 <b>Microsoft Presidio</b>	Système d'exploitation :	Multiplateforme
	Développé par :	Microsoft
Open source, MIT License	Personne de contact :	katy.fokou@smals.be

[Presidio](#) est un outil en python qui permet la protection et l'anonymisation des informations personnelles identifiables (PII) telles que le nom, l'adresse, le numéro de carte de crédit, le numéro de téléphone, ... L'anonymisation se fait en deux étapes (figure 1): l'identification des PII et l'anonymisation de celles-ci. Pour cela, Presidio est composé de 2 modules principaux:

- *Presidio Analyzer* pour **l'identification des PII**. Ce module est constitué de sous-modules appelés *Recognizers* qui utilisent diverses méthodes pour détecter les PII : le [NER](#) (*machine learning*), les expressions régulières, les terminologies. De base, le système contient plusieurs *recognizers* qui permettent de détecter des PII dans un texte et un modèle de langue de l'outil [spaCy](#) pour le NER. Le module est personnalisable, on peut y ajouter des regex et des modèles NER propres.
- *Presidio Anonymizer* pour **l'anonymisation des PII identifiées**. L'anonymisation se fait de plusieurs manières, de la suppression totale des PII au remplacement de celles-ci par un code *hash*. Tout comme *l'Analyzer*, ce module est personnalisable.

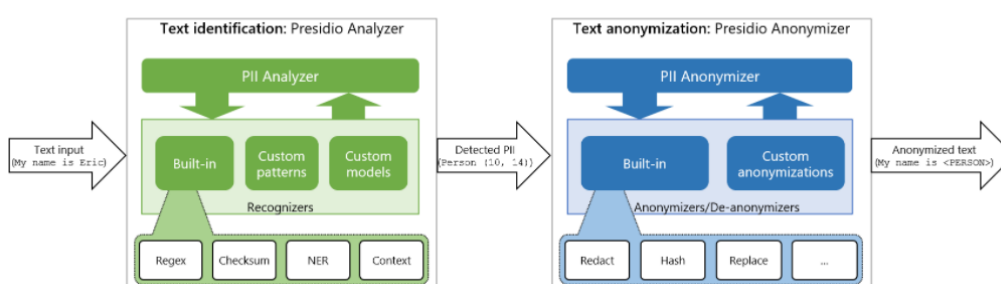
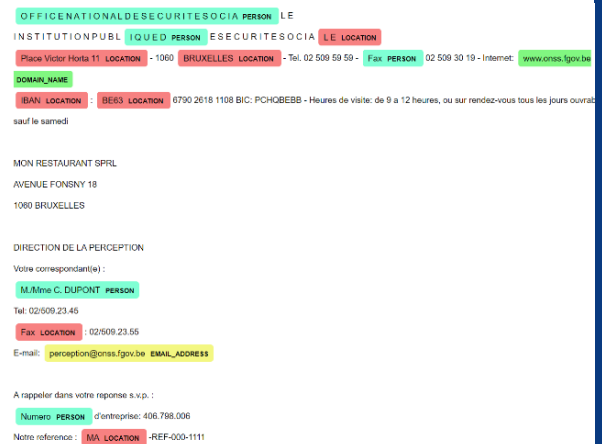


Fig. 1: Flux d'anonymisation

Presidio est un produit open source à recommander pour l'anonymisation des données textuelles. Presidio comme beaucoup de produits ne permet pas *off-the-shelf* d'identifier les données PII dans les documents de nos clients. Cependant, la plateforme est facile à customiser pour intégrer nos spécificités.

Les premiers tests portent sur l'identification des PII. En premier, nous testons les fonctionnalités de base de l'Analyzer avec un texte en français, ce qui nécessite d'installer le modèle de langue français de spaCy (fr\_core\_news\_md). On utilise ensuite l'outil de visualisation de spaCy « displacy » pour visualiser les résultats (figure 2). On constate que les résultats contiennent beaucoup de faux positifs pour les PII tels que les noms et les lieux détectés avec le Recognizer de type NER. Certaines PII comme le numéro de téléphone ne sont pas reconnus car, par défaut, la librairie propose des expressions régulières compatible avec le format américain. Néanmoins, il est possible en quelques lignes de code d'ajouter un Recognizer propre.



OFFICINATIONALDESECURITESOCIA PERSON LE  
 INSTITUTIONPUBLIQUE PERSON ESECURITESOCIA LE LOCATION  
 Place Victor Horta 11 LOCATION - 1060 BRUXELLES LOCATION - Tel. 02 509 59 59 - Fax PERSON 02 509 30 19 - Internet: www.crisis.fgov.be  
 DOMAIN\_NAME  
 IBAN LOCATION : BE63 LOCATION 0790 2618 1108 BIC: PCHOBE33 - Heures de visite: de 9 à 12 heures, ou sur rendez-vous tous les jours ouvrés  
 sauf le samedi  
 MON RESTAURANT SPIRL  
 AVENUE FONSNY 18  
 1080 BRUXELLES  
 DIRECTION DE LA PERCEPTION  
 Votre correspondant(e) :  
 M./Mme C. DUPONT PERSON  
 Tel: 02/509.23.45  
 Fax LOCATION : 02/509.23.55  
 E-mail: perception@crisis.fgov.be EMAIL\_ADDRESS  
 A rappeler dans votre réponse s.v.p. :  
 Numéro PERSON d'entreprise: 406.798.006  
 Notre référence : MA LOCATION -REF-000-1111

Fig. 2: Données personnelles identifiées par Presidio

Pour ce test, nous avons ajouté un nouveau Recognizer suivant un template prédéfini pour détecter le numéro KBO sur base d'une expression régulière: KboRecognizer. On y ajoute une liste de mots (bce, kbo, entreprise) de contexte qui renforce le score de détection s'ils sont présents dans l'entourage du numéro ainsi qu'une fonction de validation basée sur le calcul d'un checksum. Le numéro KBO est maintenant détecté (figure 3).

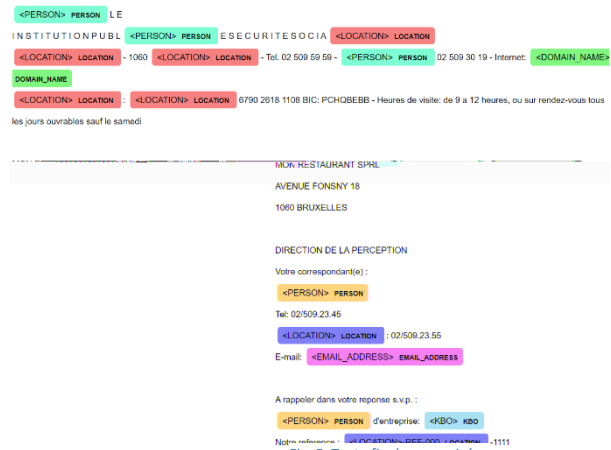
A rappeler dans votre réponse s.v.p. :  
 Numéro PERSON d'entreprise: 406.798.006 KBO  
 Notre référence : MA LOCATION -REF-000-1111

Fig. 3: KBO détecté en ajoutant un "Custom Recognizer"

```
{'recognizer': 'KboRecognizer', 'pattern_name': 'KBO/BCE number', 'pattern': '(?!\\d)[0]{0,1}\\d{3}[\\-\\.\\s]\\d{3}[\\-\\.\\s]\\d{3}(?!\\d)', 'original_score': 0.7, 'score': 1.0, 'textual_explanation': None, 'score_context_improvement': 0.30000000000000004, 'supportive_context_word': 'entreprise', 'validation_result': True}
```

Fig. 4: Explication produite par Presidio Analyzer pour la détection du numéro du KBO

Presidio Analyzer permet de générer une explication pour chaque PII détectée dont le numéro KBO comme illustré en figure 4.



<PERSON> PERSON LE  
 INSTITUTIONPUBLI <PERSON> PERSON ESECURITESOCIA <LOCATION> LOCATION  
 <LOCATION> LOCATION - 1060 <LOCATION> LOCATION - Tel. 02 509 59 59 - <PERSON> PERSON 02 509 30 19 - Internet: <DOMAIN\_NAME>  
 DOMAIN\_NAME  
 <LOCATION> LOCATION : <LOCATION> LOCATION 0790 2618 1108 BIC: PCHOBE33 - Heures de visite: de 9 à 12 heures, ou sur rendez-vous tous  
 les jours ouvrables sauf le samedi  
 MON RESTAURANT SPIRL  
 AVENUE FONSNY 18  
 1080 BRUXELLES  
 DIRECTION DE LA PERCEPTION  
 Votre correspondant(e) :  
 <PERSON> PERSON  
 Tel: 02/509.23.45  
 <LOCATION> LOCATION : 02/509.23.55  
 E-mail: <EMAIL\_ADDRESS> EMAIL\_ADDRESS  
 A rappeler dans votre réponse s.v.p. :  
 <PERSON> PERSON d'entreprise: <KBO> KBO  
 Notre référence : <PERSONAL\_IDENTIFICATION\_NUMBER> -1111

Fig. 5: Texte final anonymisé

Il est possible d'ajouter un Recognizer de type NER spécialement entraîné sur nos données mais cette fonctionnalité n'a pas été testée dans le cadre de cette quick review.

Une fois les données personnelles identifiées, celles-ci sont traitées par l'Anonymizer et remplacées par leur type (figure 5).

Presidio est open source et sous licence MIT, à utiliser dans un environnement Python (>=3.6). La suite Presidio s'installe comme un package Python (pip) ou via Docker.