# M2 MIASHS : cas d'étude

Sitthida Samath, UAR Persée
Julien Velcin, Université Lyon 2, Laboratoire ERIC

Février 2024

La commande qui vous est fixée consiste à développer un système de recherche d'information qui permet de naviguer efficacement dans un grand corpus de données textuelles. Il s'agit des métadonnées descriptives des documents disponibles sur www.persee.fr, programme et portail de numérisation et de diffusion du patrimoine scientifique.

Les documents sont de tous types (articles, comptes-rendus etc.), majoritairement issus de revues en sciences humaines et sociales, francophones, et couvrent une production du dix-neuvième siècle à aujourd'hui. A l'origine, un document est contenu dans un fascicule qui appartient lui-même à une collection correspondant généralement à une revue scientifique. Pour permettre la navigation dans le portail Persée, les collections sont associées à une discipline principale https://www.persee.fr/disciplines.

Le jeu de données décrit plus de 900 000 documents. Il a été produit à l'occasion du cas d'étude, à partir d'un sous-ensemble des fichiers de dumps de données liées `.rdf` du triplestore Persée, disponibles à l'adresse https://data.persee.fr/explorer/demander-un-dump/dumps-collections/ et à leur état d'octobre 2021. Les données rdf-xml originales ont été aplaties et préparées sous forme de tableaux (dataframes pandas Python). Le jeu de données comprend également un tableau de correspondance collection-discipline principale.

Pour chaque entrée de document, en plus du titre et sous-titre, des auteurs, de la date de publication, on peut trouver, quand ils existent, un résumé, des mots clés, une table des matières. Les données comprennent aussi, sans caractère exhaustif, des relations de citation entre documents de Persée. Pour les champs multi-valués, on trouvera une colonne par valeur, avec un indice entre crochets dans le nom des colonnes répétées. Par ailleurs, l'identifiant du document contient le code de collection (`collection_id`) qui permet d'exploiter la discipline principale. Pour en savoir plus sur le modèle de données et les intitulés des colonnes, issus des standards de métadonnées de description bibliographique, se reporter à l'adresse https://data.persee.fr/explorer/schemas-de-donnees/ ("niveau document"). Le jeu de données mis à disposition pour le case d'étude est soumis aux conditions générales d'utilisation de Persée à l'adresse https://www.persee.fr/cgu.