**SmartExtract : une plateforme de capture de données inspirée par le data-mining, un outil révolutionnaire**

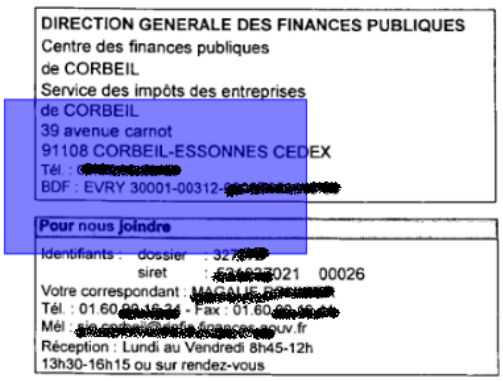
Par François Chahuneau
Directeur des technologies chez Numen

*SmartExtract* est la plateforme de capture de données développée par Numen pour couvrir tous les besoins de son activé BPO (Business Process Outsourcing)/DPO (Document Process Outsourcing). Cette plateforme, qui intègre le moteur OCR ABBYY, combine de multiples technologies « made by Numen » dans les domaines du traitement et de l'analyse d'image, de la classification automatique sur critères mixtes (contenus/mise en page), de la recherche floue et de l'apprentissage machine. *SmartExtract* s'appuie sur *SmartGED*, le moteur de GED CMIS développé par Numen qui couvre tous les besoins liés au stockage, au workflow et au reporting de production. Une des particularités remarquables de *SmartExtract* est son IHM *WebCapture* 100% Web, développée en technologie « client-riche » AngularJS.

L'originalité de la conception de ***SmartExtract*** réside notamment dans l'omniprésence de techniques empruntées au data-mining, jusque dans le fonctionnement des outils interactifs manipulés par les opérateurs. Pour illustrer ce point, nous prendrons pour exemple l'application de cette plateforme au traitement BPO des Procédures Civiles Exécutoires (Avis à Tiers Détenteurs, Oppositions Administratives, etc.) que Numen assure quotidiennement pour le compte de 22 banques.

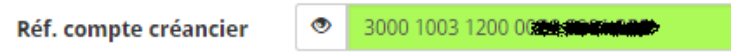
L'une des opérations d'extraction de données réalisée dans ce cadre est l'obtention du « **RIB créancier** », à confronter à un référentiel des trésoreries et intermédiaires de justice opérant le recouvrement des créances. Dans les cas où la mise en page des documents ne correspond pas à un modèle stable, ou dans les rares cas où les tâches automatiques de localisation de l'information recherchée ont échoué, l'opérateur doit effectuer une « capture au lasso » pour extraire la donnée en la *désignant* sur la page, ce qui permet de récupérer le texte OCR correspondant. Toutefois, il serait totalement superflu pour l'opérateur de perdre du temps à effectuer une localisation précise, du fait de mécanismes de data-mining embarqués dans l'outil.

Dans l'exemple ci-dessous, l'opérateur a désigné à la souris une large zone (en bleu) qui intercepte l'information recherchée :

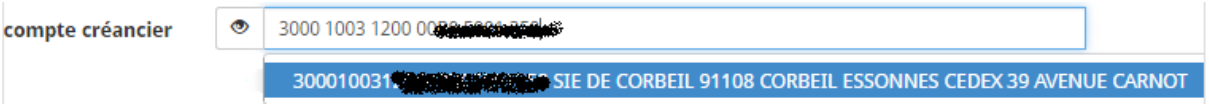

DIRECTION GENERALE DES FINANCES PUBLIQUES
Centre des finances publiques
de CORBEIL
Service des impôts des entreprises
de CORBEIL
39 avenue carnot
91108 CORBEIL-ESSONNES CEDEX
Tél. : [redacted]
BDF : EVRY 30001-00312-[redacted]

Pour nous joindre
Identifiants : dossier : 327[redacted]
siret : [redacted]021 00026
Votre correspondant : MAGALIE [redacted]
Tél. : 01.60.[redacted] - Fax : 01.60.[redacted]
Mél : [redacted]
Réception : Lundi au Vendredi 8h45-12h
13h30-16h15 ou sur rendez-vous

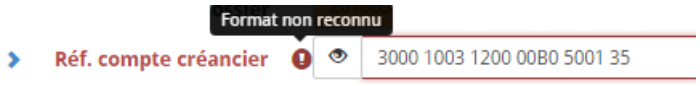
**Au moment où le bouton de la souris est relâché, le RIB extrait apparaît directement dans le formulaire de capture associé au document**. Il s'agit ici d'un RIB valide, comme l'indique l'absence de coloration rouge sur le contour du champ de formulaire.




En appuyant sur la touche RETURN, on affiche un menu déroulant qui permet de constater que ce RIB est bien connu du référentiel des trésoreries. **La validation de l'unique proposition permet d'importer instantanément dans le formulaire l'ensemble des champs associés (nom, adresse du service, etc.)**




Inversement, tout écart d'un caractère ou plus déclencherait instantanément une alerte de ce type :




**Si la confrontation d'une donnée à un référentiel est un mécanisme familier, l'ensemble des opérations qui permettent de récupérer la bonne valeur depuis le texte caché OCR l'est moins**.
En effet, dans le dixième de seconde qui s'écoule entre le moment où l'on désigne la zone et celui où la valeur formatée du RIB apparaît dans le formulaire, les opérations suivantes se déroulent :

- Détermination des chaînes de caractères OCR en intersection avec la sélection
- Recherche de la donnée selon les principaux motifs alphanumériques de représentation des RIB rencontrés dans les PCE, ceux-ci présentant une grande variabilité (présence ou non d'un préfixe IBAN, répartition variable d'espaces ou de séparateurs). Ces motifs multiples sont décrits de manière très compacte par une *expression régulière*, du type de celles utilisées en data-mining.
- Suppression des espaces et séparateurs
- Transcodification d'autorité pour certains caractères jamais présents dans un RIB (O => 0, I => 1), ceci corrigeant statistiquement certaines erreurs OCR
- Calcul de la clé RIB et confrontation aux deux derniers caractères du code
- Insertion d'espaces dans la valeur affichée pour plus de lisibilité

***SmartExtract* implémente ainsi le concept de « masque d'extraction intelligent », pour lequel le zonage géométrique ne constitue qu'une première indication, largement complétée par les mécanismes de data-mining qui se déroulent en temps réel.**

Cette conception tire parti de l'expérience de Numen dans de nombreux projets de data-mining complexe sur archives patrimoniales ou administratives, aujourd'hui transposée dans le domaine du BPO.