Comment exploiter un corpus à l'aide des technologies du Web sémantique ? Le cas de la correspondance d'Henri Poincaré.

Nicolas Lasolle*[1,2] and Pierre Willaime*[1]

[1]Archives Henri-Poincaré - Philosophie et Recherches sur les Sciences et les Technologies – université de Strasbourg, Université de Lorraine, Centre National de la Recherche Scientifique : UMR7117 – France
[2]Laboratoire Lorrain de Recherche en Informatique et ses Applications – Institut National de Recherche en Informatique et en Automatique, Université de Lorraine, Centre National de la Recherche Scientifique : UMR7503 – France

Résumé

Le Web sémantique peut être vu comme particulièrement adapté à la description d'un corpus dans le cadre d'un projet d'humanités numériques. En effet, cette extension du Web permet de modéliser les
entités, relations, concepts, ou autres métadonnées pouvant décrire les documents d'archives, les acteurs historiques en question et leur contexte. Cette structuration s'appuie sur des technologies
standardisées (W3C) constituant le Web sémantique (Berners-Lee et al, 2001) telles que RDF (Resource Description Framework), RDFS (RDF Schema) et OWL (Web Ontology Language). La " sémantisation " d'un corpus ouvre la voie à son traitement automatique par des méthodes computationnelles

Dans cette communication, nous nous proposons de décrire notre usage des technologies du Web sémantique dans le cadre du projet d'édition de la correspondance d'Henri Poincaré (1854-1912). Ce corpus est composé d'environ 2100 lettres pour lesquelles une numérisation du document original, une transcription, un apparat critique un ensemble de métadonnées descriptives sont disponibles. Ce projet est porté collectivement par les Archives Henri Poincaré (UMR7117 - CNRS / Université de Lorraine / Université de Strasbourg). Plus précisément, nous souhaitons développer les méthodes et réflexions entamées pour répondre à deux problématiques rencontrées lors de ces travaux. La première problématique est liée à la représentation des connaissances de ce corpus. La seconde concerne l'exploration des données du corpus avec des outils computationnels.

Le premier enjeu de représentation des connaissances permet de mesurer l'expressivité permise par les ontologies (ici à comprendre comme des schémas de métadonnées structurés sur plusieurs niveaux). Les métadonnées décrivant finement le corpus ont pu être ainsi piochées dans
différentes ontologies (FOAF, BIO, BIBO, Relationship, Dublin Core). Les manques d'expressivité sur des points précis, problématiques pour les enjeux scientifiques, ont été comblés par la création d'une nouvelle ontologie, nommée ahpo. Celle-ci décrit la structuration du modèle de

*Intervenant

connaissances liant les différents schémas de métadonnées préexistants et précise l'introduction de métadonnées propres. L'alignement partiel permet de conserver une bonne interopérabilité. Nous nous intéresserons tout particulièrement à deux difficultés : la prise en compte d'une validité temporelle dans le renseignement des métadonnées (Gutierrez et al, 2005) et la gestion de l'incertitude sur leur contenu (Stoilos et al, 2006). La question de la temporalité peut être exemplifiée par la ville de résidence d'une personne. Il est souhaitable de pouvoir conditionner la validité des contenus de cette métadonnée à des contraintes temporelles. L'incertitude quant à elle peut concerner la datation des lettres. Il peut être possible de spécifier une date butoir d'envoi, Poincaré faisant référence à la lettre en question dans une autre, sans pour autant être en mesure de préciser. Ces deux questions, la représentation temporalisée des connaissances et la gestion de l'incertitude et de l'approximation, sont tout particulièrement intéressantes pour l'exploitation scientifique du corpus.

Ce second enjeu d'exploitation se propose d'utiliser, et d'étendre, les capacités d'interrogation permises par le Web sémantique. Dans le domaine, le langage SPARQL est le standard pour interroger des graphes de données. C'est un langage expressif qui permet de formuler des requêtes complexes pour exploiter les liens entre les éléments d'un graphe de données. Par exemple, un historien pourrait formuler une requête afin de récupérer les lettres échangées entre David Hilbert et Henri Poincaré en 1890 et qui traitent de géométrie non-euclidienne. Cependant, les résultats de l'exécution de requêtes ne sont parfois pas suffisants pour apporter des réponses à une problématique de recherche et cela oblige ainsi les chercheurs à devoir formuler de nouvelles requêtes. L'idée du mécanisme de recherche approchée est de créer des règles de transformation de requêtes afin d'assouplir ou modifier certaines contraintes de façon automatique. Par exemple, dans le cas de
la requête présentée ci-dessus, il est possible de relâcher des contraintes en étendant les bornes temporelles liées à l'expédition de la lettre, en recherchant les lettres avec des thèmes scientifiques
proches, ou en s'intéressant aux échanges avec les correspondants de Poincaré qui ont collaboré avec David Hilbert. Une autre idée pour exploiter les connaissances du corpus est de mettre en place des règles d'inférences de manière à compléter automatiquement des bases de
connaissances. L'objectif est de compléter l'édition des données du corpus et de dégager de nouvelles connaissances pour s'assurer que les recherches sur le corpus retournent l'ensemble des résultats attendus.

Ces travaux de recherche et méthodes ne sont pas spécifiques au corpus d'Henri Poincaré et ont pour objectif d'être réutilisés pour exploiter d'autres corpus en SHS et de proposer des extensions aux outils traditionnels du Web sémantique (Bruneau et al, 2021).

Références

Bruneau, Olivier, Nicolas Lasolle, Jean Lieber, Emmanuel Nauer, Siyana Pavlova, et Laurent Rollet 2021. " Applying and Developing Semantic web Technologies for Exploiting a Corpus in History of Science: The Case Study of the Henri Poincaré Correspondence ". Semantic Web, 12(2), IOS Press

Berners-Lee, Tim, James Hendler, et Ora Lassila. 2001. " The Semantic Web ". Scientific American, 284(5), 34-43.

Ghorbel, Fatma, Faycal Hamdi, Elisabeth Métais, Nebrasse Ellouze, et Faiez Gargouri. 2018 " A Fuzzy-Based Approach for Representing and Reasoning on Imprecise Time Intervals in Fuzzy-OWL 2 Ontology ". International Conference on Applications of Natural Language to Information Systems. 167-178.

Gutierrez, Claudio, Carlos Hurtado, et Alejandro Vaisman. 2005. " Temporal RDF ". European Semantic Web Conference. 93-107.

Stoilos, Giorgos, Nikos Simou, Giorgos Stamou, et Stefanos Kollias. 2006. " Uncertainty and the semantic Web ". IEEE Intelligent Systems, 21(5), 84-87.