# Résumé de l'article "Constant-Delay Enumeration for Nondeterministic Document Spanners"

Cet article a été publié en 2021 dans le journal "ACM Transactions on Database Systems" (TODS), qui est une revue de référence pour la recherche en base de données, même si elle n'est malheureusement pas en libre accès. Il fait suite à un article que nous avons publié en 2019 dans les actes du congrès "International Conference on Database Theory" (ICDT), la conférence européenne de référence en théorie des bases de données : ses actes sont publiés en libre accès. Le travail ICDT'19 avait été primé par la récompense "SIGMOD Research Highlights" qui distingue les meilleurs travaux en recherche théorique et pratique en bases de données.

Nous nous intéressons dans ce travail aux "document spanners", qui sont un formalisme déclaratif pour spécifier des tâches d'extraction d'information à partir du texte. Plus exactement, un document spanner peut être défini par une expression régulière avec des captures indiquant quelles parties du texte il faut extraire, ou par un automate fini étendu avec des variables. Par exemple, un document spanner peut indiquer qu'il faut extraire toutes les paires d'une adresse email et d'un numéro de téléphone, définies par des contraintes régulières. Nous nous intéressons à la complexité de produire tous les résultats d'un document spanner sur un document textuel fourni en entrée. Nous étudions cela dans le modèle des algorithmes d'énumération : c'est-à-dire que nous distinguons d'une part le temps de prétraitement, nécessaire pour calculer une représentation concise des résultats à énumérer, et d'autre part le délai maximal qu'il faut potentiellement attendre ensuite entre deux résultats.

Nous démontrons dans cet article que cette tâche peut être résolue après un prétraitement linéaire en le document et avec un délai constant en le document, c'est-à-dire qui reste le même indépendamment de sa longueur. Ce résultat est déjà connu dans le contexte théorique de l'évaluation de requêtes de la logique monadique du second ordre sur des mots et des arbres, mais nous montrons ici son applicabilité à l'extraction d'information déclarative. La nouveauté technique de notre résultat par rapport à l'état de l'art est qu'il fonctionne pour des document spanners décrits par des expressions régulières ou automates non-déterministes, avec une complexité polynomiale en le document spanner, là où les travaux précédents nécessitaient une représentation déterministe ou l'application d'une procédure de déterminisation dont la complexité est exponentielle. Nous présentons également des résultats expérimentaux sur de la recherche de motifs sur des données génétiques, obtenus à l'aide d'un prototype.

Ce travail est le fruit d'une collaboration sur l'énumération démarrée deux ans auparavant avec Pierre Bourhis (Université de Lille) et Stefan Mengel (CNRS CRIL), auquel s'est ensuite joint Matthias Niewerth de l'université de Bayreuth. Nous avons obtenu les résultats ensemble lors de discussions avec Pierre puis avec Stefan, j'ai écrit la majeure partie de l'article sauf la partie expérimentale qui est le travail de Matthias.