

FAIR Data

in Medical Research

Incorporating the FAIR Principles
in the Research Data Life Cycle



Martijn G. Kersloot

Promoveren?

De afgelopen viereneenhalf jaar was ik een **promovendus** en deed ik promotieonderzoek naar hoe we andere onderzoekers kunnen helpen bij het herbruikbaar maken van de gegevens die ze verzamelen in hun onderzoek. Een hele mond vol!

In dit **lekenboekje** leg ik meer uit over wat promoveren inhoudt en wat ik de afgelopen jaren onderzocht heb.

Wie zijn er betrokken bij een promotietraject?

De **promovendus**, degene die gaat promoveren, wordt begeleid door een of twee **promotor(en)**. Dat zijn hoogleraren (professoren) of universitaire hoofddocenten die eindverantwoordelijk zijn voor de onderzoekslijn. De promovendus wordt meestal dagelijks begeleid door een of twee **co-promotor(en)**. Zelf zijn zij ook gepromoveerd en meestal zeer nauw betrokken bij het onderzoek.

Voorafgaand aan de verdediging beoordeelt een **promotiecommissie** of het proefschrift van voldoende niveau is. De commissie bestaat uit een aantal hoogleraren of experts die verstand hebben van het onderwerp waarop gepromoveerd gaat worden. Pas na hun toestemming mag de verdediging plaatsvinden. Tijdens die verdediging stellen zij aan de promovendus een aantal vragen over het geschreven proefschrift.

De promovendus wordt bijgestaan door **paranimfen**: meestal vrienden of collega's van de promovendus. Vroeger hielpen ze promovendus bij moeilijke vragen of vielen zelfs voor hem in bij ziekte, maar nu vervullen ze vooral een ceremoniële rol. Ze helpen bij de voorbereidingen en zitten tijdens de verdediging naast de promovendus om eventueel een stelling of passage uit het proefschrift voor te lezen.

Als laatste is er nog de **pedel**, die zorgt dat iedereen zich tijdens de verdediging netjes aan het protocol houdt.

Wat is promoveren?

Tijdens een promotietraject doe je **wetenschappelijk onderzoek**. De resultaten van de verschillende onderzoeken die je doet worden opgeschreven in artikelen die vervolgens gepubliceerd worden in wetenschappelijke tijdschriften.

Deze **publicaties** worden vervolgens gebundeld tot een **proefschrift**, waarin ook een algemene inleiding en discussie over het onderzoek worden toegevoegd.

Vervolgens wordt dit proefschrift verdedigd ten overstaande van een **commissie**. Na een succesvolle verdediging mag iemand zich gepromoveerd noemen en de titel '**doctor**' (dr. of PhD) voeren.

Hoe verloopt de ceremonie?

Promoties beginnen altijd op **het hele uur**. Vlak voordat de ceremonie begint doen de paranimfen huishoudelijke mededelingen (o.a. niet klappen tijdens de ceremonie!). Op het hele uur begint de promovendus met zijn **lekenpraatje**, een korte presentatie waarin wordt uitgelegd waar het proefschrift over gaat. Een gesproken versie van dit lekenboekje dus!

Om **kwart over het uur** begint de verdediging officieel: de commissie komt binnen (hoogleraren in toga) en de voorzitter opent de bijeenkomst. De promovendus leest een formele tekst voor waarna de discussie wordt geopend. Commissieleden worden tijdens de verdediging aangesproken als **hooggeleerde** (hoogleraar) of **zeergeleerde** (doctor) **opponent**. Na drie kwartier komt de pedel binnen, precies op het uur, en spreekt "**hora est**" (het is tijd) uit. De promovendus leest hierna nog een formele tekst voor en de commissie trekt zich hierna terug voor **beraad**.

Na 5 tot 10 minuten komt de commissie, onder aanvoering van de pedel, terug met het oordeel en (hopelijk!) de **bul**. De bul wordt overhandigd en hierna volgt de **laudatio**, een persoonlijk woord van de promotor. Na deze felicitatie wordt de zitting beëindigd (nu mag applaus wel!) en is het tijd voor de receptie.



Achtergrond

Hoofdstuk 1

Evidence-based medicine

De geneeskunde van nu is grotendeels gebaseerd op de **uitkomsten van klinisch onderzoek**. In zo'n onderzoek wordt bijvoorbeeld onderzocht welk medicijn het beste werkt tegen een bepaalde ziekte.

Als je naar een arts gaat met klachten, combineert hij zijn klinische kennis, kennis verkregen uit verschillende onderzoeken en jouw individuele wensen om zo goed mogelijke zorg te leveren. Dit heet **evidence-based medicine**.



Gegevens 'verdwijnen op de plank'

De laatste stap van de Research Data Life Cycle, het **vinden en hergebruiken** van gegevens is erg belangrijk: het brengt de wetenschap verder en zorgt er vooral voor dat andere onderzoekers het wiel niet opnieuw uitvinden.

We zien alleen dat nadat veel onderzoeken afgerond zijn, de verzamelde gegevens **op de plank verdwijnen** en niet hergebruikt worden. Omdat het grootste gedeelte van die gegevens slecht omschreven is (bijvoorbeeld wat er verzameld is onder welke omstandigheden), wordt er zelfs geschat dat **80%** van de verzamelde gegevens überhaupt niet hergebruikt kan worden.



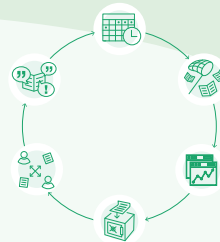
De FAIR Principles

Om dit herbruikbaarheidsprobleem op te lossen, zijn in 2016 de FAIR Principles opgesteld. Die geven aan dat onderzoeksgegevens en informatie over het onderzoek en de bijbehorende gegevens (metadata) FAIR moeten zijn voor **andere onderzoekers** (mensen) en voor **programma's** die de gegevens analyseren (computers).



Research Data Life Cycle

Een klinisch onderzoeksproject kun je opdelen in verschillende fasen. In alle fasen van het project staat **Research Data Management** centraal: hoe ga je om met de gegevens die je verzamelt in je onderzoek? In de **Research Data Life Cycle** wordt omschreven welke fasen een onderzoekstraject kent en welke data management-taken daarbij horen.



Allereerst **plan en ontwerp** je je project: hoe gaat het onderzoek eruit zien en wat voor gegevens wil je gaan verzamelen? Vervolgens **verzamel je gegevens** van de patiënten die meedoen aan het onderzoek. Denk bijvoorbeeld aan de bloeddruk of het gewicht en de lengte van de patiënt. Nadat die gegevens verzameld zijn, **analyseer** je ze. In het voorbeeld van een medicijnonderzoek: zit er verschil in de groep patiënten die wel een medicijn kreeg en de groep die dat niet kreeg? Die gegevens en uitkomsten **sla je op**, zodat er later nog naar gekeken kan worden. Uiteindelijk **deel en publiceer** je je bevindingen in een wetenschappelijk artikel, met idealiter ook de gegevens erbij die je verzameld hebt. Die gegevens kunnen anderen dan weer **vinden en hergebruiken** voor hun onderzoek.

Findable

Goed beschreven en dus *vindbaar*

Accessible

Duidelijk of en hoe je bij de gegevens kan komen en dus *toegankelijk*

Interoperable

Uitwisselbaar met andere verzamelde gegevens en dus *interoperabel*

Reusable

Duidelijk omschreven hoe de gegevens kunnen worden gebruikt en dus *herbruikbaar*

Doel: de stappen om onderzoeksgegevens FAIR te maken opnemen in de Research Data Life Cycle

Veel van de taken die een onderzoeker op een dagelijkse basis uitvoert hebben raakvlakken met de stappen die nodig zijn om onderzoeksgegevens meer FAIR te maken (FAIRificatie). In mijn promotieonderzoek onderzochten we daarom hoe we deze stappen kunnen **opnemen in de huidige manier van werken** van onderzoekers: de Research Data Life Cycle.



Inhoud

Mijn proefschrift bestaat uit drie verschillende delen.

In **Deel I** (*De huidige staat van FAIR*) onderzochten we wat onderzoekers weten en vinden van het FAIR maken van hun verzamelde gegevens.

In **Deel II** (*NLP en FAIR*) onderzochten we hoe Natural Language Processing (een manier om een computer 'vrije' tekst te laten analyseren) kan helpen bij het FAIR maken van gegevens.

In **Deel III** (*FAIR by design*) ontwikkelden we een proces om gegevens FAIR te maken, direct wanneer ze verzameld worden.

Deel I

De huidige staat van FAIR

Hoofdstuk 2 en 3

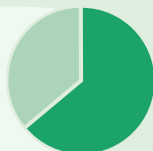
De hoofdstukken in dit deel beschrijven de resultaten van een vragenlijst die is uitgezet onder klinisch onderzoekers. In de vragenlijst werd gevraagd wat ze weten van de FAIR Principles en welke stappen ze zetten om hun gegevens meer FAIR te maken.

164 onderzoekers hebben de vragenlijst ingevuld

Training, (financiële) ondersteuning en tools

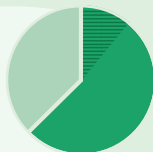
zijn nodig zodat onderzoekers hun gegevens meer en beter FAIR kunnen maken

64,0% van de ondervraagde onderzoekers had **eerder van de FAIR Principles gehoord**.



62,8% doet op zijn minst enige moeite om **enig aspect** van FAIR te bereiken. 11,0% behandelt **alle aspecten**.

(vindbaar, toegankelijk, interoperabel en herbruikbaar)



93,9% richt zich op het realiseren van FAIR **voor mensen**, en maar 32,1% op FAIR **voor computers**.



35,1% geeft aan **zelf gegevens FAIR te kunnen maken** en 81,6% geeft aan hier **hulp voor nodig te hebben**.



Deel II

NLP en FAIR

Dit deel bevat een hoofdstuk met literatuuronderzoek naar het gebruik van Natural Language Processing (NLP: het 'begrijpen' van tekst door een computer) in de gezondheidszorg en een hoofdstuk dat het ontwikkelen en testen van zo'n NLP toepassing beschrijft.

Hoofdstuk 4 en 5

77 artikelen
samengevat

Groot verschil in testen van algoritmen en beschrijving daarvan in artikelen

16 daarom
aanbevelingen opgesteld voor
toekomstig onderzoek

teruggekeerde

255227004

niet-kleincellige

longkanker

254637007

De ontwikkelde toepassing detecteert
medische termen en hun onderlinge
relaties in tekst, wat de noodzaak voor
handmatig doorlezen van patiëntendossiers
voor onderzoeksdoeleinden kan verminderen.

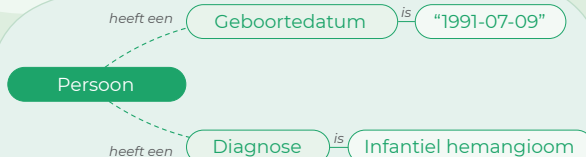
Deel III

FAIR by design

Hoofdstuk 6, 7 en 8

In dit deel wordt een nieuw proces omschreven dat gegevens automatisch meer FAIR maakt. Het proces is verweven in het opzetten van een onderzoeksproject, zodat onderzoekers uiteindelijk minder tijd kwijt zijn aan het FAIR maken van hun gegevens.

Nadat patiëntgegevens zijn ingevoerd in de registratie, worden deze gegevens automatisch omgezet in een formaat dat FAIR en leesbaar is voor computers. Het formaat is een soort mindmap waar de gegevens van de patiënt en de relatie van die gegevens tot de patiënt worden opgenomen.



Voorbeeld van gegevens die leesbaar zijn voor een computer
Elk blokje bevat een door een computer-leesbare definitie

Registry of Vascular Anomalies

Het FAIRificatieproces ontwikkeld in dit deel, is toegepast op een registratie (een type onderzoek dat over de lange termijn gegevens van een bepaalde groep patiënten verzamelt) voor vasculaire afwijkingen. Dit soort afwijkingen is zeldzaam, daarom is het erg belangrijk om daar gegevens over te verzamelen. Zo wordt er meer kennis vergaard. Met die kennis en gegevens kunnen er hopelijk betere behandelingen worden ontwikkeld.



Het proces is geïntegreerd in de software die onderzoekers gebruiken om gegevens te verzamelen, gegevens worden dus FAIR gemaakt 'aan de bron'.



Doordat computers het formaat kunnen lezen, kan er op grote schaal analyse van de gegevens gedaan worden en kunnen verschillende gegevensbronnen met elkaar gecombineerd worden.

Discussie & Conclusie

In de zeven hoofdstukken van het proefschrift werd beschreven hoe onderzoekers de gegevens die ze verzamelen meer FAIR kunnen maken. Een aantal conclusies beschreven in het proefschrift worden hieronder genoemd.

Veel onderzoekers zijn momenteel **niet op de hoogte** van de (betekenis van de) FAIR Principles.

Processen om onderzoeksgegevens FAIR te maken kunnen worden **opgezet voordat het onderzoek start**.

Software kan gegevens, direct wanneer ze verzameld worden, **automatisch FAIR maken**.

Tekst uit patiëntendossiers kan **herbruikbaar gemaakt worden** via Natural Language Processing.

FAIR-gemaakte onderzoeksgegevens kunnen **automatisch beschikbaar worden gesteld**, zodat anderen deze kunnen gebruiken.

Om het FAIR maken van onderzoeksgegevens schaalbaar te maken zijn er **standaarden, methoden, ondersteuning, tools en financiering** nodig. **Beleidsmakers, subsidieverstrekkers en onderzoeksinstituten** moeten samenwerken om dit aan te bieden aan de onderzoeksgemeenschap.



Stellingen

Stellingen bij een proefschrift zijn een aantal beweringen, door de promovendus gedaan, die hij wil verdedigen tegenover de promotiecommissie. Het merendeel heeft betrekking op het onderzoek dat gedaan is. Naast de serieuze stellingen zijn er ook een aantal ludiek, de zogenaamde schertsstellingen

1. In plaats van onderzoekers te verplichten hun data FAIR te maken, zouden subsidieverstrekkers en instellingen zich eerst moeten richten op het creëren van FAIR bewustwording en het beschikbaar stellen van ondersteuning en tools.
Dit proefschrift, hoofdstuk 2 en 3
2. Het is belangrijk om het praktisch nut van FAIR data aan onderzoekers te laten zien, om zo hun houding tegenover en intentie tot FAIR maken van hun eigen data positief te beïnvloeden.
Dit proefschrift, hoofdstuk 2
3. Onderzoekers moeten prioriteit geven aan het machine-leesbaar maken van hun data in plaats van deze alleen leesbaar te maken voor mensen.
Dit proefschrift, hoofdstuk 3
4. Consistente evaluatie en eenduidige rapportage over NLP-algoritmen gebruikt in de geneeskunde zouden de adoptie van deze algoritmen erg ten goede komen.
Dit proefschrift, hoofdstuk 4
5. Om FAIRificatie schaalbaar te maken, moeten de-novo FAIRificatie workflows de standaard worden.
Dit proefschrift, hoofdstuk 6 en 7
6. Om het FAIR maken van onderzoeksdata toegankelijker te maken voor onderzoekers is het belangrijk om workflows en tools te integreren in hun huidige manier van werken.
Dit proefschrift, hoofdstuk 9
7. Men zou een heel promotietraject kunnen wijden aan de woordspelingen die gemaakt kunnen worden met het acroniem "FAIR".
8. Voor veel onderzoekers, inclusief de onderzoekers die hun gehele promotietraject richten op dit onderwerp, voelt FAIR data een FAIR-van-hun-bed-show.
9. Verspilling in onderzoek tegengaan: geen woorden, maar (meta)data.