Visualizing University Students Collected & Missing Demographics Data

Rahel Gunaratne*
Carleton University

Gananatha Subrahmanyam†
University of Ottawa

Fateme Rajabiyazdi‡
Carleton University

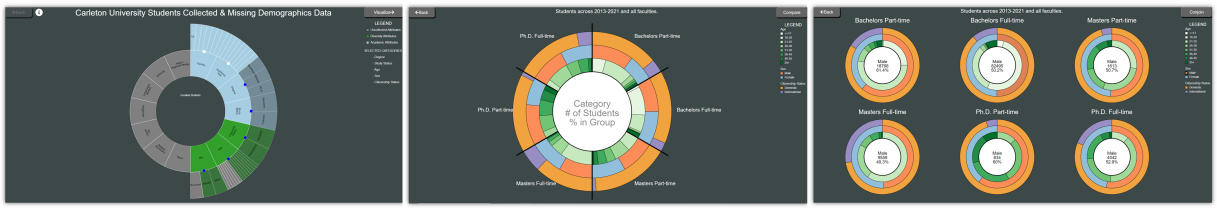

Figure 1: The designs on a web-based platform using Javascript and D3 library. The leftmost column represents the attribute selection visualization as a sunburst. As depicted on the legend, the green outlined nodes represent the diversity attributes. The blue outlined nodes represent the academic attributes. The grey filled nodes represent the uncollected attributes. The darkened nodes depict the selected attributes. The middle column represents the selected attributes visualization as a ring diagram. Each slice of the diagram represents every combination of the selected academic attributes. The slices contain all selected diversity attributes with each demographic represented as arcs. Each arc represents the ratio of the selected demographic to the total number of students in that respective category. The right column represents each slice as a ring diagram and hovering over an arc displays the corresponding population for each demographic.

Abstract

We present an interactive data visualization design representing student demographics and academic information. Our goal is to demonstrate the relationships between the diversity-related attributes collected as part of demographics information and the academic attributes. We first collected university student data from our local university. In collaboration with a researcher studying the impact of diversity, equity and inclusion, we chose relevant data attributes from our dataset. Next, we sketched as many designs as we found relevant representing our dataset and shared them with our collaborator. As a group, we selected the designs that we saw value in implementing. Lastly, we developed an interactive data visualization on the web. With our design, we draw attention to the lack of demographics data collected that describe student diversity. By showcasing the potential missing demographic data attributes, we hope to raise awareness about the importance of collecting this data in universities.

Index Terms: Human-centered computing—Visualization— Visualization application domains—Information visualization

1 Introduction

Demand for social change has been a recurring theme in human history. They brought about social policy changes in order to protect individuals and groups from discrimination on the basis of gender, dis/ability, language, religion, age, sexuality, and race. For example, the number of women in the workplace has shot up over the past century. It was the result of The Fair Labor Standard Acts, Equal Pay Act, Civil Rights act, Pregnancy Discrimination Act and The Family and Medical Leave Act. These claims for social justice often rely on collecting data on identity markers to support the call to action. Social justice issues have particular relevance in higher education and its related opportunities [6]. Thus, to increase awareness and diversify access to higher education, nowadays, many universities and colleges collect and publicly share student demographics information.

Carleton University also collects student demographics information, including citizenship status, sex, and age, in addition to student academic information such as academic year, faculties, degrees, and study status (part time, full time or co-op). These data were publicly available online on Carleton's website (Carleton Original Data Cubes (CODC)). The CODC visualizes student information in the form of line charts for all combinations of attributes. However, it does not provide an easy way for cross comparison between attributes in order to find patterns, trends, and anomalies.

Data is used to represent the real world. With a transfer of the continuous, physical world to a discrete digital world sometimes the integrity of the data is compromised as it may get lost in translation. In some cases, the data simply does not even exist. Missing data makes it challenging to solve issues [7]; how can a problem be fixed if we are not aware that a problem even exists? Many universities and colleges collect data but often miss collecting important attributes. Similarly, the dataset we had access to, the CODC, does not provide any information on race, indigeneity, dis/ability, religion or spirituality, first language, other languages, ethnicity and nation of origin. These attributes could potentially reveal patterns in terms of diversity, equity, inclusion, accessibility and academic performance that are highly important for universities to consider as they work to effect policy changes in relation to diversity, equity and inclusion.

Our primary objective was to design an interactive visualization on a web-based platform displaying the relationship between student diversity-related demographics and academic attributes. Our secondary objective was to demonstrate the lack of demographic diversity data attributes collected by the university. We designed an interactive data visualization displaying relevant data attributes and demonstrating the student demographics diversity and academic data collected by Carleton University (See Fig. 1). We used a sunburst diagram and a ring diagram to provide a complete view of the information. Finally, we incorporated detail-on-demand to improve accessibility and readability.

2 University Student Demographics and Academic Attributes Data

The CODC obtains the data from the Banner Enterprise Resource Planning (ERP) System. ERP is the primary system of record for Carleton University and provides Carleton's faculty and staff with access to financial, human resources and student information.

*e-mail: RahelGunaratne@cmail.carleton.ca
†e-mail: gananatha.subrahmanyam@uottawa.ca
‡e-mail: fateme.rajabiyazdi@carleton.ca

The data includes a variety of academic attributes: degree indicates the level of study (i.e., Bachelors, Masters and PhD), faculty (i.e., Engineering & Design, Science, Public Affairs, Business and Arts & Social Sciences), academic year (starting from 2013), and the study status of the student (i.e., part time, full time and co-op).

The demographics diversity attributes include student ages from 17 to 25, then 5 year age ranges from 26 to 35, then 10 year age ranges from 36 to 55 and then a blanket range for all students older than 55. The data contains information about student citizenship status denoting if a student is domestic or international. The dataset includes information on student biological sex (i.e., male, female).

The dataset also contains information about student convocations. The convocations data count the number of students graduated separated by sex, degree, faculty, and academic year.

3 Related Work

3.1 Existing Diversity Visualizations

Standard bar charts support finding trends, patterns and anomalies. Each bar represents one demographic with the height of the bar demonstrating the population [4]. But, they only function well with a small number of attributes. It cannot communicate the variety and richness of diversity. Many bar charts in a grid structure combats this problem. The Carleton Data Cubes used this approach before it was taken down.

Another approach is the Diversity Map [8]. The design gives an overall impression of diversity. Each rectangle corresponds to an object's attribute values. Each vertical stacking of the rectangles represents each attribute. The opacity of the rectangle depends on the proportion of population. This approach is powerful and it provides clear insight into the diversity of a particular population. But, this project requires visualizing diversity for many different populations. It would require a grid structure of many Diversity Maps.

3.2 Diversity Visualization Etiquette

Representing people with data requires careful consideration of many factors. A summary of this concept is "if I were a datapoint on this graph, would I be offended?" There are many principles to help guide us toward equity awareness [1]. Making sure to use people-first language, for e.g. "people with disabilities" instead of "disabled people". The purpose is to refer to people and not their labels. Another principle is ordering data in purposeful ways. Many national surveys list white people as the first option. This may convey the message that the first group is more important. Many solutions exist including if a study focusses on a group, then that group should be first. Other solutions include randomized order or user selected order. Finally, it is important to avoid color palettes that reinforce stereotypes.

Another useful principle is disclosing metadata in visualization [5]. There are risks and benefits for each type of metadata. Providing the data source allows replicability. But it may have negative impacts on privacy and trust if the data associates with a particular individual or organization. Describing the visual encoding challenges may help users understand the pitfalls. But, it may cause them to decide the visualization is irredeemable. Providing short biographies of the creators may lend credibility and increase trust. But, it may backfire if the user deems the creator to be unsuitable. Finally, providing the intended audience may allow readers to get a better understanding of design choices. But it may alienate some readers if they are not part of the intended audience.

All the principles described are relevant to this project. The design section covers specific ways of incorporating them.

4 Methods

We used Design Study Light Methodology (DSLM) [9] to design an interactive data visualization representing university student demographics and academic datasets.

First, we explored the dataset and discussed the attributes collected and missing in this dataset with our collaborator (a researcher in the social sciences with more than ten years experience in the field of diversity studies in the Canadian and international contexts, as well as a previous background in government advertising and public opinion research). Next, we consulted with our collaborator to discuss the attributes collected from our datasets and selected attributes, relevant to current times, for the goal of our visualization.

We sketched a series of preliminary visualization designs displaying the selected attributes (See Fig. 2). We shared our sketches with our collaborator and selected the sketches that best fit our purpose as a group. Then, we developed medium-fidelity visualization design prototypes. We shared our prototypes with our collaborator and obtained feedback and improved our designs in an iterative process. Lastly, we developed an interactive web-based data visualization tool using JavaScript and the D3 library (See Fig. 9).

5 Visualization

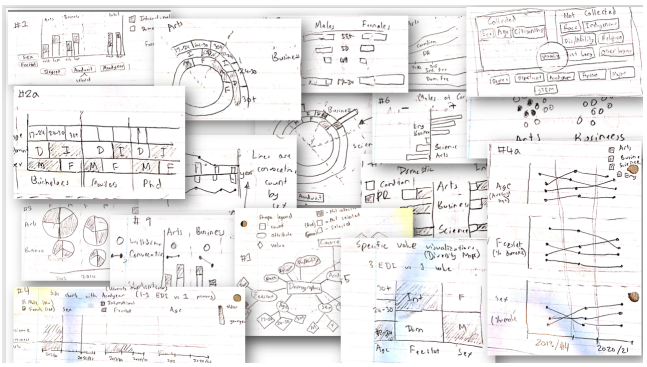

Figure 2: Collage of low fidelity potential designs using pencil and paper

5.1 Data Selection Process

As a group, with our collaborator, we selected several attributes that showed relevant information to diversity, equity and inclusion. The CODC dataset contains a large number of attributes. Since our goal was to showcase diversity attributes and their relationships with academic attributes, we focused on data that had both attributes. Our dataset had other attributes, such as the courses report, which includes the number of individuals withdrawn from a course, but it did not include any diversity attributes, so they were deemed irrelevant to this project.

In our dataset, the representation of the sex attribute is ambiguous because 'Male' and 'Female' are biological terms, but the forms provided to students to update this attribute are labelled as Gender Assignment. Due to this lack of coherency in use of terminology and the fact that it is not appropriate to ask for a person's biological sex in an academic setting, students, when filling out this attribute, may have been referring to either their sex or their gender and the database may reflect that ambiguity.

Additionally, due to the importance of being able to illustrate the relationship between diversity attributes and academic performance in science, technology, engineering, and mathematics (STEM) fields, we added a STEM and non-STEM branch in our design. Since two faculties (Faculty of Engineering and Design, Faculty of Science) in our university offer programs in STEM, we included them as categories for faculties.

5.2 Design Process

Upon selecting the attributes related to diversity and academic attributes, we started the design process.

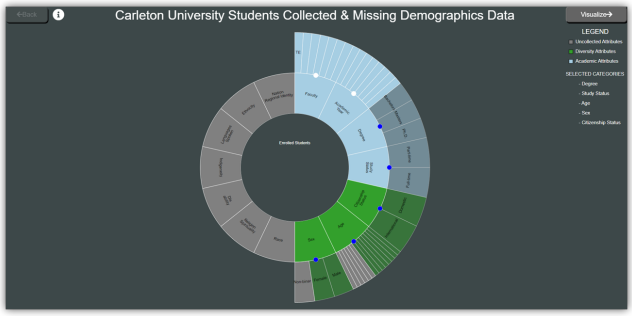

Figure 3: Visualization of all attributes in a sunburst diagram. The darkened slices indicate selected values.

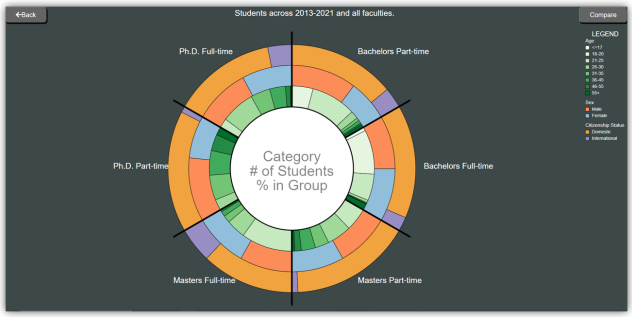

Figure 4: Visualization of all selected attributes in a ring diagram.

Our design has two main components: 1) selection of the diversity and academic attributes (See Fig. 3) and 2) visualizing the selected attributes (See Fig. 4).

The sunburst diagram was chosen based on the following design requirements. It displays all the possible attributes from a hierarchical data set.In addition, the diagram highlights the proportion of relevant uncollected data to collected data with color.

This diagram also supports interactivity to provide more detail if they desire.

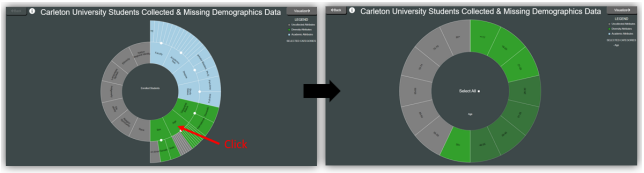

Figure 5: Clicking the Age slice changes the visualization to show all the categories in a zoomed in format.

Hovering over a node in the attribute selection visualization reveals more information about the attribute (See Fig. 6).

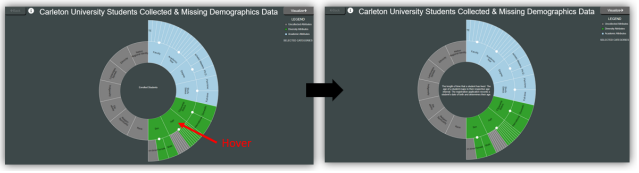

Figure 6: Hovering over the Age slice displays a description and context of the attribute in the center.

Compared with standards promoted by the government [3], the dataset is lacking many diversity attributes that could be collected. Since one of our goals was to communicate the lack of available diversity attributes, we displayed data categories that are currently unavailable in the dataset as greyed-out slices in the sunburst diagram (See Fig. 3).

These identity categories include, but are not limited to, race, indigeneity (with specific categories of First Nations, Métis and Inuit), dis/ability, religion or spirituality, first language, other languages, ethnicity and nation of origin. In addition to the unavailable identity categories, values within the categories such as age intervals beyond 55+ and all genders are also listed as uncollected. We chose to show unavailable data to raise awareness about potential data attributes that could be collected.

The novelty of the ring diagram design is that can visualize any combination of data (See Fig. 4). The selected academic attributes are depicted as slices. The slices contain every selected diversity attribute with each demographic represented as arcs. Each arc represents the ratio of the diversity value to the total number of students in the selected category. Hovering over an arc displays the count and percentage in the center circle (See Fig. 7).

Figure 7: Hovering over the arc of domestic students in 2013/14 Arts & Social Sciences Full-time displays a percentage of 96% (out of domestic and international) and a count of 5811 domestic students in the center.

Clicking on an arc transitions the visualization to display diversity of the demographic (See Fig. 8).

Figure 8: Clicking on the arc of domestic students in 2013/14 Arts & Social Sciences Full-time transitions the visualization to display the other diversity values (age, sex) of students in that demographic.

The compare mode was inspired by the related works. The grid format allows comparison of diversity between academic demographics (See Fig. 10).

We used colors that are color-blind friendly to people with Deuteranopia to enhance accessibility of our design. The palettes were sampled from the website ColorBrewer [2].

To encourage racial equity awareness, it is important not to reinforce any stereotypes. We randomized the colors within sets of categories. Also, the selection order determines the order of the displayed arcs.

Our design supports scaling. If in the future, more attributes and categories are collected, then all one needs to do is update the preliminary data file and the design will be updated accordingly. The additional attributes will be displayed as slices in the sunburst diagram visualization. In the ring visualization, since the diversity attributes correspond to the number of rings, there will be additional rings displayed.

6 Discussion

Institutions must understand the difference between sex and gender and devise appropriate ways for asking individuals to provide information on their gender when collecting demographics information instead of biological sex. They must clearly specify that they are asking for gender and should provide appropriate categories such

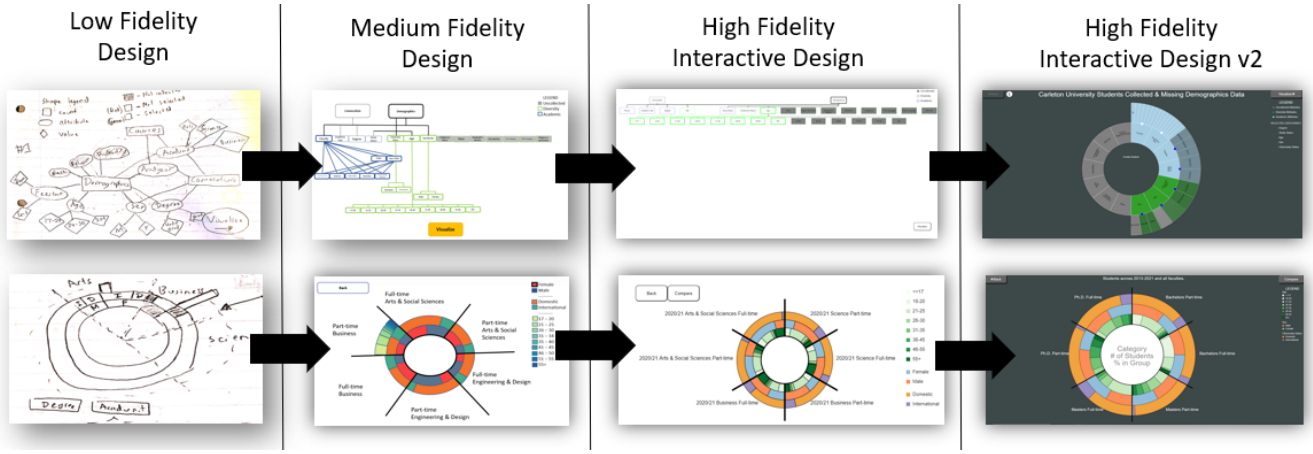

Figure 9: Four iterations of designing student demographics data visualization. The visualization design has two main components: selection of attributes to visualize (top row) and the visualization displaying student demographics (bottom row). The first column shows the low fidelity designs drawn as sketches on paper. The second column shows the medium-fidelity designs with colors and potential interactions. The third column represents the ongoing development of the designs on a web-based platform using Javascript and D3 library. The fourth column shows the change in design in attribute selection and many more features.

Figure 10: The conjoined form is pictured on the left. Clicking the compare button transitions the visualization to what is seen on the right. Several ring diagrams represents the slices. Hovering over an arc displays each demographics respective ratios and values.

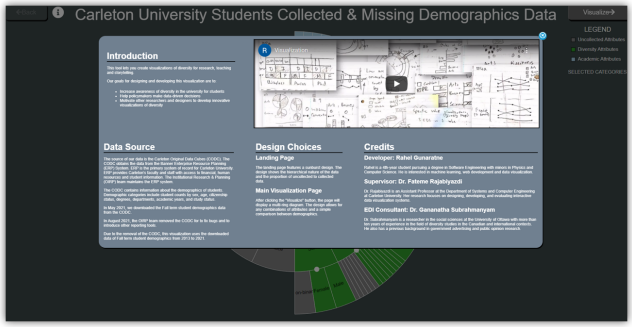

Figure 11: We provided an information section that provides meta-data on the visualization in the form of a pop-up. An introduction, data source, design choices and credit section is provided. A video describing the project is also available.

as male, female, transgender, gender neutral, non-binary, agender, pangender, genderqueer, two-spirit, third gender, and all, none or a combination of these possible genders.

Although the diversity categories listed in this report may be appropriate to collect now, it may not be appropriate in the future. Language and standards change over time. As such, organizations and institutions may need to revisit their collected categories to ensure they stay up to date.

It should be noted that blind data collected could pose risks. If the data were so fine-grained, it could lead to discriminatory practices or media attacks. A possible solution is to display 'less than 5' instead of any specific numbers below that to prevent individuals from being identified.

The frequency of collection may also have an effect on the data. For example, if data for students is only collected in first year, then that data may become outdated in subsequent years.

As stated in the results, there are a myriad of missing attributes and attribute categories. If this data were collected, there would be a better understanding of the demographics of the students.

The dataset, used in this project, is missing viable diversity attributes; however, this dataset is not the only one that is lacking in this way. For example, Ryerson University has a webpage dedicated to visualizing diversity data of students and faculty [4]. They include several diversity attributes (i.e., women, aboriginal, disabilities, 2SLGBTQ+, South Asian, Black and Chinese), but there is a need to include additional attributes that could reveal significant information about how to increase diversity in the universities.

We would like to make a call to universities to be aware of the importance of collecting these diversity attributes. Fair representation is a stepping stone for a better society and that is only achieved by first understanding the obstacles to diversity, equity and inclusion. In the hope that more attributes will be collected to increase awareness and promote Equity, Diversity and Inclusion, we designed our visualization to be scalable as more data is collected.

7 Conclusion and Future Work

In this paper, we present an interactive scalable data visualization representing university students' demographics and academic data. We used the dataset collected in our local university and selected relevant attributes. In an iterative process, with our collaborator, we ideated, sketched, and designed data visualization prototypes. We implemented an interactive data visualization that best represented the relationship between students' diversity and academic attributes.

During this design study, we got in touch with our university administrators in charge of student data collection and raised our concerns on inaccuracy and missing demographics data attributes with the university. In future, we plan to demo the visualization to university administrators and get their input and incorporate it into the next iteration of the design. Lastly, we plan to disseminate the design and make it available to the public and evaluate their reactions.

Looking to the future, we hope that our design study motivates other researchers and designers to design and develop innovative visualizations demonstrating diversity attributes in demographics

data. We hope that with the power of visualizing existing and missing diversity related demographic data attributes, we can draw more attention to the value of diversity, equity and inclusion within institutions of higher education and beyond.

ACKNOWLEDGMENTS

The authors wish to thank James Moretton (Assistant Registrar (Carleton Central Academic Records)) and Nathasha Macdonald (Assistant Vice-President (Carleton Institutional Research and Planning). This work was supported in part by Internship-Carleton University Research Experience for Undergraduate Students program and NSERC Discovery grant.

REFERENCES

[1] Applying racial equity awareness in data visualization.
[2] Colorbrewer 2.0, colorblindness pallettes.
[3] Diversity and inclusion in the public service.
[4] Diversity in ryerson university.
[5] A. Burns, T. On, C. Lee, R. Shapiro, and C. Xiong. Making the invisible visible: Risks and benefits of disclosing metadata in visualization.
[6] S. R. Gordon, P. Elmore-Sanders, and D. R. Gordon. Everyday practices of social justice: Examples and suggestions for administrators and practitioners in higher education. *Journal of Critical Thought and Praxis*, 6(1), 2017.
[7] M. Onuoha. Missing data sets.
[8] T. Pham, R. Hess, C. Ju, E. Zhang, and R. Metoyer. Visualization of diversity in large multivariate data sets, 2010.
[9] U. H. Syeda, P. Murali, L. Roe, B. Berkey, and M. Borkin. Design study "lite" methodology: Expediting design studies and enabling the synergy of visualization pedagogy and social good, Feb 2020. doi: 10.31219/osf.io/mghj3