Indexing and Exploring a Digital Humanities Corpus thanks to Elastic Representations

PhD subject proposed by Olivier Bruneau (AHP-PReST) and Jean Lieber (LORIA)

Keywords: digital humanities, semantic web, history of sciences, elastic queries, scientific texts, approximate and explained search, XAI (explainable artificial intelligence)

Context. Henri Poincaré's Archives (AHP-PReST - UMR7117) work for more than twenty years on Poincaré's letters and documents. Recently, researchers and engineers included tools from semantic web and digital humanities to improve the search capabilities and to link historical documents amongst themselves [1]. These new tools are reasons for joint works of AHP-PReST and LORIA. Several interlinked corpora about Henri Poincaré are maintained: his publications, his correspondence, and some secondary sources. In particular, his correspondence contains in a digitalized form the known letters he sent or received [2]. The main information about each letter — sender, recipient, date, subject, etc. — has been indexed in the semantic web standard RDFS and can be interrogated.[1] Two issues related to the current querying system have to be addressed in the framework of this PhD, issues whose impact goes beyond this particular corpus.

First issue: Content indexing. In order to interrogate the corpus about topics of the letters, these letters have to be indexed (or annotated) with concepts of available ontologies (domain ontologies, Dublin Core, etc.). For example, let us consider the term *fonction fuchsienne* that occurs in some letters of the H. Poincaré correspondence. This term has disappeared from the Mathematics literature since, where it has been replaced with *forme automorphe*. Both terms can be indexed by the same concept (e.g., `AutomorphicForm`). Furthermore, some correspondents to Henri Poincaré use the term *fonction kleinienne* instead of *fonction fuchsienne*. Therefore, the indexing of historical corpora with concepts of an ontology has to take into account the variation over time and space of the terminologies during the text processing: managing synonyms have to take into account these dimensions.

Second issue: Approximate and explainable search. A limit of the system is that it is restricted to exact queries. For instance, if the letters of the end of the XIX$^{\text{th}}$ century are requested, this can be modeled by a SPARQL[2] query on a chosen interval of time, say $[1890, 1900]$, though a letter of 1882 or even 1902 would be acceptable. Thus, to retrieve such letters, an approximate search has to be implemented. So, an approach based on fuzzy representations [3] seems to be appropriate: it would return the letters ordered by decreasing membership degree to a fuzzy interval representing the end of the XIX$^{\text{th}}$ century. Now, consider the query for letters sent at the end of the XIX$^{\text{th}}$ century to David Hilbert *or* his close colleagues. The execution of the fuzzy query would return results ranked by their membership degrees, that could put at the same level a letter to D. Hilbert of 1870 and a letter to a rather close colleague of D. Hilbert of 1897. Some extra explanations are requested there, explaining in what respect these results are approximate matchings. For this purpose, the notion of *elastic query*, that extends the notion of fuzzy query, is introduced: the execution of an elastic query gives triples (r, d, e) where r is the result, d is the degree of matching of r to the query, and e gives explanations on the mismatch between the result and the query. The idea is then to use a tool for managing SPARQL query transformation rules. A first version of this tool has already been implemented [4], that covers a variety of useful transformations. Another tool, designed for a different purpose by another team, has been developed that is also a candidate for managing these transformations [5]. The PhD student will have to collect queries that users want to ask to the system and represent them as elastic queries, which should raise limitations of the current tool, that will have to be overcome. An example of current limitation is the choice of appropriate widening of the time interval $[1890, 1900]$ taking into account *relevant* milestones, such as 1870 (year of publication of Maxwell equations in Electromagnetism).

Integration to OLKi IMPACT project. The methods and tools that are beeing developed in the project to which this PhD thesis participates can be applied to other historical corpora, provided that these corpora are indexed using RDFS format and are open (reading access through the Web).

[1] `http://e-hp.ahp-numerique.fr/s/hppapers/page/sparql`

[2] SPARQL is a query language for RDFS and other semantic web languages.

Furthermore, these tools enable an interrogation that is at the same time accurate (for the answers that are not a perfect match, the mismatch is explained) and flexible, thus providing an intermediate way for getting information between exact querying (that can be inappropriate to corpora of digital humanities) and information retrieval (with sets of keywords, giving results that are ranked but not explained).

This PhD will enable concept indexing (first issue) and approximate search (second issue) for corpora. These two features are especially designed to allow citizens to have access to powerful search capabilities for public corpora's analysis. The interrogation tool must be usable via a user-friendly interface, with views suited to different kinds of citizens. For a historian of science, this view has to be relevant for his/her research and has to include a precise terminology as well as suited transformation rules. For a non specialist citizen, this view is typically narrower. Moreover, these features will be explained and the general audience will be able to see what rules were applied. The PhD contributes to OLKi project by allowing a deeper and more relevant search and by explaining it.

This PhD will also contribute to OLKi project by using and enhancing the OLKi decentralized platform. Henri Poincaré's corpus could be integrated to this platform and the two features developed during this PhD will be published online and available for everyone. The OLKi platform will also enable concerned citizens to give feedbacks and to contribute in this manner to history of science. It is also planned to broadcast on a regular basis on this platform some short highlights about the corpora, e.g., anniversaries of some important events related to them or news about their enrichment.

Supervision.

Olivier Bruneau	**Jean Lieber**

Associate Professor
(Maître de conférence)
History of Science
`olivier.bruneau@univ-lorraine.fr`
`http://bruneauolive.free.fr`
	Associate Professor with habilitation
	(Maître de conférence HDR)
	Computer Science
	`jean.lieber@loria.fr`
	`https://members.loria.fr/JLieber`

O. Bruneau and J. Lieber previously worked together on subjects related to the application and extension of semantic web technologies for digital humanities.	

Candidates to this PhD. The ideal candidate is either a historian of sciences having some taste and practice for computer tools or a computer scientist open to humanities in general and history of sciences, in particular. She/He will have his office in AHP-PReST where one of her/his co-advisors and domain experts on Henri Poincaré letters work but will also spend some time in LORIA, working place of her/his other co-advisor.

References

[1] A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. van Harmelen. Semantic technologies for historical research: A survey. *Semantic Web Journal*, pages 1–27, 2015.

[2] L. Rollet, editor. *La correspondance de jeunesse d'Henri Poincaré: Les années de formation. De l'École polytechnique à l'École des Mines (1873-1878)*. Publications of the Henri Poincaré Archives. Springer International Publishing, Basel, 2017.

[3] D. Dubois and H. Prade. *Fundamentals of fuzzy sets*, volume 7. Springer Science & Business Media, 2012.

[4] O. Bruneau, E. Gaillard, N. Lasolle, J. Lieber, E. Nauer, and J. Reynaud. A SPARQL Query Transformation Rule Language - Application to Retrieval and Adaptation in Case-Based Reasoning. In *ICCBR 2017 - Case-Based Reasoning Research and Development, 25th International Conference on Case-Based Reasoning*, Trondheim, Norway, June 2017.

[5] O. Corby and C. Faron Zucker. A Transformation Language for RDF based on SPARQL. In W. van der Aalst, J. Mylopoulos, M. Rosemann, M.J. Shaw, and C. Szyperski, editors, *Web Information Systems and Technologies*, Lecture Notes in Business Information Processing. Springer, 2015.