COST Action STSM 2017

Extended Report by Mikkel Munthe Jensen

Table of Content

1.0: Introductory Note and Purpose of the STSM

The content of this report is based on the Cost Action IS 1310 STSM to Oxford done in the period January 23rd to 27th 2016.

This one-week STSM had the main purpose of helping to plan the project and AHRC research grant proposal, which currently has the working title 'The Thirty Years' War and the Golden Age of the Dutch Universities'. In order to assist in the planning, three goals were set for this STSM.

1: Outlining workflows for the digital parts of the AHRC proposal.
2: Outlining a draft for the AHRC proposal.
3: Participating and present VIA in AHRC Planning meeting January 24th 2017.

During my STSM, and in relation to point 1, I worked together with Glauco Mantegari, a digital expert from Milan and fellow STSM grantee, in digitalising and visualising a sample of matriculation registers. In order to get most out the short research stay and the competences of Mantegari, majority of my time and focus was given to outlining matriculation registers in the first point.

2:0: Outlining workflows for the digital parts of the AHRC proposal.

Since a large part of the AHRC research project relies on well-developed and well-conceived digitalisations, visualisations and digitals tools, which all demand proper workflows, I have outlined such workflows for the two major digital parts of the project; i.e.

1: The digitalisation of matriculations registers (mainly student movement)
2: The digitalisation of professor prosopographies.

In this regard, a few important points need to be stated: Firstly, the following workflows are outlines of identified tasks and should thus be seen as a planning tool for the project and a preliminary workflow and not as a complete finalised workflow to launch. Secondly, the overall approach to these digital workflows cannot be and is not a linear approach, in the sense that one task needs to be 100 % completed before another begins. The workflow will most likely be initiated in different flows, meaning when one task or subtask (e.g. a register or part thereof) is finished another can be initiated, which then again refuels the first steps (which often are the more complicated matters); i.e. an **iterative approach**.

Each task that has been identified comes with a short description of the task, what needs to be done and potential problems and possibilities. Moreover, to each task, a section called management considerations has been added, which includes the following points:

- Time estimate: a (at this stage) rough estimate on how much time there is needed to complete the task.
- Risk: Stating potential risk that might occur.
- Risk of occurring: Stating the probability that the above-mentioned risk would occur.
- Consequence if occurring: Stating what would happen if the risk occurs.
- Risk prevention: Stating what could be done in order to prevent or lower the probability of the risk to occur.
- Additional budget consideration: Stating what might increase the budget

2.1: Workflow outline for digitalizing matriculation registers

It should be noted that the time estimates for the entire part very much depends on the number of registers and entries that we decide to use. The first part of the workflow – the planning stage (2.1.1) is thus very decisive for the rest. In order to deal with this yet-not-taken-decision on exact time (for registers) and number of studied institutions (i.e. at the end of the day the total number of registers and entries) two approaches has been taken. The first is to estimate all tasks according to one register and, where fitting, to each entry (or 1000 entries). Each register has furthermore been estimated to contain an average of 6000 entries, which indeed might be too high.

The second approach has been to create a 'Registers of Reformed Uni 1540-1660 - Initial Planning by MMJ' (attached to the extended report), which would help not only to take informed decisions on the scope of the project (time and space) relating to the matriculation registers but also provide an important aid for comparing matriculation registers and the entire process of beginning the OCR (points 2.1.2-2.1.3). This list contains the name of 39 reformed institutions and information on availability of sources, period of sources and research period, OCR issues and similarity (organisational content and order, and textual appearance like commas).

The workflow outline for the research project's digital part 1, matriculation registers, consists of 5 main headings, of which many contain several tasks. Lined according to headings, the identified tasks and preliminary workflow has been outlined in the following way:

2.1.1: Initial Planning

2.1.1.1: Essential decisions on the scope of the project

The essential decisions on the scope of the project (time and space) will have great impact on the time that need to be invested in each task. It might not change the workflow, as such, but the assessment of time required will be highly influenced.

A: Deciding specific which universities

- Reformed universities, but which specific?

B: Deciding the specific time period

- From university foundation/conversion to 1660?

- Decision could be based on the universities and the availability of the data – see the attached excel file: "Registers of Reformed Uni 1540-1660 - Initial Planning by MMJ". This file contains a list of all the names, I could find, of reformed universities in the period 1540-1660. The first spreadsheet (named OCR Prepared list) contains information on whether the printed version of the matriculation registers exist, whether they have been digitalised and open access available and OCR related information, such as total number of entries, total number of pages, number of entries in research period, separators, OCR issues, complexity, matriculation information (data, name, place, faculty etc.) and whether the registers also contain list of professors and promotions. The second spreadsheet (named Matri. Reg. 1540-1660) contains information on each register related to each specific university – mainly basic university information (year and place) and bibliographical information for each register.

2.1.1.2: Calculating Number of Entries

- A rough estimate of the number of entries is important for the manual validation process after OCR Data transformation.

Based on HH numbers:

Heidelberg	8691
Herborn	2806
Marburg	5163
Basel	7105
Subtotal:	*23765*
Leiden	24315 (*Note*: Have been digitalised: Zoeteman's data)
Franeker	6556
Groningen	6240
Subtotal:	*37111*
Total(approx.):	61000

Total (approx.) Leiden excluded: 37,000

Other Reformed Universities

In the excel file, 'Registers of Reformed Uni 1540-1660 - Initial Planning by MMJ' all other reformed institutions are listed and were it has been possible, information added. In total I have found 39 institutions, which in the period 1540-1660 at some point were Reformed. I have found the printed sources and online versions. However the list is not 100 % completed, and would need more time to complete.

2.1.1.3: Compare Matriculation Register

- Although matriculation registers to a high degree are similar, they are, however, unfortunately not identical. In order to determine estimated time that needs to be invested in digitalising these register, the diversity and similarity among them are necessary to do.
- This similarity refers to two issues:
    - Organisational order: To what extent is the order of data (temporal, personal, geographical and disciplinary data) similar?
    - OCR related: To what extent are the placement of data (big spaces between the data points?) and the division of data (commas, dash, etc.) similar?
- In the "Registers of Reformed Uni 1540-1660 - Initial Planning by MMJ" these similarities and differences are noted.
- From the point of view of transforming OCR into structured text, the organisational order is not problematic.
- OCR-related differences, however, can pose large problems for the OCR development. So far several of the matriculations register do not have demarcations between the data (for instance between name and date, or name and place), which possible makes it difficult for the OCR to separate the different data categories.
    - *Note*: Nick (OCR Expert) told us that as long at OCR can recognise the characters, the separation between data etc., is not so problematic.
    - *Note*: In order to document/provide evidence for this, I have asked DWork at Heidelberg Universitätsbibliothek (Dr. Thomas Wolf) for a higher quality issue of the university register – they have provided me with a high-resolution version, which I have send to Nick White. Nick does not think, it will make a huge different, but for the sake of evidence and documentation we will try anyway.

2.1.2: From printed matriculations to structured data

This step is the actual digitalisation workflow and contains four major tasks. As stated in the introduction, the approach to these tasks should be an **iterative approach**, meaning that the successes and failures of one task should improve the other and that the workflow is not linear but circular.

2.1.2.1: Transforming Printed Matriculation Registers into OCR data

Task: Each individual printed matriculation register need to be transformed into digital text using OCR

- For technical development: see STSM report from Glauco Mantegari
  - *NOTE*: A circular improvement process should be implemented:
    - Validate a sample (e.g. a page) of each specific OCR-transformed matriculation register
    - Feed the validated sample to the OCR for it to learn (optimise) and circulate the process
- Each matriculation register – although they might contain similar or identical data, are not 100 % identical structured, which might require specific work from OCR/digital-specialist on each register (see also point 2.1.1.3)
- Note: Estimates made by Nick White after first rounds of test (Very positive!)

Time estimates for OCR processes per register (Nick White)

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?**

- Note: Prioritise the 'easy' registers to have a list of places, names etc. and start the entire workflow – which then again would help out on the more complex registers.

Management considerations

- Time estimate per register (average): 1 Week (Allocation of time from one to another – Estimate confirmed by Glauco and Nick)
  - *Note*: For 'easy' registers this process will take a day or two (according to Nick) and is an automated process (speed depends on computer power and the time here are not taking from the OCR expert's time)
- Risk: Since the registers are different (but the difficulties are related to issues in the text) some complex registers will take longer time to complete.
- Risk assessment of occurring: Medium
- Consequence if occurring: Very High: Slow down all other activities related to the data in the register (Thus begin with easy registers, so the workflow will begin)
- Risk prevention:
  - Be conservative in the time estimate
  - Hire OCR experts
  - Complete preliminary test on different approaches (machine learning on complicated issues, 'other approaches' on other problems (parentheses for instance))
- Additional budget consideration:

- Hire OCR expert
- Pay for storage and resources for computer power (The tests/OCR takes up a lot of computer power – increased power - faster results - faster workflow)

Note: Perhaps tests with manuscripts should be performed for matriculations registers, which are either not published or where they are published alphabetically. This could be relevant for the following institutions, to which I so far have found no printed registers:
- Gymnasium Beuthen an der Oder
- Hooghe School en Oranjisch Collegie te Breda
- Die University
- Gymnasium Hammonense
- Hohe Landesschule Hanau
- Markgrafen-Gymnasium (Durlach/Karlsruhe)
- Montauban University
- Montpellier University
- Pädagogium Casimirianum Neustadt
- Orange University
- Orléans University
- Orthez University
- Saumur University
- Sedan University
- Gymnasium Arnoldinum
Note: many of these are French universities; perhaps I have searched the wrong places to find these! (perhaps this is less important if focus stay at the Dutch-Reformed area)

2.1.2.2: Mass validation of OCR data

Task 1: Each OCR-transformed entry need to be manually validated
- This is a very time-consuming task and is highly dependent on the success of the OCR process; i.e. the better OCR the less work is needed to validate.
- Validation: That each data set is prober separated, that all letters are prober written and all numbers/data are transformed correctly.
- This validation task could be performed by students

Task (additional)
- Create a guide for students to use to validate the OCR data
- What to look for, what to be aware of, what to change, what to register etc.

Management assessments
- Time estimate per entry: 20 seconds (Note: This needs to be confirmed and is based on high quality OCR results)
- Time estimate per page (30 entries): 10 minutes (20x30/60)
- Time estimate per 1000 entries: 5-6 hours (20x1000/60/60) (Note: intense work)
- Risk 1: That each validation take up more time than anticipated
- Risk 2: That the validation is flawed
- Risk of occurring 1: Medium
- Risk of occurring 2: Medium (But it is almost impossible to have 100 %, but takes the workflow further)
- Consequence if occurring 1: High: Further processes will be delayed + extra cost to students
- Consequence if occurring 2: Low: Correction process will begin in stage 2: Visualisation (point 2.2)
- Risk prevention 1: Be conservative in the time estimate/Depends a lot on step 2.1.2.1
- Risk prevention 2: Be conservative in the time estimate (more time more quality)
- Additional budget consideration: Employment of students

2.1.2.3: Delimitate OCR data

- Task: Transform validated OCR-text into delimited OCR-data
- For technical development: see STSM report from Glauco Mantegari
- The validated OCR-text needs to be delimited accurate and according to the categories (columns) of the 'Table of structured matriculation data' (see below 2.1.2.4)
- This is a time-consuming task, especially for 'complicated' registers where no clear demarcation between data exists.

Note: This could be done simultaneously with the mass validation process, since each entry should be checked.

Management assessments

- Time estimate per entry: 20 seconds
- Time estimate per page (30 entries): 10 minutes (20x30/60)
- Time estimate per 1000 entries: 5-6 hours (20x1000/60/60) (Note: intense work)

- Risk 1: Encountering difficulties of demarcating between data (e.g. name and 'adjective' place of origin)
- Risk 2: Mistakes (due to time pressure/fatigue)
- Risk of occurring 1: Medium
- Risk of occurring 2: High
- Consequence if occurring: High - Further processes will be delayed + extra cost to students
- Consequence if occurring: Medium – Corrections need to be made, additional work for later validation/correction phases will increase.
- Risk prevention:
    - 1: Strong: Make good guide/introduction to the students, who need to do the demarcations. As the individual register is structured the same way throughout the register a pattern could be found
    - 2: Ensure that the students, who do this, not are working 5-6 hours in a row, but perhaps only 2-3 hours, breaks in between etc. – i.e. optimising their concentration.
    - 1+2: Be conservative in the time estimate
- Additional budget consideration: Employment of students

2.1.2.4: Transforming OCR-delimited-data into structured data

- Task: Transform OCR-delimited data into structured data in a table (Open Refine)
- For technical development: see STSM report from Glauco Mantegari
- Create a data table (Table of structured matriculation data) in Open-Refine in which the OCR-delimited data is imported.

Additional validating process

- Use explorative analysis to identify mistakes – using clustering and similar tools to identify mistakes
- Do statistic on values that are missing, mistakes etc. If a significant occurs, use it to optimise previously steps

Preliminary draft of 'Table of Structured Matriculation Data'

- University*
- Year*
- Month
- Month Normalised
- Day
- Day Normalised

- Name*
- Name Normalised*
- Academic Status
- Academic Status Normalised > Dictionary
- Origin
- Origin Normalised > Gazetteer
- Faculty
- Faculty Normalised > Dictionary
- Notes

*Mandatory fields (Data that we have in all registers)

Management assessments
- Time estimate per register: 1 week
- Risk: Too many errors/mistakes in the imported data would slow the process
- Risk of occurring: Low (due to the previous the validation processes).
- Consequence if occurring: Low (Corrections will be done in the Open Refine).
- Risk prevention:
    - Be conservative in the time estimate.
    - Each time the statistics of failure would improve the imported 'delimited OCR-data'.
- Additional budget consideration: None

2.1.3: Adapting structured data into actionable data

Task: Adjust and adapt all the data in Open Refine into applicable data
- Fill out the added 'normalised' columns in the table
    - Subdivided task 1: Normalise basic data: dates and month to numbers (E.g. May 3rd = 05 – 03).
    - Subdivided task 2: Normalise/Merge complex data: (e.g. place or origin: gron, grönni. Gronni = Grönningen).
- The second subdivided task are rather complex and for its completion a comprehensive 'Gazetteer of Places' and 'A Dictionary of Latin abbreviations' have to be created.

Management considerations (NOTE: Based on the idea that the Gazetteer and Dictionary both are in place, not complete but a part of the process)
- Time estimate per register: 1 week
- Risk: Too many unknown or complicated 'places' would slow the process
- Risk of occurring: High (but decreasing with the expanding gazetteer of places)
- Consequence if occurring: Medium (will take up more time, but is at the last step in this stage).
- Risk prevention:
    - Be conservative in the time estimate.
    - Each time, the Gazetteer of places would improve our knowledge and optimise our routine in this task
- Additional budget consideration: None

2.1.3.1: Creating a 'Gazetteer of Places'

A gazetteer of places needs to be developed and created, since
- There exist many variants of each place name in the matriculations registers.
- Many of the places are/or can be in obscure or not well known places
- The use of Latin places names (need to transcribed into modern English)
- Place of Origin sometimes appear as an adjective to the matriculant's name

- Sometimes the place of origin is not a city/village but a region or county

Tasks (main):
- Create a first scheme for the Gazetteer (We will figure out the granularity – historical hierarchies)
- Clustering variants
- Identifying primary form for each cluster in English
- Flagging the things that are unknown, uncertain etc., which need to be treated in some ways

Tasks (Refinements)
- Managing vague indications; e.g. "Ex Veteri Praefecture"
- Concern the granularity: Setting hierarchies (number of types)

- For technical development: see STSM report from Glauco Mantegari

Draft Scheme of the Gazetteer
- ID
- Normalised Name in English
- Name in text (Exactly how it is written in the text)
- Type: Town, Shire, County, Region
- Years attested: E.g. Gron. can be dated in this, this and this year.
- Latitude of a representative point
- Longitude of a representative point

Note: Every entry in the Gazetteer is a single variant of a place name (Basic entry of origin)
Note: We can leverage this with Orbis Latinus, Thesaurus, EMLO Gazetteer, etc.

Management considerations
- Time estimate: High, 3-4 month full time (8 hours a day with skilled works)
- Risk:
  - We do not know how different the place names are – look like they are very different
  - Too much time to figure out the historical place names
- Risk of occurring: High
- Consequence if occurring:
  - 1: Slow down the process of adapting the data (2.1.3)
- Risk prevention:
    - 1: Be conservative in the time estimate.
    - 1: Start early and start small. (Iterative approach)
- Additional budget consideration: None

2.1.3.2: Creating 'A Dictionary of Latin Abbreviations'

Task: Create a dictionary of Latin abbreviations
- The matriculation registers contains many Latin abbreviations of the subject of study, the academic degree they hold and/or the faculty to which they matriculate. This needs to be normalised to which purpose two additional columns have been added to the table scheme (see 2.1.2.4: 'Academic Status Normalised' and 'Faculty Normalised').
- In order to optimise and standardise the adapting of data process a 'Dictionary of Latin Abbreviations' should be created.

Preliminary example of such a Dictionary

Latin Abbreviation	Meaning	Discipline	Academic Status
J. U. D.	Juris Utriusque Doctor	Law	Doctor of Law
Jur.	Juris	Law	Student
Jur. Utr.	Juris Utriusque	Law	Student
L. L.	Legum	Law	Student
S. S. L. L.	Sensu Stricto Legum	Law	Student
Med.	Medicine	Medicine	Student
Humanitatis	Humanities	Philosophy	Student
Litt.	Litteraria	Philosophy	Student
Mathemat.	Mathmatics	Philosophy	Student
Phil.	Philosophia	Philosophy	Student
Sacr. Litt.	Sacra Litteraria	Philosophy	Student
Th.	Theologia	Theology	Student
Theol.	Theologia	Theology	Student
(blank)			
Med. et Phil.			
Phil. et Jur.			
Phil. et Med.			
Phil. et Theol.			
Theol. et Phil.			

Note: Issues that need to be addressed (marked with red boxes)

- *Doubles*: Sometimes matriculants have been inscribed in two faculties (how to solve this in the data table) Note: Be careful, many problems arises if doubles exist
- *Blanks*: There are several registers where indication of faculty does not appear in the registers or are missing in other matriculation entries.
- *Academic status*: Sometimes matriculants holding academic titles such as J.U.D. (Doctor of Law). This needs to be separated and indicated to which the column 'academic status normalised' have been added. However, should matriculant without any degree (written in the registers) simply be listed as 'students' or should they preferable stay blank?

Management considerations

- Time estimate: Low, 2-3 days (A week max, but the dictionary will be extended as more matriculations are digitalised)
- Risk: Unknown abbreviations that need to researched
- Risk of occurring: Medium
- Consequence if occurring: Low – (Research or ask the academic network to help out)
- Risk prevention: None
- Additional budget consideration: None

2.1.4: Match and Link people between different registers – creating a merged list.

Task: Match and link people between the different registers
- Comprehensive task to match and link data relating to each unique individual with the result of having one merged document, where the matched and linked data are connected
- In order to start this process, individuals need to be identified and given a KEY ID. The difficulty is that persons with similar names appear in the data or that the names a misspelled, but in reality are the same person. To identify unique individuals, a matching process should take place through the following criteria:

Matching and Linking Process

Match by name (use also fuzzy-matching (show similar/misspelled names)).

↓

- Results (list of matching names/similar names).

↓

Match this list by 'Place of Origin', since place of origin is a stable fact and 'most likely' will appear the same in all registers.

↓

- Results (list of matching names/with same place of origin).

↓

Match this list with year and date of the registers (If the time period is too far from each other (e.g. 1588 and the next is 1640, these two entries does not match, and we are probably taking about two different individuals, who bear the same name and are born the same place but in a different time period. Think also on the universities as institutions).

↓

Identify and distinguish each unique individual based on the matching process.

↓

Link each data entry with identical KEY ID to each other in a merged data table of all the matriculation registers. In this merged version identical persons should be flagged with the KEY ID.

Management considerations

- Time estimate: 1-2 weeks per register (Very uncertain)

Note The numbers of identical persons that need to be matched and linked are unknown until we have the first results of a couple of registers. Then we can begin to have an overview of the numbers and adjust the time estimates per match. - Risk 1:

- Risk: Unable to match/decide on some of the individuals
- Risk of occurring: High
- Consequence if occurring: Need to decide what to do with those cases/entries (flag them, use inferred/uncertain/unknown markers)
- Risk prevention: None
- Additional budget consideration:

2.1.5: General quality control processes

Task 1: Create a general quality control process to optimise both workflow and quality of data
- As we have more experience with the different task in the workflow, we should create a quality control process.
- We need to define what and how we need to control (which parameters) – what are the difficulties in the data.
- Define a workflow for the quality control and implement in subsequent rounds of work.

Task 2: Create a general quality control process to optimise both workflow and quality of data

- During the workflow, document encountered issues, failures, mistakes and successes for each step. What are the difficulties with the data? Are there specific issues and recurrent problems?
- With this information, optimise the process and workflows and re-calculate the time estimates and risks for each section.
- Disadvantage: Take time to document
- Advantage: Optimise process, counter problems, and give more precise estimates.

Management considerations
- Time estimate: these are running task alongside the workflow

2.1.6: Summary of management consideration

2.1.6.1: Time Estimates (Preliminary - at this stage of project planning)

2.1.2.1. Transforming Printed MR into OCR Data:	1 week per Register
2.1.2.2. Mass validation of OCR Data: (1 Register average = 6000 entries, 5-6 hours per 1000 entries)	1 week per Register
2.1.2.3: Delimitate OCR Data: (1 Register average = 6000 entries, 5-6 hours per 1000 entries)	1 week per Register
2.1.2.4. OCR data to structured data	1 week per Register
2.1.3. Adapting structured data	1 week per Register
2.1.3.1 Gazetteer of places (running development)	14-16 weeks in total
2.1.3.2 Dictionary of Latin Abbreviations (running development)	1 week in total
2.1.4. Matching and linking people	1-2 weeks per Register
2.1.5. Quality control process (running development)	1 week in total

Total
Per register: 6-7 weeks
Other (running development): 16-18 weeks

Note: This is an optimistic estimate and does not include additional time spent on meetings, planning, supervision, management etc., but is strictly related to the task above.

Total divided among specialists
Per register: 6-7 weeks
- OCR expert: 1 week (point 2.1.2.1)
- Student work: 2 weeks (points 2.12.2 + 2.1.2.3)
- Scholar + Digital expert: 3-4 weeks (2.1.2.4+2.1.3+2.1.4)

Other (running development): 16-18 weeks
- Digital expert: 14-16 weeks (2.1.3.1)
- Scholar: 2 weeks (2.1.3.2+2.1.5)

Note: this is a very rough division and should not as such be taken for granted, as the tasks do not stand alone by one person. The OCR expert need input help from Digital expert and scholar, the last two need to work a lot together with the different steps, the students need supervision by the scholar/digital expert etc.

Total by example of 10 registers (60000 entries)
For the registers: 60-70 weeks
Other (running development): 16-18 weeks

Total: 76-88 weeks or roughly 1½ years of work

2.1.6.2: Risk Estimates (Preliminary - at this stage of project planning)

- Risk: That unanticipated problems occur.
- Risk of occurring: High, since we are dealing with digital workflows and data models that needs to be developed.
- Consequence if occurring: High – delay results and breach the time schedule and budget
- Risk prevention: Have as large a margin as possible – introduce more conservative/less optimistic time estimates.

2.1.6.3: Needed personnel

One OCR Expert
One Digital Expert
One Scholar
Students

2.2: Rudimentary workflow outline for visualisation of matriculation registers

There are two audiences that could/should be targeted, but each requires a different approach and thus also different budget assessments. However, in both cases the visualisation could also be used to assess and fix any mistakes that still might appear in the data set (point 2.1).

2.2.1: Visualisations of Matriculation Registers for academia

Task: Create visualisations in order to be able to explore, analyse and interpret the data

- This is a comprehensive task and requires close cooperation between scholar and digital expert
- Define: What, how and why we want to visualise this data; i.e. what do you want to display – e.g. geographical distribution of students and academics. In other words, the content of the visualisation.

Note: see the work Glauco Mantegari have done in his STSM report
Note: No preliminary workflow has been created for this

This part of the project has five major steps (Note: Not a strictly linear workflow!)

- Step 1: Data transformation – please see point 2.1.

- Step 2: Design the visualisation tools based on what, how and why to visualise (Mockups, Make user experience test (how are users using the interface, how will scholars use it, use existing tools to test prototypes e.g. tableau.)
  - *Time estimate: 2 weeks*

- Step 3: Transform and adjust data for visualisation
  - *Time estimate: 2 weeks*

- Step 4: Development tools (custom visualisation, java script etc.)
  - *Time estimate: 1½-2 month*

Management considerations

Time estimate in total: 2½-3 month

NOTE: Time estimates are very uncertain as it depends on what we want to visualise (step 2)
- Risk: That poor and uncertain data exist – how to display this (This is very much connected to the quality of the data).
- Risk of occurring: Very high – almost impossible to avoid.
- Consequence if occurring: Slowing down the process, create uncertainties in the data/visualisation output.
- Risk prevention: Combat these uncertainties in phase 1 (point 2.1)
- Additional budget consideration:
  - Infrastructure: Dedicated server and server space (Secure and efficient access) + IT support.

2.2.2: Visualisations of Matriculation Registers for the public (optional?)

Task: Create visualisations for the public to explore
- This is a less comprehensive task but requires still cooperation between scholar and digital expert.
- Define: What, how and why we want to visualise this data for the public to use/explore.

This part of the project has three major steps
- Phase 1: Data transformation (point 2.1)
- Phase 2: Design of the visualisation/exploration tool (how to visualise)
- Phase 3: The creating of tools

Note: No preliminary workflow has been created for this

Management considerations – not filled out – need expert advice!
- Time estimate:
- Risk:
- Risk of occurring:
- Consequence if occurring:
- Risk prevention:
- Additional budget consideration:

2.3: Task and (rudimentary) workflow for professorial prosopographies

Since focus was laid on the workflow outline for matriculation registers, this STSM report contains only the headings for a rudimentary outline of professor prosopographies.

2.3.1: Initial planning - Identifying data sources

2.3.2: Developing data model

2.3.3: Collecting and transforming data according to data model

2.3.4: Creation of Visualisation/exploration

2.3.4.1: Data transformation (step 5.2-5.3)

2.3.4.2: Design the visualisation tools based on what, how and why to visualise (Note: see VIA below)

2.3.4.3: Transform and adjust data for visualisation

2.3.4.4: Development tools (custom visualisation, java script etc.)

3.0: VIA Presentation and its relation to AHRC Project

For the AHRC planning meeting, I presented the digital tool VIA, Virtual Itineraries of Academics for the participants. VIA is a digital visualisation and exploration tool for scholars working with academic travel data, created by myself, Marco Quaggiotto (Politechnico Milano) and Joëlle Weis (Luxembourg University). The main function of the tool is to help scholars to explore the relationship between three categories of parameters; geography, chronology and a variety of prosopographical attributes connected to early modern academic travellers. VIA's strength lies first and foremost in its use of interconnected parameters, meaning that when one or several of the above-mentioned parameters are selected, all other parameters adjust accordingly. This instantaneously adjustment of all data in all parameters allows scholars easily to conduct both broad as well as very specific data exploration. Based on the three categories of parameters, the interface is divided into three interconnected frames, i.e. a geographical frame, a chronological frame and a prosopographical attribute frame, which provide the user with a good workable view.

In relation to the AHRC project, VIA or the data model could be used to visualise and explore the professor prosopographies. However, since the pivotal point of VIA is the individual travel and the not the individual persons (each professor), the Key ID of VIA lies at the travels, whereas the Key ID at the AHRC project lies with the person. This is a fundamental difference, which needs to be addressed early in the stage of development. For a big research project like the AHRC, my recommendation would be to use the experience we have from developing VIA and visualising prosopographical data to create a visualisation and exploration tool specific for professor prosopographies, based on the data model (point 2.3) and with each unique name (individual) as key ID. The Key ID would thus also correspond both with the matriculation registers and EMLO.

4.0: Outlining a draft for the AHRC proposal

During the STSM, I have managed to comprehend and outline the variety of points and sections an AHRC standard research grant proposal needs. An AHRC proposal consists of two main sections, each containing several subsections. The main two sections are:

1: An online form that need to be filled out as a part of the application process. This form consists of the following subsections:

- Objectives
- Summary
- Outputs
- Academic Beneficiaries
- Impact Summary
- Summary of resources required
- Proposal Classification

2: Project attachments, which are:

- Case for Support
- Curriculum Vitae
- Publication List
- Visual Evidence (optional)
- Technical Plan
- Justification of Resources
- Pathways to Impact
- International Co-Investigator Head of Department (If applicable)
- Project Partner Letter of Support (If applicable)

In the attached 'draft outline of AHRC Research Grant', I have added own notes and comments from the meetings to each sections and subsections to help facilitate the proposal writing process. I would like to highlight three subsections:

1: *The case of support*. This attachment is the actual project description. I have added some text, which mainly simply has been copy/pasted from Howard Hotson's initial outline draft. Under the heading of research methods, I have added three (optional) headings, called 'methodological approach', 'technological innovation' and 'sources and data'. Besides the research questions unique character, our strongpoints are also the digital/technological innovations and academic beneficiaries; i.e. the usefulness of data models for all historians and data 'containers' for especially other early modern historians.

2: *Technical plan*. It is important to note, that although the technical elaboration within the case of support only consists of a technical summary (a few lines), the AHRC project attachments also contains a technical plan (max 4 pages) in which the more technical issues can be elaborated. In this regard, it should also be noted that in the project attachments, an attachment named 'visual evidence' exist, where we can provide screenshots/links to our digital tools.

3: *Dissemination and impact of research*. In the case of support, I have noted all our comments on possible outcomes of the project. This consists both of a book, technical and academic papers. However, thoughts and attention should also be given public impact of the project; i.e. how and

why this project is valuable for the public. A couple of ideas have been written under this section, such as the anniversary of the Thirty Years' War, the King and Queen of Winter and ERASMUS programme. It should be noted, that the project outcome (academically and publicly) appears under several headings and sections of the project; i.e. under online form (outcome, academic beneficiaries, impact summaries – shorter versions), under the case of support (Dissemination – build on the online form sections), and individual project attachment called 'pathway to impact' – that is how these ideas for public impact would be achieved. It seems thus that both the academic and public outcome play a large role in the assessment, which therefore demands attention.