A Navigation Tool for Exploring Semantic Web Corpora

Nicolas Lasolle[1,2][0000−0002−1253−649X]

[1] Université de Lorraine, CNRS, Université de Strasbourg, AHP-PReST, F-54000 Nancy, France
[2] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
nicolas.lasolle@univ-lorraine.fr

Abstract. Semantic Web technologies provide a way to represent and to exploit data and can be particularly suitable for historical corpora. The Henri Poincaré correspondence corpus is composed of more than 2000 letters which constitute scientific, administrative and private exchanges. Several technologies have been used for this corpus: the RDF model, the RDFS knowledge representation language and the SPARQL query language. Recently, a navigation tool has been created to explore this correspondence corpus by exploiting similarities between resources. This tool can simplify corpus exploration and highlight unexpected relations between elements. It relies on the use of a flexible search mechanism based on the definition and the application of SPARQL query transformation rules. The system can be connected to any RDF database as long as underlying data is exposed through a public SPARQL endpoint.[3]

Keywords: Digital humanities · Semantic Web · Historical corpora · SPARQL query transformation · Knowledge base exploration

1 Introduction

This navigation system will be presented through a live or recorded demonstration which will introduce several use cases and present its functionalities. This tool has been imagined in the context of the exploitation of the Henri Poincaré correspondence corpus. Henri Poincaré (1854-1912) is a famous French scientist who made several significant contributions in multiple areas of mathematics, physics and philosophy. Numerous research works are dedicated to the life and career of this man of science. The study of his correspondence is a long-term project which has led to the publication of several thematic volumes. An important aspect of this project is related to the online publishing and exploitation of this corpus [2]. The letters of the corpus are available on a website and come with a scan of the original document,[4] a transcription, a critical apparatus and a set of meta-data. Semantic Web technologies have been used to offer advanced

[4] Some are unavailable due to copyright rules.

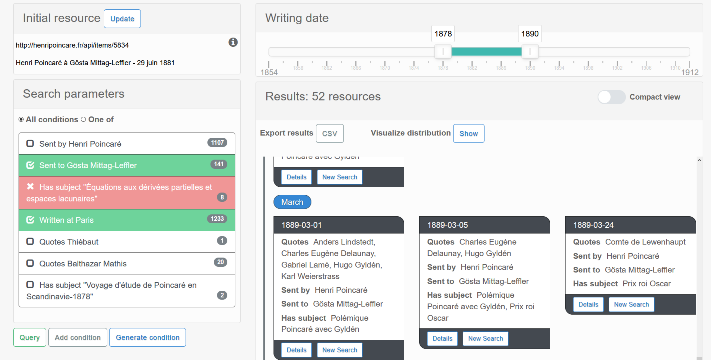
tools for the exploitation of this corpus. For this purpose, the RDF model is used to describe facts, the RDFS language is used to represent the domain knowledge and the SPARQL query language allows the interrogation of the corpus graph. When it comes to the exploration of an historical corpus, one of the main issues is to be able to put forward new and unexpected relations between individuals, institutions, scientific works, etc. A navigation tool has been developed for the correspondence corpus exploration by exploiting the similarities between documents. This system relies on the use of a SPARQL query transformation rule mechanism which can help to highlight unexpected relations between resources.

The remainder of this article is organized as follows. First, the navigation tool interface is presented (Section 2). Then, the need and the use of the flexible search mechanism for this navigation tool are presented (Section 3). A discussion related to this system architecture and reusability is provided (Section 4). Section 5 concludes and points out some future works.

2 A Navigation Tool for Exploring Corpora

This system can be used for the exploration of any Semantic Web graph. It is particularly relevant for the exploration of historical corpora because results are presented within a chronological-based interface. A demonstration video of this tool is available online. The tool is available as a Web interface which allows users to generate SPARQL queries and to visualize, to filter and to export results.

The top-left block gives some information about the initial resource related to the current search process. In our example, this resource corresponds to a letter sent by Henri Poincaré to Gösta Mittag-Leffler on June 29, 1881. The bottom-left block gathers a set of search conditions which can be used to create SPARQL queries and which are generated based on the initial resource. Next to each condition is given an integer which corresponds to the number of resources matching the given condition on the RDF graph. Users can select different conditions by clicking on them. Clicking on the "query" button updates the results presented on the bottom-right block of the interface within a chronological-based view. For each result, some information about the resource is given. For the Henri Poincaré correspondence corpus, each letter is described with its label, its sender, its recipient, the topics and the persons quoted. Above the result block, a date slider can be used to filter the set of presented results. The system proposes to export the results in a CSV file which embeds the IRI and some information about each resource. Another functionality concerns the presentation of a bar chart which expresses the distribution of the results in relation to the chosen date property. It is possible to start a new search process centered around one of the letters presented in the result block. The idea of the system is to start with an initial resource and to explore the corpus by navigating from a resource to another by taking different paths. This way of exploring the corpus could lead to the identification of unexpected relations between the elements of the corpus.



http://henripoincare.fr/api/items/5834

Henri Poincaré à Gösta Mittag-Leffler - 29 juin 1881

● All conditions ○ One of

- □ Sent by Henri Poincaré 1107
- ☑ Sent to Gösta Mittag-Leffler 141
- ☒ Has subject "Équations aux dérivées partielles et espaces lacunaires" 8
- ☑ Written at Paris 1233
- □ Quotes Thiébaut 1
- □ Quotes Balthazar Mathis 20
- □ Has subject "Voyage d'étude de Poincaré en Scandinavie-1878" 2

Results: 52 resources

Poincaré avec Gyldén

1889-03-01
Quotes Anders Lindstedt, Charles Eugène Delaunay, Gabriel Lamé, Hugo Gyldén, Karl Weierstrass
Sent by Henri Poincaré
Sent to Gösta Mittag-Leffler
Has subject Polémique Poincaré avec Gyldén

1889-03-05
Quotes Charles Eugène Delaunay, Hugo Gyldén
Sent by Henri Poincaré
Sent to Gösta Mittag-Leffler
Has subject Polémique Poincaré avec Gyldén, Prix roi Oscar

1889-03-24
Quotes Comte de Lewenhaupt
Sent by Henri Poincaré
Sent to Gösta Mittag-Leffler
Has subject Prix roi Oscar

Fig. 1. *The navigation tool interface in use for the exploration of the Henri Poincaré correspondence corpus.*

3 Going Further by Applying a SPARQL Query Transformation Mechanism

In some situations, the generated conditions may not be sufficient for providing interesting or surprising results. A first solution, which is included in the tool, is to let users manually add a new condition. Another idea to overcome this issue is to be able to generate new conditions that the initial resource would not necessary match but which are related to its characteristics. For this purpose, the system relies on the use of transformation rules that can be applied to provide new filtering conditions which are related to the conditions generated from the initial resource characteristics. These transformation rules are defined and applied using the SQTRL (SPARQL Query Transformation Rule Language) tool, which has been introduced to allow flexible querying with SPARQL [1].

On the user interface, the "More" button is used to generate new conditions in an iterative way. For the running example, the first click on the button generates two new conditions: {`sent to Henri Poincaré`} and {`sent by Gösta Mittag-Leffler`}. This is related to the application of a transformation rule which exchanges the sender and the recipient of a letter. Another action of the button adds the condition {`has topic Mathematics`}, based on the application of a topic generalization rule which replaces *Équations aux dérivées partielles et espaces lacunaires* by *Mathematics*. By applying the same transformation rule, the tool generates the condition {`has topic Travel`}. Two other rule applications generate conditions to search for letters sent of received by one of the persons quoted: {`has for correspondent Thiébaut`} and {`has for correspondent Balthazar Mathis`}.

4 System Architecture and Reusability

The system, its source code as well as a user and technical documentation are available online on a GitHub repository. When developing this tool, one of the main challenges was to ensure its reusability with other corpora. In this context, the backend application (developed with the Jena API [3]) comes with a configuration file to define: the URL of the SPARQL endpoint; the path to the transformation rule file; the list of properties to be used for condition generation; the list of properties to be displayed for each result; the property and language for labels; the date property and the temporal interval associated with results filtering. The documentation describes another use case which is related to the search of literary works by querying the DBpedia public SPARQL endpoint[5].

5 Conclusion

A navigation system, accessible through a Web user interface, has been proposed for exploring the Henri Poincaré correspondence corpus. This system benefits from the use of a flexible search mechanism for exploiting the relations between the elements of the corpus. It is a generic tool for Semantic Web graphs exploration and could thus be reused with other corpora. Several future works are considered for improving this system. Some improvements concern adding new features to the user interface such as adding new export formats or being able to remove a condition from the list. A major improvement would be to keep a trace of any action performed with the tool and thus being able to save the research process. In the context of a historical corpus, the research methodology is an important aspect of the work associated with results presentation. Another idea is to provide the user with an explanation of conditions generated by the use of the SPARQL query transformation mechanism. This could be the description of which transformation rule has been applied and for which resources.

Acknowledgement. This work is supported partly by the French PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

References

1. Bruneau, O., Gaillard, E., Lasolle, N., Lieber, J., Nauer, E., Reynaud, J.: A SPARQL Query Transformation Rule Language — Application to Retrieval and Adaptation in Case-Based Reasoning. In: Aha, D., Lieber, J. (eds.) Case-Based Reasoning Research and Development. ICCBR 2017. pp. 76–91. Lecture Notes in Computer Science, Springer, Cham (2017)
2. Bruneau, O., Lasolle, N., Lieber, J., Nauer, E., Pavlova, S., Rollet, L.: Applying and Developing Semantic Web Technologies for Exploiting a Corpus in History of Science: the Case Study of the Henri Poincaré Correspondence. Semantic Web – Interoperability, Usability, Applicability **12**(2), 359–378 (2021)
3. McBride, B.: Jena: a Semantic Web toolkit. IEEE Internet Computing **6**(6), 55–59 (2002). https://doi.org/10.1109/MIC.2002.1067737

[5] https://dbpedia.org/sparql/.