INTRODUCING OPEN SYLLABUS EXPLORERS

What are the most important books worthy of study at university? What does it mean to be in the ‘Western Canon’ and what books does it include? What are some of the pedagogical differences between universities? What do the books we read tell us about what a university education values today? How much can we really learn from what is included on a syllabus?

These are just some of the questions that we will be discussing in this project. We are greatly indebted to [The Open Syllabus Project](http://opensyllabusproject.org/) which first inspired the project and supplied us with our initial dataset. We have obtained our data from beta version 0.4 between January and March of 2016. While still in its early stages, this project has already made accessible great quantities of data that has enabled projects such as this. If you are interested in our findings, we strongly encourage you to learn more about The Open Syllabus Project and see if you would like to [get involved](http://opensyllabusproject.org/get-involved/).

Open Syllabus Explorers is a summative project completed for an undergraduate level Digital History class at [Carleton University](http://carleton.ca/) taught by Professor [Shawn Graham](http://electricarchaeology.ca/). Due to its nature, this project takes every effort to explain its methodology and show the work process. We welcome any feedback or comments you might have and sincerely hope you find this project enjoyable, engaging, and educational.

Explore away!

THE PROCESS

Our project can be summarized by the obtention, cleaning and manipulation, and visualization of the data we collected, followed by analysis, reflection and presentation of our findings. As we mentioned in the introduction, our initial data was obtained from [The Open Syllabus Project](http://opensyllabusproject.org/) beta version 0.4. We started with a simple copy and paste of the general top 1000 texts into an open text editor (in this case [Notepad++](https://notepad-plus-plus.org/)). At this point, the data looked like [this](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-top1000.txt).

The above image shows a small sample of the dataset of 1000 entries we collected from the general top 1000 texts from The Open Syllabus Project. With this data obtained, we turned to the far more difficult process of cleaning the data.

*Cleaning Process*

The format of this copied data was not easy to work with and required cleaning to convert the data into a workable form. For a small dataset such as one with 10 entries as shown above, it would be quickest to reformat data manually. For a large dataset of 1000 entries however, manually cleaning this data would take several hours. As we planned to obtain seven additional top 1000 datasets for specific countries or institutions for our [comparative analysis](http://opensyllabusexplorers.thenewhistorian.com/blog/comparative-analysis/), it was clear we needed a more efficient method of cleaning. For this, we used Regular Expressions that are available in many open text editors, including Notepad++.

Regular Expressions work like a control-F search function on steroids. They can be tricky to learn how to design, but once completed they can make data cleaning very efficient. For those who are unfamiliar to Regular Expressions, an excellent introduction can be found [here](http://www.themacroscope.org/?page_id=521), an instructive tool can be found [here](http://www.regexr.com/), and a useful cheat sheet can be found [here](http://regexlib.com/CheatSheet.aspx). It took some time for us to write a good Regex for this project and it might not be the most efficient in the world, but it certainly does the job. The process outlined below is applicable to any dataset taken from The Open Syllabus Project in the beta stage.

| Step | Regex | Explanation |
| --- | --- | --- |
| 1 | Manual deletion | Remove unnecessary headings at the beginning of the page. |
| 2 | $ replace with , | Places a comma at the end of each line (this will help avoid data bleeding into other columns later) |
| 3 | alt-shift ~ | Place cursor at start of first line of text, hold alt-shift and place cursor at the beginning of the last line of text, type ~ character and it should appear throughout the document. |
| 4 | (^~\d) replace with ~\1 | Adds a second ~ to the beginning of any line starting with a number character. This identifies the separators between the data. |
| 5. | ^~ replace with ~, | Adds a comma after the first ~ in a line. This will be useful later. |
| 6. | ^~, replace with nothing | Removes all ~ from any line of text that does not begin with a number character. |
| 7. | (\<^\w.+>) replace with ,\1 | Places a comma at the beginning of every line without a ~. |
| 8. | \r\n replace with nothing | This moves all data onto a single line. |
| 9. | ~ replace with \r\n | This removes every ~ character and replaces it with a new line. This will arrange all data by ranking in your table. |
| 10. | alt-shift , | Repeat step 3 but this time insert commas after each column of numbers and remove comma from any four digit number. This step will be a little finicky but will work. |
| 11. | ,, replace with , | This replaces repeated commas with a single comma. |
| 12. | ,\_ replace with , AND \_, replace with , | Note: \_ represents a space character. This removes spaces between columns. |
| 13. | alt-shift remove column spaces | Repeat step 3 process to remove extra space between different number columns. |
| 14. | Add Headings | Headings are: Rank,Count,Score,Text,Author-last,Author-first,Date |
| 15. | Save | Save as txt and csv files for easy editing. You may choose to clean up the text further by individual line in Microsoft Excel. |

The most difficult part of this process was constructing this formula. Once the formula was obtained, the cleaning process only took about 5 minutes to complete. At this point, the full set of our cleaned data can be found [here](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-top1000-cleaned.txt) and appears in the following form:

*Supplementary Information*

After obtaining our original dataset and cleaning it for easy usage, we were then able to turn to adding supplemental information. The organic work process was nowhere near as clean and tidy as we will be presenting. In reality, there was a great deal of trial and error, failure, redundant work, and multiplying files. While some of the following narrative will address this process, we are most interested in portraying the supplementary information that was added to the freshly cleaned dataset. That process can be defined by the following elements: final manual cleaning of the dataset, removing duplicates from the dataset, adding a gender category to the top 100, adding a nationality category to the top 100 with a corresponding coordinate category, and adding dates in a machine-readable format. When taken together, this was the [final dataset](https://github.com/prefaces/Final-Project/blob/master/OSE-General-Top100-final.csv) produced.

MANUAL CLEANING

This step is unavoidable when cleaning data into usable form. Regex performs a phenomenal function through its ability to quickly and efficiently clean large tracts of data, but it is only as strong as the original data. Additional commas in the titles of text or missing data columns (i.e. author’s first name) could often throw off the formatting of the data. We found it easiest to view the data in Microsoft Excel after saving the previous dataset as a csv file. Here, we could easily visualize improper formatting and rearrange columns where required. This stage was time-intensive, but it did complete the data cleaning process. For our purposes, this was time well spent.

DUPLICATE REMOVAL

An additional consequence of using data from a project in its beta stage was that there were many flaws within the data. Some of these issues have been acknowledged by The Open Syllabus Project, while others have not:

The Explorer is very much a work in progress. As you may discover, it gets a lot of things wrong. Fixes and improvements will be incremental. But it also gets a lot right, and makes curricula visible and navigable in ways that we think can become valuable to authors, teachers, researchers, administrators, publishers, and students. We hope that this beta version of the tool convinces you of that potential…  
Our matching algorithms also have some difficulty with short titles based on commonly-used word, and more so when these lack an author. This affects a very small number of works, but some significant ones like The Bible and “The Constitution.”

Clearly, this is still a project in its early stage, though one with a high degree of potential nonetheless. Due to some of these flaws in the data, final results have to be greeted as a best approximation rather than authoritative. Returning to the data flaws, we identified some of the repeating data and tried to consolidate. There were certainly limitations to our method, such as only considering the top 1000 texts, but it was a step in the right direction. Using Microsoft Excel, we sorted the data by author and identified obvious duplicates in the data (i.e. Aristotle’s Ethics and Aristotle’s Nicomachean Ethics are clearly the same text). After identifying duplicates, we added the text count values together, reordered the list according to the count values, and created a new ranking list. It should be noted that we removed the Teaching Score column as we could not recreate this with the limited available data after reordering the data. In the end, 71 entries were removed as our top 1000 turned into a top 929 and can be seen [here](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-top1000-withgeo-norepeats.csv).

We offer no guarantees that these changes more closely resemble the input data. We followed the following principle for some of the more complicated instances: whenever we were faced with a single instances of a text and a compilation of the text with others, we chose to count the compilation text as an instance of both single texts. For example, the top 1000 show three instances of conflicting works by William Blake: Songs of Innocence, Songs of Experience, and Songs of Innocence and Experience. In this instance, we counted each instance of Songs of Innocence and Experience as both separate texts. While we can make no assertions for how these were taught in class, this principle is most consistent with the data available to us. Our method certainly was not perfect as we saw the unlikely occurrence of Sophocles’ Oedipus vault to first place (we consider this unlikely, but plausible). Furthermore, our method was poor in addressing issues with author distribution (i.e. Tocqueville vs De Tocqueville). In the end, we identified duplicates within the following [texts](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-top1000-duplicate-removal.txt) and made changes accordingly. Just like the makers of The Online Syllabus Project, we too tried to interpret the data to the best of our ability, but acknowledge that there are instances of unavoidable flaws.

GENDER

At this point, the decision was made to only consider the top 100 of the general top 1000 for our analysis. This was primarily determined for the purpose of visualization of our data as even the top 100 could be difficult to identify. After this decision was made, the remaining elements (Gender, Nationality, and Date) became simpler as we only considered 100 entries rather than 1000. Identifying genders was done manually and was a fairly quick process. Whenever uncertain, Google was a useful resource to quickly determine the gender of an individual. Though a simple step, this allowed us to do some valuable gender analysis of the Western Canon and allowed us to contrast the genders of authors of popular texts between institutions in later sections.

We would like to briefly acknowledge that we understand that making assumptions on the basis of a gender binary has its flaws. We encourage you to keep our chosen methodology in mind when considering our findings. With this warning in mind, we believe analysis on the basis of the gender of authors in a large sample can offer valuable insights both into the perceived importance of an individual’s gender at a certain historical time period **and** modern treatment of the Western Canon.

DATES

Determining the dates of a text turned out to be much more difficult than we anticipated. As we considered this an important component of our analysis, we put a lot of extra work in to identify the dates of a text wherever possible. The Open Syllabus Project acknowledges the incompleteness of the date field in the project, saying, “Our best estimate is that our current approaches successfully map 70% of dates and 60% of fields.” Recognizing this limitation, we attempted to fill in the missing information.

For this, we found a [resource](http://sonic.net/~rteeter/greatbks.html) containing data on the so-called ‘Great Books’ of the Western Canon with accompanying date and nationality fields. Though not a perfect or complete list, we recognized the value of this data. As such, we gathered it, cleaned, and used it to update our data. After spending several hours trying to find a way to automate this data reconciliation (we were unsuccessful), we decided to do this manually. Again, this took a couple hours but was necessary. After this process, we had identified dates and author nationalities for many texts, but had many still remaining. At this point, we used Google to try to identify dates (and nationalities) of all authors in the top 100 texts (we were successful in identifying approximately 95%). [Here](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-top1000-withgeo-norepeats.csv) is a link to this stage in progress.

Finally, before we move on, it is useful to identify a couple of methodological decisions made. As birth and death date data is more readily available than publishing date data, we decided to follow The Open Syllabus Project and include the birth date of the author as our primary focus. While it would be clearly preferable to know the year each text was published, this information is sadly not widely available. In our final formatting, we rendered this data as birthyear-birthmonth-birthday (with the latter two entries as 01-01 for simplicity). For this method, we rendered BCE years as negative values (though none of our timeline or graphing software responding particularly well to dates before the turn of the Common Era). While there are obvious limitations to our final rendering of dates, we believe it is sufficient to conduct the analysis in which we are interested.

NATIONALITY AND COORDINATES

Identifying the nationality of authors was very much intertwined with the previous element so we will not go into great detail about the process here. An additional element to the previous process of assigning nationalities to the top 100 texts was the corresponding coordinate values, rendered in the form latitude,longitude. We were not interested in particular city locations when that information was available, but instead, for consistency sake, used the generic country coordinates that we found on the following [site](http://www.latlong.net/). Though a relatively simple addition to our data, this was a valuable element for later visualizations.

Some interesting methodological problems arose when we considered assigning nationalities to our authors, considering the relatively recent historical emergence of the modern nation state. For the purposes of later visualizations, we used modern designations for older locations (i.e. Rome became Italy, despite Italy not existing until the 19th century). Again, this information is not included for complete accuracy, but rather to serve our interest in geographical representation of the texts. After these elements were considered, the complete dataset for the top 100 can be seen [here](https://github.com/prefaces/Final-Project/blob/master/OSE-General-Top100-final.csv).

*Applying Methodology*

To serve our interest in conducting a comparative analysis, we decided to obtain additional datasets. We identified three countries of particular interest and four institutions of particular interest. Each cleaned dataset can be seen below:

[Canada](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-Canada-top1000-cleaned.csv)

[United States](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-US-Top1000-cleaned.csv)

[United Kingdom](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-UK-Top1000-cleaned.csv)

[Carleton University](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-Carleton-top1000-cleaned.csv)

[Harvard University](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-Harvard-Top1000-cleaned.csv)

[Cambridge University](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-Cambridge-Top1000-cleaned.csv)

[London School of Economics](https://github.com/prefaces/Final-Project/blob/master/OSE-raw-LSE-Top1000-cleaned.csv)

The Open Syllabus Project is growing, but continues to have a limited amount of data available (i.e. only a handful of Canadian schools are included). Based on these limitations, we believe these sources will yield us the most interesting results. While there would certainly be value in comparing all countries and institutions with data available, this was not within the scope of our exercise.

The decision was made to use the top 25 texts from each dataset for our comparative analysis. While this is not ideal, it allows us to visualize each institution effectively and minimizes the considerable workload of having to prepare a dataset with 1000 texts. After identifying the top 25 texts from each data source, gender, dates, nationality, and coordinate elements were added to each dataset. Surprisingly, several datasets showed a high amount of variation from the top 100 general dataset so we were required to turn to Google to try to identify additional elements. It should be noted that no attempt was made on our end to remove potential duplicates from these datasets, largely due to time constraints and the focus of this project. For our comparative analysis, perfect datasets were less important to us. There were a couple instances where two of the same text made it into the top 25. In this instance, we did remove the duplicate and reassign the ‘Count’ values. Furthermore, the top ‘text’ for Harvard University was a reference to an on campus library, so that was also removed. The final section provides these final datasets as a resource,

*Presentable Datasets*

The following datasets are used for some of the more specific analysis conducted later in this website. We would like to draw your attention to the first set – General Top 100 – as it will be used for the majority of the canon analysis work.

[General Top 100](https://github.com/prefaces/Final-Project/blob/master/OSE-General-Top100-final.csv)

[General Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-General-Top25-final.csv)

[Canada Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-Canada-Top25-final.csv)

[United States Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-US-Top25-final.csv)

[United Kingdom Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-UK-Top25-final.csv)

[Carleton University Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-Carleton-Top25-final.csv)

[Harvard University Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-Harvard-Top25-final.csv)

[Cambridge University Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-Cambridge-Top25-final.csv)

[London School of Economics Top 25](https://github.com/prefaces/Final-Project/blob/master/OSE-LSE-Top25-final.csv)

[Carleton Humanities](https://github.com/prefaces/Final-Project/blob/master/Humanities-dataset-cleaned-final.csv)

(Note: the final dataset for Carleton Humanities will be explained in later analysis.)

*Conclusion*

While long and fairly dry, we consider this page to be particularly important to our project. We hope our process and methodology was clear and that the resources we provide are helpful. We are more than happy to consider any comments, questions, or suggestions you may have. We hope you enjoy the more interesting analysis sections to come!

Canon Analysis

Syllabus data such as this immediately lends itself to analysis of the Western Canon. While there are many arguments that have been made about the issues with treating the Western Canon as an objective entity and the issues with what sources get admitted into the canon, there is still considerable value to considering the most prominently assigned texts at universities across the world. Furthermore, though limitations with the data have been pointed out on the Process page, the data available from [The Open Syllabus Project](http://opensyllabusproject.org/) can offer a window into the Academy. The following analysis will be based on the top 100 texts as identified through the [dataset](https://github.com/prefaces/Final-Project/blob/master/OSE-General-Top100-final.csv) we created.

We believe many interesting conclusions can be coaxed out of this data. This page will concentrate on three primary elements of our dataset (Timeline, Geography, and Gender), before turning into some more general observations that can be made. An important part of the digital humanities is in recognizing the limitations of digital analysis. In keeping with this limitation, we will not be making any grand assertions based on the available data, but rather will consider, analyze, and explore. We hope to make progress toward understanding why certain texts were historically and continue to be considered part of the Western Canon. Furthermore, we hope to start to address the question of what this data implies for what we consciously or unconsciously consider to be part of the Western Canon today. If we are only successful in beginning to ask these questions, inspire more research, and encourage our readers to consider these questions for themselves, we will be very happy. We hope you join us in exploring the Western Canon below.

*Timeline*

First, we turn to a historical timeline of the top 100 texts we have identified. This information will help us understand what time periods of history yielded the texts and ideas our culture values the most. We start with an interactive timeline of these texts for you to consider for yourself, before we begin to analyze what we see. This feature was created using [TimeGlider](http://timeglider.com/). We have displayed the first 80 texts in the top 100 on this timeline, sized according to their prominence in the top 100 (i.e. how frequently the text was assigned). This decision was made for readability, as displaying all 100 makes it difficult to interpret the data. Please note that all dates given on this timeline are for the known birth date of their author. We invite you to interact with this feature below.

Perhaps unsurprisingly, this timeline shows a concentration of texts generated by Greek authors between 800-400 BCE, followed by only a few texts between 400 BCE and the 16th century (note that the Bible was not identified in our dataset – please see the Process page for an explanation). Following a small decrease of the number of texts in the 17th century, the 18th, 19th, and 20th centuries show a steady increase in the number of texts located within each century.

Particularly noteworthy in this timeline is the enormous length of largely uninterrupted time between the texts from Classical Greece to a couple representatives of the Renaissance and early contract theorists in the 16th century. With the obvious exception of the Hebrew and Christian Bibles which were not identified in this dataset, it is clear that modern universities do not value the texts or ideas of this period as highly as they value texts of Ancient Greece or from post-Renaissance Europe. When you contrast this with a typical Medieval school or university, the difference is further emphasized. From a historical standpoint, it is quite clear that Saint Augustine and Chaucer are not representative of 2000 years of history between Ancient Greece and the Renaissance. This is not to say that this historical time period is neglected in modern universities, but rather to say that an interest in reading primary texts from this time period is not as pronounced as reading primary texts from either before or after this rather significant gap.

The lack of texts by authors associated with the Enlightenment is somewhat surprising. Of these authors, only Voltaire, Franklin, (debatably) Rousseau, and Jefferson are represented. Again, we believe this points less prominently toward a lack of value placed on this time period by modern universities, but rather a lack of interest with reading primary text materials. The significant increase in number of texts assigned from the 19th and 20th centuries suggests that historical representation is less important to the selection of primary sources taught at universities than other factors such as modern subjects of study, the interest in ideologies, or perhaps even accessibility to students. Though these are all hypothetical suggestions without any firm grounding, it would be fascinating to compare this timeline with a timeline of texts considered most important for education even a century ago.

Finally, we would like to highlight three periods of time on this timeline that are particularly significant for the modern university curriculum. These periods include: 428-324 BCE, 1797-1819 CE, and 1922-1946 CE. The first period recognizes the important contributions of Plato and Aristotle to philosophy. The second period includes are remarkable clustering of authors including: Shelly, Mill, Darwin, Melville, Tocqueville, Jacobs, and Marx. This points to a particular interest in authors productive roughly between 1830-1870. Finally, the third period includes another noticeable clustering of authors including: Kuhn, Foucault, Huntington, Achebe, Morrison, and King, as well as several textbooks and writing guides. This period points to a particular interest in authors productive roughly between 1950 and 1980. While this analysis would require significant research, this data points to the particular importance of the birth of Greek philosophy, the emergence of the modern political world, and the transition to early post-modernism and the corresponding critique of modernity. Again, while these observations do not carry significant weight, these trends are noticeable and worthy of future research. If they are correct, this may say a great deal of the pedagogical priorities of the modern university.

While these are merely observations from the data, they do point to the utility of digital history to identify these patterns and point the way toward future research. These identified patterns and suggested trends offer many points of departure from this macro analysis toward more detailed investigations.

*Geographic Distribution*

Similar to the timeline analysis, we can also consider the geographic distribution of texts. This analysis lends itself well to visual representation. This section will showcase a three different ways to represent the geographic distribution of texts. Some of these representations are simple, while others are more complex. Ultimately, all point to the prominence of texts from the United States and the United Kingdom, with Greece and France also having a respectable showing. At first glance, this demonstrates a clear prioritization for Western sources from the last couple centuries. While Greece is an obvious exception to this analysis, it is not difficult to argue that Classical Greece is the source of much of Western philosophy and intellectual culture.

The following visuals represent this data from the top 100 texts generated from syllabi. The first example was generated using the mapping visualization platform Palladio. Large circles indicate a high frequency in number of occurrences.

From this visual, the Western orientation of the top 100 is clearly visible. This visual highlights the lack of considerable diversity within this dataset. While there are a few outliers outside the general rule, the primacy of Western Europe and the United States is quite clear. The identification of these outliers is easy to see with the following Circle Packing graph, generated by the vector graphic software [RAW](http://raw.densitydesign.org/).

Though these are helpful visuals, they do not tell us a great deal beyond what we already know. We already know the Western Canon is primarily oriented on western texts written in intellectual centres such as Greece, the United States, the United Kingdom, France, Italy, and Germany. Even the outlier texts in the top 100 are largely written for a western audience. Limited conclusions can therefore be drawn from this geographical analysis. This geographical analysis is better suited to the comparative approach. The value of digital history is demonstrated nonetheless through visualizing clear trends and quantifying assumptions. While there will always be dangers associated with taking its analysis too far, analysis in moderation can offer a highly useful supplement to a more focused research paper. For the purposes of clarity, we have chosen to close this section with a Reingold-Tilford Tree visualization from the vector graphic software [RAW](http://raw.densitydesign.org/). Even though this section did not generate any earth-shattering insights, it should display the potential of visualizing data, as seen below.

GreeceSophoclesAristotlePlatoHomerThucydidesHesiodHerodotusEuripidesUnited StatesStrunkTurabianCampbellTwainHuntingtonMariebHallidayKingKuhnMankiwGilmanThoreauDeitelGibaldiHackerFitzgeraldMelvilleFranklinMorrisonCozbyJacobsChopinCreswellHardinJeffersonRawlsDueEllisonBushHurstonWalkerFukuyamaKrugmanTanenbaumGermanyMarxWeberNietzscheFranceTocquevilleDescartesRousseauFoucaultSartreVoltaireUnited KingdomShelleyHobbesMiltonChaucerShakespeareDarwinConradWollstonecraftMillSmithDefoeLockeMoreWordsworthMalthusBergerGellnerAustenHuxleyPaineEliotBlakeWoolfItalyMachiavelliVirgilOvidPalestineSaidNigeriaAchebeLibyaAugustineRussiaStravinskyChinaConfuciusCanadaStewart

*Gender and the Canon*

As with the geographic analysis, this section will not reveal anything that is not already known. It is widely acknowledged that there is a distinct lack of female authors considered to be part of the canon. Our data shows no different. Before we comment on this rather stark gender divide in the texts we read, it is useful to take a look at two visuals of this data. The bar graph was created using Microsoft Excel while the Circle Packing visual was created using [RAW](http://raw.densitydesign.org/).

While these findings are not particularly shocking, they are somewhat saddening. The 20th and 21st centuries have seen positive steps forward regarding equal treatment of genders, though there is much work still to be done. We are certainly not advocating for an equal gender distribution on every syllabus as this compromises the integrity of a syllabus. Instead we pose questions on a historical basis. How else can we understand a historical time period other than through a diversity of views? We would like to highlight some of the work currently being done at Simon Fraser University, McGill University, and the University of Pennsylvania on something called [New Narratives in the History of Philosophy](http://www.newnarrativesinphilosophy.net/about.html). The premise of this project is to include women thinkers into the philosophical canon of the early modern period, recognizing the role the played in the development of ideas. Projects such as these do not take away from the debt Western society owes to its acknowledged intellectual greats, but rather enriches our understanding of the conversation that has taken place throughout history between a variety of individuals. We encourage you to take a look at this project and consider the views being offered.

While we are certainly straying beyond a simple analysis of our findings, we believe it is important to consider the implications of the data presented to us. This syllabus data does not merely reflect the historical realities of publication and dominance of male intellectual figures. It also reflects the modern reception and treatment of our inherited history. As such, this data can be understood to reflect the historical development and modern reception of the Western Canon.

*Additional Observations*

Analyzing our data through temporal, geographical, and gendered lenses produce considerable insights into the composition of the modern Western Canon. Again, we do not claim to make assertions on what the canon is, but rather make observations based on the data we have presented. The following page will explore a comparative analysis that will push some of these ideas further. In the meantime, it is productive to briefly consider what this analysis does not show.

One noteworthy feature of these methods of analysis is that they do not address what is missing. Based on the curriculum of a ‘Great Books’ program (more on the next page), the following texts are particularly noticeable in their absence from the top 100: Dante’s Divine Comedy (134), Nietzsche’s Beyond Good and Evil (364), Hegel’s Phenomenology of Spirit (535), Cervantes’ Don Quixote (not in top 1000), Luther’s Freedom of a Christian (not in top 1000), Montaigne (not in top 1000), and the aforementioned Bible (not in top 1000). This is just a small sample of what is not present in the top 100 (and sometimes even top 1000). There is no easy way to visualize this information, thus stressing the importance that this type of work through the digital humanities be accompanied by closer analysis and critical thinking. While the timeline analysis can suggest apparent gaps, a grounding in the subject of analysis is necessary to understand the data produced, specifically to describe what is not present. Finally, human analysis is also necessary to understand that the presence of a biology textbook or style guide in the top 100 is an outlier. While its presence as part of the top 100 prescribed texts is notable, it is not particularly useful for canon analysis. All this to say – there are limits to what digital humanities techniques can produce on their own.

In closing this section, it is useful to consider the merits of attempting to use this data to analyze the Western Canon. As stated at the beginning, it is important to acknowledge the limitations of a project and recognize that observations are being offered based on a dataset with its own methodology. After understanding these limitations, this project is an excellent starting point for consideration of modern treatment of the Western Canon. The observations made above reflect some quantifiable realities. We hope this analysis can be a backbone or starting point for some more in depth research conducted in the future.

*Comparative Analysis*

In undertaking a comparative analysis of the obtained data, our analysis focused on two key aspects: author gender and author nationality.

A breakdown by gender binary is flawed, as noted in our process section, but can provide a revealing look at how often texts written by women are prominently featured in syllabi. The imbalance here is quite apparent: texts authored by men are, by a significant margin, more regularly assigned than texts authored by women. A comparative breakdown of the top 25 books from a variety of institutions and countries by gender can, however, tell us which of these bodies are doing well to provide female representation in their literature.

An analysis of nationality and origin of these texts can reveal inherent biases in the academic preferences of our selected institutions and countries. Unsurprisingly, some nations – and the institutions within them – may reinforce a study of their own works above others. Some nations’ texts, however, may be regarded as such seminal works that they cross borders on a consistent basis. Our breakdowns by nationality have revealed where the studied texts will likely originate from based on where one studies.

Each of these breakdowns and visualizations followed a similar process. Working with cleaned datasets of the top 25 books from the London School of Economics, Harvard University, Carleton University and Cambridge University, as well as from Canada, the United States and the United Kingdom, counts of male- and female-authored texts were plotted on bar graphs to show relative differences. These same datasets included a breakdown by nationality which were plotted on bar graphs accordingly, with counts for each country of origin. Graphs were generated using Microsoft Excel.

Finally, we conclude with an investigation into gender and nationality for the general canon obtained from the Open Syllabus Project as it compares to Carleton University’s Humanities degree, or as it has been dubbed, the “Great Books” program.

*Country Comparison*

**Nationality Distribution**

Canada’s top 25 books, by nationality:  
The most compelling observation when it comes to the nationalities of origin of Canada’s top 25 books is that Canada itself does not appear in the data. While both the United Kingdom and the United States engage in a fair amount of “local bias” in their authors’ nationalities, Canada is significantly less indulgent in its own literature.

There are a few viable explanations for these observations. One is undoubtedly historic: Canada, comparatively, does not have the centuries-long history that the United States and United Kingdom. Given that, Canada has had less of an opportunity to contribute to the world’s literature and build its academic reputation. The United Kingdom, the older of the two other nations, represents the largest contribution to Canada’s syllabi, while the United States ranks slightly lower and tied with France and Greece.

Additionally, Canada’s history and cultural development has often been marked as a tapestry of nations rather than a distinct one in itself. Following this narrative, it may be more natural for Canadian scholars to seek a diverse range of voices rather than study their own.

United Kingdom’s top 25 books, by nationality:  
The United Kingdom shows a clear propensity for its own literature. While other nations, from Canada to Palestine, make brief appearances in the country’s top 25, the overwhelming portion of the list is dedicated to the United Kingdom’s own authors. The United States comes in a far-off second place with a few entries on the list.

As explained in the Canadian section above, the United Kingdom’s long history lends itself to a rich source of literature for study. These data show is that the United Kingdom’s institutions do not, perhaps, have a strong intention to study the works of international authors.

United States’ top 25 books, by nationality:  
The United States shows a similar preference to the United Kingdom in its own works, though it also shows an appreciation for other nations’ authors. Greece and the United Kingdom both rank well on the United States’ syllabi, demonstrating a clear appreciation in the US for western thinkers.

As opposed to Canada, the United States has the lowest distribution of author nationalities, with only six different nations comprising its top 25. The US shows a great propensity for the works of certain nations, but perhaps little regard for others.

**Gender Distribution**

Canada’s top 25 books, by gender:  
United Kingdom’s top 25 books, by gender:  
United States’ top 25 books, by gender:

A direct comparison of the three examined nations shows the United Kingdom has a higher proportion of its top 25 books written by women than either the United States or Canada. Correlating this data with the analysis of nationalities, whereby the United Kingdom had an overwhelming majority of its top works coming from the UK itself, one could infer that the United Kingdom itself has a more significant number of well-regarded female authors. This observation is speculation, but is a possible explanation for the trend.

The overall observation from this data is that each of the countries demonstrates poor representation of women authors. These are certainly in-part the consequences of a historic oppression of women in academia. The question today is whether nations and their institutions are doing enough to bring women to the forefront of scholarly conversations.

*Institution Comparison*

**Nationality Distribution**

Cambridge University’s top 25 books, by nationality:  
The United Kingdom’s Cambridge University yielded some of the more interesting results in terms of text nationalities. Whereas the other universities on this list demonstrate a bias towards authors from their own country, Cambridge breaks this trend with most of its top texts coming from the United States and the United Kingdom coming in second. Why this is the case is not readily clear. Cambridge, one of the world’s oldest surviving universities, has perhaps in recent decades made it a priority to expand its literature to include more of a North American scope, with both Canada and the United States receiving fair representation on its list.

Cambridge also demonstrates the least variety of nationalities throughout its top 25, with only five countries comprising the list. Whether or not this has been done on a conscious basis remains to be seen, but data from Cambridge seems to suggest a preference of authors of certain nationalities.

Carleton University’s top 25 books, by nationality:  
Carleton University, one of the few available Canadian universities in the Open Syllabus Project’s beta phase, demonstrates a distinct tendency towards Canadian authors. Carleton, however, also demonstrates perhaps the best distribution of authors’ nationalities of the four institutions we have examined. While Canada ranks highest, a total of nine nationalities are represented in Carleton’s top 25, with the United Kingdom, the United States, France and Ireland making multiple appearances throughout.

To relate these observations to our analysis of Canada’s nationality distribution, it may be a symptom of Canadensis that Carleton scholars experience a diverse range of authors.

Harvard University’s top 25 books, by nationality:  
Harvard University presented the clearest nationality trend in its top 25. Overwhelmingly, Harvard demonstrates a preference for American authors in its syllabi. As one of the United States’ premiere universities, it is perhaps not surprising that Harvard may be nationalistic in its academic disposition. Harvard is undoubtedly an American institution, and the academic experience of students appears geared to reflect that.

London School of Economics’ top 25 books, by nationality:  
The London School of Economics also demonstrates a clear propensity for its home country’s own authors, but to a lesser degree than that of Harvard University. Greek and American authors make an impact on LSE’s top 25 list, but a local bias is quite evident by looking at the chart.

There are numerous general takeaways from this collection of data. While limited, these cross-sections of universities demonstrate a general bias towards authors’ from the institution’s home country. While this may not come as a surprise, it is also not, as we have seen in the case of Cambridge, a sweeping generalization. There are exceptions to this rule, though the reason for those discrepancies would require further investigation.

Additionally, it is clear that the location of a university affects, but is rarely an exact reflection of, what students are assigned. Contrasting country data and institution data reveals discrepancies in this regard: Carleton University does have a wide distribution of authors’ nationalities as does Canada, but features a significant number of Canadian authors in its top 25 where the country has none. Harvard also over represents its assignment of American authors as compared to the country itself.

**Gender Distribution**

Cambridge University’s top 25 books, by gender:  
Carleton University’s top 25 books, by gender:  
Harvard University’s top 25 books, by gender:  
London School of Economics’ top 25 books, by gender:

We can draw similar conclusions in regards to author gender for institutions as we did for countries. None of the schools we have examined did well in representing female authors, though Carleton University had the best showing and Cambridge University had the poorest.

Though this analysis would benefit from an examination of a greater variety of institutions, we have not observed any institution actively taking upon itself the task of providing a greater representation of female authors.

*Canon vs Great Books Comparison*

As a final extension to our comparative analysis, we have decided to contrast the Canon with a Great Books program. This section was originally planned out of a sense of personal indulgence, but we soon saw value to the comparison (disclaimer: one of the authors of this project is enrolled in this program at the time of this project’s composition). As such, this section will contrast our dataset with Carleton University’s [Humanities program](http://carleton.ca/bhum/).

We created a Great Books dataset manually based on the syllabi from throughout the course and a personal knowledge of the texts studied within anthologies. Based on this information, we were able to assemble a new dataset and add in similar extending features such as gender, nationality, and dates. As we were looking at a sample of a single program or tailored set of courses that did not tend to repeat text usage, we did not have a ‘Count’ feature like the other data. Finally, we compared our new dataset to the top 1000 syllabi dataset, transposed the rankings, and produced our [final version](https://github.com/prefaces/Final-Project/blob/master/Humanities-dataset-cleaned-final.csv) of the Great Books dataset cross-referenced to the rankings of [The Open Syllabus Project](http://opensyllabusproject.org/). We were able to make the following comparisons.

NATIONALITY

The nationality comparison between the two datasets shows considerable difference as seen below.

General Nationality:  
Great Books Nationality:

Part of this difference can be attributable to the intentionality behind the curriculum of the Great Books program to showcase non-western sources. While part of the variety of sources for the Great Books program is attributable to the different sample size of the comparison (this is due to the nature of the data), a couple elements jump out to support the conclusions drawn. For one, the strong presence of countries such as India and Spain (absent from all other datasets used), highlights the Great Book program’s interest in pushing beyond traditionally ‘Western’ sources. Furthermore, the proportionally smaller influence of the United States and Great Britain also point to a greater diversity of sources. From these findings, we observe that the General dataset cannot necessarily make the same claims to representing the Western Canon as a Great Books program. Again, this is merely an observation and not particularly definite, but significant discrepancies between the nationalities of authors is indicative of significant difference.

GENDER

Unlike the previous comparison, a comparison of gender usage between the two datasets does not differ significantly, as seen below.

General Gender:  
Great Books Gender:

If anything, the proportional disparity between the General dataset is less significant than that of the Great Books dataset. We will not dwell long on these findings other than to say that this Great Books program represents the Western Canon’s tendency to not include a high percentage of female representation in its curriculum. This is not to speak to the substance of what is taught in this particular Great Books program, but rather to make this assertion based on the syllabus. This finding points to both the strength and weakness of using digital history techniques – trends can be identified but a qualitative analysis is typically lacking.

REFLECTION

This comparison is useful to acknowledge the difference between a program purporting to be a Great Books program and treating the syllabus collection from largely Western universities as Canon. A great deal more work could be done on the quality of a Great Books program to determine whether or not its curriculum represents the Western Canon. Again, this points to the largely macro-level nature of this project as none of this analysis takes a close look at any of its subjects. Nonetheless, some fascinating observations are made and we hope this can be a platform on which future research is conducted. We have chosen to conclude this section with another Reingold-Tilford Tree visualization from the vector graphic software [RAW](http://raw.densitydesign.org/) to visualize the Great Books dataset.

Reingold-Tilford Tree Visualization: Great Books  
Spainde VegaCervantesGreeceAristotlePlatoSophoclesHomerHesiodEuripidesAristophanesAeschylusUnited StatesWrightKripalHamilton Madison and JayStunkFranklinSaudi ArabiaIsraelPalestineEusebiusSaidIndiaRushdieJayadevaGandhiIraqGermanyWulfSchillerNietzscheLutherKarlstadtHeideggerGoetheMarxKantHegelArendtItalyWeissTassoPetrarchAriostoMachiavelliVirgilOvidThomas AquinasDanteBoccaccioBoethiusCanadaTaylorRuvinskyGrantUnited KingdomSidneyMaloryKnoxHumeHerbertAddison and SteeleBarbonShellyHobbesMiltonSmithDefoeLockeShakespeareFranceRousseau Jean-JacquesVoltaireFoucaultRabelaisEgyptPlotinusChinaLao TzuConfuciusNetherlandsErasmusAustraliaBullLibyaAugustineIrelandBurke

*Concluding Observations*

This comparative approach has generated a few general observations about the nature of academia between countries and institutions.

First, in regards to author nationality, countries and their institutions have clear and perhaps unsurprising propensities for authors of their own culture. With varying degrees and with some exceptions, institutions tend to assign a high proportion of authors from their own country’s history, providing a local bias to the nature of academia in most countries we examined. Canada appears to be an exception to this rule, though Carleton University in Ottawa does reflect these findings to some degree.

Across the board, authors from the United States, the United Kingdom, and to a lesser–degree, Greece, all make regular appearances on top 25 lists. This perhaps indicates a general or historical recognition of the quality of these nations’ authors in the eyes of academic institutions.

In terms of gender, our analysis indicates that no institution or nation is doing well in terms of representation of female authors. While this is no doubt an unfortunate consequence of historic oppression of women, we would argue that more can be done at this time to improve modern representation of women in academia.

REFLECTION

*How Much Can We Learn from a Syllabus?*

The very premise of this project hangs on this question to which there is not a simple answer. As was stressed by [The Open Syllabus Project](http://opensyllabusproject.org/faq/) the purpose of this exercise is not to evaluate quality of teaching or to make claims that certain texts are inherently superior to others. This is an important observation if for no other purpose than to emphasize the importance of limiting assertions. Following the same logic, our project does not try to make objective assumptions about the superiority of texts or what belongs in the Western Canon (though we do have our own subjective opinions on what belongs). Instead, we attempt to make observations based on the syllabus data collected. This returns us to the question posed at the outset.

We can learn a great deal from a syllabus, but there are certain limitations to what we can learn. Syllabus data can point to both the historical construction of a Western Canon and to modern pedagogical reception of said canon. In this sense, we are not merely limited to asking historical questions such as what factors contributed to a great proliferation of texts during a particular period. We can examine how this information is received and prioritized today. Additional information such as text pairings would go a long way toward understanding how these texts are treated in academic courses. Based on the data gathered in our datasets, we are left to make assumptions and observations based on the simple indicator of frequency in which a text appears on various collections of syllabi.

There is no easy answer to the question posed at the start of this section. As a sort of contract between instructor and student, a syllabus should purport to represent the material taught in a class. Required or recommended literature accompanying these objectives can speak to the direction of a course and the priorities of an instructor. As such, a syllabi can tell you a great deal. With the data in its current form, we believe we have pushed our observations as far as they can go. Additional information could open a wealth of potential explorations. For the time being, we will content ourselves with making observations from our data based on historical development of the Western Canon and modern treatment of these sources.

*Most Interesting Findings*

Our research, while productive, was generally not surprising. We knew, for example, that heading into this project we would find a large gap between works from the Greek philosophers of old and the thinkers of the renaissance and beyond. And yet, seeing that gap on our timelines was profoundly interesting.

It has been revealing to see such clustered (and often distant) periods of time when a great deal of revolutionary works were published. Three 20-year time periods ended up being concentrated moments of scholarly work: Plato and Aristotle’s era of 428-324 BCE; the impressions left by Marx and Darwin in 1797-1819 CE; modern authors like Foucault and Achebe writing from 1922-1946 CE. Reflecting on these moments of historic importance shows that we may not necessarily recognize authors’ contributions on a consistent basis, but rather in bursts and in context of their eras.

We found few surprises in our comparisons of institutions and countries. Countries and universities tend to prefer their own authors, with a few notable exceptions. The representation of women in syllabi across the world is lacking, and as of yet no university or country in our study has successfully made it their goal to work towards adequate female representation.

Ultimately, it is not what we have discovered that brings value to this exercise: it is what we have shown. The gaps in history and in representation are most notable in this project, and clearly illustrate the current nature of academia and the conscious or unconscious value we set to periods of history and the authors we choose to study.

*Strengths and Weaknesses of Project*

The content of this heading has been addressed a couple times throughout this project, but it bears repeating. The strengths and weaknesses of this project are largely linked to the field of digital history (or digital humanities). This project is particularly successful at taking large datasets and manipulating them for visual presentation. It is further successful at studying this data, exploring its implications, and making observations based on what the data includes. Finally, and forgive us for being immodest, this project is particularly successful in bringing this data to life, while clearly describing our process in the interest of clarity, replication, and even to introduce new students to the digital humanities. As a whole we believe the project was successful in achieving what it set out to accomplish, and we are pleased with our final results.

There are weaknesses to the model we have employed and we have tried to identify them throughout this project. These weaknesses include the use of second-hand data from another project in its beta stage, the limitations behind the assumptions we can make on the basis of the data we have, and the lack of depth to the project. This project was always dependent on outside sources of data and we were handcuffed according to the form in which it was presented (that being said, our data source was excellent). We discuss throughout the project that this sort of study can only make observations and is limited in the assumptions that can be made. Finally, this project treated its data on a macro level, using only prior knowledge and personal academic study to interpret. Additional iterations of this project or future research using it as a source would benefit greatly from a closer analysis of the texts or time periods in question. We recognize the importance of clearly defining the scope and merits of this project. We hope these limitations have been well communicated to our reader.

*Digital History*

The nature of this project reflected the strengths of engaging in digital history. We owe a great deal to the work of the Open Syllabus Project for undertaking their work on this ever-developing database. It is thanks to them that we were able to access this trove of information and manipulate the data to show the trends we’ve discussed here.

In working with our data, the skills we have gained throughout our studies in digital history have been invaluable. The open source nature of GitHub has made a two-man job flow seamlessly. From cleaning messy data with regex functions to the visualisation tools we’ve used to demonstrate our findings, digital history has turned what could have been a lifeless essay to a dynamic and interactive exploration.

Our thanks to Dr. Shawn Graham for his instruction and consultation on our work as we sought to learn more about this distinct field.

*Concluding Remarks*

This work has been carefully put together but is by no means exhaustive. The Open Syllabus Project continues to evolve as it works through its beta phase. Many of our findings work from selective datasets of countries and institutions, and further exploration of schools and other filters of the database can only add more valuable to the exercise. If you are inspired by the work done here, we urge you to consider expanding it yourself. Our data is available on [GitHub](https://github.com/prefaces/Final-Project) and is, of course, open-source.

In short, thank you for exploring the Open Syllabus Project with us. Our findings and their presentation has been a time-consuming and rewarding effort. We hope you have gained something from the time spent pouring over a few thousand syllabi.

REFERENCES

*Data Sources*

Primary data was copied from filtered searches of the [Open Syllabus Project](http://opensyllabusproject.org/).  
Supplementary data extracted from [Great Books](http://opensyllabusexplorers.thenewhistorian.com/http:/sonic.net/~rteeter/greatbks.html), by Bloom.

*Tools Used*

Data committed to [GitHub repository](https://github.com/prefaces/Final-Project/).  
Data cleaned through regex functions in [Notepad++](http://notepad-plus-plus.org/).  
Data cleaned and charted with [Microsoft Excel](https://products.office.com/en-ca/excel).  
Interactive timeline developed with [TimeGlider](http://timeglider.com/).  
Mapping visualization with [Palladio](http://palladio.designhumanities.org/#/).  
[RAW](http://http/raw.densitydesign.org/) used for data visualization.  
Website on [WordPress](http://wordpress.com/).

About

Open Syllabus Explorers is a project by Benjamin Doyle and Craig Lord. It was designed to meet the requirements of an undergraduate university level course in the digital humanities at Carleton University. This project was supervised by Professor [Shawn Graham](http://electricarchaeology.ca/).

*Benjamin Doyle*

Benjamin Doyle is a fourth-year Bachelor of Humanities candidate at Carleton University in Ottawa, where he is pursuing a combined honours degree in History. Following graduation in June 2016, he will be starting a graduate degree in Public Policy at the School of Public Policy and Governance at the University of Toronto. He has found Open Syllabus Explorers to be a rewarding research project that has continually challenged him. Benjamin’s online profile can be found at [Prefaces](http://prefaces.ca/).

*Craig Lord*

Craig Lord is a fourth-year Bachelor of Journalism candidate at Carleton University in Ottawa. Though his primary studies revolve around modern storytelling and reporting, his minors in political science and history have reflected his academic interests while at university. Open Syllabus Explorers has been a challenging and fun endeavour to cap off his academic career by reflecting on academia around the world. After graduation, Craig will be pursuing a career in journalism including work this upcoming summer with the Globe and Mail in Toronto. Craig’s online profiles can be found at [The New Historian](http://thenewhistorian.com/) and his [personal portfolio](http://craiglord.ca/).