L’expérimentation web sémantique du projet ISTEX

Le projet ISTEX est un investissement d'avenir soutenu par l’Agence Nationale de la Recherche visant à doter la France d'une bibliothèque numérique scientifique reposant sur deux axes complémentaires : d'une part, une acquisition massive de publications scientifiques (17 millions) couvrant l'ensemble des disciplines et d'autre part, la mise en place d'une plateforme unique d'hébergement, de gestion et d'accès à ces ressources. En janvier 2016, l’Inist-CNRS a lancé une expérimentation visant à publier, selon les normes du web sémantique, des données extraites du projet ISTEX. Cette expérience vise à développer une méthode pour mettre en ligne des jeux de données dans le respect des normes et standards du W3C. L’objectif est de répondre aux demandes des documentalistes et des chercheurs, en utilisant la structuration sémantique comme un moyen pour répondre à plusieurs besoins :

- proposer une documentation structurée et interopérable du fond ISTEX pour les utilisateurs de portail documentaire comme pour les chercheurs
- mettre à disposition des équipes de recherche des jeux de données très spécifiques permettant d’alimenter leurs travaux de recherche sur du machine learning ou du data alignement
- valoriser les jeux de données produits par des travaux de recherche
- rendre compatible le fond ISTEX avec des entrepôts de données présents dans le web sémantique
- faciliter d’avantage les travaux de recherche dédiés à la fouille de textes (bibliométrie, scientométrie, ...).

Les jeux de données sont là pour venir compléter, enrichir, consolider et lier toutes les informations présentes dans la plateforme. L’objectif est de proposer un graphe de jeux de données structurées reliés à des ressources extérieures ou à des référentiels d’autorité. In fine, ce lacis de données conduira toujours à un retour vers les documents plein texte présents dans ISTEX. C’est une autre façon pour diffuser et exploiter les ressources acquises.

La structuration sémantique proposée impose la modélisation des informations à publier au travers d’une ou plusieurs ontologies existantes. Ce postulat révèle des difficultés plus ou moins attendues : choix et appropriation des ontologies, rigueur d’une structuration sur des données hétérogènes. Elle a également permis de vérifier des attentes concrètes ouvrant la voie à un passage à l'échelle plus compatible avec le volume des données présent dans le fond ISTEX.

Nicolas Thouvenin

Je suis responsable du service “R&D et expérimentation” de l’Inist-CNRS. Je m'intéresse aux technologies et standards du web sémantique depuis 2011, lors de mes premiers travaux sur la publication de terminologies scientifiques au format SKOS. En 2013 et 2014, j’ai participé au groupe web sémantique du GFII. Actuellement, je pilote les travaux de plusieurs équipes autour de 2 thématiques : le Text Data Mining et l’interopérabilité des données.