Building knowledge graphs to access and understand historical scientific manuscripts: A case study for manuscripts from Ferdinand de Saussure (1857 – 1913)

Case study presented by MIRALab, University of Geneva

Abstract. More than 40.000 handwritten texts from Ferdinand de Saussure were left unpublish, waiting for scientist to understand them. The chaotic aspect of these texts lead to a primary need for classification and dating issues. This case study presents a way to extract scientific content, historical and bibliographic context and terminology context from such manuscripts. A knowledge graph constituted by entity classes and relationships, has been used in order to propose an implementation on the Semantic Web.

Keywords. Digitization, Digital Humanities, Historical ontologies, Semantic Web Interface

Foreword

This presentation is based on documents [1, 2, 3, 4, 5] published by the Knowledge engineering research group KE@CUI as output of the SNF project *Knowledge engineering models and tools for the digital scholarly publishing of manuscript* (grant CR21I2_159747).

Introduction

The field of digital humanities is a collaborative and transdisciplinary area between digital technologies and the discipline of humanities. Ferdinand de Saussure was one of the fathers of modern linguistics (general linguistics, comparative grammar, social sciences). He has made, however, a few publications and he never published in General Linguistics, having only given a course between 1906 and 1911. The famous *Cours de Linguistique Générale* has been published posthumously (by Bally and Sechehay),, based on the notes taken by his students.. As a consequence, more than 40.000 pages of handwritten texts remain, for the most part, unpublished and unexploited.

So, which are the needs for taking advantage of these manuscripts? First of all, to retrieve and access the data. That is, to visualize the manuscripts and their transcription and to gain access through thematic classification plans, people, place, events and concept indexes. Secondly, the understanding process is very important, which means that the exact meaning of each term has to be determined. Moreover, the dating is essential so that the manuscripts can be placed on a chronological order. Lastly, the disclosure of the author's work, in different forms, is needed. However, these needs face some typical problems like difficulties in the thematic classification, as there are multiple themes in each manuscript, as well as dates and chronological issues. Another limitation is the text order problem and, therefore, there is a need to rebuild the intended text in order to be understandable.

Thus, in order to satisfy these needs and gain an holistic knowledge, a multidimensional approach needs to be constructed, including scientific, historical, bibliographical and terminological context. These evolving interconnected knowledge resources can constitute an advanced model to help humanists deal with the knowledge-intensive tasks they must perform when studying historical scientific manuscripts.

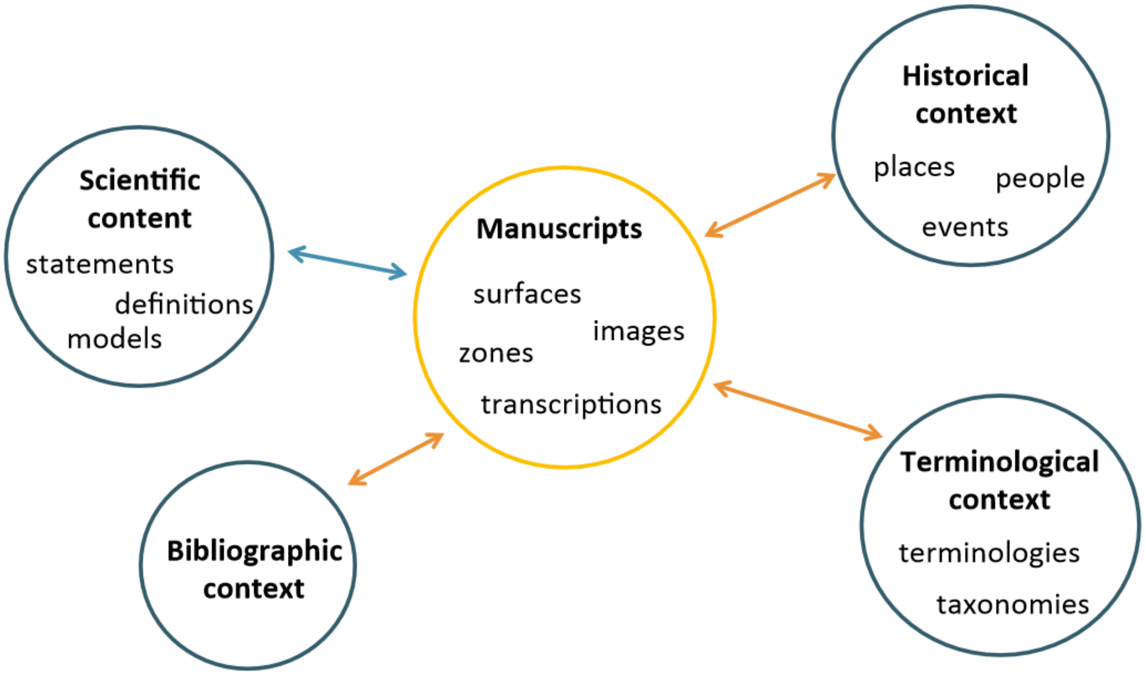


Figure 1: Knowledge representation needs

Requirements and Implementation

In order to take advantage of the state of the current knowledge about the manuscripts, the modeling of each context mentioned above is mandatory.

Concerning the historical context, there is a need to correlate the time-varying entities and relationships.

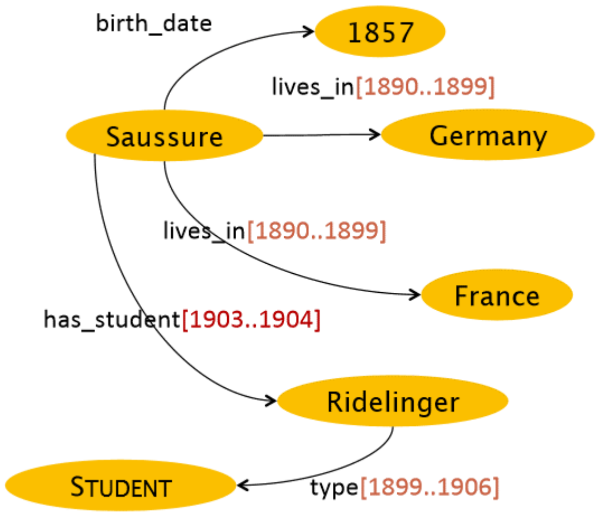


Figure 2: Time-varying entities and relationships

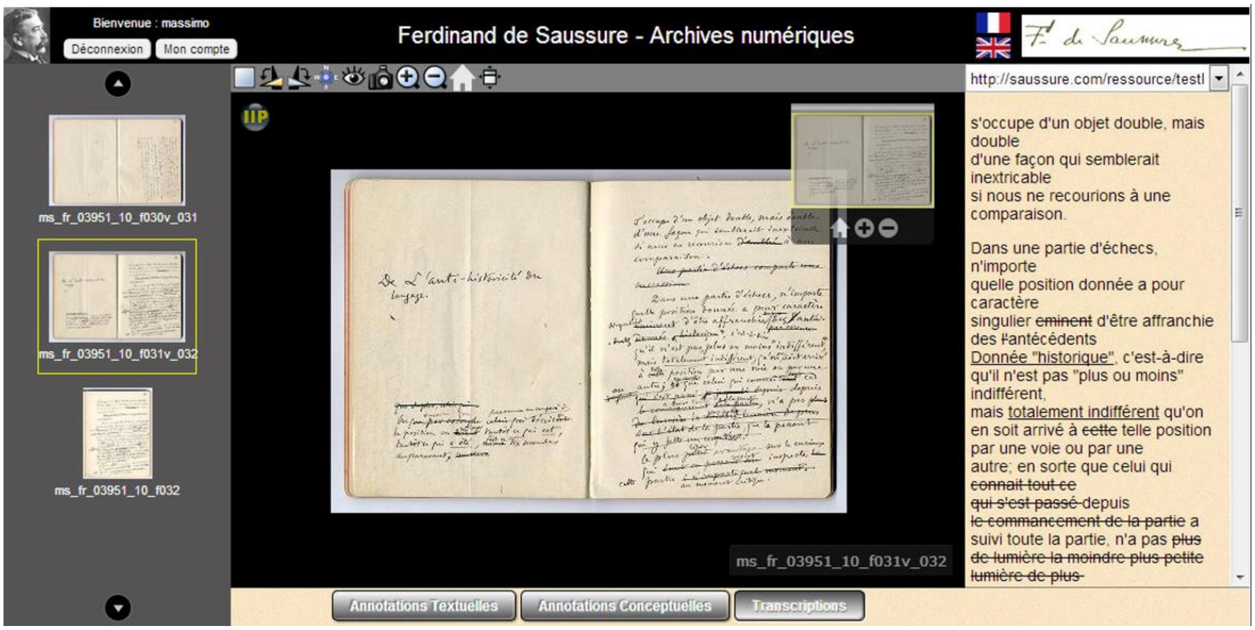
By entity classes, we refer to people, event, places etc. These classes can be divided into rigid, e.g people, and non-rigid, e.g student. On the other hand, there are the relationships, as shown in Figure 2, that can be separated between the fluent ones, like the variable lives_in and the non-fluent, like the variable place_of_birth. Based on these terms, historical reasoning rules are applied to infer temporal relations. As an example of the latter, we can present the following sentences:

- If text X refers to event E then writing-time(X) > time (E)
- If A sends a letter to B at time T then A knows B at time T (and thereafter)

Regarding the terminological context, a lot of work has already been done by scientists who have invented new concepts, redefined terms and worked on unstable terminologies, showing that terminology evolves over time. Cosenza et al, as an example, identifies 14 terminologies in Saussure's work. As a consequence, multiple terminologies have been incorporated into the same knowledge graph, created by different researchers and expressed in different formalisms, like TBX terminologies, SKOS schemes, OWL ontologies, tables, texts, etc. There is also a conceptual evolution over time, which means that different definitions can be used for the same term. This can evoke, in general, a global inconsistency.

Thus, the challenge for the terminological context modeling is to understand the text and then to date it. The understanding process includes the determination of the meaning of the terms in a text and then the determination of the terminology that has been used by the author. Afterward, in order to estimate the date, the determination of the terminology used in the text is needed, as well as the indication of a possible writing period.

Implementation of all the aforementioned has already been done with the help of the semantic web technologies. There have been techniques for temporal modeling and reasoning as well as techniques for dynamic terminology representation and processing. Moreover, manuscripts and knowledge graph have been finally represented and stored. In Fig. 3 there is an example of such a manuscript.



s'occupe d'un objet double, mais double
d'une façon qui semblerait inextricable
si nous ne recourions à une comparaison.

Dans une partie d'échecs, n'importe
quelle position donnée a pour caractère
singulier ~~eminent~~ d'être affranchie des ~~l~~'antécédents
Donnée "historique", c'est-à-dire qu'il n'est pas "plus ou moins" indifférent,
mais totalement indifférent qu'on en soit arrivé à ~~cette~~ telle position
par une voie ou par une autre; en sorte que celui qui ~~connait tout ce~~
~~qui s'est passé~~ depuis
~~le commancement de la partie~~ a suivi toute la partie, n'a pas ~~plus~~
~~de lumière la moindre plus petite lumière de plus~~

Figure 3. Example of a manuscript with the Semantic Web techniques

In Fig.4 the architecture of the described system is presented. In the second row of this cyclical procedure, we meet the "client" at the first box, the Web Server/Front-end at the second and the Back-end as well as the storage control at the third one. The storage control includes the updates and the authentication.

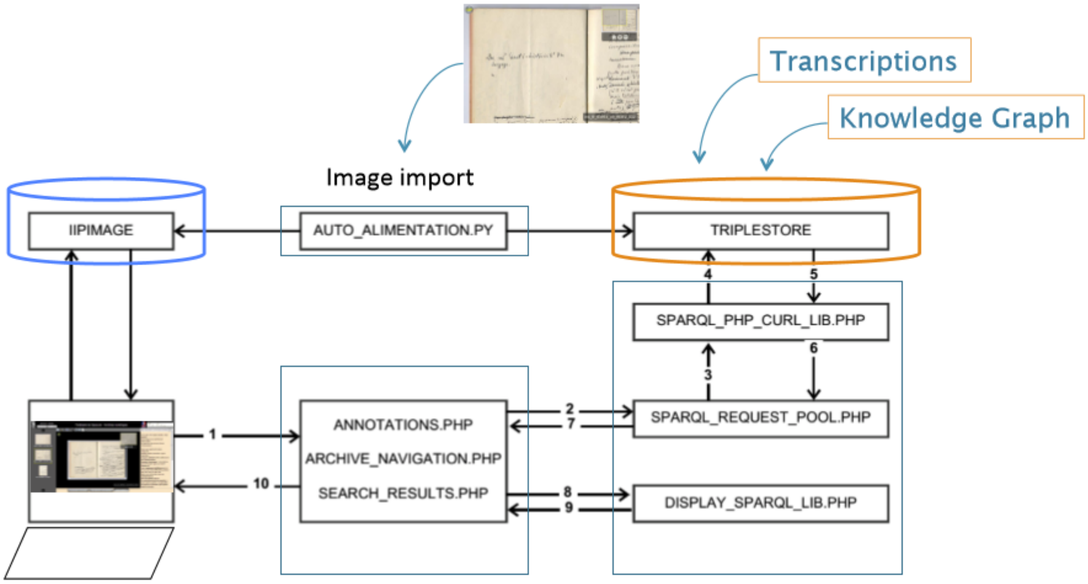


Figure 4. System architecture.

To sum up this procedure, the steps of the knowledge graph services are the following:

1. Adding manuscripts and transcriptions
2. Importing knowledge resources (historical context and terminologies)

3. Temporal reasoning
4. Semantic indexing of texts with multiple ontologies/terminologies
    - Terminology finding
    - Terminology generation (by correlation detection)
    - Finding relationships (similarity, (dis)agreement)

Terminologies and Semantic Indexing

The aspect of the terminology is maybe the most important in this attempt of deciphering old manuscripts. The general purpose is to map all the terminological models to a common generic model In order to avoid any logical inconsistencies, we have first of all to consider each terminology in isolation. Moreover, we need to connect the equivalent concepts (semantic alignment).

Before concluding to the terminology identification, there is a need of semantic indexing. This, first of all, means that there will be an association between the text elements (multi-) words and the concepts in the terminologies. The most common problem though is the one of polysemy, where one word can correspond to several possible meanings. In order to solve this kind of problems, the current approach is based on a similarity score, depending on the term's context in the transcription and the term references in the terminology. This constitutes a distributional semantics approach and can lead to the terminology identification by computing similarity scores for all the terminologies. Precisely, score of a terminology = f(score of each term).

Conclusion and Future work

To conclude, this work has presented a model illustrating how semantics technologies can be applied to historical datasets. It has built the infrastructure for the storage, the semantic enrichment, the visualization and the publication of a corpus of scientific manuscripts. The implementation of these steps is achieved through the Semantic Web techniques. Although this infrastructure has been applied to Ferdinand de Saussure's work, it is ready to be used on any other corpus of manuscripts as it proposes a multi-knowledge resource structure to represent the evolving nature of an author's terminology.

As an evolution of this work, new steps are developed. Firstly, crowdsourcing for the transcription of the manuscripts has already been taken into consideration. Moreover, new ways to test the temporal (contextual) inferences and new tools for the extraction of scientific contents are under investigation. The ultimate purpose, however, is to finally create a unified interface for the digital humanists.

References

1. Aljalbout, S., Cosenza, G., Falquet, G., Nerima, L. A (2016) Semantic Infrastructure for Scientific Manuscripts. In Federico Boschetti (Ed.) proc. International conference 2016: "Digital Edition: Representation, interoperability, text analysis, Venice.

2. Aljalbout, S., Falquet, G. (2017) A Semantic Model for Historical Manuscripts. In proc. Third International Workshop on Semantic Web for Scientific Heritage at ESWC'17. Portorož, Slovenia, May 2017.

3. Aljalbout, S., Falquet, G. (2017) Un modèle pour la représentation des connaissances temporelles dans les documents historiques : Applications sur les manuscrits de F.Saussure. In Proc. 28es Journées francophones d'Ingénierie des Connaissances (IC 2017), Caen, July 2017.
4. Cosenza, G. (2017) Les projets de Digital Humanities relatifs à l'œuvre de Ferdinand de Saussure. Cahiers Ferdinand de Saussure, no. 70. Droz, Genève.
5. Cosenza, G. (2016) Entre terminologie et lexique : les chemins de la pensée de F. de Saussure. Cahiers Ferdinand de Saussure, no. 69. Droz, Genève.
6. Reichling A. (1949), What is general linguistics?, 1:8-24, Lingua, Elsevier
7. Cosenza G. Dalle parole ai termini: I percosi di pensiero fi F. de Saussure. Edizioni dell'Orso (2016)
8. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F. (2015), Semantic Technologies for Historical Research: A Survey. Semantic Web Journal, 6(6): 539-564