Gratis online training helpt onderzoekers bij datamanagement

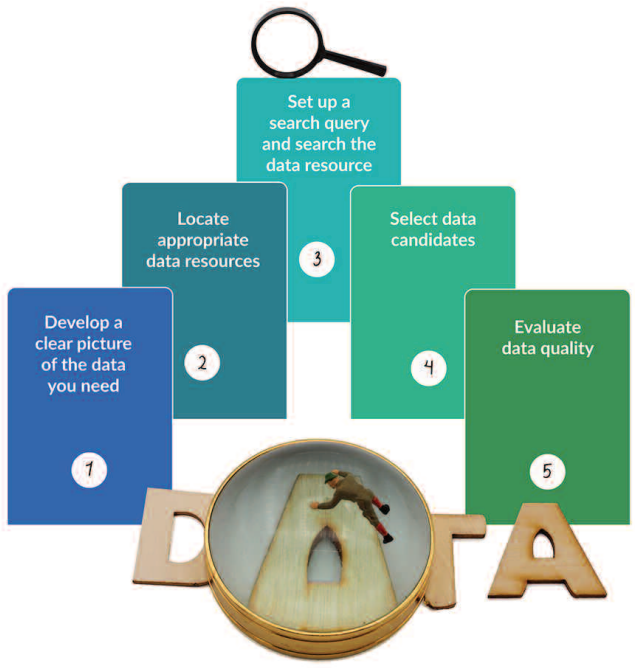
Leren hoe je data kunt vinden

CESSDA ERIC, het consortium van Europese sociaal-wetenschappelijke data-archieven, heeft aan haar gratis online training over datamanagement een nieuw hoofdstuk toegevoegd: data discovery.
Ricarda Braukmann

Vanaf 2017 telt de Data Management Expert Guide zes hoofdstukken over het management en hergebruik van sociaalwetenschappelijke data. Het bevat praktische tips voor het hele onderzoeksproces: over de planning van een onderzoeksproject, het organiseren van de dataverzameling, het verwerken van gegevens alsook het archiveren en publiceren van de onderzoeksdata (zo FAIR mogelijk).

Data life cycle

Aan de bestaande training is onlangs een hoofdstuk over het vinden van data (data discovery) toegevoegd, waardoor de training nu alle stappen van de data life cycle omvat. Dit nieuwe hoofdstuk biedt onderzoekers tips en trucs over het vinden van bestaande data, data die zij kunnen hergebruiken om nieuwe onderzoeksvragen te beantwoorden.



Bij het vinden van data zijn vijf stappen belangrijk: verkrijg een duidelijk beeld van de benodigde data, bedenk welke bronnen interessant kunnen zijn, zoek actief binnen dataverzamelingen, selecteer interessante datasets en als laatste stap: evalueer de kwaliteit en toepasbaarheid van de gevonden data.
credits Verbeeldingskr8t / CESSDA ERIC

Vijf stappen

Het hoofdstuk beschrijft vijf stappen in het vinden van data. De eerste stap gaat over de uitdaging om een duidelijk beeld te krijgen van de soort data die men wil vinden. Het hoofdstuk presenteert een aantal vragen om goed te definiëren naar welke data gezocht wordt. Vervolgens wordt een overzicht van mogelijke bronnen waar data gevonden kunnen worden, gegeven. Naast de CESSDA ERIC data-archieven die sociaalwetenschappelijke data vindbaar en toegankelijk maken, worden ook andere Europese en internationale databronnen toegelicht. Stap drie in het vinden van data is het actief zoeken binnen een archief of dataverzameling. Het hoofdstuk geeft een aantal tips voor het formuleren van effectieve zoekopdrachten waarmee bruikbare data gevonden kunnen worden. Als laatste wordt aandacht besteed aan het selecteren van datasets en aan het evalueren van de kwaliteit en toepasbaarheid van de data. Wat zijn vaak voorkomende toegangscategorieën? Waar mag ik de data voor gebruiken? Zijn ze beschikbaar in het juiste formaat? Zijn er kosten verbonden aan hergebruik? Hoe refereer ik aan de data? Dit zijn allemaal vragen waar het hoofdstuk onderzoekers mee op weg helpt om uiteindelijk de juiste dataset te kunnen selecteren.

De Data Management Expert Guide is online gratis beschikbaar.

cessda.eu/DMEG

Dr Ricarda Braukmann is programmaleider sociale wetenschappen bij DANS. DANS heeft als Nederlandse Service Provider van CESSDA ERIC bijgedragen aan de ontwikkeling van de Data Management Expert Guide.

GELEZEN

ESFRI Roadmap

Maarten Heerlien

In het najaar van 2018 publiceerde ESFRI, het European Strategic Forum on Research Infrastructures, een nieuw strategisch rapport, met daarin de landschapsanalyse van elk van de zes ESFRI-kennisdomeinen. Voor het SSH-domein (Social & Cultural Innovation) ziet het ESFRI-forum kansen in de verdere ontwikkeling van big-data-analyse in taaltechnologie. Godsdienstwetenschappen en doorontwikkeling van digitale diensten voor open science worden aangewezen als strategisch belangrijk voor SSH.
Aan de Roadmap 2018 zijn zes projecten toegevoegd, wat het totaal aan lopende ESFRI-projecten op 18 brengt. Voor twee van deze nieuwkomers fungeert Nederland als lead country. European Holocaust Research Infrastructure beoogt een onderzoeksinfrastructuur te ontwikkelen voor eenduidige toegang tot en analyse van geografisch verspreide bronnen over de Holocaust en wordt gecoördineerd door het NIOD. Distributed System of Scientific Collections richt zich op virtuele integratie en ontsluiting van Europese natuurhistorische collecties. Coördinatie van DiSSCo ligt bij Naturalis Biodiversity Center.
roadmap2018.esfri.eu

COLUMN

Een beleefde revolutie: differential privacy

Reden waarom ik van de statistiek houd, nummer 433: statistici zijn van die heerlijk ingetogen mensen. De hoogste lof die je als statisticus kan ontvangen is dat je 'voorzichtig' bent. En de meest negatieve reactie waar ik getuige van ben geweest, van een statisticus op een tochniet-zo-heel-zinnig plan van onze groep: een bedachtzame pauze - gevolgd door 'kán je doen...'. De drie puntjes waren hoorbaar, maar vergevingsgezind.
Dus als je in een vakblad leest, 'de bezorgdheid is reëel en het gevaar is ook reëel', dan let je op. Nu is dit citaat al uit 1972. Maar wel van Ivan Fellegi, een statisticus die zijn tijd ver vooruit was. Fellegi was een Hongaarse immigrant die na de opstand noodgedwongen naar Canada vluchtte, om daar allerlei briljante artikelen te schrijven over onderwerpen die vandaag de dag opeens zeer actueel zijn, zoals het koppelen van verschillende databestanden met onzekerheid. Hij werd ook de 'Hoofdstatisticus' van Canada, een titel die ik persoonlijk veel mooier vind dan 'dichter/theoloog/ramenlapper des vaderlands'. Wat dit vooral betekende is dat hij zich bemoeide met de officiële statistiek. En het gevaar waar hij zich druk om maakte? Privacy.

In de jaren '70 bestond de bezorgdheid over privacy slechts bij een paar helderziende individuen, die toen al inzagen dat de opkomst van computers en grote databestanden een nieuwe tijd inluidde. Sla nu maar eens een krant open zonder dat allerlei privacy horror stories je bespringen. Zelfs de politiek is wakker geworden, dus dan weet je zeker dat het allang uit de hand is gelopen.

Waarschijnlijk het gevaar, maar zeker de bezorgdheid, zijn nu uitgegroeid tot zo'n groot probleem, dat de Census Bureau (het Amerikaanse CBS) een drastische beslissing heeft genomen. Vanaf nu worden de resultaten van de volkstelling uitsluitend gepubliceerd met behulp van een statistische databeschermingstechniek genaamd "differential privacy".

Wat is dat nou weer? Welnu. Zelfs als je duidelijke 'identificatoren' - variabelen zoals naam, adres, postcode - uit een bestand verwijdert, blijkt het toch vaak mogelijk om personen te herleiden. Dit kan bijvoorbeeld door zo'n 'opgeschoond' databestand te koppelen aan andere bestanden die her en der te vinden zijn. Een beroemd voorbeeld is de heridentificatie van een aantal mensen uit een dataset met miljoenen Amerikanen die Netflix online beschikbaar had gesteld voor onderzoeksdoeleinden. Het CBS beschermt ons al sinds jaar en dag tegen dit soort praktijken, en speelt internationaal zelfs een leidende rol in het ontwikkelen van het soort databeschermingstechnieken waar Fellegi in 1972 over schreef.

Maar een groep informatici, aangevoerd door Cynthia Dwork van Harvard, was toch ontevreden. Ze bedachten een strenge, formele definitie van privacy en een set methoden om die te waarborgen: differential privacy. Het idee is simpel: stel, er moet een 'uitkomst' gepubliceerd worden. Dat kan een tabel zijn of een correlatie, maar ook een volledige dataset. Deze data worden niet lukraak op het internet geplempt, maar moeten eerst een verstoring ondergaan, bijvoorbeeld door er willekeurige ruis bij op te tellen. Als je uit deze verstoorde uitkomst niet met voldoende zekerheid kan bepalen hoe de oorspronkelijke dataset er uit zag, dan is er ook bijna geen kans op het herleiden van individuen. Je kunt zelfs niet goed bepalen óf een bepaalde persoon wel of niet in de oorspronkelijke dataset zat, ook al weet je verder letterlijk alles over die persoon.

Differential privacy is een fascinerend, maar controversieel, begrip. Open data wordt een fluitje van een cent, als je eraan kunt voldoen. Het nadeel is natuurlijk dat je door de verstoringen ook minder kunt met de data: er moet een balans gevonden worden tussen de bruikbaarheid en de bescherming van de data. Daarover wordt nu dan ook (voor statistische begrippen) fel gedebatteerd in Amerika. Is John Abowd, het hoofd van de Census, wel 'voorzichtig'?

De discussie komt ook naar ons land. Gevaar en bezorgdheid zijn er al. Differential privacy dient zich binnenkort vast ook aan in de Europese officiële statistiek, en in software voor onderzoeksdatabeheer zoals iRods, Dataverse, of Figshare. In het slechtste geval moet de sociale wetenschap op de schop: iets lastigere data-analyses, grotere steekproeven, meer preregistratie, en nieuwe onderzoeksontwerpen. In het beste geval zijn er binnenkort geen excuses meer om onderzoeksgegevens over mensen niet open te delen. "Goed te doen…"

Daniel Oberski

Daniel geeft de volgende column graag aan Pearl Dykstra.

Daniel Oberski
Daniel Oberski is universitair hoofddocent in methodologie van data science en statistiek aan de Universiteit Utrecht. Hij promoveerde in Tilburg en Barcelona en was visiting professor in Maryland. In 2014 ontving hij een Veni-subsidie voor het ontwikkelen van methoden die meetfouten in administratieve registerdata opsporen en corrigeren.