DEMC Journal

N° 3 | 2025
2025

Outiller informatiquement l’usage d’un grand corpus de textes juridiques

Opérationnaliser la base de données du Sabin Center

Valentin BRUNEL *Ingénieur de Recherche*
MSH SUD
Institut de Recherche pour le Développement
Camille MARTINI

Édition électronique :
URL :
https://demc-journal.org/articles/revue-3/3352-outiller-informatiquement-lusage-dun-grand-corpus-de-textes-juridiques
ISSN : 3036-5295
Date de publication : 19/12/2025

Pour **citer cette publication** : BRUNEL, V., MARTINI, C. (2025) Outiller informatiquement l’usage d’un grand corpus de textes juridiques. *DEMC Journal*, (3). https://doi.org/10.34745/

Mots-clés :

Cet article vise à présenter quelques travaux entrepris au sein du Centre d'Études et Recherches internationales et communautaires (CERIC) de l'UMR Droits international comparé et européen (UMR DICE), afin de faciliter l'exploitation d'un grand corpus de textes juridiques dans le cadre du projet de recherche de l'Agence nationale pour la Recherche (ANR) sur les expertises dans les procès climatiques : fabrique, usages et réception – Proclimex. Après avoir illustré l'apport d'une démarche de ce type par rapport aux travaux français, francophones et internationaux, l'article s'attache à décrire la base constituée ainsi que son originalité. Enfin, quelques exemples d'utilisation d'analyses automatisées dans la recherche en droit sont proposés en troisième partie, appuyés par des extraits des scripts disponibles en accès ouvert.

Au-delà du cas d'usage précis que représente notre travail sur les données compilées par le *Sabin Center for Climate Change Law*, un laboratoire de recherche de l'université de Columbia également relié au Earth Institute (ci-après, le Sabin Center), cet article souhaite illustrer la fertilité d'une démarche de recherche empirique et informatique pour la science juridique. À la croisée de la linguistique de corpus, du traitement automatique des langues, et du droit, les méthodes et travaux illustrés ci-dessous peuvent faire l'objet de développements et discussions dans de nombreux champs juridiques.

Même si les approches empiriques deviennent de plus en plus nombreuses et constituent un champ important de la recherche juridique, avec notamment des revues dédiées (*Journal of Empirical Legal Studies*, et en France la récente revue *Jurimétrie*, par exemple), ces dernières manquent souvent d'un éclaircissement concret quant aux méthodes employées. L'intérêt de cet article par rapport à l'abondante littérature présentée ci-dessous est donc de présenter un cas pratique, presque pas à pas, n'éludant pas les questions techniques souvent passées sous silence et pourtant cruciales dans la détermination des questions de recherche, et des réponses apportées.

Enfin, alors que les questions empiriques, l'interdisciplinarité et les méthodes comparatives en droit touchent de plus en plus de juristes (et de non-juristes), cet article propose un exposé des potentialités, mais aussi des limites de certaines approches, afin de favoriser leur adoption consciente et responsable, dans un contexte où l'accessibilité croissante de certains outils peut entraîner l'augmentation problématique d'usages "naïfs", même en milieu savant.

• I. L'analyse juridique empirique outillée informatiquement : une longue histoire, des usages croissants dans le monde

anglo-saxon

Les données textuelles occupent une place centrale dans la recherche en droit. Le virage numérique des 50 dernières années a naturellement entraîné un développement de méthodes d'utilisation d'outils informatiques spécialisés pour l'analyse de données juridiques sous forme de textes (Alschner et al., 2017; Blackham A., 2022; Mulcahy L. & Wheeler S., 2020; Sadl U., 2017). L'étude de la littérature utilisant ces méthodes met en lumière l'importance des technologies de fouille de textes et d'analyses lexicales et de réseau dans l'étude du discours juridique.

Jusqu'à la fin des années 2000, les analyses empiriques en droit reposaient majoritairement sur des traitements semi-manuels ou des bases de données classiques, tels que des logiciels de traitement de textes génériques, des catalogues de décisions, ou des feuilles Excel. Les chercheurs comme Peter McCormick aux Etats-Unis (P. McCormick, 2009; P. J. McCormick, 2015) ou Cynthia L. Ostberg et Matthew Wetstein au Canada (Ostberg, 2007) utilisaient des méthodes statistiques appliquées à des corpus relativement restreints et soigneusement sélectionnés. L'objectif était principalement descriptif : mesurer les tendances de citation, la charge de travail des tribunaux, ou l'influence de facteurs socio-politiques sur les jugements. Des recherches empiriques plus classiques, mais également informatisées, comme celles de Dietrich Fausten, Ingrid Nielsen et Russell Smyth en Australie (Fausten et al., 2007) illustrent l'évolution de l'usage des citations et la charge de travail des cours d'appel, en combinant analyses statistiques et traitement informatique des décisions judiciaires.

L'ouvrage de David Muttart de 2007 s'inscrit pleinement dans ce courant en mobilisant un codage systématique et manuel de l'ensemble des décisions de la Cour suprême du Canada sur une période donnée, permettant une analyse statistique exhaustive des variables juridiques et contextuelles, et illustrant la volonté de combler le déficit empirique dans l'étude doctrinale de la jurisprudence (Muttard, 2007). Enfin, des contributions comme celles de Donald R. Songer (Songer, 2008) et celle d'Ostberg et Wetstein précitée (Ostberg, 2007) montrent que l'intégration des outils informatiques et statistiques à l'analyse empirique des décisions judiciaires permet d'évaluer l'influence des facteurs politiques, régionaux ou sociodémographiques sur le comportement des juges, ouvrant la voie à une étude sociojuridique quantitativement plus nuancée et robuste (Weiden, 2008).

Plus récemment, les dynamiques de recherche en droit mobilisant des outils informatiques spécialisés, notamment pour l'analyse de la jurisprudence, se sont largement automatisées et massifiées grâce aux outils de *text mining*, d'apprentissage automatique, et d'analyses de réseaux. Dyevre (Dyevre, 2020) illustre cette évolution avec l'application de techniques de *machine learning* pour détecter des motifs discursifs et des structures de raisonnement dans de vastes corpus juridiques, rendant possibles des analyses prédictives et comparatives à grande échelle. Mathias Siems (Siems, 2024) adopte une analyse bibliométrique – méthode quantitative qui repose sur le traitement et l'examen systématique de données bibliographiques – afin de retracer l'évolution du terme « *corporate purpose* » dans la littérature académique. Cette méthode lui permet de retracer la fréquence d'utilisation du terme depuis les années 1960 jusqu'à aujourd'hui, de questionner sa pertinence comme concept juridique précis au regard de son ambiguïté, et d'identifier son glissement sémantique, notamment vers les enjeux sociaux et environnementaux des entreprises.

Dans un autre domaine du droit, Wolfgang Alschner, Julia Seiermann, et Dmitriy

Skougarevskiy (Alschner et al., 2018) ont développé un corpus structuré intitulé « *Text of Trade Agreements* », qui permet une analyse systématique des accords commerciaux préférentiels et favorisent l'étude quantitative de ces instruments commerciaux internationaux, ouvrant la voie à des recherches sociojuridiques plus fines qu'avec des méthodes traditionnelles reposant sur l'examen de documents imprimés ou de logiciels génériques. Ces travaux montrent que les outils spécialisés offrent à la fois un degré de précision, mais aussi l'identification de tendances et de relations structurelles dans le droit. Dès 2017, Alschner soulignait que la multiplication des bases de données juridiques structurées et l'essor des méthodes de traitement automatisé des textes transformaient la discipline en ouvrant la voie à de nouvelles formes de modélisation, de cartographie et de prévision des dynamiques normatives à l'échelle internationale (Alschner et al., 2017).

Enfin, ces recherches favorisent l'interdisciplinarité et la reproductibilité en combinant droit, sciences sociales, et data science. L'usage de logiciels spécialisés ou de scripts programmés permet non seulement de reproduire les analyses, mais aussi d'appliquer les méthodes à d'autres corpus ou juridictions, renforçant ainsi la rigueur scientifique et la comparabilité des résultats par rapport aux approches traditionnelles. Cette évolution méthodologique contribue à transformer l'étude du droit en un champ où les résultats sont plus systématiques, cumulables et vérifiables, favorisant un dialogue scientifique international.

La dynamique actuelle, comme le montre cette revue de littérature, reste largement portée par la doctrine anglophone, qui a su s'approprier ces outils et structurer un corpus théorique et empirique dédié à l'analyse de la jurisprudence de pays en majorité anglo-saxons. La présente contribution s'inscrit donc dans la perspective de combler cette lacune méthodologique et théorique de la littérature francophone (voir cependant Paquin & Alschner, dans (Gesualdi-Fecteau & Bernheim, 2022)), en proposant une synthèse et une mise en œuvre de ces approches adaptées aux spécificités du contexte juridique et académique francophone.

Parmi les précurseurs de ces approches en France, la revue *Jurimétrie* présente quelques travaux d'analyse lexicale (notamment Simiand, 2024). Cependant, la majorité des travaux outillés informatiquement portent encore dans le monde francophone sur des facteurs externes au texte juridique (financement de la recherche, réseaux de juges, etc.). On en trouve un bon exemple dans un article consacré à l'étude de l'utilisation de la QPC pour les dix ans de l'ouverture de cette voie de droit (Acar et al., 2021). Si le texte fait l'objet de l'analyse, son contenu a été codé de manière assez traditionnelle avant d'être traduit à nouveau statistiquement.

De manière assez représentative, un projet d'ampleur, le projet Justice Algorithmique des Élections (JADE) porté par l'Université Grenoble Alpes, s'attache à caractériser par des algorithmes les déterminants externes d'une décision. Pour le moment, les analyses réalisées sont donc encore tributaires d'une analyse externe : on caractérise une décision sans s'appuyer sur le texte de la décision, ou en essayant d'expliquer son issue par des éléments non juridiques. L'usage du *text mining* reste exploratoire et accessoire, dans la présentation du projet en ligne (Rambaud et al., 2024*)*.

Les membres du projet JADE ont utilisé des jeux de données disponibles en accès ouvert et les ont combinés au sein d'une base de travail particulière. Le projet JADE a fait le choix de diffuser la base constituée, plutôt que les outils ayant permis sa constitution. Cela a d'ailleurs fait l'objet d'un *data paper* (Bligny et al., 2025). De même que nous avons choisi de plutôt diffuser les outils permettant de constituer des corpus interrogeables, nous nous distinguons

aussi de l'approche du projet JADE en faisant le pari d'analyser le texte de droit en tant que tel, pour comprendre comment certains déterminants géographiques, historiques ou culturels expliquent les particularités de discours des requérants, des juges, ou des deux conjointement.

• II. La constitution d'une base de données interrogeable à partir d'un corpus en ligne : le cas de la *Climate Change Litigation Database*

A. Les données compilées initialement par le Sabin Center

La compilation des différents textes étudiés a été réalisée par le Sabin Center, institut de recherche ayant pour but de développer la recherche juridique sur les questions climatiques, afin d'outiller les juristes sur les questions liées à l'environnement. La base de données issues des documents compilés par le Sabin Center se présente comme un corpus de textes identifiés par une série de métadonnées d'abord contextuelles, puis sémantiques.

Le Sabin Center met en place et entretient un réseau de reporters nationaux chargés de signaler à l'organisme américain les affaires climatiques portées dans les juridictions dont ils ou elles assurent le suivi. Celles-ci sont ensuite compilées au sein d'une base de données diffusée librement sur le site (voir Bibliographie), avec les documents afférents et un bref résumé produit par les membres du Sabin Center. Ce dernier a ainsi constitué un réseau d'évaluateurs et évaluatrices des contentieux climatiques, rassemblant des praticiens et des membres du milieu universitaire du monde entier. Ces experts contribuent à l'examen de la base de données mondiale sur les contentieux climatiques, afin de garantir son exhaustivité et sa mise à jour régulière. Selon le Sabin Center, ce réseau favorise la mise en relation de chercheurs partageant des intérêts communs, et à encourager des échanges sur les argumentaires juridiques développés dans le cadre de ces litiges (Sabin Center for Climate Change Law, n.d.-a). La base de données relative aux contentieux initiés aux États-Unis est quant à elle élaborée et mise à jour en collaboration avec le cabinet d'avocats américain Arnold et Porter LLP (Sabin Center for Climate Change Law, n.d.-b).

Le Sabin Center met à disposition l'ensemble des documents et des métadonnées, à travers un portail internet qui ne permet pas la recherche à l'intérieur des documents mais seulement via des métadonnées et résumés produits au sein du Sabin. Notre travail a consisté à reprendre ces métadonnées, les trier, et dans certains cas rationaliser, puis à leur adjoindre, pour les documents le permettant, une extraction du texte afin de le rendre exploitable. En cela, elle s'approche de méthodes de *web scraping* déjà expérimentées et bien décrites dans la recherche en Sciences sociales (voir par exemple (Michon & Wiest, 2021)), mais très peu en droit.

Voici une liste des métadonnées disponibles sur le site du Sabin Center :

1. Au niveau de l'affaire :
1.1. titre: Titre de l'affaire en question
1.2. status: Statut (encore en cours ou non)
1.3. filing_date: Date de dépôt de l'affaire
1.4. reporter_info: Référence de l'affaire dans le système en question
1.5. jurisdiction: Juridiction devant laquelle l'affaire a été portée
1.6. jur_pays: Pays de ladite juridiction
1.7. jur_court: Tribunal en question
1.8. principal_law: Loi, Texte ou Régulation en question dans l'affaire (de manière générale)
1.9. pl_pays: Pays où s'applique ladite loi
1.10. pl_law: Nom précis de la loi
1.11. taxonomy: Classification du Sabin par rapport à la thématique principale de l'affaire, selon plusieurs catégories
1.12. theme_precis: Classification plus précise de l'affaire, le cas échéant
1.13. summary: Résumé réalisé par les membres du Sabin des principaux tenants et aboutissants de l'affaire.
2. Au niveau du document :
2.1. date: Date d'édition du document
2.2. type: Type du document (jugement, décision, plainte, etc.)
2.3. desc: Description ou résumé du document.

Ces différentes métadonnées s'appliquent à la fois aux affaires, qui peuvent comporter plusieurs documents, et aux documents pris individuellement. Dans la base de données que nous avons créée, l'échelle retenue est celle du document : ainsi, toutes les métadonnées présentées y figurent, d'abord celles ayant trait aux affaires dont sont tirés les documents, puis celles ayant trait à chaque document, avant le texte intégral de chaque document.

Plusieurs de ces variables sont en réalité des recodages de métadonnées du Sabin (ainsi les variables « pl_pays » et « pl_law » sont-elles tirées de la variable « principal_law » qui contient les deux informations). Malheureusement, comme le réseau de reporters du Sabin n'a pas mis en place de vérification a priori des saisies, le niveau général d'harmonisation, bien que correct, n'est pas parfait, et plusieurs informations devraient être recodées afin d'obtenir une base parfaitement propre.

B. Les besoins pour la recherche d'une base de données plus complète et plus interactive

Le champ de recherche des contentieux climatiques se prête particulièrement à l'analyse documentaire par fouille de texte. Le Sabin Center définit les contentieux climatiques, ou procès pour le climat, comme tout type d'actions qui « soulèvent des questions substantielles de droit

ou de fait concernant la limitation des effets du changement climatique, l'adaptation à celui-ci ou les aspects scientifiques du changement climatique » (Sabin Center for Climate Change Law, n.d.-c, traduction libre). Ce phénomène a émergé aux États-Unis dans les années 1980, puis dans le reste du monde dans les années 2000. Il n'a cessé de prendre de l'ampleur, leur nombre ayant presque triplé au niveau mondial depuis 2015, passant de 800 affaires initiées entre 1986 et 2015 à plus de 2 000 depuis, pour un total de 3 081 affaires référencées au 1er septembre 2025. L'identification des contentieux climatiques entrant dans la définition susvisée est facilitée par le travail du Sabin Center, dont la base de données met à disposition des documents provenant de nombreux litiges climatiques répertoriés à travers le monde.

L'analyse de ces litiges, qu'ils soient portés devant des juridictions nationales ou internationales, nécessite la recherche et la collecte de documents souvent très volumineux, parfois longs de plusieurs centaines de pages. Le Sabin Center en comptabilisait plus de 15 000 au 1er septembre 2025. Ces documents proviennent de procédures juridiques diverses et sont rédigés dans plusieurs langues (la langue officielle du pays dont émane la juridiction en cause). Ils peuvent émaner des parties au litige, prenant alors la forme de plaidoiries, mémoires ou soumissions, ou bien provenir d'experts, sous forme de rapports techniques, lorsque ceux-ci sont rendus publics. Enfin, la base est principalement composée de documents rédigés par les juridictions elles-mêmes, sous la forme de décisions de justice, qu'il s'agisse de jugements de première instance, d'arrêts rendus en appel ou en cassation, visant à trancher le différend de manière définitive ou à titre provisoire.

La base de données du Sabin Center permet un filtre très limité des affaires identifiées, permettant essentiellement d'identifier le nombre d'affaires rendues par pays ou juridiction. Cependant, il n'est pas possible de mener une analyse systématique des contentieux identifiés par le Sabin Center, ni conduire des recherches qualitatives ou quantitatives à partir de cette base de données. Cette structure restreint la possibilité de conduire une véritable recherche empirique. Il est, par exemple, impossible d'examiner l'évolution dans le temps des arguments mobilisés par les requérants, d'identifier ou cartographier les tendances jurisprudentielles du contentieux climatique, ou encore d'analyser la fréquence de certains fondements juridiques ou le recours à l'expertise selon les types de juridictions ou de litiges. Les fichiers CSV fournis par le Sabin Center ne permettent pas non plus cette analyse. En effet, ceux-ci ne présentent que les métadonnées immédiatement accessibles sur le site, sans aller chercher l'intégralité des textes.

L'outil présenté dans cet article a été élaboré pour pallier ces limites. Régulièrement actualisé afin d'intégrer les nouvelles décisions issues des contentieux climatiques recensés, cet outil permet d'effectuer des recherches textuelles sur l'ensemble de la base. Il facilite notamment l'identification de mots-clés spécifiques, tels que ceux relatifs à l'invocation du régime international du climat, que ce soit par les parties ou par les juridictions saisies. Il permet aussi l'analyse de leur fréquence et de leur répartition dans les décisions (Martini et Brunel, 2024). Parmi les autres fonctionnalités avancées de cet outil figurent l'élaboration de nuages de mots, permettant de visualiser les termes les plus récurrents dans un ensemble de décisions ; la localisation précise des occurrences dans la structure du texte afin de comprendre l'importance de leur positionnement dans la décision du juge (introduction, rappel de la position des parties, dispositif du jugement, etc.), ainsi que la comparaison lexicale entre différentes juridictions, périodes ou catégories de contentieux. Ces outils ouvrent la voie à des approches empiriques variées en contentieux climatique, en croisant les méthodes qualitatives d'analyse juridique ou pluridisciplinaires et celles d'exploration de données par fouille de texte.

En ce qui concerne les limites de l'utilisation de cette base, il existe un risque de biais ou

d'erreur lié notamment à la définition adoptée par le Sabin Center des litiges climatiques, qui peut exclure certaines affaires pertinentes ou au contraire inclure des litiges non pertinents. De plus, l'imprécision de certains textes, lorsque seule la traduction non officielle d'un jugement est disponible, peut compliquer l'analyse. En outre, le caractère confidentiel ou l'absence de publication des mémoires écrits des parties peut rendre difficile l'analyse de certaines décisions en raison du manque d'éléments contextuels sur les techniques d'interprétation et les sources de droit mobilisées par les requérants. Pour minimiser le risque de collecte de données imprécises ou incomplètes, la revue des documents publiés sur la base de données du Sabin Center doit donc être complétée par des recherches personnelles afin notamment de vérifier l'exactitude des termes employés dans chaque jugement traduit et éviter l'existence de doublons. Enfin, un risque important tient également à la fiabilité incontrôlable de la base de données produite par le Sabin : si cet organisme fait référence dans le domaine du droit international du climat, il reste seul responsable de la qualité des données qu'il diffuse. Pour donner deux exemples plus précis, nous avons déjà abordé la faible harmonisation des métadonnées au niveau de l'affaire, et pouvons ajouter la parfois faible qualité des fichiers diffusés, qui a rendu impossible dans certains cas de transcrire leur contenu automatiquement.

C. Création d'une base interrogeable : principes, étapes et résultat

Cette partie présente les principes et solutions retenus pour rendre disponible et mettre à jour le travail réalisé, afin que d'autres travaux de recherche puissent s'en saisir. La reproductibilité des méthodes est en effet une part essentielle de notre démarche.

1. Principe

La mise à disposition, sous un autre format, de corpus textuels librement disponibles sur le site du Sabin Center, première option envisagée dans le cadre de notre travail, pose problème pour plusieurs raisons. Tout d'abord, nous n'avons pas souhaité créer de doublon en dupliquant la base de données sur un entrepôt de données institutionnel.

Même si cela aurait eu le mérite de sécuriser la base en elle-même (dans un contexte où la pérennisation des travaux de recherche américains et leur libre utilisation en Europe devient de plus en plus incertaine), cette solution pose de nombreux problèmes en termes de droits, d'utilisation raisonnée des ressources et de lisibilité de la propriété des données.

De plus, une simple copie remise en forme de la base du Sabin aurait empêché deux éléments très importants de notre démarche : l'incrémentation progressive des évolutions de la base et la reproductibilité de notre travail dans d'autres contextes. En donnant accès aux résultats de notre travail sous la forme d'un corpus de textes n'ayant pas été compilé par nos soins, nous n'aurions finalement pas partagé avec la communauté scientifique ce qui a fait l'intérêt de notre démarche.

Nous avons donc préféré mettre à disposition les scripts ayant permis d'extraire et ordonner la base du Sabin Center, puis ceux permettant de l'analyser. Nous espérons ainsi qu'au-delà de la simple utilisation des données du Sabin Center, c'est la démarche dans son ensemble qui pourra être plus facilement appropriable et mobilisable dans d'autres contextes.

2. Architecture des éléments déposés en accès ouvert

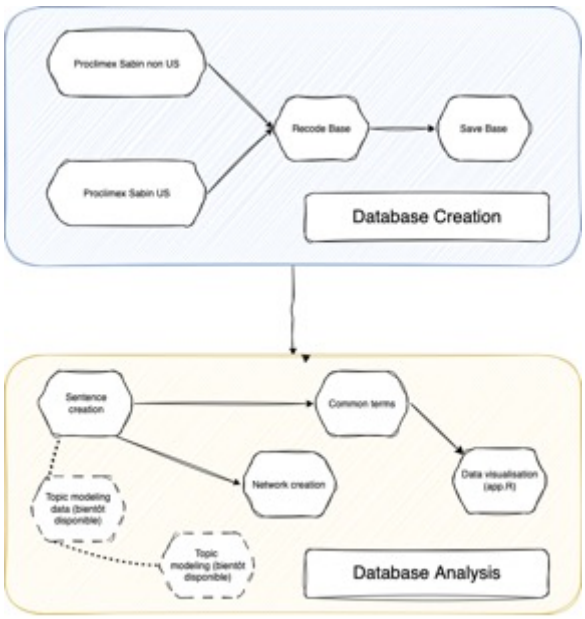
La mise à disposition préférée prend donc la forme d'un dépôt organisé au sein d'une forge logicielle (GitLab a été favorisé, par facilité professionnelle, au sein de la collection hébergée par la MSH SUD, unité d'affectation de l'un des auteurs). Voici le lien vers le dépôt : https://forge.ird.fr/mshsud/proclimex/-/tree/valentin_clean?ref_type=heads

Ce dépôt, adopte les conventions Git (un logiciel de gestion des versions de fichiers décentralisé) et se compose de deux parties principales. La première a trait aux étapes de constitution de la base de données par extraction des informations présentes en ligne. Cette partie du dépôt répond aux besoins de nombreux projets de recherche en droit : la présence massive de corpus juridiques en ligne sous forme de portails interrogeables rend l'utilisation de scripts de création automatisée de corpus potentiellement très large.

La deuxième partie du dépôt concerne les scripts d'analyse des données. Elle se veut également mobilisable dans des cas divers. Plusieurs scripts différents permettent ainsi de travailler sur la recherche d'occurrences, les associations de mots ou encore les thématiques sous-jacentes à des parties du corpus.

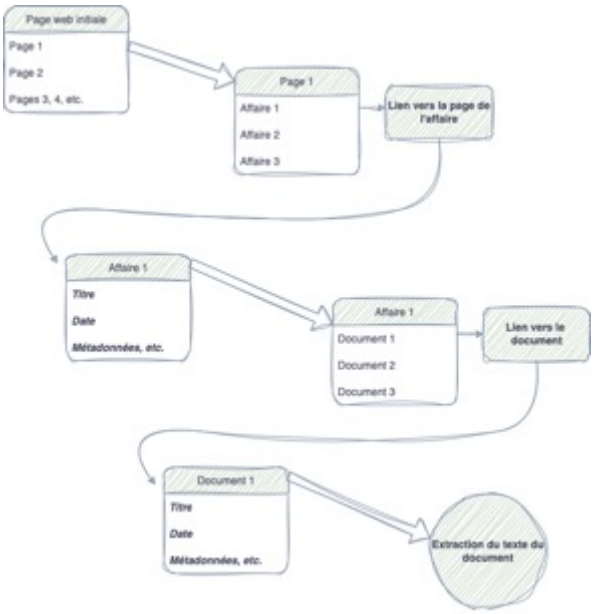
La première partie (*database_creation*) comprend donc quatre scripts distincts. Ces scripts permettent pour les deux premiers d'extraire les informations à partir du site du Sabin (base « US » et base « non US » ou « internationale »), pour le troisième de recoder les informations obtenues et enfin pour le dernier de les exporter. Ces scripts ont été constitués à partir de plusieurs packages Python (*spacy*, *pdfreader* notamment, en plus de *beautiful soup* pour les principaux (Montani et al., 2023; *pdfreader*, s. d.; Richardson, 2007).

Voici un schéma reprenant l'architecture globale du dépôt tel qu'il se trouve en ligne (branche valentin_clean sur Git) :



Graphique 1 : Schéma fonctionnel reprenant les interactions imaginées entre les scripts

Le schéma suivant reprend le contenu du premier script, qui permet l'extraction de données automatisée. Les éléments en gras et en italique se retrouvent en colonne dans la base de données constituée.



Graphique 2 : Schéma illustrant le fonctionnement des scripts de création de données

Sans entrer dans le détail des scripts présents sur le dépôt en accès ouvert et documentés, il reste possible d'expliciter un point de détail, par exemple sur la méthode du *scraping*.

Littéralement, l'inscription présentée en détail ci-dessous (Graphique 3) demande pour chaque lien vers une affaire (*for i in affaires*), d'aller ouvrir l'url associée, puis d'en lire correctement les informations (*BeautifulSoup*). On sélectionne ensuite le contenu à extraire, déclaré par des balises (*.entry-title* pour le titre, *.entry-taxonomy.non_us_case_category* pour la taxonomie), puis on ajoute ensuite à une liste adéquate (*titre.append*) le contenu de ce qu'on a extrait, en prenant soin d'ajouter une mention "*No taxonomy*" par exemple si le contenu est vide (afin de ne pas décaler la liste). L'élément « translator » permet de nettoyer le texte des caractères non souhaités. Les deux listes obtenues, titre et taxonomie, rempliront les deux premières colonnes de notre jeu de données.

```
for i in affaires:
    try:
        reqh = Request(i)
        html_pageh = urlopen(reqh)
        souph = BeautifulSoup(html_pageh, "lxml")

        titreh = souph.select('.entry-title')
        taxonomyh = souph.select('.entry-taxonomy.non_us_case_category')

        for j in titreh:
            titre.append(j.get_text())

        if bool(taxonomyh) is False:
            taxonomy.append("No taxonomy")
            print(str(affaires.index(i))+" no tax")
        else:
            for l in taxonomyh:
                taxonomy.append(l.get_text().translate(translator))

    #print(i)
    except:
        summary.append("NA")
        titre.append("NA")
```

Graphique 3 : Capture d'écran d'une partie du script d'extraction de données

Le principe est le même pour chaque métadonnée au niveau de l'affaire, et ensuite pour chaque métadonnée au niveau document (on procède alors toujours affaire par affaire, puis on sélectionne document par document).

Enfin, un dernier outil permet d'extraire le texte d'un document .pdf, à partir du lien du document en question. Ce processus est assez coûteux en temps et en puissance de calcul (chaque mise à jour complète de la base prenant environ une demi-journée).

Les deuxième et troisième scripts sont consacrés à harmoniser et nettoyer les métadonnées, puis à constituer la base globale et l'exporter. Il s'agit d'étapes assez classiques sur lesquelles beaucoup de documentation est déjà disponible. Les scripts consacrés à l'analyse peuvent maintenant être détaillés dans une troisième partie.

• III. Utilisation dans le cadre de la recherche en droit

Les travaux présentés dans le cadre de cet article s'appuient sur un corpus bien structuré de textes juridiques afin de réaliser des opérations de comptage, de repérage et d'identification de segments de texte pertinents. Ces opérations sont ensuite accompagnées d'autres opérations plus complexes visant à présenter les parties de texte en fonction de leur environnement, directement (par le biais d'un réseau de mots) ou plus indirectement (via la classification automatisée, ou *topic modeling*). Cette partie présentera rapidement ces opérations, de la plus simple à la plus complexe, afin de montrer l'usage qui peut en être fait dans le cadre de la recherche juridique.

A. Quelques usages

Les scripts présentés dans le dossier "analyse" constituent des routines permettant de récupérer, au sein de la base de données, un certain nombre de segments de texte en fonction de métadonnées ou de leurs contenus, puis de les analyser.

a) a) Tri par métadonnées

La première fonction de la présentation du corpus en base de données est de réaliser des tris des textes en fonction de leurs métadonnées. Plus concrètement, cela signifie qu'il devient possible de compter les documents et les affaires en fonction de leur thématique, de leur nombre de documents, de leur juridiction ou encore de la date de production de ces documents. De manière plus originale, cela permet également de combiner ces indicateurs. Cela est généralement peu fait dans la recherche juridique, car les bases de données ne présentent dans le meilleur des cas les affaires ou les documents qu'en fonction de quelques filtres. S'il est techniquement possible de filtrer les documents en ligne, compter le nombre de résultats, et répéter l'opération pour un autre filtre, un tel procédé n'est guère ergonomique. La base de données constituée par extraction automatisée permet en revanche de facilement comparer les catégories entre elles.

Un exemple en est fourni par les travaux réalisés sur les documents mentionnant l'Accord de Paris, en fonction de leurs auteurs, de leur date et du type d'affaires (Martini & Brunel, 2024). Dans le cadre de cette communication, les pourcentages croisés ont permis de montrer que l'Accord de Paris était relativement plus mobilisé par les requérants que par les juges, et plus mobilisé dans les affaires auprès des États qu'auprès des entreprises. Pour cela, le comptage et la répartition par métadonnées ont été nécessaires mais non suffisants, il a fallu également trier les documents suivant leur mention ou non de mots-clefs.

b) b) Exploration par mots-clefs

Le deuxième intérêt de la présentation de la base en corpus permet une exploration des phrases ou groupes de phrases contenant certains mots-clefs. Ainsi, de la même manière qu'une revue de littérature ou qu'une recherche bibliographique classique en droit permettrait d'identifier les décisions ou documents pertinents, les outils proposés identifient par une recherche en plein texte les phrases contenant certains syntagmes et permettent ensuite de les compter ou comparer.

Toujours en suivant l'exemple précédent, la sélection de documents comprenant certains mots-clefs a permis de comparer leur utilisation. De manière plus générale, il est permis ensuite de calculer des co-occurrences de mots-clefs les uns avec les autres au sein d'une même phrase, ou bien au sein d'un même ensemble. Pour l'Accord de Paris, nous avons pu comparer son utilisation à celle d'autres instruments, comme le Protocole de Kyoto.

Ainsi, il a fallu regrouper en anglais les occurrences de "*Paris Agreement*", "*Paris Accord*", ou encore "*Paris Framework*". En français, nous avons dû regrouper les occurrences de l' « Accord » ou des « Accords de Paris ». Lors de l'analyse des données textuelles qui précède la collecte et la compilation des extraits, il est également possible d'identifier la forme lemmatique d'une occurrence (verbe non conjugué, nom singulier si elle est au pluriel, adjectif sous forme générique). On retrouvera les résultats de ce type d'approche dans l'interface en ligne développée dans le cadre du projet (voir plus bas).

De manière générale, les outils d'exploration linguistique (tels que la librairie *Spacy* pour Python, dans notre cas) permettent d'opérer de très nombreux choix et tris ou filtres. Ainsi, nous pouvons donner les verbes les plus fréquents d'un corpus, les noms, les adjectifs, reconstituer le schéma de dépendance linguistique au sein d'une phrase, identifier les entités nommées (noms propres) par famille (nom, entreprise, lieu, etc.), ou encore (ce dernier cas étant moins propice à l'analyse juridique) prédire le sentiment positif ou négatif d'un texte. Afin d'aller plus loin dans la description de ces possibilités explorées au cours de notre recherche, nous présenterons l'exemple des réseaux de mots (l'exemple des mots les plus communs étant présent au sein du dépôt).

c) c) Réseau de mots

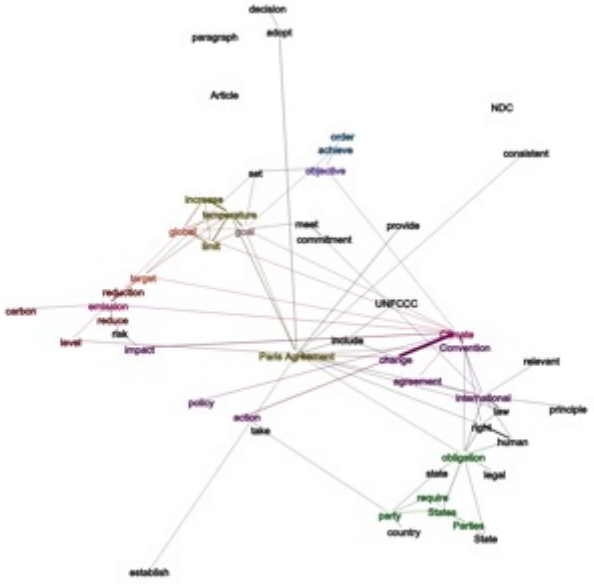
Parmi les nombreuses possibilités ouvertes par les modèles d'analyse du langage, nous avons entrepris de réaliser des réseaux de mots dans des sous-parties du corpus ou dans le corpus dans son intégralité. Avec un certain nombre de fonctions paramétrables, il est possible de présenter l'ensemble des mots comme des nœuds et leur proximité relative les uns par rapport aux autres au sein du corps comme des liens.

Cette approche permet de comprendre quel est l'environnement sémantique immédiat de certains termes, afin de vérifier par exemple si certaines notions juridiques sont associées à certains exemples ou certaines autres notions. Un exemple déjà présenté dans une conférence est disponible dans un article sur la notion d'ignorance en droit (Brunel & Delage, 2025).

Ainsi, toujours concernant l'Accord de Paris, nous avons pu établir une cartographie des termes qui lui étaient associés au sein de notre corpus, et ce pour plusieurs sous-corpus

différents. Cela permettrait par exemple en poussant la démarche de montrer de quelle manière les termes associés à l'Accord de Paris diffèrent suivant le statut de l'auteur (juge ou requérant).

Voici pour illustration un réseau représentant, pour le corpus international, les termes principaux associés au terme « Paris agreement ». On y remarque plusieurs groupes de mots, l'un ayant trait aux émissions de gaz à effets de serre, l'autre aux obligations des États, un autre encore sur les objectifs à atteindre. Ce réseau de mots permet ainsi de montrer facilement dans quel contexte sémantique intervient l'usage d'une formule ou d'un terme.



Graphique 4 : Réseau autour des termes « Paris agreement », Sabin International

Légende : Réseaux des mots cooccurrents présents au sein du corpus Sabin International, pour toutes les phrases mentionnant « Paris agreement ». Les mots filtrés sont les suivants : premier filtre pour tous les liens d'un poids supérieur ou égal à 20 (ne sont conservés que les mots apparaissant ensemble plus de 20 fois), puis deuxième filtre pour ne garder que les mots d'un degré supérieur ou égal à 150 (les mots apparaissant 150 fois en tout dans le corpus). Les couleurs sont des colorations de mots par proximité au sein du graphe de réseau.

d) d) Le *Topic modeling*

Un dernier exemple d'utilisation de la base de données concerne l'analyse thématique des textes dans le corpus. Le principe de la méthode de *topic modeling* (ou classification thématique) est de repérer les thématiques présentes au sein d'un corpus en isolant plusieurs termes intervenant souvent ensemble, d'autres, puis de contraster ces thématiques en les positionnant les unes par rapport aux autres.

La technique est donc hautement inductive en ce qu'elle fait remonter, à partir de

corrélations ou co-occurrences au sein du texte, des termes représentant ainsi des thématiques, sans préjuger de leur pertinence ou de leur signification. De manière complémentaire, elle peut se révéler assez imprévisible. Un exemple d'application de cette technique a été présenté dans un article déjà cité portant sur l'utilisation de la notion d'ignorance au cours du temps, qui revient sur les principales caractéristiques et outils mobilisés pour appliquer le *topic modeling* à un corpus (Brunel & Delage, 2025).

Pour le moment, le *topic modeling* n'a pas été employé dans le cadre des données textuelles issues du Sabin Center, mais il est envisagé de l'utiliser dans les mois à venir.

B. La création d'une interface

En plus de l'usage des scripts présenté plus haut, a été construite une interface Shiny visant à faciliter l'exploration de l'utilisation de termes spécifiques au sein du corpus. Cette interface répondait à un besoin concret de prise en main facilitée des résultats obtenus. Elle permet d'explorer la présence de syntagmes définis au sein du corpus de textes, à l'aide de plusieurs outils complémentaires. Son code est disponible au sein du dépôt sur la forge IRD, mais l'interface en elle-même se trouve à l'adresse suivante : https://analytics.huma-num.fr/Valentin.Brunel/proclimex/keywords_1/

Un premier outil comprend le comptage total d'occurrences du syntagme au sein du corpus. Le second représente la part de documents en fonction du temps et du statut de l'auteur (juge ou partie). Le troisième permet de lire ces occurrences associées à leurs métadonnées. L'interface est aussi complétée par une série de filtres simples concernant la date, le corpus et le statut de l'auteur afin de faciliter la recherche des occurrences intéressantes. Enfin, l'interface permet de compter les fréquences relatives par affaires ou par documents au sein desquel(le)s apparaissent les occurrences.

Publiée en tant que pilote via le serveur Shiny de l'IR* HumaNum, cette interface a permis de faciliter le travail d'identification des occurrences et des phrases au sein desquels certains mots-clefs apparaissaient. La comparaison des fréquences relatives d'apparition de certains mots-clefs a aussi pu être source d'interrogations (ainsi de la différence avec le temps entre 1,5° et 2°). Il est prévu un nettoyage du code de l'interface, et sa mise en production plus large dans les mois à venir.

IV. Conclusion

Cet article présente une série de dispositifs et d'intuition de recherche visant à faciliter

l'appropriation par les chercheurs et chercheuses en droit des outils informatiques. A rebours de la plupart des approches quantitatives du droit en France, les outils présentés dans cet article touchent au cœur de la science juridique : le texte en lui-même. Toutefois, loin de se substituer à l'analyse juridique plus traditionnelle, l'approche présentée vise à lui apporter un complément sous forme de focale différente.

Approcher les textes juridiques par l'analyse automatisée permet d'observer des régularités, déceler des tendances, objectiver des ressentis. Ces derniers, pour être objectivés, doivent donc dans un premier temps être décelés, ressentis, par une analyse fine du texte et une excellente connaissance du sujet. En cela, nous nous inscrivons en faux contre tout providentialisme technologique : il s'agit simplement de proposer à qui le souhaite de nouveaux verres, qu'il est toujours possible d'ôter, et qui ne remplaceront jamais les yeux.

Enfin, les travaux présentés ici sont encore, à bien des égards, des recherches en cours. Il nous a paru important de les montrer et les défendre dans cet état, afin de susciter avant toute fossilisation prématurée discussions et débats, le sel de la recherche.

• V. Bibliographie :

· Acar, T., Champeil-Desplats, V., Gelblat, A., & Hennette Vauchez, S. (2021). Physionomie générale du corpus QPC et méthodologie de la recherche. La Revue des droits de l'homme. Revue du Centre de recherches et d'études sur les droits fondamentaux, 20, Article 20. https://doi.org/10.4000/revdh.12680

· Alschner, W., Pauwelyn, J., & Puig, S. (2017). The Data-Driven Future of International Economic Law. Journal of International Economic Law, 20(2), 217-231. https://doi.org/10.1093/jiel/jgx020

· Alschner, W., Seiermann, J., & Skougarevskiy, D. (2018). Text of Trade Agreements (ToTA)—A Structured Corpus for the Text-as-Data Analysis of Preferential Trade Agreements. Journal of Empirical Legal Studies, 15(3), 648-666. https://doi.org/10.1111/jels.12189

· Blackham A. (2022). When law and data collide : The methodological challenge of conducting mixed methods research in law. J. Law Soc, 49(Suppl. 1), Article Suppl. 1. https://doi.org/10.1111/jols.12373

· Bligny, C., Letué, F., Martinez, M.-J., Rambaud, R., & Hafsaoui, A. (2025). Cross-Referenced Data on Electoral Disputes and French Legislative Election Results. Journal of Open Humanities Data, 11(1). https://doi.org/10.5334/johd.315

· Brunel, V., & Delage, A. (2025). De l'ignorance du droit au droit de l'ignorance ? Cahiers Droit, Sciences & Technologies, 19, Article 19. https://doi.org/10.4000/13weu

· Dyevre, A. (2020). Text-mining for Lawyers : How Machine Learning Techniques Can Advance our Understanding of Legal Discourse (SSRN Scholarly Paper No. 3734430).

Social Science Research Network. https://doi.org/10.2139/ssrn.3734430

· Fausten, D., Nielsen, I. L., & Smyth, R. (2007). A Century of Citation Practice on the Supreme Court of Victoria (SSRN Scholarly Paper No. 995060). Social Science Research Network. https://doi.org/10.2139/ssrn.995060

· Gesualdi-Fecteau, D., & Bernheim, E. (2022). La recherche empirique en droit : Méthodes et pratiques. Themis.

· McCormick, P. (2009). American Citations and the McLachlin Court : An Empirical Study. Osgoode Hall Law Journal, 47(1), 83-129. https://doi.org/10.60082/2817-5069.1163

· McCormick, P. J. (2015). The End of the Charter Revolution : Looking Back from the New Normal. University of Toronto Press. https://www.jstor.org/stable/10.3138/j.ctv102bjxk

· Michon, S., & Wiest, E. (2021). A database about the Members of European Parliament : Contributions and limitations of automated data collection in the study of European political elites. BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique, 152, 125-141.

· Montani, I., Honnibal, M., Honnibal, M., Boyd, A., Landeghem, S. V., & Peters, H. (2023). explosion/spaCy : V3.7.2 (Version v3.7.2) [Logiciel]. Zenodo. https://doi.org/10.5281/zenodo.10009823

· Mulcahy L. & Wheeler S. (2020). 'Couldn't You Have Got a Computer Program to Do That for You?' Reflections on the Impact that Machines Have on the Ways We Think About and Undertake Qualitative Research in the Socio-Legal Community. Journal of Law and Society, 47(1), Article 1. https://doi.org/10.1111/jols.12217

· Ostberg, C. L. (avec Wetstein, M. E.). (2007). Attitudinal decision making in the Supreme Court of Canada (1st ed.). UBC Press. https://doi.org/10.59962/9780774855846

· pdfreader : Pythonic API for parsing PDF files (Version 0.1.15). (s. d.). [Python; MacOS :: MacOS X, POSIX]. Consulté 13 décembre 2024, à l'adresse http://github.com/maxpmaxp/pdfreader

· Rambaud, R., Bligny, C., Letué, F., Martinez, M.-J., Cottin, S., Camby, J.-P., Prunier, G., Girard, D., Hafsaoui, A., & Deschamps, K. (2024). Projet Justice algorithmique des élections (JADE) : Une analyse statistique de la jurisprudence du Conseil constitutionnel relative aux élections législatives (4 oct. 1958 – 1er avril 2024):Partie 1/2 – La structuration générale du contentieux AN. Revue française de droit constitutionnel, 140(4), 863-892. https://doi.org/10.3917/rfdc.140.0863

· Richardson, L. (2007). Beautiful soup documentation. April.

· Sadl U., O. H. P. (2017). Can Quantitative Methods Complement Doctrinal Legal Studies ? Using Citation Network and Corpus Linguistic Analysis to Understand International Courts. Leiden Journal of International Law, 30, 327-349.

https://doi.org/10.1017/S0922156517000085

· Siems, M. (2024). « Corporate Purpose » as a False Friend : A Bibliometric Analysis (SSRN Scholarly Paper No. 5046087). Social Science Research Network. https://doi.org/10.2139/ssrn.5046087

· Songer, D. R. (2008). The Transformation of the Supreme Court of Canada. University of Toronto Press. https://doi.org/10.3138/9780802096890

· Weiden, D. L. (2008). Attitudinal Decision Making in the Supreme Court of Canada, C. L. Ostberg and Matthew E. Wetstein, Vancouver, BC : University of British Columbia Press, 2007, pp. 288. Canadian Journal of Political Science/Revue Canadienne de Science Politique, 41(4), 1027-1028. https://doi.org/10.1017/S0008423908081171