Verwerking van studiedata aan De HHs

Beschrijving en verantwoording

let's change
YOU. US. THE WORLD.

Handleiding, v. 0.9.0
Lectoraat Learning Technology & Analytics, De HHs | 09-07-2023

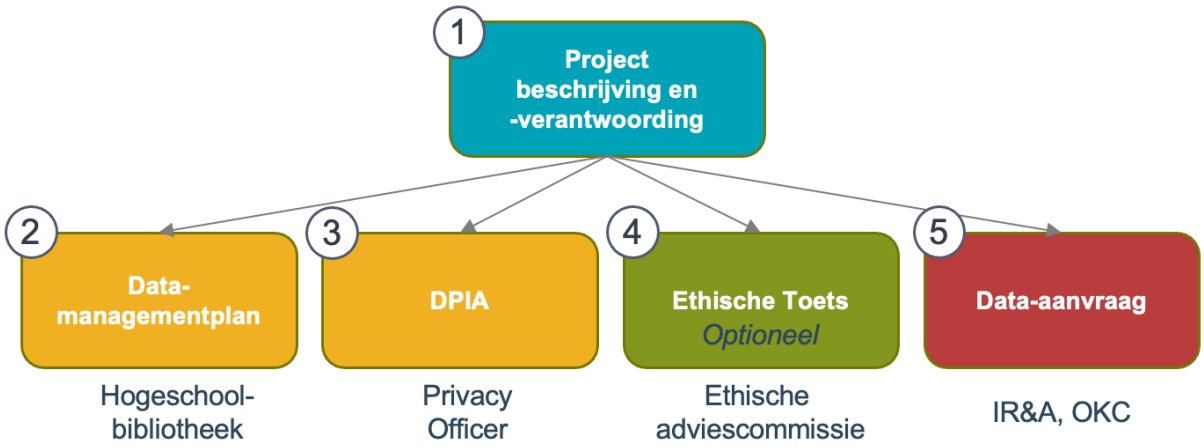
DE HAAGSE HOGESCHOOL

Inhoudsopgave

1 Inleiding

1.1 Verwerking van studiedata aan De HHs

In het lectoraat Learning Technology & Analytics onderzoeken we studiedata van De HHs. Hiervoor verwerken we studiedata die door de dienst Onderwijs, Kennis- en Communicatie (OKC) vanuit verschillende informatiesystemen gecombineerd en geleverd kunnen worden. Voor de levering aan het lectoraat dient een aantal formele stappen doorlopen te worden.



1.2 Doel van deze handleiding

Deze handleiding van het lectoraat Learning Technology & Analytics beschrijft en verantwoordt de procedures en methode om deze studiedata aan te vragen en te verwerken. Bij subsidieaanvragen dienen deze documenten meestal meegeleverd te worden.

Ook vind je een overzicht van de tools die we gebruiken, hoe die met elkaar samenhangen en hoe je ze gebruikt. De broncode van deze handleiding staat op github en kan door leden van het lectoraat bewerkt worden.

1.3 Suggesties

Heb je suggesties voor verbeteringen? Mail die dan naar de lector, Theo Bakker, t.c.bakker@hhs.nl.

2 Procedures voor levering van studiedata

2.1 Levering aan het lectoraat

Voor de levering van studiedata dienen een aantal procedures doorlopen te worden en bijpassende documenten opgesteld te worden:

1. Projectbeschrijving en -verantwoording
2. Datamanagementplan
3. DPIA
4. Eventueel een toets bij de Ethische Adviescommissie
5. Formele data-aanvraag bij het team IR & Analytics, OKC.

> ! Belangrijk
>
> Stem voordat je een of meerdere van deze documenten formeel instuurt, deze eerst inhoudelijk af met de lector.

2.2 Projectbeschrijving

De projectbeschrijving bevat:

1. Een **toelichting op het belang van het onderzoek:** voor studenten en docenten of andere stakeholders in onze hogeschool, de doelstellingen van De HHs, de onderzoekslijnen van het lectoraat, de onderzoeksthema's van het kenniscentrum Global & Inclusive Learning en de onderzoeksthema's van De HHs.
2. Een **theoretische en ethische onderbouwing**. De theoretische onderbouwing bevat een link met het vakgebied en eerdere publicaties; de ethische onderbouwing maakt gebruik van de Data Science Ethics Checklist van Deon die ingevuld als bijlage wordt toegevoegd.
3. Een uitwerking van de **onderzoeksvragen en operationalisering** daarvan.
4. Een uitwerking van de **databronnen.**
5. Een beschrijving van de **levering en verwerking**.
6. De **methoden van analyse**.
7. De **verwachte resultaten**.
8. Een **toelichting op de reproduceerbaarheid** van het onderzoek volgens de FAIR principes.
9. Bijlagen
    1. **Referenties**
    2. Een **versiegeschiedenis** met wijzigingen en verspreiding.
    3. De locatie van de **repository** voor de projectbeschrijving.
    4. De ingevulde **Data Science Ethics Checklist**

Laat de projectbeschrijving lezen door een aantal collega's zodat deze aan kwaliteit wint. Geef bij een verzoek voor een review expliciet aan in het verzoek dat het een formeel verzoek is, wanneer je een reactie hoopt te ontvangen en hoe je de feedback zult gaan gebruiken.

Template: Gebruik voor de projectbeschrijving het markdown template van het lectoraat.

2.3 Datamanagementplan

De hogeschoolbibliotheek van de dienst OKC heeft datamanagement in de portefeuille. Voor het datamangementplan gebruiken we een standaard formulier van NWO. Het datamanagementplan beschrijft:

1. **Algemene gegevens**: informatie over het project, de aanvrager, de datastewart
2. **De onderzoeksdata**: welke data worden verzameld of geproduceerd, en welke bestaande data worden hergebruikt?
3. **Metadata**: welke metadata en documentatie worden meegeleverd met de data?
4. **Opslag en back-up**: Hoe worden data en metadata opgeslagen en geback-upt tijdens het onderzoek?
5. **Bewaartermijnen en archivering**: Hoe en wanneer worden data voor langere tijd gedeeld en bewaard?
6. **Datamanagementkosten**

Omdat het hier om de beveiliging van data gaat, is het goed het datamanagementplan voorafgaand aan de DPIA op te stellen aangezien dit type vragen ook terugkomt in de DPIA.

Je bespreekt het datamanagementplan met een medewerker van de hogeschoolbibliotheek met wie je een afspraak kan inplannen via research@hhs.nl.

Template: Gebruik voor het datamanagementplan het standaard NWO formulier.

2.4 Data Privacy Impact Assessment (DPIA)

De inschrijvingen of resultaten dataset kan na positieve afronding van de DPIA geanonimiseerd aangeleverd worden door OKC. Herkenbare gegevens, zoals voor- en achternaam, geboortedatum, adres zijn verwijderd en het studentnummer is vervangen door een willekeurige tekst, waarop beide sets wel gekoppeld kunnen worden.

Geanonimiseerde gegevens vallen formeel niet onder de Algemene Verordening Gegevensbescherming (AVG). Maar omdat datasets met veel gegevens mogelijk toch herleidbaar is tot een student, spreken we liever van gepseudonimiseerde data en voeren we toch een Data Privacy Impact Assessment (DPIA) uit.

Template: De DPIA wordt uitgevoerd aan de hand van de DPIA vragenlijst van De HHs.

2.4.1 Keuze voor een grondslag

Lees voor meer informatie over privacy, ethiek en studiedata het Referentiekader privacy en ethiek voor studiedata.

De keuze voor de grondslag voor de verwerking van studiedata voor het onderzoek en de onderbouwing daarvan zijn cruciaal. De grondslag voor de verwerking door het lectoraat is het **gerechtvaardigd belang** van het onderzoek voor De Haagse Hogeschool of – uitsluitend bij onderzoek met proefpersonen of bijzondere persoonsgegevens[1] – de **toestemming** van studenten. Zodra je op basis van data een advies geeft aan studenten heb je toestemming nodig van de student om zijn/haar/diens persoonsgegevens te verwerken.

2.4.2 Grondslag: Gerechtvaardigd belang

Voor de grondslag gerechtvaardigd belang zijn 3 voorwaarden:

Voorwaarde 1: gerechtvaardigd belang

> Uw belang is echt een gerechtvaardigd belang als het ergens in het recht is opgenomen. En wordt erkend en beschermd. Dat mag ook in een ongeschreven rechtsregel of rechtsbeginsel zijn. Als het maar gaat om een belang waarvan we in de maatschappij vinden dat het door het recht beschermd moet worden.

Hier geldt het belang van wetenschappelijk onderzoek naar onderwijs in het algemeen en het onderwijs van De Haagse Hogeschool in het bijzonder.

Voorwaarde 2: noodzakelijkheid

> Heeft u daadwerkelijk een gerechtvaardigd belang? Dan moet u vervolgens kijken of de verwerking van persoonsgegevens noodzakelijk is om dit belang te behartigen. Dit doet u door na te gaan:
>
> - Of het doel van uw verwerking in verhouding staat tot de inbreuk op de privacy van de betrokkenen. In de AVG heet dit ‘proportionaliteit’.
> - Of u het doel niet op een andere manier kunt bereiken, die minder ingrijpend is voor de betrokkenen. In de AVG heet dit ‘subsidiariteit’.

- *Proportionaliteit* - Voor ons onderzoek gebruiken we gepseudonimiseerde data, waarbij we zo min mogelijk gegevens verzamelen. Bovendien proberen we met behulp van statistische analyses en machine learning het aantal variabelen te reduceren tot die variabelen die significant zijn of substantieel bijdragen aan accuraatheid van voorspellingen.

[1] Bijzondere persoonsgegevens zijn iemands gegevens over ras of etniciteit, politieke opvatting, religieuze of levensbeschouwelijke overtuigingen, lidmaatschap van een vakvereniging, gezondheid, seksueel gedrag of seksuele gerichtheid, genetische gegevens en biometrische gegevens voor unieke identificatie.

- *Subsidiariteit* - Er is geen mogelijkheid deze gegevens op een andere manier te verzamelen.

Voorwaarde 3: afweging belangen

> Heeft u een gerechtvaardigd belang en is de gegevensverwerking noodzakelijk om dit belang te behartigen? Dan moet u tot slot een afweging maken tussen uw belangen en de belangen van de betrokkenen.
>
> Bij deze afweging kijkt u naar:
>
> - de gevolgen voor de betrokkenen;
> - hoe ernstig de inbreuk is op de privacy van de betrokkenen; welke (aanvullende) maatregelen u heeft genomen om ongewenste gevolgen voor de betrokkenen te voorkomen of beperken;
> - of de betrokkenen de verwerking min of meer kunnen verwachten. Bijvoorbeeld als vervolg op een eerdere verwerking waarvoor zij toestemming hebben gegeven of als vervolg op verwerkingen die noodzakelijk zijn om een contract uit te voeren.

In de projectbeschrijving wordt de afweging van belangen beschreven.

2.4.3 Grondslag: Toestemming

Toestemming wordt ook wel *informed consent* genoemd. Hierbij geeft de student akkoord op de verwerking van diens persoonsgegevens. Dit is van toepassing bij bijzondere persoonsgegevens of bij profiling, waarbij de student op basis van persoonsgegevens een advies wordt gegeven of een interventie krijgt aangeboden. Denk aan deelname aan een gericht experiment, bijv. de toepassing van een algoritme voor een prognose.

2.5 Toets bij de Ethische Adviescommissie

Een ethische toets is altijd van belang voor een project. Dit neem je op in je projectbeschrijving. Als je een onderzoek doet met proefpersonen of er een risico bestaat op negatieve gevolgen van je onderzoek, dan is het van belang dit voor te leggen aan de Ethische Adviescommissie van De HHs.

Voor de toets is er een checklist opgesteld die je ingevuld opstuurt samen met de projectbeschrijving. De checklist van de commissie bevat 1) vragen over de aanvrager, 2) de ethische adviesvragen, 3) het onderzoeksproject in het algemeen, 4) de uitvoering en methoden van het onderzoek, 5) de deelnemers/respondenten en 6) eventueel aanvullende vragen of opmerkingen.

De commissie vergadert 1 keer per maand. De data vind je op de webpagina van de commissie op het HHs intranet. Je stuurt je ingevulde checklist naar research@hhs.nl ter attentie van de ethische commissie.

> De Ethische Adviescommissie adviseert jou als onderzoeker over ethische aspecten ten aanzien van jouw onderzoeksproject. Het advies vindt bij voorkeur vóór of bij de start van jouw onderzoeksactiviteiten plaats, maar advies kan indien nodig ook gedurende het onderzoekstraject worden ingewonnen (bijvoorbeeld wanneer een journal je vraagt om een ethisch adviesrapport).
>
> De volgende vragen staan centraal bij de commissie:
>
> - Hoe worden de belangen van de deelnemers aan jouw onderzoek beschermd?
> - Worden deelnemers in jouw onderzoek voldoende en juist geïnformeerd over deelname?

Template: Gebruik voor een aanvraag aan de Ethische Adviescommissie de Checklist aanvraag Ethische Adviescommissie.

2.6 Data-aanvraag bij IR&A, OKC

Vanuit OKC is een **analyseset** gebouwd op basis van een aantal bronnen: Osiris, Studielink, het 1CHO bestand en een reistijden dataset. Deze dataset bevat data over alle inschrijvingen van studenten van De HHs vanaf cohort 2012. De methode voor het inlezen, verwerken en combineren van data is ontwikkeld door de Vrije Universiteit (VU) en overgenomen door De HHs. Zie voor een beschrijving van de variabelen het kwaliteitsrapport (uitgelegd in de stap Documenteren).

De data voor je onderzoek vraag je parallel aan bij het team IR & Analytics van OKC. Stem dit vooraf met de lector. Zodra er een akkoord is vanuit de privacy officer kan de data aan je geleverd worden. Een aanvraag doe je via Topdesk: BIV: Rapportages over studenten en onderwijs > tegel 'Nieuw overzicht aanvragen'.

Het formulier vraagt naar 1) algemene gegevens over de aanvrager, 2) of de gegevens die je aanvraagt persoonlijke gegevens bevat, 3) algemene informatie over je aanvraag: omschrijving, doel, welke gegevens je nodig hebt, periode, selectie, overige randvoorwaarden/wensen, wie er toegang krijgt tot de data, wanneer je de dataset nodig hebt.

Bij vraag 2 is er een optie voor 'Ja' of 'Nee, ze zijn geanonimiseerd en geaggregeerd'. Dit zijn antwoordopties die niet goed aansluiten bij de gepseudonimiseerde data op studentniveau die we in het lectoraat onderzoeken. Ik heb het IR&A team gevraagd dit aan te passen. Kies voor nu de 2e optie met een toelichting.

Niet alle data zijn al beschikbaar: enkel inschrijvingen en data over uitval of diplomering van de cohorten 2012 tot en met 2022. Een resultatendataset is in voorbereiding en gepland om voor de zomer af te ronden. Data uit Brightspace of andere systemen is nog niet ontsloten.

Template: Gebruik het formele aanvraagformulier

3 Software

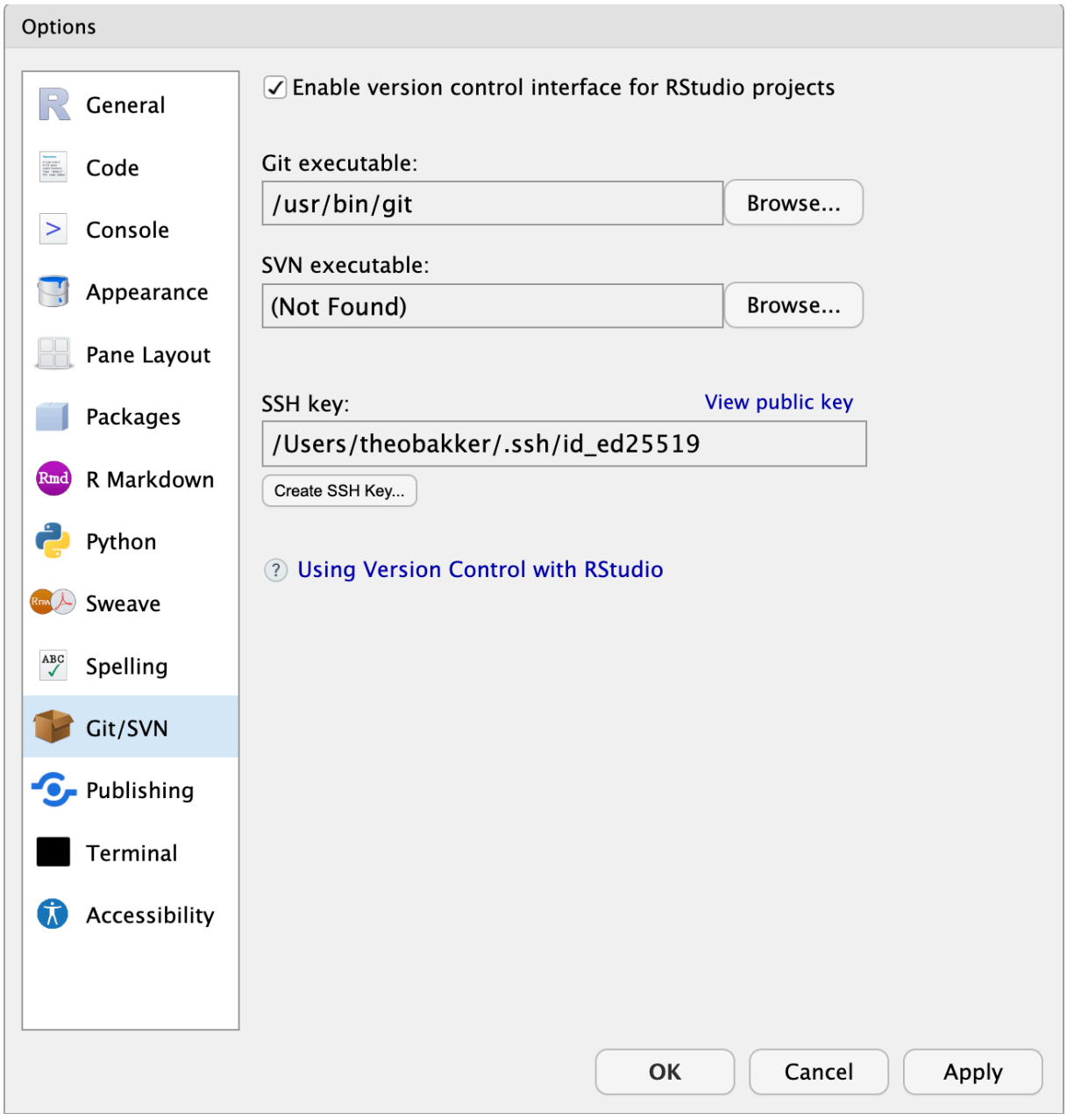
Voor het bewerken van de data en de ontwikkeling van de analyses gebruiken we open source software: R, RStudio, Git en QGIS. Voor de data van een Researchdrive.

3.1 Data

- De **onderzoeksdata** staat op Researchdrive. Voor deelonderzoeken maken we gebruik van subsets van de hoofddataset die door de lector worden gemaakt.

3.2 Broncode

- **Broncode** leggen we vast en delen we via Git (Github).
- Het **copyright** van alle broncode, analyses, rapportages en publicaties die je ontwikkelt binnen je aanstelling voor het lectoraat ligt bij De HHs en daarbinnen bij het kenniscentrum Global & Inclusive Learning en daarbinnen het lectoraat.
  - Git is een manier om de code voor de analyses van een project te scheiden van de data en de output. Neem voor je begint de volgende tutorial door: GitHub and Git Tutorial for Beginners**.**
  - Het lectoraat heeft een Github-omgeving waarop alle repositories staan: https://github.com/LTA-HHs. Alle repositories beginnen met 'lta-hhs-'.
  - Check na grote wijzigingen je wijzigingen in met een kort commentaar, zodat jijzelf of je collega's na verloop van tijd weten wat de wijziging was. Het kan zijn dat dit meerdere keren per dag gewenst is. Gebruik als vuistregel dat wat je zou moeten kunnen terugdraaien een logisch geheel is.
- We gebruiken de `ssh` standaard (Secure Shell) voor **repositories**.
  - De repository voor de inschrijvingen analyseset is 'git@github.com:LTA-HHs/lta-hhs-analyseset.git'.
  - De repository voor deze handleiding is 'git@github.com:LTA-HHs/lta-hhs-analyseset-manual.git'.
  - Maak in RStudio een eigen ssh-key aan via `Tools` > `Global Options` > `Git/SVN` > `Create SSH key`. Sla deze op op je harde schijf. Maak per computer die je gebruikt een unieke key. Zie figuur 1 voor de instellingen.
  - Deel de key met de lector, die de sleutel kan toevoegen aan de repository, zodat je via Git de code kan ophalen, bewerken, committen en uploaden. Zie de tutorials voor een verdere uitleg.



Figuur 1: Instellingen voor Git met SSH

3.3 Bewerking & analyse

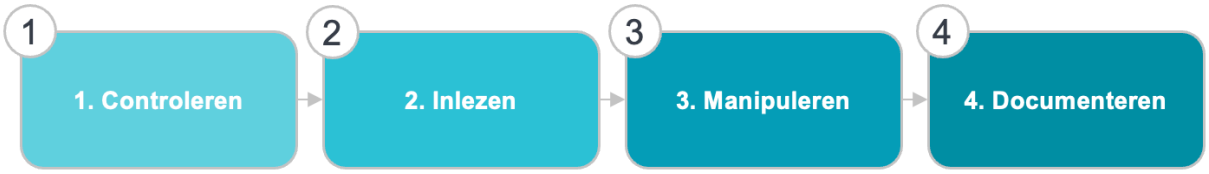
- **Studiedata** bewerken en analyseren we met R met als editor R Studio van Posit; bij uitzondering kan ook Python gebruikt worden. Iedere analyse wordt als project opgesteld.
  - Maak gebruik van de huisstijlgids voor code die het lectoraat hanteert (op te vragen bij de lector).
  - Gebruik de LTA-snippets om efficiënt te kunnen coderen (op te vragen bij de lector).
  - Neem vooraf de tutorial over RStudio door: RStudio Tutorial. Dit zal je in de uitvoering veel tijd besparen.
  - Zie voor de trainingen die je moet doen om R aan te leren **6.1 Data engineering en data analyses** van de Onboarding handleiding voor het lectoraat.
- **Geografische data** analyseren en visualiseren we met QGis. Het is mogelijk in R, maar de mogelijkheden daarin zijn beperkt.
  - QGis is een pakket met heel veel mogelijkheden. Overleg voordat je start met de lector wat je precies ermee gaat doen en wat nuttig is om aan te leren.
  - Voor het maken van geografische analyses is een extra handleiding in de maak.

3.4 Rapportages

- De **opmaak** van formele documenten, zoals handleidingen, artikelen, rapporten, automatiseren we zoveel mogelijk buiten MS Office om.
  - We maken gebruik van R, Quarto/Markdown of Latex. Al deze pakketten zijn geprogrammeerd op R of communiceren daarmee.
  - Een goede Latex editor is Texstudio.
  - Als je toe bent aan je eerste rapport, leer dan eerst Quarto aan met de tutorial: Tutorial: Hello, Quarto.

4 Prepareren van de data

Om de data te kunnen analyseren prepareren we deze eerst: 1) we vergelijken nieuwe data met eventuele eerdere leveringen, 2) lezen de nieuwe data in en passen veldnamen aan, 3) voegen velden toe en verwijderen overbodige velden en 4) maken een documentatiebestand.



Figuur 2: Stappen voor het prepareren van data

Voor de datapreparatie is een git project aangemaakt: lta-hhs-analyseset.

4.1 Voorbereidingen

Plaats het geleverde bestand in de folder 'AS_Inschrijvingen/01. Ruwe data' in een map die de datum van het bestand weergeeft: YYYYMMDD. Een levering van 5 juli 2023 wordt 'AS_Inschrijvingen/01. Ruwe data/20230705'. Voeg bij een tweede levering op dezelfde dag een volgnummer toe '-2'.

4.2 Controleren

01. Vergelijken AS_LTA.R - In deze stap en dit script vergelijken we de oude en nieuwe dataset. Het doel is verschillen in kaart te brengen en daarover te kunnen rapporteren. Als deze te groot zijn dienen die gemeld te worden aan het team IR&A van OKC.

1. Pas boven in de pagina **configuratie** van de naam en datum van levering van het oude en nieuwe bestand aan, zodat je de juiste bestanden vergelijkt.
2. Gebruik het `dataCompareR` package dat rapporteert op **inhoudelijke verschillen**, zoals namen van variabelen en hun types, inhoud van categoriale variabelen en rijen. Het corrigeert automatisch veldnamen die syntactisch onjuist zijn.
3. Test op percentages **missende waarden**. Deze dienen over het algemeen minder te worden; voor succesvariabelen, zoals BSA of Uitval, kunnen ze juist meer worden omdat die data nog niet beschikbaar is.
4. Vergelijk **aantallen studenten per faculteit, per opleiding per jaar**.
5. Sla de uitkomsten tussentijds op en koppel deze indien nodig terug aan het team IR&A van OKC.

4.3 Inlezen

02. Inlezen AS_LTA.R - In deze stap en dit script lezen we — als de data kwalitatief goed genoeg is — de data in.

1. **Lees de ruwe data in**. Pas in de configuratie de nieuwe datum van levering aan en indien nodig de nieuwe bestandsnaam.
2. **Pas veldnamen aan** waar nodig op basis van een documentatiebestand van het lectoraat.
3. **Sla het bestand op** in .fst en .rds formaat in de folder '02. Ingelezen data'.

4.4 Manipuleren

03. Manipuleren AS_LTA.R - In deze stap en dit script bewerken we de data.

1. **Lees de ingelezen data in**.
2. **Verwijder overbodige velden**.
3. **Maak nieuwe velden aan** waar nodig. Deze stap zullen we de komende maanden uitbreiden met de datasets / variabelen die we binnen het lectoraat zullen ontwikkelen.
4. **Sla het bestand op** in .fst en .rds formaat in de folder '03. Geprepareerde data'.

4.5 Documenteren

04. Documenteren AS_LTA.R - In deze stap en dit script documenteren we de data uitgebreid op basis van een achterliggend documentatiebestand en het `dataMaid` package.

1. **Lees de gemanipuleerde data in**.
2. **Lees het documentatiebestand in.** Dit is het documentatiebestand van het lectoraat met eigen veldnamen en labels.
3. **Maak de uitgebreide documentatie**. Dit wordt uitgevoerd door het `dataMaid` package. De uitkomsten worden automatisch opgeslagen in de folder 'Documentatie/Documentatie_LTA_uitgebreid'; het rapport opent zich in je browser. Zie Figuur 3.
4. **Controleer de uitkomsten.**

Kwaliteitsrapport HHs LTA Documentatie_LTA_uitgebreid

Autogenerated data summary from dataMaid

2023-07-07 21:03:27.322294

Data report overview

The dataset examined has the following dimensions:

Feature	Result
Number of observations	233527
Number of variables	186

Checks performed

The following variable checks were performed, depending on the data type of each variable:

	character	factor	labelled	haven labelled	numeric	integer	logical	Date
Identify miscoded missing values	×	×	×	×	×	×		×
Identify prefixed and suffixed whitespace	×	×	×	×				
Identify levels with < 6 obs.	×	×	×	×				
Identify case issues	×	×	×	×				
Identify misclassified numeric or integer variables	×	×	×	×				
Identify outliers					×	×		×

Please note that all numerical values in the following have been rounded to 2 decimals.

Codebook summary table

Label	Variable	Class	# unique values	Missing	Description
BSA	BSA_Advies	character	9	58.38 %	Bindend Studieadvies per collegejaar

Figuur 3: Voorbeeld output van het kwaliteitsrapport met dataMaid

Versiegeschiedenis

- 09-07-2023: versie 0.9.0 - eerste conceptversie

Repository

De code voor dit document kan bewerkt worden via GitHub. Leden van de kenniskring van het lectoraat kunnen op verzoek toegang krijgen en meewerken aan de inhoud.