L'assistant d'anonymisation NOME

Frédéric Pelletier, B.A., LL.B.
LexUM - Université de Montréal
Montréal, Québec, Canada
pelletif@lexum.umontreal.ca

Luc Plamondon, ing. jr., M.Sc.
RALI - Université de Montréal
Montréal, Québec, Canada
plamondl@iro.umontreal.ca

Guy Lapalme, prof. titulaire
RALI - Université de Montréal
Montréal, Québec, Canada
lapalme@iro.umontreal.ca

Résumé

Au Canada, la publication de la jurisprudence sur Internet ne requiert pas l'anonymisation systématique des parties nommées dans les décisions. Cependant, le Législateur exige de préserver l'anonymat de certains participants au système judiciaire. Vu les coûts supplémentaires qu'entraîne l'anonymisation de ces décisions, leur accès libre et gratuit est devenu plus rare au Canada. Afin de contribuer à la réduction des coûts de l'anonymisation, nous développons une application logicielle conçue pour automatiser certaines tâches répétitives effectuées par les éditeurs dans l'anonymisation des décisions. Dans sa version actuelle, cette application baptisée NOME est une macro fonctionnant avec le logiciel MS Word. NOME permet d'automatiser le remplacement des noms mentionnés dans un document par leurs initiales ou par d'autres signes.

Abstract

In Canada, Internet publication of case law does not require a systematic anonymization of parties named in decisions. However, there are statutory requirements that identities of certain participants in the judicial system be protected. Given the supplementary costs entailed by anonymization, free access to these decisions has become somewhat scarce in Canada. In order to contribute to lowering the costs of anonymization, we are developing a software application designed to automate some of the repetitive chores associated with the anonymization of judgments. The current version of this application, called NOME, is an MS Word macro. NOME allows for automating the replacement of names mentioned in a document by their initials or by any other characters.

Table des matières

INTRODUCTION 2

1. CONTEXTE CANADIEN DE L'ANONYMISATION 3
1.1. LA PUBLICATION DES DÉCISIONS DE JUSTICE SUR INTERNET 3
1.2. LA « FRACTURE JURISPRUDENTIELLE » 4
1.3. UN RÔLE ACCRU POUR LES TRIBUNAUX DANS L'ANONYMISATION 5

2. NOME : UNE APPLICATION D'ASSISTANCE À L'ANONYMISATION 7
2.1. DÉMONSTRATION 8
2.2. IMPLÉMENTATION TECHNIQUE 12
2.3. ÉVALUATION 15

CONCLUSION 17

Introduction

[1] Au Canada, la publication libre et gratuite des décisions de justice sur Internet est maintenant généralisée. Cependant, pour certaines juridictions, les décisions rendues dans les domaines pour lesquels l'anonymat de certaines personnes doit être protégé demeurent en grande partie inaccessibles gratuitement, en raison des coûts souvent prohibitifs reliés au travail d'édition requis pour rendre publiables ces décisions. Il est en conséquence utile de développer une application logicielle permettant d'automatiser l'anonymisation des décisions de justice.

[2] Il peut paraître utopique de vouloir développer une telle application. En effet, à supposer qu'elle puisse repérer tous les noms propres de personnes, encore faut-il que cette technologie permette d'établir une distinction entre les personnes dont l'anonymat est requis et celles dont l'identité doit demeurer publique afin de réaliser pleinement le principe de la transparence judiciaire. Ensuite, il est bien connu que la préservation réelle de l'anonymat d'une personne mentionnée dans un document exige non seulement la suppression de son nom, mais aussi de ses données nominatives et d'autres informations qui permettraient de l'identifier. Or, la nature des informations permettant d'identifier une personne varie suivant le contexte d'énonciation d'une décision. Ce contexte ne peut être analysé par une machine seule, bien que celle-ci puisse techniquement obtenir un haut taux de succès. En matière d'anonymisation, le jugement humain doit forcément intervenir, ne serait-ce que pour valider le travail de la machine.

[3] Le développement du prototype actuel de l'application NOME a tenu compte de ces contraintes. D'abord, il s'agit non pas d'un logiciel d'anonymisation automatique, mais bien d'assistance à l'anonymisation. NOME vise à aider l'éditeur qui anonymise à effectuer les tâches répétitives de remplacement des informations dans un document, lui permettant ainsi de concentrer ses efforts et son attention sur la sélection des types de données à remplacer. Dans sa version actuelle de développement, NOME ne permet qu'un remplacement des noms propres, mais les essais menés jusqu'à présent permettent d'emblée de conclure que son utilisation dans un contexte de publication permet de réaliser des gains significatifs de productivité.

[4] Avant de présenter l'application NOME elle-même (2), le présent exposé dresse d'abord un portrait de la situation particulière de l'anonymisation au Canada (1).

1. Contexte canadien de l'anonymisation

[5] Au Canada comme dans la majorité des États développés, la publicité des débats judiciaires est considérée comme conditionnant l'existence même de la vie démocratique. Malgré les pressions souvent exercées par certains justiciables désirant soustraire les aspects gênants ou humiliants de leur vie privée aux regards indiscrets, la Cour suprême du Canada a constamment réitéré la prépondérance des principes constitutionnels de liberté d'expression et de transparence des tribunaux judiciaires sur le droit à la vie privée des justiciables[1]. Évidemment, le Législateur prévoit quelques exceptions lorsque la protection de l'anonymat des participants au système judiciaire favorise une meilleure administration de la justice ou permet de leur éviter un important traumatisme en raison de leur état de vulnérabilité.

1.1. La publication des décisions de justice sur Internet

[6] Les possibilités sans précédent d'accès à distance par Internet aux données judiciaires et ses conséquences sur l'expectative de vie privée des justiciables ne manque pas de susciter des réflexions, au Canada comme ailleurs[2]. S'agissant de la diffusion des décisions de justice, y compris sur Internet, il est peu probable cependant que le principe de leur publication strictement intégrale soit remis en question au Canada, tant on le considère lié – à tort ou à raison – à la réalisation pleine et entière du principe de transparence de la justice. L'anonymisation des décisions demeure exceptionnelle, et ne vise que des personnes bien précises. Dans une décision visée par une restriction quant à la publication de l'identité d'une personne, les médias et éditeurs de jurisprudence ne peuvent publier l'information permettant d'identifier cette personne, mais peuvent rapporter tout autre fait[3]. Notons cependant au passage

---

[1] Voir notamment *Procureur général de la Nouvelle-Écosse c. McIntyre*, [1982] 1 R.C.S. 175.

[2] Voir à ce sujet le document de travail commandé par le Comité consultatif sur l'utilisation des nouvelles technologies par les juges pour le Conseil canadien de la magistrature, « La transparence de la justice, l'accès électronique aux archives judiciaires et la protection de la vie privée » , mai 2003, en ligne: <http://www.cjc-ccm.gc.ca/cmslib/general/OpenCourts-2-FR.pdf>. Au Québec, voir l'*Avis de la Commission d'accès à l'information concernant le Système intégré d'information de justice (SIIJ) présenté par le ministère de la Justice*, dossier 02 17 29, janvier 2004, en ligne : <http://www.cai.gouv.qc.ca/05_communiques_et_discours/01_pdf/a021729.pdf>.

[3] On ne saurait passer sous silence la difficulté bien concrète, pour certaines affaires, de déterminer les informations permettant d'identifier une personne. Le juge Dickson de la Cour suprême du Canada affirmait à ce sujet : « L'intimée a fait valoir en outre que le par. 442(3) [maintenant 486(3) du *Code criminel*] risque de museler les médias parce que, dans certains cas, il est difficile de déterminer quels éléments de preuve permettraient de découvrir l'identité du plaignant. Il y

que plusieurs diffuseurs canadiens de jurisprudence sur Internet, dont IIJCan, atténuent l'effet attentatoire à la vie privée des personnes nommées dans les décisions en bloquant l'indexation de ces décisions par les robots des moteurs de recherche généralistes tels Yahoo! et Google[4]. Ce blocage a pour effet d'empêcher les internautes de trouver accidentellement des décisions judiciaires lorsqu'ils recherchent le nom d'une personne sur ces moteurs généralistes.

[7] La révolution des technologies de l'information permise par l'arrivée du Web dans la vie des citoyens provoque une pression sans précédent pour que les tribunaux prennent en charge la diffusion de leurs propres décisions sur Internet. Puisque les décisions de justice se préparent maintenant directement au format électronique par la magistrature, il est tout à fait naturel pour les tribunaux de diffuser l'ensemble de leurs décisions sur leur site Web. Si la situation à cet égard n'était qu'embryonnaire en 2000[5], il est maintenant possible d'accéder gratuitement par Internet aux décisions de plusieurs dizaines de tribunaux judiciaires et quasi-judiciaires canadiens, soit par le site Web de chaque tribunal, soit par l'entremise du site Web de l'Institut canadien d'information juridique (IIJCan), lequel regroupe l'ensemble des décisions provenant de plusieurs tribunaux canadiens[6].

1.2. La « fracture jurisprudentielle »

[8] Bien avant l'avènement du Web, les fournisseurs commerciaux de banques de données jurisprudentielles informatisées au Canada[7] ont toujours scrupuleusement respecté les restrictions à la publication, tout en publiant l'ensemble des décisions rendues. Les décisions visées par des restrictions à la publication sont identifiées et soigneusement éditées avant leur

---

a donc danger que la presse choisisse de ne pas publier de reportage valable sur certains procès. À mon avis, il suffit de dire à cet égard que les journalistes sont certainement assez compétents pour décider quels renseignements font l'objet de l'interdiction et, dans l'hypothèse contraire, le juge peut préciser dans son ordonnance ce qui ne doit pas être publié. » (*Canadian Newspaper Co. c. Canada (P. G.)*, [1988] 2 R.C.S. 122, ¶ 22).

4 Ce blocage est effectué au moyen d'une norme très simple d'application, volontairement adoptée par l'industrie des moteurs de recherche, en ligne : <http://www.robotstxt.org/wc/exclusion.html>.

5 Daniel POULIN, Frédéric PELLETIER et Bertrand SALVAS, « La diffusion du droit canadien sur Internet », (2000) 102 *R.N.* 189.

6 En ligne : <http://www.iijcan.org>.

7 Voir notamment LexisNexis/Quicklaw, en ligne : <http://www.lexisnexis.ca>, et Westlaw*e*Carswell, en ligne : <http://www.westlawecarswell.com>.

publication, et ce par du personnel éditorial expérimenté et formé en droit. Ces éditeurs commerciaux ont ainsi développé une expertise propre et intégré à leur tarification globale les coûts supplémentaires reliés au travail d'anonymisation.

[9] Cette expertise et ces ressources font cruellement défaut au sein de plusieurs tribunaux[8], lesquels s'abstiennent de diffuser sur leur site Web ou de distribuer à IIJCan les décisions rendues dans les domaines normalement visés par des restrictions à la publication. C'est en particulier le cas pour les décisions rendues dans les affaires de droit de la famille, de protection de la jeunesse et de jeunes contrevenants. Pourtant, ces mêmes décisions peuvent être consultées dans les bases de données commerciales.

[10] Cette situation entraîne en quelque sorte une « fracture jurisprudentielle », c'est-à-dire que l'accès à la jurisprudence récente est généralement libre et gratuit sur Internet, alors pour certains domaines de droit pour lesquels les parties et leurs avocats ont pourtant moins de ressources, l'accès est payant. Compte tenu du rôle primordial joué par le droit prétorien dans le système juridique canadien, il s'agit là pour les justiciables d'une entrave considérable quant à l'accès au droit. Les tribunaux canadiens ont exprimé leur volonté d'y remédier. Mais comment?

1.3. Un rôle accru pour les tribunaux dans l'anonymisation

[11] Au nombre des solutions envisagées afin de permettre aux tribunaux d'assumer pleinement le rôle accru qui leur est imparti dans la diffusion de leurs décisions, nous en retenons deux.

[12] En premier lieu, le Conseil canadien de la magistrature travaille actuellement, en collaboration avec le laboratoire LexUM du Centre de recherche en droit public de l'Université de Montréal (LexUM) et les autorités de divers tribunaux canadiens, à une norme d'anonymisation qui serait commune à tous les tribunaux[9]. Cette norme permettrait de combler le manque d'expertise en

---

[8] Outre les tribunaux du Québec, dont la diffusion des décisions est assurée par une société d'État indépendante (SOQUIJ, en ligne : <http://www.soquij.qc.ca>) et les tribunaux supérieurs fédéraux, lesquels bénéficient du support d'éditeurs officiels de la Couronne, la majorité des tribunaux canadiens qui diffusent leurs décisions le font à même leurs ressources propres.

[9] Une ébauche de cette norme intitulée *Guidelines for the Protection of Identities in Published Decisions* est disponible en langue anglaise, en ligne : <http://externe.lexum.umontreal.ca/pelletif/docs/id_prot.pdf>.

matière éditoriale souffert par le personnel des tribunaux. Elle viserait deux objectifs principaux à cet égard. Tout d'abord, elle permettrait au personnel des tribunaux de savoir précisément quels types de données nominatives ou autres renseignements personnels devraient être retirés afin de protéger l'anonymat d'une personne nommée dans une décision avant sa distribution au public. Ensuite, à plus long terme, elle permettrait de conscientiser la magistrature sur la nécessité de rédiger les motifs de manière à éviter d'y inclure des informations superflues ou inutiles à la bonne compréhension du raisonnement juridique qui sous-tend la décision.

[13] La seconde solution envisagée est le développement d'une application informatique permettant d'automatiser l'anonymisation. C'est de cette application dont il sera maintenant question.

2. NOME : une application d'assistance à l'anonymisation

[14] À l'automne 2003, le LexUM et le laboratoire de Recherche appliquée en linguistique informatique de l'Université de Montréal (RALI) entreprenaient un ambitieux projet de recherche dans le but de développer un logiciel d'anonymisation automatique[10]. Ce logiciel devait permettre d'automatiser les tâches suivantes :

1. Déterminer, parmi les personnes nommées dans une décision, lesquelles voient leur anonymat protégé par la Loi;
2. Identifier les passages du texte qui constituent des données nominatives sur chaque personne déterminée en (1) : nom, date de naissance, coordonnées uniques, etc.;
3. Identifier les passages du texte qui constituent des renseignements personnels sur chaque personne déterminée en (1), et qui présentent un risque significatif quant à la divulgation de l'identité de cette personne;
4. Supprimer les données nominatives et les renseignements personnels identifiés en (2) et (3) en les remplaçant par des initiales, des points de suspension, etc.

[15] L'orientation prise dans ce projet n'est pas de développer une application à 100% automatique, vu la quasi-impossibilité pour la machine de tenir compte des particularités de chaque cas quant aux combinaisons de renseignements qui peuvent mener à l'identification d'une personne. C'est pourquoi il est plus juste de désigner NOME comme une application d'assistance à l'anonymisation.

[16] Dans sa version actuelle, NOME détecte les noms propres de personnes dans un document puis suggère des initiales susceptibles de remplacer ces noms[11]. L'utilisateur peut choisir quels noms, parmi ceux suggérés, seront effectivement remplacés par leurs initiales. D'un seul clic, l'utilisateur commande ensuite à NOME d'effectuer tous les remplacements dans le document. Enfin, NOME permet à l'utilisateur d'accepter ou refuser chaque remplacement en mettant à profit la fonction de suivi des modifications du logiciel Word. Ce prototype a été financé par la Société québécoise d'information juridique (SOQUIJ), laquelle effectue l'anonymisation des

---

[10] Luc PLAMONDON, Guy LAPALME et Frédéric PELLETIER, « Anonymisation de décisions de justice », *In* Bernard Bel et Isabelle Martin (éd.), TALN 2004, *XIe Conférence sur le Traitement Automatique des Langues Naturelles*, p. 367-376, Fez, Maroc, Avril 2004.

[11] Le remplacement des noms par leurs initiales est une pratique imparfaite mais universellement répandue au Canada. Elle s'avère surtout utile pour désigner informellement les affaires par un intitulé distinct (on dira p. ex. « l'affaire *M.B. c. Québec* »).

décisions émanant des tribunaux québécois. Nous dressons ici le bilan de cette première phase du projet, qui ne touche que les noms de personnes.

2.1. Démonstration

2.1.1. Remplacement des noms dans un document

[17] Afin d'illustrer les possibilités offertes par NOME, décrivons les étapes d'une session type d'anonymisation effectuée avec cette application, à l'aide de la Figure 1.

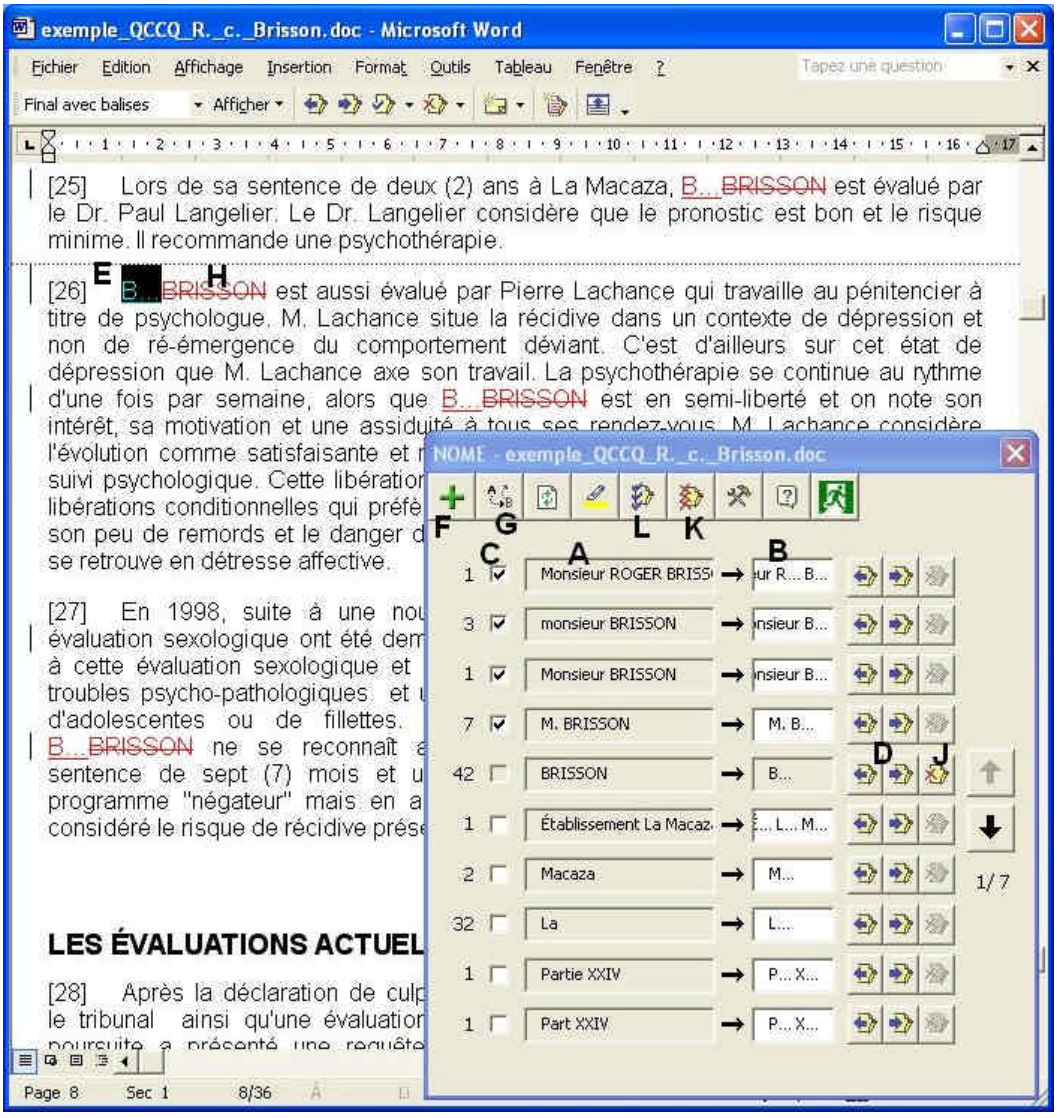

Figure 1 : Fenêtre principale de NOME

1. Après avoir ouvert le document avec Word, l'utilisateur exécute la macro NOME. Celle-ci lance une analyse sur le document pour identifier les noms propres qui y sont mentionnés (voir Figure 1, point **A**).
2. NOME ouvre une fenêtre contenant une liste des noms propres qui ont été identifiés. Pour chaque nom, NOME suggère automatiquement une chaîne de remplacement (**B**). Si l'utilisateur désire modifier les initiales suggérées, par exemple lorsque deux personnes différentes portent les mêmes initiales, il peut le faire directement dans les boîtes de texte correspondantes;
3. L'utilisateur coche les noms des personnes dont la protection de l'anonymat est requis (**C**);
4. Si l'utilisateur désire connaître le contexte dans lequel une personne est nommée afin d'être en mesure de décider s'il y a lieu de protéger son anonymat, il appuie sur les boutons de navigation (**D**) puis NOME défile le document jusqu'à l'occurrence précédente/suivante du nom dans le texte du document (**E**);
5. En cas d'oubli par NOME d'une instance du nom d'une personne dont l'anonymat est requis, l'utilisateur peut sélectionner cette instance dans le document et l'ajouter à la liste de NOME en appuyant sur le bouton « Ajouter », sur lequel figure un signe **+** (**F**);
6. Lorsque tous les noms à remplacer ont été cochés, l'utilisateur appuie sur le bouton « Remplacer » (**G**) et NOME remplace les occurrences des noms cochés par les chaînes de remplacement contenues dans les boîtes de texte correspondantes. Ce faisant, NOME utilise la fonctionnalité de « Suivi des modifications » de Word, de sorte que les modifications apportées au document soient bien marquées (**H**);
7. Si l'utilisateur désire vérifier les remplacements effectués, il appuie sur les flèches de navigation (**D**) et NOME défile le document jusqu'à l'occurrence précédente/suivante (**E**) du nom remplacé. L'utilisateur peut annuler les remplacements un à un en appuyant sur le bouton d'annulation (**J**). L'utilisateur peut aussi rejeter en bloc toutes les modifications apportées au document (**K**);
8. Une fois le travail terminé, l'utilisateur accepte toutes les modifications (**L**) et il peut sauvegarder et distribuer le document anonymisé.

2.1.2. Adaptation des paramètres de détection des noms

[18] Les règles et pratiques de rédaction, ainsi que le vocabulaire propre aux décisions de justice, demeurent relativement uniformes pour un tribunal donné, mais ils peuvent varier d'un tribunal à l'autre et d'une langue à l'autre. En conséquence, il est essentiel que les paramètres de détection des noms propres utilisés par NOME soient facilement adaptables par l'utilisateur.

[19] L'adaptation des paramètres de détection des noms propres prend la forme de 3 listes de mots stockés dans des fichiers Word. Il s'agit de la liste d'inclusion, de la liste d'exclusion et de la liste des titres de civilité. Ces trois fichiers se modifient comme tout document Word, par le préposé à l'anonymisation ou par ses superviseurs. Plusieurs listes peuvent être créées afin de tenir

compte des particularités de traitement d'un ensemble particulier de décisions. L'utilisateur n'a qu'à choisir la liste appropriée avant de traiter un document. La Figure 2 illustre la fenêtre de réglage qui apparaît lorsque l'utilisateur clique sur le bouton « Réglages ».

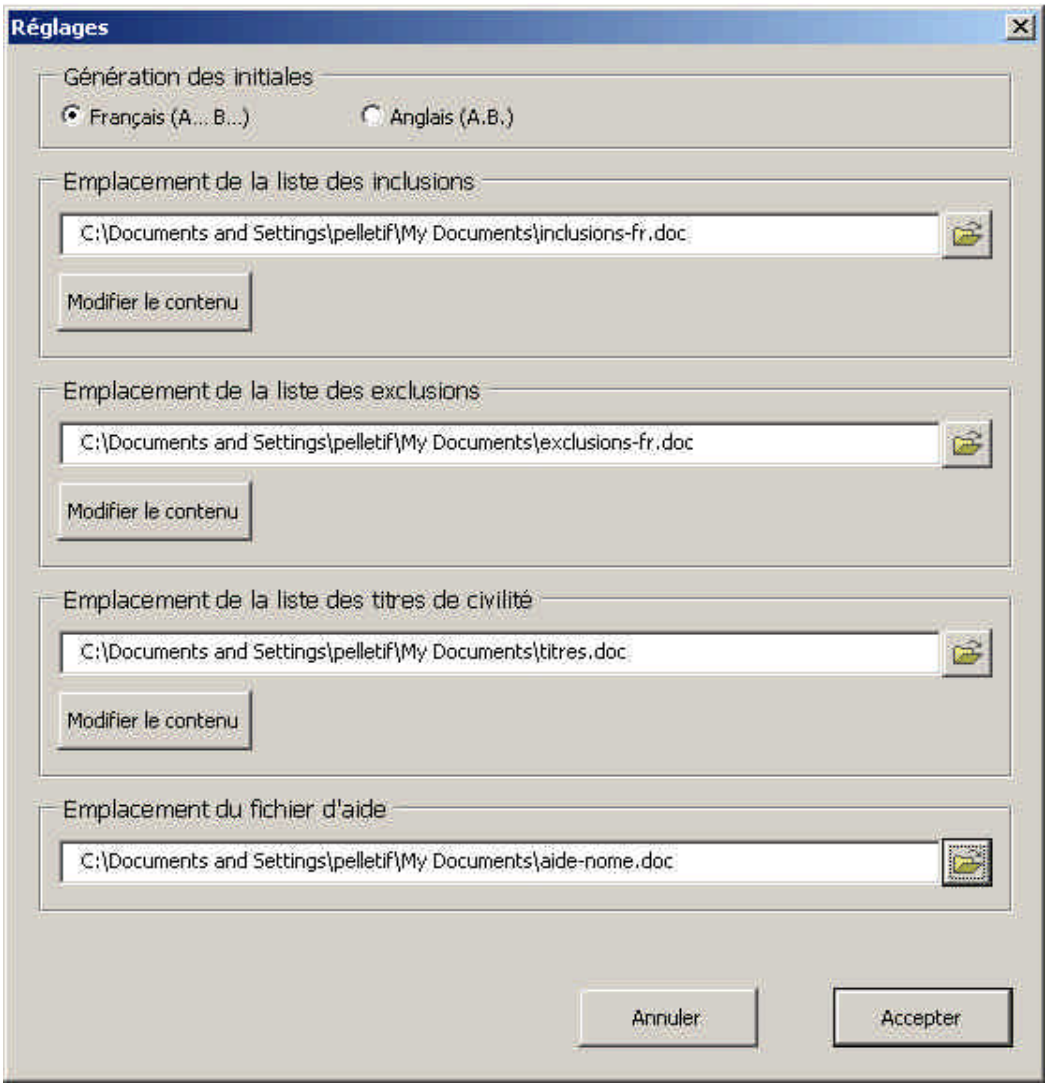

Figure 2 : Fenêtre de réglages de NOME

[20] Cette fenêtre permet notamment à l'utilisateur de choisir et modifier les fichiers de la liste des inclusions, des exclusions et des titres de civilité.

Liste des inclusions

[21] La liste des inclusions contient les mots qui devraient toujours être détectés par NOME. Ainsi, lors de l'analyse de chaque nouveau document, NOME ajoutera systématiquement ces mots à la

liste de noms présentée à l'utilisateur. Cette liste peut être utilisée pour caviarder des renseignements autres que des noms de personnes. Par exemple, l'ajout de « né le » à la liste d'inclusion provoquera l'ajout de « né le » dans la liste des noms présentée à l'utilisateur si cette expression est présente dans le document. L'utilisateur saura alors que le document contient probablement une date de naissance – information qui doit normalement être remplacée pour protéger l'identité d'une personne. L'utilisateur pourra alors utiliser les boutons de navigation pour accéder directement au passage détecté et caviarder manuellement la date. La détection automatique plus raffinée de tels renseignements figure parmi les projets de développement futurs de NOME.

[22] Rappelons que si un nom qui n'est pas détecté par NOME n'est pas susceptible d'apparaître dans plusieurs documents, l'utilisateur ne devrait pas utiliser la liste d'inclusion, mais bien le bouton « Ajouter » pour ajouter ce nom à la liste en cours (Figure 1, point **F**), sans que cette inclusion n'affecte l'analyse d'autres documents.

Liste des exclusions

[23] La liste des exclusions contient les mots qui ne devraient jamais être détectés comme des noms de personnes à anonymiser par NOME. Pour un tribunal donné, par exemple, le nom du tribunal ou du district judiciaire devrait être systématiquement exclu, tout comme les noms géographiques. L'ajout d'un mot à la liste d'exclusion permanente empêche NOME de suggérer toute expression contenant un de ces mots. Par exemple, l'ajout de « Canadian » dans la liste d'exclusion exclut « Canadian Embassy », « Canadian Legal Information Institute », etc. L'ajout de « maître » dans la liste exclut tous les noms d'avocats, à condition qu'ils soient précédés de ce titre, par exemple « maître Roger Grenier », « maître Grenier », etc.

Liste des titres de civilité

[24] La liste des titres de civilité contient les mots qui précèdent généralement des noms de personnes dans un contexte formel, tels « monsieur », « Mrs. », « docteur » ou « maître ». Cette liste est utilisée par NOME pour améliorer la qualité de la détection des noms. La présence de ces indicateurs signale que le ou les mots qui suivent constituent un nom de personne, et que ces

mots ne doivent pas faire partie des initiales suggérées par NOME comme le sont les prénoms et noms de famille.

2.2. Implémentation technique

2.2.1. Contraintes de conception

[25] Vu la nécessité de développer un prototype de logiciel d'assistance à l'anonymisation rapidement utilisable en contexte de production, nous nous sommes fixé comme contrainte qu'il soit facilement compatible avec Word, l'un des logiciels de traitement de texte les plus utilisés. Nous avons donc utilisé le langage Visual Basic et ses bibliothèques standard pour créer une macro Word. La macro s'installe facilement puisqu'elle tient en un seul fichier modèle Word (« .dot »). En contrepartie, Visual Basic est un langage interprété qui nous limite dans la complexité des algorithmes que nous pouvons utiliser, principalement en raison de la lenteur d'exécution qui en résulterait. De plus, par souci de simplicité et de portabilité, nous avons exclu le recours à des ressources externes comme des dictionnaires, des lexiques de plusieurs milliers de mots, des modèles statistiques et des logiciels de traitement automatique de la langue.

[26] De ce fait, nous n'avons pas utilisé les techniques habituelles propres au domaine du traitement de la langue, techniques qui auraient permis de comprendre partiellement le sens des mots. Cela explique que NOME repose essentiellement sur la présence de majuscules au début des mots. Tous les mots débutant par une majuscule n'étant pas des noms propres, nous appliquons des règles simples d'élimination pour ne présenter à l'utilisateur que les mots les plus susceptibles d'être des noms propres. Bien que cette approche basée sur les majuscules produit du « bruit » ou des « faux positifs » dans la détection de mots, elle a l'avantage de pouvoir être utilisée directement pour toutes les langues qui adoptent les mêmes conventions d'écriture que le français et l'anglais en regard des noms propres.

2.2.2. Détection des noms de personnes

[27] Pour les raisons exposées à la section précédente, NOME se base principalement sur la présence de majuscules au début des mots pour déterminer quelles séquences de mots du document forment des noms propres. Tous les mots débutant par une majuscule n'étant pas

nécessairement des noms propres (c'est le cas du premier mot de chaque phrase, de certaines abréviations, etc.), des règles simples d'élimination sont appliquées afin de présenter à l'utilisateur une liste de noms la plus courte possible. La prudence est cependant de mise car il est de loin préférable de présenter trop de noms et de laisser l'utilisateur faire le tri, plutôt que d'ignorer des mots susceptibles d'être des noms propres.

[28] L'algorithme de détection des noms s'exécute en 5 étapes :

1. le balayage du document pour en extraire les séquences pertinentes;
2. le filtrage de ces séquences pour ne conserver que les noms valides (dits noms longs);
3. la récupération des noms courts par la décomposition des noms longs;
4. la récupération des noms entièrement en majuscules à partir des noms longs;
5. l'ajout des inclusions explicites.

Balayage du document pour en extraire les séquences pertinentes

[29] La règle de base est que toute séquence du document d'au moins un mot débutant par une majuscule constitue un nom potentiel. Le ou les premiers mots de la séquence peuvent commencer par des minuscules s'ils font partie de la liste des titres de civilité. Tant que la séquence n'est pas brisée par autre chose qu'un mot débutant par une majuscule, le mot suivant est rattaché à la séquence formant le nom propre. À ce stade, l'algorithme accepte des séquences comme « Jean Leduc Levasseur », « Georges », « LA REINE », « madame François », « Lilly D'Arcy », le premier mot de chacune des phrases, etc. Les mots peuvent être séparés par certains signes de ponctuation : « - » pour les noms composés, « . » après une initiale, etc. Donc, en plus des séquences sans ponctuation données plus haut, l'algorithme accepte aussi les séquences comme « Dr. Walter », « Jean-Louis C. Garon », « A.G. », etc.

Filtrage des noms longs

[30] Seules les séquences les plus prometteuses parmi celles extraites à l'étape précédente sont conservées. Si la séquence commence par un ou plusieurs titres de civilité, elle doit satisfaire les conditions suivantes :

- Ne pas contenir une expression de la liste d'exclusion – par exemple, « Maître Charles » est rejeté si « Maître » ou « Charles » fait partie de la liste d'exclusion;
- Contenir au moins deux lettres contiguës, ce qui exclut les initiales – par exemple « monsieur A.G. » est rejeté, mais « monsieur A.G. Yu » est retenu.

[31] Si la séquence ne commence pas par un titre de civilité, les conditions d'acceptation sont plus contraignantes :

- Ne pas contenir une expression de la liste d'exclusion;
- Être formé d'au moins deux mots séparés par une espace, ce qui élimine tous les mots seuls, principalement le premier mot d'une phrase;
- Contenir au moins deux lettres contiguës, ce qui élimine les initiales – par exemple, « A.G. » est rejeté;
- Ne pas être entièrement en majuscules, ce qui élimine les expressions procédurales se trouvant en-tête de certaines décisions – par exemple, « SOUS LA PRÉSIDENCE DE » est rejeté.

[32] Les séquences qui satisfont aux critères de filtrage sont appelées « noms longs » car ils sont tous composés d'au moins deux mots.

Récupération des noms courts

[33] Les critères imposés aux noms longs obligent à ignorer les noms courts comme les prénoms employés seuls (p. ex. « Georges ») et les noms de famille non précédés d'un titre de civilité (p. ex. « les Langlois »). Considérer tous les mots débutant par une majuscule comme étant des noms produirait trop de « bruit » dans les mots détectés car le premier mot de chaque phrase serait inclus. NOME fait le compromis suivant : accepter les noms courts formés d'un seul mot à condition qu'ils fassent partie d'un nom plus long cité ailleurs dans le document. Ainsi, chaque nom long est décomposé en ses parties autres que les titres de civilité (p. ex. « Monsieur Jean-Marie Lavoie » est décomposé en « Jean-Marie » et « Lavoie ») et si une partie a été vue parmi les séquences prometteuses, elle est incluse dans la liste des noms (p. ex. si « Lavoie » a été vu seul dans le document, il est ajouté; si « Jean-Marie » n'a pas été vu seul, il n'est pas ajouté).

Récupération des noms entièrement en majuscules

[34] Les critères imposés aux noms longs obligent à ignorer les suites de mots entièrement en majuscules car les expressions procédurales placées en en-tête des décisions sont parfois écrites entièrement en majuscules, sans toutefois constituer des noms propres (p. ex. « SOUS LA PRÉSIDENCE DE »). Si toutefois une séquence entièrement en majuscules est présente sous une forme acceptable parmi les noms longs ou courts validés jusqu'à maintenant, elle est ajoutée à la liste des noms valides (p. ex. « JEAN PELLETIER » est accepté si « Jean Pelletier » figure déjà dans la liste des noms validés).

Ajout des inclusions

[35] Les mots compris dans la liste des inclusions et qui sont présents dans le document sont ajoutés à la suite des noms.

2.3. Évaluation

[36] Nous avons évalué NOME de façon intrinsèque, puis extrinsèque.

2.3.1. Évaluation intrinsèque

[37] Nous avons effectué une comparaison systématique entre l'anonymisation humaine et les suggestions initiales faites par NOME après un premier balayage du document. L'échantillon choisi était quatre décisions rendues en langue anglaise par la Cour supérieure de justice de l'Ontario en matière de droit de la famille, pour un total de 18 000 mots. Nous disposions des versions originales et des versions anonymisées par des humains. L'anonymisation humaine avait entraîné un total de 197 modifications aux quatre documents. Ces modifications portaient sur 27 éléments d'information différents, dont 24 étaient des noms de personnes. Les trois autres éléments étaient des dates de naissance. Après le premier balayage effectué sur les documents originaux, NOME a présenté 122 des 197 occurrences de noms à l'utilisateur (62 %) et a correctement identifié 13 des 24 noms de personnes (54 %).

[38] La moitié des noms propres ignorés par NOME sont des prénoms d'enfants formés d'un seul mot débutant par une majuscule (p. ex. « Rodney ») et qui n'apparaissent pas dans des noms longs. Par exemple, si « Rodney Dovers » avait figuré dans la décision, « Rodney » apparaissant seul aurait été détecté. L'autre moitié des noms propres ignorés apparaissent dans les en-têtes des décisions et sont écrits entièrement en majuscules, ce qui est une pratique traditionnelle des tribunaux (p. ex. « RONALD APPLETON, demandeur »). Ce type d'oubli commis par NOME n'est pas critique en soi car le préposé à l'anonymisation peut facilement détecter ces noms en examinant l'en-tête des décisions. Cependant, l'oubli des prénoms d'enfants est plus sérieux car ceux-ci peuvent être dissimulés n'importe où dans une décision comportant souvent des dizaines de pages. Une relecture attentive du document s'impose. L'ajout de prénoms connus à la liste

d'inclusion pourrait remédier temporairement à ce problème dans la mesure où la vitesse d'exécution de NOME demeure acceptable.

[39] Au demeurant, même si une relecture de vérification est toujours nécessaire après avoir effectué les remplacements avec NOME, le temps nécessaire à l'anonymisation des décisions est significativement réduit, comme l'a démontré notre évaluation extrinsèque.

2.3.2. Évaluation extrinsèque

[40] Nous avons évalué NOME en contexte réel de production, dans le cadre des activités de l'Institut canadien d'information juridique (IIJCan), dont les éditeurs anonymisent en moyenne 10 décisions par jour. Les préposés à l'anonymisation sont d'avis que le logiciel est très utile pour anonymiser les décisions motivées, en particulier celles qui comportent plus de deux ou trois pages, ce qui constitue la norme en droit canadien. NOME est devenu indispensable pour anonymiser les décisions comportant plusieurs dizaines de pages, lesquelles contiennent souvent un nombre impressionnant d'occurrences de noms propres. Par contre, NOME s'avère moins utile pour les décisions peu ou pas motivées, c'est-à-dire qui ne comportent qu'un en-tête, une ou deux lignes puis le dispositif, puisque dans ces cas, les noms propres se repèrent et se remplacent facilement sans automatisation.

[41] Le fait que NOME balaie le document pour en détecter automatiquement les noms de personnes, que ces noms soient présentés sous la forme d'une liste à cocher et que les initiales soient automatiquement suggérées évite les fastidieuses opérations de recherche et remplacement traditionnellement effectuées par les préposés à l'anonymisation, qui utilisent NOME quotidiennement à l'IIJCan. Le gain de productivité provient du fait que le risque d'oublier des noms propres ou des occurrences de ces noms est considérablement réduit, et que les remplacements sont automatiquement exécutés. L'attention humaine n'est pas divertie par des opérations triviales et répétitives; elle peut conséquemment se concentrer sur la détermination des personnes à anonymiser et des renseignements de tous types permettant d'identifier cette personne.

Conclusion

[42] La phase actuelle de développement de NOME visait principalement à évaluer s'il était possible d'automatiser le processus d'anonymisation des décisions de justice, compte tenu des contraintes techniques et humaines imposées par un tel travail. Il est maintenant démontré qu'une telle application est utile, en particulier pour les décisions comportant plus de 2 ou 3 pages, comme c'est très souvent le cas pour les jugements motivés rendus au Canada.

[43] Il est permis de souhaiter que les prochaines versions de NOME, intégrant l'identification des dates, adresses et autres informations nominatives, seront d'autant plus utiles pour assister les éditeurs dans l'anonymisation. Cependant, le fait de n'utiliser que le langage Visual Basic dans la programmation de NOME limite la possibilité d'utiliser des algorithmes plus complexes et raffinés. Pour les développements ultérieurs de l'application, il faudra sans doute envisager l'utilisation d'autres langages complémentaires de programmation afin de bénéficier de tout le potentiel offert par les techniques actuelles de traitement automatisé de la langue.