PROCÉDURE D'ÉCHANTILLONNAGE

Par Michelle Cumyn et Charles Tremblay-Potvin

1. Objectifs

Notre projet de recherche a pour but d'élaborer un nouveau modèle d'indexation de la documentation juridique afin de développer une interface permettant d'effectuer une recherche par facettes. Les facettes devront représenter les différents aspects du contenu des documents indexés. En exploitant l'association de termes de recherche relevant de différentes facettes au sein des documents indexés, l'interface sera conçu pour : 1) accompagner l'utilisateur dans la qualification juridique du problème à résoudre; 2) l'assister dans la découverte d'analogies possibles; 3) l'aider à développer une vision transversale du droit; et 4) repérer de manière efficace la documentation pertinente.

L'objectif principal du projet est de développer un prototype de l'interface de recherche envisagé afin d'évaluer la qualité des résultats de recherche qu'il permettra d'obtenir. Nous devons donc procéder par échantillonnage puisqu'il serait impossible d'indexer l'ensemble des documents que le projet a vocation à couvrir.

Les facettes envisagées, qui sont au nombre de six, sont les suivantes : Personne, Action, Chose, Contexte, Régime juridique et Remède ou sanction.

2. Contenu de la banque de données prototype

Dans la demande de subvention, nous avions prévu y inclure des décisions de justice (jurisprudence) et des fiches qui résument des articles ou des ouvrages de référence en droit (doctrine), documents qui figurent tous dans la banque de données de SOQUIJ. Nous avons finalement décidé de retenir seulement des décisions de justice. En effet, la recherche dans les banques de données juridiques se fait principalement dans la jurisprudence, et la recherche multi-sources demeure peu développée. La conception d'une interface de recherche conviviale et efficace pour la recherche jurisprudentielle constituerait une avancée importante, même en supposant qu'elle ne soit pas immédiatement transposable à la recherche doctrinale. Des considérations pratiques s'ajoutent aux précédentes : nous avons éprouvé de la difficulté, lors du test d'indexation, à indexer les décisions à partir des résumés seulement. Or, la banque de SOQUIJ contient les textes intégraux des décisions, mais une grande partie de la doctrine n'est pas disponible en texte intégral. Pour pouvoir les indexer, il faudrait se les procurer indépendamment.

3. Méthode d'échantillonnage

Nous réaliserons un échantillonnage à deux niveaux (échantillonnage en grappe). Au premier niveau, nous avons choisi de manière raisonnée trois domaines du droit. Au

deuxième niveau, nous sélectionnerons de manière aléatoire des documents au sein de ceux-ci.

Nous avons décidé d'effectuer une sélection raisonnée des domaines de droit pour trois raisons. Premièrement, nous voulons restreindre le nombre de domaines, parce que chacun d'eux nécessite un travail important de mise à plat et de classification des régimes juridiques qui le composent. Deuxièmement, nous souhaitons, par le choix des domaines, favoriser un certain chevauchement, avec des décisions (ou des faits) pouvant appartenir à plus d'un domaine, de façon à dégager des vues transversales du droit. Troisièmement, le choix d'un nombre limité de domaines fera augmenter le nombre de décisions par domaine, de façon qu'il y ait une densité ou une redondance suffisantes de l'information, ce qui est nécessaire pour créer des conditions de recherche réalistes lors de l'essai du prototype et à l'étape de l'étude d'utilisabilité.

Au deuxième niveau d'échantillonnage, nous choisirons de façon aléatoire des documents au sein des domaines sélectionnés en respectant la proportion suivant laquelle ils apparaissent dans la banque de données de SOQUIJ. Cependant, nous devrons nous assurer qu'un domaine qui contient un nombre très important de documents ne nuira pas à la représentation des autres domaines au sein du prototype.

4. Période d'échantillonnage

Compte tenu de l'évolution rapide de la terminologie juridique, nous avons décidé de limiter notre corpus de décisions à une période de temps relativement courte. D'ailleurs, en date du 1er janvier 2016, une importante réforme des organismes administratifs en matière de travail entrait en vigueur. La Commission des normes du travail (CNT) a été fusionnée avec la Commission de la santé et la sécurité au travail (CSST) ainsi qu'avec la Commission de l'équité salariale pour former la Commission des normes, de l'équité, de la santé et de la sécurité au travail (CNESST). La Commission des relations du travail a également été fusionnée avec la Commission des lésions professionnelles pour former le Tribunal administratif du travail. Ces modifications risqueraient de complexifier l'analyse du vocabulaire et l'indexation des décisions. Or, l'adaptation aux changements institutionnels ou terminologiques ne fait pas partie des objectifs du projet. Nous sélectionnerons donc des documents antérieurs à cette réforme.

Nous ne souhaitons pas non plus que notre échantillon soit trop limité dans le temps, car l'observation des courants jurisprudentiels est un aspect important du travail des juristes. Nous estimons qu'une période de cinq ans permet d'atteindre un juste équilibre. Nous avons donc décidé de retenir les décisions rendues du 1er janvier 2011 au 31 décembre 2015 inclusivement.

5. Taille de l'échantillon

En ce qui concerne la taille de l'échantillon, elle doit être suffisante pour qu'il soit possible de faire l'essai du prototype et d'en évaluer la performance lors de l'étude d'utilisabilité. Pour ce faire, l'échantillon doit être suffisamment diversifié tout en comportant une certaine redondance de l'information. Nous croyons pouvoir atteindre ce résultat en retenant entre 2 000 et 2500 décisions au sein de l'échantillon. Si le choix d'un nombre restreint de domaines du droit favorise une certaine redondance ou densité de l'information, leur sélection doit aussi permettre une certaine diversité, afin que le prototype soit représentatif du droit dans son ensemble. En outre, nous sommes limités par les ressources financières dont nous disposons pour réaliser l'indexation des documents de la banque.

6. Domaines du droit (premier niveau)

Notre choix s'est arrêté sur les trois domaines suivants : (1) le droit administratif, (2) le droit des obligations (ce qui comprend les contrats et la responsabilité civile) et (3) le droit du travail. Ces domaines offrent une représentation diversifiée du champ juridique, couvrant à la fois le droit privé et le droit public, le droit civil et la *common law*, le droit statutaire et le droit jurisprudentiel, le droit fédéral et le droit provincial. De plus, ces domaines sont susceptibles de se chevaucher, par exemple lorsqu'une décision concerne la responsabilité de l'État, le contrat de travail, le contrat administratif ou le contrôle judiciaire d'une décision en droit du travail.

7. Échantillonnage proportionnel corrigé ou théorique (deuxième niveau)

En limitant le corpus à la période de temps indiquée précédemment, nous obtenons l'aperçu suivant de sa composition, à partir des rubriques du plan de classification de SOQUIJ qui correspondent aux domaines de droit choisis :

- ADMINISTRATIF (DROIT) (**14 353**)
- CONTRAT (**1 596**)
- CONTRAT D'ENTREPRISE (**5 039**)
- CONTRAT DE SERVICES (**8 837**)
- CONTRATS SPÉCIAUX (**1 968**)
- LOUAGE DE CHOSES (**224 899**)
- OBLIGATIONS (**1 543**)
- PRÊT (**1 796**)
- MANDAT (**566**)
- RESPONSABILITÉ (**6 128**)
- TRAVAIL (**55 793**)
- VENTE (**10 936**)
- **TOTAL : 324 596**

À noter que le total indiqué a été ajusté pour tenir compte du fait que certaines décisions puissent se retrouver sous plus d'une rubrique.

Nous observons qu'il existe une disparité importante entre le nombre de décisions appartenant aux différentes rubriques que nous avons sélectionnées. En particulier, le louage de choses et le travail paraissent surreprésentés par rapport à la responsabilité ou au droit administratif. Faut-il chercher à rééquilibrer la composition de l'échantillon?

Nous avons examiné de plus près la composition du corpus en essayant d'évaluer la redondance ou au contraire la diversité des décisions relevant de différentes rubriques et sous-rubriques du plan de classification. En droit du travail et en matière de louage, l'abondance des décisions est due notamment au fait que SOQUIJ a intégré à sa banque les décisions d'organismes spécialisés avec lesquels elle a conclu des ententes de service pour la gestion de leur documentation. À l'origine, ces décisions étaient rapportées dans des banques spécialisées, et seules les décisions en appel ou en contrôle judiciaire de ces organismes étaient rapportées dans la banque générale. Maintenant, toutes ces décisions se retrouvent dans la banque générale.

Parmi les sous-rubriques du plan de classification de SOQUIJ dans le domaine du travail, celle des accidents du travail et des maladies professionnelles contient 39 172 décisions pour la période retenue, soit plus de 70% des décisions en droit du travail. Parmi les sous-rubriques du louage de choses, c'est le bail d'habitation qui est surreprésenté avec 222 827 décisions, soit près de 90% de toutes les décisions relatives aux contrats. Le nombre très important des décisions rendues par la Régie du logement en matière de bail d'habitation et par la Commission des lésions professionnelles en matière de santé et sécurité du travail explique ce déséquilibre.

Nous avons envisagé d'exclure entièrement certaines sous-rubriques, soit celle des accidents du travail et des maladies professionnelles en ce qui concerne le domaine du travail et celle du bail d'habitation en ce qui concerne les contrats. L'utilisation de cette méthode permet d'en arriver aux proportions suivantes (à noter que dans les tableaux, le total et les pourcentages ne tiennent pas compte des chevauchements) :

Rubrique	Période 2011-2015
Administratif	14 446 (21%)
Contrats et obligations	30 460 (45%)
Responsabilité	6 131 (9%)
Travail	16 670 (25%)
TOTAL :	**67 562**

Si nous souhaitions conserver une certaine représentation des sous-rubriques que nous avons exclues précédemment, une autre possibilité serait d'en admettre un nombre limité au sein de l'échantillon. Nous pourrions en retenir un nombre équivalent à une sous-rubrique similaire. Concernant les accidents du travail, il s'agirait d'en retenir le même nombre qu'en matière de santé et de sécurité au travail. Les décisions découlant d'un accident du travail ou d'une maladie professionnelle seraient ainsi représentées au sein de l'échantillon, sans être surreprésentés par rapport aux autres. De même, pour obtenir une représentation adéquate du louage, nous pourrions retenir le même nombre de décisions portant sur le bail d'habitation qu'il y en a sur le bail commercial. En utilisant cette méthode, nous obtenons presque les mêmes proportions que si nous excluons complètement les sous-rubriques en question :

Rubrique	Période 2011-2015
Administratif	14 446 (20%)
Contrats et obligations	31 733 (45%)
Responsabilité	6 131 (10%)
Travail	17 538 (25%)
TOTAL :	**69 703**

Une autre difficulté a attiré notre attention. Il existe une grande variété de contrats (appelés « contrats nommés »), dont il faudrait faire l'analyse pour les inclure au plan de classification de la facette Régime juridique. Certains de ces contrats comptent un très petit nombre de décisions au sein du corpus (ex. le mandat (566) et le prêt (1 796)). Il se pourrait donc qu'ils soient peu ou pas représentés au sein de l'échantillon. De plus, certains contrats nommés figurent dans d'autres rubriques du plan de classification que celles relevées plus haut. Ils sont classifiés dans des rubriques à caractère sectoriel qui contiennent aussi des décisions relatives à des matières non contractuelles. Par exemple, la rubrique « Transport et affrètement », qui contient des décisions sur le contrat de transport, inclut aussi des décisions en droit du transport qui ne concernent pas les contrats. Si nous voulions inclure tous les contrats nommés du Code civil, il nous faudrait donc nous livrer à des manipulations compliquées lors de l'échantillonnage, ce qui pourrait occasionner des pertes de temps, voire des erreurs.

Pour éviter ces complications, qui découlent de la structure du plan de classification de SOQUIJ, et pour limiter le travail de mise-à-plat et d'analyse des différents contrats, nous avons donc décidé de nous en tenir à quatre contrats principaux, bien représentés au sein de la banque et qui présentent des similarités entre eux (leur sélection est donc propice aux chevauchements) : le contrat d'entreprise, le contrat de service, le contrat de travail et la vente. Les autres contrats nommés, y compris le louage, seront exclus de l'échantillon. Le domaine des obligations sera donc représenté par les rubriques suivantes : contrat, contrat d'entreprise, contrat de service, obligations, responsabilité, vente. À noter que le contrat de travail fait partie de la rubrique Travail dans la classification de SOQUIJ. Nous parvenons alors aux proportions suivantes :

Domaine de droit	**Période 2011-2015**
Administratif	14 639 (23%)
Contrats et obligations	26 411 (41%)
Responsabilité	6 132 (9%)
Travail	17 541 (27%)
TOTAL :	**64 726**

L'échantillon respectera ces proportions, et sera composé de 2 500 décisions.

8. Procédure d'échantillonnage

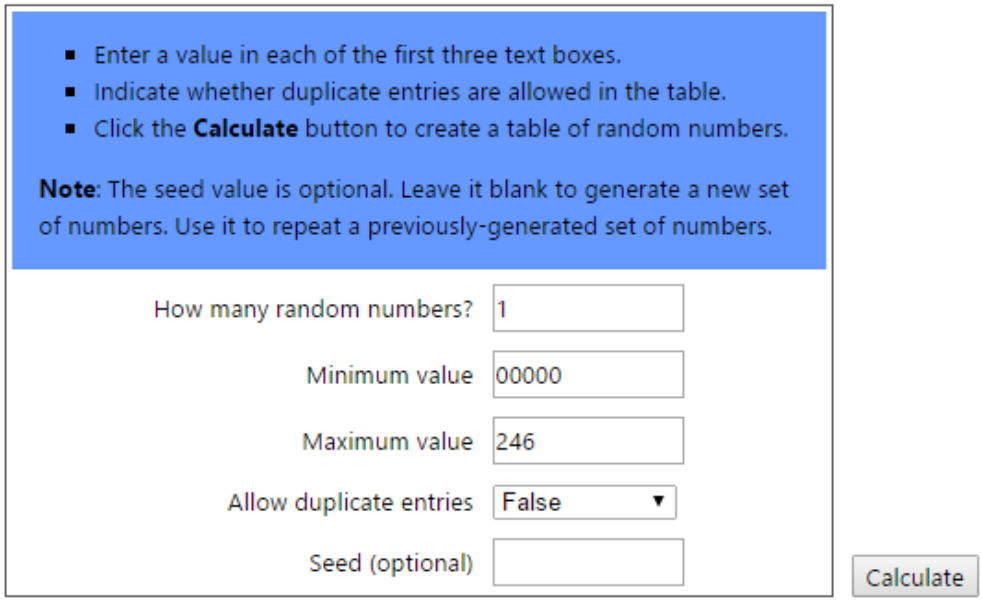
Pour sélectionner les décisions qui composeront l'échantillon, utiliser les filtres qui apparaissent à gauche de l'écran dans le moteur « Recherche juridique » de SOQUIJ. **Commencez toujours par cliquer sur « Nouvelle recherche »** en haut à droite de votre écran.

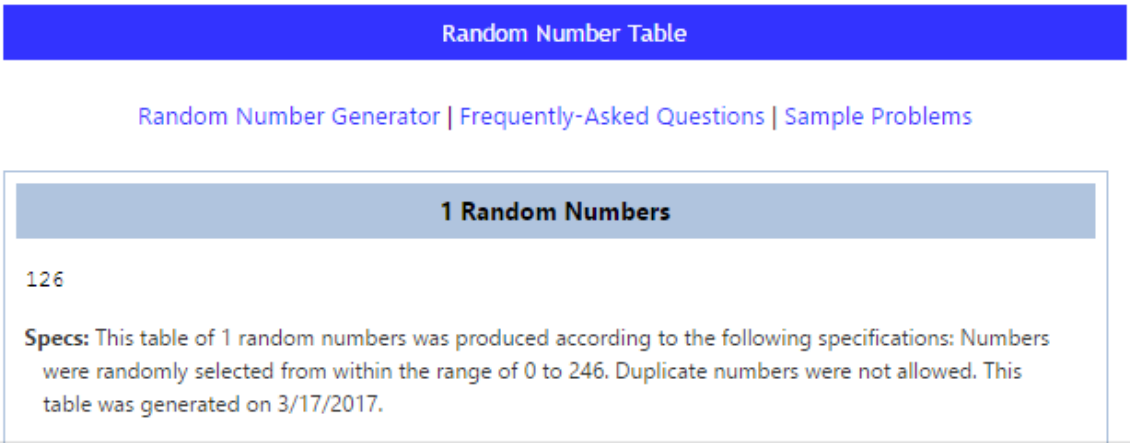
La première étape consiste à sélectionner la **période de temps couverte par l'échantillon**. À partir de l'icône en forme de calendrier, sélectionner « date de décision/ entre le/ 2011-01-01/ et le/ 2015-12-31 ».

Ensuite, **à partir du plan de classification, sélectionnez les rubriques visées par l'échantillonnage**. Pour faire apparaître le plan de classification, il faut cliquer sur « Afficher plus d'éléments » dans le pavé « plan de classification » situé à gauche de l'écran. Lorsqu'il faut sélectionner plus d'une rubrique du plan de classification, cliquez sur la case « Multi » en bas à droite de l'écran et cochez les rubriques désirées. Celles-ci seront automatiquement regroupées à l'aide de l'opérateur « OU ».

Chaque page de recherche affiche 25 résultats. Pour connaître le nombre de pages de résultats, cliquez sur « Dernière ». **Entrez le nombre de pages de résultats dans un**

générateur de nombres aléatoires (http://stattrek.com/statistics/random-number-generator.aspx) comme dans l'exemple suivant, en supposant qu'il y a 246 pages de résultats :





Le générateur choisit une page au hasard (ici : la page 126). (Veuillez conserver une copie.) Il s'agira de sélectionner le premier résultat de chaque page, jusqu'au nombre de décisions désiré, en commençant par la page sélectionnée au hasard. Supposons qu'ici nous voulons 240 décisions. Il faudra sélectionner le premier résultat des pages 126 à 246, puis des pages 1 à 120. Si le nombre de décisions à sélectionner est supérieur au nombre de pages, il faudra sélectionner la première décision de chaque page, puis la deuxième décision à partir de la page choisie au hasard, jusqu'à l'obtention du nombre de décisions désiré.

Avant de commencer à prélever les décisions qui composeront l'échantillon, il faut **préparer la feuille Excel** comme suit. Dans une première colonne, numérotez chaque ligne jusqu'au nombre de décisions souhaité. Dans une deuxième colonne, indiquez le numéro de la page de résultats d'où proviendra chacune des décisions sélectionnées. Vous verrez qu'il est très laborieux, dans la banque de Soquij, de se rendre directement à la page de résultats 126. Il s'avère plus simple de débuter l'échantillonnage sur la première page, tout en respectant le principe que l'échantillonnage se fait à partir de la page sélectionnée au hasard (ce qui garantit son caractère aléatoire). Dans notre exemple, il faudrait sélectionner les premières décisions des pages 1 à 120 et ensuite des pages 126 à 216 (voir le modèle ci-dessous). Ajoutez à la feuille Excel une colonne intitulée « Référence » et une colonne intitulée « Résumé ».

Vous pouvez maintenant procéder à l'échantillonnage. **Cochez la première décision de chaque page** en faisant défiler les 10 premières pages de résultats. 10 décisions ont été sélectionnées. **Cliquez sur « Consulter »** (en bas de l'écran). **Copiez la référence AZ et collez là dans la feuille Excel** à l'endroit approprié. **Précisez si la décision comporte un résumé (1) ou non (0)**. Faites de même pour chacune des 10 décisions sélectionnées, puis cliquez sur « Retour ». **Cliquez sur « Désélectionner tout » et « Annuler la sélection » au besoin.** Puis, sélectionnez la première décision des 10 pages suivantes, et ainsi de suite. Assurez-vous que le numéro de la page de résultats d'où provient chaque décision correspond à celui indiqué dans la colonne « Page » de la feuille Excel.

	A	B	C	D	E	F	G
1	RESPONSABILITÉ; nombre de décisions à sélectionner: 240; nombre de pages de résultats: 246;						
2	Page sélectionnée au hasard: 126						
3							
4	**Décision**	**Page**	**Référence AZ**	**Résumé**			
5	1	1	AZ-51186223	1			
6	2	2	AZ-51096316	1			
7	3	3	AZ-50765495	1			
8	4	4	AZ-51104205	1			
9	5	5	AZ-51020994	1			
10	6	6	AZ-50860768	1			
11	7	7	AZ-50787355	1			
12	...jusqu'à 120	...jusqu'à 120					
13	121	126	AZ-51186355	0			
14	122	127	AZ-51096768	0			
15	123	128	AZ-50765994	1			
16	124	129	AZ-51104495	0			
17	125	130	AZ-51020223	0			
18	...jusqu'à 240	...jusqu'à 246					
19							

Cela prend 3-4 minutes pour prélever les références de 10 décisions. Pour constituer un échantillon de 2 500 décisions, il faut donc prévoir au moins 17 heures de travail. Il vaut mieux travailler avec 2 écrans, et ne pas interrompre une séance d'échantillonnage en cours. (Deux échantillonneurs pourraient se relayer au besoin.)

À noter que pour le droit du travail, il faudra exclure les décisions en matière d'accidents du travail à l'aide de l'opérateur « SAUF ». Par la suite, il faudra faire un échantillonnage des décisions au sein de cette sous-rubrique, en suivant la même méthode que précédemment et tel qu'indiqué ci-dessous.

Veuillez créer des feuilles séparées dans Excel pour chaque échantillonnage.

9. Critères de recherche et nombre de décisions à sélectionner

Plus précisément, voici les critères de recherche à utiliser pour chaque échantillonnage :

Droit administratif :

- Entre le : « 2011-01-01 et le 2015-12-31 » ET Plan de classification : « ADMINISTRATIF (DROIT) » = 14 639 résultats. **Retenir 550 décisions**

Contrats :

- Entre le : « 2011-01-01 et le 2015-12-31 » ET Plan de classification : « CONTRAT » OU « CONTRAT D'ENTREPRISE » OU « CONTRAT DE SERVICES » OU « OBLIGATIONS » OU « VENTE » = 26 411 résultats. **Retenir 1 025 décisions**

Responsabilité civile :

- Entre le : « 2011-01-01 et le 2015-12-31 » ET Plan de classification : « RESPONSABILITÉ » = 6 132 résultats. **Retenir 250 décisions**

Droit du travail :

- Entre le : « 2011-01-01 et le 2015-12-31 » ET Plan de classification : « TRAVAIL » SAUF « accidents du travail et maladies professionnelles » = 16 672 résultats. **Retenir 645 décisions**

Accidents du travail :

- Recherche entre le : « 2011-01-01 et le 2015-12-31 » ET Plan de classification : « accidents du travail et maladies professionnelles » = 39 172 résultats. **Retenir 30 décisions**.