Offre de stage Master 2

Modélisation et fouille d'une base de données sur les pharmacopées anciennes

Encadrement : Agnès Braud (ICube – équipe SDC), Florence Le Ber (ICube – équipe SDC), Nathalie Hernandez (IRIT)
Lieu : Laboratoire ICube (site d'Illkirch : Pôle API, 300 bd Sébastien Brant)
Gratification : tarif légal en vigueur, environ 630 euros par mois

Ce stage s'inscrit dans le cadre du projet de recherche TRACES financé par l'ANRS Maladies infectieuses émergentes (ANRS MIE) et qui a pour objectif le développement d'outils pour la recherche de pistes originales de lutte contre la tuberculose. La tuberculose demeure en effet l'une des maladies infectieuses les plus meurtrières au monde et l'apparition de souches résistantes aux traitements fait du développement de nouvelles approches thérapeutiques un enjeu majeur. Le projet est lui-même lié au groupe de recherche interdisciplinaire national sur les pharmacopées anciennes IRGAP [1]. L'étude des pharmacopées anciennes nécessite des compétences dans de nombreuses disciplines qu'IRGAP vise à rassembler. Pour ce travail interdisciplinaire, la mise en place d'une base de données est essentielle pour, d'une part, stocker les informations produites par les chercheurs du groupe ou issues d'autres sources et utiles pour leurs travaux, d'autre part, permettre une analyse des données pouvant intégrer les informations de toutes les disciplines pertinentes [2,3]. Le modèle retenu est celui de la Wikibase [4] sur laquelle repose la base de connaissances Wikidata [5]. Wikidata stocke les données structurées de projets de la Fondation Wikimédia sous la forme de triplets et est notamment utilisé pour alimenter les Infobox des pages Wikipédia. Cette base a été retenue pour différentes raisons, en particulier son caractère collaboratif, les possibilités de liaison avec des sources de données existantes reconnues, la possibilité de représenter les incertitudes sur les données (incertitudes inhérentes aux données historiques car des informations ont pu être perdues ou être difficiles à interpréter), son caractère multilingue, la possibilité d'exporter les informations vers Wikidata afin de diffuser les résultats des recherches au grand public, et le grand nombre d'outils disponibles en particulier pour l'interrogation et la visualisation des données. Une instance Wikibase a été mise en place pour le projet. Il est nécessaire de mener une réflexion sur la modélisation des données dans cette base : quelles données ? Décrites par quelles propriétés ? Ces propriétés sont-elles déjà présentes dans Wikidata ?

Le sujet de ce stage s'inscrit dans ce cadre et vise à mettre en place l'ensemble du processus allant de la modélisation à la fouille des données, voire leur visualisation. Il s'agira de réfléchir dans un premier temps à la mise en place d'un modèle de stockage des informations sur des remèdes (préparations médicales) : médecin ayant proposé le remède et manuscrit dans lequel il l'a fait, remède lui-même, sa composition, les maladies et symptômes traités, les molécules présentes dans les ingrédients et les principes actifs identifiés. Ce modèle devra intégrer les incertitudes possibles sur les traductions faites des textes anciens (ingrédients, maladies et symptômes) et les résultats des chercheurs. Il reprendra autant que possible les propriétés utilisées dans Wikidata si celles-ci décrivent correctement ce qu'est la donnée, en définira de nouvelles sinon.

Ensuite des outils de fouille de données seront développés afin d'aider les chercheurs du projet travaillant sur la recherche de pistes pour de nouveaux médicaments. Il s'agira notamment de proposer des outils de découverte de sous-graphes dans la base Wikidata afin de mettre en évidence des récurrences dans les compositions/préparations des remèdes, avec la possibilité de considérer les ingrédients (végétaux, minéraux, animaux) à différents niveaux de leur taxonomie, voire au niveau moléculaire. Ces outils devront permettre de gérer les éventuelles incertitudes sur les données.
Ce travail sera appliqué sur des données liées aux recherches sur la tuberculose.

Profil recherché :
- Master 2 en Informatique
- Compétences en modélisation et fouille de données
- Intérêt pour le travail en lien avec d'autres disciplines (histoire et biochimie)

Candidature : envoyer un CV, une lettre de motivation et les notes de M1 à Agnès Braud (agnes.braud@unistra.fr) et Florence Le Ber (florence.leber@engees.unistra.fr)

Références
[1] https://irgap.unistra.fr/
[2] Agnès Braud et Xavier Dolques. À la recherche de l'origine des collyres. La science des données à la rescousse ?. Dans M. Pardon-Labonnelie, P. Fechter, V. Pitchon (Dirs.). *Le Temps des collyres. Approche interdisciplinaire des pharmacopées*, EUD, à paraître
[3] Vanessa Fokou, Karim El Haff, Agnès Braud, Xavier Dolques, Florence Le Ber, et al.. Exploring Old Arabic Remedies with Formal and Relational Concept Analysis. *Concepts 2024, Cadiz, Spain, septembre 2024*, Sep 2024, Cadiz, Spain.
[4] https://wikiba.se/
[5] https://www.wikidata.org/wiki/Wikidata:Main_Page