Systèmes d'information pour les chercheurs en droit

Alex Chauvet, Annie Foret

alex.chauvet@u-bordeaux.fr, univ. Bordeaux
foret@irisa.fr, univ. Rennes 1 & IRISA

Convergences du Droit et du Numérique, 2017

Résumé

Cette étude concerne les systèmes d'information en droit et propose un nouveau prototype. Nous discutons les systèmes d'information actuellement utilisés par les chercheurs en droit, en précisant leurs limites. Nous présentons les principes d'un nouveau prototype pour un meilleur système. Ce travail s'accompagne d'une première réalisation concrète, un système à facettes sémantiques, résultat de notre chaîne de traitement sur un ensemble de décisions du Conseil Constitutionnel.

Contenu :

1 Introduction

La recherche juridique est aujourd'hui largement fondée sur un positivisme : il s'agit d'étudier le droit tel qu'il est. Il s'ensuit que les études menées sont non seulement assises sur les textes mais surtout sur leur application. A ce titre, les arrêts et décisions de justice occupent une place prédominante dans les sources des travaux.

Cependant, ce prisme contentieux n'est pas nouveau. Ainsi, l'importance de la compilation des décisions de justice pour en favoriser l'étude était déjà à l'origine de la création de recueils de jurisprudence dont certains ont allègrement dépassé le siècle d'existence (recueil Lebon, Bulletin de jurisprudence de la Cour de cassation).

Pour autant, la problématique est aujourd'hui renouvelée par au moins deux facteurs : l'explosion du nombre de décisions depuis une trentaine d'années[1] ; l'ouverture et l'accès croissant aux décisions de justice. Il y a donc une aubaine pour la recherche mais aussi une source de difficultés qu'elle n'a pas encore su appréhender de manière satisfaisante. S'il existe des systèmes d'informations compilant la jurisprudence, ceux utilisés par les chercheurs en droit sont largement perfectibles et demeurent limités dans leurs possibilités (voir Figure 4). En particulier, ils n'exploitent pas suffisamment les caractéristiques de ces textes juridiques qui se prêtent pourtant assez bien à un traitement informatique, surtout dans un pays où la motivation des décisions de justice est caractérisée par une rédaction sèche et reprenant la forme d'un syllogisme. En effet, plus que des énoncés littéraires, ils sont des propositions articulant des informations (énoncés porteurs de structures, redondance des termes et formules stéréotypées, marqueurs du discours, articulation logique).

Pourtant, la recherche juridique n'a pas investi ce champ consistant à produire elle-même ses bases de données et systèmes d'information. Elle ne produit complètement ni ses données ni les outils permettant de les exploiter. Sur ce point, elle s'oppose par exemple à l'économie, qui est pourtant une science humaine voisine et qui a longtemps été enseignée avec le droit. Ainsi, les chercheurs en droit subissent les systèmes d'information plus qu'ils ne participent à leur conception, alors même que ceux-ci ne sont pas toujours adaptés. Les raisons sont multiples. La première réside sans doute dans un déficit de compétence et de formation aux techniques afférentes. Mais plus fondamentalement, la doctrine juridique n'est pas assez sensibilisée sur les potentialités de tels outils. En particulier, elle ne perçoit pas forcément que ceux-ci ne sont pas seulement un moyen d'archiver des arrêts et décisions, mais aussi un moyen de les présenter selon une perspective. Ces systèmes peuvent éventuellement permettre une superposition des perspectives doctrinales sur un même objet. Par exemple, une décision du Conseil constitutionnel sur une loi relative au Code du travail intéressera les spécialistes du droit constitutionnel mais aussi les spécialistes du droit du travail. Leurs visions seront probablement différentes mais complémentaires et aboutiront à des classifications différentes. En l'état, il peut être assez difficiles pour l'un des groupes d'accéder rapidement à la perspective de l'autre. Or les juristes affectionnant tout particulièrement les classifications et le droit étant de moins en moins compartimenté et plus en plus transversal, il y aurait là matière à gagner en efficacité des recherches.

Par ailleurs, d'autres arguments d'ordre méthodologique devraient aussi attirer l'attention des juristes. Ils ressortissent de biais épistémologiques trop peu souvent examinés.

Un premier biais épistémologique a trait à l'autonomie et à l'indépendance de la recherche. La plupart des systèmes d'information aujourd'hui utilisés par la doctrine sont construits par les juridictions elles-mêmes. Les chercheurs sont donc dans une relative dépendance vis à vis des organes qu'ils sont chargés d'étudier voire parfois de critiquer... Pendant longtemps, cette dépendance était nécessaire dans la mesure où la diffusion des décisions de justice était limitée voire verrouillée. Les juridictions ne donnaient vraiment à voir que ce qu'elles voulaient. Aujourd'hui, elle n'a plus de raison d'être. Ainsi, continuer d'utiliser uniquement ces systèmes d'information, c'est refuser de prendre de la distance avec ce qui constitue régulièrement un objet de recherche. La chose est connue des chercheurs en droit pour ce qui est de

[1] http://www.justice.gouv.fr/statistiques-10054/chiffres-cles-de-la-justice-10303/

l'autonomie conceptuelle. Car c'est là une spécificité du droit que d'avoir un objet qui n'est pas muet. Le soleil ne se qualifie pas lui même d'étoile ; la terre ne se désigne pas comme une planète. Or les objets juridiques prétendent très fréquemment appartenir à des catégories. Par exemple, lorsqu'une juridiction refuse de qualifier un mécanisme de peine, elle ne le fait pas forcément selon des critères objectifs et scientifiques. Elle peut le faire pour ne pas attacher à ce mécanisme le régime juridique qui va avec une peine (sa proportionnalité, son absence de caractère rétroactif notamment). Cela ne signifie pas que le chercheur ne peut pas qualifier lui ce mécanisme de peine. Aussi contre-intuitive qu'elle puisse paraître, cette autonomie des concepts est nécessaire car doctrine et juridictions (plus généralement autorités normatives) sont dans des situations radicalement différentes : la juridiction applique le droit et tout son travail de classement est guidé par cet impératif ; le chercheur vise à rendre le chaos du monde un peu plus intelligible et cohérent. Ainsi, de la même manière qu'il n'est pas scientifique de ne pas s'affranchir d'une notion produite par un juge lorsque celle-ci ne convient pas, il n'est pas scientifique non plus de continuer de se reposer sur les systèmes d'information fournis par les juridictions sans envisager de s'en affranchir. Il y a là un biais de la recherche. En effet, ces derniers ne présentent pas des données brutes mais bien des informations, avec une perspective. Celle ci est très remarquable lorsque sont reprises les classifications des anciens recueils papiers ou lorsqu'il n'est possible de filtrer les décisions que selon des propriétés limitées et choisies par le service de documentation de la juridiction.

Un autre biais épistémologique vient d'une tendance assez naturelle des études doctrinales à se focaliser sur des corpus limités de décisions. Le plus souvent, il s'agit d'étudier la jurisprudence des juridictions les plus importantes du pays (Conseil d'Etat, Cour de cassation, Conseil constitutionnel) et par forcément en intégralité. La raison est assez simple : ces juridictions contribuent plus activement à la production du droit que les juridictions inférieures dont la mission est plutôt de juger des affaires "du quotidien". A cela s'ajoute que les décisions de ces juridictions sont plus facilement accessibles car leur diffusion est mieux assurée et plus systématique. Mais procéder ainsi consiste à ne s'intéresser qu'à un spectre très restreint du droit en postulant éventuellement qu'il est identique au droit appliqué par les autres juridictions. Pour les chercheurs du début du XXème siècle, ce mode opératoire était assez justifié car il y avait alors beaucoup moins et de juridictions et de décisions. Mais aujourd'hui, compte tenu des délais, de la complexité croissante des procédures et de leur coût, nombreuses sont les affaires qui ne parviennent pas à ces juridictions. Or la meilleure connaissance des arrêts et décisions de principe des juridictions les plus prestigieuses du pays ne permet pas forcément de savoir comment une affaire sera tranchée en première instance... L'intérêt d'utiliser des systèmes d'information est alors de dépasser le champ d'étude actuel des décisions contentieuses et d'élargir le spectre : d'abord en intégrant plus largement les décisions des juridictions suprêmes, puis en y intégrant les décisions des juges "ordinaires". Ce faisant, on aurait des indications sur les masses et ordres de grandeur et une modélisation sans doute plus fidèle du droit tel qu'il est appliqué concrètement.

Dernière enjeu épistémologique : la pérennité et l'accessibilité des recherches entreprises. Dans toute étude contentieuse, la démarche du chercheur commence nécessairement par un recensement et une classification des décisions à étudier selon des critères propres à sa perspective. Actuellement, ce travail est souvent perdu à l'issue de la recherche ou ne fait l'objet que d'un tableau plus ou moins précis dans l'appareil critique. Le conserver à l'aide d'un système d'information permettrait qu'il soit directement utilisé par d'autres qui ne perdraient donc pas de temps ou beaucoup moins à retrouver le corpus de décisions pertinent. Accessoirement, elle faciliterait aussi la revue par les pairs et autoriserait le développement de méthodes quantitatives et statistiques sur les décisions de justice.

Si les perspectives sont nombreuses pour le juriste, le développement de ces outils n'en demeure pas moins une tâche difficile qui requiert donc une aide technique. A cet égard, pour les chercheurs en informatique concernés par la représentation des connaissances et le traitement automatique des langues naturelles, travailler sur un objet comme le droit permet de tester la pertinence de certains modèles et de démarches associées. Nous considérons ici l'approche des systèmes d'information logique avec l'objectif de redonner du pouvoir à l'utilisateur. Dans le domaine du droit, nous envisageons des systèmes

d'aide à un usager et une chaîne de traitement allant des textes bruts ou partiellement structurés, à des représentations formelles et à facettes logico-sémantiques, pour traiter et valoriser des données.

Au-delà de la validation d'une approche, ce type d'étude peut aussi mener à raffiner et à adapter les modèles initiaux et ceci à différents niveaux d'analyse (logico-formel, linguistique, traitement des données, interactions).

L'étude suivante se veut donc une démonstration par l'exemple. Son objet est volontairement limité et consistera à présenter des décisions du Conseil constitutionnel pour un juriste se donnant pour objet de recherche les élections législatives.

2 Illustration avec un système compilant les décisions du Conseil constitutionnel relatives aux élections législatives.

2.1 Choix effectués

Il serait illusoire de vouloir immédiatement produire un système d'information abordant un très grand nombre de décisions de justice. En effet, en dépit d'une certaine proximité dans leurs structures et formulations, celles-ci présentent une assez grande variabilité qui rend leur traitement moins facile qu'il peut paraître de prime abord. Dans un premier temps, il est donc nécessaire de faire un choix répondant à des impératifs multiples.

D'abord, puisque l'objet principal est juridique, le vocabulaire peut être un obstacle. En effet, selon le domaine concerné, les termes et expressions employés peuvent être plus ou moins abscons et constituer des facteurs de ralentissement. Le choix de la thématique est donc important. Par ailleurs, celle-ci doit autoriser une amélioration des outils existants ou au moins des ajouts propres au chercheur. Ce point n'est pas anodin car si le corpus de décisions étudié est trop hétérogène ou inversement trop homogène, il ne permet aucune systématisation et donc aucune présentation rationnelle. Là encore, il y a donc un choix à opérer, notamment de la part du juriste. Il consiste à déterminer ce qu'il faut archiver et comment. Enfin, vient le problème majeur, celui de la difficulté technique, certains critères pouvant être extraits ou récupérés de sources ouvertes, d'autres devant être complètement construits.

Dans le cadre de ce projet, le choix de la thématique s'est ainsi porté sur les élections législatives. La juridiction concernée était donc principalement le Conseil constitutionnel.

Bien que sa conception soit conforme à une utilisation standard, le site de cette juridiction peut présenter des limites pour un chercheur (voir annexe). Les données qu'il contient peuvent faire l'objet d'ajouts et d'une présentation différente de celle dont l'utilisateur est captif. En effet, sur la question des élections législatives, il apparaît rapidement que les critères formels retenus sont insuffisants. Ces élections peuvent être évoquées dans des décisions DC, QPC (pour leurs règles), AN (pour le déroulement des élections), I et D (pour les fins avant terme des mandats). De même, pour le juriste, cette classification ne permet par forcément d'accéder à l'information recherchée. Par exemple, pour les décisions I et D relatives aux incompatibilités et déchéances de mandat, il importe surtout de savoir quelle est la cause de cette fin de mandat avant terme. Souvent, il s'agit de l'exercice d'une profession ou d'une fonction incompatible. Idem pour les décisions AN relatives au contrôle de l'élection : ce contrôle peut porter sur le déroulement du scrutin, sur les comptes de campagne ; il peut aboutir à une validation ou une invalidation etc. Dans toutes ces situations, il n'existe pas de propriété qui permettrait d'ordonner les décisions étudiées et la limitation de la recherche textuelle devient alors handicapante.

D'un point de vue technique, le choix de cette thématique et de cette juridiction était aussi assez opportun. En effet, les sites des juridictions et légifrance, qui sont les sources principales de décisions, sont assez inégaux dans la façon dont les données sont structurées. La plupart du temps, les décisions sont archivées comme du texte brut avec un balisage html qui vise principalement à leur mise en forme sur la page internet. En revanche, le site du Conseil constitutionnel va un peu plus loin en offrant des décisions légèrement structurées permettant une extraction un peu plus simple de certaines informations.

2.2 Scenarios d'utilisation.

La conception du système d'information impose de s'entendre sur un cahier des charges. Il s'agit ici de définir ce que l'on veut pouvoir faire et que le site du Conseil constitutionnel ne permet pas.

Scenario 1 : amélioration de l'ergonomie.

> Le système d'information proposé doit d'abord permettre une navigation équivalente à ce qu'autorise la recherche experte sur le site du Conseil constitutionnel. Il en reprend donc les critères (type, date, résultat etc). Dans le cadre d'une telle utilisation, son intérêt réside principalement dans le caractère intuitif des questions posées et dans le regroupement de certaines propriétés que le site du Conseil n'exploite pas complètement. Par exemple, les dates des décisions peuvent être regroupées par années (comme le site du Conseil), mais aussi par mois voire jours.

Scenario 2 : gain en précision des recherches.

> Le système d'information proposé doit ensuite permettre une navigation plus fine sur les décisions, notamment en ne traitant plus une décision comme une unité d'information insécable. Il s'agit ici de descendre au niveau du "considérant" (paragraphe). Le système permettant l'étiquetage des éléments (l'ajout d'une nouvelle propriété), le chercheur peut ainsi identifier les "considérants" importants des décisions et les regrouper.

Scenario 3 : choix du point de départ de la recherche ; ajout de perspectives propres au chercheur.

> Le système d'information proposé doit enfin permettre, en fonction des critères intégrés, de commencer la recherche depuis un point qui ne serait pas autorisé par le site du Conseil constitutionnel. Il doit permettre au chercheur d'organiser la recherche en fonction de la perspective qu'il s'est donnée. En s'appuyant sur des catégories établies par le juriste, il doit ainsi faciliter l'isolation d'un corpus de décisions présentant la même propriété.
>
> Si, par exemple, on s'intéresse à l'organisation et aux principes des élections législatives, il faut pouvoir accéder aux décisions présentant la propriété identifiant cette thématique, *indépendamment de la structure du site du Conseil constitutionnel* (notamment en décisions DC, QPC, I, AN etc.). En l'occurrence, une telle requête consiste à sélection la question sur `SousDomaine ?` puis Organisation et principes des élections législatives.
>
> Idem, par exemple, pour les décisions I (incompatibilité) qui sont souvent présentées selon un critère personnel : incompatibilité de M. ou Mme X. Or s'il travaille sur les incompatibilités en général, le chercheur préférera une information sur la profession en cause plutôt que sur la personne en cause. Ici, une telle requête consistera à naviguer dans les décisions I en fonction des questions `ProfessionCompatible ?` et `ProfessionVisee ?`

Afin de répondre à ce cahier des charges, le prototype requiert le choix de critères qui dépendent de la nature des données et des relations des éléments à classer (contenant/contenu, un à plusieurs, plusieurs à plusieurs). Ils sont détaillés dans la section 3.4 sur le choix des descripteurs.

3 Chaîne de traitement

3.1 Vue d'ensemble

Dans cette expérience, nous ciblons la réalisation d'un prototype, permettant une recherche flexible, portant sur un ensemble de *décisions de justice*.

Nous suivons pour cela la chaîne de traitement illustrée dans la Figure 1 pour une mise au point à partir d'une sélection d'exemples, appliquée ensuite à un ensemble plus complet de documents.

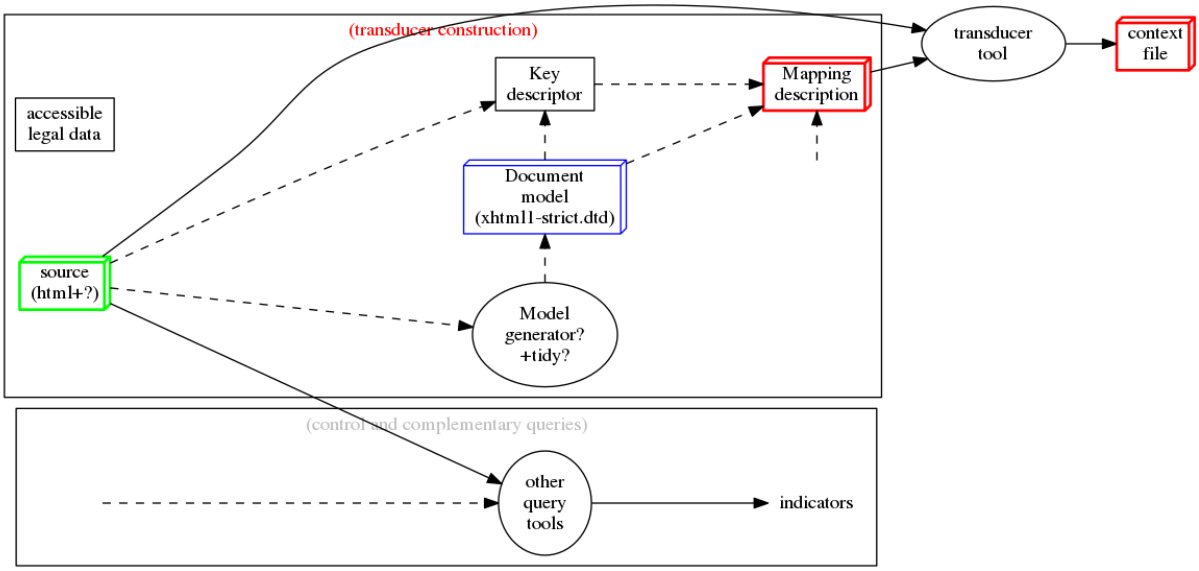

Figure 1: schéma de développement

La chaine vise à obtenir un "fichier de contexte logique" (cadre à droite de la Figure 1), qui pourra être chargé dans un "système de gestion de contexte logique". *Ceci est analogue au chargement d'une base de données dans un système de gestion de base de données, mais un contexte logique offre d'autres possibilités.* Nous avons utilisé Camelis (version 1) http://www.irisa.fr/LIS/ferre/camelis/ basé sur l'analyse de concept logique (S. Ferré and O. Ridoux), une approche qui étend l'analyse de concept formel (B. Ganter and R. Wille).

Cette réalisation, s'est appuyée sur des données légales accessibles au format XML / HTML, avec un modèle de document associé. Ce format permet alors un certain type de traitement uniforme. La construction est guidée à la fois par un modèle d'entrée et un modèle de contexte logique visé avec : un choix de types d'objets ; une sélection de propriétés ; un lien entre les deux modèles et la spécification d'une clé dans le source.

3.2 Systèmes d'information logique

3.2.1 Définitions

Formellement, un *contexte logique* est défini par un ensemble fini O d'objets o_i (de label l_i) et pour chaque o_i, un ensemble fini de descriptions logiques $d(o_i)$ pour un langage logique L.

Un *système de gestion de contexte logique* permet de charger et d'exploiter un tel contexte, d'interroger par des requêtes logiques (explicites dans L, ou interactives) ; la réponse est alors un sous-contexte d'objets satisfaisant cette requête.

3.2.2 Utilisation pour l'exploration de contextes.

Une vue en contexte logique permet d'explorer des informations de manière flexible, avec des facettes sémantiques, des possibilités d'inférences logiques, sans rédaction de requête a priori et d'obtenir aussi des indications sur la qualité des données. Un tel contexte peut être enrichi par d'autres informations

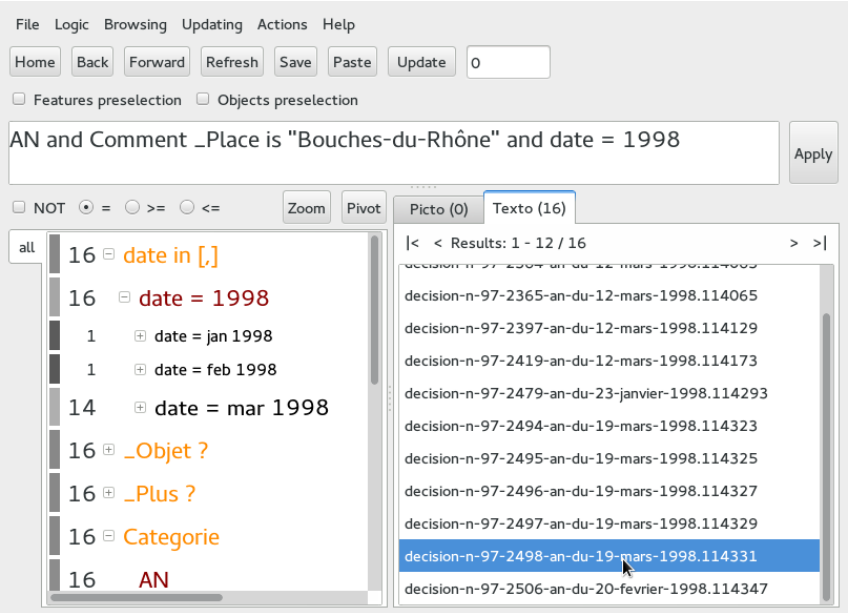

Figure 2: Sous-contexte sélectionné dans Camelis (3 fenêtres synchronisées)

(de natures diverses), en le reliant à d'autres applications (par des actions associées selon des arguments dans le contexte).

Présentation de contexte. Un contexte est présenté avec trois fenêtres synchronisées (voir Figure 2) : une question (une requête éditable) dans la fenêtre en haut, un index multi-hiérarchique de propriétés (les liens cliquables) dans la fenêtre à gauche, une présentation des objets (des exemples sélectionnables) dans la fenêtre à droite.

Navigation dans des sous-contextes. L'utilisation typique du système est une navigation dans un contexte menant du contexte initial *All* (comprenant tous les objets prévus dans le contexte) à un sous-contexte C_i ($i > 0$) comprenant les objets satisfaisant un ensemble de propriétés particulières (cf video[2]). Les propriétés du sous-contexte C_i sont résumées par la formule en haut (requête éditable), cette requête n'a pas besoin d'être rédigée, elle peut être construite par clics et sélections dans l'index de propriétés dans la fenêtre à gauche (les objets présentés seront alors ceux vérifiant la propriété cliquée) ou par sélection dans la fenêtre des objets.

Langage de requêtes. La connaissance du langage de requête n'est pas nécessaire. Cependant elle peut être lue ou éditée pour un usage en mode expert.

Le langage permet une combinaison avec les opérations booléennes : and, or, not de propriétés selon le type d'attribut ; il permet d'intégrer des questions de la forme :

- sur les chaînes de caractères : ... is "une chaîne" ;

  et aussi (où l'initiale du mot-clé suffit, c pour contains etc.) :

[2] http://www.irisa.fr/prive/foret/DEMO/VIDEOS/Screencast7sep2017-LIS-DecisionsConseilConsitutionnel.webm

... contains "une chaîne" ; ... beginswith "une chaîne" ; ... endswith "une chaîne" ;

... match "expression régulière (de type unix)" ;

exemple (nombre avec signe optionnel) : ... match "[+-] ?[0-9]+"

- sur les entiers : ... = un_entier ; ... in un_intervalle

  exemple : in 2000 .. 2015 ; ex : in 2000 ..

- sur les dates : ... = une_date ... in un_intervalle

  exemple : date in [may 2016,jul 2017]

Le langage de requête permet aussi des questions sur les heures, les droits sur fichier.

3.2.3 Nouveau prototype

Notre réalisation concrète sur des données du Conseil Constitutionnel, illustre les avantages d'une telle approche pour des informations juridiques. D'autres formats cibles sont aussi possibles. Le contexte que nous avons construit présente les caractéristiques suivantes.

- Les objets sont des décisions du conseil constitutionnel sur les élections.
- Les facettes sont des propriétés choisies, avec extraction de valeurs automatisée ; un ajout de tags personnels est possible.
- Il s'agit d'un contexte modulaire, pouvant comporter des données hétérogènes.
- Le système permet une navigation/sélection flexible et sûre, même sans connaissance du langage.

Nous détaillons par la suite certaines étapes de la construction.

3.3 Descripteurs de clés pour l'accès et la représentation

3.3.1 Codes de document existants.

Au niveau document, plusieurs identifiants peuvent être utilisés.

Numéro de document. Pour identifier un document lors des traitements, nous privilégions le numéro de décision. Cet identifiant est un texte avec un ou plusieurs numéros, un indicateur et une date, comme dans cet exemple : `Décision n° 2017-651 QPC du 31 mai 2017`

La même information apparaît de façon légèrement différente, dans le lien html :

`decision-n-2017-651-qpc-du-31-mai-2017.149036.html`

Code ECLI. Toutefois, les juridictions françaises utilisent depuis quelques temps un numéro visant à servir d'index au niveau européen dans le but de la constitution de bases de données. Ce numéro s'appelle ECLI (European Case Law Identifier[3]). Dans les décisions du Conseil constitutionnel, on le trouve souvent à la fin, en pied de page. Chez 99% des juristes, ce numéro n'est pas naturellement recherché mais il existe et finira sans doute par s'imposer ou à coexister de manière pérenne. La Figure 3 montre comment le numéro ECLI est généré pour le Conseil constitutionnel.

La double représentation ECLI et numéro de décision est en fait possible dans le prototype et nous proposons les deux comme facettes de recherche dans l'index.

[3] https://e-justice.europa.eu/content_european_case_law_identifier_ecli-175-fr.do

Identifiant France + Conseil constitutionnel	Année du jour de la décision	Numéro de la décision	Type de la décision
ECLI :FR :CC :	2017 :	2017.640.	QPC

Figure 3: Composants du Code ECLI

3.3.2 Codes choisis

Nous avons choisi le numéro de décision comme descripteur principal pour la représentation.

Au niveau source. Au niveau des données, l'unité de base est le document, dont nous considérons comme identifiant principal, le "numéro de document" (un texte avec un ou plusieurs numéros, un indicateur et une date, comme décrit ci-dessus).

Au niveau prototype. Au niveau du contexte produit (système d'information), l'unité de base est un objet désigné par cette information. Nous lui associons des propriétés le décrivant et qui semblent utiles à la recherche parmi ces objets et ces propriétés.

La clé privilégiée est ainsi la même dans le source et dans sa transformation. Les traitements s'appuient sur cet identifiant.

3.3.3 Unité d'information et balisage XML/html.

La décision entière en tant que document est l'unité d'information minimale en l'état. Mais conserver la décision entière comme seule unité d'information, impose des résultats mixtes et certaines limitations.

Il serait possible de gagner en précision en exploitant un balisage formel du document html, certes rudimentaire mais déjà plus fin, qui distingue les grands blocs, dont les fameux motifs.

Pour prendre en compte ces différents niveaux dans le système d'information, il serait possible d'ajouter des objets repérant des composants de document. il faudrait alors affiner la clé relative à ces éléments.

3.4 Choix des descripteurs de propriétés pour les facettes

3.4.1 Descripteurs temporels

Au niveau source. La date est indiquée dans le source comme dans les exemples de descripteurs de clés ci-dessus.

Au niveau prototype. Nous pouvons la représenter dans le contexte de plusieurs façons : initialement comme un type chaine ; mieux comme un type date, permettant une granularité de recherche (par année, par année et mois, etc.).

Un avantage de cette dernière représentation est aussi de pouvoir considérer les sous-contextes par année, avec leur nombre d'objets (et éventuellement d'ordonner les années dans l'index par nombre d'objets décroissant).

3.4.2 Descripteurs pour le type de document

Les décisions sont de différents types, nous en considérons quatre : QPC, DC, AN et I. Il s'agit du premier critère formel à prendre en compte. On y distingue trois groupes qui n´ont pas la même structure : 1) QPC et DC ; 2) AN ; 3) I.

Le type de décision est mentionné après le numéro de la décision.

L'extraction automatique est simple, à partir de ce numéro (figurant à plusieurs endroits, dont le nom de fichier).

Nous les représentons comme des attributs de base, ils sont regroupés en hiérarchie taxonomique (sous Categorie), par des axiomes ajoutés au contexte tels que : `axiom QPC, Categorie`

On peut traduire le groupe intermédiaire par une hiérarchie à 3 niveaux ainsi :

```
axiom QPC, QPC-DC
axiom DC, QPC-DC
axiom AN, Categorie
axiom I, Categorie
axiom QPC-DC, Categorie
```

3.4.3 Descripteurs envisagés selon le type de document

Pour des critères plus substantiels, il faudrait distinguer selon le type de décisions:

- pour celles du premier groupe (QPC-DC), on pourrait retenir comme première critère l'objet de la décision (la loi contrôlée), mais aussi les principes constitutionnels invoqués (égalité par exemple) voire le résultat de la décision (que l'on trouve au dispositif: conforme, non conforme).

- pour celles du deuxième groupe (AN), il s'agit du contrôle des élections "concrètes", les critères pourraient reposer sur l'objet de la contestation: propagande de campagne (bulletins, affiches etc.), opérations de vote et comptes de campagne.

- pour celles du dernier groupe (I), il s'agit de déterminer si la profession d'un individu est compatible avec les fonctions de député. Il existe des textes à cet égard mais ils sont souvent précisés par des décisions de justice. En l'occurrence, il y aurait un critère de sélection reposant sur les professions sur lesquelles le Conseil constitutionnel se prononce.

Les décisions de justice ont souvent une structure formelle qui facilite les choses. Cependant, elle n'est pas systématique et est souvent variable...

3.4.4 Descripteurs envisagés selon le « problème juridique»

Procéder par «problème juridique» permettrait de réduire à des résultats quasi binaires certains des résultats mixtes. Un balisage dédié pourrait être utile. Nous listons des caractéristiques à prendre en compte selon le type de décision.

Pour une décision QPC-DC : Les décisions DC et QPC sont comparables et sont les plus complexes. Pour une décision DC :

- Une décision DC peut trancher sur plusieurs « dispositions » ;

- Une disposition peut être contestée au regard de 1 ou plusieurs principes constitutionnels ;

- Pour chaque couple disposition + principe constitutionnel, le résultat peut être « non conforme », « conforme », conforme « sous réserve » (qui correspond à conforme à condition d'être appliqué ou interprété de telle ou telle manière). Les variantes « contraire »/ « non-contraire » peuvent être assimilées à conforme/non conforme. Pour une représentation peut-être plus simple, on pourrait casser ce résultat en deux résultats binaires si cela peut rendre les choses plus simples : « non-conforme »/ « conforme » puis « conforme »/ « conforme sous réserve ».

Pour une décision AN : - Une décision AN n'évoque jamais qu'une seule élection. Comme pour les décisions I, la clé unique est en fait la « personne » ou l'élection jugée. Là aussi l'information est non pertinente d'un point de vue scientifique. Pour gagner en précision, peut-être pourrions-nous descendre ici encore pour plus de précision.

Une décision AN peut évoquer plusieurs problèmes (comptes de campagnes, opération électorales, propagande par ex.)

- Un problème a un résultat binaire : « régulier » / « irrégulier » (et si dans la décision, il y a des irrégularités substantielles), il y a annulation, inéligibilité etc.

Pour une décision I : - une décision I peut théoriquement évoquer plusieurs professions, même si le plus souvent il n'y en a qu'une. La clé est la personne « jugée » mais cette information n'est pas pertinente d'un point de vue scientifique ;

- pour chaque profession, le résultat est binaire « compatible »/ « incompatible ».

3.5 Observations et difficultés

3.5.1 Format (html,..) et modèle du document

Une première difficulté de traitement automatisé tient au fait qu'un document source ne respecte pas toujours strictement le modèle attendu. Cet échec se manifeste dès le début, si on tente d'appliquer un outil reposant sur les technologies XML/html, au document html fourni.

Des outils informatiques existent pour "nettoyer" un fichier dans le but de le rendre compatible avec le format attendu. Mais nous observons lors d'un premier test, que la commande "tidy" reste silencieuse.

La solution provisoire est alors de recourir à des outils plus rudimentaires (comme "sed", procédant par motif sur chaque ligne) et ne reposant pas sur la technologie XML. Une solution plus satisfaisante serait d'appliquer et d'ajuster des méthodes de *nettoyage de données* (en anglais "data cleansing").

3.5.2 Rendu d'information sémantique

Le rendu d'une information particulière par une facette peut poser plusieurs problèmes. C'est le cas du résultat de décision de conformité que nous voulons indiquer pour les documents de type QPC et DC.

Nature du résultat. Le résultat attendu est a priori "conforme" ou "non conforme", cependant un résultat peut être :
- mixte, comme dans la décision n° 2008-573 DC du 8 janvier 2009
- avec des réserves, comme dans la décision n° 2011-628 DC du 12 avril 2011

Expression du résultat. On observe des variantes pour exprimer la conformité, par "... est conforme..." / "... n'est pas conforme ...", mais aussi :
- "... n'est pas contraire à ...", comme dans la décision n° 2010-602 DC du 18 février 2010

3.5.3 Représentation d'entités

La désignation d'une entité, d'un lieu varie selon le temps, par exemple : la désignation de département (facette *Comment_Place* dans notre prototype) est *Loire* ou *Loire Atlantique* selon la période.

Il s'agit là d'un problème non spécifique, et plus généralement la désignation varie selon la terminologie, l'expression employée (expressions synonymes, approximations).

Une grande variabilité, explicitée ou non, existe dans les données, sur le texte brut ou les codes utilisés.

3.6 Qualités attendues

Pour un système de qualité, certaines difficultés de différents ordres devront être prises en compte, nous en avons listé ci-dessus. Nous complétons avec d'autres aspects relevant de l'informatique qui concernent notre prototype et les données de l'expérience.

3.6.1 Traitement

Sur des données de nature ouverte et évolutive (avec de nouvelles décisions publiées), le traitement devrait être facilement reproductible (et aussi ouvert et adaptable).

En même temps, la complexité du système doit être "raisonnable" : en temps et place de calcul (pour une bonne interface) ; on peut aussi prendre en compte l'énergie (par exemple limiter les accès web).

Le traitement devrait produire un résultat à la fois *conforme* aux sources (et permettre d'y accéder) et *couvrant* (éviter de laisser des documents de côté).

Pour cela, un apport important du traitement automatique des langues (TAL) est à prévoir. On trouve par exemple : « n'est pas incompatible » dans la decision n^o 96-16 I du 19 décembre 1996

On trouve aussi cette écriture inhabituelle ailleurs : D É C I D E (avec des espaces).

3.6.2 Acceptabilité

Pour une meilleure aide à l'usager, il est important de s'assurer de l'intelligibilité du résultat, de choisir des structures et des nommages du contexte pertinents : pour les objets (ou unités) et pour les facettes (ou critères). Au-delà du contexte lui-même, les questions des manipulations permises (expressivité) et du temps de réponse (à une requête) avec le nombre de clics (pour une recherche) peuvent être déterminants pour l'acceptabilité du système.

4 Conclusion

Nous proposons un prototype avec d'une part des facettes générales (date, etc.) de nature objectives, factuelles et des facettes choisies par le juriste de nature subjectives, qui ont un impact sur une navigation orientée, simplifiée, sûre et la mise en évidence de caractéristiques.

La qualité/facilité de construction dépend cependant d'une certaine structuration dès le départ qui repose en partie sur un travail des juridictions elles-mêmes. Même si les décisions de justice présentent une certaine structuration formelle repérable grâce à des mots-clés, une structuration plus explicite faciliterait grandement l'exploitation de ces sources (xml notamment). En outre, des difficultés de rendu peuvent apparaître selon la nature et la représentation de l'information (permettant une classification aisée ou non, avec des données régulières ou non, une terminologie disponible ou non). C'est le cas par exemple pour le concept de fonction (activité) ... Là encore, si on ne peut demander aux juridictions de toujours employer les mêmes termes, une structuration plus explicite dès la rédaction permettrait de grandement améliorer les systèmes.

Du côté informatique, ce type d'étude permet de tester des modèles formels pour la représentation et l'extraction de l'information et l'intelligibilité des données.

Pour les travaux futurs, il faudrait prendre en compte les points suivants :

- l'évolution du style de rédaction (FR recul de l'imperatoria brevitas), perte des mots-clés typiques pour la structure ("considérant", "attendu") qui nécessiterait un appui du traitement automatique des langues (TAL).
- l'européanisation et l'internationalisation des sources du droit (droit comparé, problématique transnationale) et le référencement des décisions au niveau européen (ECLI (European Case Law Identifier))
- une terminologie ouverte et complète des termes juridiques de référence (avec leurs variantes).

Du côté droit, une telle étude permet d'envisager une amélioration des méthodes de travail des chercheurs et augure d'un possible gain de productivité. Elle oblige aussi à une certaine réflexivité sur les objets étudiés afin de permettre la construction de systèmes d'information adaptés. Dans cette perspective, l'association de chercheurs en droit et d'informaticiens est une étape intermédiaire avant que les juristes n'acquièrent un peu plus d'autonomie.

	Conseil d'état	Cour de Cassation	Conseil Constitutionnel	Legifrance [et Dalloz , Lexis Nexis]
interface dédiée	oui (Ariane)	non	oui (rech.exp.)	oui (rech.exp.)
arrêts et décisions compilés	jurisprudence interne (incomplet)	jurisprudence interne (incomplet)	jurisprudence interne (complet)	toutes juridictions (séparé pour Légifrance)
recherche transversale (types de décision)	sans objet	sans objet	oui	oui
(juridictions)	sans objet	sans objet	sans objet	non (Legifrance) oui (Dalloz, Lexis Nexis)
filtrage sur critères formels (N^o, date etc.)	oui	limité	oui	oui
"perspective" (critères de fond)	classification interne (Leb.)	classification interne (Bull.)	version "papier"	reprise des classifications internes des juridictions
recherche textuelle ("langage" de requêtes)	très limitée aucun	non	très limitée aucun	$\leq$ 5 termes (et/ou, 1 sauf) − err. sur caract. spéc. [$\leq$ 4 (et/ou/sauf) Dalloz] [$\leq$ 5 (et/ou/sauf) Lexi360]
données structurées	non	non	structuration légère	non

Figure 4: Caractéristiques et limites des systèmes actuels

A Annexe - Présentation des systèmes d'information actuellement utilisés par les chercheurs en Droit

Pour l'essentiel, les outils utilisés par les chercheurs en droit ne sont pas produits par eux ou très indirectement. En tout état de cause, il n'en ont pas la maîtrise. En effet, ceux-ci sont principalement constitués des sites officiels des juridictions, des bases de données d'éditeurs juridiques (Dalloz, Lexis Nexis, Francis Lefebvre etc.) et du site Légifrance. D'un point de vue épistémologique, ces sources présentent toutes des inconvénients :

- elles sont d'avantage orientées vers la diffusion des décisions de justice plutôt que vers l'intelligibilité de l'information juridique ;
- elles sont incomplètes ou parcellaires ;
- elles ne compilent pas des données brutes mais plutôt des textes ou des classifications internes des juridictions ;
- leurs caractéristiques techniques peuvent limiter le chercheur dans ses démarches etc.

A.1 Limites techniques des sites officiels des juridictions.

Les sites officiels des juridictions permettent l'accès aux arrêts et décisions de justice. Force est cependant de constater que pour au moins le Conseil d'Etat et la Cour de cassation, ils n'ont pas vraiment vocation à assurer cette fonction de diffusion et de publication.

Conseil d'Etat et Arianeweb. Sur le site du Conseil d'Etat, l'interface dédiée est "ArianeWeb"[4]. Le mode avancé permet principalement de filtrer selon des critères formels (numéro de décision, date etc.). La limitation principale se manifeste lors de recherches fondées sur des critères substantiels qu'elles soient transversales et/ou thématiques. Pour que le système lui retourne des résultats fondés sur des critères de fond, l'utilisateur peut utiliser la classification du Conseil d'Etat. Toutefois, n'est classée qu'une petite fraction des arrêts (ceux publiés ou mentionnés au Recueil Lebon) et la logique de la classification échappe au chercheur voire est parfois incohérente. L'autre possibilité pour l'utilisateur est de recourir à la recherche textuelle très limitée pour laquelle l'outil n'est manifestement pas conçu (pas d'équation de recherche, seulement pluriels et synonymes).

[4] http://www.conseil-etat.fr/Decisions-Avis-Publications/Decisions/ArianeWeb

ArianeWeb n'est en fait qu'une version bridée du logiciel professionnel utilisé par les personnels du Conseil d'Etat eux-mêmes. Il répond donc avant tout à leurs besoins spécifiques et s'adresse donc plutôt à des utilisateurs ayant une idée très précise de ce qu'ils cherchent ou ayant déjà consulté une autre source documentaire et qui viendraient compléter leurs informations grâce à cet outil.

Site officiel de la Cour de cassation. Sur le site de la Cour de cassation[5], il n'y a pas d'interface spécifique, simplement une page "jurisprudence". Le site classe les arrêts en fonction de critères formels (formations de jugement, dates, numéros de requête, objet normalisé du litige, résultat). Aucune recherche textuelle n'est possible. Pour une recherche thématique ou transversale, l'utilisateur doit utiliser la classification "par rubrique" effectuée par la Cour et qui est inachevée (au mieux les arrêts remontent aux années 2000). Il peut aussi naviguer à tâtons s'il a une certaine connaissance des compétences de chacune des formations et chambres de la Cour. Mais pour une recherche plus précise, il doit s'appuyer sur les publications aux bulletins de la Cour qui ne font l'objet que d'une numérisation très insuffisante (un pdf de la version papier est téléchargeable sur le site; le prétendu bulletin numérique ne fait que reprendre les arrêts cités dans la version papier selon des critères temporels et sans permettre de recherche par mots clés par exemple).

Site officiel du Conseil constitutionnel. Le site du Conseil constitutionnel fait presque exception[6]. Ayant fait l'objet d'une réfection récente, il donne accès à des fonctionnalités un peu plus avancées. Le classement se fait toujours selon des critères formels mais le mode avancé permet la recherche selective ou transversale selon le type de décision. Il permet aussi la recherche textuelle sur des blocs spécifiques des décisions (visas, motifs, dispositifs), signe de la présence d'un balisage plus fin des décisions. La recherche textuelle reste cependant limitée (pas d'équation de recherche) et la normalisation des décisions n'est pas achevée (certaines décisions du même type ne sont pas classées ensembles). De plus, si la recherche transversale sur différents types de décision est permise, il n'existe sur le site aucune classification fondée sur des critères substantiels. Pour y accéder, l'utilisateur doit télécharger les tables analytiques du Conseil constitutionnel qui n'existent qu'en pdf.

A.2 Limites techniques de Légifrance.

Le site Légifrance[7] a vocation à permettre une meilleure diffusion du droit en général. Concernant la question plus spécifique des arrêts et décisions de justice, il permet un recensement souvent plus complet que les sites officiels des juridictions car il donne aussi accès à des arrêts mineurs. Son intérêt principal vient surtout de la recherche textuelle qu'il autorise alors que les sites officiels des juridictions ne la permettent pas ou seulement de manière très limitée. Il permet aussi d'utiliser la classification interne des juridictions.

S'il est un complément utile, Légifrance présente toutefois des défauts importants pour un chercheur:

- il n'autorise pas la recherche simultanée sur les fonds de la juridiction administrative, de la juridiction judiciaire voire éventuellement du Conseil constitutionnel alors même qu'une des caractéristiques de la législation moderne est de générer des problématiques ressortissant de plusieurs branche du droit ;

- les critères de recherche sont principalement formels (date, numéro de requête, publication etc.)

- les filtres de recherche textuelle sont souvent limités dans leur efficacité; l'absence de balisage formel des arrêts et décisions empêche la limitation du champ à des blocs spécifiques des arrêts; l'équation de recherche est limitée à 5 termes (4 et/ou et 1 exclusion); certains caractères "spéciaux" génèrent des erreurs.

- puisqu'il s'adresse à un public large, Légifrance adopte une présentation sans perspective ou presque.

[5] https://www.courdecassation.fr/

[6] http://www.conseil-constitutionnel.fr/
http://recherche.conseil-constitutionnel.fr/?expert=2

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

A.3 Limites techniques des systèmes d'information payants.

La recherche juridique s'appuie par ailleurs sur des systèmes d'information produits par des éditeurs juridiques, les principaux étant Dalloz et Lexis Nexis. La plus-value de ces outils vient surtout du fait qu'ils permettent d'accéder en même temps aux décisions de justice et aux articles de doctrine qui peuvent en constituer le commentaire. Mais il est rare que leurs systèmes d'information ajoutent beaucoup à ceux déjà existants et en particulier à Légifrance.

Le site Dalloz[8] reprend globalement les fonctionnalités du site Légifrance en y ajoutant la recherche transversale sur toutes les juridictions mais réduit l'équation de recherche à 4 membres et les articulations et/ou/sauf.

Lexis360 permet la recherche textuelle transversale, une équation de recherche à 5 membres avec les articulations et/ou/sauf et des critères de proximité. Il superpose par ailleurs les analyses effectuées par l'éditeur lui même dans sa base de donnée propre appelée jurisdata.

A.4 Bilan

Globalement, les systèmes d'information actuellement utilisés par les juristes fonctionnent comme des banques d'arrêts et décisions. Tant que la recherche repose sur des critères formels basiques et référencés (date, numéro de décision etc.), ils sont relativement satisfaisants bien que perfectibles. Mais précisément, tous les critères formels ne sont pas retenus (par ex. composition de la juridiction, origine de l'appel, date de saisine etc.).

Reste cependant que le défaut majeur des systèmes d'information actuels vient de l'absence de critères de recherche substantiels et de l'insuffisance des moyens de compensation proposés. Les critères de classification substantiels par les juridictions posent des problèmes épistémologiques et interroge sur les rapports entre doctrine universitaire et doctrine organique des juridictions; la recherche textuelle demeure aléatoire et trop limitée.

A l'évidence, il ne peut y avoir une seule et unique classification substantielle. Toutes auront potentiellement des "défauts". La critères retenus révèlent en effet la perspective du chercheur. Mais au moins, cette perspective est explicite. Mieux, elle constitue la plus-value d'une étude qui réside justement dans un travail de classification et de modélisation qui ne se borne pas à reprendre des critères formels non pertinents ou substantiels mais imposés par d'autres.

[8] http://www.dalloz.fr/
http://www.dalloz.fr/documentation/Recherche?ctxt=0_dCRzMD1ET0NUwqdkJG5UZXh0ZTI9My8xMQ==