Scientific Report

Working Group 4, COST action IS1310

Mandi Astola

From January 15th to March 15th of 2017, I had the opportunity to carry out a short term scientific mission (STSM) for Working Group 4 of COST action IS310. The purpose of my project was to expand a database of early modern epistolaries, EROL. Digitizing the titles of epistolaries and contributing them into a single database will allow researchers to map out the amount and diversity of such published works. My aim was to focus on Italian epistolaries and digitize all the titles named in Corrado Viola's *Epistolari Italiani del Settecento* (2004). This report consists of a brief description of my process, the problems I encountered and my solutions, and some future recommendations. It was a great experience to dive into the world of digitizing Italian epistolaries. I am glad to have become confident in digitizing work and to have learned so much from the people I met during this project.

The process of digitizing Italian Epistolaries

I began my project with a visit to Oxford. In preparation for my visit I had learned to use Zotero, which was very easy to use. On my first day at Oxford, I received helpful and informative instruction from Miranda Lewis about the work of the Cultures of Knowledge team, the construction of EMLO and some of the common problems one encounters in digitizing correspondence. I ordered up the copy of Corrado Viola's *Epistolari* which is held in the Bodleian Library. From Tuesday to Friday, as well as some hours on Saturday and the following Monday, I worked in the library and typed the titles from this publication into EROL.

I worked each day in two 4 hour blocks with a lunch break. While in Oxford I worked roughly 30 hours on collecting titles and during this time managed to enter about 700 titles into Zotero. I also photographed the book so that I could carry on working at home. The photography took approximately two hours.

I continued my work for a few hours daily when I got home. Having made myself familiar with Justine Walden's EMIL database of Italian epistolaries, I began looking into ways of making it usable by importing it into Zotero. This took me quite some time as the process is not straightforward. With assistance from Laura Astola, my mother, I managed to import EMIL into Zotero. Fuelled by this success, I spent some time contacting other institutions and collectors of epistolaries to see if they possess digital databases which we could make use of, with the appropriate accreditation.

I received by email from Corrado Viola a copy of *Epistolari Italiano del Settecento: Repertorio Bibliografico* and *Primo Supplemento* as a PDF. Together with Laura, I imported the data from the PDF into Zotero. We used an OCR program to convert the PDF to text. Thus, we had a text with many bibliographical details. In every title, the author is written first. The subsequent categories (year, publisher, edition etc.) are separated by commas. Using a script written in Python we divided the parts of the titles separated by commas and made categories out of them so that they could be exported as Bib files. This was a success as both books could be imported quickly. We did not manage to separate all categories such as "place of publication" and "publisher" into separate categories, however. Many are still lumped into a single category.

What is still left to do

1. Cleaning up the data: The portion of data that I imported from Justine's database or from the PDF of *Epistolari* using a script is not perfectly categorized. There is still work to be done in making sure that every bibliographical piece of information (date, publisher, edition) is in the correct category in the Zotero catalogue. It is very hard to make a computer recognize which comma-separated parts of a bibliography are which. Teaching a computer to tell publishers from places of publication would take a long time. Bibliographical details are not rendered in a uniform style. My suggestion would be therefore to separate these details manually.
2. Digitizing Corrado Viola's Secondo Supplemento: Viola could not send me a PDF of the *Secondo Supplemento*, which is the second supplement to *Epistolari.* This is because the book is still in print and is being sold, so giving out a PDF would disadvantage the publisher. The best way to digitize the volume would be to order an e-book. If only a physical copy is available, I would strongly recommend scanning it and using appropriate software to convert it to text and import the titles into Zotero.

Problems I encountered and my solutions

I quickly encountered the problem that not all titles were findable on the internet. The fastest way to digitize titles is to locate them on a catalogue site like WorldCat using keywords, and then import the titles into Zotero using the Zotero Firefox plugin. I had trouble finding some of the works initially. The site SNB.OPAC.it proved very useful as it contained most of the titles I was looking for. I saw the site listed in Justine Walden's database and it was recommended also to me by Dirk. I quickly developed a workflow for my searches. If a title was not found on SNB, then I would search it in WorldCat; if this didn't work then I would search for it in Google and see if I could find a catalog site on the first page. There were only very few works that could not be located at all. Those which did not appear in a catalogue site when searching on Google would usually appear at least in a reference. Then I would copy and paste as much as I could of the reference into Zotero, which is faster and more accurate than typing it manually. One or two works were findable only on Amazon, in which case I would import them.

I lost some time because I misunderstood or did not pay attention to the types of works that the titles referenced. For instance, many titles referred to sections of books. In these cases it was much faster to search in Google Books. This realization made my work faster.

Upon leaving Oxford, I could not take the *Epistolari* with me or photocopy it in its entirety. Therefore I took pictures of it with my tablet. Spreading this photography work over two days allowed me to look over the pictures carefully and determine whether I needed to re-take some of them. Indeed, some of the first batch of pictures did not turn out very sharp! I also noticed that the pictures which I took in the morning, with daylight, were much clearer than those taken in artificial light.

Back at home, I added the titles from the photographs I had taken. Many of the images were still quite blurry and hard to read, which made my work slower. I tried SimpleOCR to see if I could automatize the reading of the text, but this was unsuccessful. I believe it was mostly because the lighting was uneven and the page was slightly convex, making the lines of text curvy rather than straight and thus impossible for the program to read.

Due to the tedious and slow nature of the work I began to look into the possibilities of importing titles digitally. This was a very good decision as it was much faster. My skills with Excel (needed for digitizing Justine's database) were limited, however, and it took some time for me to find out how to import the database. I found the instruction by Beatrice Penati, on Academia.com, particularly helpful. (link: https://www.academia.edu/1747759/From_spreadsheet_Excel_to_.bib_file_-_a_simple_solution.). As for the conversion of the PDFs into Zotero entries, Laura Astola's help was absolutely essential.

Recommendations

Searching for titles on the web:
For Italian titles, use SNB.OPAC.it and Worldcat.org. These are the sites that have the best catalogue of Italian titles. I also noticed that sometimes the name of the collector, which at least in *Epistolari* is sometimes listed first, is not the best key word to search with. If a title does not come up in a search, try different keywords. Sometimes this does the trick.

Photographing a book:
If you manage to make good quality scans or photographs, the reading of the text can be automatized using a program like SimpleOCR. Try to photograph in good lighting and lay the book as flat as possible so the lines of text are straight. If it is possible to scan the text, this would be a preferable option.

Working ergonomically:
I recommend using a book rest and a book snake to keep the pages in position. If you are working with photographs of a book, I recommend using two screens, with a picture of the book on one screen and the browser and Zotero open on the other.

Digitizing:
I would try to do as much as possible digitally and automatically if possible. This is often much faster. Even without programming skills, it is much faster to scan a book and automatically digitize the text. There is also a great deal of information on the internet on how to do operations with a text editor and how to convert files from one format to the other.

Zotero:
I found that it is possible to search a title in one's own library in Zotero, but once the results come up, the collection of the titles is not visible. For me this was not problematic, but I can imagine it being a hindrance to someone who is searching and re-categorizing titles. I have learned that Mendeley shows the collection of a title when searched.