# AI for Enhanced Access To Legal Texts Network
# -
# French Legislation Use Case

March 29, 2021

Titre : L'IA pour un meilleur accès au réseau de textes juridiques - Application à la législation française.

Mots clés : IA, Contenu sémantique, Réseau de documents, Textes juridiques, Législation française

## 1 Contexte et motivations

Dans un cadre de transparence envers ses citoyens et afin de faciliter leur participation à la vie démocratique, plusieurs pays ont opté pour le partage de l'information publique et adopté des lois favorisant l'accès à cette information sous toutes ses formes.

En particulier, l'accès au droit est rendu possible via des outils en ligne comme Legifrance [1], un service public de diffusion du droit créé en 2002 en France. Il propose une base très complète constituée des codes officiels et textes consolidés en vigueur, textes du journal officiel et de jurisprudence.

Cet accès doit permettre au citoyen de tracer le cadre de ces droits et devoirs face aux situations auxquelles il est confronté tous les jours. Or, tel qu'il est conçu le droit peut s'avérer complexe et inaccessible pour un simple citoyen.

En effet, malgré l'encadrement dont bénéficie l'utilisateur de Legifrance pour interroger son contenu, un usage optimal suppose de maîtriser le mode d'élaboration des textes, leurs structures et chainages dans le temps, la hiérarchie des normes ainsi que le langage utilisé.

D'autres sites publics offrent des versions explicatives du droit présenté sous sa forme brute sur Legifrance. Leurs consultation s'impose au spécialiste du secteur de droit, tout comme au simple utilisateur, lorsqu'il s'agit d'interpréter les règles de droit. Ils

[1] https://www.legifrance.gouv.fr/

seront souvent amenés à naviguer parmi les pages de ces sites et à travers les différents corpus (législation, jurisprudence, etc.) pour pouvoir construire une réponse à un besoin spécifique.

Dans ce contexte, l'accès à l'information juridique est la première grande question dans l'accès au droit. Les textes dans le domaine juridique possèdent des caractéristiques spécifiques qui sont importantes à prendre en compte pour améliorer l'accès à l'information. D'un côté, le contenu sémantique de ces textes est souvent exprimé par un vocabulaire et sous des formes linguistiques complexes. D'un autre côté, les documents sont de différents types avec une structure particulière à chacun de ces types et ils contiennent des références de différentes natures vers d'autres textes qui définissent le contexte dans lequel ils doivent être interprétés.

Les systèmes d'accès à l'information juridique existants ne proposent pas de solutions directes pour prendre en compte une recherche d'information qui porte aussi bien sur le contenu sémantique que sur les liens intertextuels. Ils contournent cette difficulté avec des techniques simples, par exemple, en modélisant les liens comme des attributs qui sont intégrés dans la base (par exemple "modifié par", "abrogé par") et qui peuvent être interrogés. Ils représentent aussi le contenu sémantique riche et complexe comme un sac de mots indexant les textes. Les résultats retournés ne se présentent pas comme des graphes et l'utilisateur est amené à parcourir les liens hypertextes pour construire le contexte de la réponse (un utilisateur peut facilement s'y perdre sans trouver ce qu'il cherche).

## 2 Objectifs et hypothèses de recherche

La limitation des systèmes d'accès actuels est problématique pour les professionnels de droit du fait de l'abondance et de la diversité des relations qui lient entre elles les sources de loi. Ceci a été confirmé par l'analyse des besoins de ces professionnels qui cherchent à formuler des requêtes complexes qui portent aussi bien sur le contenu des documents que sur les relations intertextuelles qu'ils entretiennent.

Ces besoins s'expriment souvent sous la forme de requêtes complexes, dites requêtes relationnelles, qui portent à la fois sur le contenu sémantique des documents et sur les liens intertextuels. Par exemple : "*Quelles sont les décisions de jurisprudence qui citent l'article 1382 du code civil*?", "*Quels sont les textes qui modifient les articles du Code de l'environnement qui concernent les véhicules à moteur et les chemin ruraux*?".

Les récentes avancées dans les différents domaines de l'IA ont mis en avant de nouvelles approches et méthodes de traitement de données de différentes natures (brutes, formatées, etc.) et types (numériques, textuelles, etc.). En particulier, dans le contexte d'accès à l'information juridique, nous jugeons très pertinent la mise en place d'une approche pluridisciplinaire d'analyse et de fouille dans ces collections de textes pour la prise en compte de ces dimensions sémantique et intertextuelle. L'objectif à terme est de proposer un système qui utilise des méthodes d'IA (méthodes statistique, symboliques, de traitement automatique de textes) combinées avec des techniques d'analyse de

graphes pour mieux répondre aux besoins des utilisateurs.

Le projet a un double objectif qui se décline en deux étapes :

- Explorer l'utilisation des nouvelles technologies de l'IA pour la fouille de gros volumes de textes juridiques disponibles en ligne (modélisation sémantique, résumé automatique, traduction automatique, classification, régression, etc.).
- Intégration des résultats de la première étape dans un modèle sémantique plus complet pour un système riche d'accès à l'information juridique.

## 3 Méthodologie

Afin d'atteindre le premier objectif du projet, nous proposons de se baser sur les nouvelles techniques de traitement automatique de langue (plongement de mots et phrases dans des espaces vectoriels) pour la fouille de la grande masse de textes juridiques disponible. Ces technologies doivent permettre l'extraction d'information, la recherche d'information par similarité sur les mots, la construction de modèles thématiques (topic modeling) statiques mais aussi dynamiques et la génération de clusters de textes selon des caractéristiques que nous pouvons extraire de leurs métadonnées (auteurs, dates, lieu, etc.).

Une approche d'intégration des techniques de NLP utilisés à la première étape avec le modèle sémantique et de graphe décrite dans l'état de l'art sera proposée pour atteindre le deuxième objectif. Cette méthodologie doit pouvoir tirer profit des outils sémantiques disponibles proposant des moteurs d'indexation basiques (ex. Solar) afin de proposer un outil complet enrichi avec les nouveaux modèles linguistiques.

Les données sur lesquelles cette étude peut se baser proviennent de différentes sources:

- Documents de Legifrance : plusieurs types (textes nationaux, codes, jurisprudence), textes structurés selon un format DTD complexe, différentes métadonnées selon le type et références entre les textes construisant un réseau dense.
- Documents d'EUR-Lex [2] (droit de l'Union Européenne) : plusieurs types (journal officiel : législation et communications, jurisprudence), métadonnées et références entre les textes.
- Textes locaux : collectés sur les sites des collectivités territoriales, plusieurs types (documents locaux, documents éditoriaux), textes non-structurés (pdf, rdf, html, doc), plusieurs irrégularités peuvent exister dans les textes.

**Calendrier** Afin d'atteindre les objectifs, le travail sera divisé en trois grandes étapes:

1. Revue de la littérature autour de l'utilisation des nouvelles technologies de l'IA pour l'analyse des textes juridiques (5 mois).

[2] https://eur-lex.europa.eu/homepage.html

2. Sélection et adaptation des approches aux particularités des textes français, implémentations et tests (24 mois).

3. Rédaction du manuscrit final et préparation de la soutenance (7 mois).

## 4 Infrastructures et outils existants

Le travail se basera sur une ontologie développée pour les textes juridiques français et un outil d'instanciation automatique de l'ontologie avec des textes codifiés (version de test, à élargir pour les autres types de textes).

## 5 Profil du candidat et candidatures

Le (la) candidat(e) devra avoir de très bonnes aptitudes en apprentissage automatique et traitement de données. Une connaissance en ingénierie de connaissances et techniques du web sémantique sera très appréciée. Le (la) candidat(e) devra avoir de très bonnes aptitudes linguistiques en français et en anglais (parlé et écrit). Les candidats intéressés sont priés de contacter Nada Mimouni (nada.mimouni@lecnam.net) et Elisabeth Metais (elisabeth.metais@lecnam.net).

## References


[1] Mimouni, N., Nazarenko, A. and Salotti, S. Answering Complex Queries on Legal Networks: A Direct and a Structured IR Approaches. In AI Approaches to the Complexity of Legal Systems, Lecture Notes in Computer Science, vol 10791. Springer., pages 451-464, 2018.

[2] Mimouni, N., Nazarenko, A., Paul, E. and Salotti, S. Towards Graph-based and Semantic Search In Legal Information Access Systems. In Twenty-Seventh Annual Conference on Legal Knowledge and Information Systems (JURIX 2014), pages 163-168, Krakow, Poland, 2014.

[3] Alschner, Wolfgang, AI and Legal Analytics (November 2, 2020). in Florian Martin-Bariteau & Teresa Scassa, eds., Artificial Intelligence and the Law in Canada (Toronto: LexisNexis Canada, 2021), Available at SSRN: https://ssrn.com/abstract=3733957.

[4] Tarasconi, F., Botros, M., Caserio, M., Sportelli, G., Giacalone, G., Uttini, C., Vignati, L., and Zanetta, F. (2020). Natural Language Processing Applications in Case-Law Text Publishing. JURIX.

[5] Chalkidis, I., Kampas, D. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law 27, 171–198 (2019). https://doi.org/10.1007/s10506-018-9238-9.

[6] Leone V, Di Caro L, Villata S. Taking stock of legal ontologies: a feature-based comparative analysis. Artificial Intelligence and Law. 2019 Jun 13:1-29.

[7] Bibal, A., Lognoul, M., de Streel, A. et al. Legal requirements on explainability in machine learning. Artif Intell Law (2020).

[8] B. Waltl, G. Bonczek, E. Scepankova and F. Matthes, Semantic types of legal norms in German laws: classification and analysis using local linear explanations. Artificial Intelligence and Law. 2019 27 (1), 43-71.

[9] Agnoloni, Tommaso, Lorenzo Bacci, Ginevra Peruginelli, Marc van Opijnen, Jos van den Oever, Monica Palmirani, Luca Cervone et al. "Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links." In Jurix, pp. 113-118. 2017.

[10] Medvedeva, M., Vols, M. and Wieling, M. Using machine learning to predict decisions of the European Court of Human Rights. Artif Intell Law 28, 237–266 (2020). https://doi.org/10.1007/s10506-019-09255-y.