Résumé du projet INTERREG DECRIPT en vue des entretiens qualitatifs dans le cadre de l'étude de marché

21 juillet 2020

Objectif général du projet

Traiter automatiquement les données textuelles pour leur gouvernance, leur sécurité et leur utilisation dans le domaine de la prospective, notamment en pouvant identifier et protéger automatiquement les données personnelles dans les textes en langage naturel (documents bureautiques, messages, etc.). La compétence spécifique développée est de comprendre la sémantique d'un texte grâce à l'intelligence artificielle linguistique. Le modèle sémantique qui sera proposé va ainsi plus loin qu'un moteur de règles.

Dans ce cadre, deux objectifs spécifiques sont fixés et trois applications pilotes seront développées.

Enjeu du projet : Problématique de la sécurité des données

- Assurer la sécurité des données est devenu incontournable pour leur collecte et leur exploitation, y compris pour les textes susceptibles de contenir des données personnelles.
- Une référence importante est la nouvelle règlementation RGPD (« *Règlement Général sur la Protection des Données* »), qui est entrée en vigueur en mai 2018. Le RGPD exige que toute entreprise (européenne et étrangère) doit être en mesure de prouver que les données à caractère personnel qu'elle détient sont protégées et surtout inexploitables en cas de vol. Exemples de données protégées : IBAN, numéros de téléphone, identifiants divers, données biométriques, enregistrements caméras, etc.
- D'autres législations en matière de protection des données concernent les entreprises et les organisations. La Suisse attend sa future loi révisée de protection des données. Aux Etats-Unis, le California Consumer Privacy Act (CCPA) a été repris dans d'autres Etats.
- La mise en oeuvre des réglementations en matière de protection des données crée de nouveaux besoins d'identification des données personnelles. Ainsi, un nouveau métier se crée : "Data Protection Officer" (*DPO, Délégué à la Protection des Données*), qui consiste notamment à supprimer ou à masquer/offusquer les données des personnes repérées dans les documents. Plus globalement, le DPO assure le respect de la conformité de son entreprise aux réglementations applicables.
- Il en découle un nouveau besoin d'outils automatiques permettant l'identification et le masquage/offuscation des données protégées, y compris dans les textes, afin de faciliter la mise en conformité des entreprises avec la législation. Un traçage devient obligatoire.
- La mise en conformité des acteurs économiques avec le RGPD, au-delà de respecter le règlement, peut être à l'origine d'un nouveau "contrat de confiance" avec les clients, partenaires, collaborateurs… satisfaits de la protection accordée à leurs données.

Spécificité du projet : Approche par modélisation linguistique

- Des travaux linguistiques permettront de déterminer comment les données propres aux personnes et ayant une valeur pour les utilisateurs sont exprimées dans les documents textuels.
- Un modèle linguistique -méta-modèle- de repérage des données sera élaboré pour servir à l'automatisation de l'identification de ces données.

Application pilote 1 : La gouvernance des données

- L'application de gouvernance des données doit permettre notamment de s'assurer de la conformité des procédures avec le RGPD, tout en identifiant les données de valeur. Les applications de gouvernance des données sont conçues afin de permettre à l'entreprise numérique de minimiser le risque d'exposition.
- Un premier but spécifique de l'application de gouvernance des données est d'étendre la gouvernance des données actuelles aux données textuelles, par un "mapping" entre la sémantique des données textuelles et la sémantique des données structurées de l'entreprise. Un deuxième but spécifique de l'application de gouvernance des données est de rendre possibles des audits sur l'usage des données personnelles et de garantir la protection des données de personnes.

Application pilote 2 : L'identification et le masquage de données personnelles

- L'objectif de l'application est de faciliter et automatiser l'anonymisation (ou offuscation) de tous types de données à caractère personnel au cœur de zones de texte (numériques), d'origines et de formes variées (avis client, document, courrier électronique, etc). Cette problématique concerne tous les organismes (privés comme publics) dans leur démarche de protection des données (notamment dans l'application du RGPD), que ce soit en interne, en vue d'externalisation d'un traitement, ou pour leurs flux sortants. Les problématiques principales sont que les données personnelles peuvent être de différentes natures et origines, et sont complexes à localiser et typer lorsqu'elle se trouvent au sein d'un texte. L'application doit donc assurer la détection, la localisation, le typage, puis le masquage des données personnelles au sein d'une zone de texte et selon diverses procédures sélectionnables (substitution, chiffrement ou encore offuscation). En résultat, l'application produit des textes anonymisés, en remplaçant les données personnelles par des valeurs qui ne portent plus aucune information, de plus, ces valeurs ne permettant pas de retrouver les informations originelles.
- Termes utilisés dans la pratique pour ce type d'opération : offuscation, anonymisation, dé-identification, dépersonnalisation, masquage ou maquillage de données, pseudonymisation, *data masking*, *anonymization*, *data cloaking*, *data masquerading*, *tokenisation*, *data obfuscation*, *data masking*, *data privacy*, *data sanitization*, *data scrambling*, *data deidentification*. L'opération inverse s'intitule ré-identification ou dé-anonymisation.

Application pilote 3 : La détection des entités non nommées

- L'objectif de l'application d'identification des entités non nommées est de détecter, dans un texte, une entité : un nom de personne (physique ou morale), d'organisme, de lieu ou de produit/service, et de le désambiguïser, même quand cette entité n'est pas nommée explicitement. Exemple : "Ville du bout du lac" doit pouvoir renvoyer à "Genève".
- Un but spécifique de cette application de détection des entités non nommées est d'intégrer l'identification d'entités non-nommées explicitement dans un processus d'analyse de l'information, ainsi que d'en évaluer la qualité et l'impact réel. A cette fin, il s'agit d'utiliser un modèle et une approche linguistique, qui permettront d'obtenir des meilleurs résultats en termes de précision. Au final, l'analyse de volumes importants d'informations sera accélérée.

Organisation et planification du projet

- Projet INTERREG franco-suisse financé par les collectivités publiques concernées (FEDER, Interreg fédéral suisse, Canton de Berne, Canton de Genève, Canton du Valais) et les partenaires du projet.
- Partenaires du projet :
  - Chef de file France : Université de Franche-Comté à Besançon.
  - Chef de file Suisse : Haute école de Gestion à Genève.
  - Autre partenaire France : ERDIL à Besançon.
  - Autres partenaires Suisse : Haute école de Gestion Arc à Neuchâtel et Global Data Excellence SA à Genève.
- Début du projet : 1er janvier 2020.
  Fin planifiée du projet : 31 décembre 2021.
- Une collaboration est assurée avec l'IBL (Institute for International Business Law de la Faculté de droit de l'Université de Fribourg), dont la recherche actuelle porte sur "Sharing Economy & Law and Law & Artificial Intelligence", avec des applications directes au RGPD.

Etude de marché

L'étude de marché vise à identifier les besoins et les contraintes en matière de détection, de gouvernance et d'offuscation des données, auprès des entreprises et des organisations privées, para-publiques et publiques utilisatrices, ainsi que des prestataires de service en matière d'analyse des données. Toutes les organisations manipulant des données au sens large sont concernées, ces données incluant non seulement des noms, mais également des adresses et des identifiants de toutes sortes.

Les besoins et les contraintes dont l'identification est visée par l'étude de marché concernent la reconnaissance et le traitement des données se rapportant aux personnes, que ces données soient explicitement nommées comme par exemple un numéro de téléphone ou qu'elles soient implicites, une donnée pouvant être représentée de diverses façons, par un synonyme, une description, une abréviation, etc.

L'étude de marché porte également sur les modèles d'affaires permettant d'exploiter au mieux les valeurs économiques qui sont créées par les données.