Half-time for the ADOCHS project!

On 1 November 2016, the CegeSoma launched the ADOCHS project in collaboration with the Royal Library (*KBR*), and the Universities of Brussels (*Vrije Universiteit Brussel* (*VUB*), *Université Libre de Bruxelles* (*ULB*)). The project aims to improve quality control processes for digital cultural heritage collections. After having worked for two years at CegeSoma, Anne Chardonnens now prepares to join the *ULB*-team. We interviewed her to reflect on the results she has accomplished so far.

Can you introduce yourself to our readers?

I am a PhD student in Sciences et Technologies de l'Information et de la Communication at the *Université libre de Bruxelles*. Before being involved in the ADOCHS project, I first worked for nearly one year on the MADDLAIN project.
This project aimed to improve digital access to collections, by studying the behaviour and needs of the users of the State Archives, CegeSoma, and the Royal Library.

What does ADOCHS mean?

It is an acronym for Auditing Digitization Outputs in the Cultural Heritage Sector. The objective is to develop new methodologies to analyse and improve the quality of metadata and images in the context of digitization processes. I am in charge of the metadata part, while a PhD student in Digital Mathematics is responsible for the images part.

CegeSoma is one of the partners of the ADOCHS project; in which ways can the project be useful to the Centre?

The CegeSoma collections are documented by the help of descriptive metadata. It is this 'data about data' that then allows users to search Pallas - the digital catalog. Thus, metadata containing titles, legends or keywords will help users to find one of the 300.000 photographs held by the Centre.

However, it turns out that the quality of these metadata is very uneven. As explained one year ago, the keywords used to describe the collections present us with different types of problems. The ADOCHS project provides an opportunity to take the time to identify these problems and test new solutions, in order to create better access to the collections.

Does it mean that you spend your days rectifying misspelled keywords?

To manually correct metadata that is inaccurate, incomplete, or inconsistent is a very time-consuming task. It is a painstaking work and it would take several fulltime people dedicating themselves exclusively to this task to obtain significant results.
I prefer methods allowing me to automate the entire process, or part of it. For example, the free software OpenRefine makes is easy to find duplicates, empty fields, similar terms containing a slightly different spelling, and then to process a thousand anomalies with a single click.

You are about to join the Université libre de Bruxelles (another partner of the ADOCHS project) after having worked for two years at CegeSoma. What is your feedback halfway through?

The first months allowed me to better understand the institution, its mission, its history, its collections, and how it works. It was a particular context: the recent integration in the State Archives, the arrival of a new director, the departure of a long-time computer scientist (who knew all the ?secrets? of the main database), and finally, the prospect of data migration to the collection management system of the State Archives. In addition to this better understanding of the context in which the project takes place, the time spent inside the institution has allowed me to better understand the issues related to metadata quality, which also turns out to be critical for the success of other projects like EHRI or UGESCO.

After having delved into the state of the art, the needs of the institution and its users (see for example the needs of researchers), and carried out some preliminary analyses, I have decided to tighten the scope of my work around authority data for person entities. In the case of CegeSoma, a lot of names of personal names are contained in the keywords used to index doucments. My work has focused on exploring how the Linked

Open Data can be used to disambiguate the personal names, link them to other collections around the world, or enrich them using contextual elements from knowledge bases such as Wikidata. But it is not over, it is a work in progress ...

What's next?

As an EHRI fellow, I currently have the opportunity to stay at the *CDEC* (Fondazione *Centro Di Documentazione Ebraica Contemporanea*) in Milan. I am conducting tests to see to what extent parts of the CegeSoma collections could be linked to theirs, through the Person authority records. It is very exciting and lets me imagine new possibilities!

Of course, there is still much to be done and it would be vain to believe that everything can be ?cleaned up? (unfortunately, one cannot miraculously deduce the identity of a person whose only first name has been encoded, without date or any other additional information), but it is exciting to be able to analyse the possibilities and limits of new tools with the help of empirical data, knowing that this will be of direct benefit to staff and end-users.