Brief description of the project

Student: Erica Andreose

As a research fellow I am working on the process of preparing and uploading documents for the Digital Edition of Aldo Moro's Works.
The project was born in 2017 and involves several Italian universities along with other public and private entities. The main work of creating and populating the platform is managed by UNIBO. In 2021 there was the last upload of this first phase, which saw about half of the total documents (808 documents) present on the portal.
The project, due to the covid-19 pandemic and other logistical issues, took a break of about 2 years and was only resumed in late 2023 with the creation of two new 6-month research grants and the hiring of new technical staff.
Due to the long hiatus, fragmented documentation and the inconsistent nature of the project workflow, the resumption of work was difficult and time-consuming.
Bugs and issues related to the KwicKwocKwac source platform, through which researchers (domain experts) mark up documents and compile related metadata, were immediately noticed.
KwicKwocKwac plays a key role in the work process. It allows experts to tag searchable information within texts (people, places, organizations, mentions and citations) and to enter a long list of metadata related to each document worked on. This metadata will populate a database hosted on MongoDB that will then be used to generate the first part of the RDF. This first RDF file (data.ttl) is the key part for the search and indexing operation of the works on the final ENOAM platform. While as far as the tags inserted in the documents are concerned, KwicKwocKwac does not directly return a clean html file (or even an xml one) but it is necessary to download these files via a server (we use FileZilla to speed up this phase) and then process them with a long cleaning path via python scripts that rely on external libraries. Only at the end of this process, can the already existing RDF file (data.ttl) be aligned with a new complete RDF file that also contains all the information tagged in the text htmls. So during the alignment phase, mentions, citations, places, people, organizations, and keywords are entered into the RDF.
We had to study well how KwickKwocKwac works in order to understand how to avoid running into saving errors or word processing bugs.
As for the whole stage of processing documents downloaded from KwicKwocKwac, we go through a list of python scripts.
The scripts were initially found to be non-functional because of the many updates that had occurred in the external libraries, and on python itself, over the years of

hiatus. it was necessary to get hands on functions and fix these errors, as well as improve some of the steps that were gross or fallacious.

Once the documents are processed you get for each: a clean html, a pdf with a standardized style, and an xml, all renamed and placed in specific folder paths. After that, you can move on to the stage of generating and aligning the RDF file, which will always have to be recreated from scratch, re-processing even old documents already in the final site. This procedure can create major management problems, since it is complex to check the correct shape of the file and a small error can compromise the functionality of the whole digital edition (since the RDF file allows indexing and parsing of texts and related metadata). During compilation of the RDF, some missing data related to geolocation (and thus directly related to map operation) and keyword search (related to advanced search filters) emerged. These parts of the RDF are not filled in because they are not present within the scripts used so far. it turned out that these steps were being processed by an external entity. Therefore, I am trying to reconstruct these parts by writing two new python functions that can make up for this lack. I developed a python script (location.py) for retrieving and storing geolocation data related to the locations named in Aldo Moro's documents. This script was then fed into the RDF generation process and allows for proper compilation of the data that enable the interactive map present in the location search to function. As far as keywords are concerned, I am planning to improve the current state of the search filter by clustering the main topics in the corpus of documents by making use of machine learning. As for the process of uploading new documents online, it is done through the use of some node js scripts and the vue framework. Even at this stage it seems that the whole site has to be recreated from scratch, starting from the new rdf, with the addition of the new documents that you want to insert. There are also numerous filters within the scripts that need to be identified and changed in order to unlock the pages of the new volumes and tomes that one wants to upload.

Parallel to this work, the creation of html and pdf files for the introductory notes and historical critical essays that are provided to us directly raw by the researchers is carried on. The html files for these specific documents have to be created by hand, just as all the menus that allow division into paragraphs have to be entered by hand, given the usually very long nature of these texts.

The project, as can be understood, is complex and multifaceted, with many layers of work and "patches" in which manual intervention must be made from time to time.

The principle of these problems (in addition to the architecture developed, which does not appear to be elastic for carrying out the work) is related to the poor documentation and visible approximations present in the workflow.