VR 2020 0412 DOC.0379/2

Fonds Wetenschappelijk Onderzoek - Vlaanderen

Open Data and research data infrastructure in the Flemish research environment

12 oktober 2020

pwc

Table of content

Managementsamenvatting 4

Introductie 4

Maturiteit 4

Vereisten 5

Potentiële oplossingen 7

Conclusies & aanbevelingen 8

Volgende stappen 9

Glossary 11

Introduction 13

Context and purpose of this document 14

Structure of this document 14

Process and method used 14

Maturity of Open Science and Open Data 15

Highlights 15

Demography of the research 17

Maturity levels 19

Requirements - Common themes 27

Technology 27

Governance 30

Standardisation 32

Legal 33

Support and training 34

Funding 35

Potential Solutions 36

Central Research Data Repository 36

Intermediate Layer 37

Standardisation 39

Non-architectural solutions 39

Conclusions & Recommendations 42

One size does not fit all 42

Do not reinvent the wheel again 42

Increase data awareness and clarity for researchers 43

Thematic and standardised approach 43

Further automation for improved efficiency and workload 43

Maturity needs to be further put into practice 44

Open where possible, securely closed where needed 44

Bring efforts made and successes to broader audience 44

Priorities 44

Next steps 45

Annex I: Questionnaire 46

Annex II: Interviews 51

Managementsamenvatting

Introductie

In de Vlaamse onderzoekswereld bestaan reeds verschillende initiatieven, standaarden en processen om op een zo uniform mogelijke manier met metadata en onderzoeksdata om te gaan. Dit rapport beschrijft de huidige stand van zaken rond beschikbare data infrastructuur alsook de maturiteit van verschillende onderzoeksinstellingen rond hun data management processen. Hiernaast beschrijft dit rapport eveneens welke noden deze organisaties hebben om op een betere manier met hun data om te gaan. Enkele potentiële oplossingen worden voorgesteld samen met de nog relevante openstaande vragen. Ten slotte worden algemene conclusies en aanbevelingen naar beleidsmaatregelen als volgende stappen voorgesteld.

Maturiteit

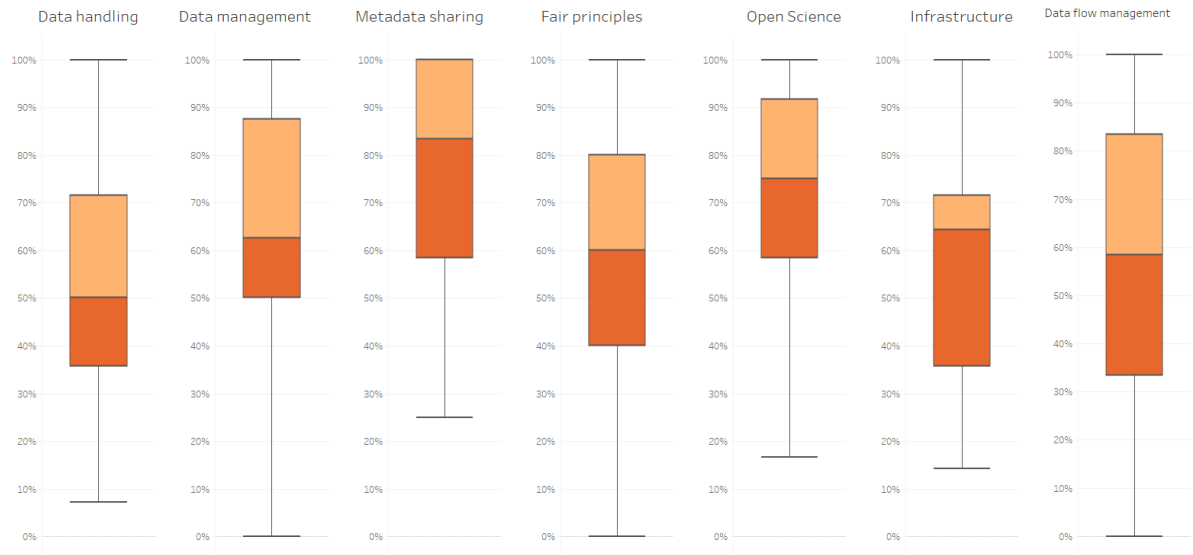
Om een overzicht te krijgen van welke Vlaamse onderzoeksinstellingen groeien in hun data management maturiteit, al metadata en onderzoeksdata delen, het juiste technologische landschap hebben, ... werden interviews uitgevoerd die gebaseerd waren op een vragenlijst. Deze gesprekken werden gevoerd met 35 instituten, waarvan er 33 in deze studie werden opgenomen.

Vervolgens werd een scoringsmodel ontwikkeld om de verschillende onderzoeksinstituten te beoordelen op hun maturiteit met betrekking tot databeheer, Open Data, het delen van gegevens, enz. De scores die aan elk van de instituten worden toegekend, zijn gebaseerd op hun eigen zelfevaluatie in de vragenlijst.

Zeven categorieën werden gekozen om de bestaande data-infrastructuur en het maturiteitsniveau van elk instituut te peilen. Deze resultaten voor deze categorieën zijn als volgt:

1. **Data behandeling:** Data behandeling belichaamt de processen om de onderzoeksgegevens te verwerken, de kwaliteit van de gegevens te controleren en problemen aan te pakken. De volwassenheid rond data handling is verdeeld, zoals te zien is in de onderstaande figuur. Van de 33 instituten hebben er 11 processen om met hun gegevens om te gaan, het datamodel en de kwaliteit van de gegevens, 15 instituten geven aan hieraan te werken en 7 instituten hebben niet de middelen of kennis om dit zelf te doen.
2. **Data management:** Het Vlaamse onderzoekslandschap is zeer gediversifieerd als het gaat om de governance en het beheer van de gegevens. Sommige instituten werken aan een volledig bestuursmodel, andere zetten nieuwe processen in voor de verschillende rollen en verantwoordelijkheden die in de organisatie zijn beschreven, en sommige maken alleen gebruik van de beschikbare diensten voor hun Research Data Management (RDM) / Data Management Plan (DMP). Wat echter duidelijk werd tijdens de interviews is het feit dat de meeste instituten willen groeien en hun governance model verder willen ontwikkelen naar de toekomst toe.
3. **Metadata delen:** 24 instellingen geven aan dat ze hun metadata delen met FRIS en andere organisaties via een beveiligde cloud omgeving, of via API's, en 8 geven aan dat ze van plan zijn om hun data in de nabije toekomst te delen. Slechts één instelling deelt zijn gegevens nog niet.
4. **FAIR-principes:** Over het algemeen zijn de meeste instellingen goed op de hoogte van de FAIR-principes en de richtlijnen. Ze zijn vaak ook ver gevorderd met de implementatie van de FAIR-principes voor metadata. Als we echter naar de onderzoeksgegevens zelf kijken, is het moeilijker om deze FAIR te maken. De trend in Vlaanderen is dat onderzoeksgegevens nog niet openlijk worden gedeeld, maar wel opgevraagd kunnen worden.

5. **Open Data / Open Science:** Bijna alle instituten zien de voordelen van Open Data en Open Science, maar de meeste weten nog niet hoe ze hun Open Datastrategie moeten omschrijven in de mix van gediversifieerde initiatieven van FOSB en EOSC. Deze doelstellingen moeten eerst worden verduidelijkt en verfijnd. Zoals in de visual te zien is, scoorden de meeste instituten goed. Het grote bewustzijn van Open Science en Open Data betekent echter niet dat er een groot Open Science en Open Data beleid is. Uit ons onderzoek kwam één verrassend kenmerk naar voren: de meeste instituten hebben Open Science en Open Data op hun agenda staan, maar het beleid en de praktijk worden niet afgedwongen.
6. **Infrastructuur:** In de categorie infrastructuur zien we dat niet elk instituut zijn eigen hardware beschikbaar heeft om zijn metadata en onderzoeksgegevens op te slaan en te archiveren. Veel instituten zijn afhankelijk van externe oplossingen en repositories wat de lagere score in maturiteit verklaart. Daarnaast is er ook een trend waarbij instituten samenwerken om technologische infrastructuur en hardware te delen.
7. **Beheer van gegevensstromen:** We zien dat veel instituten wachten op beslissingen van FRIS, FWO en FOSB alvorens verder te gaan met het beheer van de datastromen, bijvoorbeeld over het nieuwe metadatamodel, workflows, best practices richting standaarden, enz. De meeste instituten hebben een roadmap klaar voor de toekomst, maar wachten op de verdere implementatie ervan.



Om onze eerste bevindingen te valideren en consensus te bereiken tussen de onderzoeksinstituten hebben we twee interactieve workshops gehouden.

Vereisten

Dit stelde ons in staat om ten eerste de belangrijkste en meest noodzakelijke vereisten te identificeren en, waar nodig, discussies uit te lokken en te faciliteren en ten tweede om mogelijke oplossingen uit te werken om aan deze vereisten te voldoen. Deze vereisten kunnen worden samengevat in zes hoofddomeinen, namelijk:

- **Technologie:** Metadata en onderzoeksdata moeten best een enkele keer ingevuld worden en nadien zoveel mogelijk hergebruikt worden. De technologie stelt in staat om recurrente taken te automatiseren en interoperabiliteit tussen platformen mogelijk te maken.

  Naarmate de tijd vordert, zal de noodzaak om gegevens op te slaan, te verwerken en te archiveren exponentieel blijven toenemen. De verwachting is dat de hoeveelheid data die wordt gegenereerd over de overgrote meerderheid van de onderzoeksthema's enorm zal toenemen. Instellingen uiten daarom

hun bezorgdheid over hun toekomstige opslagcapaciteit om aan deze eisen te voldoen, inclusief de nodige rekenkracht.

Naast de noodzaak om de administratieve lasten te verminderen en de noodzaak van een gecentraliseerde opslag-, verwerkings- en archiveringsoplossing, werden enkele andere aandachtspunten met betrekking tot technologie en architectuur aangehaald, nl. een beperkte behoefte aan nieuwe platformen, en het afhandelen van de dubbele affiliaties in FRIS.

- **Governance:** In de huidige opzet ervaren veel onderzoekers de overvloed aan richtlijnen, regels en voorschriften als een last die hun mogelijkheden om zich te concentreren op hun dagelijkse werk beperkt. De administratieve overlast veroorzaakt frustratie en het idee werd meerdere malen genoemd dat er een manier zou moeten zijn om generieke informatie, die is ingevoerd voor verschillende andere doeleinden, te hergebruiken. Een evenwicht tussen administratie en de meerwaarde voor de onderzoeker moet hier gevonden worden.

  Naast de administratieve overlast, lijkt iedereen het eens te zijn over de complexiteit van het onderzoekslandschap zoals het is. Veel initiatieven worden gelanceerd op regionaal, nationaal, Europees en soms internationaal niveau. Dit zorgt voor uitdagingen, vooral omdat sommige initiatieven niet noodzakelijkerwijs op elkaar afgestemd zijn. De verschillende beleidsniveaus moeten meer op elkaar worden afgestemd en er moet met name een hiërarchie worden gedefinieerd.

- **Standaardisatie:** Het is duidelijk dat een 'one solution fits all' niet mogelijk is voor de Vlaamse onderzoekswereld. Hoe ingewikkeld de afstemming en standaardisatie ook blijkt te zijn, er bestaat een enorme hoeveelheid deskundige experts die de ins en outs van hun onderzoeksdomein kennen. Deze expertise moet worden benut bij het definiëren van (meta)datastandaarden en andere processen.

  Om de duplicatie van gegevensinvoer in verschillende repositories / toepassingen te vermijden, is er behoefte aan een gecentraliseerde laag die gegevens over meerdere repositories en applicaties verzamelt.

- **Wettelijk:** In een steeds meer geglobaliseerde en multidisciplinaire wereld wordt de mogelijkheid om samen te werken en gegevens te delen steeds belangrijker. Met de toenemende centralisatie moet men echter voorzichtig zijn met het intellectuele eigendom en de privacy van deze opslag gegevens. Bovendien verhogen de GDPR-richtlijnen de complexiteit van deze eisen, omdat ze een nieuwe laag van privacyregels toevoegen.

- **Ondersteuning en training:** Begeleiding en opleiding in het omgaan met data is eveneens een vereiste van de researchers. Het data management plan (DMP) vereist dat de onderzoeker de gegevens, die naar verwachting in een project worden verworven, op schrift stelt en hoe daarmee wordt omgegaan. Maar omdat DMP begon als een top-down uitnodiging, gaat het momenteel niet in op de specifieke behoeften van de verschillende disciplines en domeinen. Het hebben van DMP-tools is een belangrijke doorbraak, maar als de onderzoeker het niet goed invult, wordt het nut en het doel van de DMP aanzienlijk beperkt. Daarom is het nodig om onderzoekers te coachen bij het opzetten van DMP's.

  FAIR en Open Data vormen echter de basis voor herbruikbaarheid door andere domeinen. Het combineren van verschillende academische disciplines om wetenschappelijk onderzoek te bevorderen kan alleen worden gedaan dankzij Open en FAIR-gegevens. Dit moet als zodanig worden geadverteerd en gestimuleerd. Onderzoekers moeten veel tijd besteden aan het maken van FAIR data, maar zijn terughoudend omdat de voordelen van deze inspanning niet meteen duidelijk zijn. Daarom kwamen verschillende instellingen op het idee om onderzoekers of organisaties die hun datasets consequent openstellen op een open en gestandaardiseerde manier te stimuleren.

- **Financiering:** Er werd vermeld dat in het huidige financieringsmodel de verdeling en het mechanisme van het financieringsmodel als vrij generiek wordt ervaren. Bij de herziening van het model kan rekening worden gehouden met elementen zoals langlopende onderzoeksprojecten of zeer specifiek of geavanceerd onderzoek, en kan een passend evenwicht tussen grotere en kleinere instellingen worden gewaarborgd.

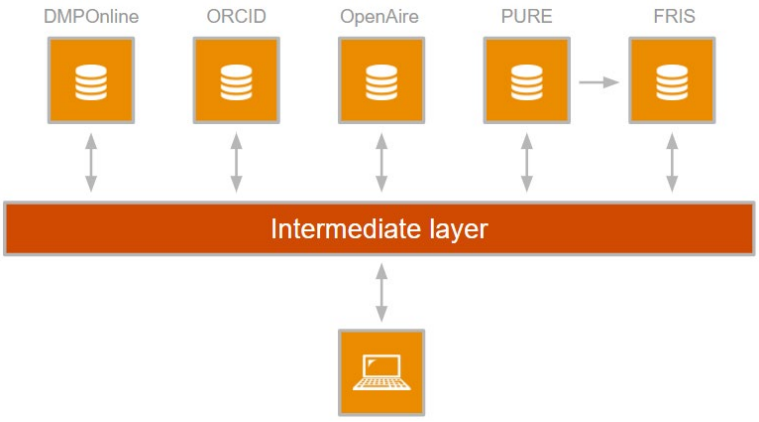
  Daarnaast geven veel instellingen aan dat ze problemen ondervinden met het openstellen van datasets met een opmerkelijke omvang. Het zou een grote toegevoegde waarde hebben voor de Open Science gemeenschap als de barrières om de grotere datasets te publiceren verdwijnen.

Potentiële oplossingen

De verschillende vereisten kunnen aangepakt worden door bepaalde oplossingen aan het gehele onderzoekslandschap te voorzien. Drie potentiële oplossingen werden uitgewerkt en voorgesteld. Bij elke oplossingen dienen nog verdere analyses te gebeuren om nog openstaande vragen verder uit te werken en te beantwoorden.

Een eerste mogelijke oplossing komt tegemoet aan de vereiste naar archiverings- en opslagcapaciteit van data. Een mogelijke oplossing hier is om een **centrale opslagplaats voor onderzoeksdata** te voorzien. Vaak worden onderzoeksgegevens opgeslagen in clouddiensten zoals SharePoint of Google Drive of op minder toegankelijke bronnen zoals een externe harde schijf. Daar waar er in bepaalde domeinen een thematische opslagplaats aangeboden wordt, wordt deze eveneens vaak gebruikt.

Dergelijke oplossing kan idealiter vorm gegeven worden door een *federated* structuur op te zetten door wat reeds bestaat met elkaar te verbinden, aangevuld met waar nodig nieuwe ontwikkelingen. Hierdoor worden dubbele investeringen vermeden en de kennis in het netwerk wordt aangesproken en gebruikt. Idealiter worden bestaande oplossingen binnen deze structuur opgenomen zodat minder mature organisaties op dit gebied van deze diensten en oplossingen gebruik kunnen maken. Dit zal organisaties helpen om hun onderzoeksgegevens op een efficiëntere en meer gestandaardiseerde manier te beheren.



Een tweede mogelijke oplossing pakt de grote overlapping in beschikbare data aan binnen bestaande platformen en die vanuit het perspectief van de onderzoekers veel vergelijkbaar administratief werk met zich meebrengen. Hierbij kan een **verdere automatisering** en het linken van de bestaande platformen met bijvoorbeeld een intermediaire laag die alle informatie met elkaar verbindt een oplossing bieden. Deze oplossing, die op Vlaams niveau kan voorzien worden, zorgt er dan voor dat informatie slechts eenmalig moet ingevoerd worden en kan hergebruikt worden in de andere platformen alsook in het genereren van documenten en formulieren die eerder

manueel werk vragen zoals CV's, informatie belangrijk voor de ethische commissie, en dergelijke. Deze intermediaire laag kan een overzicht bieden van welke informatie reeds beschikbaar is en welke informatie nog noodzakelijk is om te voldoen aan bepaalde vragen vanuit onder andere het FAIR perspectief. De belangrijkste functie van deze oplossing is dat er geweten is waar welk type informatie beschikbaar is en dat deze informatie op een zo efficiënt mogelijke manier hergebruikt wordt. Door dergelijke oplossing te voorzien op Vlaams niveau, krijgt de onderzoeker opnieuw controle over welke informatie men beschikbaar stelt aan welke interne en externe instanties. Door dergelijke oplossing niet overgecompliceerd te ontwikkelen, en puur te focussen op het verbinden en hergebruiken van bestaande informatie, kan dit toegepast worden binnen alle mogelijke onderzoeksdomeinen en door onderzoekers die actief zijn bij zowel kleinere als grotere organisaties.

Op het vlak van **standaardisatie** zijn eveneens potentiële oplossingen mogelijk. Om een gefedereerde oplossing te kunnen creëren is een hogere graad van standaardisatie noodzakelijk in het hele onderzoeksecosysteem zodat deze samenwerking efficiënter en makkelijker kan verlopen. Binnen bepaalde domeinen worden reeds verschillende standaarden afgesproken over bijvoorbeeld het gebruik van API's, authenticatie, datamodellen en dergelijke. Ook hier moet gebruik gemaakt worden wat er reeds bestaat en goed werkt en verder gewerkt worden aan een efficiëntere werking waar nodig en relevant.

Hiernaast kunnen een aantal oplossingen gevormd worden die een meer ondersteunende rol aan het onderzoekslandschap invullen. Allereerst kan een **opleidingsplatform** zorgen voor een groter bewustwording en uniformiteit rond data. Hierbij kunnen onderzoekers opgeleid worden maar evenzeer is het concept van 'train de trainer' belangrijk. Dit platform kan eveneens bestaande oplossingen promoten en aangeven wat de minimumeisen zijn om te voldoen aan de verwachtingen rond databeheer. Deze oplossing sluit aan bij het idee van competentiecentra die momenteel op EOSC niveau worden opgezet. Een tweede oplossing bestaat eruit om de implementatie van correct databeheer, Open Data en het volgen van de FAIR principes te **stimuleren**. Dit kan door bijvoorbeeld een slim beloningssysteem te implementeren of door organisaties een platform te geven waar ze hun succesverhalen kunnen delen met anderen. Hiernaast wordt het voorzien van data stewards gezien als een nuttige en noodzakelijke aanvulling net als het creëren van een netwerk over heel Vlaanderen waar ervaringen, expertise en technologische oplossingen met elkaar besproken en gedeeld worden.

Conclusies & aanbevelingen

Als resultaat van deze studie kunnen we stellen dat het **onderzoekslandschap in Vlaanderen zeer divers** is. Zowel wanneer gekeken wordt naar de aanwezigheid van specifieke profielen en kennis rond data management als naar IT infrastructuur. Daar waar organisaties enerzijds een eigen systeem hebben gebouwd en dit reeds jaren in gebruik hebben, vertrouwen anderen op de ondersteuning van universiteiten of Europese netwerken om hierin begeleid te worden. Er bestaat dus reeds veel expertise en oplossingen in het landschap, waardoor nieuwe oplossingen bouwen niet de eerste keuze dient te zijn. Het **maximaal hergebruiken van wat reeds bestaat** is dan ook belangrijk en mogelijk om geen dubbele investeringen te moeten uitvoeren, alsook om maximaal kennis te delen en samenwerkingen te bevorderen. Om dit in de praktijk te kunnen realiseren zal een goed governance model nodig zijn, om te afspraken te maken rond de manier waarop meer ervaren organisaties als dienstverlener of als mentor kunnen instaan voor andere.

Zoals aangegeven is het onderzoekslandschap zeer divers en kunnen de volgende stappen in het aligneren van verschillende standaarden, werkprocessen, technologische oplossingen en dergelijke gebeuren door een **thematische aanpak** toe te passen. Door allereerste te werken op de verschillende onderzoeksdomeinen kan voldoende flexibiliteit behouden blijven en tegelijkertijd gezorgd worden voor maximale afstemming en samenwerking, ook tussen de verschillende ESFRI's.

Het maximaal hergebruiken van wat reeds bestaat is niet enkel van toepassing op expertise en technologische oplossingen, maar evenzeer op het gebruiken van informatie rond het onderzoek zelf, gegevens van instellingen en onderzoekers, project informatie en dergelijke. Onderzoekers geven dergelijke informatie op dit moment vaak

manueel in verschillende systemen in wat veel manueel werk veroorzaakt. Het **verder automatiseren van administratieve processen** door bestaande informatie te hergebruiken maakt het proces om CV's, GDPR documentatie, Data Management Plannen of documentatie voor Ethische commissies veel efficiënter en zal onderzoekers toelaten meer tijd te investeren in waarde toevoeging.

Op basis van de maturiteitsscores kan gesteld worden dat er reeds belangrijke stappen in de goede richting gezet zijn. Veel organisaties hebben kennis over FAIR en Open Data, hebben een datastrategie opgemaakt, en dergelijke. De volgende stap is echter om de vergaarde **maturiteitsniveaus verder om te zetten in de praktijk**, waarbij de strategieën uitgevoerd worden, ondersteuning ontwikkeld wordt, data actief gemonitord wordt naar hun FAIR en open karakter, enz.

Het is duidelijk dat niet alle data zomaar open kan zijn. Er dient gestreefd te worden naar **zo veel mogelijk Open Data waar mogelijk, maar veilig gesloten waar dit nodig is**. Er zijn steeds situaties waarin data gesloten moet blijven en dit moet mogelijk zijn wanneer er bijvoorbeeld gewerkt wordt met zeer gevoelige data zoals medische gegevens of privacy gevoelige informatie. Ook deze datasets kunnen op een optimaal georganiseerde centrale locatie opgeslagen worden, zolang deze maar voldoende beschermd zijn door hoogwaardige identificatie- en authenticatie modules.

Hiernaast dient de **bewustwording** rond data aspecten bij onderzoekers vergroot te worden. De onderzoekers zelf zijn geen data experten en hen informeren en betrekken rond alle data gerelateerde initiatieven is dan ook belangrijk. Dit gebeurt idealiter in makkelijk verstaanbare en duidelijke te onderscheiden termen en kan ingevuld worden door informatiesessies en algemene ondersteuning te organiseren. Hierin kan een rol voor FWO, of de 'coordination hub' meer specifiek, weggelegd zijn. Eveneens kunnen hier organisaties uit het landschap in betrokken worden die onder andere hun succesverhalen en ervaringen delen met anderen. Zoals vermeld bij de maturiteitsscores zijn er reeds veel inspanningen gebeurt die zich vertalen in de kwaliteit van het geleverde werk, rekening houdend met alle data aspecten die hierbij nodig zijn. Het is dan ook belangrijk om de **geleverde inspanningen tot bij een breder publiek** te brengen. Er kunnen ondersteunende activiteiten opgezet worden die de zichtbaarheid van de geleverde inspanningen vergroot en anderen hierin stimuleert. Dit kan door succesverhalen te tonen maar evenzeer door een kwaliteitslabel te creëren of het organiseren van *proof of concepts* waarbij geëxperimenteerd wordt met data op domeinoverschrijdend niveau.

Volgende stappen

De verschillende onderzoeksinstellingen kunnen in de toekomst verder geholpen worden in eerste instantie te **bekijken welke technologische oplossingen mogelijk zijn op het vlak van archivering en opslag** van data. Hierbij dient gekeken te worden welke bestaande infrastructuur oplossingen met elkaar kunnen verbonden worden om ervoor te zorgen dat deze passen binnen een *federated* structuur. Bij het uitvoeren van een analyse welke oplossingen nog nodig zijn dient rekening gehouden te worden met het diverse landschap zodat alle soorten organisaties ondersteuning krijgen in hun meest kritische vereisten.

Hiernaast kan er ten tweede een analyse gebeuren hoe er **verder geautomatiseerd kan worden op het vlak van metadata en onderzoeksgegevens**. Hierbij dient eveneens gekeken te worden hoe databanken die administratieve en projectgegevens bevatten met elkaar kunnen gelinkt worden, wat de administratieve werklast zal doen afnemen. Hiernaast moet metadata eveneens gedistribueerd worden naar FRIS en naar EOSC. Verdere analyse dient nog te gebeuren hoe metadata uit FRIS kan doorstromen naar EOSC zonder additionele werklast te veroorzaken voor onderzoekers of instellingen. Hierbij dient eveneens rekening gehouden te worden dat de evaluatie op Vlaams niveau zal gebaseerd zijn op de in FRIS beschikbare metadata.

Een derde traject betreft het **ontwikkelen van informatie en diensten die het bewustzijn rond (meta)data vergroot**. Naast het opzetten van informatiesessies en -materiaal kan er verder gebouwd worden op de bestaande expertise in het netwerk. Het uitwerken van een kwaliteitslabel en *proof of concepts* organiseren zal

de zichtbaarheid en het begrip van bepaalde data gerelateerde aspecten vergroten. Hiernaast dienen er eveneens diensten ontwikkeld te worden, zoals het voorzien van data stewards, die organisaties kunnen helpen met meer technische uitdagingen, zoals het connecteren met platformen zoals FRIS.

Glossary

The following terms and abbreviations are used in this document.

Glossary	
API	Application programming interface
AWS	Amazon Web Services
CERIF	Common European Research Information Format
CRIS	Current Research Information System
DMP	Data management plan
DOI	Digital object identifier
ECOOM	Expertisecentrum Onderzoek en Ontwikkelingsmonitoring
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
FOSB	Flemish Open Science Board
FRIS	Flanders Research Information System
FWO	Fonds Wetenschappelijk Onderzoek - Vlaanderen (Fund for Scientific Research)
GDPR	General Data Protection Regulation
HPC	High performance computing
IoT	Internet of Things
ISO	International Organization for Standardization
LDAP	Lightweight Directory Access Protocol
ORCID	ORCID is a non-profit organisation that provides a service to uniquely register and identify all who participate in research, scholarship and innovation across disciplines and borders.
OSLO	Open Standaarden voor Linkende Organisaties

PI	Personal Investigator
PID	Persistant IDentifier
REST	Representational State Transfer - an API protocol for data exchange
RDM	Research Data Management
RFP	Request for proposal
SOAP	Simple Object Access Protocol - an API protocol for data exchange
SSO	Single Sign On
VSC	Vlaamse Supercomputer Centrum

Introduction

Context and purpose of this document

The main objective of the Fonds Wetenschappelijk Onderzoek – Vlaanderen (FWO) is to stimulate and support fundamental and strategic scientific research both at the universities of the Flemish Community and between these Flemish universities and other research institutions. One of the challenges is to allow the Flemish research community to join and actively participate in the European Open Science Cloud (EOSC).

EOSC allows scientific research data to comply with the so-called FAIR principles (Findable, Accessible, Interoperable and Reusable). Connecting to and applying the FAIR principles brings a lot of advantages:

- Reduction of research duplication, in terms of time, effort and financing.
- Better management and manageability of digital resources and information.
- Possibility of focusing research more on activities with added value, such as interpreting data instead of searching, collecting or recreating already existing data.
- Savings of the total research cost and ownership of data by reusing solutions and data sets.
- Data is becoming more and more machine readable.

In order to make scientific research, financed with public funds, publicly accessible as quickly as possible according to the principle 'as open as possible, as closed as necessary' and to unite all Flemish stakeholders in a shared vision towards Open Science and the EOSC, it was decided that in 2020 the Flemish Open Science Board (FOSB) was established. FOSB makes it possible to use the existing expertise to consolidate and integrate Flanders in this international trend towards Open Science.

Following FOSB's advice to allocate the remaining budget by the end of October 2020, with a focus on data infrastructure, FWO has released an RFP to provide an overview of the needs and wishes of the stakeholders in Flanders concerning data infrastructure. It is important that for each institution, both current and future data requirements are captured in terms of storage and data services to meet EOSC compliance requirements.

This document describes the outcome of the conducted high-level data landscape assessment, and provides a mapping of the existing data infrastructure in the Flemish research world. It can be used as a basis to develop a first roadmap for implementing improvements and allocating the budget.

Structure of this document

This document consists of the following sections:

1. Introduction (this section) contains the context, purpose and structure of this document. In addition, it briefly elaborates on the process and method used to conduct this study.
2. The next section outlines the maturity of the different institutes towards Open Data, Open Science and data management.
3. Section 3 depicts the needed requirements from all participants, categorized into six dimensions.
4. Potential solutions are described and proposed based on all captured needs.
5. Conclusions and recommendations are given on which next steps should be undertaken.

Process and method used

During September and October 2020, the project conducted a high-level assessment of the existing data infrastructure in the Flemish research world, a maturity assessment and a gap analysis. In order to exclude any potential bias during the course of the project, objective and neutral analysis were performed based solely and exclusively on the input received from the institutes during the interviews, questionnaire and workshops. Following approach was used:

1. We first analysed the predetermined questionnaire prepared by FOSB, and analysed the different needs and requirements with regard to Open Science in the Flemish government and European Commission.
2. With these requirements, a final questionnaire was developed and sent out to all participants (37 research institutions in Flanders).
3. An interview was conducted with these participants to clarify any remaining questions and requirements.
4. A scoring model was developed to assess the different research institutes on their maturity towards data management, Open Data, data sharing, etc. The scores assigned to each are based on their own self-assessment in the questionnaire.
5. To validate our initial findings and reach consensus amongst the research institutes we conducted two interactive workshops. This allowed us firstly to identify the most important and necessary requirement and, where necessary, to provoke and facilitate discussions and secondly to elaborate on potential solutions to meet these requirements.
6. All questionnaires, detailed feedback, discussion points and conclusions were captured and included in the final report.

All these steps were done with a continuous feedback loop between the PwC team and FWO. The approach of the research and the study itself is based on some important principles, which are briefly explained below:

- **Collaboration**: A clear collaboration with the different research institutes in order to create as much support as possible and to guarantee the highest possible response rate. In order to achieve this cooperation, the following action points were taken:
    - In the questionnaire, the input of the participants of the working group 'Architecture' was used to make any adjustments.
    - A webinar to instruct the participants on the purpose of this assignment and to prepare them for the questionnaire and following interview that was held.
    - The research was done on the basis of a self-questionnaire, in which the entities themselves had the opportunity to fill in their gaps, needs and requirements, in order to increase involvement. The answers provided were further elaborated and detailed during the interviews.
- **Corrections:** To increase the relevance of the study and maturity assessment some corrections have been made. These corrections include the exclusion of two institutes who made clear the study was not relevant for them.
- **Comparability**: The questionnaire will gauge the existing data infrastructure in the Flemish research world and its maturity. In order to assess the different research institutions on different dimensions (i.e. data management, sharing, Open Data, technology, …) we used a scaling as followed:
    - 0 - Not being considered.
    - 1 - Under consideration or in the planning phase.
    - 2 - Fully implemented.

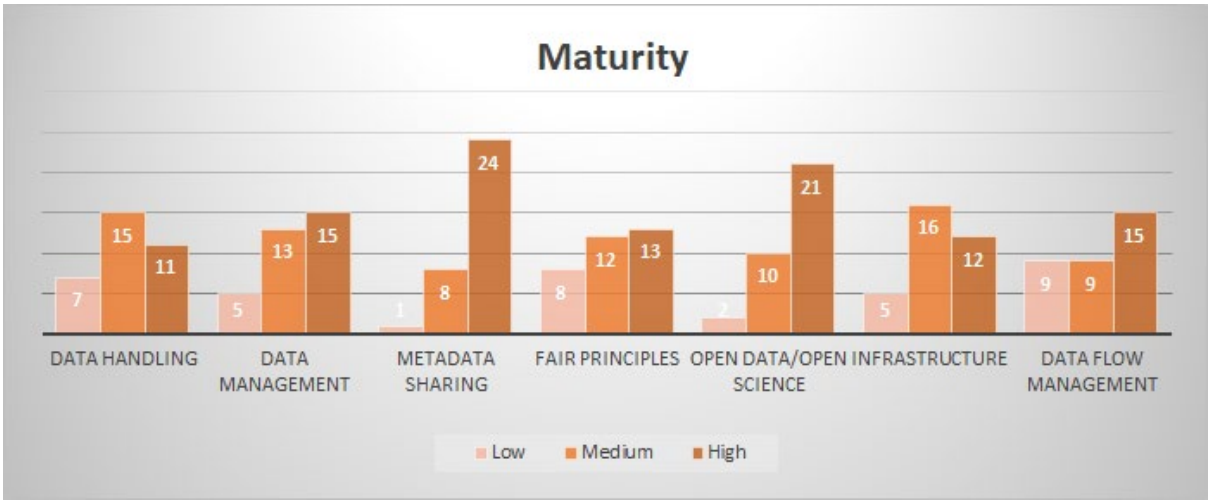
Maturity of Open Science and Open Data

This section discusses the most important findings with regards to the objective of the research, i.e. to determine and analyse the maturity and readiness of the Flemish research world towards Open Science and Open Data. This chapter is divided into the following sections:

- The highlights that emerged from the research, i.e. is the Flemish research world ready for EOSC.
- The demography of the research, i.e. which institutes contributed to the research, what is the response rate, what is the balance between the different participants, ...
- The general results of the research, as well as the results and findings on the maturity per category.

Highlights

In order to get an overview of which Flemish research institutes are growing in their data management maturity, are sharing data already, have the right technological landscape in place, ... , a self-questionnaire was conducted amongst 35 institutes, of which 33 were included in this study. The input of two institutes is missing based on the different focus they have. Including them would give wrong overall results. In this sense, the figures illustrated graphically below only relate to these 33 institutions.



In general, we see the following results:

- Data handling embodies the processes in place to handle the research data, check the quality of the data and tackle issues.The maturity around data handling is divided, as can be seen in the figure above. Out of the 33 institutes, 11 have processes in place to handle their data, the data model and the quality of data, 15 institutes indicate that they are working on improving this, and 7 do not have the resources or knowledge to do this themselves.
- The Flemish research landscape is very diversified when it comes to the governance and management of data. Some institutes are working on a full governance model, others are putting new processes in place for the different roles and responsibilities described in the organisation, and some only use the Research Data Management (RDM) / Data Management Plan (DMP) services available. However, what

became clear during the interviews is the fact that most institutes want to grow and further develop their governance model towards the future.

- 24 institutions are indicating that they are sharing their metadata with FRIS and other organisations through a secure cloud environment, or through APIs, and 8 mention that they are planning to share their data in the near future. Only two institutions do not share their data yet.
- Overall most institutes are well aware of the FAIR principles, the guidelines and are well advanced in their implementation of FAIR principles for metadata. However, when we look at the research data itself, it is more difficult to make this FAIR. The trend in Flanders is that research data is not yet shared openly, but can be accessed when requested.
- Almost all institutes are seeing the advantages of Open Data and Open Science, but most are not sure yet how to depict their Open Data strategy in the mixture of diversified initiatives by FOSB and EOSC. These goals need to be clarified and refined first.
- In the category infrastructure, we see that not every institute has their own hardware available to store and archive their metadata and research data. A lot of institutes rely on external solutions and repositories. Next to this, there is also a trend where institutes work together to share technological infrastructure and hardware.
- We see that a lot of institutes are awaiting decisions from FRIS, FWO and FOSB before going any further with data flow management, e.g. on the new metadata model, workflows, best practices towards standards, etc. Most institutes have a roadmap ready for the future, but wait to implement these further.

It is worth noting that almost all Flemish research institutes are in the process of improving their data management and data landscape in order to receive and share data with others, i.e. FRIS, EOSC, etc. However, due to the nature of the different institutes and the research they perform, the data landscape in Flanders is very diversified. This means that a 'one size fits all'-solution will not be possible, and that existing solutions throughout the landscape must be reused and integrated as much as possible to avoid double investments, both at the level of metadata and research data.

Strengths	**Weaknesses**
• The benefits and advantages of Open Data and Open Science are known within the institutes. • All institutes will be able to share (meta)data with FRIS in the future.	• Researchers themselves are not always convinced on the added value of sharing data. • Use of PIDs is not yet formalized and fully integrated in the working of the institutes. • Overall goals of FOSB and EOSC are not yet refined enough. • Data governance (incl. processes, reporting, ...) is not mature enough in the Flemish research landscape.
Opportunities	**Threats**

• A lot of technical solutions already exist throughout the landscape. • Alignment of (metadata) standards and other initiatives at European / national / local level. • Reduction of double work and administrative labor. • Economies of scale for large and long term data storage solutions.	• A 'one size fits all'-solution does not fulfill the needs and requirements from different research domains. • Standardisation without management and enough flexibility loses its purpose. • No clear governance, decision-making and policies are in place to identify the single source of truth in a federated landscape. • GDPR and IP concerns.

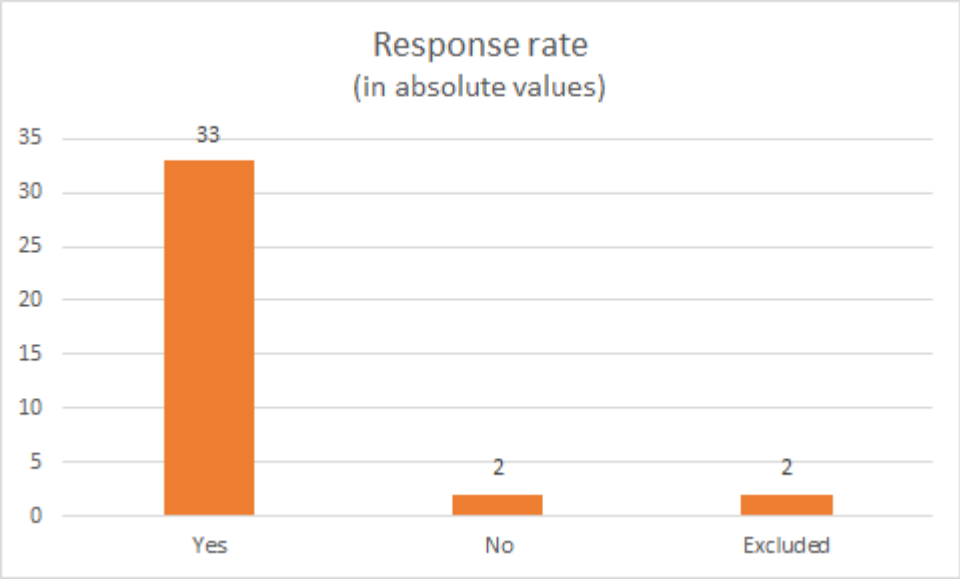
Demography of the research

The relevance of a benchmark study depends on the representativeness of the answers for the entire Flemish research world. That is why in this chapter we look at the response rate, the maturity model used and the balance between the bigger and smaller institutions.

Response rate

The response rate of this survey was very high. As described in the table below, 35 of the institutions answered the questionnaire correctly and participated in the interviews, of which two institutes, who made clear the study was not relevant for them, were excluded from the results. Only a small minority (two institutions) answered little or nothing.

This means that in total 33 institutes were included in the study.



The response rate of this study provides a representative sample of responses on the data infrastructure and allows us to draw relevant conclusions for the entire Flemish research world.

Maturity model

In order to assess the readiness of the Flemish research landscape to share their metadata and data, and to get an overview of the maturity of the different institutes towards Open Data and Open Science, amongst other aspects, a maturity scoring model was developed.

Based on the questionnaire, seven categories were chosen to gauge the existing data infrastructure and maturity level of each institute. These categories are :

1. **Data handling:** How good is the institute handling their data? This category embodies the processes in place to handle the research data, check the quality of the data and tackle quality issues
2. **Data management:** How far is the institute progressing towards a secure and complete data management programme? This category includes the different roles and responsibilities described in the organisation, the governance model place, as well as the use of RDM / DMP services.
3. **Metadata sharing:** Is the institute sharing its metadata? This category assesses how familiar the institute is with FRIS, their goals, and if they already share certain metadata (and research data) with other parties.
4. **FAIR principles:** How familiar is the institute with FAIR principles? This category explores how institutes are implementing the FAIR principles and how they are making their data interoperable and reusable.
5. **Open Data / Open Science:** Does the institute know FOSB, EOSC, their future goals and has the institute a strategy in place about Open Data? This category provides an insight into the maturity of the institutes towards Open Data and Open Science strategies.
6. **Infrastructure:** What is the technological landscape of the different institutes? This category shows how advanced the institutes are in their technology and hardware to store and archive (meta)data.
7. **Data flow management:** Does the institute have any workflows in place to handle data flowing through the lifecycle? This category answers if the institute has a metadata model in place and uses workflows and standards.

It is evident that not all the questions could be fitted in one of the above mentioned categories. Insights gained based on qualitative aspects will be discussed later in the document.

The model uses the following compliance scaling, to the extent possible, to depict how mature each institution is with regards to a specific topic (e.g. *to what extend metadata and research data is accessible today)*:

- 0 - Not being considered
- 1 - Under consideration or in the planning phase
- 2 - Fully implemented

Depending on the topic gauged, the description of the compliance scale was adapted. In addition to this scaling exercise and with the purpose to give a more granular picture of the maturity of the Flemish research landscape, the topics were weighted based on their importance.

Remark
The scores of each institute are based on their own self assessment of questionnaire. During the interviews additional questions were asked to complete the model. The maturity depicted in this study represents their current situation compared with what is expected now, not their ideal situation in the future.

Balance between research institutes

As indicated earlier, an approach with self-questioning has been chosen. This ensures a high level of involvement of the surveyed institutes but also requires some form of validation of the results which was done during the interviews. It became clear that a balance needs to be made between the different research institutes, the size of the organisation and the research (output) performed.

"We do not have the manpower or resources"

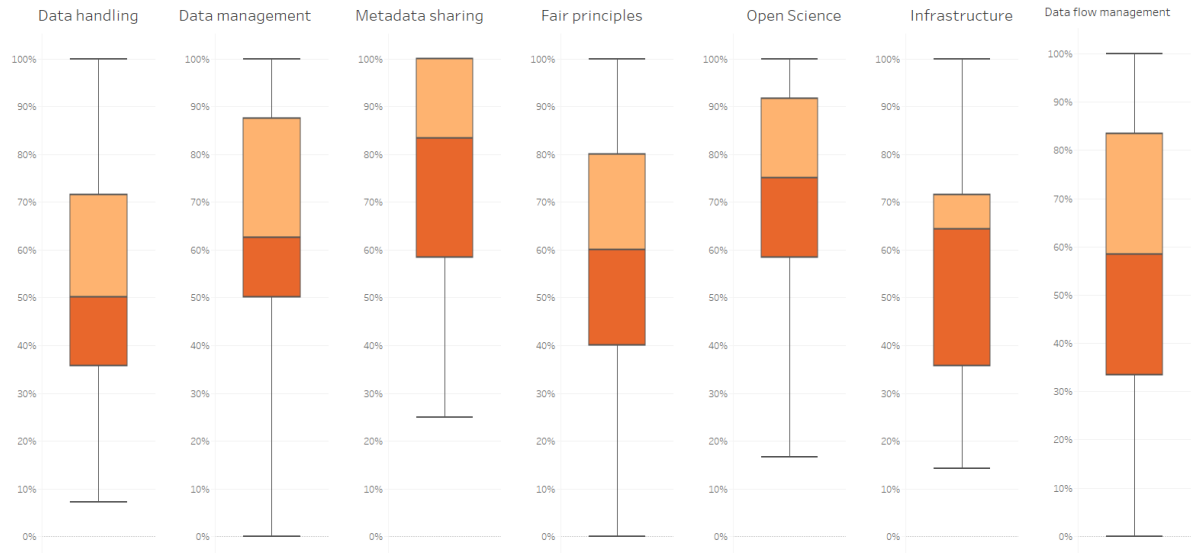
On one side, smaller research institutes often do not host their own data architecture but leverage on existing market solutions, or on the architectural landscape of bigger institutes. This shows in the results from the questionnaire and consequently also in the maturity level of these institutes, specifically in the category 'technology'. These smaller institutes often don't have the same resources available to further develop their own architectural landscape.

On the other hand, the bigger institutes also create and provide more research data and metadata compared to the smaller institutes. Therefore, their need for technological solutions to manage and store all the data is also larger, which means they have a large need for additional funding. This increased need for funding may also apply to smaller organisations that are active in a domain where they typically need to cope with large datasets.

This situation depicts a sensitive and delicate balance that has to be taken into account in the study. All requirements and needs of the different research institutes are important, but can vary depending on their current situation.

Maturity levels

The result of this maturity evaluation is translated into box plots, as shown below. Each of the categories described in the previous question has its own box plot. The percentage axis describes how well an institute scored compared to the highest score. Concretely, if an institute has a score of 25 and the highest score is 30, this is expressed as 83%, which means that the institute scored relatively well (e.g. predominantly *1 - Under consideration or in the planning phase* and *2 - Fully implemented* maturity levels for the topics identified). If we look at the full scale a score of 0% translates to the maturity for all the topics being *0 - Not being considered*.



How to interpret the results of the box plots?

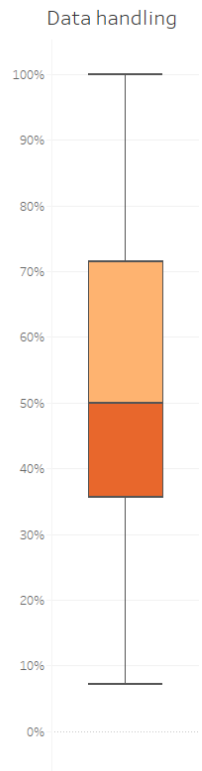
Considering the *data management category*, there are multiple insights that can be derived from the graph. First of all, if one looks at the *median*, it indicates that half of the institutes scored above ~60%. In other words half of the respondents are advanced in terms of data management. As the median is closer to the bottom of the box, the distribution is positively skewed.

When looking at the quartiles, the lower quartile – the bottom end of the box – indicates that 25% of the institutes surveyed have a score below ~50%. Concerning the upper quartile – top end of the box, it indicates that 75% of the respondents have a score below ~90%. The other way around only 25% of the respondents have a score above ~90%.

The top whisker illustrates the highest score, which is in our case 100%. On the other hand, the bottom whiskers illustrate the lowest score, which is 0%. In the case of *infrastructure* it is approximately ~15%. The box represents 50% of the institutes, one can easily deduct that 50% of the respondents score between ~50% and ~90%.

Generally speaking it seems that the Flemish landscape is well advanced in terms of metadata sharing, Open Science and Open Data as well as data management. On the contrary, infrastructure and data handling are two categories which testify of a less significant development.

Data handling



The data handling category measures the presence of processes to handle the management of research data that was looked into (e.g. maintenance of the data model, maintenance of code lists and vocabularies, etc.). Institutes did not score well on that topic. It appears that sometimes the researchers have the sole responsibility of managing the data. Some institutes don't even look into data handling as it is supposedly the responsibility of the repository, or third party managing and storing the data. Centralised processes of data handling and data governance are clearly lacking.

"The construction and maintenance of the data model (Common Data Service) is centrally provided for all Flemish universities. The update and use of the ECOOM classification, metadata standards, PID, etc. is organised centrally"

One of the most important aspects of data management is data quality. While data quality processes might not be implemented in all institutes, it is at least on the agenda of most of the institutes. For the institutes that scored well on this, most of them have automated checks on data (e.g. with the use of tools). It was already reported that data quality checks are done throughout the data life cycle and not specifically when data are about to be published. When institutes are about to publish data they are often looking at (community) standards, which is guarantee of quality.

Obviously, sometimes the checks are also manual. However, it is clear that when quality checks are performed, they are automated to the extent possible. It was also mentioned that data quality is quite difficult to check because one needs to agree on what quality is first.

Furthermore, we investigated whether processes to handle quality issues were in place. Unanimously, quality issues are addressed when they arise and if reported at a later stage, it often comes down to the personal investigator (PI) or researcher responsible for data. Sometimes quality goes hand in hand with accredited or certified repositories. That aspect is further discussed in the technology section.

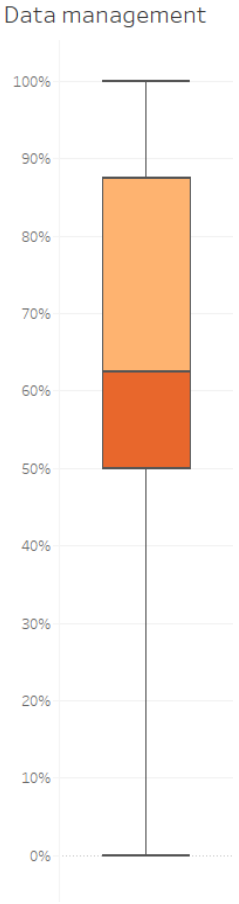
"The quality control does not take place afterwards but step by step during the execution. Of course it remains dependent on the commitment of the researcher to actively work with it"

The box plot depicts a situation where 75% of the research institutes scored less than ~75%. It is also clear that 50% of the institutes scored less than ~50%, which calls for action in terms of data handling.

Data management

Data management comprises all the activities and disciplines related to managing data as a valuable resource. In this maturity model, we measured different, yet specific aspects of data management. For starters, we tried to unveil whether all institutes had specific roles related to data management. It appears that the answers were quite disparate. Sometimes the institutes have all the roles needed to ensure proper data management (e.g. data curation, data quality, etc.) while some others are practicing data management on top of top of their actual role or even not at all.

Another equally important element is data management plans (DMP). Out of all the data management aspects, this is the one where all institutes scored best. Most of the institutes are using DMPs. Either they have an internal tool or use a third party tool (e.g. DMP online, etc.). Despites making use of them, the interviews unveiled key issues about the administrative burden and simplification, which are discussed more in detail in the requirements section.



To ensure consistent and proper data management, there is the need of policies and procedures. They serve the purpose of guiding and decisions and actions to see to a proper management of data. Once more the responses are quite diverse. Some institutes have highly standardised procedures and policies for data sharing, data life cycle while some others don't have policies nor procedures. Nonetheless, in some cases where there is no internal procedure nor policy, the institutes remarked they used someone else's (e.g. small institutes relying on the policies of a university, as they are storing and publishing their data there).

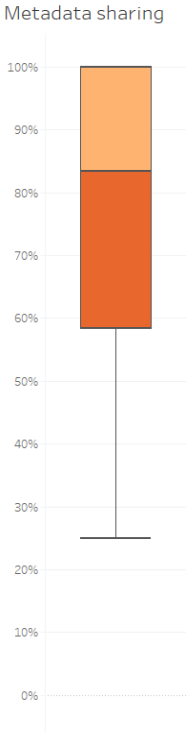
"The working methods already described will be digitally documented and included as much as possible within [the organisation] so that the users will also be helped to respect the rules during the research. In the elaboration of a process, governance roles and business rules are taken into account to steer the execution."

Lastly, it was made clear that there is no governance or supervision of the use of research data although it might be feasible for some institutes. While data sharing is quite common, the means is less common. Some institutes share data on specific cloud environments (e.g. SharePoint), some other through the portal where the data is stored, etc. To summarize, different solutions are used and data sharing is often done on an ad hoc basis.

In conclusion, compared to the other categories and topics measured, data management practices are the more disparate in the institutes surveyed. Most of the institutes scored between ~50% and ~90%, which might not be deduced when reading the above paragraphs, for one reason; the size of the institute, more often than not, impacts how well it scores in terms of data management. Smaller institutes tend to rely more on the larger institutes to get help on data management aspects. They also assume that data quality checks are done when they are depositing their data on larger institutes' storing facilities. While they may not take care of the data management related tasks,

the latter are still being done – for most of them – at another level. Nevertheless, almost all institutes indicate that governance and management of data is an important topic where they can and want to grow in the future.

Metadata sharing



Regarding metadata sharing, it seems, based on the box plot, that the institutes are well advanced in metadata sharing. For most of the institutes metadata is shared through a cloud environment with the use of APIs. One aspect of this category is whether institutes already share metadata with FRIS. Again most of them are aware of FRIS and its mandate, and most of them are already sharing their metadata with FRIS.

Lastly, as mentioned, institutes automatically share their metadata with FRIS as well as with other parties (e.g. journals, domain-specific repositories, etc.). Only a handful of institutes do not share metadata at all. For the ones that don't, there are various reasons which are detailed in the requirement section.

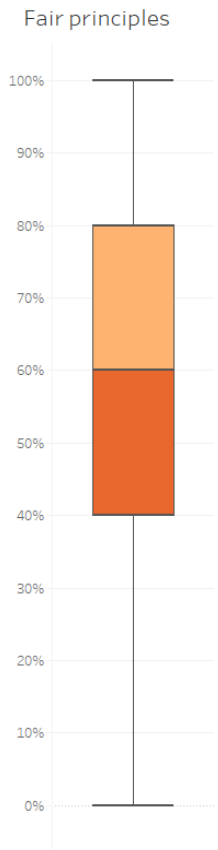
A closer look at the box plot reveals that 75% of the institutes surveyed scored more than ~60%, which results in metadata sharing being the most advanced category.

"Every researcher decides for himself when and how to share his research data"

However, in the questionnaire we also asked about the sharing of research data. There, we see that most institutes also share their research data, but not as good as the metadata. Most researchers are still in control of their own research data and are afraid to share their data too early. They want to decide for themselves when the data is ready to be shared. As a consequence research data is shared after completion of the project, and often only upon direct request.

FAIR principles

When looking at the FAIR principles, overall most institutes are well aware of the guidelines and well advanced in their implementation. It is important to outline that three different topics were measured: (i) FAIRness in terms of metadata, (ii) FAIRness in terms of research data and (iii) FAIR awareness. Concerning the last one, all institutes are conversant with the emerging trend of FAIR (meta)data. Most of them partake in European working groups, where FAIR is common practice. Besides, a lot of researchers from the institutes organise or participate in training – which are occasionally provided by the larger institutions – on this matter. However, despite the movement being underway, there is still a lot of work that needs to be done in order to translate the FAIR principles into day-to-day practices.



"Additional awareness-raising and training of researchers will certainly be necessary in this area, but we are waiting for the infrastructure"

PID is the cornerstone of the FAIR principles, enabling findability of digital resources. Most of the institutes have understood this concept. PIDs are in most cases automatically assigned (depending on the infrastructure) and well integrated in the Flemish research landscape. If not, they are added when appropriate. There is obviously some isolated cases, where data cannot be associated to a PID (e.g. continuous data).

"PID are assigned to all data and metadata by the [...] portal allowing them to be cited"

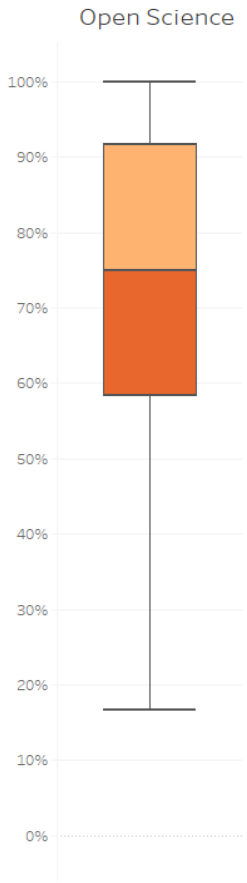
FAIR is also about accessibility of metadata and research data. The trend in Flanders is that metadata and research data can be accessed when requested. In some cases they are not available at all and in some others there are tools and processes in place to grant access (e.g. user interface to facilitate data access) to metadata and research data. The access scheme is quite differing and often done at the level of the repository.

Two equally important aspects of the FAIR principles are interoperability and reusability. Institutes try to make their data as interoperable and reusable as possible with the means they have at their disposal. Global networks, international standards and catalogues are required to realise this.

In conclusion, 50% of the institutes score more than ~60%. But an important aspect is not reflected in the box plot. While FAIR metadata is a common thing in the Flemish landscape, interviews have revealed that FAIR data is still lagging behind.

Open Science

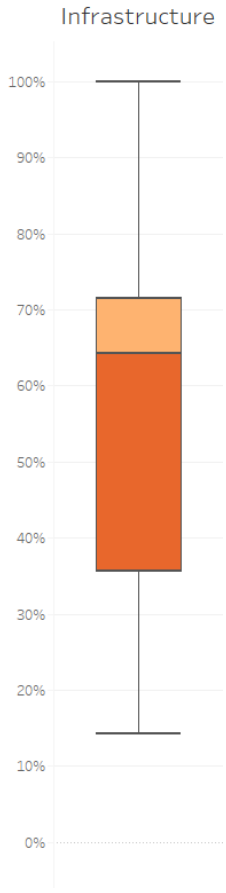
Open Science, and by extension Open Data, is the cornerstone of European policy towards reuse of research data. Open Science is the practice of science is such a way that offers the possibility to collaborate and contribute to other's research, to ensure free availability of (meta)data and that enables the reuse and reproduction of data. In the interviews, two essential elements were measured. First, the familiarity with EOSC and FOSB's goals which are all about Open Science and secondly, Open Data policies and strategies. On the first aspect, most of the institutes are quite familiar with the objectives of both EOSC and the FOSB. They live the values and try to translate them in their day-to-day operations. Nevertheless, it was reported that despite the objectives and values being clear, the implementation isn't.



Institutes often reported that they were part of European working groups having a focus on Open Data or Open Science. Europe and by extension, European working groups, are definitely identified as the ones setting the tone in terms of Open Data. To the questions on Open Data, institutes mentioned that when possible they were using the CC-BY licence. Besides, they make their data open by default, when there is no sensitivity nor restriction.

As depicted in the visual, most of the institutes scored well. 75% of them scored 60% or higher. However, the great awareness of Open Science and Open Data doesn't signify great Open Science and Open Data policies. Our investigation revealed one surprising characteristic: most institutes have Open Science and Open Data on their agenda but policies and practices are not enforced. Institutes are only taking their first steps towards Open Science and Open Data.

Infrastructure



The word infrastructure is quite broad, in this context it was narrowed down to a few concepts such as hardware, standards and accreditations.

It seems that *technology* has a dire need of improvement, on the same level than *data management*. As depicted in the picture 50% of the organisation scored less than 55%. However, this trend can be attributed to the fact that most smaller institutes don't host their own infrastructure and hardware as they don't have the resources to do so. These institutes solve this problem by leveraging and storing their data at the bigger players in the Flemish landscape.

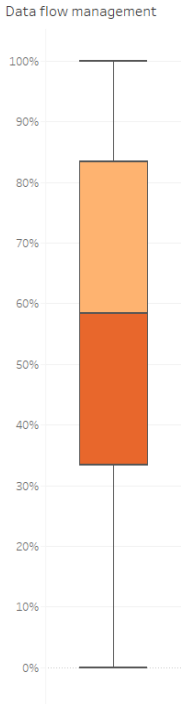
"Local storage will usually be university servers"

For the vast majority, institutes store their data on-premises. It can be either in their facilities or sometimes they are using the facilities of a bigger institute (e.g. a university). In conclusion, many institutes are outsourcing their storage and use third-party services in that regard. Few institutes reported storing data on cloud storage (e.g. AWS, Azure etc.).. It is important to clarify that when talking about storage, it is about research data not being published yet, but manipulated for the execution of the research itself. With this in mind, some institutes mentioned using the Vlaamse Supercomputer Centrum (VSC) for computation or the intention to use it.

The box plot describes a quite peculiar situation, which is, at any point, similar to the situation reported for the data management category. Smaller institutions don't have the capacity nor the resources to store their data internally. Fortunately, they have the possibility to use the facilities from larger institutes and universities. The box plots show that 50% of the institutes surveyed scored less than ~65%. However, it should be taken into account that in the current situation ad hoc solutions are found but these are in many cases not ideal or suited fully to their needs. A major requirement still exists to be able to use e.g. standardised, qualitative, secured data storage capacity.

The key take away from that category analysis is that none of the institutions have a similar manner when it comes to sharing data, and storing and archiving is done with the means available, while frequently being in need of additional or improved solutions.

Data flow management



Data flow management relates mainly to workflows, standards and data models on the use and storage as well as archiving of (meta)data during their lifecycle. To put it simply, this is about how well the data lifecycle is managed and to what extent it is documented.

It became clear that most institutes do not have a full workflow in place to follow the data from its creation to its publication. They all indicate this is very difficult to follow, often the data remains with the researcher until publication. Despite this, many institutes reportedly document their data life cycle. The latter is often formalised in documents, policies, etc.

Along the same lines, few institutes reported having documented their data models. Concerning the ones that do, mention of Darwin Core, CERIF, etc. were made. Besides, researchers are encouraged to use standards, but standards are used at the researcher's discretion.

"[...] unstructured / research data is not documented."

Institutes have reported using a wide range of repositories when it comes to storing (meta)data. Often the repositories used are domain-specific and not necessarily regional / national ones. In addition to that, the repositories used have standards that the researchers and institutions try to comply with. It was often reported that repositories were compliant with CERIF.

"Repository is CERIF compliant to a large extent"

In respect of the last step of the data lifecycle (i.e. archiving), all institutes agreed on the need to keep all research data, and metadata associated with it, as long as possible. Nonetheless, not all of them are currently doing so. They have expressed the will to keep their data as long as requested, but only a few institutes have an archiving solution in place. What the institutes are lacking are a clear and standardised policy on archiving as well as the means to do so.

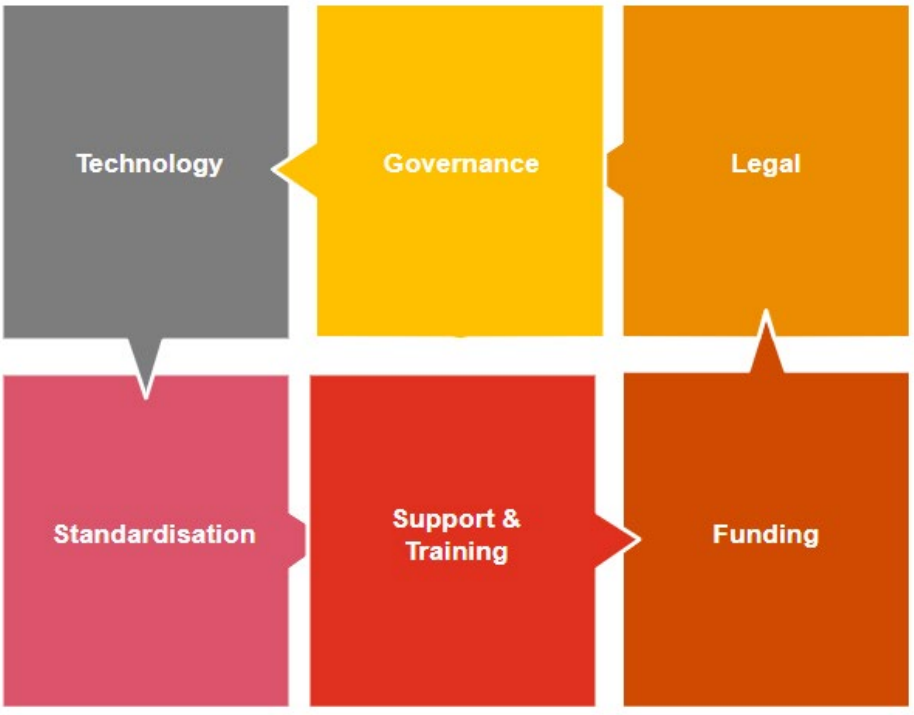
"[...] In the long term, the plan is to have a university archive. Currently, archiving is done ad hoc and is different for each institution"

"The discussion is still ongoing - what do we want to archive (e.g. versioning)? There is awareness, but no concrete policy"

Here the box plot has the greatest span of all. It means that there is no clear trend and that all institutes scored differently on the topics of the category. To conclude, 50% of the institutes survey scored less than 55% which underline the improvement needed in terms of documentation as well as archiving.

Requirements - Common themes

Based on the feedback that we have received during the interviews, we were able to identify several requirements that recurred throughout the research landscape. These requirements could be summarized into six main categories, namely:



In the following chapters, we will discuss each category and discuss some of the main requirements indicated by the organisations. On top of that, we discuss identified challenges and success factors to meet the requirements of the needs of the researchers.

Technology

Automate where possible - Enter once, reuse often

In order to meet the needs of this requirement, **FRIS should be further optimized as a (meta)data discovery hub.** On the one hand, data should automatically be fed into FRIS and on its turn, fed into the EOSC platform. On the other hand, FRIS should align with international or other existing metadata schemas, such as datacite, DDI, etc. Further optimizations with regards to findability and accessibility of datasets to ensure that FRIS becomes the best data discovery portal are also critical success factors. The possibility to automatically feed data from FRIS to Google Search and OpenAire also allows for the researcher to limit the amount of (meta)data entry points.

Moreover, **clear policies on data flows and quality control** should also be defined. It is not always clear which (meta)data is obligatory and who is responsible for the submission of this data. It is therefore important that clear policies are defined in order to identify responsibilities and ownership with regards to providing, managing and performing quality checks on data.

Increasing in importance is the concept of "**enter once, reuse often".** Metadata flows should follow the example of ORCID, generalizing as much as possible and being interoperable across different platforms. This could be done by developing a platform that enables a researcher to fill out standard forms, such as DMPs or ethics information, based on his ORCID and previously filled in information. Some of this data can also be harvested from other systems or authorities, such as EWI, VLAIO, FWO,... Funders should also be encouraged to publish all relevant metadata of each grant they fund. A possibility here is to have the OAI-PMH as an endpoint. By doing so, institutional systems can automatically harvest this information.

It should be kept in mind that **research is learning and adapting,** hence an **automated generation process of,** for example DMPs, should also be **robust to change**. In addition to that, DMPs can have different formats and information when dealing with different research themes. Even making a consistent DMP for the same research domain across different institutions is not as obvious. There should also be an emphasis on the uniqueness of a DMP, over-standardising the generation procedure can lead to pernicious effects.

Storage and processing capacity

As time progresses, the need to store, process and archive data will continue to increase exponentially. The amount of data that is generated is expected to increase tremendously across the vast majority of research themes. A logical consequence of this is that the demand in archiving capacity is increasing as well, as the hot storage of data is significantly more costly than cold storage. When combining these two trends with emerging technologies, such as machine learning and the Internet of Things (IoT), that significantly increases the demand in computing and data processing capacity, it is clear that institutions raise their concerns about their future capacity to meet these demands.

Research data storage capacity (and management of it)

In order to be future-proof in regards to data storage, emphasis should be put on **how to cope with different data sizes and types.** As the complexity and variety of research domains are increasing, the output data of this research is also diversifying. This implies that each research theme has its own requirement with regards to data storage and capacity. Some research domains require low latency, high broadband storage solutions, whilst others are struggling to cope with the storage of research data when a third party, e.g. an NGO, is involved in the research domain. Consequently, the storage solution and capacity has to be very tailored to the needs of that specific domain.

However, before dealing with the issue of coping with different research data sizes and types, a **clear definition on what research data is**, is something that is currently lacking. A clear identification of what research data is, when this should be stored (e.g. at the start, during or after the publication) and who owns this decision making is something research institutions are struggling with. The need exists to define a research-specific framework.

Besides the needs on the IT infrastructure itself, **human supervision should be present as well.** Digital fitness is not as adequate as needed for current IT solutions. Hence, human guidance should always be available for those who need it. If the designed interfaces and platforms are complicated to comprehend and use, researchers will not easily transition towards these new IT solutions. **Hence user friendliness and ease of access should be stressed.**

Supplementary to that, having a centralized storage solution is very promising, but **if the costs are unbearable** for the research institutions, not a lot will opt into this solution.

Archiving capacity

Next to active storage capacity, archiving storage is a need that frequently came up across multiple research themes. However, it is not always clear **what long term storage implies, which guidelines should be followed**

(e.g. what's the minimum duration of archiving data?) and **which solutions are out there in the market.** As the amount of data to be stored is vastly increasing, the need grows to have a clear definition on how to store data, for how long and to identify which solutions are the most cost-effective for the institution's specific case. On top of that, in some specific cases, data redundancy is an extra feature that needs to be met. This means that data should be stored in several locations in order to be accessible if one of the data storage centers fails. Institutions indicated that they are in need to have guidance for these topics.

The importance of storing and archiving data should not be neglected. Nonetheless, for the sake of cost savings and decrease of administrative burden, it should be noted that data should not be stored for the sake of being stored. Clear standards on what data is relevant and which data could be neglected should be in place.

Processing capacity (cooperation with VSC)

Many institutions already indicated during the interview that they were currently running or planning to run pilot projects with the 'Vlaamse Supercomputer Centrum' (VSC). With the growing need to be able to process vast amounts of data for specific research domains who need High Performance Computing (HPC), this demand for a centralized processing unit is expected to increase towards the future.

An important point to keep in mind with a centralized processing solution, is that **the data should preferably be stored where the processing happens.** This is because on the one hand, transferring data is always a costly process. Hence, if the storage is located near the processing location, this is already a cost-reducing effort. However, it also needs to be clear that VSC in itself is not a storage solution.

Secondly, it also reduces the processing time, increasing the researcher's time effectiveness. Besides the technical requirement that data should be stored where the processing happens, the emphasis here should again be on the user friendliness and user accessibility. The processing of data should be set up in a way that the researcher does not encounter any (technical) issues when trying to query the data or do computations on it.

Others

Besides the need to reduce administrative burden and the need of a centralized storing, processing and archiving solution, some other attention points regarding technology and architecture were raised several times during our research. In what follows, we will elaborate on these points.

Limited need on platforms as many rely on what universities provide

Some institutions indicated that they have limited needs on a centralized platform, as they are currently relying on what their corresponding universities provide. Ergo, they do not see an added value in the development of a complex centralized solution and see more opportunities in expanding the cooperation between smaller institutions and universities. However, **doing this requires the alignment between different data models across different institutions.** The multitude and variety of research institutions and themes make it a very hard exercise to align all data models across the different stakeholders. It is not clear that a standardised model is possible for an individual research theme, let alone the aligning them over several research institutions. An extra layer of complexity here is **how the systems of institutions that do not integrate with universities, will integrate with these systems in the future or with FRIS.**

Because of the fact that not all stakeholders have access to what universities provide, there should be an assessment of **which organisation should be eligible to access these services** if they would want to. On top of this, **clear ownership should be defined.** On the one hand, this federated structure implies that the different initiatives between the universities should be interoperable. On the other hand, the interface and access levels need to be clearly defined, so that the needs and concerns around data privacy and protections are met.

Mapping the platforms and functionalities that are currently available and conducting a gap-analysis on the functionalities or platforms that are currently missing, offers a clear overview of the different potential data streams within the research landscape, allowing efficient data flows, interoperability and cross-referencing. This however implies that **the needs of researchers need to be mapped** out by defining user stories in order to identify the main struggles and needs. Next to that, if there would be such a federated structure, newly added data and information should automatically be fed into FRIS and wherever possible, connections with existing platforms should be made available. This limits the amount of effort that smaller organisations need to do to be compliant with the FRIS regulations, as the universities are already connected to the FRIS platform.

In line with keeping this simple in terms of terminology and technical savviness, having a limited amount of platforms with a clear overview of which university provides which services, scientists and researchers will be more aware of where they need to go with their problems with regards to data management. Having standardised interfaces across university platforms can facilitate the adoption rate for new users. Next to that, a majority of the smaller organisations do not need to invest in an IT department, as they can rely on the infrastructure that is provided by the universities, potentially opening up additional training budget. Researchers will also be more familiar with who to contact in case they are experiencing any difficulties.

Cope with the double affiliation issue in FRIS

An additional issue that was often mentioned, was the issue of double affiliation on the FRIS platform. This issue is mainly due to the fact that in some scenarios, different stakeholders use different unique identifiers, resulting in duplicate records for the same researcher. Solving this issue is in line with the aforementioned ambition of FRIS to be the go-to search engine for research (meta)data.

Governance

Decrease administrative burden

In the current set-up, many researchers experience the abundance of guidelines, rules and regulations as a burden that limits their ability to focus on their day to day work. This causes frustration and the idea was mentioned multiple times that there should be a way to reuse generic information that has been entered for various other purposes. One example of this is the generation of DMPs. Due to the nature of this process, which implies a re-entry of a multitude of identical data, many researchers experience this task as an administrative burden. A solution to this problem would be that this process is partly automated by a system that harvests and re-uses information that was previously entered.

There are several difficulties to overcome when thinking about potential solutions to reuse generic information for multiple purposes. There should be a complete **alignment and synchronisation of administrative databases and practices.** When updating the source data, all consecutive databases should be updated simultaneously. This requires a synchronization across a multitude of different organisations and database systems

A second difficulty is the question of **which data serves as the single source of truth.** The main issues here are twofold, on the one hand, there is uncertainty on where the source data will find its origin. On the other hand, there is the difficulty of defining who the owner is of this ground truth and who will be able to make changes to this.

In addition, there should be a clear **distinction between static and dynamic data.** Static data is defined as data that will not change over the course of a research. Dynamic data will change over the course of a research. Making changes in the latter and optimizing this administration workflow would be an improvement, however,

having a clear-cut definition of the static and dynamic data types for each organisation and research theme is a complicated matter.

Having a system that gathers data over various platforms and institutions, requires **alignment on different political levels.** Due to the nature of our political landscape, standardisation and easy exchange of data is growing in importance. One solution for this matter could be an authoritative layer that acts as a template and that is linked to the researchers' profiles, which in its turn is linked to the involved institutes, funders and governments.

A last identified difficulty is that **everyone's situation should be taken into account.** Some research domains are more inclined to cooperate with each other than others. This multidisciplinary research and connection is even more complex when having research that involves non-research environments (e.g. industry, healthcare, education,...) It speaks for itself that having a streamlined process for such complex working streams is not an obvious exercise.

Alignment of the different initiatives on different levels: Flanders vs Belgium vs EOSC

Everybody seems to agree on the complexity of the research landscape as it is. Initiatives in the domain have the purpose to push things forwards and federate all research disciplines. However, the current situation doesn't actually help. Many initiatives are launched at regional, national, European and sometimes international levels. This creates challenges, especially that some initiatives are not necessarily aligned. There should be an increased alignment between the different policy levels and notably a hierarchy defined.

"Make sure focus is on science and society benefits"

Find balance between administration and value for researchers

Having systems and processes in place that ensures proper data management and governance are undoubtedly of great importance, however, one must be careful not to overwhelm the researchers with an abundance of administrative processes. The aim should be to find the sweet spot between value that the processes and policies bring and the administrative workload that this entails.

However, it should be noted that some **researchers are not aware of the added value that data management provides.** Hence, there should be some investments towards awareness campaigns that address this added value. One reason for this unawareness could be the lack of a correct incentive mechanism for good data management. Extra attention should be given to the external funders. Frequently, **the data management plans and everything surrounding the concept of data management, is governed by an external funder**. This external funder might have the potential to overrule certain admin practices, which makes it difficult to implement clear-cut guidelines.

Besides that, there also exists a **discrepancy between the value that data brings and cost of ownership of data.** In most cases, he who gains the most value out of data, i.e. the researcher, often does not have to bear the costs of ownership and storage of this data, i.e. the research institution.

Researchers should be provided with **the necessary, user-friendly tools** that aid them in administrative processes such as the generation of the DMPs, in combination with some assistance by the technical staff, who are generally more aware of certain information than the researcher himself. The added value that a DMP provides is also not always clear to a researcher. Hence, training and workshops to stress the importance and benefits of well-defined DMPs is something that should be investigated.

Standardisation

"standardisation is growing in importance, but we should be cautious to not lose the uniqueness of a research theme."

Avoid duplication of data entry in different repositories / applications

To solve this problem, there is a need to have a centralized layer that **harvests data across multiple repositories and applications.** However, making this solution possible is a tremendously convoluted exercise. First and foremost, **integration between different platforms and data models is needed.** The extreme variety of data in the research ecosystem should be defined and the common ground has to be identified. Next to that, **a centralized layer imposes the difficulty of the identification of the ground truth.** If a researcher would enter his data in two different systems in two different ways, it will be hard to define where he made the mistake. Before thinking about this unified layer, the decision needs to be made which data repository will serve as a reliable source.

Besides that, one should keep in mind that the **metadata standards are constantly evolving.** Hence, when defining an overarching solution that incorporates different metadata repositories with sometimes outdated data schemes will cause some integration difficulties.

To meet the researcher's needs, it is necessary that **a well-designed data architecture is defined.** Having a multitude of systems and processes that communicate with each other, requires a well-thought architecture that incorporates all the specific needs and requirements of each stakeholder. One of these requirements is **the usage of PIDs for each dataset that has been made available.** On top of these datasets, there should be an architecture that allows for **efficient exchange of metadata.** It is important to keep in mind that **the wheel should not be re-invented.** The solutions currently in place are already providing data aggregation tools (e.g. ORCID, FRIS, Google Scholar,...). **A link with all repositories available would be a more (cost)-efficient solution.** Over and above, compliance with OpenAire standards is also beneficial towards a more harmonized European metadata repository. Needless to say, is that as well as for other requirements, the need for **user friendliness** should not be neglected.

In order to successfully meet the expectations of this requirement, it should be noted that ease of access and ease of use is something that has been stressed upon a lot during the course of our research. Researchers and staff are not eagerly awaiting a new, over-engineered tool. Contrarily, their wish is a **straightforward solution having a streamlined (meta)data entry form**, with real time information on what (meta)data has already been provided so that this (meta)data should only be entered once.

In addition, to comply at any given point in time with FRIS/EOSC guidelines, **a dedicated staff should be appointed in order to have an efficient data handling system in the Flanders ecosystem.** The role of this dedicated staff would be to closely monitor changes in the (meta)data landscape. This allows for quick adaptations for this centralized solution, in order to be compliant to the rules and regulations at any point in time.

Thematic approaches for the alignment of systems and standards

By now it should be clear that a 'one solution fits all' is not possible for the Flemish research world. Regardless how complicated alignment and standardisation turns out to be, there exists **a tremendous amount of knowledgeable experts that know the ins and outs of their research domain.** This expertise should be exploited in the definition of (meta)data standards and other processes. In most cases, researchers that are affiliated to thematic data centers, are also very familiar with **community standards.** By incorporating these community standards into the best practices for the alignment of systems and standards.

Support in manpower (or technology) to transform data into standardised data

Not all researchers and staff are as digitally literate as others. Providing them with the **right tools and guidance to produce standardised data** will allow them to adapt to the FAIR mindset. Next to digital fitness, general guidelines on when data should be published impose themselves. It is not always clear to researchers when their research data needs to be published. Would this be during the research or after?

Standardised technologies

For the sake of faster adoption of new technologies and streamlined architecture, it is of utmost importance that the **technologies to connect with these platforms are standardised.** On the one hand, the platform to share data should all use **standardised API protocols**. Being forced to additional development every time there's a new platform with it's own API to communicate with, leads to increased inefficiency and frustrations. Another example of such a technology, is the **authentication protocol**. Many different types of authentication protocols were mentioned over the course of our interviews. Some institutions use the institution's log-in credentials to perform authentications, others use Microsoft ADFS. It speaks for itself that when building a centralized, standardised ecosystem, these authentication and API protocols should all be aligned with each other. However, doing this will require efforts from a multitude of organisations, as they will need to adapt their current systems or infrastructure.

During the course of our interviews, following standards and technologies, amongst others, were frequently mentioned:

- Metadata: CERIF, Darwin Core, DataCite Metadata Schema, DDI, Dublin Core, LIDO, Spectrum
- API: REST, SOAP
- Authentication: Azure AD, Edugain, IP address ranges, LDAP, Single Sign On

Standardised FRIS metadata model and mapping guidelines for datasets is needed

Because FRIS has the ambition to be the one stop shop for Flemish research information, **it should increase its efforts to standardise its metadata models**. When the feature is implemented that it will also provide the searcher with dataset information, research institutions will need to have the information available on **how to map their datasets in an efficient manner**. Many of the institutions are currently aware that this feature will be available on FRIS platform, nonetheless, guidance and policies are lacking.

Legal

Concerns about how to protect IP / privacy of shared solutions.

In an increasingly globalized and multidisciplinary world, being able to cooperate and to share data is vastly growing in importance. Nonetheless, with increased centralization, one should be cautious about the **intellectual property and privacy of these storage solutions.** On top of that, the **GDPR guidelines increase the complexity of these requirements**, as they add a new layer of privacy regulations.

If there would be an initiative to store data on a centralized repository, extra attention needs to be put on the **user access management and cybersecurity of this solution.** It speaks for itself that researchers who need access to this data, should be able to do this in a few simple steps. But those who wish access to this data, but do not have the proper rights, should not be given any window to abuse this centralized system. The ability to define the level of access for other researchers or organisations is of utmost importance.

Need of a standard on sharing data with other researchers while making certain being acknowledged

In an increasingly data-driven world, shareability is one essential pillar. It appeared during the course of the interviews that data gets often shared, but data sharing is as disparate and complex as it can be. Some organisations reported that *friendliness* of data sharing was to be improved. Furthermore, **there are currently too numerous ways of getting your data shared to another party, if any.** This call for **standardised practices to share data and the means allowing this**. However, there is a downside to data sharing that needs to be addressed at the same time. Research is all about citation, acknowledgement and recognition. What helps researchers and organisations thrive, is the relevancy and soundness of their work. Nevertheless, data that is shared is not often attributed, publications are. If data is to be shared and re-used, it means that a researcher(s) or an organisation might have contributed to a project without being recognised. The ability to acknowledge data is a major issue for the researchers.

Support and training

Support on connecting to FRIS

While many organisations reported being familiar with FRIS and its mission, some reported that despite being familiar with them **the infrastructure and the way how FRIS operates remains blurred**. In a landscape where FRIS is the lawful gateway to provide FAIR-compliant metadata to EOSC, it is of the utmost necessity to ensure transparency as much as clarity allowing all research institutes to comply and provide their metadata to FRIS.

Guidance and training on coping with data

A data management plan (DMP) is a convenient tool that is growing in importance in the research landscape. It requires the researcher to put in writing the data that is expected to be acquired or generated in a project and how it will be dealt with. However as DMP started as a top-down invite, **it currently fails to address the specific needs of the different disciplines and domains**. Generating data in human sciences is another story than generating data in astronomy. As the story is different, so should be the DMP.

In addition to that, a wide range of aspects relating to the data lifecycle are touched upon in the DMP (e.g. metadata, data sharing, data licencing, data storage, data archiving, etc.). **All aspects that necessitate a specific expertise and an average researcher might not have it**. Having DMP tools is a major breakthrough, but if the researcher supposed to fill it in does not do it properly, the usefulness and purpose of the DMP is significantly reduced. Therefore, there is a need to coach researchers when setting up DMPs.

Marketing plan to promote the work and FAIRness of the work of scientists enhancing reuse by other domains

Research data is not truly reusable unless it is open, i.e. available under an open licence and at marginal costs (in most cases at zero cost), and openness comes often hand in hand with the implementation of the FAIR principles. With data being said, FAIR and Open Data are the foundations of reusability by other domains. **Combining several academic disciplines to foster scientific research can only be done thanks to Open and FAIR data. This should be advertised as such**. There is an opportunity cost due to the unrealised benefit of interdisciplinarity.

Stimulation of the Open Data and FAIR culture: getting tools and methods

FAIR and Open Data is all about sharing data. The same should be held true for tools and methods. **Research depends a lot on reproducibility and transparency about tools, methods and data used**. This could lead to more reliability and quality in research but also accelerate the discovery process.

Incentivise organisations/researchers doing a lot of effort

As mentioned above, researchers tend to spend a lot of time making data FAIR. The reason that some researchers are more reluctant to do this extra effort, is sometimes because **they have the feeling that this extra effort does not serve any benefits.** This is why several institutions came up with the idea of **incentivising researchers or organisations** that are consistently opening up their datasets in an open and standardised way. Several different ideas came up during the several discussions we had. On an organisation level, **a quality label** (cfr. ISO-standards) would be something that institutions would thrive for.

For researchers on an individual level, one form of incentivisation would be a plain monetary remuneration. Another idea was making published, FAIR research datasets a valid credential on a researcher's CV. This can be done by a research data paper, which entails a paper that describes the dataset in detail, together with the steps taken to prepare the dataset in the manner that it is presented. A way to incorporate the peer-reviewing efforts of colleagues is also something that should be thought of.

"The entire Open Science movement is based on intrinsic motivation and an ethical responsibility that researchers have. Introducing rewards can help, but be careful not to overshoot. The internal 'competition' amongst researchers is a powerful driver in itself."

On a side note, one should be cautious that **the system to provide rewards is set up in a democratic way.** As was mentioned to us, the feeling exists that larger institutions often have more resources available to do that extra effort towards making data FAIR

Funding

Financing model needs to represent all types of stakeholders

It was mentioned that in the current funding model, the distribution and mechanism of the funding model is perceived as being quite generic. When reviewing the funding mechanism, elements such as long term research projects or very specific or state of the art research can be taken into account, as well as safeguarding an appropriate balance between larger and smaller institutions.

Pilot projects to make large data sets available to other institutions

Many institutions indicated that they experience troubles with opening up datasets with a notable size. It would bring great added value towards the Open Science community when the barriers to publish the bigger datasets sizes are vanishing. Hence, foreseeing POCs or a sandbox environment where institutions can try to publish data and see what the total cost would be to do this should be something to be investigated.

Potential Solutions

Central Research Data Repository

What

A solution that can benefit many organisations across the research ecosystem is a standardised central research data repository. Much in the lines of what FRIS currently offers on metadata, there is a clear outcry for the need of research data storage and archiving. Ofcourse, as previously mentioned, the maturity level of a specific institute will dictate what service is required in this particular case.

We notice the use of some standardised thematic repositories across organisations (e.g. GBIF) whenever these specific repositories are available. However, for many other institutes research data is kept in cloud services such as SharePoint or Google drives, as well as less accessible sources such as an external hard-drive or their local computers. In these cases we see a clear added value for the implementation of a federated approach that can link existing repositories with the creation of a central repository which will help these institutes manage their research data in a more efficient and standardised manner. The end goal should be to have data flowing between repositories and possibly into archiving solutions, and have as much automation as possible when it comes to FRIS and the future EOSC architecture.

How

Before we look into the requirements and extensions of what this central repository should entail, we have to focus on what is **already available**. Whenever institutes use existing standardised repositories, the added value will more likely come from API's connecting different platforms to each other. As a result, a centrally hosted repository which will be obligatory to use is not a desired solution. Therefore, a **federated** approach is far more valuable, reusing what exists, avoiding double investments in new infrastructure and using the expertise that already exists in the research ecosystem. Ideally it can be reusing existing infrastructure and providing the opportunity to smaller organisations to make use of it. As a result, a data repository can be hosted by one or a few organisations and can be reused by others depending on their domain. These repositories can be seen as 'central' service offerings, hosted by any stakeholder active in the network. No conclusions need to be made on who should host or organise this kind of solution.

Open questions

In terms of basic architecture for the central research data repository some very important questions need to be addressed such as:

- Do we need a centralized repository, or do we need a platform that connects datasets across the research ecosystem?
- Will the repository only accommodate the storage of research data or also/only have the ability to archive parts of data for long-term storage?
- Will the repository contain finalized cleaned-up data that is linked to publications (and FRIS), or will it be used as a live database? And if it's used as a live database, how will we protect it from becoming a data dump?
- Will it only store data that can be viewed by all institutes (Open Data) or also include sensitive data that needs to be restricted (artistic data)?
- Will it be used as a general repository, or is there room for discipline-specific features/sub-repositories?
- How will data be linked with FRIS/EOSC? Via DOI's or PID's?
- How will data link to other repositories across different institutes?
- How can data be exported?
- How will the cost related to storage be handled?
- How will we handle the transition between repositories? (Change management)
- What are the definitions of data stored within the repository? Is it research data, project management data, accompanying metadata, etc.
- Who will be in charge of further change requests? (Governance model)

Once these plethora of questions are answered we can start looking into the possibilities to extend its basic architecture such as:

- Linking research data to its corresponding publication in FRIS
- Adding modular infrastructure components that can interoperate through standardised protocols and link to processing power
- Connections to other European initiatives

Intermediate Layer

What

There is a huge overlap of information available between different platforms such as FRIS, PURE, ORCID, DMPOnline, OpenAire, etc. This also means that from a researchers' perspective there is a lot of similar administrative work being made across these platforms. A possible solution would be to minimize this administrative work and automate as much as possible, if possible via an intermediate layer or platform linking these instances, and their corresponding data, together. The key functionality of this intermediate layer is to know which type of information is available in which existing platform or database, and to be able to reuse the existing information for other purposes.

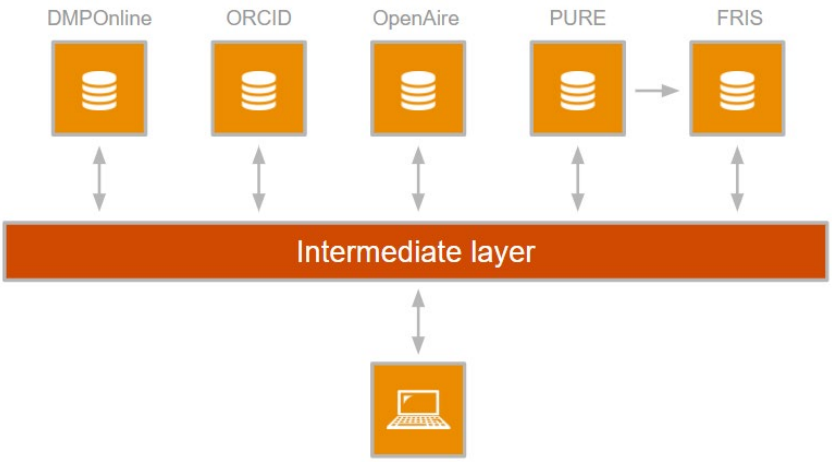
The possibility to automate most of this administrative work would be beneficial for many initiatives that are ongoing related to FAIR and Open Data such as the creation of DMP's and reusability/traceability of generic information for GDPR reasons, etc.

How

An intermediate platform would create an overview of what is available and which administrative work is absolutely necessary to complete from a FAIR perspective. This intermediate layer can be accessible via a front-end interface that allows researchers from different institutes to connect and easily track where information is already available. It can also function as a guide for researchers to track what is mandatory to complete in order to comply with the different FAIR principles and Open Data initiatives. It can give them an overview of which documentation is required or which platforms can be used based on their institutional login, like linking straight to DMPOnline, ORCID or their PURE platform. The platform could also help them reuse previously captured information for DMP's when working on e.g. co-financing projects or function as a link to the ethical committee, or even reuse information on their affiliation which can be automatically populated into FRIS or their CV. While some links between those databases might exist, it primarily makes sure a complete overview exists and reuse can happen as efficient as possible. By providing such an intermediate layer on a Flemish level, researchers stay in control of the information they (need to) provide to a variety of instances, both internally within their organisation as externally. In order to facilitate the widespread use and added value of such a solution, it should not be over engineered and can be develop in a generic way, making sure it can be applied by any type of research in any domain, for researchers being active in both smaller as well as larger institutions across the ecosystem.

While taking a deeper dive into the Flemish research ecosystem, the following applications and initiatives will benefit from an intermediate layer:

- DMP creation (DMPOnline)
- Metadata (PURE, IMIS, FRIS, ORCID, OpenAire)
- GDPR privacy register
- Ethical committee information
- Reporting to ESFRI's
- Data citations and publications
- Availability of research data



In the long run the integration of this available data will lead to plenty of advantages such as:

- Less repetitive work
- More time saved (and more time to spend on research)
- Better Open Data culture
- More standardisation
- Better traceability

Open questions

However, before we can set up this intermediate layer or platform, some burning questions will need to be addressed:

- Who will be the owner of the intermediate layer (both semantic and technically speaking)? (albeit an ESB or a platform)
- What are the priority systems to be interconnected?
- Is the scope of this layer only Flemish connections, or also broader EU connections?
- How can this be developed best, by institute or on national/Flemish level ?
- How do we handle the different maturity levels across institutes?
- How do we handle changes in the source systems?
- How will we handle the variety in research institutions and thematic approaches?

Standardisation

What

In order to create any centralized or federated solution as described in the previous topics it's necessary to bring more standardisation across the research ecosystem. When talking about architectural standardisation we've identified plenty of technologies that might need more standardisation in the future:

- Metadata standards
    - Reuse domain specific standards that are already defined
    - Remain flexible knowing they will evolve over time
    - Promote FRIS standardisation
- API standards to connect applications instead of developing more
- Authentication standards
    - Flemish / European authentication
    - Standardised SSO technologies
- Methodology standards (e.g. OSLO)
- PID standards

Within certain communities multiple standards on metadata and data have already been successfully implemented. Communities such as ESFRI's, EMBL CERN and EGI have specific standards covering API's, authentication, data models and vocabularies. As clearly stated in the previous chapter, wherever possible, we should avoid reinventing the wheel but instead work with these communities to share and further improve where possible.

How

The ownership of these standards should be kept close to the thematic networks in which they are developed, such as repositories (GBIF, OBIS, Seadatanet, Zenodo,...), ESFRI's (DiSSCo, ELIXIR, LifeWatch,...) and institutes. On top of that it is required that the creator of these standards needs certification or accreditation in order to do so, and will require a steady financing model to maintain these standards.

Non-architectural solutions

Some of the institutional requirements can be covered by solutions that do not directly require an architectural setup. Such solutions can still improve the efficiency and willingness of sharing data, reduce administrative work and can bring more alignment across the research ecosystem.

Training platform

"Let's avoid making mistakes everyone is making"

A training platform is one of these commonly requested solutions that would create awareness and uniformity. It's requested to educate the researchers, professors and even to train the trainers. Another possibility might be to introduce it to master or PhD students and make it part of their curriculum, or even assign accompanying training credit. This platform can promote existing solutions and clearly outline what the minimum requirements are in order to adhere to good practices around data management. It can be a place to share specialized thematic knowledge and best practices. On the other hand there are some outstanding uncertainties that will require attention to make the implementation successful, such as who will do the follow-up, take ownership and invest further energy and resources to maintain the platform? This need ties in with the idea of competence centers that are currently being set up at an EOSC level which will focus on digital upskilling and topics such as:

- Provision of training
- Data services or research software engineering
- Guidance resources and advisory services (including policy advice and implementation)
- Development of communities (e.g. communities of trainers)
- Catalogues of resources, services or policies
- Creation or dissemination of standards
- Evaluation and assessment services (for trainers and trainees, training materials, etc.)
- Hub for collaboration between stakeholders

Provision of data stewards

The data stewards have also been recently introduced within some institutes and will fulfill the role to bring alignment and better overall data management where applicable. These key people will be essential in the first phase to mobilise institutes and researchers and will bring clarity within specific scientific domains. We've noticed that next to more generic data stewards, mostly smaller or more specialised institutes are in need of data stewards which can help them with the specific requirements linked to their discipline, and therefore data stewards should not only be tied to a specific institute. There also seems to be concerns around the funding period which may need to be extended due to the nature of this evolving landscape and the need to keep up with its latest developments.

Incentivise Open Data

On top of the awareness created by EWI, FWO and other communities it seems beneficial to somehow incentivise the implementation of correct data management, Open Data and maximally adhering to the FAIR principles. Although we should be wary of not creating a competitive landscape by inctivising, there are some options to create more visibility and added value such as promoting data citations (e.g. data publication journal) or implementing a smart reward system that can tie back to the evaluation of the researchers. These stimuli will produce better scientific outputs and might be linked to a researchers scientific CV which will benefit their career in the long run. At EOSC level there are efforts being made to create a Knowledge Hub that can be a central platform for sharing success stories.

Network across Flanders

"Unified data management simplifies processes for all stakeholders"

The final goal is to bring awareness, best practices and clearly indicate the added value of good data management - in a sense creating a network across flanders. This goal can be reached by a combination of architectural and non-architectural changes and will require further investment of governing bodies, institutes and

the researchers themselves. On a European level this would mean a better alignment of ESFRI's, so that they can create a unified network across Flanders.

Conclusions & Recommendations

In this chapter some high level recommendations and conclusions are described. These apply to a variety of requirements and are linked with potential solutions described.

One size does not fit all

Based on the insights gathered during the interviews and as seen in the maturity analysis, we can conclude that the research landscape in Flanders is very diverse. Both in terms of actual capacities to cope and manage all data related topics as well as developing or managing underlying infrastructural assets or technological solutions. While smaller research organisations frequently rely on what third parties provide (often universities), larger institutions as well as specific thematically focused institutions use established and proven infrastructure to cope with their data needs.

A clear distinction needs to be made between how they cope with data compared to their operational research work. While these organisations may not manage their own data repositories or do not have specific data oriented profiles within their organisation, they are often very efficiently organised related to their research. Not all organisations are in need of managing all data related aspects on their own, as some organisations comply with various existing standards, guidelines or automations based on either the university they are affiliated with or based on the European network organisation they cooperate with (e.g. ESFRI).

As a result, potential solutions or support from both an architectural or non-architectural point of view needs to be targeted to a very diversified audience where all types of stakeholders perceive added value by the initiatives that can be taken. When developing policy measures or supporting services, a differentiated portfolio is key.

Do not reinvent the wheel again

A lot of initiatives and infrastructural investments are going on or have been made in the past throughout the Flemish ecosystem. Some, like IMIS or VLIZ, have already been established for many years and proven their capabilities and use in the past. While others, at e.g. KU Leuven have gone through various years of planning and will be rolled out in the coming months. This variety of existing solutions should be maximally reused. This is also stated by the note[1] of Minister Crevits to the Flemish Government, mentioning to consolidate existing expertise as much as possible, as well as facilitate the interoperability between existing databases and reusing existing infrastructure. All these three elements are ample available in the Flemish research ecosystem.

In this context, relying on the existing ecosystem will enhance sharing knowledge and cross-sectoral cooperations. In order to facilitate putting this into practice, appropriate governance structures and agreements need to be developed, e.g. to agree on legal aspects such as ownership of data, identity and access management, standards, etc. During the exploration on which organisations are willing to share either knowledge, work processes, infrastructure, services etc. the gained expertise of the organisations should be the most important element being taken into account. As such, a large number of organisations can assist others based on their specific situation. Organisations who gained certain acknowledgement can serve as a service provider

[1] "Nota aan de vlaamse regering - Departement EWI." https://www.ewi-vlaanderen.be/sites/default/files/bestanden/nota_aan_de_vlaamse_regering_-_open_science_beleid_voor_vlaanderen_en_de_oprichting_van_de_flemish_open_science_board_fosb.pdf.

or as a mentor showing how they lead by example inspiring others and developing the maturity of the whole Flemish research ecosystem.

Next to developing an appropriate governance model, describing the technical interoperability approach is crucial to be able to put these guidelines into practice. A large variety of existing archiving and storage infrastructures need to be able to be connected with each other. These interoperability challenges need to be analysed on the level of each solution and standardisation on APIs to be used should be pursued.

Increase data awareness and clarity for researchers

In order to comply with certain requirements, researchers often need to provide other organisations or other internal business units with documentation about their work and the (meta)data that is used, shared, etc. for a variety of reasons. As a result, researchers often face specific requirements related to data on e.g. filling out data management plans. Researchers are aware these processes are needed, however raising awareness on the added value and necessity of these data management plans is still needed. Researchers can not be seen as data experts and there is a clear need for clarifying the requirements they need to comply with. Informing a diverse set of researchers using clear and straightforward communication and terminology is needed in order to eliminate any unawareness that might exist due to researchers not being fully submerged in the data topic. By organising easily approachable information sessions explaining why and how certain requirements are needed and how they can be filled out will increase the data awareness and opportunities available. The content of these sessions can be coordinated by e.g. FWO, i.e. the Coordination Hub, and accompanied by success stories of research institutions which have made significant progress in a specific area in recent years. As a result, learning from each other using the existing knowledge within the network can be attained. Further alignment with EOSC should however be taken into account as some initiatives are being developed such as creating a competence center to share information and expertise.

Thematic and standardised approach

The research landscape is very diverse and complete alignment between different sectors and research domains might be difficult due to completely different topics, working methods, agreements, etc. However, standardisation and alignment on technical solutions, standards used, data management will be more easily agreed upon and adopted applying a thematic approach. As a result, sufficient flexibility can be kept while making sure maximum alignment and cooperation is reached, also between different ESFRI's.

Research institutions are often overwhelmed by the vast amount of available technologies, interfaces and protocols to cope with their data management issues. This can be an API protocol to connect with FRIS, an authentication solution to access sensitive data or a tool to prepare and publish data in an Open and FAIR way. There is an important role to be played here by the FWO, as they can act as a guide and go-to source of information that provides best practices and standards across research themes, working closely together with thematic data centers and ESFRIs to define community based standards.

Further automation for improved efficiency and workload

While learning from each other and improving the understanding of specific data management terminology and respective documentation is needed, additional automation of processes and reuse of existing information should be developed to decrease the workload of researchers. Next to data management plans, similar information needs to be provided for e.g. GDPR compliance, evaluation by the ethics committee, reusing information for CV's, etc. Typically, filling out these forms or documents need to happen in local and specific templates, requiring a lot of manual work. Reusing already existing information, applying the 'only once' principle across the Flanders research landscape should be stimulated and facilitated, not only within institutes. Standardisation of these

documents and sharing information between information systems will help significantly and improve the balance researchers can keep between complying with administrative work and their added value by working on the research itself.

Maturity needs to be further put into practice

Based on the maturity scores assigned by category, a positive conclusion can be derived. A lot of organisations are knowledgeable about what FAIR, Open Science and Open Data is. The majority is working on or have completed their data procedures or strategies, while some have information material to inform researchers on best practices. Others might rely on larger institutions to manage these aspects, both at national and international level and in frequent cases, metadata is already shared. The first steps into increasing the maturity levels are made, while the next steps of putting these principles and strategies into practice need to be taken into account for the upcoming years. The deployment at scale of those strategies, developing data processes, developing support, actively monitoring and enhancing the open and FAIR aspect of research data to increase reuse will be on the agenda for the near future.

Open where possible, securely closed where needed

While the target should be to have Open Data, sharing metadata, reusing existing datasets, etc., in some occasions and for very good reasons, these datasets might need to be closed. Using sensitive data such as medical records or privacy related information, as well as specific competitive information of businesses can be reasons. These datasets need to be sufficiently protected by high quality identification and authentication modules. When guarantees can be made about the required closeness of the data, they can be stored in any optimally organised location.

Bring efforts made and successes to broader audience

On a day to day basis, researchers and institutions are investing heavily in the quality of their work, taking into account a variety of contextual requirements to which they need to comply, such as (meta)data management. Next to focusing on the FAIR- and openness of the data, supporting activities around this topic can be elaborated to further enhance and stimulate the prominence of all data related efforts that need to be made. Inspiring others by showcasing success stories, or creating a quality label to put efforts in the picture, or organise proof of concepts to experiment with data on a cross sectoral level, will bring the efforts that are made by researchers to a broader audience.

Priorities

Based on the requirements captured and the validation received during the workshops with the representatives of the Flemish research landscape, following requirements were most supported. Firstly, using existing technological solutions to avoid double investments is critical. Secondly double work should be avoided as much as possible, both in terms of data entry in different repositories as well as double administrative work by reusing generic information for multiple purposes (eg. DMP, ethics committee, GDPR, etc.). Thirdly, there is a need for research data storage capacity, as well as the management of it resulting in researchers not having to take care of this themselves. Fourthly, further alignment on all political levels is needed, making sure initiatives are aligned and as efficient as possible. Fiftly, stimulating the open and FAIR culture is needed, by providing tools and methods to facilitate the adoption as well as incentivising organisations and researchers that comply and put a lot of effort in these topics.

Next steps

Explore potential technical solutions for archiving and storage

Explore which technical solutions can be developed linking existing infrastructural solutions with each other to make sure these fit within a federated structure. During this exercise, the requirements for archiving and data storage need to be taken into account, considering existing and planned IT investments. When conducting a gap analysis of the current and the ideal to be solution, a central solution can be suggested to make sure all types of organisations receive support in their most critical needs.

Analyse potential further automation

Reusing research metadata and research data is crucial. As a result, existing databases or platforms should be able to connect to each other, wherever relevant. In addition, metadata should be distributed to FRIS, as well as to EOSC. Further analysis needs to be conducted to verify how FRIS can be connected to EOSC without increasing workload for researchers or institutions. This analysis needs to take into account that some organisations are immediately connected to EOSC without being connected to FRIS, while evaluation on Flemish level will be based on metadata available in FRIS.

In addition, administrative information is ample available in a variety of databases. Linking these will decrease the workload and administrative burden of researchers. Aligning which information to be provided is needed as well as the automatic reuse of specific information.

Develop information and services increasing awareness around (meta)data

In order to create awareness and support the bigger picture of Open Data, FAIR, EOSC, FRIS, data management plans etc, information material and information sessions need to be developed in an easily understandable manner. Building further on expertise of the network and showing success stories will increase cooperations within the network.

Promotion of Open Science, Open Data, FAIR data etc. is important as well. Developing a framework around these topics enhancing the willingness and visibility of the efforts made will improve the understanding and appreciation to a broader audience. Creating a quality label, such as an ISO standard, will help in this respect. Appropriate agreements need to be made on when and in which cases to provide such a label as well as determining standard requirements. In addition, organising cross-sectoral proof of concepts will enable researchers to test data sets from other disciplines and might result in unexpected results. This will also improve the acknowledgement of work done by researchers and institutions and increases the visibility.

In addition, providing assistance in solving questions on the principles mentioned above or in solving technical challenges, such as connecting with FRIS can be developed.

Annex I: Questionnaire

(*) (**) (***) These asterisks indicate the weighing that we apply to these questions. A question with three asterisks is considered more important to use than a question with one asterisk. We would kindly suggest you spend more time on questions with the highest weighing.

General Information

organisation name (***)

Representatives of the organisation during the interview (**)

Date of the interview (**)

Introduction

What is/are your role(s) regarding data (management)? (***)

In which aspects are you active regarding data (management)? (*)

Are you familiar with the Flemish Open Science Board and its goals? (***)

Are you familiar with the EOSC and its goals? (***)

Is your organisation familiar with FRIS? (Do you know that FRIS will be the gateway to EOSC providing FAIR-compliant metadata?) (***)

How are you providing metadata at this moment? (***)

Do you require or desire any interfaces in order to acquire this metadata? (***)

Organisation

What kind of profiles are present to manage data, support, innovation, ...? How many FTE? (*)

Do you have the right competences to maintain your data infrastructure? Or are any third parties involved? (*)

How are ethical considerations managed within your organisation? (*)

Do you have any incentives present / rewards to manage your data properly? (remuneration model) (*)

Do you have any legal advisory present for legal support on data? (*)

FAIR & Open Data

What are you doing to implement the FAIR principles? (**)

Do you have a PID policy for metadata and research data? (***)

Is there a process to monitor it? (***)

To what extent is the metadata and research data accessible today? (***)

Which data and for whom? (***)

What is the process to access metadata and research data? (e.g. protocols) (**)

What does this process look like and who is involved? (e.g. embargoed, open, restricted or closed) (**)

Has it been documented? (**)

How is the sensitivity of metadata and research data assessed within your organisation? (e.g. public, ethical/legal restrictions) (**)

To what extent is the metadata and research data interoperable? (e.g. knowledge representation, vocabularies) (***)

To what extent is the metadata and research data reusable? How is the data being reused? (e.g. machine-understandability, community-standards) (***)

How does the authorization and authentication work to gain access to research data? (***)

What are you doing to provide Open Data? (e.g. licencing) (***)

Do you have an Open Data strategy in place? If so, can you elaborate on the strategy? (***)

Do you have any specific people in charge of FAIR data or Open Data initiatives? (***)

What are the activities and projects you do on FAIR data? or Open Data? (***)

Technology

What applications or software do you use to store research data? What is the scope of these applications? (**)

What hardware is used to support this? (e.g. on-premise, cloud, ...) (***)

Do you maintain your own infrastructure or are parts outsourced? (*)

How does data get shared? (***)

Are these open protocols available? (e.g. API's, manual exports, etc.) (*)

On what kind of technology does your interfaces run? (e.g. REST, SOAP, etc.) (*)

Who can access your data and how can they access it? (***)

Under what types of format is research data stored? (*)

If applicable, how do you handle source code and research software? (GitHub, GitLab, ...) (**)

What is the amount of data (MB, TB, PB) you store and share with others? (*)

Governance

Do you have any processes in place to handle the management of research data? (assign ownership, maintenance of the data model, maintenance of codelists or controlled vocabularies, ...) (**)

Do you have any processes in place to check the quality of your data? (***)

Is any of this automated? (***)

Do you have any processes in place to handle data quality issues? (e.g. workflows, etc) (***)

Do you have any policy, procedures or rules in place to manage data? (security, access, publishing, maintenance, ...) (***)

Do you have responsibilities assigned to different research data owners? (*)

Are these responsibilities formalized in role descriptions? (e.g. data stewards) (*)

Do you have any workflows in place to handle data flowing from the source to registration (publication)? (*)

What is the process to acquire research data? (*)

Is this process documented? (*)

Can a user do this him/herself or do they need technical support? (*)

Metadata and research data

What metadata and research data is already shared between organisations? (***)

Which restrictions do you face when sharing research data? (***)

What kind of metadata and research data do you already share with other organisations? (***)

What kind of metadata and research data is already (or can you make) available through FRIS? (**)

To what extent are existing standards already used and / or implemented for the storage and / or disclosure of the data? (e.g. CERIF) (*)

To what extent has your data model been documented? (e.g. database schema or UML model) (***)

Have checked values and code lists been defined as part of the data model? (*)

Data Services

Do you have any RDM (Research Data Management) or DMP (Data Management Planning) services in place? (*)

How is data shared with other organisations today? Are you using any services to publish this data? (**)

Is there supervision of the use of the research data? (***)

Indication costs associated with data management (cash/in-kind/FTE)? (*)

How do you expect this will progress? (*)

What do you expect centrally? (*)

Aspects where collaboration is possible across institutes and other actors in Flanders? (**)

Where not (for your institute)? (**)

What is the policy on storing metadata and research data and archiving of this data? (***)

What solutions are currently used (or planned) for storing metadata and research data? (***)

Are these centralised or decentralised? Outsourced? (***)

Do you use certified / accredited repositories? (***)

Can you give an indication of how much research data is currently archived ? (***)

How much do you expect to be able to archive ? How will this progress in 3 years? (***)

Is there a need for additional solutions for storing or archiving metadata and research data in your institute? (***)

What is the major bottleneck? (***)

What are the estimated investments around data storage and management? (Figure per year + FTE's + Expenses = rough figure) (*)

Where do you expect the money will be coming from? (e.g. research projects, internal resources, FOSB, EC) (***)

If applicable, which exploitation models would you have in mind? (e.g. project financing, organisation financing) (***)

Shared Data Services

If a centralised storing / archiving is provided, would your institute be interested in using this? (**)

What are the key requirements to be able to do this? (**)

if the shared environment does not have analytics / AI / ... capabilities, but is pure "storage", would your entity use it for that purpose? (*)

if you would use the shared environment for storage purposes, would you use it as data archiving or data storage (or both)? (*)

Do you want to use the shared environment for analytical purposes? (*)

If yes, do you require any specific analytic tools? (*)

Would you want to use a shared environment for computation on the data? (*)

If yes, do you have any specific computation use cases (e.g. HPC) (*)

Do you have specific legal / regulatory requirements to be taken into account for storing the data centrally (e.g. GDPR, encryption, privileged access restrictions, ...)? (*)

TO-BE situation

What do you think you need to make a smooth connection with the FRIS services? (e.g. applications, interfaces, ...) (**)

If you start today with metadata exchange with FRIS what should be in place to make this transition go smoothly? (e.g. technology, financial and human resources, support, documentation ...) (**)

What do you expect the benefits to be from joining the EOSC? (***)

How do you anticipate providing your research data to EOSC? (e.g. Applications, interfaces, ...) (***)

Which parts of the architecture are lacking for your institute (gap analysis) to be able to meet the requirements of government? (**)

Is further formalisation to improve the efficiency of sharing/using data needed? And how can you improve? (**)

Do you foresee changes in the current infrastructure, technologies, processes, ...? (***)

What is your roadmap for data management in the coming years? (e.g. data transfer to the cloud ...) (***)

Do you have specific legal / regulatory requirements to be taken into account for storing and archiving the data centrally (e.g. GDPR, encryption, privileged access restrictions, ...)? (*)

Are there any other points on your mind? Do you have any suggestions, concerns, needs? (*)

Annex II: Interviews

Alamire

Alamire has a structural cooperation with KUL. This means that they inherit most of their data management maturity from their parent university, such as the availability of data stewards, their RDM helpdesk, IT and infrastructure support and the upload of metadata to FRIS. They have their own thematic repository in place called IDEM (Integrated Database for Early Music) which is very extensively built out in accordance to the FAIR principles. They have no specific needs from an Alamire perspective, all their needs are voiced in the KUL interview.

AMS

The Antwerp Management School (AMS) does mostly strategic basic research and applied research on behalf of the academic business world. As the business market and organisations are not concerned about Open Data, and often block the sharing of data, it is very difficult for AMS to comply with all principles of Open Data and FAIRness. They see the benefits of Open Data and share where possible, but lack the competences and resources to further develop.

AMS doesn't host any infrastructure themselves, but is highly dependent on the University of Antwerp. They need better guidance on Open Data, what it means and how to apply it. A large benefit for them, next to funding, comes from the supply of safe storage and archiving, as well as guidance in the world of Open Data and Open Science.

AnaEE

AnaEE-Flanders, as part of AnaEE-Europe, is committed to storing data and making it available via the Data and Modelling Center (one of AnaEEs service centers). However, since AnaEE is not fully operational yet, data management activities at the European scale are not operational yet either. AnaEE-Europe is a network of analytical platforms in Europe to test the impact of global changes (e.g. climate) and provide solutions based on the data.

As AnaEE is not fully operational yet, their maturity is still a bit lacking for the moment in certain domains. Specifically, in the category of data handling and data management new improvements and decisions will have to be made in the near future.

What AnaEE mostly wants is clarity and streamlining between the different levels (Europe - Flanders - Local), and training as technical knowledge is lacking for the moment.

CERN

CERN is an international institution that is focussing on the study of the basic constituents of matter – fundamental particles. The data they generate reaches up to gigabytes per second, implying that the infrastructure and data handling capacity of this organisation is world-class.

As an institution, they embrace the FAIR principles and open up data as soon as possible. In order to do this, they rely on their advanced decentralized Worldwide LHC Computing Grid. Because of their international nature, the organisation is not really aware of the FRIS portal.

Due to their advanced mindset and infrastructure, CERN indicated that they do not have any specific needs.

CLARIAH-VL

DARIAH-BE, Digital Research Infrastructure for the Arts and Humanities in combination with CLARIN, European Research Infrastructure for Language Resources and Technology joined forces in Flanders as CLARIAH-VL, Open Humanities Services Infrastructure. They are well advanced in terms of metadata FAIRness principles. However, this turns out to be a more difficult exercise when talking about research data, mainly due to IP restrictions. The ambition does exist to evolve towards a data-sharing community. However, in order to evolve to such a mindset, the digital fitness of researchers and user friendliness of sharing data should be improved.

When looking at the future, they expect that they will be needing additional funding and they wish for a streamlined process to publish data to FRIS and EOSC, instead of a multitude of processes.

DiSSCo

DiSSCo is the Distributed System of Scientific Collections and aims to mobilise further the collection holding institutions in Belgium and to have a complete inventory of Natural Heritage Collections. They have specific data management profiles in place to align best practices between their partners. From a European level DiSSCo has a provisional DMP published. They are also providing their metadata via PURE to FRIS and are mobilising other partners to take advantage of existing repositories such as GBIF. Working with their partners they see added value in standardisation in terms of metadata models, better training and alignment, and making sure that duplication of data entry is minimalized by integration and automation.

DUBBLE ESRF

ESRF is a European institute where researchers from different countries who are members of ESRF can make use of their facilities and measurement infrastructure. DUBBLE (Dutch-Belgian beamlines) are funded by the Dutch and Flemish research councils and are as such a facility on the ESRF site. All the research data that is generated using the ESRF facilities is managed completely under the ESRF data infrastructure, which can then be exported via specific data services to the researchers' institutional data infrastructure. Therefore the DUBBLE ESRF team does not handle research data of metadata themselves, everything is managed and controlled by other parties, albeit ESRF or the specific university the researchers work for. This means DUBBLE ESRF has no specific needs and cannot be categorised among the other institutes, and therefore is excluded from the maturity models.

ELIXIR

ELIXIR is an intergovernmental organisation that brings together life science resources from across Europe. ELIXIR does not generate data itself, but aims to enable/facilitate researchers to generate FAIR data and submit to ELIXIR Deposition databases, which is built around the FAIR principles. Within life sciences, they see that awareness towards open and FAIR data culture is an area of improvement. Next to that, they believe that the researchers should be provided the tools and expertise to aid this culture adoption. They believe that the FOSB can play an important role in this. They are a strong advocate for standardisation of interfaces, APIs and authentication solutions. They also feel a need for archiving capacity. Processing capacity would be a nice-to-have feature.

EMBRC

EMBRC is the ESFRI on European Marine Biological Resources and is funded by BELSPO and FWO. They partner with many institutes across Flanders such as Ugent, VLIZ, KUL and UHasselt. At a European level they provide metadata via IMIS which flows into FRIS. In terms of research data most data is kept at the institutional level, and on a European level there are some discipline specific repositories that are used. EMBRC aims to bring harmonisation of data management activities between their partners but find it difficult due to the huge diversity of maturity levels throughout the partnerships. They see a need to provide a clear return on investment for partners when implementing better data management. And also to bring standardisation wherever possible. The implementation of data stewards across the ecosystem is a great first step into that direction, however it is one of many steps that still have to be taken.

ESS

ESS is the Belgian European Social Survey institute that is connected to the European consort called ESS ERIC (European Infrastructure Consort). ESS Belgium consists of 3 people that keep their data on the KUL premise. Data collection happens via survey offices and is processed using the KUL infrastructure, to then be transferred to the data controller ESS ERIC. This means most of ESS' data management processes are inherited from their parent institute KUL, and that research data eventually ends up at the ESS ERIC infrastructures. Their overall maturity in data management is high as it's covered by KUL. ESS is open to using any centralized repositories if it is made available, but does not have any specific further needs of their own. These needs are covered by KUL, and any further improvements are in line with those of KUL.

EuroBioImaging

EuroBioImaging Flanders has a complex structure, existing out of the different universities as consortium partners. All separate BioImaging institutions across Europe make up EuroBioImaging as a whole. As a consequence, people, architecture, policies, ... are also divided between the consortium institutions who follow their own set of rules, which means a unified data governance and approach is very difficult to maintain.

We therefore see that the maturity on infrastructure and Open Data is very high as each institute has the necessary solutions in place. If we take a look at the maturity of data management and governance, we see however that this is very diversified.

EuroBioImaging has some concerns about the role of FRIS and the extended scope that is still unclear.

Flanders Make

Flanders Make is the strategic research centre for the manufacturing industry. Their goal is to contribute to the technological development of the vehicles, machines and factories of the future. Because of the nature of the organization, they are experiencing a duality in on the one hand an increased willingness to make data FAIR, but also an increased protection of private industry's data. Their infrastructure is based on Confluence, Zoho CRM and MS Azure (in development). A lot of their processes and policies are still in development and they are actively looking for the most optimal solutions to do this in the most efficient way with regards to the complexity of their research domain.

We noticed that they have a growing need for standardizing metadata, archiving data and sharing data with external partners in a cost-effective and efficient way.

ICOS

The Integrated Carbon Observation System, ICOS provides standardised and Open Data from more than 140 measurement stations across 12 European countries.The stations observe greenhouse gas concentrations in the atmosphere as well as carbon fluxes between the atmosphere, the land surface and the oceans. Thus, ICOS is rooted in three domains: Atmosphere, Ecosystem and Ocean. ICOS is on top of the broad data management practices. To illustrate, all their data is gathered following international standards (e.g. PID assigned to data and metadata) and quality checked. On top of that, all their data are available in open access, and all their processes documented. They rely on the University of Antwerp for their infrastructure needs. ICOS is currently lacking information and resources to connect to FRIS. Besides, because of making their data accessible to anybody who wants, ICOS doesn't get the recognition it should. To allow them to share all types of data in real time, the fluxes need to be improved. It requires an upgraded infrastructure (processing capacity needed, e.g. connection with the VSC). Lastly, ICOS would like the quality of data certified by FAIR-labels.

IMEC

Imec is a world-renowned research center for nanoelectronics and digital technology. They are in the middle of the development of a new data platform to capture metadata and research data for researchers, but are awaiting on the decisions of this exercise. Depending on what will be provided centrally, the development of their own platform will differ. However,it is already certain that this platform will be used to provide (meta)data to FRIS and EOSC. In order to accomplish this, they will work in three phases: 1. Quick wins needed, i.e. interfaces towards FRIS; 2. a minimal viable product to link data and foster collaboration; 3. a full blown data platform.

The biggest need from their perspective would be to bring standardisation on metadata schemes, central API's to forward data sets, and clear communication around these initiatives.

INBO

"Keep it simple and let's maximize added value for the researchers"

INBO is the institute for nature and forest research. They are currently investing in many Open Data initiatives such as promoting DMP's, increasing interoperability on their analysis flows and use many thematic standardised metadata repositories such as GBIF, Dryad and Zenodo. Their research data is stored in the cloud, more specifically on AWS. The biggest need from their perspective would be to bring standardisation on metadata schemes, API's, authentication protocols and DMP creation. More standardisation and integration will eventually mean that less duplication of data entry is made, especially when connecting with EOSC in the future. Making more computing power available would also be very beneficial to them, looking into possibilities via AWS or partnerships with the VSC.

Instruct-ERIC

Instruct-ERIC is a pan-European distributed research infrastructure making high-end technologies and methods in structural biology available to users. Research data is in most of the cases directly uploaded to the Protein Data Bank or the Genbank, which are data repositories that are open and accessible on an international level. Their research domain is very reliant on IP and patenting. They are of the opinion that the FWO can take up the responsibility in standardisation exercises. One of these ideas could be the generation of a standard CV based on the ORCID. They also advocate the idea of having an incentivisation system for persons or institutions that put a lot of effort towards making data open and FAIR.

ITG/ITM

Research is one of the three pillars in the academic mission of the Institute of Tropical Medicine (ITM). ITM's three scientific departments perform research on the level of pathogens (Department of Biomedical Sciences), patients (Department of Clinical Sciences), and populations (Department of Public Health). All researchers in the three departments are concerned with data and data management. Data is currently being extracted from the ITM's PURE database directly into FRIS. Their expectation is that EOSC will provide storage, sharing and reuse of research data across borders and scientific disciplines. Currently, their biggest concerns are with the policy implementation and hands-on support. Additionally, ITG is familiar with the Open Science goals but it is not clear how Open Science will pan out.

KMDA CRC

At KMDA CRC research and related activities are being done at the zoo of Antwerp and Planckendael. In this global community data of individual animals and populations is registered and shared centrally. Metadata is kept on PURE and flows into FRIS thanks to the licenses provided by EWI. There is no dedicated resource working on RDM, so support is needed. A central or federated repository would be very much welcomed in order to assist in the storage of research data. As a smaller player within this large research ecosystem there is not enough visibility on possible solutions and how they can evolve in their research data management.

KMSKA

The Royal Museum of Fine Arts Antwerp (KMSKA) is the sole Flemish museum with a scientific status, as they are not only responsible for maintaining their collection, but also to conduct scientific research into the works and techniques used. They are very reliant on software that is provided by the University of Antwerp (Brocade), which will be used in a later stage to feed their research metadata to FRIS. Their needs are very diverse. On the one hand, they indicate that there is a need to improve researcher's digital fitness and their awareness for data standardisation. Besides this, extra funding will also be a crucial success factor. Next to that, extra support in terms of legal issues around IP would always be welcome. The institution also indicated during the interview that for their case, implementing a (monetary) incentivisation system for opening up datasets would be very contradictory with regards to their current collaborative culture.

KU LEUVEN

KU Leuven already has a lot of years of experience in the field of research data. They strongly believe in the power of a federated structure, and as such also have a roadmap and strategy in place for the future. Their infrastructure is founded on the belief that data should be as close to researchers as possible.

Because of the size of the KU Leuven organisation, they provide support to the researcher population at several levels, i.e. the central level of the university as a whole, but also the level of the science group and the departments, so the knowledge and support for data management within the departments and research groups is developed and supported.

As the KU Leuven is such an important player on the research market, and also provides almost half of the research done in the Flemish environment, they score very high on the maturity scale. Their biggest need is funding for the moment.

LifeWatch

The Belgian LifeWatch project is part of the European LifeWatch infrastructure. LifeWatch was established as part of the European Strategy Forum on Research Infrastructure (ESFRI) and can be seen as a virtual laboratory for biodiversity research. They are very advanced with regards to metadata and data sharing. They have a strong team of data stewards that are also related to VLIZ or INBO. Their hardware is deployed on-premise and for some specific workloads they work with specific platforms that are developed with other parties (e.g. University of Amsterdam). As this institution is so advanced in terms of data management and sharing, they do not have any pressing issues in that regard. However, archiving seems to be something that will become a difficulty, but given their current budget there is no room for additional spending on archival storage capacity.

They see a huge upside in moving towards community-based standards and expect that the demand for profiles capable of reformatting data will grow significantly. The LifeWatch project is also an advocate for an incentive towards researchers and/or organizations that do a lot of effort with regards to publishing data in a standard format.

Orpheus

Orpheus is a research institution that focuses on artistic research, more specifically media, music, performances, videos or anything that can be seen as an artistic product with the exception of articles. As FRIS does not currently provide a metadata format for artistic research they are working with ECOOM to develop a registration model. Currently, they are manually updating FRIS around 4 times a year. Their research data is kept on the Google suite platform and also on their on-premise Linux NAS server with Qnap web services. They hope to be more informed about the developments in RDM and appreciate any support in implementing interfaces, data models, and so on. They are working with KUL and UGent to help shape their RDM policies, as they lack expertise themselves.

Plantentuin Meise

The Garden (Plantentuin Meise) is active in the Biodiversity Information Standards organisation (known as TDWG), as well as numerous EU-funded research projects (e.g. ICEDIG, DiSSCo Prepare, SYNTHESYS+) in which research into data management and its optimization is performed. They are highly knowledgeable on everything that is ongoing within the research ecosystem regarding data management. Connections are made with many platforms such as GBIF, FRIS, Zenodo, GenBank and many others. There is a need to provide clarity on how data stewards will connect to each other, but also to lower the costs for IT services locally by making specialists available. The garden would happily participate whenever a repository for long term and short term storage is made available, but also require computational power in order to analyse their results. An additional DOI in FRIS would enable linking metadata to other repositories.

Share

Share (part of University of Antwerp) is quite similar to what ESS is (part of KUL). Data collection also happens via survey offices and is processed using the UA infrastructure, to then be transferred to the Share EU platform. This means most of Share's data management processes are inherited from their parent institute UA, and that research data eventually ends up at the European Share infrastructures. Their overall maturity in data management is high as it's covered by UA. In their line of research it is imperative to protect their IP, so sharing Open Data doesn't come naturally. Another need is more clear communication and or training on DMP's and other RDM related activities that are suddenly required. Share asks to keep it simple for the researcher, and make it clearer what the added value is in improving your RDM.

UAntwerpen

The University of Antwerp has a very heterogeneous landscape when it comes to data. The type of data is so different across disciplines that it is nearly impossible to have a unified method to store, handle and share data.

As they don't have a unified approach to data, or a centralized infrastructure, they struggle with the handling and management of data. However, they do have the necessary solutions in place to store and share the (meta)data.

They fear that not all researchers are currently equally motivated or incentivised to participate in Open Data. Solutions might be to raise awareness and thrive for a mentality switch.

UGent

As one of the most important universities in Belgium, the university of Ghent is well familiar with the complexity and diversity of the research data landscape. They are well aware of the importance of good data management and this is reflected in a strong data stewards team. The current infrastructure is not ready to share data to third parties, which is seen as a pain point. However, the lack of this solution is directly linked to the very diverse data which the UGent possesses. When looking at the future, the institution is looking to invest in a central data register and data vault. It sees opportunities for the FWO to support and guide the institutions with data management itself, in combination with upskilling initiatives towards researchers.

UHasselt

As one of the universities, UHasselt already has the necessary infrastructure, governance and policies in place to provide FAIR and Open Data. However, they indicate that everything can be improved and streamlined more.

UHasselt is expecting to get a lot out of the national and international initiatives if they are organised well:

- Central storage, considering ease of access, GDPR, etc.
- Skill and competence centers
- Quality increase

UNU-CRIS

The United Nations University Institute on Comparative Regional Integration Studies (UNU-CRIS) is a research and training institute of the United Nations University. Due to their international character, they have no real data management focussed solely on the Belgian landscape. On top of that, all researchers that work with UNU-CRIS, are all affiliated to other universities, meaning that they only need to comply with the guidelines and standards from their affiliated organisations. Because of these reasons, this organisation was excluded from the research.

VIB

VIB is one of the four strategic research centers active in life sciences (IMEC, VITO & Flanders Make). They have a complex interuniversity structure where there is a central headquarter but research is done at the affiliated universities. They are not yet providing information to FRIS as it was having troubles to align with their needs such as the ability to indicate double affiliations, however those issues have been resolved and they are looking into integrating with FRIS in the near future. Besides FRIS they are also in between the use of different repositories spread throughout the universities they work with, which makes Open Data sharing a challenge that requires clear IP related issues to be resolved. They feel that a centralisation of data services is lacking at the moment, and would welcome more standardisation across the ecosystem. They see a combination of both top-

down and bottom-up approaches are needed here, there must be advantages from an institutional level as well as from a researchers' perspective.

VITO

VITO is conducting research towards sustainability in following five domains:

1. Energy
2. Materials
3. Health
4. Chemistry
5. Earth observation

Their maturity on data management, Open Data, infrastructure, … is dependent on the domain, as not every domain has widely accepted standards available for the moment. VITO is working on deploying a meta catalogue to make metadata available for researchers, and to provide metadata towards FRIS and EOSC. It is in this area that VITO indicates the need for extra support, both towards infrastructure and manpower.

Vlerick

"All help is welcomed with open arms"

At Vlerick business school they focus on research that is mostly done in collaboration with external partners. They are currently using DSpace to store their publication data, and are entering data manually into FRIS. As for research data they are actively migrating data to their cloud environment, SharePoint. In general they are knowledgeable on the FAIR principles but find it difficult to implement for a couple of reasons. One of these reasons is that they have limited resources to do so, as they have no specific experts in-house and most initiatives on data management are conducted on top of their normal workload. Another very important reason is that they almost exclusively work with external partnerships and the data produced under these conditions are always linked to NDA's. It's therefore more challenging to implement the 'Open Data'-principles when it comes to international collaboration with privately owned companies. The major bottleneck on their part is a lack of resources and in-house expertise when it comes to data management, and they would happily use whatever is made available as a central solution. When it comes to storing data the most specific need would be the ability to handle sensitive data and be able to restrict access when possible.

VLHORA - DOSP

The VLHORA is an overarching organisation that represents all the higher-education institutions in Flandes. They started the initiative of the DOSP, their custom-built platform for data management which is very similar to the FRIS platform and will directly be linked to it.

They are currently in a transition phase, as they have limited to no data management in place currently, but are putting effort into changing this around. They really see the need for extra training in order to raise data management awareness. They are aware of the pressure to shift towards a FAIR and open culture, but this is counteracted by several issues, no data-storing culture, such as IP conflicts and administrative burden. They see an opportunity for the FOSB to grant extra funding, support in training and to aid them in formulating an Open Science policy.

VLIZ

The Flanders Marine institute (VLIZ) is a research organisation that promotes and supports Flemish marine research. Founded in 1999, it already started accumulating research data in order to become a precursor of the FAIR and Open Data mindset and have their own, custom-developed integrated metadata system, IMIS. They have a strong team dedicated to research data management and really invest in the community by providing training in FAIR data management for scientists and data managers. They strongly believe that standardisation and formatting towards community-based standards drastically reduces the threshold to re-use data. For this standardisation exercise, they think it is of the utmost importance that this is done with the help of thematic data centers and ESFRIs.

VUB

As other institutes the VUB provides metadata on publications and projects via the PURE portal to FRIS. On top of PURE they are already interacting with plenty of interfaces that are well known such as WoS, PubMed, Scopus, ORCID, Embase, Mendeley and ArXiv.org. Furthermore they are looking into integrating with DataCite and/or OpenAire to import metadata. They are also planning on having around 3 data stewards in the upcoming years to support their researchers and bring further alignment in their data management practices. Research data is kept on the SharePoint environment and when required, shared via the secure Belnet transfertool FileSender. At this moment their priority is to establish an archiving architecture that is in line with the recommendations of the governing bodies. An important part of this architecture is that it can take into account the GDPR and other relevant legislations. Most of their needs are in line with what we have previously detailed throughout the report and are as follows: allow easy integration with existing repositories, keep the hands-on work for researchers as low as possible, implement machine readable DMP's, make publications more visible in multiple search engines.

Waterbouwkundig Labo

At the Waterbouwkundig labo they have multiple research departments with very different needs and maturities when it comes to storing data and data management. As a metadata sharing platform they are using IMIS hosted by VLIZ (following multiple standards) and for publications they use the FHR archive. At this moment research data is not being disclosed yet or shared as Open Data, it is stored on their SharePoint environment in different project folders. This shows their need for a centralized repository both for active research data and long term archiving. On top of that there is a clear need for more training and knowledge sharing, or even funding for implementations or specialists.