Hogeschool van Amsterdam

Steeds FAIRder

Verslag van het Urban Vitality zaaigeldproject 'FAIR: geen woorden maar data'

Onderzoeksgroep Mensen in Beweging - faculteit Bewegen, Sport en Voeding, faculteit Digitale Media en Creatieve Industrie en faculteit Gezondheid

Creating Tomorrow

Steeds FAIRder

Verslag van het Urban Vitality zaaigeldproject 'FAIR: geen woorden maar data'

Author
Niek van Ulzen[a], Anne de Jong[a], Marije Kanis[a], Rajaram Kaliyaperumal[b], Núria Queralt-Rosinach[b], Marco Roos[b], Ronald Cornet[c]
a. Hogeschool van Amsterdam, Centre of Expertise Urban Vitality
b. Leiden UMC, Humane Genetica
c. Amsterdam UMC, Klinische Informatiekunde

Department
Onderzoeksgroep Mensen in Beweging - faculteit Bewegen, Sport en Voeding, faculteit Digitale Media en Creatieve Industrie en faculteit Gezondheid

Date
30 oktober 2020

Project type
Urban Vitality zaaigeld 2020
Mede mogelijk gemaakt door SIA SPRONG projectnummer SVB/SPR.VG01.003

Version
1.0
https://doi.org/10.21943/auas.13148501

Executive summary

Bij praktijkgericht onderzoek van Urban Vitality (UV) moeten vaak gegevens uit verschillende databronnen en projecten gecombineerd worden. Dat varieert van nieuw gegenereerde gegevens van applicaties, vragenlijsten, sensoren, fysieke testen en activiteitenmeters tot bestaande gegevens van oude onderzoeksprojecten en gegevens opgeslagen in systemen van praktijkpartners. Eén van de grootste knelpunten, ook voor toepassingen op het gebied van Artificial Intelligence (**AI**[1]), is dat die gegevens nauwelijks vindbaar zijn, dat de voorwaarden onduidelijk zijn waaronder die gegevens toegankelijk zijn voor hergebruik, dat verschillende data(bronnen) niet met elkaar kunnen praten en dat gegevens onbegrijpelijk zijn voor zowel mens als machine door het ontbreken van documentatie. Hierdoor is koppeling en analyse inefficiënt. Een illustrerend voorbeeld met betrekking tot de vindbaarheid van gegevens binnen Urban Vitality is dat datasets van 67 van de 73 in 2019 gepubliceerde wetenschappelijke artikelen niet te vinden zijn in de **data repository UvA/HvA figshare**. Ook op grotere schaal is er sprake van suboptimaal data management en zijn data niet klaar voor AI-toepassingen: bij de huidige maar ook bij vorige virusuitbraken zijn wereldwijd waardevolle data nauwelijks vindbaar en toegankelijk, en al helemaal niet interoperabel en herbruikbaar [1].

Als UV onderzoeksdata geschikt wil maken voor AI, dan moeten deze data zoveel mogelijk FAIR [2] gemaakt worden. FAIR staat voor een viertal principes waarbij het gaat om data vindbaar (Findable), toegankelijk (Accessible), uitwisselbaar (Interoperable) en herbruikbaar (Reusable) te maken voor zowel mens (onderzoekers) als machines (computers). Vindbaarheid is als een etalage: zowel mens als machine kan zien wat er binnen te verkrijgen is. Toegankelijkheid wil niet zeggen dat de data open zijn, maar specificeert wie op welke manier en onder welke voorwaarden 'naar binnen' mag. Uitwisselbaarheid, of (semantische) interoperabiliteit, is als 'speed dating for machines': het is handig als gegevens uit verschillende bronnen een 'match' hebben en als er een gemeenschappelijke, universele (computer) taal is. Herbruikbaarheid wil zeggen dat er extra informatie aanwezig is waardoor de gegevens te interpreteren zijn, ook door een machine zonder menselijke tussenkomst. FAIR gaat over een verantwoorde omgang met onderzoeksdata en de principes zijn opgenomen in de Nederlandse gedragscode voor wetenschappelijke integriteit [3]. FAIR gaat ook over het ontwikkelen van toekomstbestendige data voor machines zodat die onderzoekers optimaal kunnen assisteren bij het doen van analyses. Met machines worden computers bedoeld en die zijn in zekere zin dom: computers moet alles expliciet, logisch en ondubbelzinnig verteld worden, want anders begrijpen ze het niet. Er is dus een universele computertaal nodig om de vindbaarheid, toegankelijkheid, uitwisselbaarheid en herbruikbaarheid van data te specificeren.

Hoewel het oorspronkelijke artikel uit 2016 over FAIR data [2] volgens Google Scholar meer dan 3000 keer is geciteerd (d.d. 17-09-2020), de FAIR-principes leidend zijn voor de European Open Science Cloud [4] en subsidiegevers FAIR ook omarmen, zijn er maar weinig praktijkvoorbeelden van FAIR-implementatie voor onderzoeksprojecten die niet in grote consortia worden uitgevoerd. Daarom hebben we het project 'FAIR: geen woorden maar data' opgezet. Het betrof een proof of concept waarbij de focus lag op het leren toepassen van de FAIR-principes en het verkennen van de implementatie ervan binnen het Mensen in Beweging programma van UV. Het hoofddoel was om te onderzoeken hoe we de FAIR-principes in kunnen zetten, en wat daar voor nodig is, om onderzoeksdata efficiënter en effectiever te (her)gebruiken.

Tussen februari en juli 2020 hebben data stewards, docenten, onderzoekers en lectoren van drie faculteiten van de Hogeschool van Amsterdam (HvA) samen met bioinformatici van Amsterdam UMC en Leiden UMC gewerkt aan dit project. We hebben met name de implementatie van de uitwisselbaarheid (de I in FAIR) en computer leesbaarheid getoetst aan de hand van een in de literatuur beschreven FAIRifcatie-

1 Vetgedrukte woorden worden uitgelegd in de **begrippenlijst**.

workflow [5] en een aangeleverde casus. De casus betrof een aantal reeds bestaande datasets uit eerdere (onderzoeks)projecten met gegevens over de lichaamssamenstelling van kwetsbare ouderen. Deze datasets moesten gecombineerd worden om een nieuwe onderzoeksvraag te kunnen beantwoorden.

We hebben geleerd dat *FAIRificatie* – het FAIR van maken van data – niet alleen een **datasteward**-activiteit is, maar dat met de huidige technische stand van zaken het een gezamenlijke inspanning vereist van datastewards, docent-onderzoekers (domeinexperts) en ICT-experts. Het omzetten van **(meta)data** van het oorspronkelijke dataformat naar een universele computertaal vergt momenteel een behoorlijke tijdsinvestering. Ter illustratie: het wordt al vrij snel complex om eenvoudige gegevens met een kolom over het proefpersoonnummer en een kolom over het geslacht te representeren in een formele, ondubbelzinnige en universeel toepasbare computertaal:

1) <https://www.hva.nl/rdf/person/1> rdf:type <http://purl.obolibrary.org/obo/NCBITaxon_9606>.
2) <https://www.hva.nl/rdf/person/1> sio:has attribute <https://www.hva.nl/rdf/person/1/gender_attribute>.
3) <https://www.hva.nl/rdf/person/1/gender_attribute> rdf:type <https://loinc.org/76691-5/>.
4) <https://www.hva.nl/rdf/person/1/gender_attribute> sio:has quality <https://www.hva.nl/rdf/gender/2>.
5) <https://www.hva.nl/rdf/gender/2> rdf:type <http://snomed.info/id/703118005>.

Deze representatie is misschien nauwelijks te begrijpen voor een docent-onderzoeker, maar een computer wordt hiermee expliciet duidelijk gemaakt dat 1) persoon 1 van het type 'homo sapiens' is zoals gedefinieerd in de **NCBI-taxonomie**; 2) deze persoon een attribuut 'gender identiteit' heeft; 3) dit attribuut van het type is zoals gedefinieerd in de **LOINC** ontologie; 4) en een kwaliteit met waarde twee heeft; en 5) dat deze waarde overeenkomt met 'feminine' zoals gedefinieerd in de **SNOMED CT** ontologie. De relaties in dit voorbeeld worden gedefinieerd met behulp van Semanticscience Integrated Ontology (*sio*:) en het **RDF**-schema (*rdf*:). Het zoeken van definities voor concepten en relaties op basis van **ontologieën** wordt voor alle relevante gegevens in de dataset gedaan en zo wordt een semantisch datamodel van gelinkte data ontwikkeld.

Door het voor een computer expliciet te moeten maken, moet een docent-onderzoeker ook bij zichzelf nagaan wat hij/zij eigenlijk precies wil weten. In het voorbeeld ging het blijkbaar niet om het biologische geslacht van iemand, maar om de gender identiteit. Ook worden coderingen ondubbelzinnig gedefinieerd: in het voorbeeld is geen verwarring over wat waarde 2 betekent. En het is belangrijk om te realiseren dat er niet per se overeenstemming tussen verschillende onderzoeksgroepen hoeft te zijn over een (nieuwe) coderingsstandaard: het is niet erg als in de ene database waarde 1 voor 'masculine' en waarde 2 voor 'feminine' staat, terwijl dat in de andere database andersom is; zolang de twee databases die concepten maar op dezelfde universele, ondubbelzinnige computer leesbare wijze vastleggen. Hoewel het momenteel dus een behoorlijke tijdsinvestering vraagt, vergroot deze exercitie de datakwaliteit en is het idee dat datahergebruik op deze manier toekomst- en AI-bestendiger en ook tijdsefficiënter wordt.

We hebben ook geleerd dat FAIR gaat over digitalisering van onderzoek. Onze implementatie van de FAIR-principes is gedeeltelijk gebaseerd op **semantic web** technologie en zowel datastewards als docent-onderzoekers kwamen met allerlei tools in aanraking die een bepaalde digitale vaardigheid vereisen, zoals: OpenRefine, Protégé, Ontology Lookup Service, semantische 'graph' databases, **Resource Description Framework** (RDF), **Turtle**, **SPARQL**, ShEX, Open Science

Framework, **UvA/HvA figshare**, Github, **Castor EDC** en DMPonline. Niet alleen datastewards en onderzoekers krijgen in toenemende mate te maken met deze digitale transformatie, maar ook onderzoeksdeelnemers en het bredere publiek. Er wordt een web van (complexe) informatie-systemen ontwikkeld en aan elkaar gekoppeld met soms gevoelige, persoonlijke gegevens. Dat kan juist leiden tot wantrouwen, vragen over privacy en zorgen over geautomatiseerde beslissingen. Participatieve en actieve methodes zoals citizen science en **datafysicalisatie** kunnen belangrijke manieren zijn om gebruikers, onderzoeksdeelnemers en het bredere publiek bij onderzoek te betrekken en zo inzichtelijker te maken hoe onderzoeksdata worden gebruikt en onderzoeksresultaten tot stand komen.

FAIR heeft ook te maken met duidelijkheid verschaffen over privacy. Dit is ons binnen het tijdsbestek van het pilotproject niet gelukt, laat staan op een computer leesbare wijze. We hadden te maken met (gepseudonimiseerde) persoonsgegevens en dus is het de verantwoordelijkheid van een docent-onderzoeker en de HvA als instelling, om de privacy van de personen in kwestie te waarborgen. Privacy is een grondrecht zoals is vastgelegd in de Algemene Verordening Gegevensbescherming (AVG). De vraag of en op basis van welke AVG-grond de gegevens mochten worden hergebruikt en gecombineerd voor verder wetenschappelijk onderzoek, hebben wij na circa 15 mensuren die er in dit project door verschillende experts aan zijn besteed, nog niet kunnen beantwoorden. Door deze onduidelijkheid zijn de gegevens nog niet herbruikbaar. Dit probleem geldt niet alleen voor de huidige casus, maar voor vrijwel elk UV-onderzoek.

Het project heeft geresulteerd in een aantal praktische richtlijnen, handleidingen, en in een herbruikbaar en verder te ontwikkelen **semantisch datamodel** [6]. Deze zijn openbaar en te vinden in **UvA/HvA figshare** [7]. Naast enkele concrete **tips en tricks voor docent-onderzoekers**, doen we op basis van de opgedane ervaringen in dit project de volgende aanbevelingen voor bestuur en management van de HvA en UV:

Aanbevelingen voor bestuur en management van de HvA en UV

Investeer in kennis en training over FAIR data – met name over semantische interoperabiliteit, linked data en kennisrepresentaties – bij zowel docent-onderzoekers als datastewards.

Stel in het kader van het onderzoeks- en open science beleid als doel dat de HvA een **FAIR enabling** organisatie wordt.

Neem FAIR data en open science mee in het denken over digitale transformatie binnen de HvA.

Ontwikkel met HvA research IT, HvA RDM support en/of HvA expertisecentrum applied AI FAIR Data Points via bestaande initiatieven zoals UvA/HvA figshare of Castor.

Zet een tijdelijke taakgroep op met een afgebakende opdracht om concrete HvA-procedures en -standpunten te ontwikkelen m.b.t. privacy en hergebruik van onderzoeksdata.

Onderschrijf de zes punten van het **Tada-manifest** over (ethisch) gebruik van data.

Beloon docent-onderzoekers die extra tijd hebben gestoken in FAIR of open science, met een prijs of door ze te benoemen tot FAIR / open science champion.

Verken hoe FAIR toegepast kan worden bij kwalitatief onderzoek.

Bouw per UV-thema of living lab FAIR datasets/werkwijzen op door met elkaar na te denken over welke data, meetprotocollen en -instrumenten veel voorkomen of relevant zijn.

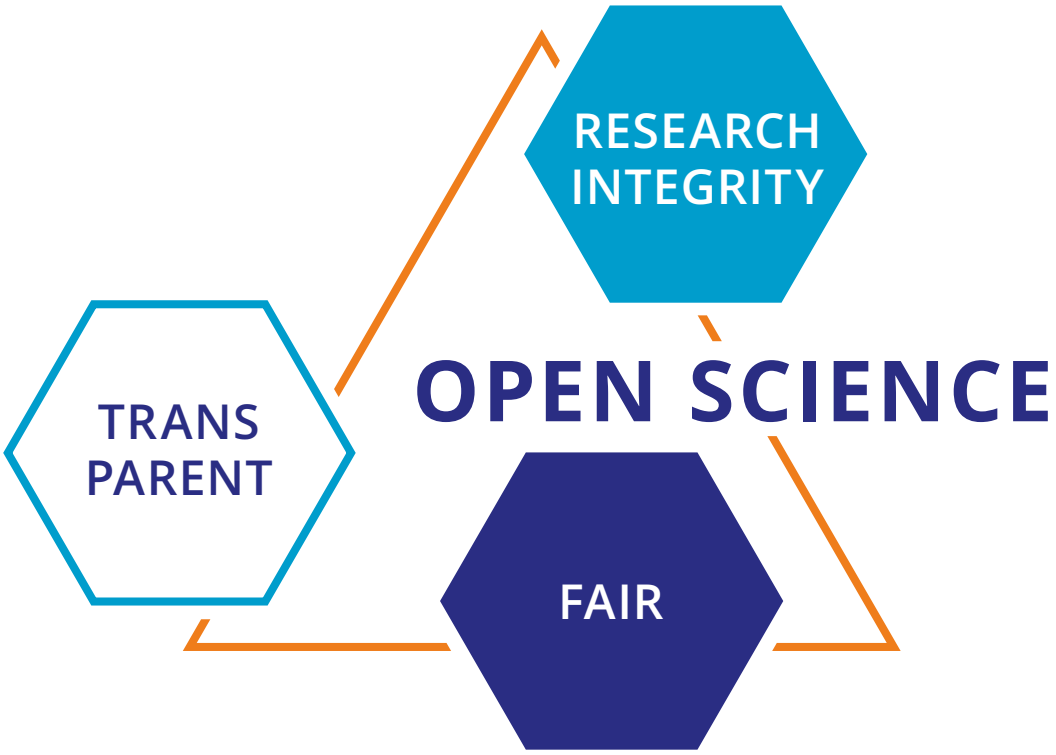
Neem in UV-beleid op dat elke dataset behorende bij een wetenschappelijke publicatie in UvA/HvA figshare gedeponeerd wordt.

Maak niet alleen data FAIR maar ook software, syntax en scripts met behulp van Git en de aanbevelingen op **fair-software.nl** van DANS en het Netherlands eScience Center.

Ontwikkel herbruikbare Castor-formulieren met FAIR metadata in de **Castor form exchange**.

Maak transparant(er) welke ‘standard operating procedures’ en meetinstrumenten Urban Vitality gebruikt en hoe dat te beheren, al dan niet i.s.m. **Meetinstrumenten in de zorg**.

Implementeer FAIR-principes op grotere schaal via (data)samenwerkingen waarbij UV al betrokken is zoals Sarphati Amsterdam, Sport Data Valley of Amdex.



Samengevat hebben we door het FAIRificatie-proces meer inzicht verkregen in onze data en dataverzamelingsmethoden en konden we beter aangeven wat bepaalde gegevens betekenden. Die betekenis hebben we ook, op een voor een computer begrijpelijke manier, kunnen representeren. Wel bleek met name het computer leesbaar maken van de data arbeidsintensief pionierswerk, mede doordat het internationaal volop in ontwikkeling is en er nog weinig gebruiks-vriendelijke tools zijn. FAIR-toepassing vergt een investering, maar maakt informatie begrijpelijk voor computers en kan koppeling en hergebruik van data voor zowel onderzoek als praktijk daarna sterk vereenvoudigen. Onduidelijkheid rondom (informed) consent en privacy bleek (nog steeds) een struikelblok voor het hergebruik van data te zijn. Het FAIR maken van data en bijbehorende services hoort bij de bredere ontwikkelingen van open, transparante wetenschap en digitalisering van onderzoek. Dit vraagt om nieuwe digitale vaardigheden voor zowel datastewards als onderzoekers, maar ook om een ethisch bewustzijn over de mens achter die data en waarom en waarvoor we die data verzamelen.

Inhoud

1. Inleiding

Bij praktijkgericht onderzoek van Urban Vitality (UV) moeten vaak gegevens uit verschillende databronnen en projecten gecombineerd worden. Dat varieert van nieuw gegenereerde gegevens van applicaties, vragenlijsten, sensoren, fysieke testen en activiteitenmeters tot bestaande gegevens van oude onderzoeksprojecten en gegevens opgeslagen in systemen van praktijkpartners. Eén van de grootste knelpunten, ook voor toepassingen op het gebied van Artificial Intelligence (**AI**[2]), is dat die gegevens nauwelijks vindbaar zijn, dat de voorwaarden onduidelijk zijn waaronder die gegevens toegankelijk zijn voor hergebruik, dat verschillende data(bronnen) niet met elkaar kunnen praten en dat gegevens onbegrijpelijk zijn voor zowel mens als machine door het ontbreken van documentatie. Hierdoor is koppeling en analyse inefficiënt. Een illustrerend voorbeeld met betrekking tot de vindbaarheid van gegevens binnen Urban Vitality is dat datasets van 67 van de 73 in 2019 gepubliceerde wetenschappelijke artikelen niet te vinden zijn in de **data repository UvA/HvA figshare**. Ook op grotere schaal is er sprake van suboptimaal data management en zijn data niet klaar voor AI-toepassingen: bij de huidige maar ook bij vorige virusuitbraken zijn wereldwijd waardevolle data nauwelijks vindbaar en toegankelijk, en al helemaal niet interoperabel en herbruikbaar [1].

Als UV onderzoeksdata geschikt wil maken voor AI, dan moeten deze data zoveel mogelijk FAIR [2] gemaakt worden. FAIR staat voor een viertal principes waarbij het gaat om data vindbaar (Findable), toegankelijk (Accessible), uitwisselbaar (Interoperable) en herbruikbaar (Reusable) te maken voor zowel mens (onderzoekers) als machines (computers). Vindbaarheid is als een etalage: zowel mens als machine kan zien wat er binnen te verkrijgen is. Toegankelijkheid wil niet zeggen dat de data open zijn, maar specificeert wie op welke manier en onder welke voorwaarden 'naar binnen' mag. Uitwisselbaarheid, of (semantische) interoperabiliteit, is als 'speed dating for machines': het is handig als gegevens uit verschillende bronnen een 'match' hebben en als er een gemeenschappelijke, universele (computer-) taal is. Herbruikbaarheid wil zeggen dat er extra informatie aanwezig is waardoor de gegevens te interpreteren zijn. FAIR gaat over een verantwoorde, transparante en reproduceerbare omgang met onderzoeksdata; de principes zijn daarom opgenomen in de Nederlandse gedragscode voor wetenschappelijke integriteit [3]. Maar FAIR maakt informatie ook begrijpelijk voor computers en kan koppeling en hergebruik van data voor zowel onderzoek als praktijk daarna sterk vereenvoudigen. Om data toekomstbestendig en klaar voor AI-toepassingen te maken, moeten die data computer leesbaar zijn en daarvoor is een universele computertaal nodig waarin informatie expliciet, logisch en ondubbelzinnig wordt gedefinieerd.

Hoewel het oorspronkelijke artikel uit 2016 over FAIR data [2] volgens Google Scholar meer dan 3000 keer is geciteerd (d.d. 17-09-2020), de FAIR-principes leidend zijn voor de European Open Science Cloud [4] en subsidiegevers FAIR ook omarmen, zijn er maar weinig praktijkvoorbeelden van FAIR-implementatie voor onderzoeksprojecten die niet in grote consortia worden uitgevoerd. Daarom hebben we het project 'FAIR: geen woorden maar data' opgezet. Het betrof een proof of concept waarbij de focus lag op het leren toepassen van de FAIR-principes en het verkennen van de implementatie ervan binnen het Mensen in Beweging programma van UV. Het hoofddoel was om te onderzoeken hoe we de FAIR-principes in kunnen zetten, en wat daar voor nodig is, om onderzoeksdata efficiënter en effectiever te (her)gebruiken. De aanpak was om met een mutlidisciplinaire projectgroep van docent-onderzoekers, datastewards en FAIR-experts:

1. Het FAIRificatie-proces beschreven door Jacobsen et al. [5] te doorlopen en te toetsen aan een casus. De focus lag met name op het definiëren van een **semantisch model** en het transformeren van de datasets in een formele, ondubbelzinnige, en universeel toepasbare computertaal;
2. Te onderzoeken welke stappen we met de huidige middelen binnen Urban Vitality kunnen zetten om zo FAIR mogelijk te komen.

2 Vetgedrukte woorden worden uitgelegd in de **begrippenlijst**.

Als casus konden we gebruik maken van 11 datasets afkomstig van Amsterdam UMC en de HvA over de lichaamssamenstelling van kwetsbare ouderen. Deze datasets moesten gecombineerd worden om twee methodes voor de schatting van eiwitbehoefte te kunnen vergelijken. Omdat het hier om een proof of concept ging hebben we data dus achteraf FAIR gemaakt; voor toekomstige projecten is het efficiënter om data meteen vanaf het begin, aan de bron, FAIR te maken. Deze rapportage beschrijft wat bereikt is (**Resultaten**), welke FAIR-onwikkelingen er zijn, FAIR-aanbevelingen voor HvA en UV en hoe docent-onderzoekers FAIR praktisch kunnen toepassen (**Hoe FAIRder?**). Als laatste gaat de rapportage dieper in op onze ervaringen met het FAIRificatie-proces en het semantisch modelleren, de onduidelijkheid over privacy en hergebruik van onderzoeksdata, en welke rol meer participatieve methoden kunnen spelen bij onderzoek dat steeds digitaler wordt (**Werkwijze en discussie**).

2. Resultaten

2.1 Deliverables

Het project heeft geresulteerd in een aantal praktische richtlijnen, handleidingen, en in een herbruikbaar en verder te ontwikkelen **semantisch datamodel** [6, 7]. Onderstaande tabel somt deze 'deliverables' op.

Tabel 1. Deliverables.

Wat	Waar
Beschrijving van het modelleer-proces en het semantische datamodel.	**Github**
Handleiding voor het deponeren van onderzoeksdata in de vorm van een data package om onderzoeksdata vindbaar en herbruikbaar te maken.	**UvA/HvA figshare**
Inventarisatielijst van gebruikte meetinstrumenten bij UV-onderzoek rondom kwetsbare ouderen.	**UvA/HvA figshare**
Een concept voor een minimum dataset voor het UV-thema Gezond ouder worden / Mensen in Beweging.	**UvA/HvA figshare**
FAIR tips en tricks voor UV-onderzoekers.	Dit document
FAIR aanbevelingen en ideeën voor HvA en UV.	Dit document
Gastles aan HvA-studenten van de minor software for science (HBO-ITC) over FAIR data en FAIR software.	Op aanvraag
Demonstratie van een lokale 'knowledge graph endpoint' met SPARQL-queries.	Op aanvraag
FAIR-voorstel voor samenwerking met Meetinstrumenten in de zorg.	Op aanvraag
Twee web-pagina's met FAIR-informatie.	**Digital Life** **UV Open science**
Een in UvA/HvA figshare gedeponeerde en vindbare dataset van de casus.	UvA/HvA figshare*

* Nog niet mogelijk vanwege onduidelijkheid over privacy.

2.2 Samen FAIRder

Ten eerste hebben we bij het **semantisch modelleren** geleerd dat multidisciplinaire samenwerking tussen onderzoekers (de domeinexperts), datastewards en FAIR-experts essentieel is. De onderzoeker is de inhoudsdeskundige en geeft aan welke informatie belangrijk is om vast te leggen en welke niet, wat de precieze betekenis is van bepaalde data en waarom voor een bepaald **meetinstrument** is gekozen. Datastewards vragen hierop door, zoeken naar ontologietermen en zorgen er samen met FAIR-experts voor dat die data in een formele, ondubbelzinnige, en universeel toepasbare

computertaal worden gerepresenteerd. Dit gaat in continue wisselwerking. Het was bijvoorbeeld niet eenduidig wat er in de datasets met 'geslacht' werd bedoeld: waren de onderzoekers geïnteresseerd in het biologische geslacht van een persoon of in de gender identiteit? Het zijn verschillende concepten die in de diverse **ontologieën** ook anders worden gedefinieerd. Dus als dit concept door een datasteward op ondubbelzinnige wijze gerepresenteerd moet worden, dan moeten onderzoekers aangeven wat voor hun onderzoek relevant is. Ook bleek dat er van eenzelfde meetinstrument geheel verschillende uitvoeringen in omloop zijn. Als dat soort informatie niet wordt vastgelegd, dan is de kans groter dat er in eventuele meta-analyses of bij het combineren van datasets appels met peren worden vergeleken.

Ten tweede geldt dat, met betrekking tot data, je ook samen verder komt bij de uitwerking van de Urban Vitality thema's. Verschillende voorheen gescheiden onderzoekdisciplines komen nu binnen een thema bij elkaar en dat leidt tot vragen/opmerkingen zoals 'hebben jullie bij jullie onderzoeksprojecten ook gegevens over x verzameld, want dat zou voor ons interessant kunnen zijn'. Als (beschrijvingen van) data vindbaarder zouden zijn, dan zou dat samenwerking en kennisuitwisseling bevorderen en zou het minder van 'toevallige' onderlinge gesprekken afhangen. Een volgende stap binnen een Urban Vitality thema zou kunnen zijn om bij onderzoeksprojecten relevante data op identieke wijze te verzamelen en te coderen om de interoperabiliteit te bevorderen (zie **Hoe FAIRder?**).

Ten derde worden onder impuls van het Dutch Techcentre for Life Sciences, ZonMw en consortia Health-RI en GO FAIR data steward communities opgezet om onderling (FAIR) kennis uit te wisselen en samen te bouwen aan FAIRdere infrastructuur en professionalisering van datastewards.

Manieren om ook onderzoeksdeelnemers en eindgebruikers bij de FAIRificatie (het FAIR maken van data) te betrekken, worden in paragraaf 4.4 omschreven.

2.3 Inzicht en overzicht

Door een diverse groep van docenten, onderzoekers, lectoren en datastewards van vier verschillende HvA-faculteiten bij het project te betrekken, hebben we met elkaar meer inzicht gekregen in wat er komt kijken bij FAIRificatie. Een inventarisatie van gebruikte **meetinstrumenten** bij het onderzoek met betrekking tot kwetsbare ouderen liet een range van meetinstrumenten zien. Van meetinstrumenten die in vrijwel elk project gebruikt werden tot meetinstrumenten die niet meer dan één keer voorkwamen. Het bleek dat soms verschillende versies of uitvoeringen van eenzelfde meetinstrument werden gebruikt, dat een test met dezelfde naam twee verschillende testen betrof, en dat er niet altijd (in het Nederlands) gevalideerde meetinstrumenten werden gebruikt. Dit gaf ons niet alleen een overzicht van de overlap tussen projecten, maar het benadrukte ook het belang van juist gebruik van en informatie over meetinstrumenten (versie, naam, referentie, status, etc.) voor het op verantwoorde manier hergebruiken van data (de R van FAIR).

2.4 FAIR in nieuwe projecten

In zeven projectaanvragen is aandacht gevraagd voor de FAIR-principes, gedeeltelijk (bij twee van de vijf aanvragen) doordat subsidiegever ZonMw specifiek om toepassing van de FAIR-principe en de inzet van datastewards vroeg. Eén van die vijf projecten is inmiddels gehonoreerd. In de MiB-projectaanvraag voor de UV stimuleringsgelden 2020 is FAIR-implementatie een afzonderlijk werkpakket. In paragraaf 3.3.3 worden aanbevelingen gedaan en ideeën aangedragen voor projecten of opdrachten om FAIR verder te ontwikkelen binnen UV.

3 Hoe FAIRder? Ontwikkelingen en aanbevelingen

3.1 (Inter)nationaal

Nationaal en internationaal is er steeds meer aandacht voor FAIR data en open science. FAIR Data en de Personal Health Train (PHT) zijn concepten die meer en meer worden toegepast voor het gebruik van gezondheidsdata. Zorginstituut Nederland en het Health-RI netwerk hebben recent PHT- en FAIR-voorbeelden van diverse praktijktoepassingen bij diverse zorg- en onderzoeksinstellingen verzameld [8]. Een wereldwijd Virus Outbreak Data Network (VODAN) van data-organisaties (waaronder **Castor EDC** waarmee UV een overeenkomst heeft afgesloten) werkt samen om data FAIR en AI-ready te maken omdat bij huidige maar ook bij vorige virusuitbraken, waardevolle data nauwelijks vindbaar en toegankelijk waren en al helemaal niet interoperabel en herbruikbaar [1, 9]. Voor ZonMw zijn FAIR data en open science twee belangrijke pijlers onder het subsidieonderzoek met betrekking tot COVID-19. Met ons pilotproject speelden we precies in op deze ontwikkelingen. En hoewel dit voorbeelden zijn met betrekking tot gezondheidsdata, is er ook buiten de life science en medische sector aandacht voor FAIR data [10].

3.2 FAIR, AI en digitalisering

FAIR data staat ook wel voor **F**ederated **AI-R**eady data. **Federated** data is een model waarbij benodigde data niet worden opgeslagen in één grote centrale database maar juist bij de bron (binnen de instelling) gehouden worden. Dit model is onder andere uitgewerkt in het Personal Health Train (PHT) concept [8]. Hierbij worden databronnen FAIR gemaakt (FAIR data stations) en analyses of algoritmen bezoeken die bronnen (FAIR data trains). Het idee is dat zo **privacy-by-design** beter mogelijk wordt gemaakt en dat bronhouders controle kunnen houden over en voorwaarden kunnen stellen aan het gebruik van de data. Deist et al. [11] beschrijven hoe ze van de PHT gebruik hebben gemaakt om databases van meer dan 20.000 longkanker-patiënten van acht zorginstellingen uit verschillende landen te koppelen en een statistisch model te trainen om de tweejaarsoverleving te voorspellen. Deze studie laat zien waarom FAIR ook voor AI-Ready staat: vaak zijn er voor (het trainen van) machine-learning algoritmen meer data nodig dan beschikbaar binnen één instelling en is koppeling van verschillende databases nodig. Deze koppeling wordt mogelijk / is efficiënter als de data bij de bron FAIR zijn. Voor AI is rekenkracht veelal niet het knelpunt, maar met name de vindbaarheid, toegankelijkheid en uitwisselbaarheid van data.

FAIR gaat ook over digitalisering van onderzoek en 'digital competence' van zowel datasteward als onderzoeker. In het project zagen we voor datastewards allerlei nieuwe tools en technieken voorbij komen, bijvoorbeeld: OpenRefine, Protégé, Ontology Lookup Service, semantische **knowledge graphs**, Resource Description Framework (**RDF**), **Turtle**, **SPARQL** en ShEX. En onderzoekers werden geconfronteerd met Open Science Framework, **UvA/HvA figshare**, Github, **Castor EDC** en DMPonline. Allerlei technieken en tools die een bepaalde digitale vaardigheid vereisen. FAIR, en in algemenere zin open science, gaat hand in hand met de verdere digitalisering van onderzoek.

3.3 Hogeschool van Amsterdam en Urban Vitality

Toch werpt de vraag zich op in hoeverre je als instelling op dit moment vol moet inzetten op FAIR data. Dat heeft vooral te maken met de tijd en inspanning die het nu nog vergt om (meta)data op een formele, ondubbelzinnige, en computer leesbare wijze te representeren. Die tijd en inspanning zijn misschien onevenredig hoog voor veel projecten [12] en het betreft een wat abstract doel waarvan de voordelen niet meteen zichtbaar zijn. Hier zijn tenminste twee argumenten tegen in te brengen. Ten eerste gaat FAIR maar voor een, weliswaar belangrijk, gedeelte over het computer leesbaar maken van (meta)data. Er is nog veel winst te behalen in het FAIR maken van data voor (andere) onderzoekers en dat heeft minder te maken met technische knelpunten. Het vindbaar en herbruikbaar maken van onderzoeksdata,

duidelijkheid scheppen over de voorwaarden waaronder data mogen worden hergebruikt en over de wijze waarop die data zijn verzameld, al dan niet op computer leesbare wijze, gaat welbeschouwd over goede onderzoekspraktijken en het reproduceer maken van onderzoeksresultaten. Mede daarom is het opgenomen in de Nederlandse gedragscode wetenschappelijke Integriteit [3]. Ten tweede is FAIR niet zwart-wit (óf wel óf niet FAIR), maar een spectrum. Per principe (F, A, I, R) kunnen data meer of minder FAIR zijn en meer of minder computer leesbaar zijn. Gegeven tijd en middelen kan een instelling dus wel zoveel mogelijk richting optimale FAIRness bewegen ('FAIR enough'). In onze optiek komt dat de kwaliteit en efficiëntie van onderzoek ten goede en beweeg je zo mee met internationale ontwikkelingen op het gebied van digitalisering van onderzoek en open science.

Hieronder hebben we kort op een rij gezet hoe, gegeven de huidige stand van zaken, datastewards ingezet kunnen worden, wat onderzoekers kunnen doen en wat UV kan doen om FAIRder te komen.

3.3.1 Inzet datastewards

In tabel 2 noemen we een aantal manieren waarop datastewards, bij projecten of binnen UV, ingezet kunnen worden om de FAIRness van data te vergroten.

3.3.2 Tips en tricks voor docent-onderzoekers

In tabel 3 staan enkele tips en tricks voor docent-onderzoekers om data FAIRder te maken. De stap om (meta)data in te zetten in een formele, on-dubbelzinnige, en universeel toepasbare computer-taal ontbreekt grotendeels, omdat dit extra inzet van datastewards met FAIR-expertise vereist.

Tabel 2. Inzet datastewards.

FAIR	Datasteward als consultant	Standaard, vraaggestuurd advies voorafgaand, tijdens en na afloop van het onderzoek. Bijvoorbeeld advies bij data management planning, archivering, openstellen van onderzoeksdata in UvA/HvA figshare en ondersteuning bij het gebruik van Castor. Op dit moment is dit de rol van datastewards binnen de UV Open science support desk.
FAIRder	Datasteward als projectlid	Meedenken en -helpen bij meer uitgebreide FAIRificatie van data als lid van het onderzoeksteam. Vaak ingegeven door subsidiegevers die verdergaande FAIRificatie en inzet van datastewards eisen (zie ZonMw) of een onderzoeksteam dat FAIRificatie van data dermate belangrijk vindt dat projectspecifieke inzet van een datasteward nodig is. Dit betekent dat tijd en middelen voor FAIRificatie in het project meegenomen moeten worden.
FAIRst	Datasteward als projectleider	Leiding geven aan een (werkpakket binnen een) project dat gericht is op FAIRificatie. Dat kan in een vorm, zoals bij dit zaaigeldproject, als specifieke opdracht van een management team of stuurgroep naar aanleiding van één van de aanbevelingen (zie onder) of als afzonderlijk werkpakket binnen een (data-intensief) onderzoeksproject.

Tabel 3. FAIR tips en tricks voor docent-onderzoekers. De laatste kolom geeft aan of het ook computer leesbaarheid betreft.

Tips en tricks	F.A.I.R.	Computer
Stel een (FAIR) Data Management Plan op in **DMPonline**.	FAIR	–
Kijk of er informatie over meetinstrumenten en bijbehorende gestandaardiseerde procedures te vinden is op **meetinstrumentenzorg.nl**.	I, R	–
Preregistreer je uitgewerkte onderzoeksprotocol/-plan en/of (statistisch) analyseplan.	F, R	–
Houd bij informed consent rekening met toekomstig hergebruik van de onderzoeksgegevens.	A, R	–
Als RDF niet tot de mogelijkheden behoort, gebruik dan zoveel mogelijk open dataformats zoals gespecificeerd door bijvoorbeeld **DANS**.	I	✶
Check de **Urban Vitality Open science research manual** om te kijken of er voor veel gebruikte variabelen standaard codering, terminologie, metadata of ontologieën zijn.	I	✶
Ga na of zelf-ontwikkelde software voldoet aan de **aanbevelingen** voor FAIR software.	FAIR*	–
Check of de **REFI-standaard** voor kwalitatieve analyse (QDA) geschikt is voor jouw project.	I	✶
Gebruik **UvA/HvA figshare** om:		
– een unieke 'persistent identifier' in de vorm van een DOI toe te voegen aan je dataset	F	✶
– metadata aan je dataset toe te voegen op basis van de '**Dublin Core**' standaard	F, I	✶
– een licentie aan je dataset toe te voegen	R	✶
– documentatie in de vorm van een **data package** aan je dataset toe te voegen	R	–
– in geval van persoonsgegevens, je dataset veilig te archiveren zonder toegang voor anderen, maar de metadata wel toegankelijk te maken.	F, A	–
Gebruik **Github** om (generiek gemaakte) scripts en syntax te delen met collega's.	R	–
Check de **Urban Vitality Open science checklist**.	FAIR**	–
Denk na welke informatie (metadata) van belang is om vast te leggen bij metingen en vragenlijsten voor replicatie van jouw resultaten of hergebruik van de data.	R	✶
Verzamel niet meer persoonsgegevens dan nodig voor het beschrijven van de onderzoekspopulatie en/of het beantwoorden van de onderzoeksvraag.	A	–
Sla direct-identificerende persoonsgegevens niet op bij de onderzoeksdata, maar apart daarvan met een code en koppeltabel.	A, R	–
Vraag de datastewards of in Castor gestandaardiseerde formulieren her te gebruiken zijn.	I, R	✶
Maak data niet alleen vindbaarder en toegankelijker voor onderzoekers en machines maar ook voor een bredere doelgroep; met name de mensen van wie je gegevens verzamelt. Bijvoorbeeld via fact sheets, infographics en datafysicalisatie (-projecten en workshops).	F, A***	–
Wees **open en transparant** over welke data je waarvoor gebruikt.	R	–

* De FAIR-principes zijn tot op zekere hoogte ook toe te passen op research software, maar de implementatie is anders [13].

** FAIR data is een onderdeel van open science.

*** FAIR is met name gericht op onderzoekers en computers, maar participatieve aanpakken als datafysicalisatie en citizen science zijn manieren om ook burgers en leken bij onderzoek te betrekken en op die manier vaak abstracte onderzoeksdata voor die doelgroep toegankelijker te maken.

3.3.3 Aanbevelingen en ideeën voor de HvA en UV

Op basis van de opgedane ervaringen in dit pilot-project komen we tot een aantal aanbevelingen en ideeën over hoe het bestuur of management van de HvA of UV het FAIR-thema verder zou kunnen oppakken, deels in de vorm van (een aanzet tot) opdrachtomschrijvingen voor datastewards.

Tabel 4. FAIR aanbevelingen en ideeën voor de HvA / UV.

Aanbeveling / idee	Voor
Investeer in kennis en training over FAIR data – met name over semantische interoperabiliteit, linked data en kennisrepresentaties – bij zowel docent-onderzoekers als datastewards.	HvA
Stel in het kader van het onderzoeks- en open science beleid als doel dat de HvA een **FAIR enabling** organisatie wordt.	HvA
Neem FAIR data en open science mee in het denken over digitale transformatie binnen de HvA.	HvA
Ontwikkel met HvA research IT, HvA RDM support en/of HvA expertisecentrum applied AI FAIR Data Points via bestaande initiatieven zoals UvA/HvA figshare of Castor.	HvA
Zet een tijdelijke taakgroep op met een afgebakende opdracht om concrete HvA-procedures en -standpunten te ontwikkelen m.b.t. privacy en hergebruik van onderzoeksdata.	HvA/UV
Onderschrijf de zes punten van het **Tada-manifest** over (ethisch) gebruik van data.	HvA/UV
Beloon docent-onderzoekers die extra tijd hebben gestoken in FAIR, of open science, met een prijs of door ze te benoemen tot FAIR / open science champion.	HvA/UV
Verken hoe FAIR toegepast kan worden bij kwalitatief onderzoek.	UV
Bouw per UV-thema of living lab FAIR datasets/werkwijzen op door met elkaar na te denken over welke data, meetprotocollen en -instrumenten veel voorkomen of relevant zijn.	UV
Neem in UV-beleid op dat elke dataset behorende bij een wetenschappelijke publicatie in UvA/HvA figshare gedeponeerd wordt.	UV
Maak niet alleen data FAIR maar ook software, syntax en scripts met behulp van Git en de aanbevelingen op **fair-software.nl** van DANS en het Netherlands eScience Center.	UV
Ontwikkel herbruikbare Castor-formulieren met FAIR-metadata in de **Castor form exchange**.	UV
Maak transparant(er) welke 'standard operating procedures' en meetinstrumenten Urban Vitality gebruikt en hoe dat te beheren, al dan niet i.s.m. **meetinstrumentenindezorg.nl**.	UV
Implementeer FAIR-principes op grotere schaal via (data)samenwerkingen waarbij UV al betrokken is zoals Sarphati Amsterdam, Sport Data Valley of Amdex.	UV

4 Werkwijze en discussie

4.1 Het team

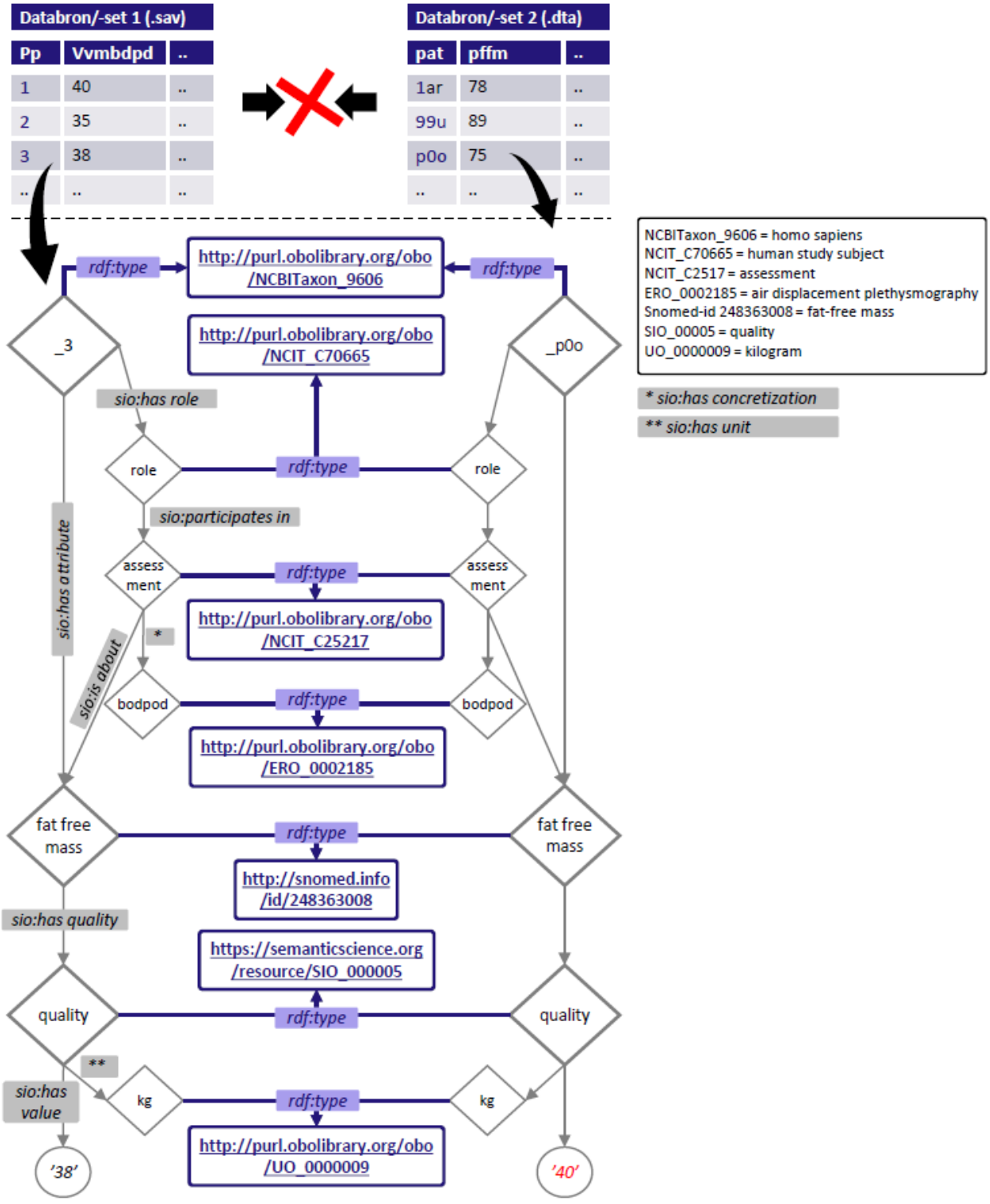
Het team bestond uit twee datastewards van de HvA en vier FAIR-experts van Amsterdam UMC en Leiden UMC die wekelijks bijeen kwamen, en een groep van domeinexperts die maandelijks werden bijgepraat en feedback gaven. Door de coronacrisis is vrijwel alles via telefoon, e-mail en online meetings gegaan. Voor de wekelijkse sessies met de FAIR-experts waren we oorspronkelijk fysiek op locatie bij Leiden UMC uitgenodigd om gezamenlijk met de BioSemantics-onderzoekgroep te kunnen brainstormen en modelleren. Als alternatief hebben we een Skype-kanaal opgezet om dit virtueel zoveel mogelijk na te bootsen (MS Teams was nog niet operationeel).

4.2 Het FAIRificatie-proces

Het gevolgde FAIRificatie-proces [5] bestaat uit een zevental stappen. In een voorbereidende fase hebben we het doel van de FAIRificatie gedefinieerd (stap 1) en de FAIR status bepaald van de data (stap 2) en **metadata** (stap 3) die als casus voor het pilotproject ter beschikking waren gesteld. In de eigenlijke FAIRificatie-fase hebben we met behulp van ontologieën een semantisch model ontwikkeld (stap 4) en de oorspronkelijke (meta)data getransformeerd naar een formele, ondubbelzinnige, en universeel toepasbare computertaal (stap 5) op basis van het in stap 4 ontwikkelde model. Vervolgens hebben we de oorspronkelijke en getransformeerde data beschikbaar gesteld in UvA/HvA figshare en in een lokaal FAIR Data Point (stap 6). In de afrondende fase hebben we het proces geëvalueerd (stap 7). Hoewel het wenselijk(er) is om data al zoveel mogelijk vanaf het begin 'aan de bron' FAIR te maken, hebben we voor de doeleinden van dit project FAIRificatie toegepast op reeds verzamelde datasets.

4.2.1 Stap 4 en 5: Semantisch modelleren en computer leesbare data

Na het goed in kaart brengen van de (FAIR status van de) data en **metadata** in stap 2 en 3 van het FAIRificatie-proces, hebben we een **semantisch model** ontwikkeld met behulp van **ontologieën** en met behulp van dit model de data omgezet in een computer leesbaar formaat. Het doel van deze stappen was om de data zodanig te representeren dat computers kunnen begrijpen wat de data betekenen (semantische interoperabiliteit). Dat houdt in dat de data, de concepten en relaties, ondubbelzinnig en op een logische manier moeten worden gedefinieerd, omdat computers niet uit de context op kunnen maken waar iets over gaat. Een computer weet bijvoorbeeld zonder expliciete definitie niet of een variabele 'HR' voor 'heart rate', 'hazard ratio' of voor 'human resources' staat. Daarom zijn er ontologieën. Het doel van een ontologie is om concepten en relaties tussen die concepten ondubbelzinnig te beschrijven met unieke 'identifiers'. Figuur 1 illustreert dit voor twee records uit twee datasets. De oorspronkelijke twee datasets (figuur 1 boven) zijn nauwelijks uitwisselbaar (twee bestandsformaten .sav en .dta van bedrijven die niet compatibel zijn) en het is onbekend wat de kolommen 'Vvmbdpd' en 'pfmm' zijn, wat de waarden voorstellen en of het betrekking heeft op mensen, dieren of auto's.
Na overleg met onderzoekers weten we dat 'Vvmbdpd' voor de vetvrije massa gemeten met een **Bodpod** staat en dat waarden in kilogram zijn, en dat 'pffm' voor fat-free mass staat, als percentage van het gewicht, en in dit geval ook gemeten met een **Bodpod**. We besluiten dat het handiger is om de vetvrije massa voor beide datasets in kilogram uit te drukken en de datastewards zoeken naar passende ontologieën om de concepten en relaties tussen die concepten op een formele, ondubbelzinnige, en universele manier te representeren. We specificeren dat de concepten 'Pp' en 'Pat' beiden om mensen gaat: 'Pp' (concept) is van het type (relatie) homo sapiens (concept), waarbij homo sapiens uniek gedefinieerd wordt door een 'Internationalized Resource Identifier' (IRI) uit de **NCBI-taxonomie**: **http://purl.obolibrary.org/obo/NCBITaxon_9606**. We specificeren ook dat de personen in kwestie (concept) een attribuut hebben (relatie) genaamd vetvrije massa (concept) die is van het type (relatie) uit de **SNOMED CT** ontologie: **http://snomed.info/id/248363008** (concept). Ook de



Figuur 1. Boven: twee datasets die nauwelijks uitwisselbaar zijn. Onder: Het conceptuele model waarbij concepten en relaties uniek zijn gedefinieerd met behulp van ontologieën. Zie tekst voor meer uitleg. De uiteindelijke syntax, het semantische model, is **hier** te vinden.

relaties zijn uniek gedefinieerd met behulp van IRI's: de relatie 'een attribuut hebben' wordt gedefinieerd met behulp van de Semanticscience Integrated Ontology: **https://semanticscience.org/resource/SIO_000008**. Zodoende wordt met behulp van ontologieën eerst een conceptueel model opgebouwd dat telkens bestaat uit trio's van concepten en relaties. Deze trio's zijn geschikt om met behulp het Resource Description Framework (**RDF**) in zogenaamde 'RDF triples' gerepresenteerd te worden in een universeel computer leesbaar formaat (het semantisch model). De syntax die we daar in dit geval voor hebben gebruikt is **Turtle** en hier is te zien hoe uiteindelijk de dataset (voor één persoon) er in Turtle-syntax uitziet.

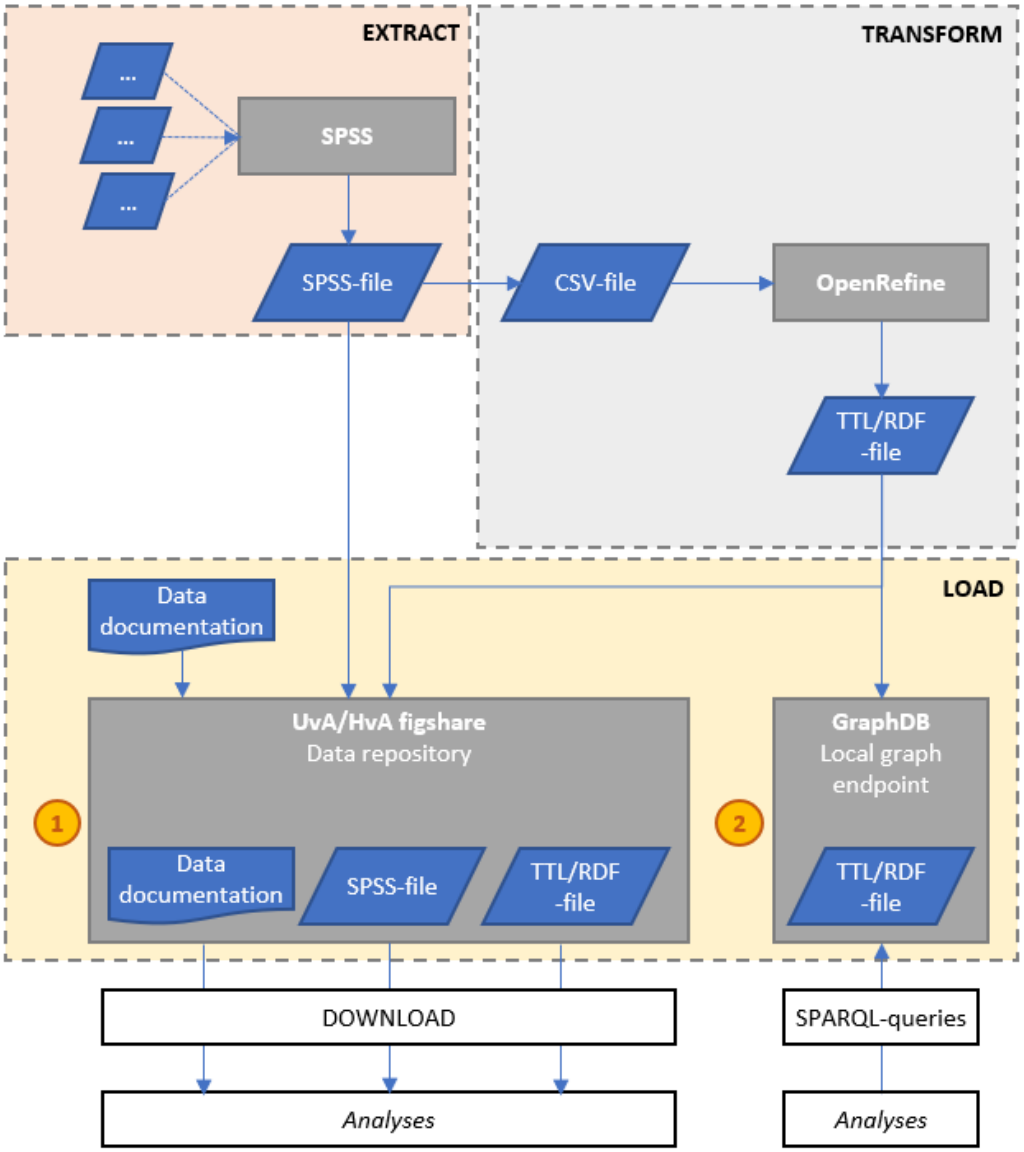
Het proces van het zoeken naar ontologieën, deze modelleren en omzetten naar computer leesbare representaties was tijdsintensief en voor ons als datastewards grotendeels nieuw. Tabel 5 demonstreert een greep uit de vragen (en resulterende beslissingen) die tijdens dit proces voorbij kwamen.

Tabel 5. Een greep uit de vragen die voorbij kwamen tijdens de modelleersessies.

Kiezen we voor een specifiek en concreet semantisch datamodel of voor een generiek en abstract model?
Hoeveel 'uitwisselbaarder' worden datasets als verschillende onderzoekers/onderzoeksgroepen verschillende ontologieën en modellen gebruiken om datasets mee te representeren? Wordt het probleem dan niet gewoon verschoven van allerlei variabele namen naar allerlei ontologietermen die voor een concept te verzinnen zijn?
Als het FAIRificeren van zo'n kleine dataset zoveel tijd in beslag neemt, is het dan wel haalbaar om veel grotere datasets FAIR te maken?
Hoe weten we in SNOMED CT wat er met een bepaalde term wordt bedoeld, want in de SNOMED-browser wordt vaak geen omschrijving gegeven?
Kunnen verschillende ontologieën zomaar door elkaar heen worden gebruikt?
Wat is het verschil tussen URI's, IRI's en URL's?
Kan UvA/HvA figshare beschouwd worden als een FAIR Data Point (FDP) en zo nee, waarom niet?
Tot hoever moeten we doorgaan met modelleren? Moeten we de gehele 'provenance' ook modelleren (i.e., waar komen data vandaan, hoe is het gemeten, wie heeft het gemeten, wanneer in de tijd, etc.)?
Wat is het verschil tussen OMOP common data model en ontologieën?
Wat is het verschil tussen ontologieën, taxonomieën en thesauri?
Welke criteria zijn er voor de kwaliteit van je semantische datamodel?
In de dataset, wat is de rol van een persoon als die persoon geen patiënt is?
In het model, hoe definieer je de relatie tussen leeftijd, geslacht, lengte, gewicht en vetvrije massa aan de ene kant en de onderzoeksdeelnemer aan de andere kant?
In het model, moeten alle instanties een klasse hebben?
Hoe verhouden RDF, Turtle en SPARQL zich tot elkaar?

4.2.2 De dataflow

Figuur 2 geeft schematisch weer welke weg de data (de dataflow) vanaf lokale, nauwelijks vindbare of toegankelijke Amsterdam UMC- en HvA-opslaglocaties hebben afgelegd om uiteindelijk FAIRder te worden gehost in UvA/HvA figshare en een lokale RDF-database. Het schema is losjes gebaseerd op een conventioneel **ETL**-proces waarbij data eerst ergens vandaan moeten worden gehaald (Extract), vervolgens geconverteerd (Transform) en ergens gehost (Load) moeten worden. De Transform-fase kan vergeleken worden met stap 4 en 5 van het FAIRificatie-proces en de Load-fase met stap 6 van het FAIRificatie-proces.



Figuur 2. Schematische weergave van de dataflow, losjes gebaseerd op een conventioneel ETL. (Extract, Transform, Load) diagram. Zie de tekst (paragraaf 4.2.2) voor verdere toelichting.

In de Extract-fase waren buiten het pilotproject om, databestanden (zie figuur 2 linksboven) van verschillende projecten al gecombineerd tot één SPSS-bestand. In de Transform-fase hebben we met behulp van OpenRefine (inclusief RDF-extensie) het ontwikkelde semantische model toegepast op de data en de SPSS-file getransformeerd in een **Turtle/RDF**-file. In de Load-fase hebben we gekozen voor twee 'data-distributies': 1) Gebruik makend van de **data repository UvA/HvA figshare**; 2) Gebruik makend, voor demonstratiedoeleinden, van een lokale RDF-database ook wel **knowledge graph endpoint** genoemd. Een belangrijk verschil is dat bij optie 1 bestanden eerst gedownload moeten worden voordat ze geanalyseerd kunnen worden, terwijl bij optie 2 in principe direct 'queries' op de database uitgevoerd kunnen worden (vergelijkbaar met het PHT-concept). Voor beide opties geldt dat dit alleen met de juiste toegangsrechten kan (de A van FAIR) vanwege privacy en dus dat de data niet 'open' zijn.

4.3 Privacy en informed consent

De casus waar we in dit project mee hebben gewerkt omvatte 11 datasets van de HvA en het Amsterdam UMC. Negen van de 11 datasets waren verzameld in het kader van wetenschappelijk onderzoek waarvoor onderzoeksdeelnemers destijds 'informed consent' hebben afgegeven. In één van de negen informed consents is expliciet iets opgenomen over toestemming voor toekomstig hergebruik van de data. De andere twee van de 11 datasets waren afkomstig uit de zorg, zijn niet verzameld met wetenschappelijk onderzoek als doel en zijn destijds verwerkt op basis van een andere AVG-grondslag (AVG: Algemene Verordening Gegevensbescherming) dan toestemming. De vraag of deze 11 datasets hergebruikt mogen worden voor een ander wetenschappelijk doeleinde dan waarvoor de datasets oorspronkelijk zijn verzameld, hebben wij en de geconsulteerde experts binnen het tijdsbestek van dit pilotproject (ca. 6 maanden, 15 mens-uren) niet kunnen beantwoorden. Een antwoord op deze vraag kan van twee (overlappende) kanten benaderd worden: ethisch en juridisch.

Ethisch gezien kunnen we onszelf afvragen wat we er zelf van vinden dat er gegevens van mensen over hun gewicht, lengte, leeftijd, BMI en vetvrije massa 'via via' bij allerlei verschillende onderzoekers terecht komen? Moeten de mensen van wie die gegevens zijn zelf zeggenschap blijven houden over wat er met hun gegevens gebeurt? Ook als die gegevens niet of nauwelijks meer tot een uniek persoon te herleiden zijn? Vinden we zelf dat de privacy van die mensen voldoende geborgd is. Privacy is een grondrecht en als als we juridisch kijken, hebben we vooral te maken met de AVG. De eerste vraag is of we te maken hebben met persoonsgegevens of met anonieme gegevens. Als we te maken hebben met persoonsgegevens, dan is de AVG van toepassing, en moet onder andere bepaald worden of het nieuwe doel van de verwerking verenigbaar is met de oorspronkelijke doeleinden en op basis van welke AVG-grondslag de nieuwe verwerking gebeurt. Dat laatste is problematisch omdat er geen toestemming is (m.u.v. één onderzoek) voor het hergebruiken van de gegevens om een nieuwe onderzoeksvraag te beantwoorden, en dat leidt dus tot de vraag of en hoe dit hergebruik te rechtvaardigen is. Als we te maken hebben met anonieme gegevens, dan is de AVG niet van toepassing, en hoeven we ons dus vanuit AVG-perspectief niet druk te maken over doelbinding, grondslagen, datalekken, bewaartermijnen, verwerkersovereenkomsten, het openstellen van data, etc.

Volgens grond 26 van de AVG zijn anonieme gegevens "gegevens die geen betrekking hebben op een geïdentificeerde of identificeerbare natuurlijke persoon of op persoonsgegevens die zodanig anoniem zijn gemaakt dat de betrokkene niet of niet meer identificeerbaar is". Om te bepalen of iemand identificeerbaar is "moet rekening worden gehouden met alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruikt door de verwerkingsverantwoordelijke of door een andere persoon om de natuurlijke persoon direct of indirect te identificeren, bijvoorbeeld selectietechnieken". Om uit te maken wat redelijkerwijs betekent "moet rekening worden gehouden met alle objectieve factoren, zoals de kosten van en de tijd benodigd voor identificatie,

met inachtneming van de beschikbare technologie op het tijdstip van verwerking en de technologische ontwikkelingen". De huidige uit elf datasets gecombineerde dataset, bevat geen direct identificerende persoonsgegevens zoals naam of geboortedatum, maar wel andere gegevens over die personen. Valt het redelijkerwijs te verwachten dat iemand met deze gegevens een persoon kan identificeren? En wat als we weten dat ergens achter slot en grendel bij het Amsterdam UMC en HvA de sleutelbestanden staan, waarmee een koppeling tussen persoon en onderzoeksgegevens mogelijk is?

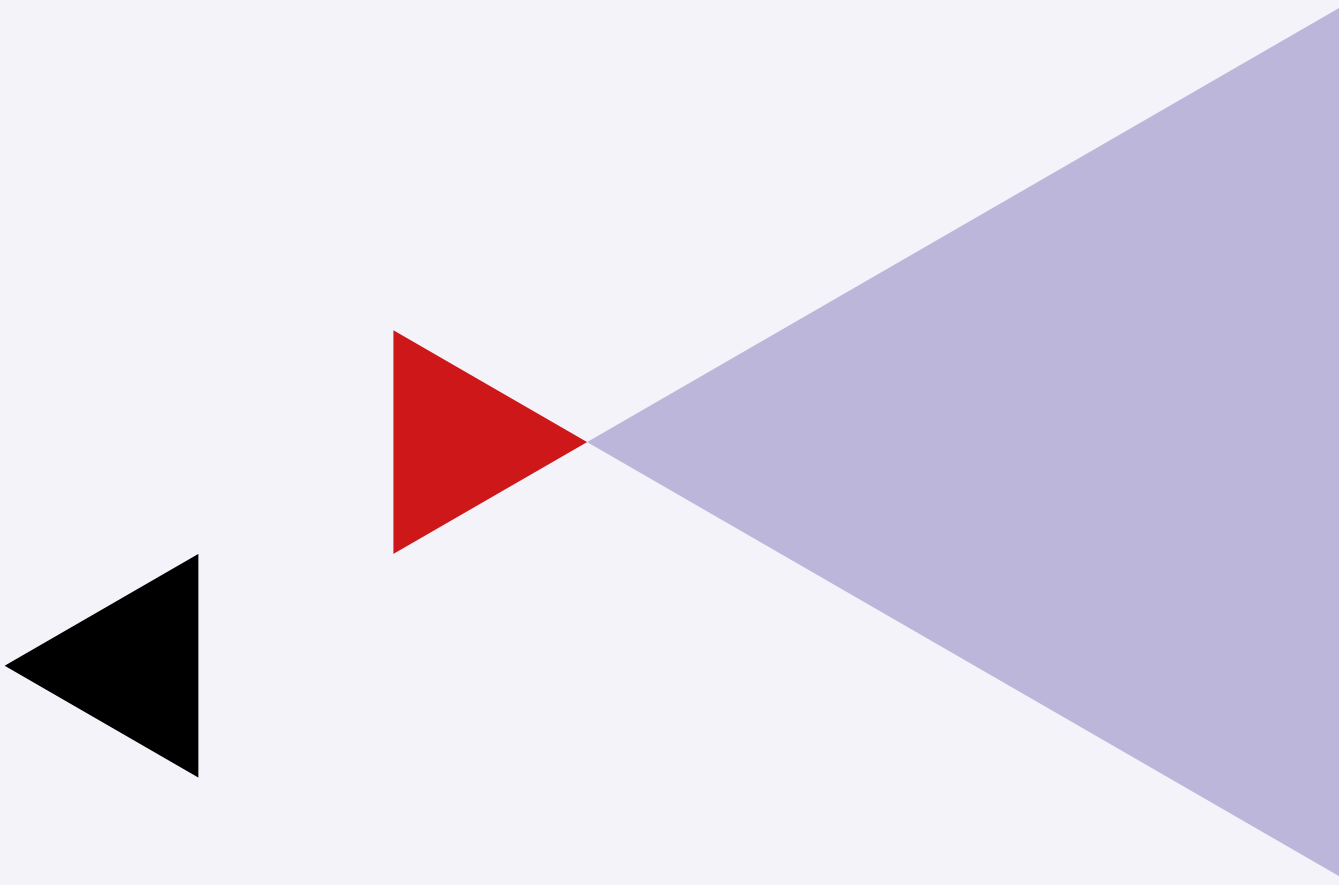
Het is niet zo dat er van de AVG niets mag, maar de hierboven geschetste onduidelijkheid geeft onzekerheid over wat wel of niet mag en wat er al dan niet geregeld moet worden. Vanwege de focus op de I van FAIR in dit project is dit punt (nog) niet verder uitgewerkt, maar de A en R van FAIR gaan wel precies over het transparant maken van en duidelijkheid scheppen over de condities waaronder data toegankelijk zijn voor anderen, voor wie en hoe. FAIR data is (dus) uitdrukkelijk niet hetzelfde als open data: data kunnen volstrekt FAIR zijn zonder open te zijn.

4.4 Grip op data

Zoals het concept van de Personal Health Train laat zien gaat FAIR niet alleen over data, maar over een ' internet of FAIR data en services' [14]: Er wordt een web van (complexe) informatie-systemen ontwikkeld en aan elkaar gekoppeld, ook wel 'machine-to-machine communication' (M2M) genoemd, waarbij menselijke tussenkomst niet altijd noodzakelijk is. Ook hebben we in ons project gezien dat data zodanig worden gerepresenteerd dat het voor computers begrijpelijker, maar voor een mens – op het eerste oog – minder begrijpelijk wordt. Naast voordelige AI-toepassingen en betere privacy-by-design mogelijkheden kan een dergelijke digitalisering bij onderzoeksdeelnemers of het bredere publiek ook leiden tot wantrouwen in informatiesystemen, vragen over privacy en zorgen over geautomiseerde beslissingen. FAIR betekent niet automatisch 'fair': het eerlijke en ethische gebruik van data voor en door iedereen. Het is daarom van belang om onderzoeks-deelnemers en het bredere publiek te betrekken bij deze ontwikkelingen en om datawijsheid bij mensen te vergroten. Participatieve en actieve methodes zoals citizen science en **datafysicalisatie** zijn manieren om dit te bewerkstelligen [15]. Om bijvoorbeeld 'ambient technology' in de thuissituatie aan te passen aan de wensen en eisen van kwetsbare ouderen, moet er eerst een begrip zijn wat dat concept betekent. Door gebruik te maken van een interactief poppenhuis op schaal met sensoren konden onderzoekers samen met de oudere mensen 'verkennen' wat ambient technology inhield en waren er indicaties dat ouderen beter konden aangeven wat ze er van vonden [16]. Wat betreft FAIR zouden op vergelijkbare tastbare wijze het concept van metadata, documentatie en herbruikbaarheid gedemonstreerd kunnen worden door middel van modellen van lego die door andere teams gereproduceerd moeten worden [17]. Ook kunnen datastromen gevisualiseerd worden om te laten zien welke 'reis' gegevens maken en wie waar op welk moment bij kan. Hansen et al. [18] proberen het FAIR-concept te illustreren aan de hand van een tekstuele 'FAIRy tale' (zie ook hier). En bij het ontwikkelen van FAIR registers voor zeldzame ziekten worden patiënten en belangenverenigingen actief betrokken bij elke implementatiefase [19]. Met dit soort meer praktijkgerichte (onderzoeks)methoden wordt FAIR ook meer fair en kunnen we proberen recht te doen aan de waarden zoals opgesteld in het manifest 'tada – duidelijk over data' [20].

5 Dankwoord

We willen de projectgroep bedanken voor hun waardevolle input, adviezen, feedback en inzet. De projectgroep bestond naast de auteurs uit: Jesse Aarden, Leontine Born, Carliene van Dronkelaar, Dennis van Erck, Marjon Kluiters, Dennis van Kooij, Ed Kuijpers, Gerben ter Riet, Josje Schoufour, Mike Tieland, Amely Verreijen, Bart Visser en Pascal Wiggers. Daarnaast willen we de Urban Vitality zaaigeldcommissie, het SIA-Sprong project Mensen in Beweging, Petra van de Moosdijk, Somaya Ben Allouch en Wilma Scholte-op-Reimer bedanken voor het mogelijk maken van dit project.



List of literature

1. Virus Outbreak Data Network (VODAN) GO FAIR Implementation Network (28 april, 2020). Retrieved from **https://www.go-fair.org/implementation-networks/overview/vodan/**.

2. Wilkinson et al. (2016). *The FAIR Guiding Principles for scientific data management and stewardship. Scientic* Data, 3, 160018. **https://doi.org/10.1038/sdata.2016.18**.

3. KNAW; NFU; NWO; TO2-federatie; Vereniging Hogescholen; VSNU (2018). *Nederlandse gedragscode wetenschappelijke integriteit*. DANS. **https://doi.org/10.17026/dans-2cj-nvwu**.

4. Mons et al. (2017). *Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud*. Information Services & Use, 37:1, 49-56. **https://doi.org/10.3233/ISU-170824**.

5. Jacobsen et al. (2020). *A Generic Workflow for the Data FAIRification Process*. Data Intelligence, 2:1-2, 56-65. **https://doi.org/10.1162/dint_a_00028**.

6. van Ulzen, de Jong (2020). *Semantic modeling process of body composition data and protein intake estimation*. University of Amsterdam / Amsterdam University of Applied Sciences. Online resource. **https://doi.org/10.21943/auas.12581336.v1**.

7. Kanis et al. (2020). FAIR: Geen woorden maar data. *University of Amsterdam / Amsterdam University of Applied Sciences*. Online resource. **https://uvaauas.figshare.com/projects/FAIR_Geen_woorden_maar_data/83840**.

8. De Personal Health Train in de zorg. Verhalen uit de praktijk (juli 2020). Retrieved from **https://pht.health-ri.nl/sites/healthtrain/files/2020-07/PHT%20in%20de%20zorgpraktijk.pdf**.

9. *Mons (2020). The VODAN IN: support of a FAIR-based infrastructure for COVID-19*. European Journal of Human Genetics, 28, 724–727. **https://doi.org/10.1038/s41431-020-0635-7**.

10. Melanie Imming. (2018, April 23). *FAIR Data Advanced Use Cases: from principles to practice in the Netherlands (Version Final)*. Zenodo. **http://doi.org/10.5281/zenodo.1250535**.

11. Deist et al. (2020). *Distributed learning on 20 000+ lung cancer patients – The Personal Health Train*. Radiotherapy and Oncology, 144, 189-200. **https://doi.org/10.1016/j.radonc.2019.11.019**.

12. Löbe et al. (2020). *Problems in FAIRifying Medical Datasets*. Studies in Health Technology and Informatics, 270, 392-396. **https://doi.org/10.3233/shti200189**.

13. Lamprecht et al. (2020). *Towards FAIR Principles for Research Software*. Data Science, 3:1, 37-59. **https://doi.org/10.3233/DS-190026**.

14. Ayris et al. (2016). *Realising the European open science cloud*. European Union. Retrieved from **https://doi.org/10.2777/940154**.

15. Kanis et al. (2019). *Zichtbaar slimmer: Resultaten in beeld*. Hogeschool van Amsterdam, Digital life. Retrieved from **http://www.digitallifecentre.nl/redactie/resources/zichtbaar-slimmer-boekje-v7-compressed-1.pdf**.

16. Kanis et al. (2013). *Sensor monitoring in the home: Giving voice to elderly people*. 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, Venice, 97-100. **https://doi.org/10.4108/icst.pervasivehealth.2013.252060**.

17. Donaldson & Mahon. (2020). *Lego: Metadata for reproducibility*. Zenodo. **http://doi.org/10.5281/zenodo.3685685**.

18. Karsten et al. (2018). *A FAIRy tale (p. 40)*. Zenodo. **http://doi.org/10.5281/zenodo.2248200**.

19. Kodra et al. (2018). *Recommendations for Improving the Quality of Rare Disease Registries*. International Journal of Environmental Research and Public Health, 15:8, 1644. **https://doi.org/10.3390/ijerph15081644**.

20. Tada.city - Met data maken we steden slimmer. Maar ook wijzer? (14 juli, 2020). Retrieved from: **https://tada.city/**.

21. Maresi (2020). *A Data Engineering's guide to semantic modelling*. Zenodo. **https://doi.org/10.5281/zenodo.3898519**.

Begrippenlijst en afkortingen

AI	Artificial Intelligence. Kunstmatige intelligentie.
AVG	Algemene Verordening Gegevensbescherming.
Bodpod	Een apparaat waarmee op een betrouwbare en valide wijze de lichaamssamenstelling van een cliënt wordt bepaald. Het apparaat meet de hoeveelheden vetmassa en vetvrije massa. Zie hier.
Castor EDC	Een online applicatie waarmee je zelf eenvoudig en veilig (klinische) data verzamelt. Ook wel een 'electronic data capture' (EDC) platform genoemd.
Datafysicalisatie	Het tastbaar maken van data, fysieke 3D-representaties die je kunt beleven en aanraken. Het tastbaar maken van informatie kan leiden tot meer inzicht wat vervolgens weer kan leiden tot kritische discussie of gedragsverandering.
Data repository	Een digitaal archief voor het archiveren, preserveren en/of publiceren van onderzoeksdata. Voor de HvA is er UvA/HvA figshare. Een register van research data repositories is **hier** te vinden.
Datasteward	Iemand die ondersteunt en adviseert bij het optimaliseren van vindbaarheid, toegankelijkheid, (her)gebruik en reproduceerbaarheid van onderzoeksdata.
DOI	Digital Object Identifier. Een DOI is een 'persistent identifier' voor een bestand, zoals een wetenschappelijk artikel of een dataset, op het world wide web.
ETL	Extract, Transform, Load. ETL is een begrip uit de informatica. Het benoemt de processen die gebruikt worden om data uit een bron te halen (Extract), de opgehaalde data om te zetten (Transform) en op een andere plek weg te schrijven (Load).
Kennisrepresentatie	Knowledge Representation. Een discipline die erop gericht is informatie te presenteren op een manier dat computersystemen van deze informatie gebruik kunnen maken [8].
Knowledge graph (endpoint)	Ook wel RDF triple-stores of Graph-database. Een database die is geoptimaliseerd voor opslag van RDF-triples en het opvragen er van door semantische (SPARQL-) queries.
LOINC-ontologie	Logical Observation Identifiers Names and Codes (LOINC) is een universele standaard voor het identificeren en coderen van medische laboratorium observaties.
Meetinstrument	Vragenlijsten, fysieke (perfomance) testen, cognitieve testen, observatie testen, etc. Zie **hier** voor een collectie van meetinstrumenten in de zorg.
Metadata	Data over data. Metadata geven informatie over de data. Metadata om een dataset op studieniveau te kunnen beschrijven zijn bijvoorbeeld: titel, creator, omschrijving, identifier en aanmaakdatum van de dataset. Metadata op dataniveau zijn bijvoorbeeld: eenheden, labels en datadefinities.
NCBI-taxonomie	Een 'curated classification and nomenclature for all of the organisms in the public sequence databases' [**https://www.ncbi.nlm.nih.gov/taxonomy**].

Ontologie	Een woordenboek is een collectie van termen en hun betekenis. Een taxonomie is een hierarchische classificatie van concepten. Een thesaurus geeft een beschrijving van de relatie tussen concepten. Een ontologie is een combinatie van deze drie waarbij de relaties tussen concepten nog explicieter zijn gedefinieerd [21].
Persistent identifier	Een permanente verwijzing en uniek label naar een object die onafhankelijk is van de bewaarlocatie. Voorbeelden zijn een DOI (Digital Object Identifier), een ISBN-nummer en een ORCID.
Privacy-by-design	Bij de start van het ontwerpen van een informatiesysteem of product wordt rekening gehouden met privacy. De aandacht hiervoor blijft tijdens de gehele levensduur van het system of product bestaan [8].
QDA	Qualitative Data Analysis. Kwalitatieve data-analyse.
RDF	Resource Description Framework. Een framework voor het representeren en uitwisselen van data op het web, oorspronkelijk ontworpen als een metadata-model. Met RDF worden data gerepresenteerd door middel van triples, die bestaan uit een subject (persoon), predicaat (is een), en object (homo sapiens).
REFI	Rotterdam Exchange Format Initiative. Een project-/data-formaat om uitwisselbaarheid tussen verschillende QDA-softwareprogramma's (bijv. MAXQDA, Atlas.ti) te bevorderen.
Semantic web	Een uitbreiding van het world wide web om internetdata computer leesbaar te maken door een set van standaarden te gebruiken, gedefinieerd door het World Wide Web Consortium (W3C). Om te bewerkstelligen worden technieken zoals het Resource Description Framework (RDF) en Web Ontology Language (OWL) gebruikt [wikipedia].
Semantisch model	Een representatie van de impliciete betekenis van de data door concepten en relaties in die data te specificeren op een voor mensen en computers begrijpelijke manier [21].
SNOMED CT	SNOMED Clinical Terms (CT) is een medische standaard voor het documenteren en coderen van medische gegevens om door een computer verwerkt te worden en om de variatie in de manieren om gegevens vast te leggen, te coderen en te gebruiken voor zorg en onderzoek te beperken [wikipedia].
SPARQL	SPARQL Protocol And RDF Query Language is een semantische 'query language' om RDF-databases te bevragen, vergelijkbaar met SQL voor relationele databases.
Turtle	Terse RDF Triple Language (Turtle) is een syntax en bestandsformaat voor het Resource Description Framework (RDF) data model. Turtle syntax is vergelijkbaar met SPARQL, een RDF query taal [wikipedia].
UvA/HvA figshare	De data repository van de HvA en de institutionele variant van het platform Figshare.