Qualité des données complexes multi-sources : instrument de valorisation des données et d'extraction de connaissances

Les données ouvertes (Open Data) et la corrélation entre sources de données (Linked Data) posent de nombreux problèmes d'hétérogénéité, de sémantique et de droits d'usage. Le CNRS, à travers ses unités de recherche et ses grands instruments, concentre plusieurs centaines de bases de données et de corpus d'informations dont les volumes croissent de façon exponentielle et dont la valorisation se révèle un enjeu stratégique. Cette valorisation ne peut être effective que si les données et les connaissances qui en dérivent sont caractérisées qualitativement, quoi qu'il en soit du caractère subjectif ou contextuel de la notion de qualité (source CNRS : Défi Mastodons 2016).
Ce projet s'inscrit dans cette perspective, il questionne la qualité des données dont les modes d'acquisition sont variés et présentent différentes imperfections. Son but est de développer des prétraitements, des réparations et des extractions de connaissances afin de mieux analyser les données et les valoriser pour la prise de décision. Les sources utilisées traitent des migrations internationales. Par la ***création d'outils*** efficaces, le projet vise à ***améliorer la connaissance des chercheurs et des décideurs les relations entre le droit des migrants et les politiques migratoires, et plus largement les enjeux sociaux et politiques des mobilités contemporaines***.

DESCRIPTION DU PROJET

Les migrations internationales ont pris dans le monde contemporain une ampleur inédite. Cela pose de nouveaux défis à la communauté scientifique en termes d'analyse et de compréhension des phénomènes migratoires. Le premier est celui des données et de leurs qualités. En effet, nombreuses sont les bases de données statistiques sur les migrations internationales. Pour aller plus loin dans l'analyse de ce phénomène complexe et multidimensionnel, requestionner des données précédemment acquises et leur mise en synergie avec d'autres types de données est nécessaire.
Ce projet propose d'étudier les parcours migratoires d'un point de vue du migrant, considéré comme un acteur clé dont les droits doivent être respectés et la protection assurée en toutes circonstances. Grâce à une approche multidisciplinaire, l'objectif est de combiner des grandes enquêtes, des sources administratives, des collections juridiques, et des corpus d'entretiens ***pour améliorer la connaissance des phénomènes migratoires et favoriser le dialogue avec les décideurs politiques afin d'aider les États dans la définition de politiques publiques respectueuses à la fois des enjeux de souveraineté nationale et du droit des migrants***.
Notre champ d'analyse porte sur différents types de données collectées par les laboratoires Migrinter, UMR 7301 (CNRS-Université de Poitiers) et CEPED, UMR 196. (IRD-Paris Descartes) : il s'agit notamment de : (1) registres administratifs et judiciaires, (2) corpus juridique des pays d'Afrique de l'Ouest et des Balkans, carrefour important des circulations migratoires vers l'Europe, (3) récits de vie, notamment des mineurs en mobilité et des migrants de Calais, (4) recensement effectué récemment auprès des personnes déplacées en Syrie et dans le Kurdistan irakien, (5) répertoire des mobilités des communautés scientifiques mexicaines, (6) des rush-video sur les populations réfugiés en Inde et en Egypte.

A titre d'exemple, dans le domaine des migrations internationales, les données administratives permettent l'observation des relations entre plusieurs acteurs : État, migrants et groupes criminels. Le mode d'enregistrement fait apparaître des ***données manquantes***, des ***valeurs mal renseignées*** ou ***mal orthographiées et des incohérences***. Ces données peuvent être mises en perspective avec des récits de vies collectés sur les routes migratoires, riches en informations contextualisées, et des corpus juridiques, nationaux et internationaux, définissant les droits des migrants, des demandeurs d'asile et des réfugiés. L'hétérogénéité de ces bases de données et leur qualité variable posent un réel challenge quant à leurs croisements et exploitations.

A partir d'un objet commun, le ***parcours migratoire***, des ***terrains différents*** inscrits dans des ***temporalités variables, l'objectif est de tenter de repérer les points de convergence potentiels des routes migratoires qui proviennent de régions d'origine différentes et s'orientent vers l'Europe. Cette question pourra être interrogée sous plusieurs angles :***

- Les relations entre le droit des migrants et les politiques migratoires dans les pays de transit ou/et d'accueil,
- L'implication des réseaux de traite des êtres humains dans la structuration de certains parcours migratoires,
- Les enjeux sociaux et politiques de la mobilité des mineur(e)s "non accompagnés",
- Le parcours universitaire et professionnel des élites du Sud.

Lieu d'émigration vers Iles Canaries - Mauritanie

Le défi ici est de répondre à ces usages sur des bases de données complexes, hétérogènes, de qualité inégale et de représentativité variable.

Tout d'abord, les données issues des sciences sociales sont de nature et de formats différents : des bases de données fortement structurées, des corpus faiblement structurés, représentés sous forme de texte brut, en passant par des données semi-structurées, les corpus juridiques. Ainsi, nous disposons de diverses bases de données, représentant des réalités migratoires dans plusieurs régions du monde : l'Afrique de l'ouest, les Balkans, le Moyen-Orient et l'Europe.

En outre, l'hétérogénéité des bases de données est liée aux modes de production de l'information. En effet, la base de données judiciaire est conçue comme un registre administratif et évolue quotidiennement. Alors que celle sur les personnes déplacées en Syrie est le fruit d'un recensement de population ; elle nous donne une photographie à un instant T. Les corpus juridiques sont plus immuables. On peut relever aussi que le recueil de l'information est plus ou moins rigoureux ; certaines bases de données sont construites en suivant un protocole scientifique ; la qualité de la base de données judiciaire dépend du professionnalisme des personnels de justice et des éléments fournis par le justiciable. Les données des grandes enquêtes sont collectées en suivant un protocole scientifique bien défini et rigoureusement appliqué.

De plus, l'hétérogénéité des sources pose la question de la qualité des données. On peut aussi relever des données manquantes sur certains attributs cruciaux, des imprécisions ou données mal orthographiées et des incohérences. Notons également la présence de données entachées d'incertitude (e.g. information oubliée ou dissimulée). Pour les récits de vie, au-delà de la richesse des informations qui s'y trouvent et la présence d'une forme de structure (recueil selon une grille d'entretien préétablie), la saisie est souvent perfectible, sans ponctuation, avec des noms de lieux et des mots incorrects. La confidentialité des données est aussi une dimension essentielle à prendre en

compte.
Par ailleurs, les sources mises en jeu, sans relation directe et portant sur des territoires différents, admettent des questionnements semblables et complémentaires ; par exemple, la migration des mineurs “non accompagnés” et le déplacement de populations qui fuient des zones de conflits ou de violences en Afrique subsaharienne ou au Moyen-Orient. Le problème qui se pose, concerne l’interrogation d’un même phénomène avec des bases de données qui ont été constituées différemment et dont la qualité est inégale. En d’autres termes, quel est l’impact de la qualité des données sur les résultats ? La qualité est ici considérée au sens large, comme par exemple la prise en compte de la représentativité. En effet, la base de données sur les personnes déplacées en Syrie est de meilleure qualité mais représente un état des lieux à un instant donné, alors que la base de données judiciaires, de moindre qualité et évolutive, est plus exhaustive et actualisée en permanence.

LES VERROUS SCIENTIFIQUES ET TECHNIQUES :
Afin de traiter et analyser les données issues des sources précitées, nous devons faire face aux verrous scientifiques suivants :

- Hétérogénéité des données et leur complexité intrinsèque

La gestion et la complémentarité des différentes ressources constituent un verrou scientifique reconnu dans la communauté comme majeur et difficile. De nombreux travaux s'intéressent à la mise en correspondance de bases de données, de données semi-structurées et/ou d'ontologies. Dans ce projet, nous nous sommes intéressés à la mise en correspondance de données de nature très hétérogènes aussi bien syntaxiquement que sémantiquement comprenant l’identification d’entité du même monde réel enfouis dans les différentes sources de données connu par le thème de la résolution des identités. Les liens ainsi mis en évidence constituent les prémisses d’un monde « linked data ». Le format RDF est incontestablement un bon candidat pour unifier le format et structure des différentes sources de données.

Ce choix stratégique nous permettra dans le futur d’exploiter des données externes au projet de manière à enrichir les données du projet et améliorer leur qualité.

- Mesurer, intégrer et quantifier l’impact de la qualité des données sur les résultats d’analyses

Les problèmes liés à la qualité des données posent un réel défi pour la qualité des résultats. Les techniques mises en œuvre doivent prendre en considération cette dimension depuis l’étape de prétraitement jusqu’à l’interprétation et la validation des résultats. Ceci passe par la définition de mesures pertinentes de la qualité des données avant et après prétraitement, mais aussi des résultats en donnant des indicateurs de qualité/erreurs. Le challenge est d’y associer un modèle interactif et cyclique (prétraitement, analyse, et validation).

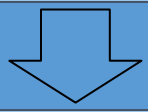
- Des données aux connaissances

Les données issues des sciences sociales posent un réel défi en termes de connaissances à extraire (profils communs, nouveaux types de relations et de motifs). Cela suppose de trouver de nouvelles formulations de ces connaissances en terme de motifs, de nouveaux prédicats et induction de règles pour juger de leurs intérêts et de nouvelles techniques d’inférence ou d’extraction. Ceci induit de nouveaux problèmes sur le plan de la représentation, du calcul et de la complexité algorithmique. Compte tenu de la faible qualité de certaines bases de données cibles, il serait intéressant de voir l’impact de cette qualité sur le niveau d'intelligibilité des connaissances extraites.

PROCESSUS DE TRAITEMENT, D'ANALYSE ET D'EXPLOITATION DES DONNEES :

Prétraitement des données : améliorer la qualité des bases de données

- Réparation et enrichissement sémantique des bases de données
- Mise en relations de bases données hétérogènes et de qualité variable

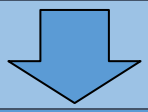


Analyse des données et découvertes de connaissances

- A partir d'un objet commun, le ***parcours migratoire***, des ***terrains différents*** inscrits dans des ***temporalités variables, repérer les points de convergence potentiels des routes migratoires qui proviennent de régions d'origine différentes et s'orientent vers l'Europe. Cette question pourra être interrogée sous plusieurs angles :***

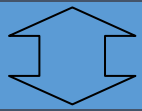
  - Les relations entre le droit des migrants et les politiques migratoires dans les pays de transit ou/et d'accueil,
  - L'implication des réseaux de traite des êtres humains dans la structuration de certains parcours migratoires,
  - Les enjeux sociaux et politiques de la mobilité des mineur(e)s
  - Le parcours universitaire et professionnel des élites du Sud

- Approches complémentaires d'extraction de connaissances :
  - Récits de vie : fouille de textes et extraction d'entités nommées.
  - Des données aux connaissances: représentation et raisonnement.
  - Découvertes de motifs, de phénomènes rares, de règles d'associations…



Analyse et interprétations des résultats

- Amélioration de la qualité des résultats retournés dans un processus itératif



Extractions de connaissances nouvelles
Construction d'un Entrepôt inédit de bases de données hétérogènes sur les migrations internationales, produites dans les pays d'origine, de transit et d'accueil

DOMAINES D'APPLICATION OU USAGES ENVISAGES :

Ce projet, par sa nature et son objet d'étude, porte naturellement sur un domaine d'application majeur : les parcours migratoires contemporains.

- En termes d'usage, le but est de fournir de nouveaux résultats d'analyse capables d'éclairer les politiques publiques sur ce phénomène complexe et multidimensionnel.
- Il ouvre aussi le dialogue entre des sciences peu habituées à ce côtoyer, les sciences sociales et les sciences de l'informatique ; les résultats méthodologiques participeront aussi à faire évoluer les métiers de chacun et à favoriser le transfert de compétences vers le monde économique pour lequel la mobilité des personnes est aussi un enjeu de première importance

PARTENAIRES IMPLIQUES DANS LE DEVELOPPEMENT DU PROJET :

Ce projet regroupe des laboratoires en informatique et en sciences sociales. Les expertises spécifiques des différents partenaires sous-tendent l'organisation du projet : la thématique fédératrice du **CRIL** (Centre de Recherche en Informatique de Lens, CNRS/INS2I, UMR 8188) concerne l'intelligence artificielle et ses applications. En particulier, le traitement d'informations incomplètes, incertaines, incohérentes, dynamiques et multi-sources, l'algorithmique et la fouille de données. Ensuite, le **LIAS** (Laboratoire d'Informatique et d'Automatique pour les Systèmes, EA 6315, Poitiers) traite de l'hétérogénéité des données multi-sources, de l'incertitude/imprécision inhérente aux données et du manque des données. **LIPADE** (Laboratoire d'Informatique Paris Descartes, EA 2517, Paris) s'intéresse à la gestion de la qualité de données incluant l'incohérence et l'incertitude, la fusion de données hétérogènes, le raisonnement logique et la fouille de données. Parallèlement, le **LIRMM** (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, CNRS/INS2I, UMR 5506, Montpellier) et **TETIS** (Territoires, Environnement, Télédétection et Information Spatiale, Cirad, Irstea, AgroParisTech, Montpellier) se focalisent sur les problématiques de fouille de données et d'extraction de connaissances à partir de données textuelles hétérogènes. **MIGRINTER,** (Migrations Internationales, Espaces et Sociétés, CNRS/INSHS, UMR 7301, Poitiers) et le **CEPED** (Centre Population & Développement, UMR 196, Paris Descartes - IRD) travaillent sur les questionnements autour des parcours migratoires et les rapports qu'ils entretiennent avec les évènements consécutifs de l'application du droit.

MOTS CLES :
Big data, Données sciences sociales et juridiques, Qualité des données, Extraction de connaissances, Raisonnement et inférence, Parcours migratoire.