

## Pilot: Research Drive / iRODS integratie pilot

*Hogeschool: Saxion Hogeschool*

*Contactpersoon: Saskia de Rijk | s.c.m.derijk@saxion.nl*

*Looptijd: februari 2021 – augustus 2021*

### **Beschrijving van de pilot**

In een door SURF aangeboden tool (een op irods gebaseerde techniek) laten zien hoe datasets van onderzoekers bij Saxion via de Research Drive kunnen worden gemetadateerd en gearhiveerd.

### **Doel van de pilot**

Kijken of dit werkproces aansluit bij de ontwikkelingen in Saxion en de wensen van gebruikers ten aanzien van het metadateren en archiveren van datasets van onderzoek(ers).

### **Welk vraagstuk of probleem heb je aangepakt en hoe (doorlopen stappen/proces)?**

Om data(sets) van onderzoekers te kunnen hergebruiken (disseminatie) moeten deze bewaard blijven en vindbaar worden voor andere tot nu toe werd in de Research Drive-data (tijdelijk) opgeslagen. Daartoe hebben we een enquête gehouden over huidige ervaringen met Research Drive en in bijeenkomsten wensen ten aanzien van onderzoeksdata geïnventariseerd en vergeleken/afgestemd met de mogelijkheden van SURF (pilot).

### **Resultaten & impact**

#### **Resultaten**

In Saxion hebben we door middel van een vooronderzoek wensen en eisen rondom onderzoeksdata (beheer, opslag, archiveren, metadateren en dissemineren) en de Research Drive geïnventariseerd. Uit het vooronderzoek blijkt dat de omgang met data binnen Saxion onderzoek erg divers is. Het type data waar onderzoekers mee werken verschilt heel erg en daarbij ook de werkprocessen en behoefte omtrent het werken met data en het delen ervan. In de bijlage is gerichte feedback te vinden op de gebruikte metadatavelden en het in te richten proces.

#### **Wat heb je geleerd?**

Momenteel ervaren gebruikers van de Research Drive nog (veel) hindernissen in het gebruik van de Research Drive. Deze basis moet eerst op orde komen willen we het systeem verder uitbreiden met extra functionaliteiten. Archiveren en publiceren van datasets zijn momenteel niet de belangrijkste issues binnen onderzoeksprojecten. Wel is duidelijk dat dit een belangrijk onderwerp gaat worden in de toekomst en moeten we als organisatie hier mogelijkheden voor aanbieden en de organisatie, processen en systemen hier klaar voor maken. In de pilot hebben we al wel feedback opgehaald met betrekking tot de ingebouwde meta-datavelden en het proces van metadateren binnen de drive. Het gaat dan onder andere om de relevantie van de gekozen metadatavelden, op welk niveau (project, folder, bestand) je kunt metadateren en hoe je dit proces binnen de drive kunt gaan inrichten (hoe vaak kun je bv een dataset met metadata voorzien en archiveren). Vanuit deze feedback kan de tooling nog worden geoptimaliseerd. Als positief wordt ervaren dat je voor het metadateren en archiveren je in de Research Drive 'kunt blijven'. Dat we meer moeten investeren in verbeterde werkprocessen, afspraken, en training en ondersteuning ten behoeve van onderzoekers en hun onderzoeksdata, voor we misschien ook wel parallel door kunnen ontwikkelen met meta dateren en archiveren (zie ook de Bijlage)

### Hoe ga je hiermee verder?

Deze pilot maakt binnen Saxion deel uit van een breder project. In dit project gaan we aandacht besteden aan bovengenoemde aspecten omtrent de Research Drive. Daarnaast willen we wel 'aangehaakt' blijven bij de roadmap van SURF op het gebied van de mogelijkheden omtrent het metadateren en archiveren van onderzoeksdata. Dus we gaan intern aan de slag met de adviezen die uit de pilot gekomen die gericht zijn op onze interne processen en werking van de Research Drive. Daarnaast willen we graag in contact blijven met SURF voor wat betreft de doorontwikkeling van dit onderwerp. Door deze pilot hebben we al veel beter zicht gekregen op de wensen en eisen van de eindgebruikers en vanuit daar kunnen we t.z.t een goede keus maken welke techniek en mogelijkheid voor ons het beste geschikt is voor het metadateren en archiveren van onderzoeksdata.

### Impact op (het faciliteren) van praktijkgericht onderzoek

Vanuit deze pilot willen we verbeteringen aanbrengen in de Research Drive en in de werkprocessen die daarom heen hangen. Dit betekent dat onderzoekers nog gebruiksvriendelijker en efficiënter hun werk kunnen doen op het gebied van datamanagement. Daarnaast is het alleen mogelijk om datasets in de toekomst te dissimineren als alle voorafgaande processtappen rondom een dataset goed worden doorlopen. Uiteindelijk zullen deze aanpassingen ervoor kunnen gaan zorgen dat we in de toekomst meer datasets zullen publiceren wat het praktijkgericht onderzoek ten goede komt. In de pilot is geconstateerd dat we eerst moeten gaan werken aan optimalisatie van de huidige werking van de Research Drive binnen Saxion.

Details over metadatavelden, feedback vanuit onderzoekers en processtappen zie de bijlage aan het einde).

### **SURF-diensten**

Integratie van de iRODS techniek op de Research Drive

De Research Drive wordt binnen Saxion momenteel gebruikt voor het databeheer (met externen) tijdens een onderzoek. We hebben gekeken naar de toepassing van de iRODS techniek in de Research Drive zodat een eindgebruiker al tijdens (en na afloop van) het onderzoek, in de Research Drive, zijn/haar datasets kan metadateren en vanuit daar kan archiveren.

### **Betrokkenen en inzet**

#### Wie waren betrokken bij de pilot (rollen/functies)?

Projectteam, medewerkers van lectoraten, waaronder ondersteuners onderzoekers en lectoren en datasteward, key-user.

Deze pilot maakt deel uit van een intern Saxion project; het preserven en archiveren van data. De projectleider van dit project is ook het aanspreekpunt voor deze pilot. Daarnaast zijn de volgende functionarissen betrokken: meerdere onderzoekers, meerdere onderzoeksondersteuners, datasteward, business informatiemanager en architect

#### Welk lectoraat(en) of kenniscentra waren betrokken bij de pilot?

We hebben niet specifiek 1 kenniscentra of lectoraat betrokken. We hebben naar alle lectoraten een enquête uitgezet en we hebben meerder brainstormsessies gehouden waar ook vanuit meerdere lectoraten mensen aanwezig waren. De enquête is door 16 lectoren, 12 onderzoekers en 2 ondersteuners ingevuld. Bij de brainstormsessies zijn in totaal ook meerdere onderzoekers en ondersteuners bij betrokken geweest.

### Welke inzet en expertise is vanuit SURF geleverd?

Vanuit SURF is een demo gegeven omtrent de werking van de iRODS functionaliteit binnen de research Drive. Daarnaast hebben we met SURF de bevindingen besproken en heeft SURF ons voorlichting gegeven over een alternatief, YODA. Voor ons is overstappen op een andere tooling voor nu niet interessant, wel kunnen we leren van de inrichtingen en mogelijkheden binnen YODA en samen met SURF kijken wat toepasbaar is voor ons.

### Lessons learned

- Binnen Saxion hebben we met het gebruik van de Research Drive nu wel last van de wet van de remmende voorsprong. Er zijn veel onderzoekers bijgekomen sinds de start bij Saxion met de Research Drive, niet iedereen wordt evengoed meegenomen en ingewerkt en er bestaan nog te veel alternatieve dataopslag mogelijkheden al dan niet gedoogd. We zullen eerst weer moeten investeren in dat gebruik van de Research Drive en de ondersteuning van onderzoekers hierbij. Op kleine schaal verder gaan met meta dateer en archiveer opties.
- Een aantal van de huidige problemen lijken opgelost te kunnen worden door trainingen te volgen over de Research Drive vanuit SURF waarmee we de inrichting en het gebruik kunnen verbeteren. Daarnaast willen we voor eindgebruikers ook kijken hoe we ze beter kunnen voorlichten en scholen omtrent het gebruik van de Research Drive
- Beter monitoren dat nieuw onderzoekers bij Saxion hulp aangeboden krijgen en getraind worden bij gebruik Research Drive. Omdat er zoveel verschillen zijn in soorten data en hoe de verschillende onderzoeksgebieden omgaan

## Bijlage: Bevindingen rondom mogelijkheden & wensen metadateren en archiveren van onderzoeksdata

### 1. Niveau van metadateren

Het is gewenst om op meerdere niveaus te kunnen metadateren.

Vanuit de onderzoekers/lectoraten gaat de voorkeur uit naar het kunnen metadateren op meerdere niveau's (project, folder en document).

- Als je kijkt naar de voorkeur als men 'moet kiezen' is de voorkeur het 'folder niveau', daarna document niveau en daarna projectniveau.
- Er zijn zoveel verschillende situaties dat je eigenlijk wil dat het mogelijk is op alle niveau's. Het 'overnemen' van metadata zou wenselijk zijn, dus dat je het alleen hoeft aan te passen als er een verschil is in de meta-data-informatie.

### 2. Gebruikte metadatavelden

Algemeen: Ons uitgangspunt/belang van metadateren:

- Omdat de onderzoekers metadateren voor - vooral hergebruik van de data - moet hier rekening mee worden gehouden in de metadatavelden.
- Data moet ook gevonden kunnen worden voor onderzoeken die andere doelen hebben, maar waar de data van een ander onderzoek ondersteunend/ bruikbaar kan zijn. Hiervoor zijn meer velden nodig dan SURF nu aanbiedt. Denk aan taggs, verbanden met andere bestanden e.d.
- Het lijkt ons verstandig om voor de velden vooral aan te sluiten op de bestaande archief en publicatieplatforms van datasets. Een deel van de benodigde metadata zal automatisch gegenereerd kunnen worden vanuit de SRCD en specifieke informatie zal moeten worden uitgevraagd. Deze 1e set is een begin, maar zeker nog niet volledig en goed algemeen te gebruiken. Er zijn voor bepaalde vakgebieden uitgewerkte richtlijnen en standaarden beschikbaar (archeologie: regelgeving; pakbon KNA en richtlijnen in E-depot voor de archeologie (DANS-EASY)).
- Dit kan wel stap voor stap. Vanuit de publicatie-platforms zie je een nog uitgebreide set aan uitgevraagde metadata. Onze zienswijze is dat lectoraten/onderzoekers 1-malig hun dataset hoeven te metadateren voor het interne archief en waar nodig voor de publicatie extern. Vandaar dat het belangrijk is dat we ook voor dit laatste, de al gewenste metadatering zoveel mogelijk uitvragen. Omdat we nu nog niet gelijk automatisch gaan publiceren, zou dit een volgende stap zijn en is geen vereiste in deze versie.

Daarnaast zie je op de platforms nog specifiekere informatie over de dataset zelf en het ontstaan. Het advies is om te kijken naar deze richtlijnen bij een fase 2 van dit project. In eerste instantie gaan we toch niet vanuit hier publiceren en is het vooral gericht op een goede archief-functie.

Als we gaan publiceren zal de zoek- en vind functie veel belangrijker worden dus ook de specifiekere informatie, zoals tijdsperiode datacollectie, soort data, databronnen, software etc.

#### Specifiek

Huidige uitvraag	Feedback
Research ID	Belangrijk voor Saxion, voor interne koppeling onderzoeksdossier

Status	Kijken in hoeverre we tegemoet kunnen komen aan variëteit in definities De standaard (ingebouwde) categorisering raw, cleaned, final voldoet voor niet voor alle lectoraten/onderzoekers. Eén indeling lijkt niet haalbaar door de verscheidenheid aan data, omgang met deze data en verschil in behoeftes. Daarnaast is dit voor de verschillende publicatie-data-platformen meestal niet van belang/hanteren ze dit niet. Wel is het van belang om de 'status' van de data mee te kunnen geven. Maar dan zou een vrije invoer en/of meer categorieën gewenst zijn en/of andere benaming.
Title	Prima, geen feedback – zie je ook terugkomen bij publicatie-platforms
PID	Prima, geen feedback – zie je ook terugkomen bij publicatie-platforms
Authors	Belangrijke uitvraag, zie je ook terugkomen bij publicatie-platforms, risico: datakwaliteit van vrij invulveld + risico bij invul meerdere auteurs
Date of Creation	Prima, geen feedback – zie je ook terugkomen bij publicatie-platforms
Ownership	Nog inhoudelijk beoordelen
Use Rights	Vrij invulveld, is het niet beter om een paar opties mee te geven? Uitvraag belangrijk, nog beoordelen wat we uitvragen hierbij.
Release date	Prima, geen feedback – zie je ook terugkomen bij publicatie-platforms

#### Extra benoemde metadata-informatie door onderzoekers en onderzoeksondersteuners bij Saxion

Trefwoorden	Vooral belangrijk als we in de toekomst gaan publiceren. MAAR: welke gebruik je? Vrij invulveld is een risico. Kijken welke thema's DANS EASY, Figshare etc gebruiken!
Samenvatting	Vooral belangrijk als we in de toekomst gaan publiceren
Doel van verzamelen	
Verband met andere bestanden	
Onderwerp (evt subklasse)	Zie je ook in publicatie-platforms
Datatype + architectuur	
Wellicht een veld met mogelijke tags toevoegen, aangezien de datasets zo verschillend kunnen zijn en per vakgebied er andere eisen worden gesteld aan de benodigde metadata.	

### 3. Tooling voor inbedden in Saxion-werkprocessen

We hebben geconstateerd dat het volgende nog uitgezocht zou moeten worden:

Ons uitgangspunt is dat content niet op allerlei verschillende plekken wordt gearchiveerd. Dus het is belangrijk om na te denken over het beleid: wanneer archiveer je je datasets, hoe vaak archiveer je een dergelijke dataset etc. Soms lijkt het wel gewenst om juist de 'werksets' /sets bewerkt en/of in progress te archiveren en niet alleen de 'finale' dataset

We constateren dat het nu mogelijk is om op allerlei momenten een kopie te maken voor in het archief. Dit is niet wenselijk ivm de hoeveelheid data-opslag die dit gaat veroorzaken. We willen hier restricties op kunnen instellen.

- In het proces bv instellen dat iemand anders nog op de 'definitief-archiveren' knop drukt
- Wel tussendoor kunnen 'saven', maar niet gelijk archiveren
- Apart proces instellen van archief naar publicatie

#### **4. Belangrijke voorwaarden voor daadwerkelijke inbedding in het proces**

Als positief punt is benoemd dat het erg fijn is, dat voor de metadatering en archivering, er niet naar een andere applicatie geswitcht hoeft te worden.

Het verder bouwen op deze oplossingsrichting is belangrijk. Als in de toekomst hier echt gebruik gemaakt van gaat worden, en dat zal dan in de loop van de tijd zeker groeien, is het wel noodzakelijk dat ook het proces aan de achterkant automatisch gaat. Voor een structurele inbedding in het werkproces is dit wel een vereiste. Dus een koppeling naar de interne repository is dan een randvoorwaarde, net zoals een koppeling tussen de interne repository en de verschillende publicatie-platforms.