Table of Contents

Information visualization

In this report we will describe how we implemented a web application for visualizing a large data set. This data set contains information about graduates on Dutch universities. Since the data set is very large, a tabular view is not very helpful when trying to see patterns in the data. Therefore we had to come up with a better interface, to make viewing and understanding the data easier.

In section 1, we will first give a description of the format of the data set. Then, in section 2, we will explain our design considerations for the interface. We use a preprocessing step to make working with the data easier; we explain this step in section 3. In section 4, we present our actual implementation; we also show screenshots there. Finally in section 5, we show for what questions our visualization can be useful.

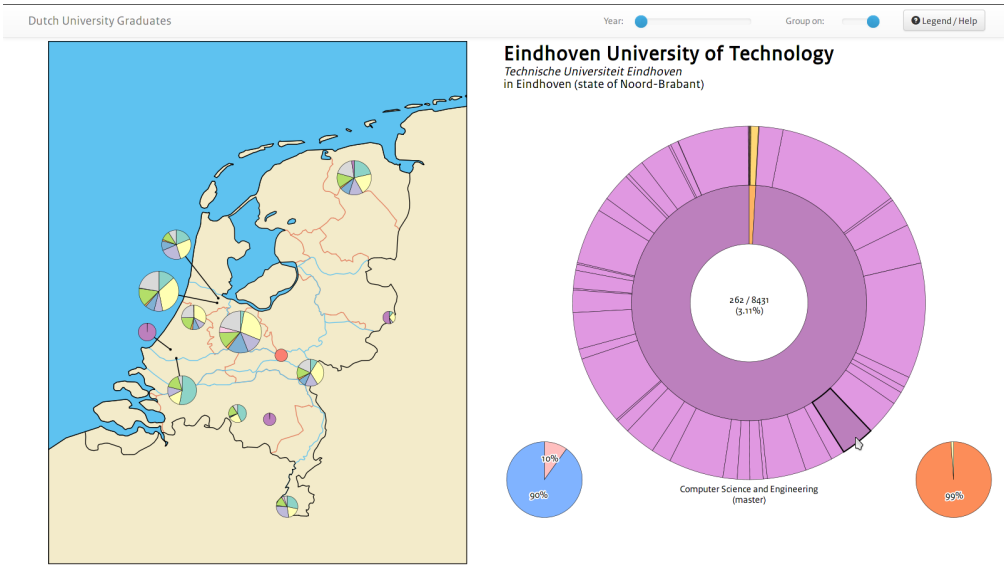

Figure 1: Overview of our application, showing two modes next to each other. Here, the Eindhoven University of Technology is selected and data is shown about the *Computer Science and Engineering (master)* in particular.

1 Description of the data set

The data set we want to visualize, is created by the Dutch organization *Dienst Uitvoering Onderwijs* (*DUO*), that is responsible for the administration of students in the Dutch higher education system. DUO makes available many data sets every year about education in the Netherlands. These data sets can be obtained from their website[1].

1.1 High-level description

Our data set is called "*Aantal wo gediplomeerden*" (*Number of university graduates*) and describes the number of graduates on Dutch universities between 2007 and 2011 (inclusive). The data is very detailed. The number of graduates is not only given for every university, but these counts are split up on

- the graduation year (2007 – 2011),
- the study,
- whether the graduate was male or female, and
- whether the graduate was a full-time, part-time or dual student.

So, for example, we can see in this data set that in 2009, on the Eindhoven University of Technology, on the master *M Computer Science and Engineering*, 54 males and 4 females graduated that were studying full-time, and 2 males and 0 females graduated that were studying part-time.

Apart from the graduate counts, quite some data about the studies themselves is given. First of all, we can see the study phase. This indicates which 'level' the study is. For example, this can be "bachelor" or "master". There are two other values in the data, namely "initiële opleiding" (that is, a program from before the bachelor / master system was introduced) and "beroepsfase" (this only occurs for health-care studies, and indicates a clerkship at a hospital). We will call those two categories *initial* and *clerkship* in the application.

Furthermore, the studies are categorized in so-called "CROHO categories". *CROHO* is a system for the registration of study programs in the Netherlands, and for this system all studies are classified as one out of ten categories. The possible categories are:

- "economie" (economy);
- "gedrag en maatschappij" (behaviour and society);
- "gezondheidszorg" (health care);
- "landbouw en natuurlijke omgeving" (agriculture and natural environment);
- "natuur" (nature);

[1] `http://www.duo.nl/organisatie/open_onderwijsdata/databestanden/default.asp`

- “onderwijs” (education);
- “recht” (law);
- “sectoroverstijgend” (various);
- “taal en culture” (language and culture);
- “techniek” (technology).

Some categories are also divided even further into “sub-categories”, but we decided not to consider these in our visualization. We did this because for most sub-categories, the description was simply “not applicable (*[main category]*)”.

1.2 Description of the format

The data is delivered as a CSV file with semicolons as the field separators and new-lines as the record separators. There is a lot of redundancy; for example for every study of a university, all information about the university is repeated.

A detailed description of the data fields follows.

- **PROVINCIE** – the state the university is located in.
- **GEMEENTENUMMER** – an ID number of the county the university is located in.
- **GEMEENTENAAM** – the name of the county the university is located in. For some reason, this is given in all-caps.
- **BRIN NUMMER ACTUEEL** – an ID number of the university.
- **INSTELLINGSNAAM ACTUEEL** – the name of the university.
- **CROHO ONDERDEEL** – the CROHO category.
- **CROHO SUBONDERDEEL** – the CROHO sub-category. Most of the time this is ”n.v.t. (...)” (which means “not applicable”).
- **OPLEIDINGSCODE ACTUEEL** – an ID number of the study program.
- **OPLEIDINGSNAAM ACTUEEL** – the name of the study program.
- **OPLEIDINGSVORM** – the education type,
  - “voltijd onderwijs” (full-time),
  - “deeltijd onderwijs” (part-time) or
  - “duaal onderwijs” (dual).
- **OPLEIDINGSFASE ACTUEEL** – the study phase, “bachelor”, “master”, “initiële opleiding” (initial) or “beroepsfase” (clerkship).

- **SOORT DIPLOMA** – the type of graduation.
- **2007 MAN** – the number of male students that graduated in 2007 on the given study, education type and study phase.
- **2007 VROUW** – the number of female students that graduated in 2007 on the given study, education type and study phase.

  ...

- **2011 MAN** – the number of male students that graduated in 2011 on the given study, education type and study phase.
- **2011 VROUW** – the number of female students that graduated in 2011 on the given study, education type and study phase.

1.3 Additional data

For our visualization we needed to know the position of the universities in the Netherlands, as we wanted to show the universities on a map (see section 2 for a more elaborate description of our visualization design). Since exact coordinates were not given in the data set, we manually entered some position data for the universities.

2 Design

In this section, we explain the general design of the visualization.

2.1 General design

We noticed that studies are not really comparable between universities. Often, different universities have different names for essentially the same studies, and it also happens that different studies have the same name across universities. Therefore, an interface which is based on study names would not be very useful.

Because of this, we designed an interface that is based on universities. First the user clicks a university, and then more information about this university is shown. The drawback of this approach is that it is not easily possible to compare the same study program between different universities. Based on the CROHO categories, the user is able to compare entire categories of studies between universities, however.

In our interface, there are two "modes". In the first mode, an overview is visible of all universities. We chose to do this on a map of the Netherlands, since such a map presents a list of universities more nicely in our opinion than a list.

Now, on the map, every university is represented by a small pie chart. This chart shows an overview of the studies on that university, based on categories. Furthermore, the size of the chart shows the number of graduates.

When the user clicks a pie chart, the interface moves to the second mode. There, detailed information about the university is shown. For this, we use a simple sunburst

diagram consisting of two rings: the inner ring shows the categories, and the outer ring shows the studies in every category. Thus, the inner ring of the sunburst diagram corresponds to the pie chart shown in the first mode. The user can switch back to the first mode, for example to choose another university, by clicking a "back" button.

Since the two modes easily fit together on the screen when using a large screen, we chose to show the two modes together indeed in that case. So if the user uses the application using a small window, only one of the modes is visible at a time. If the user uses a large screen, no screen space is wasted by showing both parts at the same time.

2.2 Sliders

The interface as described above does not yet allow the user to select a particular year. To allow this, we added a bar at the top of the screen (that is visible in all modes). In that bar, a slider is shown with which the user can select the year: "total", "2007", ..., "2011". If this slider is changed, both modes are changed: in the map mode, the pie charts are updated to only use data from the selected year, and in the detail mode, the charts are updated in the same way.

Furthermore, we added a "Group on" slider. This influences the type of pie chart shown. There are two types possible.

- With "phase", the pie charts show the fraction of graduates in certain study phases (bachelor, master, ...). For the detail mode, this means that in the inner ring of the sunburst diagram, the phases are shown, and in the outer ring the studies are ordered based on these phases.

- With "category", the pie charts show the fraction of graduates in CROHO categories. For the detail mode, now CROHO categories are shown in the inner ring, and in the outer ring the studies are ordered based on these categories.

3 Data representation

We do not use all of the fields in the source data. To be exact, we use the following fields: **BRIN NUMMER ACTUEEL**, **CROHO ONDERDEEL**, **OPLEIDINGSNAAM ACTUEEL**, **OPLEIDINGSVORM** and **OPLEIDINGSFASE ACTUEEL**, and of course the ten data fields (**2007 MAN** up to **2011 VROUW**).

As we saw before, the source data contains a lot of redundancies. Furthermore the sub-totals are not summed. So if we for example want to know the total amount of graduates on the Eindhoven University of Technology, we have to compute this total ourselves. That does not take a long time of course, but if we have to do those summations all the time, the visualization would probably become rather slow.

Therefore, we pre-process the data into a representation that is more useful for our purposes. The goals is to be able to retrieve all data we want to show (totals, subtotals, ...) immediately. Almost immediately we realized that it would be necessary to split the data in two parts: data categorized by CROHO category and data categorized by

education phase. This would enable us to implement all views described above in a fairly straightforward manner, regarding the use of data that is. To this end, we came up with the following two internal representations of data.

3.1 Categorized by CROHO category

When categorizing the data on CROHO categories, the following format worked best for us. We have a list with indexes *min*, *max* and a bunch of BRIN numbers. The *min* and *max* indexes hold the number of students for the smallest and largest university with respect to number of graduates, respectively. Each BRIN number references an array of CROHO categories. Per category, a list of years is available. This list has the following indexes:

- *2007*, *2008*, . . . , *2011*: the total number of graduates for this university and this CROHO category in that year.
- *2007-female*, . . . , *2011-female*: same as above, with only female graduates.
- *2007-male*, . . . , *2011-male*: same as above, with only male graduates.
- *2007-fulltime*, . . . , *2011-fulltime*: same as above, with only full-time graduates.
- *2007-parttime*, . . . , *2011-parttime*: same as above, with only part-time graduates.
- *2007-dualtime*, . . . , *2011-dualtime*: same as above, with only dual graduates.

There is also an index *studies*, which references a list of studies. Per study, a list with indexes from the above list holds the information per study. There is also an index *phase* per study, which indicate the education phase of that study. This came in handy in some parts of our application.

Using this data structure, we can obtain the number of male graduates in 2009 for the master *Computer Science and Engineering* at the Eindhoven University of Technology as follows. Note: the BRIN number for the Eindhoven University of Technology is 21PG.

```
data["21PG"]["techniek"]["studies"]
  ["M Computer Science and Engineering"]["2009-male"]
```

3.2 Categorized by education phase

The data structure for this categorization is very similar to the one described in section 3.1. The only differences are that instead of CROHO categories, there are education phases now and each study has an index *category* that indicates the CROHO category of the study, instead of a reference to the education phase which is known in this categorization.

Using this data structure, we can obtain the number of male graduates in 2009 for the master *Computer Science and Engineering* at the Eindhoven University of Technology as follows. Note: the BRIN number for the Eindhoven University of Technology is 21PG.

```
data["21PG"]["master"]["studies"]
  ["M Computer Science and Engineering"]["2009-male"]
```

4 Implementation

We decided that an implementation in HTML, CSS and JavaScript would be nice, since users would be able to visit the visualization easily then using their own web browser, without needing to install additional software.

For the visualization of data, we used the *D3.js*[2] library. This library allows you to handle data in JavaScript easily, by binding the data to nodes in the DOM (the tree that the browser internally uses to represent a document). It also comes with a set of default visualization techniques, such as pie charts, which we used quite a lot. Additionally, we used *jQuery*[3] and *Bootstrap*[4] (and a library[5] for sliders) for the interface.

4.1 Technical notes

We have tested the application in Firefox and Chrome (Chromium). Other modern browsers like Safari and Opera should also support our application, but we have not tested that. Internet Explorer 10 shows the application correctly, but for some reason fails to render the SVG background map. We have not been able to find why this happens.

For easy viewing, we have put the files on the internet [6] (using Dropbox). If you open the link in a supported browser, the visualization should open.

Of course it is also possible to run the visualization locally; however there are some caveats there. In Firefox, opening *graduates.html* is all you need to do to run the application. Chrome however does not allow to load data from a local file. Therefore, you need to start Chrome (Chromium) from the command line as follows:

`chrome --allow-file-access-from-files`
(or alternatively `chromium --allow-file-access-from-files`).

It should also work to deploy all files on a (local) web server and then opening that via the browser.

4.2 Background map

First of all, we needed a map of the Netherlands. Of course, we wanted not to clutter the view with (for our purposes) unnecessary items such as roads. We used *Natural Earth* data[7] to create our own simplified map.

Of course we decided to add the coast line, and we used a different color for the sea and the land areas. Furthermore we decided to include the national border and the state borders, since they make it easier to navigate the map. Finally we decided

[2] `http://d3js.org/`
[3] `http://jquery.com/`
[4] `http://getbootstrap.com/`
[5] `http://www.eyecon.ro/bootstrap-slider/`
[6] `https://dl.dropboxusercontent.com/u/14874671/Visualization%203/graduates.html`
[7] `http://naturalearthdata.com/`

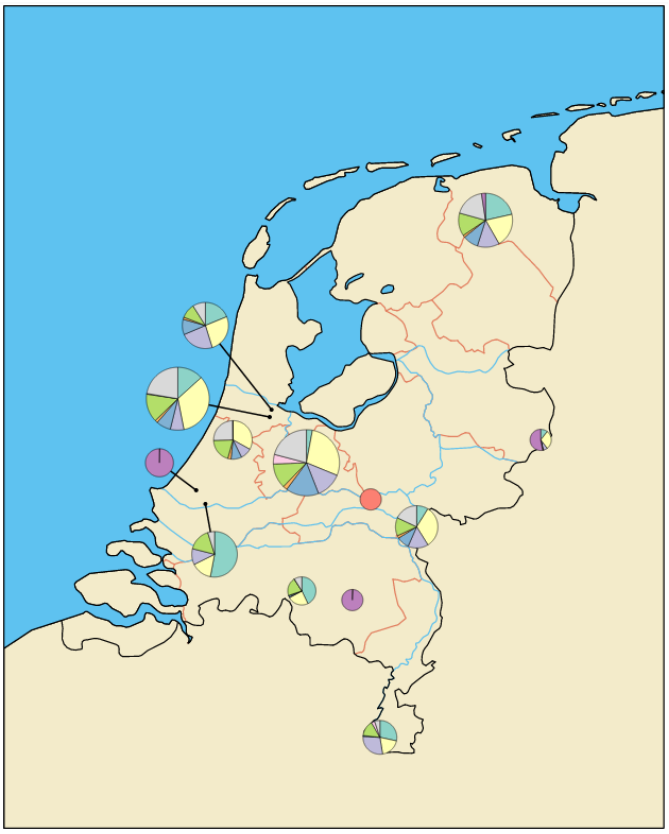

Figure 2: Screenshot of the map mode with in the background a map of the Netherlands.

to include important rivers on the map, since we felt that that also helped the user in navigating. We did not include roads or railways since they are only distracting from the actual content. Furthermore we did not include cities, since we already are going to add relatively large pie charts for the universities, that would certainly obscure the cities they are in.

See Figure 2 for the resulting background map.

4.3 Main interface

As discussed in section 2, we implemented the interface in two parts: the map mode (on the left) and the detail mode (on the right). Alternatively, if the window is small, only one of these two modes is shown at a time.

Furthermore there is a title bar on the top of the screen. In this bar (which is always visible), two sliders are included, with which the user can set the year to get data from, and the grouping (on study phase or CROHO category). See the left part of Figure 3.

Furthermore, in the title bar a button is shown to open the legend dialog (see below for more about this). See the right part of Figure 3.

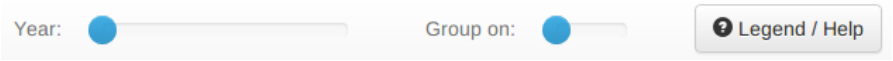

Figure 3: Screenshot of the sliders to set the year and the grouping.

4.4 Map mode

In the map mode, we wanted to have pie charts on the correct positions on the map. We manually entered coordinates for the universities in the Netherlands and implemented a system where a university can have two coordinates: a "real" coordinate and a "display" coordinate. This was needed to put for example the University of Amsterdam and the VU University Amsterdam apart from each other, because they are at the same location on the map. We draw a line from the display coordinate to the real coordinate automatically if a display coordinate is present, to show where the university actually is located.

We implemented the pie charts using the standard D3.js `d3.layout.pie()` pie chart layout. This was a bit of a hassle in the beginning, but once we got to know D3.js a bit, it was fairly straightforward to add animations when changing between grouping for example. We also implemented a small animation where one part of the pie enlarges a bit when you hover it with the mouse. This is to accentuate which category or phase the user is hovering. When hovering a part of a pie in this mode, a tooltip will pop up with some more detailed information, like absolute number of students, name of university, et cetera. To see more detailed information, you can click an arbitrary part of a pie to view that pie in more detail in the detail mode.

4.5 Detail mode

The most important part of the detail mode is of course the large chart that shows the categories (in the inner ring) and studies (in the outer ring). (If the grouping slider is set to "phase" instead of "category", the inner ring shows phases instead.) The (angular) size of the studies is linearly dependent on the number of students that graduated in that study. See Figure 4 for two screenshots.

Of course, the studies in the outer ring line up with the categories (or phases) in the inner ring. This means that when the user switches from grouping on "category" to "phase" or the other way round, the studies have to be rearranged to again line up with the inner ring. To make this easier for the user to follow, the subjects are not moved immediately. Instead, the positions are animated, such that it is more easily visible for the user where the studies are moving to. It is still quite messy if there are a lot of studies, but the user can choose ("pin") one study to follow. That study then stays visible (see the section on hovering and pinning below).

4.6 Subcharts

In the detail mode, below the large sunburst diagram, there are two smaller pie charts. On the left, there is a chart that will look something like shown in Figure 5. This chart

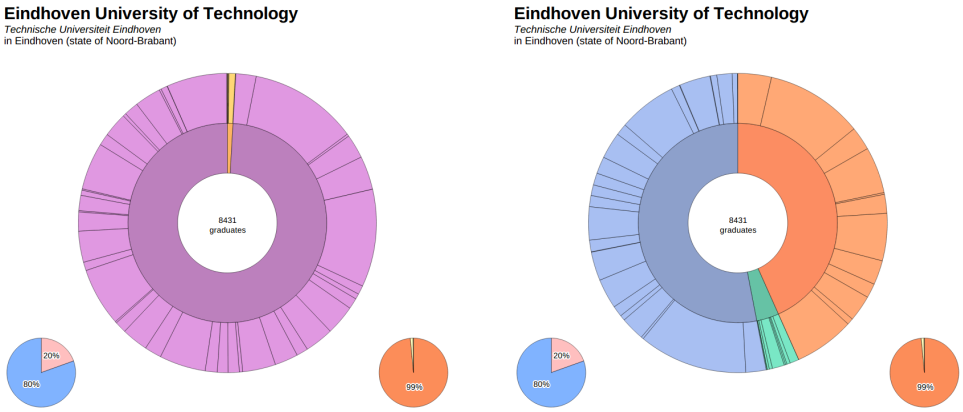

Figure 4: Screenshot of the detail mode for the Eindhoven University of Technology. In the left image, the grouping slider is set to "category"; in the right image, it is set to "phase". It can be seen in the left image for example that the TU/e (almost) only has technical studies (purple). From the right image we can conclude that the TU/e has slightly more master graduates (blue) than bachelor graduates (orange), presumably because students from other universites move to the TU/e.

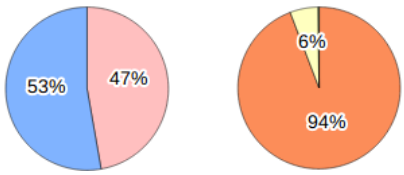

Figure 5: Screenshot of the subchart that shows the male to female ratio on the left, on the right a screenshot of the subchart that show the full-time, part-timie and dual ratio.

indicates the number of male graduates compared to the number of female graduates. Initially, this chart will show this ratio for the entire university, so it compares the total over all studies. When the user hovers a study, component or phase in the sunburst, then this subchart will update to show the ratio for that particular study, component or phase. For a component or phase, again the total of all studies in that component or phase is shown.

On the right, a similar chart will show the number of full-time studying graduates versus the number of part-time studying graduates versus the number of dual studying graduates. This chart updates in similar fashion as the male/female chart and is depicted on the right in Figure 5.

4.7 Hovering and pinning

To get more detail over a particular study, the user can hover it with the mouse. When this happens, the hovered study is highlighted and information about this study is shown. Firstly, the name of the study is shown below the sunburst chart. Secondly, the

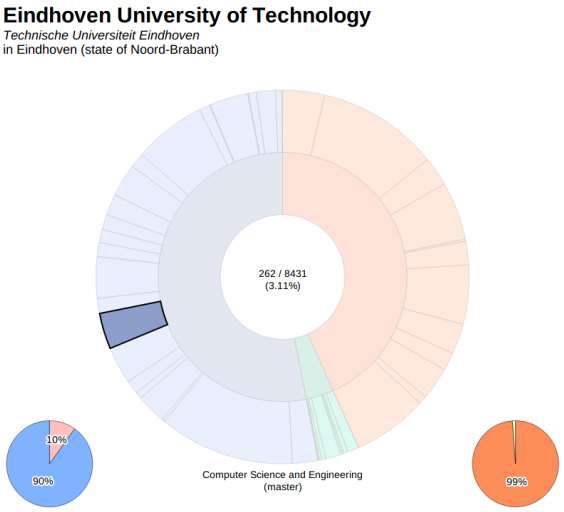

Figure 6: Screenshot of a pinned study. If this study moves or resizes now, for example because the user drags the year or grouping slider, this study stays selected and highlighted.

amount of graduates from this study is shown inside the sunburst. Furthermore, the subcharts are updated to only take into account information about the hovered study.

However, when the user leaves the study again with the mouse, the information is reset to its previous state (that is, showing information about the entire university). To prevent this, the user can also make the highlight more permanent by clicking the study. Now, the other parts of the sunburst chart are faded away a bit, and when the user moves away the mouse, the study stays highlighted. We call this process *pinning*.

Now the user can freely do other things, such as dragging the year slider, to see how the pinned study got more or less graduates over the years, for example. In this case, also the subcharts are updated correctly with the data from that year. Furthermore, the grouping slider can be dragged, and the user can now easily follow the pinned course to its new location.

The same mechanism of hovering and pinning also works for the inner ring. Then, totals for all studies in the selected category (or phase) are shown.

4.8 Legend

Of course it is important that users can look up what the colors mean. For that, we mostly use annotations when the user hovers over a color. For example, if the user hovers over the blue part in the male/female subchart, a text “male” appears below the chart.

However, we also wanted to have one place where the user can look up all colors at once. For this, we created a legend dialog. When the user clicks the “Legend / Help” button in the title bar, this dialog appears. See Figure 7 for a screenshot of the dialog.

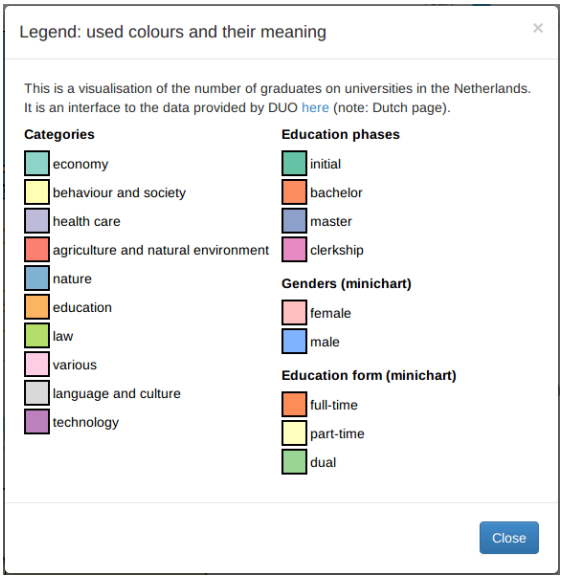

Figure 7: Screenshot of the legend dialog.

5 Trends

We have used our application to analyze the data and found a couple of trends. By this we mean remarkable changes or values in the data. In Table 1.1, we list these and describe how our application helped in spotting these trends.

6 Conclusion

We implemented a visualization for a data set containing information about graduates in the Netherlands. The implementation is a general tool for visualizing the data set; it is not tailored to a specific trend. Instead, the user can choose what needs to be shown for the question at hand. We think that a lot of questions about the data set can more easily be answered with our visualization.

Trend / remarkable value	**Our application helped in this way**
The number of graduates has been growing nationally over the years 2007 to 2011.	The overview of all universities shows pie charts with a radius that depends on the total number of graduates. When changing the year using the slider, one can clearly see that all pie charts grow.
In Utrecht, all students graduating in *Klinische Gezondheidswetenschappen (master)* study part-time. The same holds for *Nederlands recht (initial)* in Groningen.	When hovering studies, the mini chart showing the full-time/part-time/dual ratio suddenly changed color completely.
In Eindhoven, the ratio male to female is biased towards male students.	Besides our own experience, one can clearly see this in the mini chart showing this ratio. It is easy to compare this with other universities or studies.
The ratio male versus female can be more biased than it is in Eindhoven. See Nijmegen, *Pedagogische Wetenschappen (master)*.	The mini chart turns pink almost completely!
Studying dually (work and study simultaneously) seems very rare. An exception to this is the master *Redacteur/Editor* on the University of Amsterdam, which is done exclusively dually.	When hovering the inner ring of the sunburst, noticing that at least one master study had a fair amount of dual graduates is easy as all of a sudden, a green part pops up. Finding the study then is a matter of hovering the studies.

Table 1.1: Some trends that are easily visible using our visualization.