

Ontologies of Ms. Fr. 640

Celine Camps

Project aim

This project aims to develop a French and English thesaurus of Ms. Fr. 640 that enables users of the Digital Critical Edition access to terminologies on different levels. The current features of the DCE allow users to either search for specific terms they are interested in or to rely on a pre-developed list of sixteen mark-up categories, such as ‘animals’, ‘body parts’, ‘professions’, ‘measurements’, and so on. These mark-up categories provide an overview of all the instantiations of these labels as well as their frequency. The goal of my project is first to standardize or normalize these various instantiations. For example, ‘large lizards’, ‘female lizards’ and ‘Lizards’ can all be nested under the normalized label ‘lizards’ (lower cased, plural). These normalized terms will be modern (as opposed to early-modern) concepts and will be called ‘preferred label’ (or ‘prefLabel’ in my working sheet). Second, if there is sufficient time, I will develop these prefLabels further into even broader categories—in the case of lizards this would be ‘reptiles’. My focus, however, will be on the prefLabels.

Deliverables, Workflow and deadlines

I will use the generated ‘terms of usage’ lists in Github, export these into an excel or Google sheet which consists of the following columns: ‘frequency’ (how many times a specific term is used); ‘verbatimTerm’ (the term as it occurs in the manuscript – either in English or French, depending on which version I will be looking at); prefLabel-fr (As was explained to me, this is the translation of the prefLabel-en, although I personally find this problematic. I personally think it best for prefLabels-fr to be a *direct*—modern—translation of the tc); ‘prefLabel-en’ (normalized preferred English term); ‘broaderTerm-eng’ and ‘broaderTerm-fr’ (larger, umbrella categories, such as ‘reptiles’, ‘insects’, ‘birds, and their relevant translations); ‘Notes’ (remarks, problems, issues that need to be flagged). The screenshot below is an example of what such a spreadsheet will look like—this spreadsheet relates to the animal terms used in the translation. I will create a spreadsheet for each mark-up category and version (tl, tc, tcn) (the question is, do we need a spread sheet for both the tc and the tcn and which of these two version is best to work with?). I will save these spreadsheets in a folder on the Google Drive and keep a protocol document that records decisions I have made, problems that I have encountered and require fixing, etc.

Ideally I would like to finish the English and French prefLabels for all mark-up categories of the manuscript by the end of the semester (probably excluding the materials, which I anticipate as a highly problematic category), although it is difficult at this stage to say to what extent this is feasible. I will therefore start with a relatively small and (seemingly) unproblematic category like ‘body parts’, to get an indication of how long the process of normalization takes, which problems arise and how to move forward. This will also allow me to decide if it is best to start with the translation, work on both the tl and the tc, or focus on the tc and/or tcn instead. I would very much like to discuss this further with the M&K team, to see what is possible in the time that is left and what has priority (if anything).

A1 frequency							
	A	B	C	D	E	F	G
1	frequency	verbatimTerm	prefLabel-fr	prefLabel-en	broaderTerm-fr	broaderTerm-en	notes
2	1	animal		animals			
3	2	ants		ants		insects	insects - category too broad?
4	1	Aquatiliium animalium					
5	1	Barbels		barbels		fish	
6	1	Bat		bats			mammals?
7	1	Bats		bats			
8	1	beef		beef		cows	cows - category too narrow?
9	2	Bénarris					
10	1	benarris					
11	1	Benarris					
12	1	big rat		rats			
13	5	bird		birds		birds	
14	8	birds		birds		birds	
15	4	Birds		birds		birds	

Figure 1: Example of an ontology spreadsheet.

Skills

My knowledge of and familiarity with the mark-up (and its protocols) of the manuscript will hopefully aid me in developing this project. Although I will predominantly work with excel sheets (which I have experience with), I might occasionally resort to Oxygen—in case I need to search for terms (or phrases) in Ms. Fr. 640. While Tianna showed me the ropes (i.e. the basics), I might need occasional assistance and advice on how to use this tool efficiently.

2. a description of the use cases that your project addresses

The importance and usefulness of my project can be best explained by means of an example.

‘Founablesders’ (meaning founders) is one of the terms that was marked up as a profession in the English translation. The ‘ables’ is in a fact a deletion in the original page of the manuscript which has been carried over to the tc and tl. While ‘able’ was marked up as a deletion in the transcription and translation, the tool that generated the list of profession terms based on our mark up did not recognize the ‘ables’ as a deletion and read ‘founablesders’ as one word instead:

`<pro>foun<fr>ables</fr><add>ders</add></pro>`



`<pro>founablesders</pro>`

The problem with this is that if a user were interested in finding all folio numbers on which the author-practitioner mentions ‘founders’ and were to insert that word into the search bar of the DCE, fol. 36 v, on which ‘founablesders’ occurs, does not come up in the list of hits.

If, however, a user were to use the exact search term ‘founablesders’ then folio 36v does occur. Yet I think it is safe to say that not even the most creative user of the DCE would ever think of ‘founablesders’ as a different way of spelling founders.

So by normalizing ‘founablesders’ to ‘founders’ we can bridge that problematic gap between various and literal spelling variants of the same word and usability.

3. any accompanying diagrams, sketches, mock-ups, prototypes, or texts relevant to your project (i.e., your deliverables)

See attachments.

4. 300-word description of the insights you gained through your project work with regard to course content and with regard to Ms. Fr. 640 and its content, possible use, and characteristics.

I did not gain new insights of the manuscript, but my work has made me reconsider some of the markup decisions we have made in previous years. Why did we, for example, include qualifiers in the mark up, whereas we left them out in other cases.

The case of ‘founablesders’ was interesting for several reasons. In addition to the fact that it illustrates the problems with the way in which our tool reads our mark up, it also points to one of the problems of transforming text from one format (handwritten word) to another (transcribed, translated and marked up term).

5. How did your project change from proposal to final execution? If you had unlimited time and funds, how would you build on your project?

In my original project proposal I wrote:

The goal of my project is first to standardize or normalize these various instantiations. For example, ‘large lizards’, ‘female lizards’ and ‘Lizards’ can all be nested under the normalized label ‘lizards’ (lower cased, plural).

This proved problematic. First of all because not all terms could be normalized to a plural or lower cased form. For instance: ‘the one who sings’, ‘underfoot’, or ‘commanders of Malta’. Following modern grammar and spelling rules, Malta, indicating a country, should stay capitalized. Additionally, it would be grammatically incorrect to normalize ‘underfoot’ to ‘underfeet’ and odd to standardize the ‘the one who sings’ to the plural ‘the ones who sing’. Moreover, how much would such a descriptive string of words help users of the DCE to find particular professions? While reducing strings of words to a one-word Preflabel could be helpful in theory, the question was, what to reduce it to? For example, can ‘the one who sings’ be equated with singers? Similarly, to which category do you reduce ‘white thick saliva that is found on the teeth’? Saliva? Teeth? Both?

Second, it was often difficult to know if two seemingly similar Verbatimterms could be nested under the same PrefLabel. For example: ‘perfumers’ and ‘perfume makers’. Are these simply different spelling variants of the same profession? Or do they indicate different vocations? In order to answer this question (and others like it), I would have to do a tremendous amount of research—for which I simply did not have the time.

Third, the marked-up terms were already at various levels of terminology, even before I had to begun the process of normalization. For example, the more specific ‘master goldsmiths’ versus the general term ‘artisans’.

Fourth, while some terms only had to be normalized with respect to spelling—e.g. finger → fingers—standardizing other terms involved multiple steps: inventor or rustic figulines.

I therefore changed my initial protocol to normalized different terms to larger, more general PrefLabels and decided to focus on normalizing the spelling first.

I had also hoped to finish the English and French Pref Labels for all marked-up terms. Instead, I have only been able to go through the English PrefLabels pertaining to the mark-up categories 'musical instruments', 'body parts' and 'professions'. While the former is finished, the latter two still require editing. There are still problematic cases that need to be discussed with digital lead Terry Catapano and the Making and Knowing team

If I had unlimited funding and time, I would gather an entire team of researchers, specialized in various subjects, ranging from history of art, metalworking, alchemy, etc. to aid me in developing a modern and early modern thesaurus of BnF. Ms. Fr. 640.