Report for COST-STSM-IS1310-36364

# “Extracting and visualizing structured data from university matriculation registers”

At the University of Oxford, 05-01-2017 to 26-01-2017
Host: Prof. Howard Hotson, University of Oxford

Submitted by Glauco Mantegari

# Introduction

Throughout the early modern period, members of a university community enjoyed specific rights and privileges, and also (typically) paid a matriculation fee. It was therefore in the interests of both institutions and individual students to keep a public record of membership, known as a matriculation register. Compiled in Latin, each entry typically includes the following information:

- Given name and surname of matriculant;
- City and/or region of origin;
- Day, month and year of matriculation;
- Other data (subject studied, age at matriculation, etc.)

These registers were carefully maintained for most German and Dutch universities, and preserved in the university archives. Most of these preserved registers were published in the decades before and after 1900, and equipped with indexes of people (and sometimes places). As a result, most are now out of copyright and available in digitized form online.

Transforming the semi-structured printed sources into fully standardized and structured digital data would allow the exploration of large patterns of exchange which up to now has remained invisible to scholarship, and to visualise those patterns in engaging, informative and crowd-pleasing ways.

The STSM aimed at researching on the methodological and technical aspects connected to the definition of a technical workflow to process, save in structured form, and make available for analysis, people, place and date data from OCR'd matriculation records. These activities were done in close collaboration with Prof. Howard Hotson (Faculty of History) and project staff of Cultures of Knowledge, as well as technical staff employed at other university departments.

## Long-Term Goals

We have identified a set of long term goals that will be supported by the definition of this initial workflow and subsequent activities. The points mentioned here below form part of iterative processes.

1. Creating a list of personal names and run matches with existing authority files accessible through APIs to check if they already exist in the cloud, such as VIAF[1]

[1] https://viaf.org/

And Wikidata[2]. VIAF identifiers will also help us identify people from the registers that also appear in EMLO[3]. Other potential sources of prosopographical data, such as those concerning people around Comenius and Hartlieb can be exploited as well.

2. Starting a geographical gazetteer using the names contained in the entries and connect to existing geo-gazetteers, such as GeoNames or others.
3. Designing specific visualizations for exploring distributions and networks, and their changes over time.
4. Extending the dataset with information coming from different registers using efficient OCR requiring minimal manual validation.
5. Linking these datasets based e.g. on persons and places.

## The Registers

The list of registers available (see Appendix A) offered us the opportunity to evaluate the best fit for an initial experimentation. We decided to focus mainly on the Gröningen register[4], which covers years 1615-1914 and offers a well-defined structure for each entry, where the fields are separated by commas. A few complicating factors had to be taken in consideration:

- Sometimes one or more of the fields are missing
- Sometimes additional information is inserted in the penultimate place
- Occasionally brackets are used to relate entries together
- There are footnotes as well as handwritten notes

Some pages of interest in the downloaded book are not readable (e.g. pp. 62-63)

# OCR Tests and Results

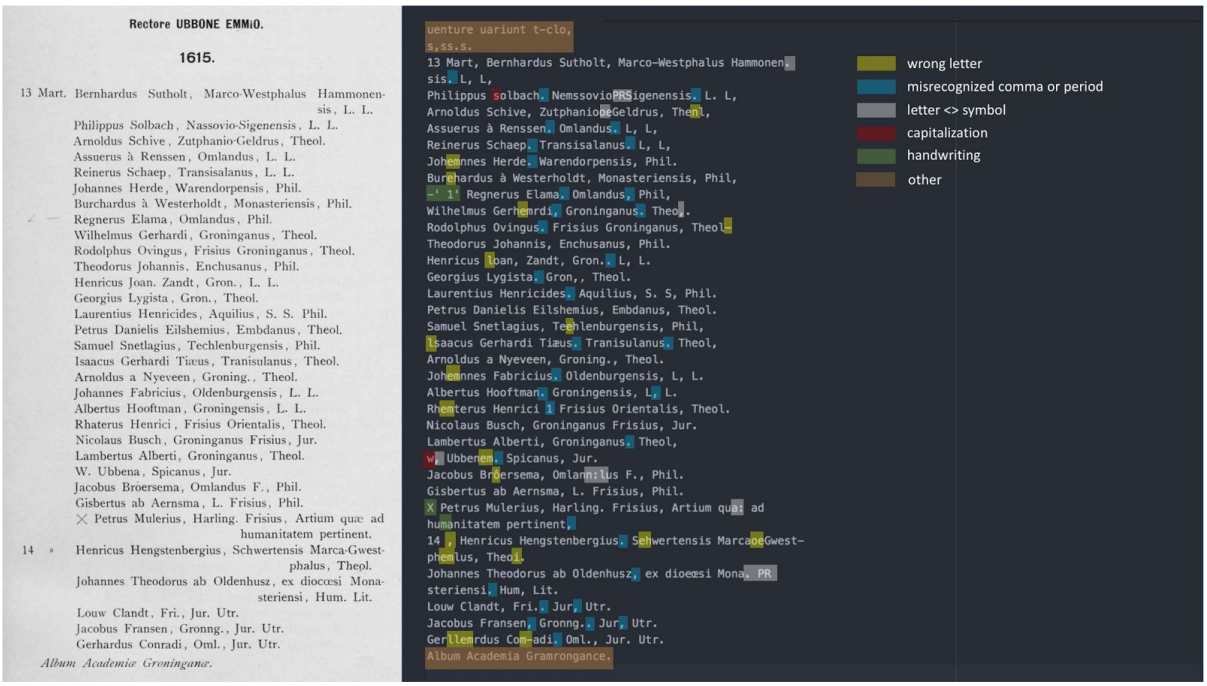
A useful reference for our work were the materials of the “OCR and postcorrection of early printings for digital humanities“ workshop: https://github.com/cisocrgroup/OCR-Workshop. The OCR expert involved in these tests was Mr. Nick White at the University of Oxford.

## Gröningen

We used the Tesseract engine (https://github.com/tesseract-ocr) and worked on the plain txt output files. The pages of the register are split into two columns that are managed separately during OCR. The related output files are named accordingly.

Figure 1 represents a visual comparison of the original text of a page (left) and the resulting OCR (right) after some tests and optimizations of the engine. The main categories of issues are highlighted using different colors.

[2] https://www.wikidata.org
[3] http://emlo.bodleian.ox.ac.uk/
[4] http://facsimile.ub.rug.nl/cdm/ref/collection/boeken/id/4060




The most crucial problem was comma recognition, since commas are used to separate the different fields in each entry. Minor problems concern misrecognition of certain symbols (e.g “a” is sometimes misrecognized as “em”), wrong capitalization and issues with handwritten notes.

Without considering the issues with commas the engine was able to identify with a 100% accuracy 15 entries on a total of 26 (60%). The PDF of the register has also an embdedded OCR text. The date numbers are often recognised as a separate column and so are displayed separately to the lines of names, which would make associating the names with the dates impossible.

## Manual correction

Manual correction of the pages for years 1617-1626 was also performed. Some problems such as comma recognition, the addition of “»”, and common misspellings could be solved easily in semi-automated ways. Others, like those related to personal and place names took more time. On average, validating a page (2 columns) required 40-45 minutes. The goal of this activity was to create a ground-truth files to train neural network-based OCR engines (see next section).

The manually corrected file consist of 613 entries:

- Personal names
  - 601 unique
  - 12 repeated. Based on the dates they are probably homonyms.
- Origin
  - 321 unique place names
  - Most of the place names are variants or abbreviations: e.g. “Domo Groninganus”,”Gron.”,”Gronigiensis”,”natione Gronninganus”, etc.
- Subject of study:

- 18 unique names
- Different terminology can be used to identify the same general subject: e.g. “Humanitatis”, ”L. L.”, ”Litt.” for Letters

# Herborn

The Herborn register was the object of a second set of tests. The font used in the text did not work well with Tesseract OCR, so we tried out Ocular, a newer neural network-based OCR engine. This approach will also likely be used for all the other registers, as it should generally provide better results, at the cost of more computing time (which is in general a price worth paying).

## Ocular OCR

Ocular[5], from the Berkeley NLP Group, needs to be trained for each register, which is a largely automatic process. It is very much a case of the more computer time we allot it, the better it gets. Therefore, it is important to do enough testing to find a sensible amount of time to allot to the training process. It looks like approximately one day will be about right.

Ocular is also quite slow in recognition, and will take about 1 day to OCR a whole register, once the training is in place. It takes significant computer resources, too, most importantly at least 64GB RAM. Using a shared computing cluster where we can specify resources needed would likely be the sensible way to go for this.

## OCRopus OCR

Another OCR engine which uses neural networks is called OCRopus[6]. It should offer broadly similar quality results to Ocular, though the training requires rather more human input. However if the OCR results are better the time investment beforehand may be worth it, versus manually correcting OCR errors. I haven’t had time to test OCRopus on the registers yet, but it would definitely be worth doing.

## Layouts

Complex layouts will need to be handled before running the OCR, as is the case with any OCR engine. This means anything which is not straightforward lines of text, so includes for example brackets incorporating multiple lines, columns. Generally the OCR of each part will need to be put back together, in some cases (as with footnotes) this will not be very straightforward. Both the splitting apart and re-integrating of the layout will be done with small custom scripts using the OpenCV image library, but a reasonable amount of human labour will be required, and they probably won’t often be re-purposable by other registers.

[5] http://nlp.cs.berkeley.edu/projects/ocular.shtml
[6] https://github.com/tmbdev/ocropy

### Comments

The precision of the OCR of herborn is under evaluation, given the time needed to train the engines, optimize them, and get quality results.

At a general level, our tests already gave us useful estimates (Table 1) about the resources that could be reasonably needed when dealing with the complete set of registers.

| | |
|---|---|
| **Layout handling** | Very dependent on register; 2-10 hours of work |
| **OCR training** | 1 day computer time, 15 minutes human labour |
| **OCR running** | 1 day computer time, 15 minutes human labour |
| **Layout re-integration** | Very dependent on register; 0-4 hours of work |
| **Manual correction** | Very dependent on register; 4-24 hours of work? |

# Data transformation

We used this subset to data transformation and create a basic schema that could be used also for other registers in the project:

- Year
- Month
- Month normalized: transformed month names into numbers
- Day
- Day normalized: transformed single digits into 0-prefixed digits
- Name
- Origin (sometimes not present)
- Origin normalized: grouped variant names of the same places using clustering algorithms. E.g. “Gröningen” is the normalized version for “Domo Groninganus”,”Gron.”,”Gronigiensis”,”natione Gronninganus”, etc. This would require further specialized work of a researcher with all the necessary knowledge of the geography of the years considered
- Subject (sometimes not present)
- Subject Normalized
- Notes

## Towards a geographical gazetteer for matriculation registers

Place names associated to each student come in a variety of forms:

- For the most part they are adjectives: e.g. “Frisius”,”Gronninganus”,”Westfalus”, etc.
- There are both complete and abbreviated variants: e.g. “Amstel.”,”Amsteldamensis”.
- There are different granularities ranging from single towns (e.g. “Amsteldamensis”) to historical regions (e.g. “Bohemus”).

- Within regions there could be additional specifications: e.g. “Frisius”, “Frisius Orientalis”,”Frisius Occidentais”, etc.
- There are vague indications: e.g. “ex Veteri Praefectura”.

The historcial gazetter for the project will build on these names, and it will record all the meaningful attested variants of place names. No such gazetteer for this domain exists so far. We designed an draft of the schema for the gazzetteer:

- ID
- Normalized name in English
- Name in text (i.e. variant)
- Type: town, region, etc.
- Year attested: the years in which this particular variant is attested in association with a matriculation entry
- Latitude of a representative point
- Longitude of a representative point

The gazetteer could also be a useful resource for the transformation of OCR data from registers that e.g. do not provide separators in the entries.

We also evaluated some external reources supporting the identification of place names in Latin and providing additional related information. In particular:

- Orbis Latinus by Graesse: http://www.columbia.edu/acis/ets/Graesse/contents.html.
    - A translation of the abbreviations and German words used by Graesse: https://rbms.info/lpn/abbreviations-and-german-words-found-in-graesse/.
- The Cerl Thesaurus: https://thesaurus.cerl.org.
- EMLO list of place names
- The Leiden matriculation register

# Visualizations

We used a list of 18 universities (mostly in Germany) that contains the absolute number of students who matriculated in each university during any given year from 1598 to 1662. Attached to every university for a particular year there are indications about the confession: Catholic, Lutheran, and Reformed/Calvinist. Minor anomalies (highlighted in the dataset) were ignored, but need to be kept track of for later inspection.

We explored possible ways of visualizing on interactive maps the data with the aim of creating proof of concepts and test interaction patterns quickly. For this we did not not program a custom system, but we used the Tableau[7] software.

We created four main visualizations:

1. Showing the distribution of confessions between the different universities and how they change over time.

[7] https://www.tableau.com/

2. Showing confessions, and the swell and shrink of the number of matriculations for each university with absolute values.
3. Showing confessions, and the swell and shrink of the number of matriculations for each university with normalized values (ratio). The ratio is defined as the number of matriculation for each year divided by the moving average for 3 years before that year.
4. Show swell and shrink as a percentage related to average matriculations in each university during the period 1616-1620 (that is, the period immediately before the Thirty Years' War began to affect matriculations rates), and creating bins to classify and style the symbols on the map accordingly:
    - Matriculations at that average for the period 1616-1620
    - Matriculations at 50% above that level
    - Matriculations at double the 1616-20 average
    - Matriculations at 50% below
    - Matriculations tending to 0%
    - At 0% of the figure for 1615-1620 the university's location are marked with a red 'X'

# Dashboard

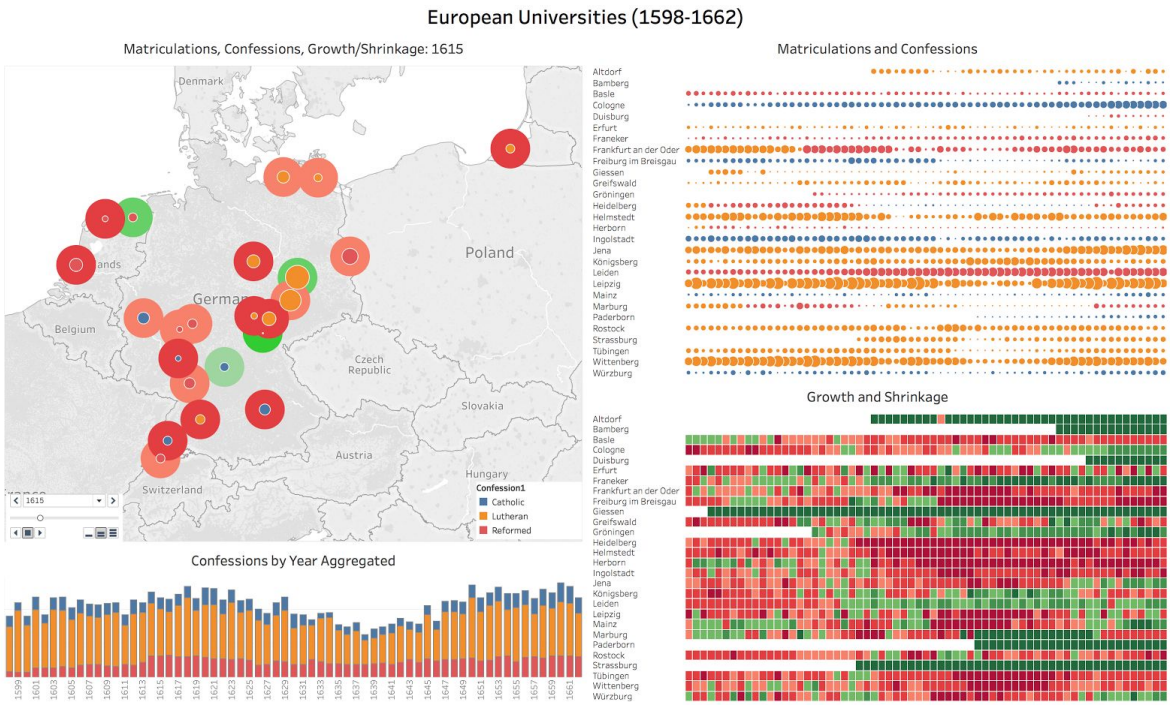
The 4 visualizations together with the related metrics helped us design a dashboard combining meaningful data points, filters, and symbols.

The swell and shrink calculated according to the criteria for visualization 4 appeared to be the most interesting. We reviewed the previously-defined bins and identified the following groups:

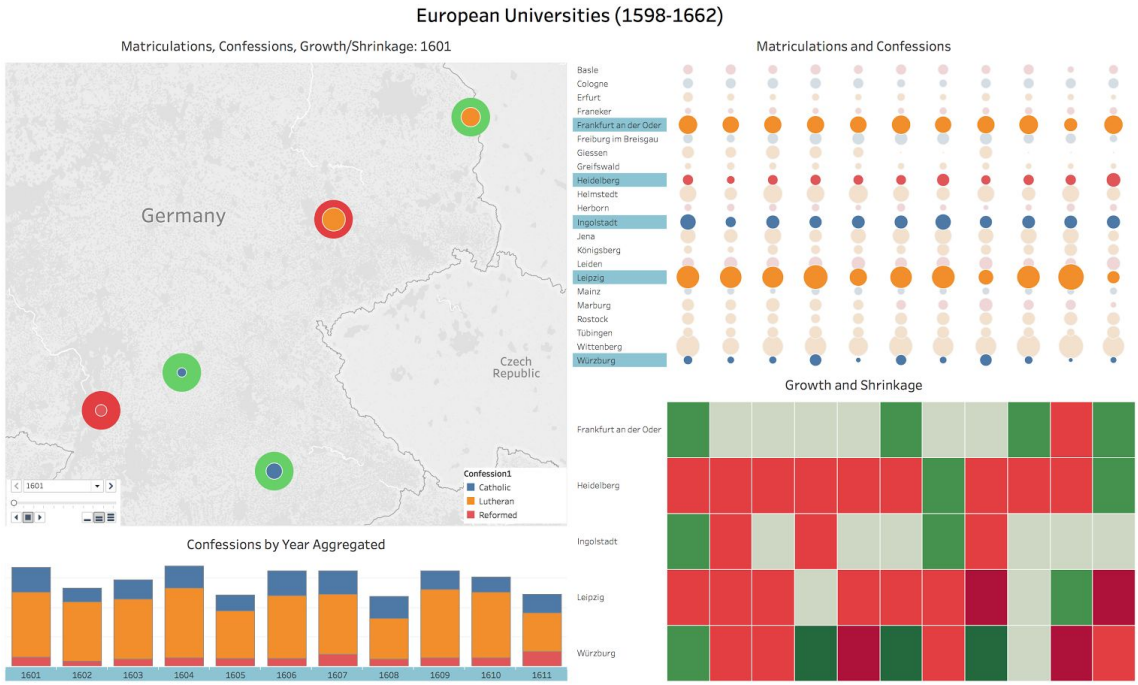
1. Less than -100%
2. -100% to less than -50%
3. -50% to less than -10%
4. -10% to less than 10%
5. 10% to less than 50%
6. 50% to less than 100%
7. More than 100%

Since some universities start after the 1616-1620 timespan we use to determine average and percentages, the symbology for those would correspond automatically to that of group 7.

The dashboard includes a map together with a timeline and two additional visualizations representing the absolute number of matriculations for each university as well as the swell and shrink groups.




The different components of the dashboard act as filters. Selecting one or more data points on one or more of them will update the dashboard accordingly. The screenshot below shows a selection of 5 universities from the “Matriculations and Confessions” view and 10 years from the “Confessions by Year Aggregated” view.



# Appendix A: List of Registers

A valuable list of matriculation registers is available at:
http://genwiki.genealogy.net/Universit%C3%A4tsmatrikel#Ver.C3.B6ffentlichungen
Most of the ones we consider for the matriculation project have already been scanned and posted online.

## Netherlands

The Dutch registers are very simple in structure, and are a good place to begin.

- **Franeker**: not available: *Album studiosorum Academiae Franekerensis* (1968).
- **Groningen**: http://facsimile.ub.rug.nl/cdm/ref/collection/boeken/id/4060 (see below)
- **Leiden**: http://digital.ub.uni-duesseldorf.de/urn/urn:nbn:de:hbz:061:1-453594 and https://babel.hathitrust.org/cgi/pt?id=mdp.39015028339888. This register has already been transformed into a database, which will be posted on the Huygens ING website (close partners in the COST Action).

## Germany

The registers of Heidelberg and Herborn contain much more information, and should be tackled later

- **Bremen**: http://brema.suub.uni-bremen.de/content/structure/70073. Includes data on subsequent studies of individual matriculants.
- **Heidelberg**: http://www.ub.uni-heidelberg.de/helios/digi/unihdmatrikel.html
- **Herborn**: https://archive.org/details/verffentlichung01nassgoog
- **Kassel**: https://archive.org/stream/zeitschriftdesv10landgoog#page/n198/mode/1up the register itself starts here: https://archive.org/stream/zeitschriftdesv10landgoog#page/n291/mode/1up.
- **Marburg**: http://www.uni-marburg.de/uniarchiv/studierende (simple entries)
- **Zerbst**: evidently digitized (and accessible in the US?): https://catalog.hathitrust.org/Record/002203290

## Switzerland

The registers for Basel, Bremen and Geneva include detailed information on the matriculants, including other universities in which they matriculated. This will be protected by copyright, but could be used to validate results obtained by other methods.

- **Basel**: digitized by Hathi Trust but not available. Iincludes data on subsequent studies of individual matriculants: https://catalog.hathitrust.org/Record/000143701
- **Geneva**: annotated edition still very much in copyright. Includes data on subsequent studies of individual matriculants:

http://www.droz.org/eur/fr/50-le-livre-du-recteur-de-l-acad%C3%A9mie-de-gen%C3%A8ve. An older edition of the bare list is available here: http://www.e-rara.ch/zut/content/titleinfo/8252674