respectivement les objectifs et les réalisations scientifiques.

Par ailleurs, comme cela a été présenté dans la Section 1.2, l'unité dispose d'un axe de recherche transversal aux deux équipes : les Humanités Numériques. De fait, les éléments liés à cet axe de recherche seront déclinés au sein de l'auto-évaluation de l'unité, laissant aux équipes uniquement les éléments qui leur sont propres.

3- 1 Autoévaluation de l'unité

Domaine 1. Objectifs scientifiques, organisation et ressources de l'unité

Référence 1. L'unité s'est assigné des objectifs scientifiques pertinents et elle s'organise en conséquence.

Outre les objectifs scientifiques fixés au niveau des équipes, pour lesquels le bilan sera présenté en équipe, le laboratoire s'était fixé quatre objectifs :

1. Améliorer la qualité des publications
2. Inciter les membres MCF du laboratoire à passer une Habilitation à Diriger des Recherches
3. Mettre en valeur notre production logicielle
4. Intensifier nos actions de médiation scientifique

Les deux premiers points ayant été identifiés comme des recommandations au sein du précédent rapport d'évaluation, les réponses apportées ont déjà été détaillées dans la section 1.4 du présent document.

Concernant la valorisation de notre **production logicielle**, notre politique fût fortement orientée vers le recrutement d'un ingénieur d'appui à la recherche, recrutement que nous pensions possible suite aux échanges avec nos tutelles en début de contrat quinquennal. Or, le recrutement espéré n'aura jamais eu lieu, et les logiciels que nous avons développés et que nous continuons à développer restent malheureusement trop souvent à l'état de prototype. Notons néanmoins que nous avons sensibilisé les membres de l'unité à la nécessité de déposer leur production logicielle sur HAL, accompagnée lorsque cela est possible du dépôt du code source.

Concernant l'aspect **médiation scientifique**, nous pouvons affirmer que l'unité est très active dans le domaine, avec une présence et une reconnaissance forte à l'échelle du site Lyon Saint-Etienne. Cette activité est portée par environ un quart des membres du laboratoire, avec une activité très significative dans le domaine. Ce point sera développé dans la référence 3 du Domaine 3.

Référence 2. L'unité dispose de ressources adaptées à son profil d'activités et à son environnement de recherche et les mobilise.

Cette section s'appuie et résume les données du Tableau de Données de Caractérisation de Production - TDCP). Le lecteur pourra s'y référer pour plus de détails.

Une grande part de financement par projet

Les ressources de l'unité proviennent en grande partie (92%) de financement par projet, pour 3 616 k€ sur le contrat quinquennal contre 317 k€ en ressources récurrentes (8%). Ainsi, l'unité dispose en moyenne **chaque année** de **602k€ de recettes sur projets** et de **53k€ de recettes récurrentes**. Ces dernières proviennent à 70% de l'Université Lyon 2 et à 30% de l'Université Lyon 1. Il faut néanmoins tempérer ces chiffres, car ils ne tiennent pas compte des charges de ressources humaines récurrentes supportées par nos tutelles, à savoir les salaires de tous les membres du laboratoire qui sont entièrement supportés par nos deux tutelles. Néanmoins, ils témoignent que l'orientation de nos recherches est fortement guidée par les projets que nous portons.
Concernant nos ressources récurrentes, 30% sont utilisés par le laboratoire pour l'équipement matériel, la gestion du fonctionnement courant du laboratoire, ainsi que les missions des chercheurs invités dans le cadre de séminaires ou de jurys de thèse ou HDR. Les 70% restant sont affectés aux équipes, au prorata de leur effectif, et sont utilisés pour des financements de missions et de stage.

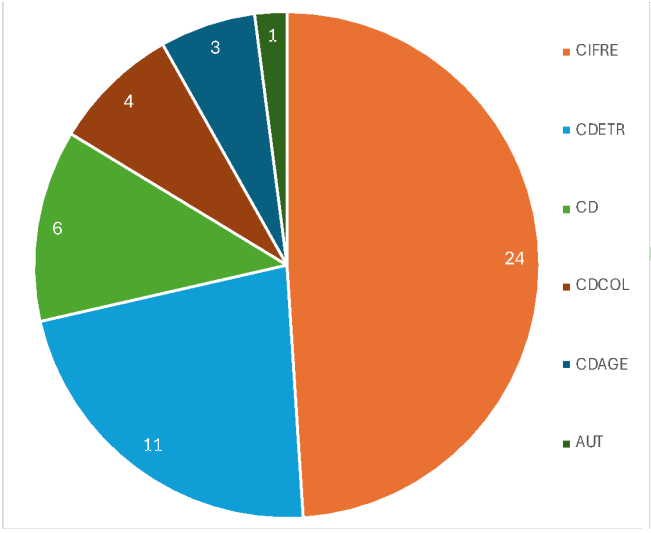

Figure 5 : Sources de financement des 49 thèses (nombre)

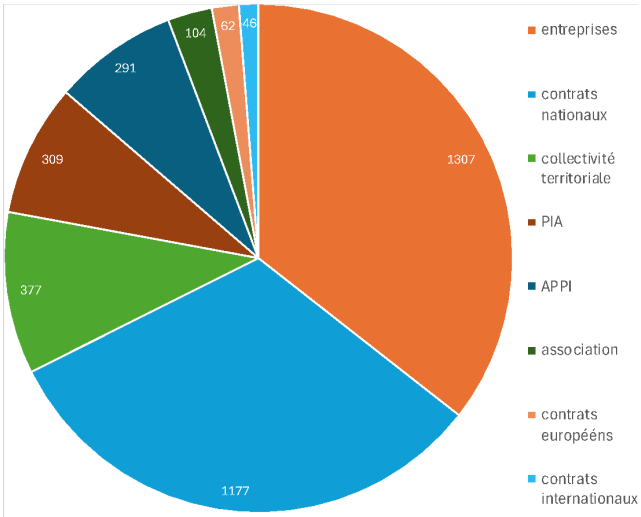

Figure 6 : Sources de financement par projet (en K€)

Pour ce qui est des ressources par projets, elles sont naturellement dédiées aux projets qu'elles financent. Comme cela peut être vu sur la figure 6, un gros tiers de nos ressources sur contrat provient de collaborations avec des **entreprises**, principalement liées à l'encadrement de thèses CIFRE. Cela se retrouve notamment dans les sources de financement de thèses, avec environ 50% des thèses financées par des entreprises. C'est un des points forts du laboratoire, reconnu – tant du point de vue des entreprises que du point de vue de ses partenaires académiques – pour ses compétences dans le domaine (au sens large) de la science des données, et pour la capacité de ses chercheurs à identifier dans les problématiques en entreprises les besoins en développement d'activité de recherche. Les demandes en collaboration que nous recevons étant supérieures à nos capacités de réponse, cela nous permet notamment de nous engager que dans des collaborations dont le sujet est très en adéquation avec les thématiques de recherche que l'on souhaite développer. Ainsi, même si une grande part de nos financements provient d'entreprises, cela n'altère pas nos choix d'orientation scientifique.
La seconde plus importante source de financement provient des **appels à projets nationaux** (principalement de type ANR), ce dont nous pouvons nous féliciter car c'était un des objectifs que nous nous étions fixés. Pour cela, une stratégie a été mise en place au sein du laboratoire, de sorte que chacune des équipes puisse être impliquée chaque année dans le dépôt d'un projet ANR, en incitant un roulement entre les collègues afin de partager les efforts, la charge de soumission de tel projet étant lourde. Il est à noter également, que plusieurs membres de l'unité sont systématiquement associés à chaque projet de ce type, mêlant souvent chercheurs seniors et juniors, et favorisant ainsi la mutualisation des ressources.
Enfin, même si c'est moins important d'un point de vue financier, un effort important a été entrepris pour favoriser le dépôt, par les membres juniors du laboratoire, de projets internes, notamment des projets de la Fédération Informatique de Lyon, afin de multiplier les collaborations avec les chercheurs du site Lyon Saint-Etienne. Cette démarche a été couronnée de succès avec 11 projets FIL financés sur le contrat quinquennal, tous en collaboration avec différents laboratoires d'informatique du site.

Un nombre important de doctorants…

L'unité accueille et forme un nombre important de doctorants : 49 étudiants ont réalisé une thèse au cours du contrat quinquennal, pour un total de 26 soutenances et 21 thèses toujours en cours (et 2 abandons). Ce nombre est relativement important, quand on le compare au nombre de chercheurs HDR (12 au 31/12/2024), sachant qu'une grande majorité des thèses est co-encadrée par un binôme HDR/non HDR.
Comme cela est visible sur la figure 5, la moitié de ces doctorants est financée par des entreprises, un petit quart par des gouvernements étrangers ; le reste provenant d'allocation doctorale de nos tutelles (avec en moyenne une allocation chaque année), ainsi que de projets nationaux ou régionaux.
Du fait de cette pluralité de sources de financement, les doctorants débutent leur thèse tout au long de l'année. Lors de leur arrivée, ils sont accueillis et installés par leur direction de thèse et par la direction du laboratoire. Une présentation aux équipes et réalisée officiellement lors des réunions d'équipe mensuelles.
Nous sommes également bien impliqués dans la formation par et pour la recherche. Parmi les membres de l'unité, nous comptons les responsables du master informatique (et de 4 de ses parcours), du master humanités numériques et du master MIASHS de l'Université Lyon 2. Par ailleurs, nous entretenons également d'étroites collaborations avec les masters d'informatique et de mathématiques appliquées de l'Université Lyon 1. Ces différentes implications dans les masters nous permettent de recruter une grande partie de nos doctorants (39%) dans l'une de nos deux universités de tutelle. Enfin, nous pouvons noter que les membres du laboratoire ERIC assurent pour l'Université de Lyon plusieurs formations transversales liées au machine learning, d'un niveau basique à un niveau plus avancé.

… et de stagiaires

La liste de tous les stagiaires n'a pas été dressée dans l'onglet ressources humaines dans le TDCP. Néanmoins, nous en accueillons en moyenne 18 par an, dont environ les ¾ pour un stage de master de plus de 3 mois. Ces stagiaires sont financés pour les trois quarts par des APP internes (FIL, Lyon 2), et pour un quart directement par les équipes de recherche sur leur budget propre. Les travaux menés par ces stagiaires concernent très souvent l'exploration de pistes de recherches nouvelles, ce qui est notamment permis par ces deux types de financement. Concernant l'origine des étudiants en stage, il est assez similaire à celui des doctorants en thèses, avec une moitié provenant de formations du site Lyon Saint-Etienne.

Référence 3. L'unité dispose de locaux, d'équipements et de compétences techniques adaptés à sa politique scientifique et à ses objets de recherche.

Locaux

Les locaux du laboratoire ERIC sont principalement situés au sein du Campus Porte des Alpes de l'Université Lyon 2, où nous disposons de 17 bureaux répartis sur 2 étages dans le bâtiment K, d'un espace de réception, d'une salle de séminaire et d'un espace café/repas. En outre, nous disposons d'un demi-bureau sur le Campus de la Doua de l'Université Lyon 1. Ces locaux sont anciens, relativement dégradés, et nous intégrerons de nouveaux locaux dans « la Ruche » en 2026. Nous serons situés au dernier étage du nouveau bâtiment qui accueillera notamment le nouveau learning center de l'Université Lyon 2.
Les EC de l'Université Lyon 2 sont répartis à 2 ou 3 par bureau, rendant parfois difficile la cohabitation en cas de réunion en présentiel ou en visio-conférence. Des places sont également disponibles pour les collègues de l'Université Lyon 1 lors de leur passage au laboratoire.
La gestion des bureaux pour les doctorants (et ATER et post-doc) est compliquée du fait du manque de bureaux disponibles, et de la présence que partielle des doctorants en thèse CIFRE (qui représentent une part importante de nos doctorants). Nous avons instauré fin 2024 un système de flex-office, de sorte que chacun puisse s'installer où il le souhaite lorsqu'il arrive. Ce système de fonctionnement, qui semble convenir aux doctorants, augurera de notre utilisation de nos futurs locaux en 2026.

Équipements

Concernant les ressources de calculs, même si de nombreuses ressources mutualisées sont disponibles et utilisées par le laboratoire (IN2P3, Jean Zay…), il nous est indispensable de disposer de ressources locales plus faciles à utiliser, notamment en phase préliminaire de tests. Nous disposons de ce fait de trois serveurs de calculs :

- Neo : serveur pour le calcul sur CPU, 40 threads, 92Go RAM,
- Cholula : serveur pour le calcul GPU, 3 NVidia GTX 1080 Ti 11Go VRAM, 3 Nvidia GTX 1080 8Go VRAM, 32 threads, 128Go RAM,
- Chimichurri : serveur pour le calcul GPU, 1 Nvidia RTX A6000 48Go VRAM, 40 threads, 176Go RAM.

Nous disposons également de 3 serveurs de virtualisation PowerEdge R640 (Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz, 40 To stockage).
En l'absence de personnel d'appui à la recherche, c'est une cellule informatique constituée de 4 ECs qui gère ces serveurs, et cela bénévolement, c'est-à-dire sans aucune reconnaissance de nos tutelles.
Si ces serveurs étaient historiquement situés dans nos locaux, nous les avons déplacés vers les locaux de la Direction des Services Informatiques de l'Université Lyon 2. Cela permet d'alléger la cellule informatique d'une partie de la gestion de ces serveurs (notamment liés à la gestion matérielle), sans coût supplémentaire, l'Université Lyon 2 nous offrant ce service.

Référence 4. Les pratiques de l'unité sont conformes aux règles et aux directives définies par ses tutelles en matière de gestion des ressources humaines, de sécurité, d'environnement et de protection des données ainsi que du patrimoine scientifique.

Gestion des ressources humaines

La politique de recrutement de l'unité est une politique pluriannuelle, établissant en début de quinquennat une liste de recrutements de membres permanents par ordre de priorité, en veillant à un équilibre entre équipe ainsi qu'à un équilibre PR / MCF. Néanmoins, tous nos recrutements étant sur des postes d'enseignants-chercheurs, il est nécessaire de trouver des équilibres avec les composantes d'enseignement, afin de proposer des postes cohérents d'un point de vue enseignement et recherche. En effet, les postes en informatique concernent différentes composantes telles que l'Institut de la Communication, l'IUT, l'UFR Anthropologie, Sociologie et Science Politique, etc.

Concernant les questions de parité et de non-discrimination, nous suivons totalement les politiques mises en place par nos tutelles, et utilisons notamment les outils de sensibilisation mis en place, notamment à destination des comités de sélection. Étant dans un domaine de recherche où la présence des femmes est moindre, nous agissons à travers nos différentes activités (médiation, thématique de recherche sur le genre) pour inciter ces dernières à se diriger vers le domaine de l'informatique notamment.
La tenue systématique de comité de suivi de thèse pour les doctorants, avec notamment des échanges entre le doctorant et les membres du comité sans le directeur de thèse, permet de prévenir un maximum de risques psycho-sociaux. Concernant les membres permanents, aucun dispositif n'est mis en place au niveau du laboratoire, mais ceux-ci peuvent bénéficier des moyens disponibles au niveau des universités tutelles.
Enfin, de nombreux moments de convivialité sont organisés pour favoriser les échanges entre les membres de l'unité, contribuant ainsi à améliorer le bien-être au travail et ainsi à prévenir les risques psycho-sociaux.

Protection des données et du patrimoine scientifique

Pendant le déroulement de nos activités de recherche, nous utilisons fortement le GitLab mis à disposition par la Maison des Sciences sociales et des Humanités, qui nous assure une protection avancée de nos travaux. Autant que faire se peut, nous déposons nos résultats de recherche sous la forme de prépublications, en attendant qu'ils soient publiés dans des revues ou conférences.

Protection de l'environnement

La protection de l'environnement est un double sujet pour notre unité. Tout d'abord dans notre activité professionnelle de tous les jours. Malgré des locaux vétustes, que nous pouvons qualifier de passoires thermiques, nous nous efforçons de préserver l'environnement. Pour cela, l'unité a équipé ses membres en doudounes et sweats afin d'affronter le froid pendant l'hiver et de minimiser le recours à des systèmes de chauffage d'appoint. Notons également les efforts de l'Université Lyon 2 sur le campus Porte des Alpes et de la Métropole de Lyon pour favoriser les déplacements domicile-travail à vélo ou en transport en commun. Ainsi, assez peu de membres utilisent la voiture pour se rendre au laboratoire.
Le second aspect concerne notre objet de recherche, notamment en ce qui concerne les travaux en machine learning de l'équipe DMD. A l'heure du deep learning et de la course en avant vers des modèles toujours plus énergivores, une partie de nos activités de recherche traite de l'IA frugale, c'est-à-dire peu gourmande en ressources énergétiques, que ce soit en termes de volume de données et de ressources de calculs nécessaires pour l'optimisation des modèles. Ce point sera développé dans la section 3.2.

Domaine 2. Les résultats, le rayonnement et l'attractivité scientifiques de l'unité

Référence 1. L'unité est reconnue pour ses réalisations scientifiques qui satisfont à des critères de qualité.

Réalisations scientifiques des deux équipes DMD et SID

Nous résumons ici les principaux résultats des deux équipes de recherche, qui seront détaillés dans la section 3.2. Au sein de chacune des équipes, la présentation des thématiques de recherche est faite selon un regroupement en axe thématique. Cette notion d'*axe thématique* n'est pas la notion formelle définie au sens de l'HCERES, le laboratoire ERIC étant organisé en équipes.

L'équipe de recherche DMD a produit plusieurs résultats scientifiques significatifs dans les domaines de l'apprentissage automatique, de l'apprentissage statistique et de la représentation des connaissances.
Dans l'axe "Apprentissage automatique, apprentissage statistique et optimisation", l'équipe a développé des modèles de mélanges probabilistes pour traiter l'hétérogénéité des données, notamment pour des données fonctionnelles et ordinales, pour des tâches non supervisées (clustering) et supervisées (régression). Les travaux sur les signatures de Terry Lyons, définissant une représentation en dimension finie d'une courbe, ont permis de détecter des anomalies dans les données temporelles multidimensionnelles, avec des applications en médecine et en sciences cognitives. L'équipe a également exploré des techniques d'optimisation combinatoire et d'apprentissage hybride, notamment pour la planification dynamique de tournées de véhicules et la minimisation de la consommation d'énergie dans les bâtiments.
L'apprentissage par transfert a également été une thématique clé, avec des contributions originales sur l'adaptation de domaine et l'utilisation de modèles de mélanges pour estimer les distributions des données sources et cibles. Des approches bayésiennes pour le transfert de processus gaussiens et de modèles d'équations différentielles, avec des applications industrielles, ont également été développées.
Dans l'axe "Représentation des connaissances et recherche d'information sur des corpus textuels", l'équipe a développé des méthodes de représentation de documents, d'auteurs et de graphes de réseaux de documents. Les modèles développés ont permis d'améliorer la classification et le résumé de documents textuels. La modélisation des interactions dans les réseaux d'information a également été explorée, avec des contributions significatives sur les Stochastic Block Models et les Processus de Dirichlet.

En matière de prévision et d'incertitude, l'équipe a développé des outils théoriques pour analyser la stabilité des modèles dans un contexte de fairness, garantissant l'équité des prises de décisions.
Ces résultats scientifiques démontrent l'impact et la pertinence des recherches de l'équipe DMD, tant sur le plan théorique que pratique, avec des applications concrètes dans divers domaines industriels et académiques.

L'équipe de recherche SID a produit plusieurs résultats scientifiques notables dans le domaine de la gestion et de l'analyse des données massives. Dans l'axe "Big Data Management", elle a développé des modèles NoSQL optimisés pour le stockage et l'interrogation des données hétérogènes, ainsi que des solutions pour les lacs de données, permettant de centraliser de grands volumes de données brutes tout en préservant leur souplesse d'exploitation. Les travaux sur les entrepôts de données orientés graphes et les schémas évolutifs multi-versions ont amélioré la flexibilité et l'adaptabilité des systèmes de gestion de données.
En matière de sécurité des données, l'équipe SID a proposé des méthodes innovantes pour la gestion sécurisée des clés cryptographiques dans les réseaux IoT et des solutions pour la traçabilité et la conformité réglementaire des données. Les recherches sur l'étiquetage des données et l'annotation automatique ont renforcé la protection des données tout au long de leur cycle de vie.
Dans l'axe "BI & Analytics", l'équipe SID a optimisé les performances des requêtes sur Hadoop avec l'approche SkipSJoin et développé des opérateurs OLAP adaptés aux bases NoSQL, améliorant ainsi les capacités d'analyse décisionnelle. Les applications concrètes incluent la médecine de précision, la détection des rumeurs sur les réseaux sociaux et l'analyse des interactions humaines avec l'environnement. Ces travaux ont été publiés dans des revues et conférences internationales, démontrant l'impact et la pertinence des solutions proposées par l'équipe SID.

La figure 7 affiche les mots clefs les plus utilisés dans les publications de l'unité.

Figure 7 : Mots clefs des publications du laboratoire ERIC. La taille et la centralité de la position sont proportionnelles au nombre d'utilisations du mot clef.

Axe de recherche transversale en Humanités Numériques

Depuis plus de 15 ans, nous avons souhaité affirmer une identité non seulement pluridisciplinaire informatique et mathématiques appliquées, présente depuis la création du laboratoire, mais également interdisciplinaire, en structurant et en développant nos collaborations avec les disciplines des ALLSHS en un axe stratégique du laboratoire. Pour cette raison, nous développons cet axe de manière séparée dans le présent rapport. Il est transversal aux équipes DMD et SID ; il permet en outre de développer la synergie entre les deux équipes.

En réponse à une demande des évaluateurs du laboratoire en 2020 de bien identifier les questions de recherche en HN, notre objectif a été de ne pas seulement trouver des terrains d'application à nos recherches, mais surtout d'hybrider les méthodologies informatiques et statistiques à celles des ALLSHS pour aboutir à des approches originales. Nous nous inscrivons également dans des collaborations sur le temps long, nécessaire à un travail interdisciplinaire qui porte réellement ses fruits.

Une grande partie de nos travaux dans le domaine des Humanités Numériques a été possible grâce à des financements de nombreux projets de recherche. Nous décrivons ici les principaux projets, qui donne un nouvel aperçu de nos activités dans le domaine.

Le projet ANR *LIFRANUM*, porté par le laboratoire MARGE avec la Bibliothèque nationale de France comme partenaire, visait à développer une plateforme d'analyse de la création francophone nativement numérique. Dans ce cadre, nous avons travaillé sur la mise au point d'un système d'information et à des nouveaux modèles d'apprentissage de représentation d'auteurs (cf. thèse de Enzo Terreau).

Le projet ANR *BI4people*, porté par notre unité, visait à rendre l'informatique décisionnelle disponible en ligne à des usagers disposant de peu de ressources financières et de connaissances techniques, en automatisant des processus actuellement aux mieux semi-automatiques. Par ailleurs, nous avons insisté sur l'importance de l'appropriation des visualisations fournies par l'outil par les usagers, ce qui impliquait une collaboration interdisciplinaire entre l'informatique et les sciences de l'information et de la communication (laboratoire ELICO).
Le projet ANR *HisArc-RDF* (Partage et réutilisation de données archéologiques et historiques : une description en RDF appuyée sur les référentiels et les normes du web sémantique), porté par le laboratoire Archéorient, visait à prototyper une chaîne opératoire FAIR (Findable, Accessible, Interoperable, Reusable) sur des jeux de données archéologico-historiques structurellement et sémantiquement hétérogènes. Outre le laboratoire Archéorient, nous avons collaboré avec de nombreux partenaires dans le cadre de ce projet : LAHRHA, HiSoMa, Bibracte EPCC, Chrono-environnement Besançon, Archimède Strasbourg, AOrOc Paris et les entreprises ABES et Archéodunum.

Nous pouvons également citer, sans les détailler, les projets région *DataLAC* (Données, Archives et Textes Archéologiques : création et exploitation d'un Lac de données sémantiques pour l'Archéologie de la Catalogne) et PMI (Transformation numérique, servicisation et mutations des modèles d'affaires des PME industrielles), ainsi que de nombreux projets locaux, financés soit par la MSH-LSE (*PicLetters* - Picasso en toutes lettres), l'Université Lyon 2 (*GéoDOAD* - Géolocalisation dynamique des données archéologiques datées ; *SO-COEQUAL* - COmputer EQUALity), le Labex IMU (*HyperThésau* - Hyper thésaurus et lacs de données : fouiller la ville et ses archives archéologiques ; IDENUM - Identités numériques urbaines), l'IDEX Université de Lyon (*CartoWeb* - Cartographie du Web littéraire francophone) ou l'Institut du Genre (*GRADIENT* - vers une méthode centrée sur les données pour estimer l'écart entre les sexes dans les emplois en science des données et en intelligence artificielle dans l'industrie et le milieu universitaire français).

Pour illustrer nos propos, nous décrivons ci-dessous quelques exemples de nos travaux dans le domaine des Humanités Numériques.
Dans [M. Selosse et al. 2019], un outil d'analyse non supervisée de questionnaires longitudinaux d'évaluations de la qualité de vie de patients est proposé. Ces questionnaires comportant de nombreuses questions, et les réponses attendues étant sur une échelle ordinale, leur traitement nécessitait le développement d'un modèle statistique dédié. De fait, un modèle de co-clustering génératif à base de modèle des blocs latents a été proposé. Ce modèle permet de résumer l'information en créant des groupes de patients homogènes du point de vue de leur qualité de vie ressentie, mais également des groupes de questions retraçant les mêmes aspects de cette qualité de vie.
Dans [Liu et al. 2020], un prototype de lac de données structuré en neuf couches est proposé : ingestion, stockage, application, gouvernance et sécurité des données. Ce système de gestion des métadonnées et le modèle de métadonnées est alors utilisé pour gérer les données archéologiques au sein du lac de données.
Dans [C. Favre et al. 2023], une restitution détaillée du regard réflexif porté sur la construction de catégories d'analyse et leurs usages en science des données. Plus précisément, cet article aborde la non-neutralité dans l'analyse selon les catégories et leurs modalités, se questionne sur qui les construit et comment, et aborde les questions des contributions multidisciplinaires et de la posture du data scientist.
[C. Cote et al 2023] présente un corpus de littérature francophone ainsi que sa méthode d'identification et de collecte. Il aborde également la caractérisation et l'indexation des contenus en vue de leur usage dans un cadre à la fois de recherche et d'enseignement.

Un de nos axes de recherche dans le domaine des Humanités Numériques est la scientométrie, qui porte sur l'analyse des données de la science ? Cet axe est à la fois abordé dans une perspective d'informatique seule mais également d'un point de vue interdisciplinaire permettant une analyse enrichie grâces aux sciences humaines et sociales, notamment la sociologie. C'est le cas des travaux menés notamment dans le cadre du défi 2020 de la conférence EGC, pour lequel l'association a mis à disposition les données concernant la communauté scientifique d'EGC. Ces travaux ont amené un regard original sur les matériaux fournis en abordant l'articulation des temps de travail, la place des femmes dans la communauté et la dimension écologique. Ce travail a remporté le prix du défi d'EGC 2020.
Finalement, le laboratoire ERIC s'inscrit aussi dans une démarche favorisant l'interdisciplinarité du point de vue du champ par essence interdisciplinaire que sont les études de genre. Ceci s'ancre à la fois en termes de projets, de recherches qui ont donné lieu à des communications et des articles sur le sujet de la masculinisation de l'informatique et la question des carrières de femmes dans ce domaine, mais aussi en termes d'animation scientifique, avec l'organisation ou la participation à des événements et des projets dont l'ancrage informatique a été développé. L'Université Lumière Lyon 2 fait partie des partenaires du Groupe d'Intérêt Scientifique « Institut du Genre » et le laboratoire en était partenaire jusqu'à ce que les modalités d'affichage de partenariat pour se maintenir au niveau des établissements (il est à noter qu'une des membres du laboratoire a intégré le conseil scientifique pour la prochaine mandature).

Implication dans des programmes d'investissements nationaux

Nous avons été impliqués dans un certain nombre de dépôts de projets d'investissement nationaux, pour lesquels des réponses au niveau du site Lyon Saint-Etienne ont été soumises. Dans la majorité de ces projets, nous intervenons en tant que contributeurs sans occuper de responsabilité particulière.
Certains de ces projets ont été lauréats, ce qui nous a permis par la suite de décrocher des financements.
Nous pouvons citer le projet SMAD-CC financé par ShapeMed, lauréat de l'APP Excellences France 2030, le projet TIGA dans le cadre du programme Territoires d'Innovation (PIA 3), les projets IDENUM et HyperThesau financés par le Labex Intelligence des Mondes Urbains (IMU), le projet Numérique et Inégalité dans le cadre du labex ASLAN et le projet CartoWeb dans le cadre de l'Idex Université de Lyon. Nous pouvons également citer notre participation au programme IADoc@UdL de l'Université de Lyon en réponse à AAP Contrats doctoraux en Intelligence artificielle de l'ANR.
Nous pouvons également citer notre forte implication dans le projet AILyS, malheureusement non retenu lors de l'AAP IA Cluster de France 2030, avec plusieurs membres du laboratoire impliqué dans le comité de direction.
Dans le domaine des Humanités Numériques, nous pouvons citer le projet PIA franco-canadien *Numérique et inégalités*, qui associe la faculté des Arts d'Ottawa, la chaire Humanités numériques du Canada, le laboratoire ICAR et le laboratoire ERIC. L'objectif du projet est d'une part d'examiner les façons dont la technologie numérique renforce et même augmente des inégalités déjà existantes dans les sociétés canadiennes et françaises et d'autre part de se pencher sur les manières dont la technologie numérique peut fournir une réponse à ces inégalités en augmentant l'inclusion ou en diminuant la marginalisation.

Rayonnement et indices de reconnaissances

L'unité n'a pas reçu de distinctions scientifiques majeures, et aucun de ses membres n'est IUF. Néanmoins, à notre niveau, nous avons eu quelques reconnaissances qui témoignent de la qualité de notre travail.
Un de nos brillants doctorants, Gaël Poux-Médard, a reçu le prix de thèse (accessit) de l'Université Lyon 2 (2023) ainsi que le prix de thèse EGC en 2024.
Un collègue, Antoine Rolland, fortement impliqué dans la diffusion scientifique, a reçu le prix du meilleur article dans la revue Tangente (en 2019), qui est une revue de mathématiques à destination des lycéens.
Enfin, Cécile Favre a reçu le prix du défi EGC 2020.
Par ailleurs, plusieurs collègues ont exercé des responsabilités dans des sociétés savantes : Trésorier de la Société Française de Statistique (SFdS) et Vice-Présidence d'INFORSID.

Référence 2. Les activités de recherche de l'unité donnent lieu à une production scientifique de qualité.

Production scientifique

La Table 1 détaille les nombres de publications de l'unité par type de publication sur la période évaluée (ainsi que les pourcentages lorsqu'il dépassent les 5%). La figure 8 présente l'évolution, par rapport aux deux précédents contrats quinquennaux, des nombres de publications dans des revues et des conférences, qui sont les deux principaux types de publications de l'unité. Enfin, la figure 9 présente l'évolution des publications de très haute qualité, c'est-à-dire les articles dans des revues SJR Q1 et dans des conférences CORE A ou A*.

Du fait de la spécificité thématique de l'unité, regroupant à la fois des chercheurs de la CNU 26 et de la CNU 27, les supports de publications privilégiés diffèrent suivant les membres : les chercheurs relevant de la CNU 27 publient majoritairement dans des conférences, tandis que les chercheurs de la CNU 26 publient majoritairement dans des revues tout en participant à des conférences non classées dans une optique de communication et d'échange avec les membres de la communauté. Cette particularité engendre une part de conférences non classées relativement importante, et nous n'avons pas souhaité, par soucis d'unité et de simplicité, séparer les conférences relevant de la CNU 26 des autres.

Article dans une revue	90 (21%)	Communication dans un congrès	203 (47%)	Poster	17
Proceedings/Recueil des communications	7	No spécial de revue/special issue	6	Ouvrage (y compris édition critique et traduction)	4
Chapitres d'ouvrage	14	Article de blog scientifique	2	Autre publication	8
Pré-publication, Document de travail	16	Brevets	1	Thèse	26 (6%)
HDR	4	Rapport	4	Logiciel	24 (6%)
Media	3				

Table 1 : Production scientifique du laboratoire ERIC

Un de nos objectifs était d'augmenter nos publications dans des revues et conférences de haut niveau, ce qui a bien été réalisé (figure 9), **avec une augmentation de 28% du nombre de revues Q1 et de 110% du nombre de conférences A/A***. Ces chiffres sont notamment à tempérer du fait que les nombres totaux de publications (figure 8) ont eu aussi augmenté, et les efforts sur la qualité des publications devront être poursuivis.
Les conférences les plus sélectives (A/A*) dans lesquelles nous publions régulièrement sont les grandes conférences dans le domaine de l'apprentissage automatique et de l'IA (IJCAI, ECML PKDD, ICDM), de la recherche d'information (ECIR, SIGIR), des technologies web (WWW) et du traitement du langage naturel (NAACL, EMNLP). Concernant les revues les plus sélectives (Q1), nous publions principalement dans des revues de machine learning (Journal of Machine Learning Research, Pattern Recognition), de statistique (Statistics and Computing, Computational Statistics and Data Analysis, Annals of Applied Statistics, Journal of the Royal Statistical Society: Series C), d'informatique et systèmes d'information (Knowledge and Information Systems, Information Systems Frontiers, Information Sciences, Journal of Information Security and Applications, Ad Hoc Networks, ACM Transactions on Internet Technology), de mathématiques et physique (Physical Review Research, Journal of Fourier Analysis and Applications, Fuzzy Sets and Systems), mais également de sciences sociales et économiques (Social Choice and Welfare, Social Indicators Research).
Nous pouvons noter également un nombre relativement important de logiciels produits sur la période (la liste et leur description est disponible dans le TDCP). Il est sûr que parmi ces logiciels, certains mériteraient d'être déployés à un niveau plus professionnel et pourraient doper nos résultats en transfert et valorisation. Malheureusement, sans ingénieur d'appui à la recherche, nous ne sommes pas en mesure d'aller plus loin que le déploiement de prototypes.
A noter également que le premier brevet du laboratoire a été publié sur la période, en collaboration avec une PME locale (et est un des éléments du Portfolio).

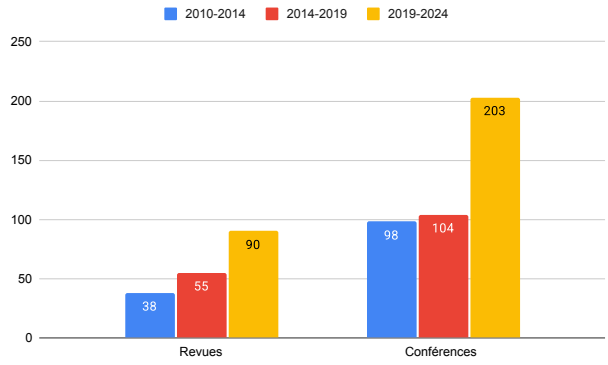

Figure 8 : Évolution du nombre de publications dans des revues et conférences

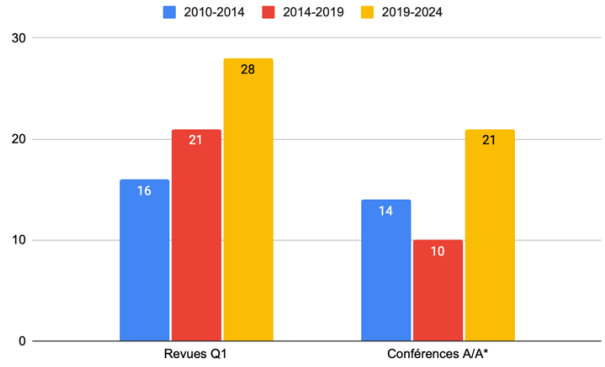

*Figure 9 : Évolution du nombre de revues Q1 et conférences A/A**

La figure 10 illustre le **nombre moyen de publications par année** des membres permanents de l'unité. Seules les publications de types revues et conférences sont comptabilisées. La moyenne de publications est de 2.5 publications par an, avec un écart-type de 2. Quatre membres de l'unité n'ont pas publié pendant cette période : deux membres en fin de carrière, dont un qui est parti au cours de la période ; un membre de retour de détachement, qui nous a quitté peu après son retour ; un collègue très investi dans la diffusion de matériel pédagogique, auteur d'un grand nombre d'ouvrages et cours en ligne.

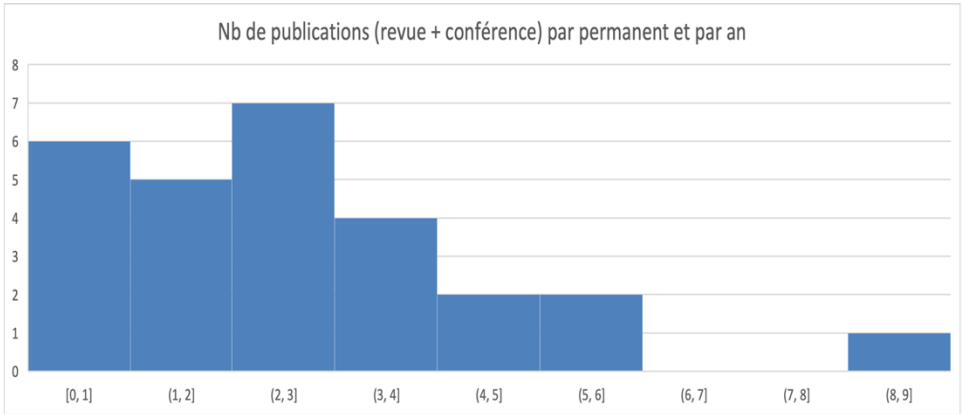

Figure 10 : Nombre moyen de publications (revues et conférences) par permanent et par an

Les **deux équipes** ont une production très **équilibrée** : le nombre de publications (revue et conférence) est en moyenne de 36.6 pour l'équipe DMD et 32.2 pour l'équipe SID, avec un nombre moyen de publications par membre et par année de 2.4 pour DMD et 2.7 pour SID[2].

Concernant les **doctorants**, si on ne comptabilise que ceux ayant soutenus à partir de 2021 (pour avoir à minima 2 années de présence pendant la période évaluée), le nombre moyen de publications est de 1 article de revue et 4 communications dans des conférences. A noter que les doctorants publient systématiquement avec leur directeur de thèse. Les doctorants sont ainsi co-auteurs de 30% des publications dans des revues et de 53% des publications dans des conférences.

Enfin, la figure 11 présente les disciplines de nos publications HAL. Comme on peut le voir, outre nos publications en informatique, mathématiques et statistique, une part significative est dans le domaine des Sciences de l'Homme et de la Société (38 publications, soit 7.5% du total), ce qui témoigne de l'activité de collaborations avec d'autres disciplines, notamment dans le cadre des Humanités Numériques.

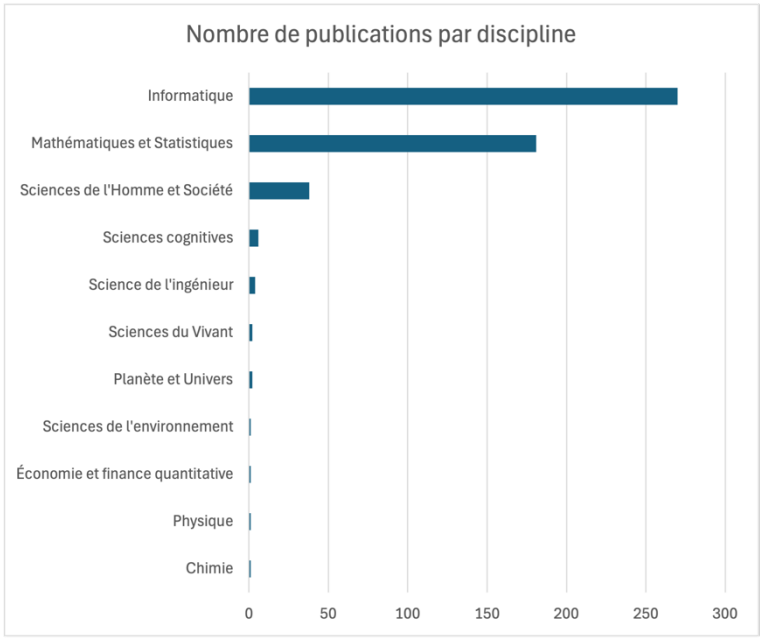

Figure 11 : Nombre de publications par discipline.

Accompagnement des chercheurs débutants

Outre les systèmes d'allégement de la charge d'enseignement pour les nouveaux MCF mis en place par nos tutelles, ainsi que les budgets recherche qui leur sont alloués, le laboratoire accompagne ses nouvelles recrues dans leur activité. Tout d'abord, en les associant systématiquement dans les réponses d'AAP portés par les chercheurs seniors. Mais également, en co-encadrant des thèses avec eux. Ensuite, nous les incitons fortement à déposer des projets internes (FIL, APPI Lyon 2, AAP Lyon 1), pour lesquels le taux de succès est très important. Également, nous les incitons à déposer des ANR JCJC, même si l'accompagnement dans ces tâches n'est que peu réalisé au niveau de l'unité, par manque de moyen et de compétence. Enfin, nous pouvons noter que les ressources budgétaires de l'unité étant en grande partie issues de projets, les membres seniors du laboratoire n'utilisent généralement pas les ressources récurrentes de nos tutelles, laissant ainsi celles-ci à disposition des chercheurs juniors.

Référence 3. L'unité participe à l'animation et au pilotage de sa communauté.

Organisation de conférences

Le laboratoire ERIC a coorganisé sur la période évaluée trois conférences nationales ou internationales majeures dans ses domaines de recherche :

- La conférence European Conference on Advances in Databases and Information Systems (ADBIS 2020) conjointement à la 24e International Conference on Theory and Practice of Digital Libraries (TPDL 2020) et aux 16e journées Business Intelligence & Big Data (EDA), qui ont eu lieu en ligne (période Covid) du

[2] Pour ces chiffres par équipe, lorsqu'une publication comporte des auteurs des deux équipes, elle est compté dans chaque équipe, et donc en double.

25 au 28 août 2020, et qui ont rassemblé près de 500 participants. Ces conférences d'audience européenne rassemblent les chercheurs dans le domaine des bases de données, des systèmes d'information, de l'informatique décisionnelle et des big data. Cet événement scientifique est cité dans le portfolio de l'équipe SID.

- Les journées de Statistique de la Société Française de Statistiques, qui ont eu lieu sur le campus de la Doua (Villeurbanne) du 13 au 17 juin 2022, et qui ont rassemblé près de 350 participants. Cette conférence, qui a lieu tous les ans dans une ville différente en France, rassemble chaque année tous les chercheurs dans le domaine de la statistique. C'est l'ensemble de la communauté statistique Lyonnaise qui s'est mobilisé pour organiser l'édition 2022, comptant notamment l'ensemble des membres statisticiens du laboratoire dans le comité d'organisation.
- La conférence Extraction et Gestion des Connaissances (EGC), qui a eu lieu sur le campus Berges du Rhône de l'Université Lumière Lyon 2, du 16 au 20 janvier 2023, et a rassemblé près de 220 participants. Cette conférence est un événement annuel réunissant des chercheurs et praticiens de disciplines relevant des sciences des données et des connaissances. Le comité d'organisation, présidé par Sabine Loudcher, était constitué d'une grande partie des membres permanents du laboratoire ainsi que de quelques collègues d'autres laboratoires du site (LIRIS, LRDE, LabHC).

Responsabilités éditoriales

Un membre du laboratoire est éditeur associé de deux revues internationales en statistique (*Journal of Classification* et *Statistical Analysis and Data Mining*). Un autre membre est éditeur de la revue *Open Journal of Databases*. Plusieurs membres ont également des activités dans des revues nationales (membre du conseil scientifique de la *Revue ouverte d'ingénierie des systèmes d'information*, secrétaire de rédaction de *Statistique et Société*), et ont également participé à l'édition de numéros spéciaux de revues internationales (*pour plus de détail voir TDCP, onglet Indice de reconnaissance*).
Notons également que plusieurs membres du laboratoire font partie du comité de pilotage de conférence nationales : EGC, European Conference on Advances in Databases and Information Systems (ADBIS), Business Intelligence & Big Data (EDA), Statlearn.

Participation à des instances de pilotage et d'expertise

Quatre membres du laboratoire ont été sur la période évaluée membres du CNU 27.
Plusieurs membres ont fait partie de commission d'évaluation pour l'HCERES, dont une présidence.
Un membre a été président d'un comité ANR lors de l'AAPG 2025.
Par ailleurs, sur la période évalue, les membres de l'unité ont participé à 7 jurys d'HDR, 89 jurys de thèse en France (dont 15 à l'international), et également à 44 comités de sélection externes.

Invitation de personnalités scientifiques

Le Collegium de Lyon, qui est l'Institut d'études avancées de l'Université de Lyon, offre la possibilité d'inviter des chercheurs étrangers renommés pour un séjour d'une année. Ces visites sont une excellente opportunité pour nouer des liens forts et initier ou confirmer des collaborations durables. De ce fait, nous collaborons toujours avec Ian Davidson, Professeur à University California Davis, qui avait fait un séjour dans ce cadre lors du précédent quinquennat.
Au cours de la période actuelle, nous avons reçu deux chercheurs dans ce cadre : Taylor Arnold (2019-2020), Professeur à University of Richmond (USA), et Brendan Murphy (2021-2022), Professeur à University College Dublin. A noter que ce dernier fût invité dans le cadre d'une collaboration avec le laboratoire COACTIS, notamment dans le cadre des travaux de recherche menés dans le cadre de la thèse de Francesco Amato.

Animation de la communauté dans le domaine des Humanités Numériques

Plusieurs membres du laboratoire ont des activités d'animation de la communauté dans le cadre des Humanités Numériques.
L'axe Sociétés et Humanités Numériques de la MSH-LSE est coanimé par un membre du laboratoire depuis sa création en 2015. Cet axe vise à mettre en dialogue, valoriser et faire émerger les initiatives sur le périmètre de la MSH dans une optique de pluridisciplinarité, d'interconnaissance et de mise en évidence de l'identité du site. Dans ce cadre, plusieurs séminaires, conférences et journées d'études ont permis d'explorer des thématiques variées, axées sur les apports disciplinaires, les méthodologies et l'actualité de la recherche.
Le pôle de spécialité HuNIS (Humanités Numériques, Individus et Sociétés connectées) de l'Université Lyon 2 est également animé depuis sa création en 2021 par un membre du laboratoire. Ce pôle a conduit plusieurs actions qui concernent à la fois les aspects recherche et enseignement de l'Université Lyon 2 : organisation de journées grand public, mise en place de modules d'enseignement transversaux et d'un cycle de conférences destiné aux doctorants.
Notons également l'animation de nombreux ateliers dans le cadre des conférences EGC, Inforsid, ou du GdR Madics, de séminaires, ainsi que de nombreuses participations à des événements de médiation sur la thématique du genre.

Référence 4. La production scientifique de l'unité respecte les principes de l'intégrité scientifique, de l'éthique et de la science ouverte. Elle est conforme aux directives applicables dans ce domaine.

L'unité veille à respecter rigoureusement les principes de l'intégrité scientifique, de l'éthique et de la science ouverte dans l'ensemble de ses productions scientifiques.
Les doctorants de l'unité suivent une formation à l'intégrité scientifique, dans le cadre des formations doctorales obligatoires. De plus, les directeurs de thèse les sensibilisent à ces enjeux, en leur fournissant des conseils et en les accompagnant tout au long de leur parcours.
L'unité fait également un effort constant pour alerter ses doctorants et collègues sur les dangers des revues dites « prédatrices ». Des listes actualisées des revues à éviter sont régulièrement diffusées afin de guider les chercheurs dans le choix des supports de diffusion.
En ce qui concerne la gestion des publications, les travaux des doctorants sont généralement cosignés avec leur directeur de thèse, en mettant systématiquement le doctorant en premier auteur, conformément à une politique qui valorise la contribution de chacun. Lorsque les supports de publication le permettent, le rôle spécifique de chaque auteur dans la réalisation de l'article est précisé de manière transparente.
Pour garantir la reproductibilité des résultats scientifiques, l'unité encourage vivement l'utilisation de Jupyter Notebook et de R Markdown pour la réalisation des expériences numériques. Ces outils permettent de rendre les travaux plus transparents et reproductibles, conformément aux exigences croissantes de nombreuses conférences et revues scientifiques. De plus, afin de garantir la traçabilité, tous les codes utilisés dans nos recherches sont systématiquement déposés sur des dépôts Git, assurant ainsi leur disponibilité et leur partage avec la communauté scientifique.
Enfin, l'unité suit une politique de science ouverte rigoureuse en déposant systématiquement ses publications et prépublications sur la plateforme HAL, et en mettant à disposition les contenus dans la mesure du possible. Cette démarche permet non seulement de garantir une large diffusion des résultats de recherche, mais aussi de soutenir l'accessibilité des connaissances scientifiques.

Domaine 3. Inscription des activités de recherche dans la société

Référence 1. L'unité se distingue par la qualité de ses interactions avec le monde culturel, économique et social

Comme nous l'avons déjà souligné, un des points forts de l'unité réside en sa capacité à monter des partenariats avec des entreprises. Ces collaborations prennent souvent la forme de thèse CIFRE (23 sur la période, pour un montant de contrats associés de 848 k€) et parfois de contrats de recherche et développement (9, pour un montant de 79 k€). Les entreprises avec qui nous travaillons sont à la fois des grands groupes (Orange, Essilor, EDF, Wordline, TOTAL, IFPEN), mais également des plus petites entreprises, souvent locales. Les secteurs d'activité sont liés à la technologie, l'énergie, la santé ou les services financiers. Nous pouvons également citer le projet WASAS avec l'entreprise Witekio, spécialisée de l'internet des objets, soutenu par BPI France.
Les sollicitations d'entreprises que nous recevons sont très nombreuses. A une grande partie d'entre elles, un temps est consacré pour échanger avec l'entreprise, souvent par un chercheur senior. Suivant la problématique de l'entreprise, il est décidé d'orienter cette dernière soit vers nos formations dans le cadre d'un stage, soit vers un contrat de collaboration quand le travail demandé est proche d'une problématique d'intérêt d'un chercheur de l'unité, soit vers une thèse CIFRE. Souvent, les collaborations débutent par un stage, ce qui permet de déchiffrer les problématiques et à chaque partie de mieux se connaitre, avant de mettre à jour une problématique de recherche. Notons également que plusieurs partenaires le sont de longue date, avec plusieurs thèses CIFRE financées (Orange, EDF, IFPEN, BIAL-X…).
Citons également notre participation chaque année aux Entretiens Jacques Cartier, qui sont des rencontres internationales visant à renforcer les partenariats entre la France et le Québec, avec un focus sur la recherche, l'innovation et le développement économique avec un ancrage au niveau sociétal.
L'unité a également un impact sur notre société à travers les docteurs qu'elle forme et qui s'y insèrent professionnellement à l'issue de leur doctorat. Nous présentons ici des statistiques sur l'activité au 31/12/2024 des 26 doctorants qui ont soutenus durant la période : 11 sont actuellement en poste en entreprise (data scientist, ingénieur machine learning, … ; dans les domaines de l'IA et de la data, de la santé, …); 6 sont actuellement en activité dans l'enseignement supérieur et la recherche (3 enseignants-chercheurs, 1 en France et 2 à l'étranger, 2 ingénieurs de recherche et 1 post-doc) ; 4 ont monté leur propre activité dans le domaine de la data (start-up, freelance…) ; 2 qui viennent de soutenir fin 2024 étaient en recherche d'emploi. Enfin, nous n'avons pas d'informations concernant les 3 restants (financement de thèse de l'étranger).

Référence 2. L'unité développe des produits et des services à destination du monde culturel, économique et social.

Une grande majorité de nos collaborations avec des entreprises débouchent sur des publications co-signées, accompagnées bien souvent de prototype logiciel. Suivant les entreprises et les enjeux de propriété

intellectuelle, ces prototypes sont diffusés librement sur un dépôt Git (par exemple dans le cadre des collaborations avec Orange) ou non. Si les discussions de propriété intellectuelle étaient relativement légères il y a quelques années, et largement en faveur des entreprises, nos tutelles se sont dotées de services permettant une négociation plus juste des répartitions des propriétés.
Comme cela a déjà été mentionné dans la partie production scientifique, le premier brevet du laboratoire a été déposé avec une entreprise partenaire, Arpège Master K, en collaboration avec le laboratoire. Avec cette PME locale, spécialisée dans le pesage industriel, nous avons développé un système de détection d'anomalies dans leur système de pesées numériques. Ce brevet a couronné une collaboration de longue date, avec un premier contrat de collaboration en 2017 puis une thèse CIFRE de 2019 à 2023.

Référence 3. L'unité partage ses connaissances avec le grand public et intervient dans des débats de société.

L'unité, en rapport avec sa taille, est très présente dans le domaine de la médiation scientifique et du partage des connaissances avec le grand public.
Chaque année, plusieurs membres de l'unité participent à la Fête de la Science en y animant des ateliers. Une participation très régulière à la nuit européenne des chercheurs est également à noter. Également, nous animons régulièrement des ateliers à la Maison des Mathématiques et de l'Informatique située à Lyon.
Plusieurs membres participent également régulièrement aux Saventuriers, à MATh.en.Jeans, aux actions de l'association Femmes et Sciences et donnent des conférences dans les lycées. Des publications régulières dans la revue Tangente, destinée aux lycéens, contribuent également à vulgariser et à faire connaître nos activités de recherche.
Un membre de l'unité est l'organisateur des Cafés de la Statistique, une manifestation grand public organisant des débats sur des thématiques d'actualité et de société. Avec environ un événement tous les deux mois, le cinquantième événement vient tout juste d'avoir lieu début 2025.
Un autre membre a participé à quatre éditions des Entretiens Jacques Cartier en étroite collaboration avec la faculté des Arts de l'Université d'Ottawa. Ces entretiens rassemblent alternativement à Lyon ou au Canada francophone des académiques, des experts du secteur privé, associatif ou culturel pour créer ensemble de nouvelles opportunités et contribuer à l'évolution de la société.
Le laboratoire a également organisé une conférence intitulée « 50 ans de traitement des données avec et pour les SHS » dans le cadre des 50 ans de l'Université Lyon 2 en 2023.
Nous pouvons enfin citer les travaux d'un membre du laboratoire, qui publie un nombre très important de cours en ligne, tutoriels, vidéo, sur le machine learning et l'apprentissage statistique. Ces publications sont très reconnues et utilisées par un grand nombre d'étudiants mais également de chercheurs d'autres disciplines, et constituent des références nationales en la matière. C'est également une vitrine importante pour notre laboratoire.

3- 2 Autoévaluation des équipes

3- 2- 1 Autoévaluation de l'équipe DMD

Domaine 1. Objectifs scientifiques, organisation et ressources de l'unité

Référence 1. L'unité s'est assigné des objectifs scientifiques pertinents et elle s'organise en conséquence.

L'équipe DMD est une équipe composée d'enseignants-chercheurs dans les disciplines de l'informatique (CNU 27) et des mathématiques appliquées (CNU 26), plus précisément en s. Historiquement concentrée sur l'étude de données complexes, l'équipe DMD a peu à peu orienté ses thématiques de recherches autour du machine learning, un domaine omniprésent dans notre société actuelle et dont l'intérêt et les applications ne sont plus à démontrer tant pour les activités de recherche que pour la société.
L'étude des données complexes et hétérogènes n'est cependant pas abandonnée par l'équipe mais elle devient maintenant un contexte de travail pour le développement de nouvelles méthodes d'apprentissage. Son implantation au sein d'une université de ALLSHS et ses collaborations avec ses partenaires amènent les différents membres de l'équipe à poursuivre leurs travaux sur ce type de données. Sur l'aspect apprentissage, l'équipe a développé une expertise fortement reconnue sur le site lyonnais, mais aussi à l'échelle nationale ou internationale sur des thématiques centrées autour de l'analyse et du développement de modèles pour l'étude de données textuelles, plus précisément dans la représentation de données textuelles, mais aussi dans le développement de modèles novateurs en statistique et apprentissage automatique. Elle choisit de présenter ses travaux selon ces deux axes. Les compétences de l'équipe sur ces thématiques sont également reconnues dans le domaine industriel, en témoignent les nombreuses collaborations dans le cadre de thèses CIFRE.

L'équipe DMD s'était fixée comme objectif de développer des outils plus fondamentaux. Cela a en partie été favorisé par l'intégration de la FIL, qui a permis d'initier de nouvelles collaborations avec d'autres laboratoires de recherche lyonnais dans le cadre de projets locaux ou nationaux (ANR. L'équipe a également renforcé ses collaborations en interne, notamment à travers l'implication systématiques de plusieurs membres dans les grands projets de recherche, ou encore le co-encadrement de thèses où la complémentarité des expertises des différents membres est exploitée.

Domaine 2. Les résultats, le rayonnement et l'attractivité scientifiques de l'unité

Référence 1. L'unité est reconnue pour ses réalisations scientifiques qui satisfont à des critères de qualité.

Au cours de ces dernières années, l'équipe DMD s'est tournée vers l'étude des problématiques majeures tant dans la recherche en apprentissage que par ses applications par la société. C'est donc tout naturellement que les recherches de l'équipe se concentrent autour de l'apprentissage de représentation de l'information sous sa forme la plus générale (par réseaux de neurones ou encore à l'aide de signatures) pour gérer la complexité des données avec lesquelles elle travaille. Le développement de modèles de mélanges permet ensuite à l'équipe de traiter l'hétérogénéité des données. Elle développe également des modèles de prévision (que cela soit pour les systèmes de vote ou sur des séries temporelles) et s'attache à mesurer l'incertitude ou les capacités prédictives de ces modèles sur de nouvelles données. L'équipe a également commencé à mobiliser de nouvelles facettes de l'apprentissage automatique comme l'apprentissage par transfert ou par renforcement afin d'élargir son champ de compétences. Sur des aspects plus sociétaux et dans la mouvance actuelle concernant les aspects éthiques et environnementaux, elle s'intéresse à l'étude et au développement de méthodes de compressions pour les grands modèles de langues (LLM). Elle développe ainsi une expertise sur les problématiques de fairness en IA tant sur le plan théorique que pratique.

Que cela soit travers l'originalité de ses contributions, son implication dans l'organisation de conférences nationales comme CAp 2021, JdS 2022 et EGC 2024 (cf section 3.1), ou ses collaborations internationales avec Ian Davidson (*University California Davis)* sur l'explicabilité des modèles d'IA (XAI), de Brendan Murphy *(University College Dublin)* sur les modèles de mélanges probabilistes, l'équipe DMD est présente et impliquée à différentes échelles de la recherche. L'ensemble de ces éléments font que, les membres de l'équipe DMD sont présents tant sur la scène locale, nationale, qu'internationale. Elle participe et partage le fruit de ses recherches dans les évènements organisés par la FIL et monte des projets financés par la fédération (à raison de deux par an en moyenne) pour tisser des liens avec les partenaires locaux (comme le *LabHC* ou le *LIRIS*) et construire des projets

de plus grande envergure (thèse ou projets ANR) ou encore des partenariats avec des médecins (projet SMAD-CC via ShapeMed avec le *Centre Léon Bérard*).

Le bilan se concentre sur les thématiques représentatives de l'équipe. Les travaux sont présentés selon ses deux grands axes de recherche : *(i)* apprentissage automatique, apprentissage statistique et optimisation ; *(ii)* représentation des connaissances et recherche d'information sur des corpus textuels.

Apprentissage automatique, apprentissage statistique et optimisation

Ce premier axe se décompose en deux fondamentales avec leurs applications sur des données complexes : l'apprentissage de modèles statistiques et machines et l'optimisation.

Modélisation et apprentissage statistique et machine
L'équipe a grandement contribué dans le domaine de l'apprentissage statistique/machine dans sa plus grande variété. Elle a notamment proposé de nouveaux modèles d'apprentissage statistique pour des données fonctionnelles ou ordinales sur des tâches supervisée (régression) et non supervisée (clustering, détection d'anomalies voire de fraudes) dans le cadre de plusieurs thèses. Concernant les données fonctionnelles (lorsque les observations sont des courbes), plusieurs modèles probabilistes de clustering et de co-clustering (clustering simultané des lignes et des colonnes d'une matrice) pour données fonctionnelles ont été développés dans le cadre de la thèse d'Amandine Schmutz. La thèse de Martial Amovin a quant à elle permis de définir un algorithme qui réalise à la fois clustering et détection d'outliers. Ce travail a abouti sur le dépôt d'un brevet, présenté dans le portfolio de l'équipe DMD. Enfin, la thèse de Jean Steve Tamo Tchomgui a permis de développer plusieurs modèles de régression linéaire avec entrées et sortie fonctionnelles. L'originalité des travaux de l'équipe réside également dans la prise en compte du caractère hétérogène des données avec le développement de modèles de mélanges probabilistes Notamment, les thèses de Margot Selosse et de Francesco Amato (soutenance prévue en 2025) ont permis de développer plusieurs algorithmes de clustering de données mixtes longitudinales, données que l'on rencontre souvent lors de l'analyse de questionnaires utilisés en marketing ou en psychologie. Dans ces travaux, l'utilisation de modèles probabilistes facilite également, du fait de leur formalisme mathématique, certaines tâches comme la sélection de modèles. Les contributions liées à cette thématique trouvent des applications dans le domaine médical [E. Peyraud et al., 2024] avec le développement de modèles de mélanges de Cox pénalisés ou dans les études marketing [F.Amato et al., 2024]. Une autre direction pour le clustering a également été explorée en tenant compte de la complexité algorithmique du problème, au moyen de techniques du type Programmation semi-définie [S. Chrétien et al. 2021], où le rang de la matrice solution correspond au nombre de clusters.

Si les modèles développés font preuve de grandes performances et un d'un grand intérêt du champ disciplinaire de recherche de l'équipe, ces derniers, comme une très grande majorité des modèles d'apprentissage présentent une incertitude qu'il est nécessaire de quantifier pour s'assurer des performances sur de nouvelles données. En ce sens l'équipe a développé des outils théoriques pour l'analyse de stabilité des modèles dans un contexte de fairness, *i.e.* pour garantir l'équité des prises de décisions indépendamment d'attributs sensibles [H. Atbir et al., 2024]. Ces travaux reposent sur le développement d'une borne en généralisation PAC-Bayésienne qui permet de garantir le bon comportement d'une mesure utilisée en économétrie, la *CVaR*, cette dernière pouvant être vue comme une mesure de fairness. Cette même étude de la stabilité a également pu être effectuée dans des modèles de prévisions sur des séries temporelles. Outre la quantification d'incertitude dans les modèles prévisions de séries temporelles, l'équipe a exploré, depuis 2022, une nouvelle méthode d'extraction de features dans les données temporelles multidimensionnelle à l'aide de la théorie des Signatures de Terry Lyons. La thèse CIFRE de Rémi Vaucher soutenue en janvier 2025 développe des algorithmes exploitant les signatures pour détecter des anomalies et construire des objets topologiques de type complexes simpliciaux dont certaines des caractéristiques algébriques ont démontré leur efficacité en médecine et en sciences cognitives (prévision de crises d'épilepsies, modélisation des signaux complexes dans le cerveau, etc …) [Vaucher et al., 2023]. L'équipe a également développé un outil permettant de quantifier l'incertitude sous la forme d'un agrégateur linéaire imprécis, appelé *Macsum*. Cet agrégateur est basé sur deux intégrales de Choquet paramétriques facilitant ainsi l'apprentissage de cet agrégateur [Hmidy et al., 2023]. Cet opérateur est utilisé chez Renault pour modéliser l'incertitude de la prédiction des prix de leurs futures voitures électriques.

Une autre facette de l'apprentissage explorée par l'équipe concerne l'apprentissage par transfert et les applications sur les données complexes. L'originalité des contributions de l'équipe sur cette thématique réside sur la nature des modèles pour lesquels le transfert est employé, mais aussi sur la méthodologie employée pour faire de l'adaptation de domaine. Par exemple, des modèles de mélanges sont utilisés pour estimer les distributions des données sources et cibles et optimiser l'appairage entre ces deux distributions [T. Martinet et al., 2023]. Nous pouvons citer également dans ce domaine les thèses de Loïc Iapteff et Youba Abed (toujours en cours), financées par IFPEN, qui proposent des approches bayésiennes pour le transfert de processus gaussiens et de modèle d'équations différentielles. Des applications de ces méthodes par transfert sont notamment mises en place dans des contextes industriels sur des données fonctionnelles ou sur des données

textuelles, *i.e.*, sur des données complexes. Cette application aux données complexes se retrouve également dans le développement de modèles temporels et dynamiques pour l'analyse de l'évolution des embeddings de documents et d'auteurs. En particulier, des modèles d'apprentissage de représentation temporelles ont été proposés pour étudier l'évolution des auteurs au cours du temps (modèle temporel gaussien [A. Gourru et al., 2022], trajectoire d'auteurs avec des ponts browniens [E. Terreau et al., 2024]) mais aussi de clustering (MMSBM dynamique [G. Poux-Médard et al., 2023]) basé sur un nouveau Powered Dirichlet-Hawkes process [G. Poux-Médard et al., 2023]). Elle a également développé des méthodes de détection de nouveauté dans les flux textuels [C. Christophe et al., 2021] ou pour étudier la dynamique des espaces de plongement [C. Christophe et al., 2021].

Optimisation, apprentissage hybride
L'équipe s'intéresse à la combinaison de l'optimisation combinatoire et du machine learning. C'est un domaine de recherche qui vise à exploiter l'apprentissage automatique pour améliorer la résolution de problèmes NP-difficiles. Dans le cadre d'une thèse CIFRE l'équipe a par exemple développé des modèles basés sur les Transformers avec des mécanismes d'attention spécialisés pour optimiser la planification dynamique de tournées de véhicules. L'hybridation peut également se faire entre des modèles mathématiques issus de la physique et ceux issus du machine learning. Des travaux de thèses CIFRE sont en cours sur ces sujets pour, par exemple, incorporer des modèles issus de systèmes énergétiques dans des modèles de machine learning afin de minimiser la consommation d'énergie dans des bâtiments. Une autre thèse CIFRE (Ugo Zennaro, Verso-Optim) étudie les problèmes d'optimisation de tournées proposés par l'entreprise et les méthodes de relaxation convexe pour les résoudre efficacement, avec une attention particulière sur l'impact des incertitudes sur les données servant à la modélisation. En complément, des travaux préliminaires récents de l'équipe [S. Chrétien et al. 2020] se sont aussi portés sur de nouvelles techniques de choix d'hyper-paramètres avec succès dans un cadre de relaxation de sparsité.

Représentation des connaissances et recherche d'information sur des corpus textuels

La représentation des données textuelles dans les modèles d'apprentissage et leur étude via les modèles de langues constituent un deuxième axe de recherche fondamental des membres de l'équipe. Ces travaux se rapprochent du traitement automatique du langage (TAL). Les travaux entrepris dans cet axe ont permis d'établir plusieurs collaborations avec la communauté ALLSHS propre à l'environnement de travail de l'équipe. Les détails de ces collaborations et les recherches associées seront décrits dans la partie 3.1 (Domaine 2, Référence 1) du présent document.

Représentation des documents et auteurs
Avoir une bonne représentation des données est essentielle au bon fonctionnement des algorithmes d'apprentissage. En ce sens, l'équipe a consacré une partie importante de son travail de recherche au développement de méthodes de représentation de documents, d'auteurs ou encore de graphes de réseaux de documents. Cette thématique de l'apprentissage a été investiguée à travers de nombreuses thèses (5) au cours du présent contrat et ont donné lieu à des publications dans de très grandes conférences comme WWW, IJCAI, ICDM ou encore ECIR. Plusieurs modèles ont développés dans le contexte de ces thèses académiques et industrielles, comme la méthode *GVNR-t* [R. Brochier et al., 2019] basée sur sur la factorisation de matrice, une méthode *IDNE* basée sur un réseau de neurones [R. Brochier et al., 2020], et aussi *RLE* pour l'apprentissage de représentations de documents [A. Gourru et al., 2020] ou encore via la méthode *GELD* pour de l'apprentissage de représentation de documents dans un cadre probabiliste [A. Gourru et al., 2020].
Une autre piste explorée par l'équipe consiste à modéliser la donnée textuelle sous la forme de graphes pour des résolutions de tâche de classification en TAL. En ce sens elle a développé une méthode de classification de documents par un réseaux de neurones graphes (GNN) hiérarchique dans l'espace hyperbolique [A. Guille et al., 2023]. La combinaison des approches GNN avec les réseaux de neurones récurrents a également permis une technique permettant de résumer des documents [R. Said et al., 2024], Enfin, au cours de la thèse de Frédéric Charpentier, une méthode originale classification de documents par l'intermédiaire des graphes sémantiques abstraits a été développée [F. Charpentier et al., 2024].

Modélisation des interactions
Cet thématique regroupe plusieurs travaux issus de la thèse de Gaël Poux-Médard (2022) sur la manière de modéliser/capturer l'interaction dans les réseaux d'information. Plusieurs modèles ont été proposés, notamment dérivés des *Stochastic Block Models* (SBM) comme le modèle *Interacting Mixed Membership Stochastic Block Model* [G. Poux-Médard et al., 2022] et un nouvel a priori dérivé des Processus de Dirichlet : Powered Dirichlet Process [G. Poux-Médard et al., 2021].

Synthèse de l'auto-évaluation de l'équipe DMD

L'équipe DMD se concentre sur l'apprentissage automatique et la représentation de l'information, en développant des modèles pour gérer la complexité et l'hétérogénéité des données. Elle développe de nouvelles approches pour l'apprentissage de représentation et/ou pour données. Elle s'intéresse aux aspects

éthiques et environnementaux, notamment la compression des grands modèles de langues et la fairness en IA. L'équipe est active à l'échelle locale, nationale et internationale, organisant des conférences et collaborant avec des universités étrangères.

Les recherches de l'équipe DMD sont structurées autour de deux axes principaux : (i) apprentissage automatique, apprentissage statistique et optimisation ; *(ii)* représentation des connaissances et recherche d'information sur des corpus textuels.

Dans le premier axe, l'équipe a développé des modèles statistiques et d'apprentissage automatique pour des données fonctionnelles et ordinales, ainsi que des modèles de mélanges probabilistes pour traiter l'hétérogénéité des données. Elle a également travaillé sur des modèles de prévision et des outils pour quantifier l'incertitude et la stabilité des modèles. L'équipe a également exploré l'apprentissage par transfert et l'optimisation combinatoire, en développant des algorithmes pour des applications industrielles et médicales. Elle a également proposé des méthodes pour la détection d'anomalies et la modélisation des interactions dans les réseaux d'information.

Dans le second axe, l'équipe DMD a développé des méthodes de représentation de documents, d'auteurs et de graphes de réseaux de documents, en utilisant des techniques comme la factorisation de matrice et les réseaux de neurones. Elle a également modélisé les données textuelles sous forme de graphes pour des tâches de classification en traitement automatique du langage.

L'équipe est impliquée dans des projets ANR, des projets PIA, ainsi que de nombreux projets et collabore avec des partenaires du site comme le LabHC et le LIRIS, ainsi qu'avec des médecins pour des projets de recherche en santé. Elle participe activement à des événements scientifiques et publie régulièrement dans des conférences et revues internationales de haut niveau.

Trajectoire de l'équipe DMD

Comme le montre le bilan, l'équipe DMD a pleinement réussi à poursuivre son développement sur les thématiques de l'apprentissage, de prévision et d'aide à la décision, pour des données complexes. Elle laisse ainsi de côté son activité historique qu'est le Data Mining. De plus, sa stratégie de recrutement a permis de développer des techniques d'apprentissage efficaces, se traduisant par le développement et l'étude d'algorithmes d'optimisation, mais aussi par l'analyse théorique des algorithmes développés, comme l'étude de garanties en généralisation.

Néanmoins, l'équipe n'a pu mener à bien toutes ses ambitions. Elle s'était également fixée comme objectif d'accroître ses collaborations avec le domaine des ALLSHS afin de répondre à leurs problématiques concrètes via le développement et le transfert de modèles interprétables. Si cet objectif n'est que partiellement traité dans le cadre du projet LIFRANUM, il faut tout de même noter qu'un poste d'ingénieur de recherche devait normalement être créé pour le laboratoire afin de nous accompagner dans cet objectif. Ce dernier n'a, à ce jour, pas encore été créé.

Enfin, l'équipe ambitionnait également d'améliorer les synergies entre les thématiques d'apprentissage et d'aide à la décision basée sur des approches multi-critères. Si ces thématiques restent encore une thématique de l'équipe, il est à noter que les contributions ou les collaborations entre les chercheurs de ces deux thématiques n'ont conduit à aucune publication. Ce constat est cependant à nuancer avec le recrutement de nouveaux profils au sein de l'équipe sur ces dernières années qui ont permis le développement de nouveaux axes de recherches plus centraux autour de l'apprentissage.

Via le renforcement de sa recherche autour de l'apprentissage, l'équipe DMD a pu continuer d'accroître son nombre de partenaires industriels et le nombre projets dans lesquels elle est impliquée en travaillant sur des thématiques plus en phase avec les sujets actuels pour les industries. Elle souhaite maintenant faire de même en se concentrant sur les thématiques majeures de la communauté en apprentissage mais aussi sociétales en développant de nouveaux projets de recherches mais aussi de nouvelles thèses.

Comme lors de la période passée, l'équipe souhaite continuer de travailler sur les différentes phases de l'élaboration des modèles, en travaillant à la fois sur les données peu importe la nature de ces dernières, mais aussi sur de nouvelles techniques d'apprentissage qui sont développées. Elle souhaite enfin y inclure un troisième axe qui lui semble être un critère déterminant pour des publications de qualité mais aussi pour répondre à des problématiques sociétales : étudier la validité des modèles. Pour refléter ce changement de thématique, l'équipe souhaite également changer de nom pour devenir ***Learning and Decision (LeaD)***.

De nouvelles thématiques en machine learning

L'équipe DMD, après avoir accentué sa transition vers la thématique de l'apprentissage statistique et de l'apprentissage automatique, s'engage désormais à pleinement consacrer ses recherches dans des domaines

qui répondent aux enjeux actuels tant industriels que sociétaux. En se concentrant sur des approches alternatives aux architectures profondes, l'équipe explore des méthodes d'apprentissage automatique plus légères et efficaces, particulièrement adaptées à des contextes spécifiques tels que le traitement des séries temporelles. Ce travail s'inscrit dans une logique de frugalité des modèles visant à proposer des solutions plus adaptées et performantes sans recourir systématiquement à des techniques computationnellement coûteuses, réduisant ainsi leur impact environnemental. De tels travaux sont déjà initiés à travers le projet ANR DIKé visant à développer de nouvelles méthodes de compression non biaisées. L'équipe s'efforce ainsi d'élargir les possibilités de l'apprentissage automatique dans des domaines industriels où ces nouvelles approches peuvent avoir un impact concret et immédiat.

Elle souhaite également renforcer l'axe apprentissage sur graphe, notamment appliqué au traitement automatique des langues, où les relations complexes entre les éléments doivent être capturées pour offrir des modèles plus performants. L'équipe souhaite également s'ouvrir au domaine du méta-apprentissage, cherchant à développer des modèles capables d'apprendre à s'adapter à de nouvelles tâches de manière plus efficace. Cet axe est particulièrement pertinent dans des environnements nécessitant une autonomie des modèles, un domaine qui s'étend vers des applications d'apprentissage par renforcement. En outre, l'équipe se penchera sur des défis propres aux données hétérogènes, en développant des approches adaptées aux systèmes multi-agents ou encore à l'apprentissage fédéré, et à l'optimisation des ressources dans des contextes sociétaux où la confidentialité des données doit être préservée.

Étude théorique des modèles

L'équipe DMD s'attache également à une étude théorique approfondie des algorithmes, afin de garantir la robustesse et la validité des modèles mais aussi pour quantifier l'incertitude de ces derniers dans des environnements complexes. Les recherches théoriques seront orientées vers la validité des algorithmes, avec un intérêt particulier pour les bornes de généralisation et l'analyse de la convergence des méthodes développées. La théorie PAC-Bayésienne et la prédiction conforme seront de nouveaux outils sur lesquels l'équipe souhaite développer une expertise et qui seront mobilisés pour fournir un cadre rigoureux garantissant les performances des modèles. Cela permettra d'assurer que ces derniers peuvent être appliqués de manière fiable à des données réelles, tout en restant capables de gérer les incertitudes inhérentes à ces dernières. Un projet ANR a d'ailleurs été déposé en collaboration avec l'INRIA de Rennes et le Laboratoire Hubert Curien.

En complément, l'équipe mettra un accent particulier sur l'optimisation des algorithmes afin de garantir leur stabilité dans des environnements fluctuants et incertains, tout en intégrant des contraintes sociétales liées à la fairness et à l'équité des décisions algorithmiques. Le travail sur le transfert de connaissances (transfer learning) permet aux modèles de s'adapter à de nouveaux contextes tout en restant robustes. L'objectif est de permettre à l'apprentissage automatique de s'ajuster aux évolutions et aux particularités des données traitées, tout en maintenant des garanties théoriques solides sur la performance des modèles.

Diversité et pluralité des données

Depuis toujours, l'équipe DMD a travaillé sur une large gamme de données variées, allant des séries temporelles aux données relationnelles complexes, en passant par des ensembles hétérogènes issus de systèmes multi-agents ou de réseaux distribués. Cette diversité des données a constitué un axe central de ses recherches et continue de l'être, car l'équipe souhaite appliquer les nouvelles méthodes d'apprentissage qu'elle développe à ce large éventail de données. En particulier, les graphes et les données issues du TAL restent au cœur de ses projets, avec l'objectif de mieux comprendre et modéliser les relations complexes entre les éléments de ces ensembles.

L'équipe entend ainsi renforcer son approche sur des données de plus en plus hétérogènes, en mettant en œuvre des techniques novatrices telles que l'apprentissage fédéré pour traiter des données distribuées de manière sécurisée, tout en préservant la confidentialité. Cette démarche s'inscrit dans une volonté d'adopter des solutions robustes et équitables, particulièrement dans un contexte où les biais algorithmiques et les enjeux d'équité deviennent essentiels. De plus, l'équipe souhaite appliquer ses recherches en optimisation combinatoire à des environnements incertains, où la qualité des données peut varier, tout en maintenant la performance des modèles malgré les imprécisions inhérentes.
Dans cette optique, la variété des données, qu'elles soient structurées, relationnelles ou distribuées, demeure une priorité stratégique, car elle constitue le terrain d'application naturel pour les nouvelles méthodes développées. Ainsi, l'équipe DMD s'engage à poursuivre son travail sur des données complexes et diversifiées, tout en veillant à ce que les solutions proposées soient adaptées à la diversité des contextes industriels et sociétaux.

3- 2- 2 Autoévaluation de l'équipe SID

Domaine 1. Objectifs scientifiques, organisation et ressources de l'unité

Référence 1. L'unité s'est assigné des objectifs scientifiques pertinents et elle s'organise en conséquence.

L'équipe SID a été créée en 2001 au laboratoire ERIC dans la mouvance de l'informatique décisionnelle (Business Intelligence - BI), des entrepôts de données (Data Warehouses - DW) et de l'analyse en ligne OLAP. Depuis plus de 20 ans, elle est reconnue en France comme une des équipes leader dans la communauté des systèmes d'information décisionnels. L'informatique décisionnelle est un domaine de recherche qui a connu de profondes mutations ces dernières années, dues notamment à la facilité d'accès aux outils décisionnels pour tous les usagers, ainsi qu'à l'avènement des données massives ((Big Data) combiné aux innovations technologiques. Ces mutations l'ont amené depuis le précédent contrat à étendre ses travaux à la gestion et à l'analyse des mégadonnées qu'elles soient massives ou hétérogènes. L'équipe SID s'était donnée en 2020 comme objectifs de concevoir une nouvelle génération d'architectures de gestion de données, architectures centralisées et/ou distribuées incluant un processus d'intégration intelligent des données massives, d'analyses avancées tout en tenant compte de la variété des données, et de la sécurité du processus (protection des données et des accès). Elle a alors décidé d'organiser ses travaux selon les deux axes de recherche : big data management et BI & analytics.

L'équipe SID a également voulu conforter son positionnement leader dans le domaine de la BI en France et asseoir sa visibilité à l'international. En 2020, elle avait annoncé vouloir organiser une ou deux conférences nationales ou internationales reconnues dans sa communauté scientifique, répondre à des appels à projet nationaux avec des équipes françaises et développer ses collaborations internationales. Par ailleurs, elle voulait poursuivre sa politique de collaboration scientifique avec des laboratoires de ALLSHS, notamment via des projets en humanités numériques.

Domaine 2. Les résultats, le rayonnement et l'attractivité scientifiques de l'unité

Référence 1. L'unité est reconnue pour ses réalisations scientifiques qui satisfont à des critères de qualité.

Pour pouvoir proposer des solutions innovantes à la gestion et à l'analyse des données massives, une particularité des travaux de l'équipe SID est de combiner des concepts et technologies issus du domaine des bases de données, de l'informatique décisionnelle, du Big Data tant au niveau théorique (modèles conceptuels, logiques et physiques de données) que système (techniques de stockage, indexation et optimisation de requêtes). Elle mobilise également d'autres domaines de recherche tels que l'apprentissage automatique, la recherche d'information, le cloud computing pour proposer des solutions qui répondent aux enjeux de la gestion des données massives.

L'équipe a également beaucoup investi les questions liées à la sécurité des données (dans le cloud, dans l'Internet des Objets, dans leur cycle de vie) pour la détection dynamique et en temps réel des altérations des données afin de garantir leur véracité et leur intégrité.

Les données gérées par l'équipe SID sont produites par des sources diverses (sources de données classiques, données issues de réseaux sociaux, données de Sciences Humaines ou Sociales, données de capteurs, données numériques, textuelles ou multimédia) et les applications s'appuient sur des modèles et des supports de stockage variés (SGBD relationnels, SGBD NoSQL, lacs de données, entrepôts de données agiles).

Tous ces éléments permettent à l'équipe SID d'avoir un positionnement original, clair et assumé d'abord en interne puis dans le paysage lyonnais, et enfin au niveau national voire international. L'équipe SID affiche une forte attractivité internationale comme en témoignent ses multiples collaborations et ses nombreux doctorants étrangers. Parallèlement, et comme en témoigne la liste des projets et des contrats dans le fichier tableau_donnees_caracterisation_production.xlsx, elle est impliquée de manière très forte dans des projets en humanités numériques qui s'inscrivent dans le projet global du laboratoire ERIC. Dans ce contexte, plusieurs coopérations se sont développées entre les deux équipes du laboratoire et avec d'autres laboratoires autour de projets de recherche et de co-encadrements de thèse.

Et enfin, l'équipe SID a pris en charge l'organisation de deux événements scientifiques majeurs dans sa communauté : en 2020, les conférences conjointes ADBIS-TPLD-EDA et en 2023 la conférence EGC. Ces deux évènements sont déjà décrits à la page 15 du DAE ou dans le portfolio de l'équipe SID ; nous les rappelons ici car dans les deux cas, un membre de l'équipe SID a présidé le comité d'organisation.

Les travaux de l'équipe sont présentés ci-dessous selon les deux axes de recherche : (1) big data management et (2) BI & analytics. Seuls les principaux travaux et les publications les plus significatives sont cités dans ce bilan.

Big Data Management

Les réalisations scientifiques dans l'axe Big Data Management se déclinent en deux thèmes : nouvelles architectures décisionnelles et sécurité des données.

Nouvelles architectures décisionnelles

Dans le cadre du thème "nouvelles architectures décisionnelles", nous avons étudié les modèles NoSQL, en particulier les modèles orientés colonnes et les modèles orientés graphes, afin d'optimiser le stockage et l'interrogation des données hétérogènes. Nous avons également exploré le concept de lac de données, qui permet de centraliser de grands volumes de données brutes et hétérogènes tout en préservant leur souplesse d'exploitation. Enfin, nous avons analysé différentes approches de BI pour rendre l'analyse des données plus accessible et favoriser une prise de décision éclairée à tous les niveaux de l'organisation.

Avec l'essor des données massives, la gestion des entrepôts de données pose des défis majeurs en termes de passage à l'échelle, d'optimisation des performances et d'évolution des modèles analytiques. Les bases NoSQL orientées colonnes apparaissent comme une solution efficace pour le stockage et l'analyse de données massives. Toutefois, l'exploitation optimale de ces technologies nécessite des stratégies avancées de gestion des données. Dans la thèse de Mohamed Boussahoua (2020), nous avons ainsi exploré l'optimisation de ces entrepôts en travaillant sur le placement et la distribution des données dans un environnement distribué. L'objectif était d'améliorer les temps de réponse aux requêtes analytiques complexes, un enjeu crucial pour tirer pleinement parti des capacités des entrepôts NoSQL.

Cependant, au-delà des questions de performance, l'évolution rapide des sources et des besoins en analyse pose un défi supplémentaire aux entrepôts de données traditionnels, souvent rigides et peu adaptatifs. Les modèles multidimensionnels classiques montrent leurs limites face à la nécessité d'intégrer de nouvelles sources et de s'adapter aux évolutions des usages. C'est dans cette optique que s'inscrit la thèse de Redha Benhissen (2023), qui propose une approche agile pour l'évolution des schémas des entrepôts de données. Notre solution repose sur un modèle de schéma évolutif multi-versions, où les différentes instances de données sont stockées dans un entrepôt orienté graphe. Un métamodèle assure la gestion des versions du schéma, et nous avons défini des fonctions d'évolution au niveau du schéma et des instances, offrant ainsi une plus grande flexibilité d'adaptation. Ce travail a été publié en 2023 dans le workshop DOLAP de la conférence EDBT (ce dernier workshop valant pour une conférence).

Pour lever les verrous liés au stockage, à l'interrogation, à l'analyse et à la visualisation des données hétérogènes, nous travaillons également depuis quelques années avec le concept de lac de données. Un lac de données propose de stocker les données dans leur format d'origine et sans schéma prédéfini. Avec un tel principe, tous types de données peuvent cohabiter dans un lac de données, qu'elles soient structurées ou non. Toutefois, pour être exploitable, un lac de données a besoin de métadonnées qui permettent de décrire et de lier les données stockées dans le lac, ainsi qu'un système efficace de gestion de ces métadonnées. Dans les thèses de Nicolas Sawadogo (2021) et d'Etienne Scholly (2022) nous avons étendu la définition du concept de lacs de données, ainsi que les fonctionnalités qu'un système de gestion de métadonnées devrait avoir pour être complet et efficace. De plus, nous avons proposé un premier modèle de métadonnées (MEDAL), puis son évolution en un métamodèle (goldMEDAL) qui comprend trois niveaux de modélisation (conceptuel, logique et physique) et quatre concepts principaux : entité de données, groupement, lien et processus. Nous avons également conçu des bancs d'essais qui permettent d'évaluer la performance quantitative et qualitative de lacs de données hétérogènes et d'autres types d'architectures de gestion de données similaires. L'ensemble de ces travaux a été publié dans les revues Intelligent Information Systems 2021, Big Data Research 2021, Information Systems Frontiers 2021 et Data & Knowledge Engineering 2023, dans les conférences IDEAS 2024, DaWak 2021, ADBIS 2021, IDEAS 2021 et dans les workshops BBIGAP de la conférence ADBIS 2019, DOLAP 2021. Les différents projets que nous avons menés en archéologie (HyperTheseau, DataLAC), littérature (Picletters, Lifranum) et science de gestion (Aura PMI et COREL) nous ont permis de démontrer que le concept de lac de données est une bonne solution pour gérer l'hétérogénéité des données dans les projets de recherche liés aux sciences sociales ou humaines. Dans chaque projet, nous avons conçu et développé un système de gestion de métadonnées basé sur le métamodèle goldMEDAL. Cependant, chaque système est développé indépendamment de celui des autres projets et nécessite un temps de conception et d'implémentation conséquent. De plus, la création de métadonnées pour décrire et lier les entités de données présentes dans le lac reste une tâche chronophage et fastidieuse. Et enfin l'accessibilité de ces nouveaux modes d'organisation, de requêtage et d'analyse des données à des non spécialistes de l'informatique ou de la science des données, tels que nos partenaires en ALLSHS, pose des problèmes de fond. Les thèses de Ahlame Diouan et de Rajae El Idrissi, en cours, s'attèlent à faire des propositions pour ces problèmes.

Une autre facette de nos travaux porte sur l'accessibilité de l'informatique décisionnelle. L'objectif du projet ANR BI4people (2020-2024) est de rendre accessible la puissance de l'analyse interactive OLAP à la plus large

audience possible, en mettant en œuvre le processus d'entreposage de données en mode software-as-a-service, de l'intégration de données multisource, hétérogènes (typiquement sous la forme de tableaux issus de tableurs, de documents textuels ou semi-structurés, ou encore du Web) à une analyse OLAP et une visualisation très simple. Pour atteindre cet objectif, il a fallu aussi inclure la privacy by design, être autonome, extrêmement simple, ergonomique et intelligible. Dans ce contexte, les étapes classiques de l'entreposage de données s'appliquaient, mais devaient être complètement automatisées. Nous avons également comparé les systèmes de chiffrement et les protocoles de calcul sécurisé pour évaluer leur pertinence dans l'analyse collaborative de données. Nous avons développé des scénarios pour sécuriser des cas d'usage et analysé leurs performances. Enfin, il faut souligner que l'évaluation (en lien avec les sciences de l'information et de la communication) de nos prototypes vis-à-vis des usagers et usagères a été mise en œuvre tout au long du projet et non uniquement à la fin. La thèse de Yuzhao Yang (2022) et les postdocs de Fahad Muhammad et d'Olga Cherednichenko ont permis d'obtenir des résultats probants qui ont été publiés dans les conférences internationales IDEAS 2021, DEXA 2021, ADBIS 2022, DaWak 2022 et DATA 2023 et dans la revue The Journal of Supercomputing. Ce projet fait partie des éléments cités dans le portfolio de l'équipe SID.

Sécurité des données

Dans un environnement numérique en constante évolution, la sécurité des données est cruciale pour garantir leur confidentialité, intégrité et disponibilité. Les solutions actuelles peinent à répondre aux défis liés à la traçabilité, à la conformité réglementaire (RGPD, HIPAA) et à la gestion du cycle de vie des données (de la collecte à la destruction). Ces travaux abordent des solutions pour assurer une gestion sécurisée des données, avec une attention particulière portée à la traçabilité et à la conformité réglementaire lors de chaque phase du cycle de vie des données. Dans la thèse de Kenza Chaoui (2024) est proposée une méthode d'étiquetage des données en deux niveaux, s'appuyant sur des outils avancés et des régulations strictes pour assurer une protection robuste. Le premier niveau repose sur l'algorithme des k plus proches voisins (k-NN) et l'outil ExifTool. ExifTool nous permet d'extraire les métadonnées de façon fiable et automatique, et grâce à l'algorithme k-NN, nous effectuons un étiquetage initial des données en fonction de leur proximité avec des points de référence prédéfinis. Le deuxième niveau utilise une annotation automatique par segments pour l'étiquetage du contenu textuel. Ce processus est enrichi par des bibliothèques légales contenant les régulations de protection des données, notamment la HIPAA, la FERPA, la PIPEDA et le GDPR. Les travaux ont été publiés dans la conférence internationale AICCSA 2024.

Depuis quelques années, nous nous intéressons à l'Internet des objets (IoT), un domaine clé où la prolifération des objets connectés génère une immense quantité de données exploitables à des fins analytiques. Cependant, cette expansion soulève des enjeux majeurs en matière de sécurité des données, notamment face aux cyber-attaques ciblant ces réseaux. Conscients de ces défis, nous avons mené plusieurs travaux sur la sécurisation des données dans les réseaux IoT, notamment à travers l'encadrement d'une thèse et des projets CyberSecGraph et ROMANCE. L'un des enjeux majeurs réside dans la sécurisation des échanges entre objets connectés tout en optimisant la consommation des ressources (énergie, calcul, mémoire). La thèse de Sami Bettayeb (2024) qui a porté sur la gestion sécurisée des clés cryptographiques dans les réseaux IoT a débouché sur deux contributions majeures : EVKMS, un système de gestion de clés basé sur des vecteurs prédistribués publié dans la revue Ad Hoc Networks 2023 (cette publication fait partie des éléments cités dans le portfolio de l'équipe SID), et BKRSC-IoT une solution exploitant la blockchain et les contrats intelligents pour la révocation sécurisée des clés compromises présentée à la conférence MEDES 2023. Par ailleurs, dans le cadre de la thèse en cours de Floribert Katembo, nous approfondissons cette problématique en développant de nouveaux modèles de protection des données avec des approches intégrant l'apprentissage automatique, permettant d'anticiper et de contrer l'évolution des cyberattaques. Nos premiers résultats, très prometteurs, ont déjà abouti à une publication dans la conférence internationale IWCMC 2023.

Une autre dimension de nos travaux concerne l'amélioration de la sécurité à l'accès des systèmes embarqués. L'objectif du projet Wasas, soutenu par BPI France en collaboration avec l'entreprise « The Embedded Kit », est de développer une solution clé en main et accessible pour le développement, la mise en œuvre et la gestion de solutions IoT sécurisées. Cette solution permettra aux PMEs et ETIs un accès facilité aux technologies de l'IOT et de la cybersécurité, du hardware jusqu'à la plateforme cloud, afin de dynamiser la création de nouvelles solutions connectées et sécurisées. Le financement, de 3,2 millions d'euros, obtenu dans le cadre de la Stratégie nationale de cybersécurité 2030 de la France, nous a permis de recruter, pour l'instant, une postdoctorante, Meriem Zouzou.

BI & Analytics

L'Analytics et la Business Intelligence jouent un rôle clé dans l'exploitation des données massives pour la prise de décision. Dans ce cadre, nous abordons plusieurs contributions majeures, telles que l'amélioration des performances des requêtes sur Hadoop, le développement d'opérateurs OLAP adaptés aux bases NoSQL, ainsi que des applications concrètes dans la médecine de précision, la détection des rumeurs et l'analyse des interactions humaines avec l'environnement.
Dans la thèse de Yassine Ramdane (2019), nous avons abordé l'optimisation des entrepôts de données massives sur Hadoop, un enjeu majeur pour améliorer les performances des requêtes décisionnelles,

notamment face aux défis de gestion du trafic réseau lors des jointures complexes. Bien que Hadoop utilise le partitionnement horizontal pour répartir les données et accélérer l'accès, la gestion du réseau reste problématique, surtout pour les jointures en étoile qui nécessitent souvent plusieurs cycles MapReduce. Une solution courante consiste à partitionner les tables selon leurs clés de jointure, mais cela n'optimise pas toujours les performances. Pour pallier cette limitation, nous avons conçu SkipSJoin, une approche innovante qui allie un modèle basé sur les données et un autre sur la charge des requêtes pour optimiser le schéma de partitionnement. SkipSJoin permet d'exécuter les jointures en étoile en une seule étape via Spark, ce qui élimine le chargement de blocs HDFS inutiles et améliore considérablement le temps d'exécution des requêtes. Ce travail a été publié dans la conférence internationale Conceptual Modeling (ER 2019) ; cette référence fait partie des éléments cités dans le portfolio de l'équipe SID.

Les bases NoSQL, bien qu'efficaces pour le stockage et l'analyse des données massives, ne disposent pas d'opérateurs OLAP natifs, limitant ainsi les capacités d'agrégation et d'analyse décisionnelle. Les solutions existantes, comme Hive ou Kylin, adoptent une approche orientée lignes, ce qui pose des défis en termes de performances et de scalabilité. Pour pallier cette limite, nous avons développé MC-CUBE (MapReduce Columnar CUBE), un opérateur OLAP basé sur une approche purement colonne, exploitant MapReduce et la jointure invisible pour optimiser les agrégations. Son évaluation sur le benchmark Star Schema Benchmark (SSB), avec HBase et Hadoop, a démontré des améliorations significatives en efficacité et scalabilité, confirmant ainsi la pertinence des bases NoSQL pour l'analyse décisionnelle. Ce travail a été publié dans l'International Journal of Decision Support System Technology 2020.

Dans la thèse de Walid Zeghdaoui (2022), nous avons exploré l'analyse des big data en médecine de précision, un domaine où l'identification des patients résistants aux traitements anticancéreux est essentielle pour améliorer les soins. L'un des défis majeurs réside dans l'extraction et l'analyse des comptes rendus médicaux, souvent non structurés, afin de détecter des résistances aux traitements. Une meilleure identification de ces résistances permettrait de prédire les risques pendant le traitement, d'individualiser les soins et de renforcer la prévention en fonction du profil du patient. Pour relever ces défis, nous avons proposé plusieurs contributions basées sur des modèles de machine learning et des techniques de traitement du langage naturel (NLP). Ces approches permettent d'automatiser l'extraction d'informations médicales, d'identifier des facteurs de résistance et de développer des modèles prédictifs pour une prise de décision plus fine. Cette méthodologie ouvre la voie à des traitements personnalisés, mieux adaptés aux caractéristiques du cancer et aux profils des patients, contribuant ainsi à l'optimisation des stratégies thérapeutiques. Ce travail a été publié à la conférence DEXA 2021.

Le développement rapide des réseaux sociaux a favorisé l'échange d'une masse de données importante, mais aussi la propagation de fausses informations. De nombreux travaux se sont intéressés à la détection des rumeurs, basés principalement sur l'analyse du contenu textuel des messages. Cependant, le contenu visuel, notamment les images, demeure ignoré ou peu exploité. Dans la thèse d'Abderrazek Azri (2022) est traité l'aspect contrôle d'intégrité des données visuelles de type image qui sont très répandues sur les médias sociaux et dont l'exploitation s'avère être importante pour analyser les rumeurs. Nous nous sommes focalisés plus particulièrement sur les techniques adoptées pour vérifier la véracité des images en développant des méthodes permettant d'associer les contenus textuel et visuel afin d'évaluer la véracité des messages. Ces travaux ont été publiés dans la revue Information Systems Frontiers 2022 (cette publication fait partie des éléments cités dans le portfolio de l'équipe SID) et dans les conférences DaWak 2023, ECML PKDD 2021 et ADBIS 2021.

Dans le cadre d'une collaboration avec des chercheurs de l'INRAE, nous avons étudié l'impact des loisirs en plein air sur les interactions entre les humains et les systèmes écologiques à travers l'analyse des données issues des réseaux sociaux. Ces loisirs apportent des bénéfices physiques et mentaux aux individus, mais les perceptions et émotions associées aux paysages restent peu quantifiées dans la littérature scientifique. Comprendre ces perceptions permettrait pourtant d'optimiser l'aménagement des espaces naturels et d'adapter les offres récréatives aux attentes des usagers. Pour répondre à cette problématique, nous avons exploité des données textuelles issues de Wikiloc, une plateforme dédiée aux activités de plein air. En appliquant des techniques de traitement du langage naturel (NLP) et une analyse de corrélation, nous avons pu identifier les caractéristiques paysagères et les éléments influençant le ressenti des randonneurs. Nos résultats, centrés sur la région Auvergne, offrent des pistes d'amélioration pour les planificateurs de loisirs en mettant en évidence les aspects les plus valorisés par les usagers. Ce travail, qui souligne le potentiel de Wikiloc comme source de données environnementales, a été publié dans la revue Ecological Informatics 2023.

Dans le contexte de la maintenance prédictive, le projet CLEAN4SED (Apprentissage continu et évolutif pour l'Edge devices du traitement des données en temps réel / Continuous LEArning and Neuroevolution for Supporting Edge Devices and predictive maintenance) vise à proposer une solution innovante pour l'IA périphérique (Edge-AI), où le modèle d'apprentissage adaptatif & évolutif et les tâches de maintenance prédictive s'effectuent sur le dispositif périphérique (l'« Iot » ou « device »). Ainsi, la solution se situe à la croisée de quatre domaines à savoir : les données massives, l'apprentissage automatique, l'apprentissage continu

(Lifelong Learning - LL) et le Tiny Machine Learning (TinyML). Le soutien financier de la Région Auvergne Rhône-Alpes permet le financement de la thèse d'Anaïs Lavorel qui est en cours.

Synthèse de l'auto-évaluation de l'équipe SID

L'équipe SID se distingue par sa capacité à proposer des solutions innovantes pour la gestion, la protection et l'analyse des données massives, en combinant des concepts issus des bases de données, de l'informatique décisionnelle, du Big Data, de l'apprentissage automatique, et du *cloud computing*. Ses travaux sont structurés autour de deux axes : *Big Data Management* et *BI & Analytics*.

Dans le domaine du *Big Data Management*, l'équipe a développé de nouvelles architectures adaptées aux données massives et à leurs caractéristiques. Elle a travaillé sur des modèles et des supports de stockage variés (SGBD relationnels, SGBD NoSQL, lacs de données, entrepôts de données agiles) ainsi que sur l'optimisation des entrepôts de données pour améliorer les performances des requêtes analytiques. Elle s'est également intéressée à la sécurité des données massives notamment dans le *cloud* et l'Internet des Objets, pour garantir leur intégrité et véracité en temps réel. Les projets incluent des applications en archéologie, littérature, et science de gestion, avec un focus sur l'accessibilité des systèmes pour les non-spécialistes.

En *BI & Analytics*, l'équipe a contribué à l'optimisation des requêtes sur Hadoop, au développement d'opérateurs OLAP pour les bases NoSQL, et à des applications concrètes dans la médecine de précision, la détection des rumeurs et l'analyse des interactions humaines avec l'environnement. Les travaux incluent des collaborations avec l'INRAE et des projets financés par la Région Auvergne Rhône-Alpes.

L'équipe a une bonne attractivité internationale, avec de nombreuses collaborations et doctorants étrangers, Elle a organisé des événements scientifiques majeurs dans sa communauté.

Trajectoire de l'équipe SID

Le bilan des travaux menés par les membres de l'équipe montre que SID a réussi à atteindre globalement tous les objectifs qu'elle s'était fixés en 2020, notamment elle a su faire évoluer ses thématiques de recherche (initialement en BI) vers la gestion, la sécurité et l'analyse des données massives tout en tenant compte des défis scientifiques et sociétaux. Elle compte poursuivre dans cette direction scientifique. De plus, elle veut développer son réseau national et international pour accroître sa visibilité, son attractivité et recruter de nouveaux doctorants. En outre, les partenariats industriels (thèses CIFRE) seront aussi recherchés de manière proactive. Les efforts de l'équipe pour cibler des supports de publications de qualité les mieux référencés du domaine seront également reconduits. Par ailleurs, l'équipe SID veut maintenir sa forte implication dans des recherches et des projets liés aux humanités numériques. Les projets ANR Cartas, EVA et le projet européen AMIS qui viennent de démarrer témoignent de cette volonté.

Autant les résultats scientifiques de SID sur la période 2019-2024 sont réels et probants dans le domaine de la BI et de la gestion intelligente des données massives, autant l'équipe fait le constat qu'une organisation en deux axes ne reflète pas sa dynamique scientifique. Régulièrement des sujets de thèse et des projets relèvent des deux axes. L'organisation en axe n'est pas non plus structurante. Pour le prochain contrat, le projet scientifique de l'équipe s'articule autour de trois thématiques sans pour autant parler d'axe : Smart Data Management, Data Protection, Data Analytics. Pour suivre l'évolution de ses thématiques, l'équipe a également décidé de changer de nom pour adopter celui de ***Data Intelligence (DataIn)***.

Smart Data Management

Le *smart data management* est un domaine qui combine des techniques avancées pour gérer, exploiter et analyser des données de manière plus efficace et intelligente.

Concernant l'optimisation des performances de système de gestion des données massives, l'équipe pense que la mobilisation de certains modèles d'apprentissage automatique devrait lui permettre d'optimiser les processus d'indexation, de partitionnement dans les modèles de données et ce en améliorant les temps de réponse des requêtes et en réduisant la consommation de ressources. Elle compte travailler sur l'amélioration des algorithmes de stockage de données sur des infrastructures distribuées (comme Hadoop, Spark, ou cloud) pour réduire les coûts tout en augmentant la performance d'accès aux données massives.
La gouvernance des données joue un rôle crucial dans les systèmes d'information décisionnels. L'équipe compte développer des approches pour garantir la cohérence, la fiabilité et la traçabilité. Pour être efficients, les lacs de données reposent sur un système de gestion des métadonnées. Nous avons déjà identifié que la création des métadonnées pour décrire et lier les entités de données présentes dans le lac était un verrou scientifique. Ce verrou s'accompagne de deux constats partagés dans la communauté des chercheurs des humanités numériques : les métadonnées associées aux entités de données sont souvent "pauvres" et avec

peu de sémantique ; la saisie des métadonnées est une tâche manuelle, chronophage et coûteuse. Notre ambition est de proposer des méthodes et outils "intelligents" pour assister l'utilisateur lors de la création de métadonnées "riches" décrivant les entités de données et leurs relations dans le lac de données. Pour cela l'équipe compte tirer parti et combiner des principes, méthodes et algorithmes d'apprentissage automatique, des graphes de connaissances et des ontologies.

L'équipe souhaite également étudier les possibilités offertes par le nouveau concept de data mesh pour l'industrialisation des processus de gestion des données massives.

Data Protection

L'intelligence artificielle joue également un rôle croissant pour la sécurité des données des réseaux IoT et en cybersécurité, permettant de détecter et prévenir les menaces grâce à l'analyse des données.

En analysant de grandes quantités de données, qui peuvent être structurées sous forme de graphes de connaissance, l'IA permet d'identifier des modèles anormaux et des comportements suspects qui pourraient passer inaperçus avec les méthodes traditionnelles. Par exemple, des modèles d'IA comme les Graph Neural Networks (GNN) peuvent être mobilisés pour détecter les attaques de compromission de nœuds. Ces réseaux de neurones, spécialement conçus pour traiter des données en graphe, sont capables d'analyser les relations complexes entre les entités et de repérer des communications inhabituelles ou des anomalies structurelles dans les données captées, suggérant ainsi une prise de contrôle non autorisée. Grâce à leur capacité à modéliser les interactions entre les nœuds et les arêtes, les GNN offrent une approche puissante pour anticiper et neutraliser les menaces de manière proactive.

En sécurité des données, de nombreuses approches s'appuient sur des modèles d'apprentissage automatique pour la détection d'anomalies ou le filtrage. Cependant certains modèles comme les modèles d'ensembles, le boosting, les perceptrons, les réseaux neuronaux binaires (BNNs) et surtout les réseaux neuronaux profonds sont souvent considérés comme des « boîtes noires » en raison de leur complexité et de leur opacité. Leur manque d'interprétabilité rend difficile la compréhension des décisions prises et limite la confiance des utilisateurs dans ces modèles. Par exemple, le contrôle d'accès repose généralement sur des règles formelles prédéfinies. Cependant, il est difficile de définir un ensemble de règles garantissant à la fois la cohérence et la sécurité du système. Ces dernières années, l'extraction de règles (ou "fouille de politiques", policy mining) a connu un essor important au sein de la communauté scientifique, notamment grâce à l'intégration de modèles d'apprentissage automatique. Néanmoins, un problème persistant réside dans l'absence d'explications rigoureuses sur la manière dont les résultats produits par ces modèles peuvent être considérés comme un ensemble complet et correct, assurant ainsi une sécurité optimale.

De plus, l'IA représente aussi un défi en sécurité des données, car les cybercriminels exploitent ses capacités pour automatiser les attaques et contourner les systèmes de défense traditionnels.

Les jumeaux numériques (Digital Twin) offrent une approche prometteuse pour améliorer la sécurité des réseaux émergents en permettant la simulation et l'analyse des vulnérabilités avant qu'elles ne soient exploitées. En reproduisant fidèlement l'environnement réseau, ces modèles virtuels devraient faciliter la détection proactive des menaces et l'optimisation des stratégies de défense.

Data Analytics

L'équipe compte concevoir de nouvelles stratégies d'optimisation pour les jointures complexes et les requêtes multi-sources dans les entrepôts et lacs de données, afin d'améliorer l'efficacité des analyses sur des ensembles de données massives.

Compte tenu de l'usage croissant de l'IA et des algorithmes d'apprentissage pour le stockage et la gestion des données massives, l'équipe souhaite étudier l'impact de cet usage sur l'analyse décisionnelle.

L'équipe va continuer à avoir un positionnement original dans l'environnement lyonnais du fait de ses thématiques de recherche en BI et autour des données massives. L'engagement de l'équipe dans son projet scientifique va lui permettre de renforcer les liens avec les autres laboratoires traitant de la BI en France et d'initier de nouvelles collaborations avec les entreprises. Les partenariats internationaux de l'équipe seront également confortés, que ce soit en termes d'animation scientifique, de co-encadrement de thèses ou de projets de recherche.

3- 3 Synthèse de l'autoévaluation

Le laboratoire ERIC a connu une période fructueuse de 2019 à 2024, marquée par une augmentation des ressources humaines, une production scientifique accrue et de qualité, ainsi qu'une forte implication dans les humanités numériques et la médiation scientifique. Le laboratoire a accueilli 49 doctorants sur la période, avec 26 soutenances de thèse, et a vu la création de 2 postes permanents. Le nombre d'Habilitations à Diriger des Recherches (HDR) a également augmenté. Les ressources financières du laboratoire se portent bien, avec notamment une part importante de ressources sur projets (PIA, ANR, …) et une proportion importante de thèses financées par des entreprises (50%).

La production scientifique a été significative avec 89 articles dans des revues et 197 communications dans des congrès, dont une augmentation de 28% des publications dans des revues Q1 et de 110% dans des conférences A/A*. Le laboratoire a par ailleurs organisé trois grandes conférences : EGC 2023, JdS 2022, et ADBIS-TPDL-EDA 2020.

L'équipe DMD s'est distinguée dans les domaines de l'apprentissage automatique, de l'apprentissage statistique et de la représentation des connaissances, développant des modèles pour gérer la complexité et l'hétérogénéité des données, et s'intéressant aux aspects éthiques et environnementaux.

L'équipe SID a proposé des solutions innovantes pour la gestion et l'analyse des données massives, en combinant divers concepts et en se concentrant sur la sécurité des données dans le cloud et l'Internet des Objets.

Le laboratoire ERIC est un acteur majeur des humanités numériques, avec de nombreuses collaborations interdisciplinaires et projets financés, et une forte présence sur la scène locale en médiation scientifique. Il a également de nombreuses collaborations internationales (figure 12). Cependant, il reste un défi concernant la valorisation de la production logicielle en raison de l'absence d'ingénieur dédié.

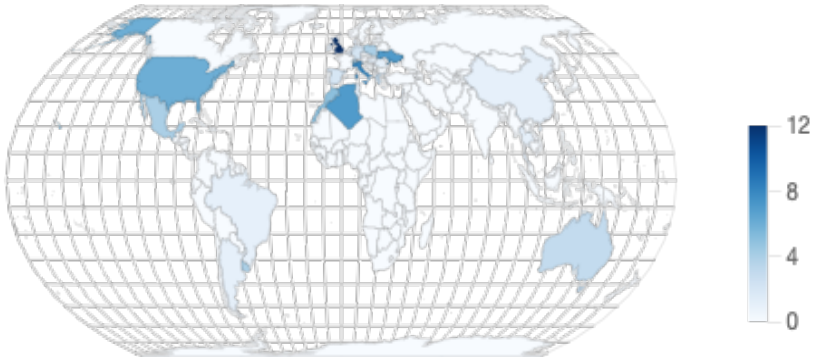

Figure 12 : Affiliations géographiques des co-auteurs internationaux des membres de l'unité.

4. TRAJECTOIRE DE L'UNITÉ

ERIC, de 1995 à 2024

Le laboratoire ERIC est une unité de recherche des universités Lyon 2 et Lyon 1, créé en 1995, qui effectue des recherches dans les disciplines de l'informatique et des mathématiques appliquées. Si à l'heure où l'Intelligence Artificielle est omniprésente, la synergie entre ces deux disciplines apparaît naturelle, le laboratoire ERIC fût précurseur dans le domaine dès le début du siècle, tant d'un point de vue recherche qu'enseignement.
Forte de ses 23 enseignants chercheurs (et environ autant de doctorants), le laboratoire ERIC mène des recherches dans le domaine du machine learning, de l'aide à la décision, de la gestion et de l'analyse des données massives. Par ailleurs, du fait de son fort ancrage dans l'université Lyon 2, il est également un acteur majeur dans le domaine des Humanités Numériques.

Outre ses objectifs scientifiques discutés au niveau des équipes, le laboratoire s'était fixé comme objectifs :
1. Améliorer la qualité des publications.
2. Inciter les membres MCF du laboratoire à passer une Habilitation à Diriger des Recherches.
3. Mettre en valeur notre production logicielle.
4. Intensifier nos actions de médiation scientifique.
Les points 2 et 4 ont été clairement respectés, avec 3 HDR passées au cours de la période et un nombre important d'activités de médiation scientifique. Concernant la qualité des publications, le ratio de publications dans les conférences et revues les plus sélectives restent stables, et les efforts entrepris dans la période, doivent être poursuivis. Néanmoins, l'unité n'a à ce jour pas trouvé d'autre moyen pour améliorer ce ratio qu'une sensibilisation de ses membres à l'importance de privilégier la qualité des publications. Enfin, concernant la valorisation de la production logicielle, cet objectif n'est clairement pas atteint. Mais tant que cette valorisation ne reposera que sur des enseignants-chercheurs (nous n'avons aucun ingénieur d'appui à la recherche), cet objectif sera difficile à réaliser.

Positionnement

L'unité est reconnue dans ses activités sur le plan local, tant par les autres laboratoires d'informatique et de mathématiques appliquées du site Lyon Saint-Etienne que par les laboratoires des disciplines Arts, Lettres, Langues et Sciences Humaines et Sociales. L'unité est également reconnue sur le plan national, comme un des piliers de la communauté EGC, mais aussi pour ses activités au sein de la Société Française de Statistique.

L'unité est également reconnue du fait de son activité pédagogique, à la fois par les différents parcours du master informatique de l'université Lyon 2 qui lui sont adossés, mais également par les nombreux documents pédagogiques produits par un des membres de l'unité. L'unité est également un acteur important sur le site lyonnais pour son activité de médiation scientifique et diffusion auprès du grand public.

L'activité internationale de l'unité est relativement importante en regard de sa taille. Outre de nombreuses collaborations individuelles de chercheurs, l'unité entretien des liens particuliers avec le Canada autour de questions d'humanités numériques, et également avec les pays du Maghreb avec lesquels nous formons un certain nombre de docteurs.

Projet scientifique

L'équipe DMD a réorienté ses recherches vers l'apprentissage automatique, réduisant son activité en Data Mining, pour mieux répondre aux besoins industriels actuels. L'équipe se concentre sur trois thématiques :
1. Développement de nouvelles méthodes d'apprentissage : l'équipe explorera des approches alternatives aux architectures profondes, en mettant l'accent sur des méthodes plus légères et efficaces. Ces travaux visent à réduire l'impact environnemental en évitant des techniques computationnellement coûteuses.
2. Étude théorique des modèles : l'équipe mènera des recherches théoriques pour garantir la robustesse et la validité des modèles, en se concentrant sur les bornes de généralisation et l'analyse de la convergence. Elle utilise des outils comme la théorie PAC-Bayésienne et la prédiction conforme pour fournir un cadre rigoureux garantissant les performances des modèles.
3. Diversité et pluralité des données : l'équipe travaillera sur une large gamme de données variées, incluant les séries temporelles, les données relationnelles complexes et les ensembles hétérogènes. Elle renforce son approche sur les données hétérogènes en utilisant des techniques comme l'apprentissage fédéré pour traiter des données distribuées de manière sécurisée, tout en préservant la confidentialité.
Ces axes permettent à l'équipe DMD de proposer des solutions adaptées aux contextes industriels et sociétaux, tout en maintenant la performance des modèles malgré les imprécisions des données.

L'équipe SID, renommée Data Intelligence (DataIn), prévoit de renforcer son réseau pour accroître sa visibilité et recruter de nouveaux doctorants, tout en recherchant des partenariats industriels. Elle continuera à cibler des publications de qualité et à s'impliquer dans les humanités numériques, comme le montrent les projets ANR Cartas, EVA et AMIS.

L'équipe se concentre sur trois thématiques :
1. Smart Data Management : l'équipe utilisera l'apprentissage automatique pour optimiser la gestion des données massives et améliorer la gouvernance des données.
2. Data Protection : l'IA sera utilisée pour détecter les menaces dans les réseaux IoT et en cybersécurité, avec l'aide des jumeaux numériques pour simuler les vulnérabilités.
3. Data Analytics : l'équipe développera des stratégies pour optimiser les analyses sur des ensembles de données massives et étudiera l'impact de l'IA sur l'analyse décisionnelle.
L'équipe maintiendra son positionnement à Lyon, renforcera ses liens avec d'autres laboratoires en France et initiera de nouvelles collaborations avec les entreprises, tout en consolidant ses partenariats internationaux.

Organisation

L'organisation de l'unité ne va pas évoluer pour le prochain contrat quinquennal, avec notamment une équipe de direction qui restera identique. Les responsabilités d'équipe continueront à être soumises à élection tous les deux ans. La seule évolution concernera l'axe de recherche transversal en humanités numériques que nous allons structurer, en le dotant d'un animateur et d'un budget propre.

Les travaux en humanités numériques du laboratoire ERIC ont permis de créer des liens avec différentes structures liées aux humanités numériques à l'extérieur du laboratoire, avec des parcours de formations, des axes institutionnels et scientifiques au niveau local, national, voire international. Ces différents liens sont bénéfiques à la vitalité des projets et à la reconnaissance des travaux du laboratoire dans ce champ des humanités numériques. Pour autant, il apparaît que pour donner encore plus de vitalité à cet axe, il s'agit de pouvoir lui donner une dynamique en interne, au-delà d'un affichage des travaux et projets regroupés sous cette bannière. Ainsi, le projet est de faire en sorte que cet axe bénéficie d'une animation scientifique qui pourra réfléchir avec les personnes concernées à la mise en place de cette dynamique.
Si de nombreux travaux ont été développés autour des ALLSHS, il s'agira pour l'animateur d'envisager le périmètre de cet axe, pour peut-être inclure une dimension pluri/interdisciplinaire au-delà des ALLSHS. Un point important est de pouvoir doter cet axe de moyens financiers. En effet, le travail avec d'autres disciplines amène parfois à des espaces de publications spécifiques, ne relevant pas nécessairement de l'informatique ; ces activités de publication méritent de pouvoir être soutenues, car elles peuvent notamment participer à la dynamique scientifique grâce au décloisonnement disciplinaire. Des moyens financiers permettront également de soutenir une dynamique de séminaires (qui pourrait s'articuler avec la dynamique des séminaires du laboratoire), en envisageant par exemple des séminaires croisés avec d'autres structures, d'autres disciplines. L'aspect mise en œuvre de travaux croisant plusieurs disciplines pourrait également donner lieu à des espaces de réflexion plutôt tournés sur le volet épistémologique : comment faire ce type de recherche ? quelles difficultés sont rencontrées et comment les surmonter ? Ceci pourrait être largement bénéfique d'un point de vue scientifique.

Stratégie partenariale

L'unité va continuer d'être un des acteurs du site Lyon Saint-Etienne dans le domaine de la recherche en informatique. Elle souhaite également être reconnue pour ses activités de recherche en mathématiques appliquées, en nouant plus de liens avec les laboratoires du site dans le domaine. Pour ce faire, l'unité soutiendra financièrement les collaborations avec l'Institut Camille Jordan notamment.

L'unité souhaite continuer à être un acteur national majeur dans le domaine des humanités numériques. La structuration de cet axe de recherche sera notamment un outil pour cela.

L'unité va également continuer à mener des projets de recherche en collaboration avec des entreprises, en privilégiant les collaborations sur le long terme, comme cela est déjà le cas avec plusieurs partenaires avec qui nous avons déjà encadré plusieurs thèses CIFRE.

Enfin, l'unité va chercher à intensifier ses collaborations internationales, en cherchant à structurer des collaborations pérennes qui sont pour l'instant uniquement des collaborations individuelles de membres de l'unité, sur le modèle de ce que nous avons développé avec l'Université d'Ottawa dans le domaine des humanités numériques.