Numen et le Panama Paperless au quotidien

Retours sur les enjeux techniques des Panama Papers

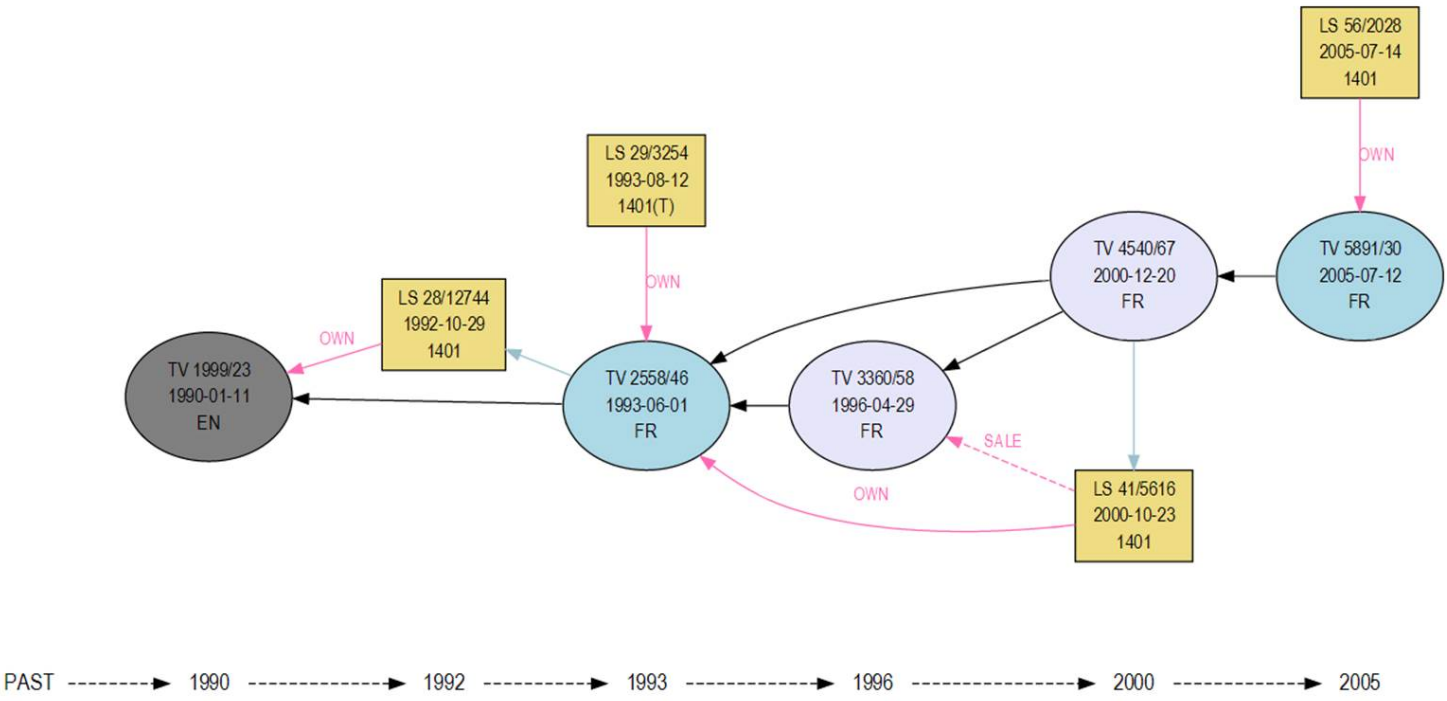
Par Cédric Sylvestre, Directeur général de Numen Digital

De nombreux articles sont parus récemment sur les dessous techniques de l'affaire dite des *Panama papers*. Où l'on apprend que l'exploitation d'un corpus de 2.6 téraoctets de données brutes rassemblant la correspondance interne du cabinet Mossack Fonseca, des copies de documents réglementaires liés à la création de sociétés, et quelques registres récapitulatifs a nécessité l'emploi de nombreuses compétences et outils : capacité à exploiter des documents sources de formats très divers (PDF image et texte, format word, mails…) ; récupération de la couche texte des contenus image par des procédés d'OCR ; création de métadonnées associées à chaque document pour faciliter le tri ; extraction de données spécifiques dans le contenu des documents grâce à des procédés de *text mining* avec recherche d'entités nommées ; recherche de motifs textuels en utilisant des techniques tolérantes aux erreurs de transcription humaines ou OCR ; utilisation de référentiels de sources externes (registres de commerce, sources en accès libre sur internet…) confrontés aux documents du cabinet afin de croiser et recouper les données extraites ; établissement de liens entre des noms de familles, d'administrateurs, d'entreprises ; représentation graphique des liens ainsi établis pour révéler et illustrer toute la structure intentionnellement dissimulée par le mécanisme des sociétés écran.

Ces publications ont eu pour effet de révéler au grand public l'existence d'un ensemble de techniques dites d'investigation numérique (*digital forensics*) utilisables sur corpus massifs, qui constituent une branche particulière des technologies « Big Data ».

L'ensemble de ces procédés est en fait quotidiennement utilisé par la société Numen pour le compte de nombreux clients.

Par exemple, dans le cadre du projet de mise en place d'un cadastre à l'Ile Maurice, Numen a travaillé aux côtés d'Infoterra, filiale d'Airbus Defence & Space, afin de constituer un corpus numérique de l'ensemble des sources relatives à la propriété foncière (actes notariés, rapports d'arpentage, registres…) qui n'avaient jamais été jusque-là rapprochées. Numen a ainsi numérisé plus de 12 millions de pages et de plans aux formats multiples et exotiques (jusqu'au format A0) ; traitement des images (redressement, binarisation, reconstitution de caractères fragmentés…) ; récupération du contenu par tous procédés (saisie, OCR…). Une fois effectuée la récupération de la couche texte, son traitement transverse par text mining et le rapprochement des motifs identifiés avec les données extraites des registres numérisés et des quelques bases de données externes existantes ont permis le rapprochement automatique de « familles » d'actes notariés et de rapports d'arpentage, regroupant ainsi toute l'information disponible sur l'histoire d'une même parcelle originelle et de ses divisions au cours du temps. Les résultats ont été présentés sous forme d'une collection de 300 000 « graphes d'historique » qui totalisent plus de 1 500 000 liens qualifiés. Ces graphes sont affichés dans un navigateur Web et permettent par simple clic d'accéder aux documents numérisés qui décrivent les transactions. Outre la similitude des techniques utilisées, l'analogie frappante avec l'affaire *des Panama papers* réside dans l'identification d'une structure cachée, qui ne peut être révélée que par l'analyse transverse et automatisée d'un vaste corpus documentaire.



La mise à disposition d'un tel outil auprès de la profession notariale par le Gouvernement Mauricien a révolutionné la pratique de « l'établissement de propriété ». Jusqu'à récemment, trois mois d'enquête étaient en moyenne nécessaires pour traiter un dossier. Depuis la mise en œuvre de ce nouveau système « zéro papier », les clercs de notaires accèdent désormais via une interface Web à l'intégralité de l'information disponible. Ils peuvent consulter les documents à l'écran et imprimer à la demande les extraits intéressants, ce qui réduit le processus à quelques heures de travail sur support dématérialisé.

D'une manière plus générale, l'avènement du traitement de masse de l'information permet à Numen de mettre en avant son agilité à traiter des documents non-structurés de formats très divers, dont l'analyse transverse permet d'extraire sur la masse des informations indécelables jusque-là. D'autres exemples concernent l'analyse de la cartographie des fournisseurs du groupe Foncia et de ses conditions d'achat dans le secteur de l'énergie, réalisé par text-mining sur un corpus de 3 millions de factures issues de 220 cabinets distincts. Ou, plus simplement, la constitution d'un référentiel à jour des trésoreries françaises par analyse transverse de plusieurs centaines de milliers de procédures civiles exécutoires concernant plus de 20 banques distinctes des groupes BPCE et Crédit Agricole.

A chaque fois, on retrouve le même contexte (documents non-structurés, sources diverses, formats hétérogènes, données bruitées ou approximatives mais heureusement massives et redondantes) et le même arsenal de technologies (traitement du texte et de la langue, maîtrise du rapprochement de données approximatives, analyses heuristiques, filtrage et apprentissage statistique) qui constituent l'un des cœurs de métier de Numen.

Numen met ainsi tous les jours à disposition de ses clients ses outils et ses experts afin de les aider à extraire des informations stratégiques pour le développement de leur entreprise.