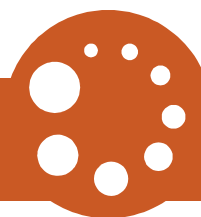




# NOGAuto: Automatisation de la codification des unités dans le Registre des entreprises et établissements

## Résumé

EXPERIMENTAL STATISTICS



Neuchâtel, 2019

**Éditeur:** Office fédéral de la statistique (OFS)  
**Rédaction:** Claude Macchi  
**Domaine:** 00 Bases statistiques des généralités  
**Langue du texte original:** Français  
**Traduction:** Services linguistiques de l'OFS

**Concept de mise en page:** Section DIAM  
**Téléchargement:** [www.statistique.ch](http://www.statistique.ch)  
**Copyright:** OFS, Neuchâtel 2019  
La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée



## Projet pilote NOGAuto : résumé

30 juin 2019

Les nomenclatures sont des éléments de base de la production statistique. La qualité du codage des unités observées a un impact direct sur l'ensemble du processus de production des données, sur la crédibilité et sur la qualité des résultats statistiques. C'est d'autant plus important dans le contexte des registres et des données administratives, qui sont le point de départ d'innombrables statistiques, la base des cadres de sondage et d'analyses des données.

Afin d'améliorer en permanence la qualité de la codification des unités dans le Registre des entreprises et établissements (SBER) et de réduire la charge pesant sur les entreprises dans leurs obligations de fournir des informations à l'office fédéral de la statistique (OFS), la section « Données des registres des entreprises » (URD) de l'OFS a lancé, dans le cadre du programme de la stratégie d'innovation sur les données de l'OFS, un projet visant à automatiser l'attribution du code des activités économiques aux entreprises.

Les procédures de codage sont actuellement assez standardisées. Les codeurs analysent et interprètent les informations sur les activités des entreprises, telles que les données fournies par les entreprises elles-mêmes, des informations en provenance d'enquêtes ainsi que les descriptions figurant dans les registres des sociétés et différentes données administratives. Sur cette base ils définissent des mots-clés qui sont comparés à une liste de mots-clés et de concepts liés à chacune des positions de la classification et à leurs notes explicatives et choisissent ainsi le code à attribuer à l'entreprise observée.

Grâce à des méthodes d'innovation, la section URD envisage de mettre au point un système d'apprentissage automatique pour automatiser la procédure de codage manuel. Ce système (« NOGAuto ») effectue les étapes de lecture et d'interprétation des descriptions des activités économiques (phase « préparation ») et associe automatiquement l'entreprise à un code de classification (phase « modélisation »). Il s'agit en fait d'apprendre à partir d'un ensemble de données existant qui a déjà été codé manuellement. Par la suite, l'horizon du système pourra être élargi en cherchant directement sur le Web d'autres sources d'information sur l'entreprise à coder.

Les résultats atteints dans cette première phase du projet sont très prometteurs. Bien qu'avec des ressources très limitées et une infrastructure encore à améliorer, un taux de prédiction de 89% a été atteint par le système dans le cadre des tests effectués sur une sélection de codes.

Il est de plus important de souligner que les efforts investis pour la réalisation de NOGAuto ne se limiteraient pas seulement à la codification des activités économiques des entreprises, mais qu'une légère adaptation du système permettrait de l'utiliser pour la codification des professions, des maladies et causes de mort, des produits etc.

### Méthodes d'analyse complémentaires utilisées:

- pour la phase « **préparation** » c'est-à-dire la phase qui comprend la lecture, la compréhension et la fouille de textes (« text mining ») on a implémenté des technologies de compréhension du langage naturel (« natural language processing », NLP) basées p.ex. sur des méthodes « naïve Bayes » pour la « compact language detection » ;
- pour la phase « **modélisation** », qui culmine avec l'association des entreprises à codifier avec un code de la nomenclature, le système utilise différentes architectures de réseaux de neurones artificiels (« artificial neural networks ») et des forêts aléatoires (« random forests »).