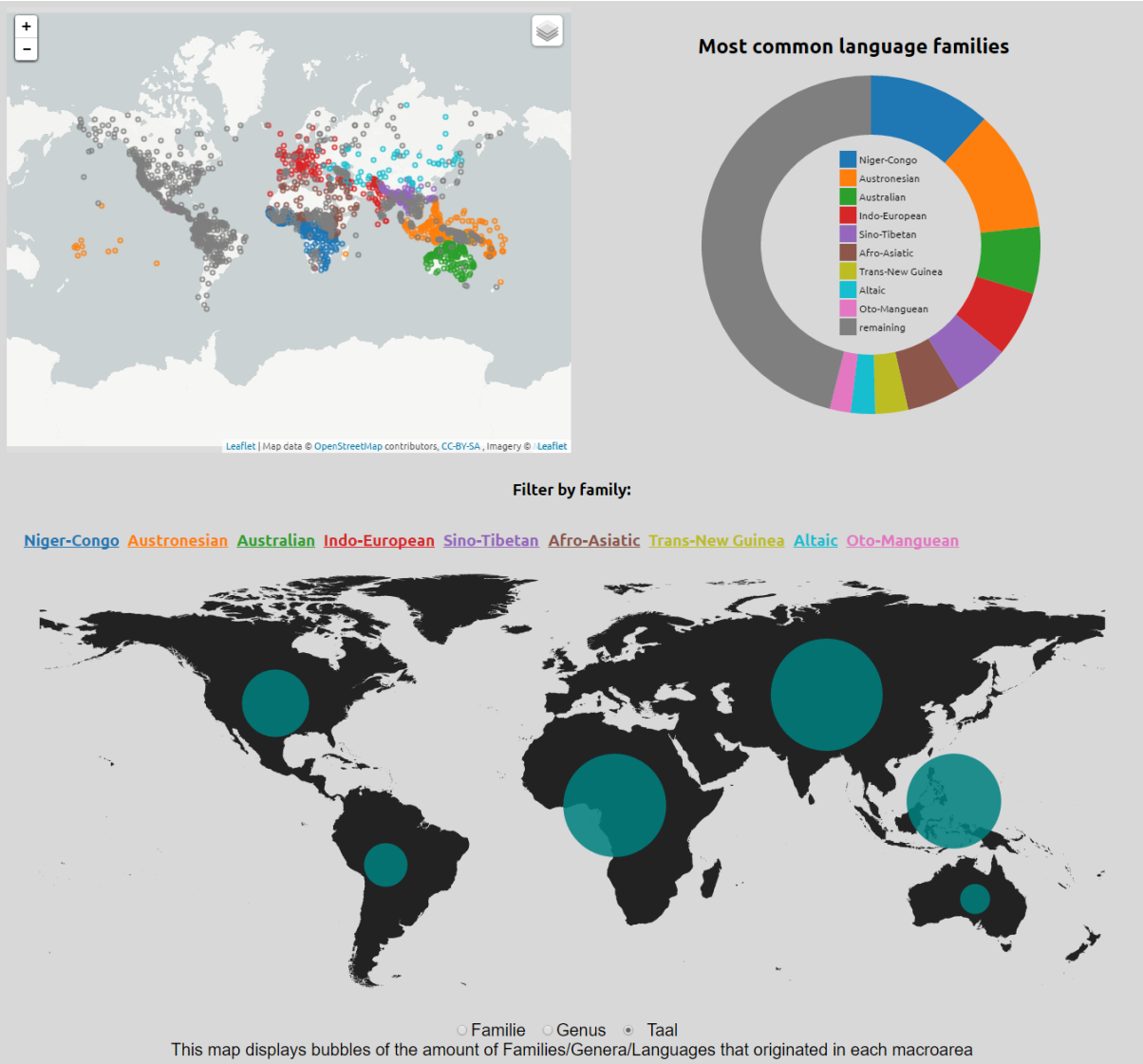
Informatievisualisatie (5072INFO6Y)

D3 Project Eindverslag

Siyawash Estanekzay	10379932
Alrian Kamdhi	11030682
Lex Poon	11031530
Britt Ruigrok	10780491
Muhammad Shuduyev	11136626
Marty Star	10215387

23-06-2016



1. Inleiding

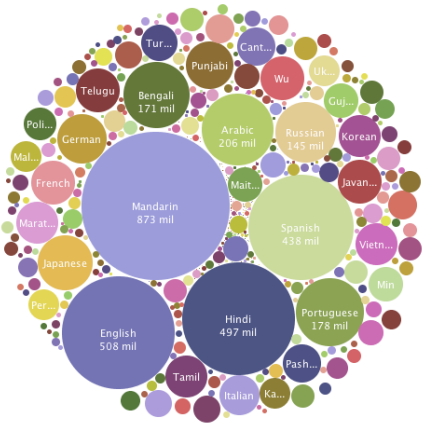
Er zijn meer dan zevenduizend talen over de hele wereld. Slechts een paar worden als wereldtaal beschouwd en over de hele wereld gesproken. Hieronder valt uiteraard Engels, Frans, Spaans, Russisch en Arabisch. Echter zal er in dit onderzoek voornamelijk de focus liggen op de top drie verspreide talen in de wereld en niet naar de top drie gesproken talen. Hiervoor is gekozen zodat het inwonersaantal geen bias kan vormen voor het onderzoek. De dataset die wij willen visualiseren is een .csv bestand bestaande uit 2679 talen en afkomstig van de World Atlas of the Language Structures (WALS). De dataset bevat de geboorteplaatsen van die talen en data als latitude en longitude die samen de coördinaten vormen om uit te zoeken waar op de wereld welke taal gelokaliseerd is. In de huidige staat bevat de dataset dus veel informatie. Deze informatie is echter niet overzichtelijk, waardoor het minder functioneel is voor de gebruikers die niet alle informatie nodig hebben. Zodoende heeft het onderzoeksteam de dataset met behulp van Python telkens aangepast naar eigen wens. Op deze manier kon het onderzoeksteam effectief en snel te werk. Volgens Fekete et al. (2008) hebben interactieve visuele representaties van data als functie om de cognitie te versterken. Hiermee wordt bedoeld dat de mens een dataset visueel beter kan verwerken dan wanneer het in een tabel staat. Aan de hand van de WALS-dataset hebben wij een visualisatie gemaakt van de 2679 talen op een wereldkaart, daarbij kunnen gebruikers per macroarea in- en uitzoomen om zodoende een visualisatie van de talen en de bijbehorende locatie ervan te beschouwen. Door middel van deze visualisatie kunnen gebruikers gemakkelijk de oorsprong van verschillende talen vinden en hoe deze zijn verspreid. Wat interessant kan zijn is wanneer de gebruiker de gevisualiseerde data analyseert, hij kan achterhalen hoe een taal zich mogelijkerwijs over de wereld heeft verspreid. Een taal kan op verschillende manier uitgroeien tot een wereldtaal. De locatie van oorsprong van een taal zou hier mogelijk invloed op hebben gehad. Door middel van de visualisatie en een literatuuronderzoek zal getracht worden om te achterhalen of de locatie van oorsprong van een taal een relatie met de verspreiding van een taal heeft?

2. Gerelateerd werk

Bestaande visualisaties

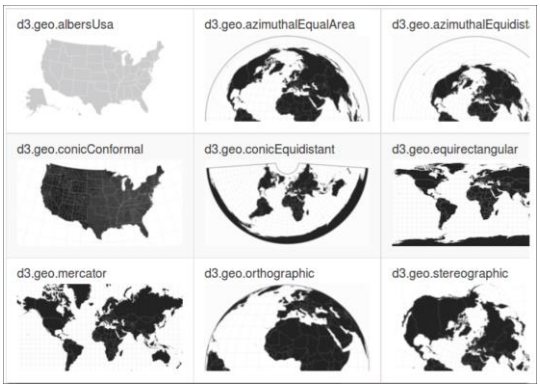
Voordat aan het onderzoek daadwerkelijk is gewerkt, is er eerst gekeken naar voorafgaande visualisaties van dezelfde database. Op deze manier kan men analyseren hoe anderen het hebben aangepakt. Door de analyse kan dus achterhaald worden wat wel en niet goed is om te gebruiken in de visualisatie. De volgende visualisaties zijn een inspiratiebron geweest voor de uiteindelijke visualisatie. Uit de bubble chart (Figuur 1) is naar voren gekomen dat het erg handig kan zijn om door middel van bellen een visualisatie te maken. De grootte van elke bel representeert de grootte van de desbetreffende taal dus des te groter de bel, des te groter de taal. Op deze manier wordt in een oogopslag duidelijk welke taal in de wereld dominerend is en welke taal niet. Er kan dus gemakkelijk

onderscheid in de data gemaakt worden zonder deze uit te werken. Door deze redenen is er voor gekozen om de bubble chart mee te nemen in de uiteindelijke visualisatie.



Figuur 1 Bubble chart van de 23 grootste talen in de wereld

Verder is er gekeken naar een tweede visualisatie die voor het onderzoek relevant zou kunnen zijn. In Figuur 2 kan men de verschillende manieren beschouwen die gebruikt worden om een wereldkaart te visualiseren. De mercatorprojectie sprong ertussen uit, de reden hier achter was dat een goed overzicht van alle landen wordt gegeven die in de WALS-dataset staan. Zo wordt er niet onnodig ruimte weggeven, denk hierbij aan het feit dat de polen niet op de mercatorprojectie worden weergegeven. Deze ruimte kan nuttig worden gebruikt voor alternatieve componenten.

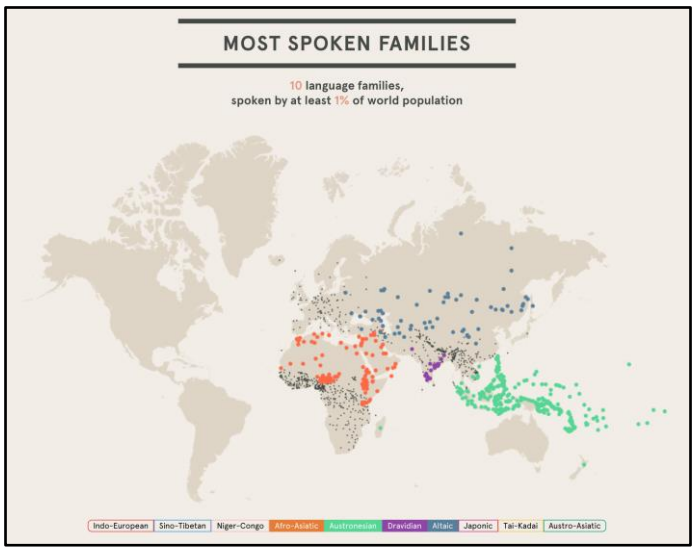


Figuur 2 De verschillende mogelijkheden om een wereldkaart te visualiseren in d3

Zoekend naar een visualisatie die gebruik maakt van een mercatorprojectie, zijn we uitgekomen op de gegeven visualisatie in Figuur 3. In Figuur 3 kan men zien dat 3 belangrijke aspecten in de visualisatie zitten. Allereerst de mercatorprojectie en ten tweede is er gebruik gemaakt van verschillende kleuren om een onderscheid te kunnen creëren tussen de families van de talen. Tot slot kan men de families selecteren door op de juiste knop onderaan de visualisatie te klikken, hierdoor wordt de desbetreffende kleur geactiveerd. Hieruit is voortgekomen dat voor de mercatorprojectie het kliksysteem met onderscheid van kleuren ook voor ons onderzoek van toepassing is.

De doeltreffendheid van de visualisaties tegenover de aanpak van onze visualisatie

De visualisaties in Figuur 1 en 3 hebben als gemeenschappelijk doel om de corresponderende data gemakkelijk en overzichtelijk te weergeven. Zoals aangegeven in 2.1 was het doel van Figuur 1 om een onderscheid te maken tussen de dominerende talen in de wereld. De doelstelling van deze visualisatie is bereikt, aangezien men snel een onderscheid kan maken in hoeverre de talen dominant zijn. Wat hieruit gehaald kan worden is dat er de bubble chart methode een goede manier is om overzichtelijk en snel aan de gebruikers iets duidelijk te maken. In Figuur 3 was gebruik gemaakt van de mercatorprojectie om de spreiding van een familie te weergeven. Wanneer men geen van de families selecteert op het kliksysteem, dan worden alle families in het zwart weergeven. Dit zorgt ervoor dat er een globaal beeld van het geheel wordt gevormd. Wanneer er een van de families wordt geselecteerd krijgt elk corresponderend stip de juist kleur, dit valt de gebruik gelijk op. Deze methode is dus gemakkelijk te begrijpen en snel te gebruiken. Kortom, er kan dus vastgesteld worden dat de visualisatie zeer doeltreffend is.



Figuur 3 Most Spoken Families

Uit de analyse van de voorgaande visualisaties is duidelijk geworden dat gebruik van een mercatorprojectie essentieel is om ons doel, de onderzoeksvraag, te bereiken. Dit komt omdat deze manier een goed overzicht kan geven van de spreiding van de geboorteplekken van een taal. Door hiervan een goed beeld te hebben, kan men betere analyses op de visualisatie loslaten. Om tot een ideale visualisatie te komen, is er gekozen voor een graduated symbol map. De graduated symbol map is een combinatie van zowel een bubble chart als een mercatorprojectie. Dit zorgt er dus voor dat er optimaal overzicht van de gewenste data wordt weergegeven.

Onderzoek

Paul, Simons & Fennig (2016) hebben een ranglijst gepubliceerd waarin per taal de spreiding is aangegeven in de hoeveelheid landen dat het gesproken wordt. Engels steekt erboven uit met 106 landen waarin de taal wordt gesproken. Daarna volgt Arabisch met 58 landen en op de derde plaats Frans met 53 landen. Op de gemaakte visualisaties is te zien dat de oorsprong van zowel Engels als Frans teruggevonden kan worden in de eigen landen. Arabisch vind zijn oorsprong terug in Afro-Azië. Wat opvalt is dat alle drie de landgebieden aan het water liggen. Hierdoor is het aannemelijk dat de spreiding van de talen zowel over water als over land heeft plaats gevonden. Volgens Garcia (2010) zijn er vijf beleidsredenen voor het verspreiden van een taal:

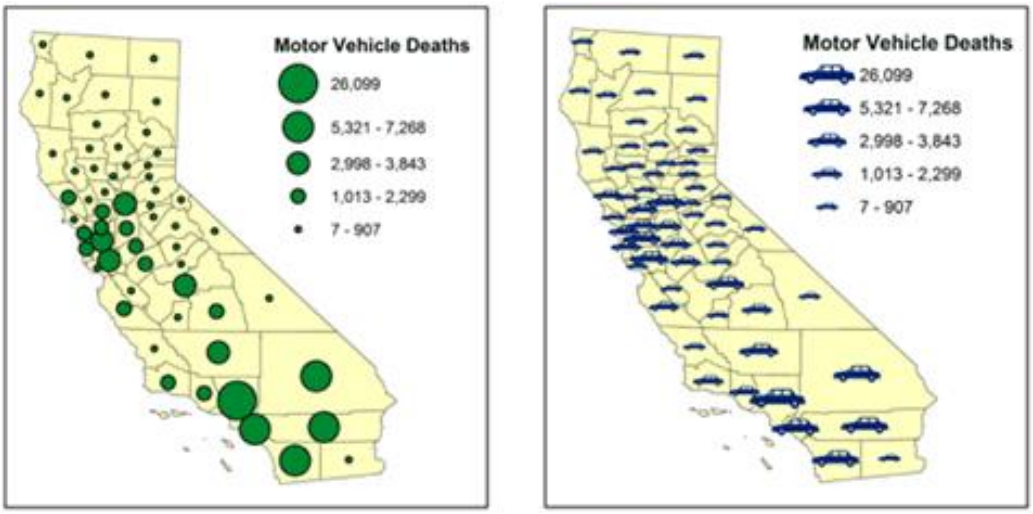
- De communicatie te verhogen
- Om een ideologie te verspreiden
- Om economische banden te ontwikkelen
- Om inkomsten uit taalstudies en producten te krijgen
- Om nationale identiteit en trots te behouden

De verspreiding van de talen Engels, Frans en Arabisch vallen onder algemeen bekende gevallen, zo is Arabisch via de Islamitische expansie verspreid. Engels en Frans zijn via koloniale overheersingen in Azië en Afrika verspreid (Garcia, 2010). Zo kan er gesteld worden dat wellicht de locatie van oorsprong van het Engelse taal en Franse taal een voordeel had ten opzichte van de andere talen. Aangezien beide landen erg ambitieus waren met handel drijven, zijn beide landen destijds gaan groeien. Zo hadden ze een voordeel dat ze via het water andere landen makkelijker konden bereiken. Op deze wijze zijn beide talen verspreid geraakt en achtergebleven bij de volkeren van de koloniën. Geconcludeerd kan worden dat de locatie van oorsprong van de talen een ondersteuning hebben gevormd voor de spreiding van de talen. Echter kan er niet uitgesloten worden dat er andere invloeden ook van toepassing waren, als het verspreiden van de ideologie en via koloniale expansie economische banden ontwikkelen.

3. Voorgestelde visualisatie

Voor de statische landkaart en de donut diagram zijn er datasets gebruikt die afgeleid zijn van de dataset afkomstig van de *World Atlas of Language Structures*. Zo is er voor de statische landkaart een dataset gemaakt dat bestaat uit de coördinaten van de macrogebieden (Afrika, Australië, Eurazië, Noord-Amerika, Zuid-Amerika en Papunesië) en het aantal families, genera en talen per macrogebied. Deze dataset is als een .csv file in eladen en de gegevens van elke kolom is uit de .csv file verkregen door de d.kolomnaam syntax van D3. Deze gegevens zijn nodig voor het visualiseren van een *graduated symbol map*. Een graduated symbol map is een landkaart waarbij de grootte van vormen worden gebruikt om iets te visualiseren (Heer, Bostock & Ogievetsky, 2014). Hier zijn voorbeelden gegeven van graduated symbol maps (figuur 4), waarbij er gebruikt wordt gemaakt van cirkels en

auto's. De grootte van de vormen van de cirkels en de auto's is afhankelijk van hoeveel mensen er zijn overleden door motorvoertuigen. Deze zelfde principe is gebruikt bij onze statische landkaart. De grootte van de cirkels op onze visualisatie wordt bepaald door het aantal families/genera/talen die zijn ontstaan in elke macrogebied. Deze cirkels zijn dan weer geplaatst op het middelpunt van hun respectievelijke macrogebied. Verder kan er bij onze graduated symbol map gefilterd worden op categorie (familie, genus of taal). Het soort wereldkaart dat gebruikt is voor de visualisatie een equirectangular projectie. Het voordeel van deze projectie is dat de belangrijkste gebieden het grootst zijn afgebeeld. Zo is er minimale verspilling van ruimte, doordat gebieden als Antarctica en Groenland relatief klein zijn afgebeeld (deze gebieden zijn volgens de dataset onbelangrijk, dus moet hun grootte geminimaliseerd worden). Het enige wat er mist bij onze visualisatie wat normaliter aanwezig is op graduated symbol maps is een legenda dat aangeeft welke waarde bij welke grootte hoort. Dit is onnodig bij onze visualisatie, omdat het bij ons niet gaat om de absolute waarden, maar meer om de waarden van de macrogebieden in vergelijking met elkaar.



Figuur 4 Graduated symbol maps

De dataset voor onze donut diagram is ook een aangepaste versie van de dataset van de WALS. Deze dataset bestaat uit de grootste families van talen en uit hoeveel talen die families bestaan. De donut diagram die uit deze dataset ontstaat laat zien uit welke familie de meeste moderne talen zijn ontstaan. De grootste aandeel ligt echter bij de categorie "remaining". Deze categorie is een collectie van alle families die niet expliciet in de dataset worden genoemd. Aangezien deze categorie het grootste is kan het in de weg staan als men alleen de grootste families met elkaar wilt vergelijken. Het is om deze reden mogelijk om categorieën weg te laten uit het diagram door bij de legenda op het vierkantje links van elke categorie te klikken. Het diagram vult zich dan met de categorieën die men wel wilt bekijken. Ook kan men zien hoeveel talen elke familie heeft en hoeveel procent van het totaal dat aantal is. Door op een van de categorieën op de donut diagram te zweven met de muis kan men een

pop up zien waarin staat hoe de familie heet, hoeveel talen die familie heeft en hoeveel procent dit van het totaal is.

De laatste visualisatie is nog een wereldkaart, maar dan anders geïmplementeerd. De data voor deze visualisatie komt van het originele dataset van de WALS en is als .csv bestand met D3 geopend. Uit die dataset zijn de longitude, latitude, familie en naam verkregen. Al die gegevens zijn gebruikt om de visualisatie te maken. Op deze visualisatie is te zien waar de moderne talen op de wereld zijn ontstaan. Ook zijn de talen ingedeeld per familie. Dat wil zeggen dat alle talen van dezelfde familie dezelfde kleur hebben. De kleuren die zijn gebruikt om de talen per familie in te delen zijn dezelfde kleuren als die zijn gebruikt in de donut diagram. Dit is gedaan om een eenheid te creëren tussen de visualisaties. Het is ook mogelijk om de datapunten te filteren door families aan/uit te vinken. Zo kan men duidelijker zien waar bepaalde families zijn ontstaan. Voor deze kaart is er gebruik gemaakt van een mercatorprojectie. Het is mogelijk om in te zoomen zodat de verschillende punten beter te zien zijn en makkelijk te klikken zijn. Aangezien het mogelijk is om in/uit te zoomen is verspilling van ruimte een minder grote zorg dan wanneer een statische kaart wordt gebruikt. Een mercatorprojectie pakt in dit geval ook beter uit. Dit komt doordat de window voor deze kaart een gelijkzijdige vierhoek is en een mercatorprojectie past dus beter in de window. Als er weer een equirectangular projectie werd gebruikt, dan zou de kaart te wijd worden, waardoor er veel naar links en rechts gesleept of veel uitgezoomd moest worden om alle datapunten te kunnen zien. De mercatorprojectie is meer compact van formaat, waardoor de datapunten makkelijker zichtbaar zijn (de projecties zijn te zien in figuur 2). De zichtbaarheid (en klikbaarheid) van de punten is een belangrijk aspect van deze visualisatie, omdat men van elk punt kan zien welke taal het is door op het punt te klikken.

Alle kleurencombinaties die zijn gebruikt in de visualisaties zijn gemaakt door ColorBrewer 2.0. De data die een kleur hebben gekregen zijn van nature kwalitatief, dus de kleuren moeten contrasten, maar niet complementair zijn om een indruk van tegenover gesteldheid te voorkomen.

4. Reflectie op het Teamwerk

4.1 Voorbereidingsfase

Elk groepslid moet voldoende D3 kennis bezitten. Zo moet iedereen de basis kennen en voldoende weten om de visualisaties maken die passen bij de gekozen dataset. De visualisaties moeten duidelijk zijn en hun boodschap effectief overbrengen. Dit moet gedaan worden aan de hand van literatuur voor een betrouwbare en wetenschappelijke benadering. De interacties die de visualisaties veroorloven moeten duidelijk zichtbaar zijn, zodat de lezer/eindgebruiker weet wat hij wel en niet met de visualisatie kan doen. De visualisaties moeten dus een goede affordance hebben voor hun interacties. Dit zijn de voorwaarden waaraan geprobeerd is om aan te houden. Ter voorbereiding voor het uiteindelijke programmeerwerk, heeft ieder groepslid een eigen visualisatie gemaakt om zo ideeën op te doen voor het eindvisualisatie. Tijdens deze fase zijn er veel ideeën opgedaan, maar ook veel ideeën afgewezen. Het maken van een tijdlijn was één van die ideeën. Er waren geen relevante datasets beschikbaar op het internet om hier een visualisatie van te maken. De ideeën die goed genoeg waren om uit werken was het maken van een wereldmap, met daarop de locaties waar elk taal is ontstaan. Ook was het idee om een donutchart te maken met daarop alle talen die gesproken worden, ingedeeld per macroniveau, een idee om uit te werken.

4.2 Ontwikkelfase

Het team bestaat uit studenten die goed met elkaar om kunnen gaan en goed kunnen samenwerken. Het eerste week ging wat moeizaam. De planning was nog niet goed uit bedacht waardoor er niet optimaal aan het visualisatie was gewerkt. De weken daarna ging het steeds beter. Door de juiste planning te hebben opgesteld wist iedereen wat er van hem of haar verwacht werd. Door deadlines op te stellen werd de team echt aangespoord om nauwkeurig en precies te werken. Per tweetal werd er aan één visualisatie gewerkt en als er iemand vastliep, dan waren we altijd samen om elkaar te helpen.

4.3 Taakverdeling

Wie?	Taak
Marty Star	**World Map programmeren, World Map dataset schoonmaken, report aanvullen**
Siyawash Estanekzay	**World Map programmeren, Planning maken, report aanvullen**
Alrian Kamdhi	**World Map programmeren, World Map dataset schoonmaken, report aanvullen**
Britt Ruigrok	**World Map programmeren, zoeken dataset, report aanvullen**
Lex Poon	**World Map prgrammeren, dataset zoeken, report aanvullen**
Muhammad Shuduyev	**Donutchart programmeren, Planning maken, report aanvullen**

4.4 Individuele reflectie op het teamwork

Alrian - Teamwork ging naar mijn mening best goed. We spoorden elkaar aan om te werken en niemand was contraproductief bezig. Wat wel miste was een initiatiefnemer die de groep in één richting moet sturen, hierdoor was iedereen met van alles bezig wat uiteindelijk een beetje een rommel werd. In de laatste paar dagen hadden we dit gecorrigeerd, waardoor we uiteindelijk als een echt team aan de slag gingen.

Britt - Het teamwork was aan begin jammer genoeg wel stroef gelopen, maar dit is in de laatste weken rechtgetrokken. De planning hadden we aangepast waardoor het beter ging. Verder was e sfeer binnen de groep was wel goed, erg aangenaam en ook gezellig. Daarnaast heeft iedereen elkaar goed geholpen.

Lex - Het werken met deze mensen was voor mij nieuw. Ik kende alleen Marty en met Muhammad, maar het teamwork met de anderen gingen terugkijkend redelijk tot goed. Ik heb het erg gezellig met ze gehad. Het samenwerken was bij het begin rommelig, iedereen was namelijk met zijn eigen visualisatie/opdracht bezig en communiceerde weinig. We merkten na een tijd dat dit niet meer werkte en hebben verschillende stappenplannen gemaakt om dit te verbeteren. Dit is uiteindelijk gelukt en ik kijk terug op een (redelijk) geslaagd project.

Marty - Ik was zelf af en toe afwezig. Dat komt doordat ik naast het project ook druk bezig met mijn werk was. Maar dit weerhield me er niet van om in de weekenden extra hard door te werken, dit was echter wel lastig te combineren, maar ik denk dat het me uiteindelijk wel is gelukt. Over het teamwork was het voor mij af en toe lastig. Ik communiceerde eerst niet zo goed met wat ik deed, ik was voornamelijk gefocussed op mijn "eigen" design, maar ik heb dit "eigen" design uiteindelijk losgelaten en anderen toegelaten om er samen aan te werken. Dit is naar mijn mening samen met het proces van teamwork geslaagd.

Muhammad - In het begin verliep het wat langzaam. Dit kwam doordat we geen planning hadden. In de derde week verliep de samenwerking beter en kwam er vooruitgang in het werk, doordat iedereen wist wat er van hem of haar verwacht werd. Door het opstellen van deadlines en een werkverdeling kon iedereen aan zijn of haar stukje werken zonder dat er overbodig werk werd verricht. Ik ben zeer tevreden met de samenwerking met deze groep.

Siyawash - Het begon allemaal wat stoef. Er werd niet goed gecommuniceerd en hierdoor was de samenwerking niet optimaal. Gelukkig is hier na mate we veder kwamen in het project steeds meer verbetering in gekomen. In de laatste week is alles perfect gegaan. Iedereen wist wat zijn of haar taken waren en hielp zo nodig andere.

Referentielijst

1. García, O. (2010). Language Spread and Its Study in the Twenty-First Century.
2. Girelli, S., Grotto. E., Lodi, P., Lupatini, D. & Patuzzo, E. (2013). Most Spoken Families. Op internet: http://www.puffpuffproject.com/languages.html, geraadpleegd op 22 juni 2016.
3. Gruver, A., & Dutton, J. A. (2014). Graduated and Proportional Symbol Maps. Op internet: https://www.e-education.psu.edu/geog486/node/1869, geraadpleegd op 22 juni 2016.
4. Heer, J., Bostock, M., & Ogievetsky, V. (2010). A Tour Through the Visualization Zoo. *Communications of the ACM*, *53*, 59-67.
5. Jean-Daniel Fekete, Jarke J. Wijk, John T. Stasko, and Chris North. 2008. The Value of Information Visualization. In Information Visualization, Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North (Eds.). Lecture Notes In Computer Science, Vol. 4950. Springer-Verlag, Berlin, Heidelberg 1-18.
6. WildInWoods. (2011). Infographics and data visualization: new look on the world. Op internet: https://wildinwoods.wordpress.com/2011/01/07/infographics-and-data-visualization-new-look-on-the-world/, geraadpleegd op 22 juni 2016.
7. Paul, L. M., Simons, G. F., & Fennig, C. D. (2016). *Ethnologue: Languages of the World*. Geraadpleegd op 22 juni 2016, van http://www.ethnologue.com.