

# **Bonification de taxonomie par le traitement de la langue naturelle**

## **soumis par Radio-Canada**

Radio-Canada publie entre 450 et 600 nouveaux contenus chaque jour. Ces contenus sont créés et catégorisés par nos édimestres et journalistes aux quatre coins du pays. La taxonomie actuelle est utilisée afin de positionner (catégoriser) le contenu sur nos diverses propriétés numériques et non pas pour décrire la nature du contenu. Dans l'objectif de développer une meilleure compréhension de la composition de notre contenu ainsi que des intérêts de notre auditoire pour ce contenu (afin, par exemple, de mettre en place des algorithmes de recommandation de contenus et des outils de recherche avancés), nous désirons enrichir notre taxonomie tout en limitant la charge de travail de nos équipes.

Nous cherchons donc à bonifier notre taxonomie : (1) en utilisant des techniques de traitement de la langue naturelle (« Natural Language Processing », en anglais), afin d'extraire l'essence d'un contenu \*textuel\* et d'obtenir une représentation des entités dont il traite ; (2) en créant des regroupements logiques d'entités similaires.

Nous possédons actuellement un catalogue de plus de 17 000 contenus textuels catégorisés pouvant être utilisés dans le cadre de l'atelier. Nous avons également plusieurs artefacts (recherche, stratégie, modèles) qui peuvent être partagés afin d'appuyer les participants dans leurs démarches.