




Project acronym: CS3MESH4EOSC

Deliverable D4.5:

Report on future FAIR data architecture in ScienceMesh within the FAIR landscape in Europe

Contractual delivery date	31-09-2022
Actual delivery date	14-11-2022
Grant Agreement no.	863353
Work Package	WP4
Nature of Deliverable	R (Report)
Dissemination Level	PU (Public)
Lead Partner	DTU
Document ID	CS3MESH4EOSC-22-10
Authors	Frederik Orellana (DTU), Guido Aben (AARNet), Pedro Ferreira (CERN), Jakub Mościcki (CERN) 

Disclaimer:

The document reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863353

Versioning and Contributions History

Version	Date	Authors	Notes
0.1	23.10.2022	Frederik Orellana (DTU)	Initial draft
0.2	1.11.2021	Guido Aben (AARNet)	Proofing and extensions
0.3	4.11.2021	Maciej Brzeźniak (PSNC), Jakub Mościcki (CERN)	Proofing
1.0	10.11.2021	Frederik Orellana (DTU)	Final version

Table of Contents

Versioning and Contributions History	2
1 Introduction	4
2 Background	4
2.1 Motivation and drivers	4
2.2 Original T4.2 draft plan.....	5
2.3 Creation of T4.7	6
2.4 T4.7 objectives, methodology and plan	7
3 Premises and stakeholders.....	8
3.1 Can we make a difference – can the EU?	8
3.2 Stakeholders.....	9
• The National Archives.....	9
• DTU University Library.....	9
• The DeIC DM Advisory Forum.....	10
• Zenodo	Error! Bookmark not defined.
• European FAIR initiatives.....	Error! Bookmark not defined.
3.3 Market analysis	13
• Needs	13
• Competition.....	17
• CS3 value proposition	18
• Market analysis recap.....	18
4 Building a FAIR repository	19
4.1 Choosing a platform	19
4.2 Establishing a hosting platform	20
4.3 Installing and running Zenodo.....	20
5 EFSS expunge functionality	21
5.1 ownCloud	22
• Functionality walk-through	22
5.2 Nextcloud	25
6 ORCID integration	26
7 Discussion	27
8 Recommendations	28

Document positioning

This deliverable reports on work carried out in task 4.7 (T4.7). An account is given of engagement efforts with FAIR initiatives and use cases. It complements the reports of deliverables 4.3 (D4.3) and 5.2 (D5.2) with a discussion of identified challenges and opportunities in national- and institution-level FAIR deployments.

1 Introduction

In this document we report on the background and implications of implementing a FAIR data architecture in connection with the Danish EFSS service sciencedata.dk.

Task 4.7 is the result of a project restructuring in M8-M10 of the project period. This report starts with an account of the background and motivation for the task. Next, we describe the endeavors in T4.7 of enabling FAIR data depositing with a CS3 technology stack, and the viability of this in a real-world setting of a call for tender for a national data repository. We conclude by discussing lessons learned. Central questions include:

- Why is it hard to be open?
- Does FAIR depend on open access, open data and open source?
- Is there a viable role for EU projects in building FAIR IT infrastructure on the national level?
- On a university level?
- How prevalent is “ivory tower” development and can it be avoided?

2 Background

2.1 Motivation and drivers

CS3MESH4EOSC is funded by the European Commission under the Horizon 2020 programme, Excellent Science, Research Infrastructures, e-Infrastructures. From the [work programme](#):

“Research infrastructures play an increasing role in the advancement of knowledge and technology and their exploitation. By offering high quality research services to users from different countries, by attracting young women and men to science and by networking facilities, research infrastructures help to structure the scientific community and play a key role in the construction of an efficient research and innovation environment.”

Clearly, an important motivation from the commission is supporting European science and research by helping to create high-quality research data infrastructure.

From our approved project proposal:

“The objective of CS3MESH4EOSC is to [...] enable friction-free collaboration between European researchers, without requiring these researches to relocate their data”

An important motivation for our specific project is indeed to help establish high-quality research data infrastructure - for *collaboration* across borders in European research.

From D4.1:

“These applications aim to cover as exhaustively as possible the workflow of activities in the daily life of researchers and help scientists and other users of data-driven environments at various stages of research and the publication process and various points of the research data lifecycle.”

In fact, the motivation behind WP4 can be said to distill and exemplify the motivation behind the project and the programme: WP4 aims to support researchers in their daily work – with a focus on collaboration. In this light, the remaining work packages are what make this possible.

Specific focus areas of this support, which we have deemed important (see our proposal and D4.1 for details), include services for:

- 1) Sharing access to datasets for browsing, viewing, editing and downloading
- 2) Sharing access to working with datasets without physically moving the data files, i.e. sharing access to analyzing and processing facilities
- 3) Sharing datasets on a publication platform offering search, indexing and bibliometric functionality
- 4) Bulk data transfers

The present task under WP4, T4.7 deals with 3) i.e. publication of research data.

To drive home the above chain of arguments:

The overall motivation behind T4.7 is to enhance the data publication functionality available to European researchers.

Other important priorities of both the EU commission (stated in the above programme and in more recent publications, e.g. this [note](#)) and our project include supporting European innovation/IT-industry and Open Science, Open Access, Open Data – and GDPR.

A strong driver behind T4.7 is an ambition to create European, open alternatives in the area of research data services, to commercial closed-source offerings, hosted on overseas cloud platforms.

2.2 Original T4.2 draft plan

Originally, T4.7 did not exist. In the project proposal, data publishing was part of T4.2:

“Task 4.2: Open Data Systems (DTU, PSNC, AARNET) -This task will integrate open data repositories (focusing on OpenAIRE standards such as OAI-PMH) in the Science Mesh. The application components developed by this task will provide users of the EFSS services with the ability to organize work-data via tagging and metadata assignment, turn a set of work-data into a published, referable dataset and finally expunge a valuable dataset to an open data repository, archive or library for curation and long-term archiving. EFSS user’s home service account will be associated with a persistent identifier (such as ORCID) and this identity

will follow the data when publishing to open data repositories. User will be able to tag datasets as public and make them public and searchable on the home service. Open data packaging archive formats such as Dublin Core, the Library of Congress BagIt, etc. will be investigated, as well as the ingestion into third-party archival services such as national archives or libraries. AARNet and DTU will provide components based on previous prototypes for ownCloud and Nextcloud and they will extend them as needed. This task will also build on Open Data prototypes at AARNet currently in internal testing. The task will perform validation with the user community and early-adopter testing.”

DTU was the task leader and main contributor. The above was fleshed out in spring 2020 by DTU in a draft plan, [still available](#).

2.3 Creation of T4.7

During spring and summer 2020, it became clear that the T4.2 draft plan was not realistic: DTU was not able to fill the almost 36 PMs with the necessary highly skilled manpower. Finally, it was decided to transfer the PMs to other partners and sharpen the focus to the Describo/RO-Crate metadata and packaging platform championed by AARnet - with which DTU had no experience, but for which relevant manpower was available via AARnet, WWU and PSNC.

While work on metadata, archiving/packaging formats, data expunge and EFSS integration would remain in T4.2, now lead by AARnet, it was decided to create a new task 4.7 dedicated to focus on the final, receiving end of the research workflow – the data repository. The work would consist in exploring possible roles of CS3 technologies and infrastructure in building, operating and/or interfacing with FAIR research data repositories.

T4.7 Future FAIR Data Architecture (DTU) 7PMs

Identify and reach out to research groups, archives/libraries and establish rationale for future FAIR Data architecture and its relation to external FAIR services and initiatives in Europe and beyond. Focus on Danish National Archives use-case and test with the [sciencedata.dk](#) platform.

- Engage with Danish National Archives on prototyping the concepts for future FAIR architecture
- Engage with university libraries on understanding their needs
- Engage with Zenodo on communities support and distributed curation concept
- Compile list of expunge and registration destinations besides Zenodo ([re3data.org](#), [datacatalogue.cessda.eu](#), Australian archive (?), ...)
- Prototype concepts based on [sciencedata.dk](#)
- Engage with Danish National Archives on validating functionality and flow
- Provide input for Task 4.2

The DTU role in the StC would evidence high level role in FAIR and would include:

- Reach out and engage with relevant FAIR initiatives in Europe (FAIRsFAIR [www.fairsfair.eu](#), GO-FAIR [www.go-fair.org](#),...)
- Engage with RDA (<https://www.rd-alliance.org/>)

Milestone M4.5; DTU, M18

“Use-case overview, engagement with FAIR initiatives; identified challenges and opportunities”

Deliverable D4.5; DTU; M32

“Report on Future FAIR Data architecture in ScienceMesh within the FAIR landscape in Europe”

Explicitly not in scope of Task 4.7 is the implementation, at the metadata layer, of integrations between the IOP and existing 3rd party subsystems. This involves EFSS vendors but also cloud providers of research relevant applications, e.g., office365, OSF.

Project description agreed by the StC, following the StC meeting on 30/07/2020.

2.4 T4.7 objectives, methodology and plan

As part of the above restructuring, DTU engaged in dialogue both internally in the consortium and with external, local players in the research data arena, notably with [Rigsarkivet](#) - the Danish national archives, and the data management group under [DeIC](#) - the Danish e-Infrastructure Cooperation. The outcome of the interactions was the following high-level objective and methodology.

Objective

- Assess the feasibility of building FAIR compliant repositories for long-time storage and curation of research datasets building on and integrating with CS3MESH4EOSC and related technology and infrastructure.

Methodology

- Investigating embedding strategies in the research data cycle: regional, national, institutional or local levels
- Matching this to distributed curation strategies
- Assessing advantages and disadvantages in a FAIR perspective

T4.7 high-level objective and methodology.

During Q3 and Q4 of 2020, more concrete action items were agreed on with the national archives:

- Aim: Build a national research related archive with FAIR support and ScienceMesh integration
- Collaboration between DTU and the Danish National Archives
- Integration between sciencedata.dk and the new archive
- The Danish National Archives delivers the governance and the legal framework for national research archiving
- DTU delivers technical expertise in building and maintaining the archive

First T4.7 first draft plan.

The idea was to build a repository and use both the planning of this and the actual implementation as tangible talking points when engaging with stakeholders. This draft plan also never saw actual execution. In the end, a formal agreement with the national archives was never reached. In Q1 2021, a plan was agreed on - with the national archives in a consulting role.

- Aim: Build a research related repository and archive for DTU with FAIR support and ScienceMesh integration.
- Involve DTU research IT dept. and the Danish National Archives.
- The archive will integrate with sciencedata.dk and CS3MESH4EOSC.
- DTU will deliver governance and the framework for DTU research projects using the platform.
- Formal project agreement planned to be signed during Q3/2021.
- Detailed milestone planning and manning by end of Q4/2021

Final T4.7 plan.

This plan still includes the actual construction of a repository; without up-front national ambitions, but still with an aim of focusing the dialogue with stakeholders on a concrete implementation.

3 Premises and stakeholders

The purpose of T4.7 was, in one specific country, at one specific university, to explore the tangible challenges involved in our European FAIR mission – and to extract lessons of a general nature.

In this section we'll look at the general situation on general-purpose FAIR public data infrastructure in Denmark. We start by a quick summary and give more background and details in the subsections below.

In 2020 at project start:

- There was no national general-purpose research data repository in operation.
- EOSC was still in an upstart phase, largely irrelevant in Danish academia, and known primarily by a handful of stakeholders in and around the Danish e-Infrastructure Cooperation, DeIC.
- Zenodo was known in research library- and some research communities, but not in widespread use.
- There was no national EFSS service.
- A university-level data repository (erda.dk) was in operation at the University of Copenhagen - based on in-house software.
- A university-level EFSS service (sciencedata.dk) was in operation at DTU.

In November 2022:

- The use and penetration of EOSC services, including B2SHARE, B2DROP etc., in Danish academia is still marginal.
- Zenodo remains known and generally condoned, but not actively promoted by university libraries/managements.
- DTU has launched a university-level figshare repository service.
- The Danish e-Infrastructure Cooperation, DeIC, has started two projects with the aim of becoming national data store and data repository services – based on erda.dk and DataVerse respectively.

Summary of FAIR data infrastructure stakeholder situation in Denmark in 2020 and 2022.

3.1 Can we make a difference – can the EU?

Building and operating a university-level research data repository with open-source tooling, involves not just technical, but also political and cultural challenges. The steps involved can be summarized by:

- 1) Hiring a team of developers
- 2) Building a technically compelling product/service
- 3) Convincing an IT department to run/host it
- 4) Convincing a university administration and IT support to recommend it
- 5) Convincing researchers to use it

The above points should preferably be executed in the order written. In reality, that's rarely the case. Also, for completeness, we may prepend a few points:

- a. Conceive great ideas for a research data services

- b. Assemble a group of like-minded people
- c. Apply for and obtain (EU) funding

These lists summarize the playbook of many EU-funded research IT projects. Ideally, however, the points, a-c, should be preceded by:

- A. University managements realize they need to take better ownership of their research data and provide better and more open data services to their researchers
- B. They ask the commission for help
- C. The commission realizes many universities have the same challenge and issues a call for tender on possible solutions

It would seem that chances of success depend largely on the alignment of a-c with A-C, i.e. on whether your particular product/service is actually requested by your particular university. In the absence of this, it will be up to you to convince your local stakeholder of either replacing existing services or embarking on new ventures.

In the absence of A-C, i.e., a situation where you find yourself externally funded, developing and operating products/services without active blessing and support from your administration, there is a risk of these products/services not being used by researchers, regardless of their quality.

In the following sections, we will zoom in on our specific project - in a specific country and university.

3.2 Stakeholders

In this section, we will look at third parties relevant to our endeavor and our interactions with these.

- **The National Archives**

From a Danish FAIR perspective, the National Archives are an important and interesting player:

- In April 2020, a government [executive order](#) was issued, specifying that all Danish universities and public state research institutions are obliged to report all produced research data (with exceptions specified in the text).
- The reporting must include a specification of the authority responsible the archival of the data.
- After the reporting, the National Archives will decide whether or not the data must be delivered to the archives.
- The archives do not operate a general-purpose data repository.

From the national archives, the contact person also joined the CS3MESH4EOSC advisory board. Unfortunately, the contact person, after initial meetings, went on leave of absence for personal reasons - making it difficult to establish a role for the archives in the project.

- **DTU University Library**

In 2019 the DTU University Library procured a [figshare](#) instance, under the URL [data.dtu.dk](#). In the preceding years meetings had been attempted with the DTU/sciencedata.dk team with limited tangible outcome – to

some extent mirroring the development at the Dutch library referred to [below](#), and likely other libraries across Europe.

- **The DeIC Data-management Advisory Forum**

The [DM Forum](#) is an advisory body under the Danish e-Infrastructure Cooperation, [DeIC](#), which aims to increase dialogue and user involvement in the development of digital infrastructure. It was tasked with writing the call for tender described below. It has also been active in attempts at establishing the role of “data steward” at Danish universities. An account of this work can be found in a report from 2020:

[“National Coordination of Data Steward Education in Denmark: Final report to the National Forum for Research Data Management \(DM Forum\)”](#)

This work was carried out in parallel with [similar efforts](#) in Holland.

Recently, we prepared a questionnaire on FAIR infrastructure in Denmark.

This questionnaire is addressed to stakeholders from research and infrastructure enablers, with the purpose of charting existing FAIR offerings in European countries.

- **What is your role in your organization?**
- **Which European FAIR projects and organizations, if any, does your work relate to?**
- **Can you list one or a few examples of general-purpose FAIR data services (repositories, storage services) available to researchers in your country?**
 - **Are these, to your knowledge, used by a) most researchers, b) some researchers or c) few researchers?**
 - **Are they offered on an a) institutional, b) national or c) European level?**
 - **Do the services, to your knowledge, abide to the GDPR?**
- **Specifically, do you see the metadata functionality offered by existing repositories as adequate from a FAIR perspective?**
 - **Is per-deposit metadata sufficient?**
 - **Should schemas be user-customizable?**
- **Do you see a need for direct transfers/deposits from storage services to repositories, and if so, is this need covered by existing solutions?**
- **Generally, from a FAIR perspective, are the data storage, sharing and publication needs of researchers in your country covered?**
- **Should these needs be covered by a) pan-European services, b) national services, c) institutional services, d) commercial providers – or by a mixture?**
- **Specifically, when pan-European general-purpose services like Zenodo, B2SHARE and B2DROP exist,**

do you see a need to offer similar or identical services on national and institutional levels, and if so, why?

Questionnaire on availability of FAIR data services.

Time permitted only one interview to be conducted - with the former head of the DeIC DM advisory forum; the interviewee, however provided written answers which are reproduced (with his permission) below.

- What is your role in your organization?

Senior consultant, previously Head of Data Management

- Which European FAIR projects and organizations, if any, does your work relate to?

EOSC-Nordic

FAIR Impact

Skills4EOSC (DeiC, not me personally)

GO FAIR International Support Office/GO FAIR Denmark Office

RDA

GÉANT

EOSC Association (DeiC is "Mandated Member" for Denmark)

- Can you list one or a few examples of general-purpose FAIR data services (repositories, storage services) available to researchers in your country?

ScienceData (by Technical University of Denmark)

UCloud (by University of Southern Denmark)

are to my knowledge broadly available, although not "officially" branded as national services (by DeiC)

Royal Library Dspace repository.

DTU Data repository and ERDA/SIF storage are such general-purpose services available for respectively DTU and KU researchers and possibly their research project affiliates.

National repository and storage services are in project stage.

- o Are these, to your knowledge, used by a) most researchers, b) some researchers or c) few researchers?

I would say b) some researchers without knowing any more precise details. Great differences between different faculties' use of the institutional services.

- o Are they offered on an a) institutional, b) national or c) European level?

Institutional level, although some are open to national use. Business models for that are however only partially developed.

National services in preparation.

I am not aware of European availability.

- o Do the services, to your knowledge, abide to the GDPR?

To some extent, depending on policy, as some only allow open access data.

Some services, such as SIF from University of Copenhagen, is dedicated to sensitive data with special procedures in place.

Most services maintain a level of security management, logging, etc., some are ISO27001 certified.

- Specifically, do you see the metadata functionality offered by existing repositories as adequate from a FAIR perspective?

The short answer is largely not. Their focus is on human actionable, rather than machine actionable. Metadata are often limited, with respect to the Reuse part of FAIR. And especially the semantic properties are poorly implemented or not at all ([I1. \(Meta\)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.](#), links to controlled vocabularies or ontologies are rare). PID strategies are also to some extent deficient, for instance with respect to relevant PIDs for data files (need not be same as for datasets).

- Is per-deposit metadata sufficient?

As an infrastructure person, I am not really the one to be judging, it would depend on actual use cases. As the general-purpose repositories often only support very limited bibliographical metadata, I am not sure who these repositories serve. Do scientific users look for data this way? Or are they rather supporting political purposes. Pointing to the question: what actually constitutes the value of FAIR, the answer probably differing between scientific communities.

- Should schemas be user-customizable?

That could be a possibility for scientific users to secure that relevant discovery metadata are always included - for actual scientific use. It would require mechanism for support of community metadata standards ([R1.3. \(Meta\)data meet domain-relevant community standards](#))

- Do you see a need for direct transfers/deposits from storage services to repositories, and if so, is this need covered by existing solutions?

I am not aware of current level of implementation. And not sure what the question exactly means: like transfer of large files through specialized ingest protocols? Some repository software implements that, but not sure about the already established solutions. In general, for the national upcoming repository, they need to answer for how to handle large datasets (=lots of files) and large files. If not ingested through repository software, then how to reference the files in a FAIR way, how to organize the ingest workflows, etc. These are classical repository issues, but still relevant!

- Generally, from a FAIR perspective, are the data storage, sharing and publication needs of researchers in your country covered?

No, implementation is just starting, and the actual practical issues arising from the FAIR principles are still very poorly understood. We need more practical experience with e.g. upcoming national services to understand what is needed for FAIR to actually become an asset to research, and not just a cost.

- Should these needs be covered by a) pan-European services, b) national services, c) institutional services, d) commercial providers – or by a mixture?

I think a mixture. The EOSC SRIA envisions a combination of institutional, national and pan-European services, the crucial question is how to divide the tasks between these parties, this is still to be developed. Commercial providers are to me okay, provided the governance remains in public hands. Research outputs from public research organisations is a public good. So no repetition of the scandal of the publication industry taking over scientific publication.

- Specifically, when pan-European general-purpose services like Zenodo, B2SHARE and B2DROP exist, do you see a need to offer similar or identical services on national and institutional levels, and if so, why?

Governance, support, and control of data could be reasons to still provide national services, also it would seem that national funding might call for a tighter adaptation to specific user or organisational needs. On the other hand, we could probably make more use of pan-European general-purpose services than is currently the case.

Answered questionnaire (December 1, 2022) on availability of FAIR data services.

3.3 Market analysis

In this section, we will look at the supply and demand situation in the area of research-data repository services at our specific university and in our specific country.

- Needs

Summary

The general situation regarding FAIR data depositing in Denmark we saw in 2020 (project start) can be summarized by:

- 1) **Functionality for expunging from EFSS services to research data repositories is requested by researchers.**
- 2) **There was a clear case for a national research data repository for Danish research.**
- 3) **It was unclear on the political level, who should or would build and operate such a service.**

Identified FAIR data needs in Denmark (2020).

Details and background are given below.

2019: National needs

In 2019 a report, "[Data management in Denmark](#)" ([attachments](#)), was published by [DeIC](#) - the Danish e-Infrastructure Cooperation. The purpose of the report was to provide recommendations to the research ministry and university managements on national data management services. In the document it is stated that only a limited information gathering at the universities was performed:

- Communication with relevant persons at the universities
- A smaller investigation conducted at the University of Copenhagen (see the attachment above)
- An earlier investigation at the University of Aarhus (not published)
- The personal experiences of the authors
- Discussions in DeIC a DeIC report from 2017: "[World-class digital infrastructure in 2025](#)"
- A report from the research ministry: "[Preanalysis: Introducing FAIR in Denmark](#)"

From the list of recommendations for new services – the ones relevant for us are:

13

- A. **Development of a fundamental [data management] services for researchers with limited [storage] needs:** This is to cater to the very large group of researchers from especially the humanities and social sciences who at the time [2019] saw no solutions covering their data management needs – typically easy storage, sharing and early DOI assignment.
- B. **Establishment of a common search service for [Danish] research data** across various existing registries of research data. Accompanied by work on common metadata standards.
- C. **Establishment of a national trusted repository** to allow permanent deposits of data accompanying articles published to research journals – with DOI and ORCID support and acting as alternative to repositories operated by the journals.
- D. **National sync&share offering:** Sharing of small and large amounts of research data must be supported – both as storage with full administrative rights to each user and as safe file sharing over the Internet. Solutions exist [?], but it is recommended to investigate using commercial services like Dropbox for smaller data volumes.
- E. **Establishment of sufficient volume of nationally accessible storage:** It is important to negotiate good terms for access to storage – which meet various demands like quick vs. slow access, big vs. small capacity and security. It must be made easy for researchers to migrate from one platform to another. It is recommended to first carry out a thorough investigation of needs.
- F. **Postponing of activities related to long-term storage** till later in the strategy period [2020-2025]

Comments: The existence of the upcoming EOSC is acknowledged, but no recommendations for Danish participation are given. Two of the above points (C, D) are very close matches to what CS3MESH4EOSC is providing - D precisely matches the aims of T4.7.

2019: Local needs, University of Copenhagen

One attachment of the report discussed above contains answered questionnaires and a summary of oral feedback from 3 professors, 1 professor MSO, 2 associate professors, 1 engineer and 1 research IT coordinator at 4 Faculties across the University of Copenhagen, on their views on the needs for new national data management infrastructure. The feedback could not be taken over one-to-one, as the situation at DTU differs from the situation at the University of Copenhagen – primarily by the fact that the latter had a data storage service, operated by the institute of physics, sanctioned by the faculty administration (faculty of science). The feedback relevant for T4.7 was similar to that obtained in our informal interactions with DTU researchers. Conclusions from the report:

1) A great diversity in research warrants a range of data management solutions

There is a great diversity in research projects conducted at UCPH, in terms of: How data are obtained. The volume of data generated and what formats they are stored in. Whether the data sets contain personal information, confidential information, publicly available information, etc. This diversity means that while one data management solution works for research group X, it may not be suitable for research group Y.

2) The need for new technical solutions is relatively small

The majority of respondents indicate to be happy carrying out their research with the technical solutions available to them now. When asked specifically for new infrastructure that could be useful: 3 indicate to want to have tools that make data processing and analyzing easier, 2 of these specifically relating to personal data projects. 7 indicate to want to have data collaboration solutions that help them work with externals in active research projects, such as a secure Dropbox equivalent both for personal/ confidential data and non- sensitive data, and a project management/communication tool. 4 indicate to want a better way of preserving data, either by improving an already available solution, or by creating a repository for

long term storage of personal data.

3) The need for improving existing technical solutions is bigger

Most respondents indicate to be content with the data management solutions they currently work with, if only: It could handle more data (for free). It was suitable for handling and storing of personal/confidential data.

- The interface would be updated and made more user friendly. The solution would be more aligned with FAIR (better metadata, a DOI option).

- There was better support. Many of the respondents suggest that some of the future efforts are directed towards improving existing solutions, for example data archiving offered by the National Archives.

4) There is a need for alignment between existing solutions

Some researchers express the wish for better alignment between already existing solutions, for example so that data sets can more easily be migrated from one solution (e.g. a data collaboration platform) to the next (e.g. a data repository). There is also a need for a better overview of what technical solutions are available locally, nationally and internationally and for support to help determine the best suitable solution.

5) A national data management network for information exchange between universities is important

It is suggested that the National Forum for Data Management (DM Forum) continues, in order to facilitate information exchange between data management support staff at the universities, and to follow national and international developments.

Conclusions from user survey on data management infrastructure, University of Copenhagen, August 2019.

Data management need	CS3 coverage
1. Tools that make data processing and analyzing easier	Jupyter notebooks integrated with EFSS service (T4.1)
2. Data collaboration solutions that help them work with externals in active research projects, such as a secure Dropbox equivalent	Explicitly covered by the DTU CS3MESH4EOSC node – sciencedata.dk
3. A better way of preserving data	T4.7
4. A repository for long term storage of personal [sensitive] data	T4.7
5. Improving data archiving offered by the National Archives	T4.7
6. Easy migration from one solution (e.g. a data collaboration platform) to the next (e.g. a data repository)	T4.7

Identified needs and their coverage by CS3 services.

Comment: The needs reported by researchers show a complete coverage by CS3MESH4EOSC.

2020: Local needs, DTU, AU

In June 2020, we conducted a user survey among users of the DTU EFSS service, sciencedata.dk. A questionnaire was sent in advance and interviews conducted with 6 major users.

What wishes and requirements did your project have for storing, sharing and archiving of data?
Was the service capable of meeting your wishes and requirements?
Did any issues occur during the use of the service and how well were they resolved?
Can you name new integrations or functionality that should be considered for the service or the CS3MESH4EOSC project?

Questionnaire of 2020 ScienceData user survey.

The respondents were 2 professors, 3 associate professors and a consultant from DTU (1), Aarhus University (4) and VIA University College (1), Several requirements and wishes were shared by multiple respondents.

Requirement/wish	Number of respondents	Supported in 2020	Supported in 2022
Data on national (Danish) premises	2	Yes	Yes
Support for sensitive data (high security, data processor agreement)	3	Yes	Yes
External sharing (outside eduGAIN)	1	Yes	Yes
Synchronization of shared folders	2	No	Yes
Easier onboarding (e.g. sync client installation wizard)	3	No	No, but documentation improved
Persistent accounts	1	Yes	Yes
ORCID support	1	Yes	Yes
Collaborative office documents	2	No	No
Long-term storage guarantees	1	Yes	Yes
Continued free of charge operation, or low cost	3	Yes	Yes
Expunge to data repositories	3	No	Yes
Metadata	1	Yes	
Viewers beyond standard set	2	No	Jupyter viewer
Compute integration	1	No	Yes – Jupyter service
Documented FAIR support	1	Yes	Yes
DOI support	1	No	Partly, via expunge to Zenodo

Requirements/wishes for EFSS service in 2020 ScienceData user survey. Extra column with 2022 situation.

Comments:

- Zenodo was mentioned explicitly by 2 respondents
- Two mentioned expunging to a national repository, one voiced doubts about the need for a such
- Jupyter notebooks (on Google colab) were mentioned explicitly by one respondent
- 4 rejected Microsoft solutions plus Dropbox out of security/privacy and functionality concerns

2021: National needs

In September 2021 a call for tender for a “national trusted repository” was issued by [DeIC](#) - the Danish e-Infrastructure Cooperation. Bidders were restricted to the Danish universities. The specifications of the call closely matched the feature set of repositories like Zenodo and B2SHARE – further confirming that a national repository was (is) indeed needed.

- **Competition**

At the project start, there was no general-purpose research data repository service in operation at the national level in Denmark (there still isn't). Zenodo and B2SHARE were in operation at the European level. At the university level, as already mentioned, a figshare instance was in early operation at DTU. None of these services supported importing from EFSS services ¹ and none were in widespread use in Danish academia.

Zenodo and B2SHARE

On an international level, Zenodo had already been in operation for almost a decade. It was not easy to judge how widespread the use was among Danish researchers. Our impression was that some had heard of it, but few had deposited research data. The Zenodo search interface allows searching the contributors field via a search pattern like e.g. "contributors.*: denmark|aalborg|aarhus|dtu|roskilde". Today (November 3, 2022), this yields 270 records. Thus, it appears that only a small fraction of the Danish research data has been deposited to Zenodo. The search "contributors.*: *" yields 43048 records. If we assume this corresponds to all European deposits, although it is a large number, scaled by population (Denmark/Europe) it corresponds to $5.857/746.4 \times 43048 = 338$ records – and thus (remember author lists generally span multiple countries) Denmark appears to be below average when it comes to Zenodo usage, but not dramatically so.

The EU project EUDAT operated, and still operates, the service [B2SHARE](#), which is based on the same software as Zenodo (Invenio). The above two searches yield 1 and 247 respectively. Thus, the service does not appear to play a role in Danish or European research.

Despite the similarity of the Zenodo and B2SHARE services, they do exhibit some notable differences (see table below).

figshare

[figshare](#) is a company which sells data repositories – i.e. establishes and runs repositories, typically for universities and libraries. The repositories are hosted (on Amazon's AWS), and the software powering them is closed source. figshare offers much the same functionality as Zenodo and B2SHARE. In particular it also offers an API for creating deposits. In 2019 the DTU library started offering a figshare instance under the URL [data.dtu.dk](#). In late 2021 the service hosted 14 records, currently it hosts 575 records.

Feature	Zenodo	B2SHARE	figshare
eduGAIN login	no	yes	no
Github login	yes	no	no
Github integration	yes	no	yes
EFSS integration	no	no	no
Funding/grants integration	yes	no	yes
Community/projects support	yes	yes	yes
Dynamic metadata-schemas	no	yes – community dependent	no

Zenodo/B2SHARE/figshare feature differences, June 2020.

¹ At present B2SHARE does support importing from B2DROP (Nextcloud) and B2DROP does support expunging to B2SHARE.

- CS3 value proposition

An important value proposition of the CS3 technology stack is connecting services, i.e. to become part of the “ScienceMesh”. This is a long-term proposition, and will be become more compelling as more services join. At this point, the relevant CS3 service to integrate with a research data repository is the EFSS service itself. Below we’ll look at what specific edges our endeavor might or might not have over the competition.

Advantages

- GDPR compliance for PII data (on premise service)
- Being part of the C3MESH4EOSC project and the CS3 community allows us to draw on the expertise of international experts

Possible impediments

- We restrict ourselves open-source software
- We restrict ourselves to on-premise operation
- Our software stack requires highly skilled labor to deploy, customize, operate and support
- Operating a research data repository at a university requires agreements and clearance with university management
- Meeting tight security requirements can be perceived as challenging for a smaller group: the small team operating on-premise services vs. e.g. the Microsoft team operating OneDrive, the team operating figshare and the Amazon team operating the platform on which it is running (AWS)

- Market analysis recap

The user requirements and needs analyses above, the general push for FAIR by policy makers and the interviews and analysis conducted in D4.3 all indicate that, at present, one main FAIR value proposition our project could offer, is functionality for expunging datasets from EFSS services to research data repositories in an easy and organized fashion, with special attention to metadata.

The above forms the background for what we decided to build in a T4.7 context:

- **A FAIR repository prototype**
- **Expunge functionality on the DTU EFSS service (sciencedata.dk) to²:**
 - **Zenodo**
 - **The new repository**

² Limiting ourselves to two expunge destinations is a matter of priorities; it should certainly be extended, both with repositories operated by ESFRI projects and with national domain-specific repositories. A list can be extracted from re3data.org. Prioritizing such a list and building support for the selected repositories requires interactions with research environments and repository operators and entails a significant technical effort. Also beyond the scope of this task.

4 Building a FAIR repository

4.1 Choosing a platform

We've localized several lists of comparison of various data repository software solutions:

<https://unesdoc.unesco.org/ark:/48223/pf0000227115>

<https://zenodo.org/record/263823>

<https://hal.archives-ouvertes.fr/hal-00746713/document>

<https://open.library.ubc.ca/soa/cIRcle/collections/graduateresearch/42591/items/1.0075768>

<http://www.rsp.ac.uk/start/software-survey/results-2010/>

While all the software products listed in these comparisons are excellent, the main challenge for most libraries and even for a technical university group, is probably finding and keeping the skilled labor required to establish, operate and maintain a service based on any of them. This is discussed in a blog post from 2020 by two Dutch library professionals:

<https://openworking.wordpress.com/2020/08/18/why-figshare-choosing-a-new-technical-infrastructure-for-4tu-researchdata/>

To quote from the post:

“After over 10 years of using [Fedora](#), an open source repository system, to run 4TU.ResearchData, we have made a decision to migrate a significant part of our technical infrastructure to a commercial solution offered by [figshare](#).”

In the end, we decided to deploy a clone of the Zenodo service. A variety of considerations were behind this decision:

- We've observed that Zenodo generally enjoys goodwill in both research library and scientific communities
- A fraction of Danish researchers were already familiar with the Zenodo UI
- The fraction was small, but to our knowledge unmatched by other any other comparable service
- The possibility for a scalable deployment as a collection of prebuilt Docker images (core services + support services) was unique
- This was a good match to the general plan to roll out a Kubernetes service

More information behind this decision can be found in the DTU [bid](#) on the 2021 call for tender referenced above. In summary DTU intended to both be able to provide a polished user interface and bypass some of the above-mentioned deployment and operational challenges by directly using the container images provided by the Zenodo team.

Despite the images being prebuilt, we foresaw needing to customize:

- Layout/theming
- Description/help texts modified/replaced
- The same base storage (ZFS) leveraged as for the DTU EFSS service (sciencedata.dk)

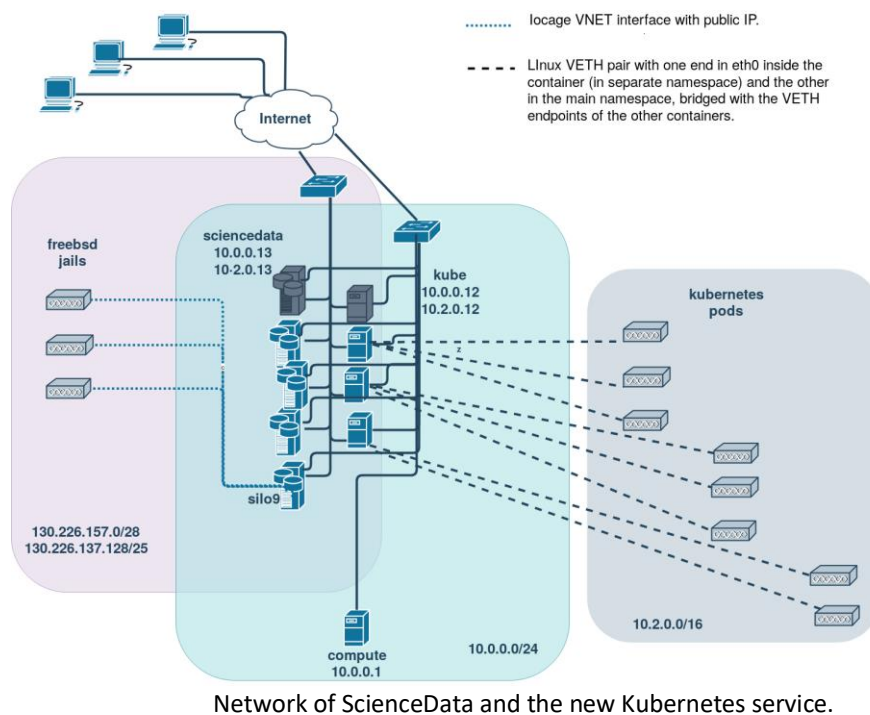
- eduGAIN login

We rebuilt the three front-end service images and kept the 6 support service images unmodified.

A related, but independent account of the challenges behind establishing a Zenodo-like, Invenio-based service can be found in the paper, "[Stripping down Zenodo to build an Invenio repository – lessons learned](#)", from the Center for Sustainable Research Data Management, University of Hamburg.

4.2 Establishing a hosting platform

Since the plan was to leverage Docker images provided by Zenodo, step number one was to establish a platform to run them on. As already mentioned, we chose Kubernetes with the cri-o runtime. Persistent block storage is provided via NFS-v4.1 from the ScienceData backend. The physical worker nodes and the pods of the Kubernetes control plan live on internal ScienceData networks, allowing the processing of sensitive data, stored without public access on ScienceData. Future plans include similar workflows for the new repository.



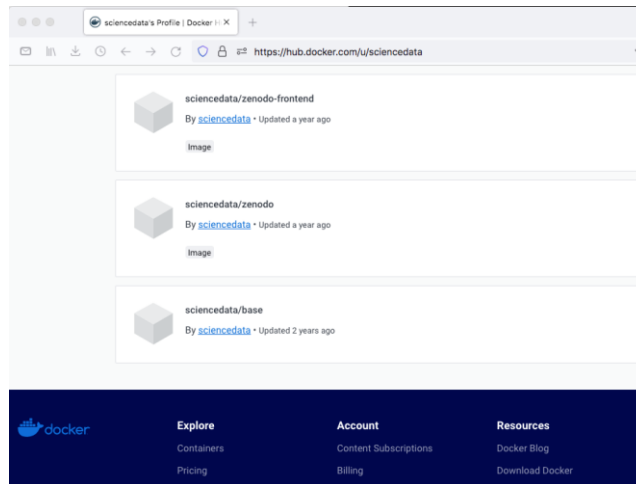
4.3 Installing and running Zenodo

We followed the official installation guide at:

<https://github.com/zenodo/zenodo/blob/master/INSTALL.rst>

With adaptations to match the DTU Kubernetes setup. In particular, we built the 3 front-end images with

these adaptations, and deposited them on Docker Hub: <https://hub.docker.com/u/sciencedata>



Moreover, we created a manifest (yaml) file for running these images, plus another manifest for running the 6 support services.

- Redis cache
- Celery task queue
- Rabbit message broker
- Elastic search
- Kibana
- Postresql database

The support images are unmodified images - also loaded off Docker Hub.

Our manifest files are kept in a public GitHub repository:

https://github.com/deic-dk/zenodo_deployment

A fork of CERN's Zenodo Docker build recipe, with ScienceData customizations, is also kept in a public GitHub repository:

<https://github.com/deic-dk/zenodo>

One of the minor adaptations was to enable viewing Jupyter notebooks (which for some reason was disabled on CERN's Zenodo service and in the unmodified images).

5 EFSS expunge functionality

As discussed [above](#), expunging research datasets from EFSS services to FAIR repositories is a main priority for T4.7. In this section we'll look at how we've implemented it on two platforms.

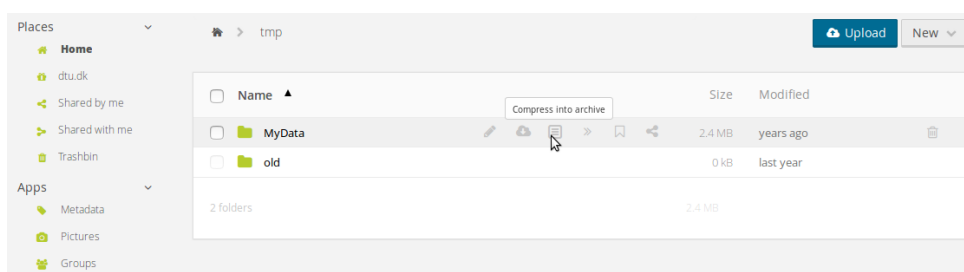
5.1 ownCloud

The DTU EFSS service, sciencedata.dk, runs a customized version of ownCloud. For this service, we've implemented an app for depositing files to zenodo.org and other repositories supporting the Zenodo APIs – including the own newly deployed repository. The work is based on an existing prototype; the code for is kept in a public [GitHub repository](#).

- **Functionality walk-through**

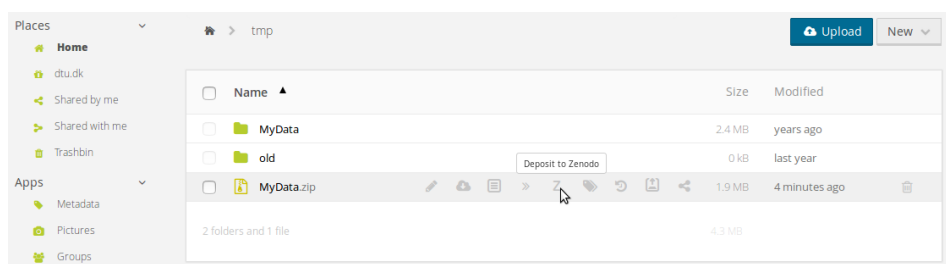
Creating a Zip archive

You can upload as many files as you wish as a Zenodo dataset deposition, but depositing all your data files as one Zip archive is typically more convenient. To create a Zip archive, simply hover a folder and click "Compress into archive":

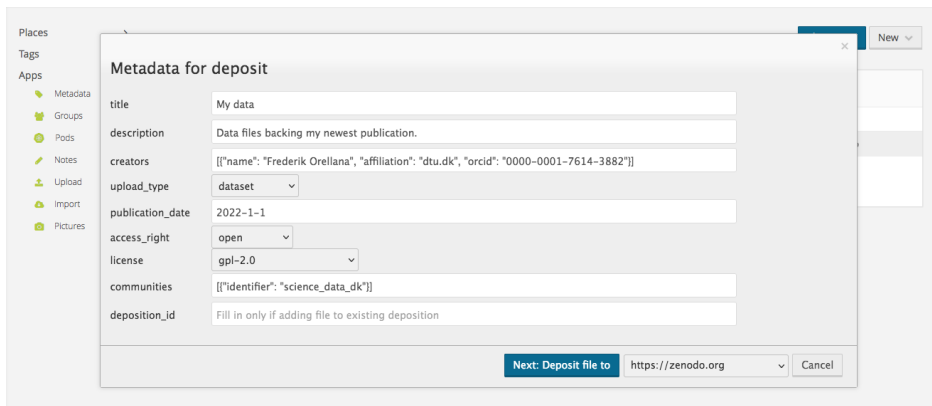


Creating a deposit

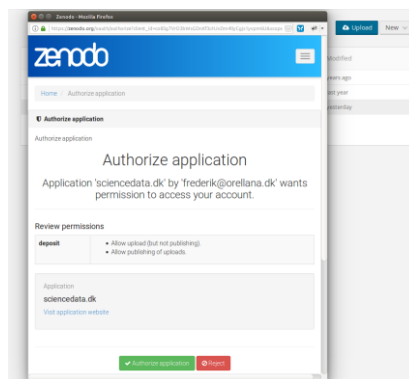
When you hover over the zip archive you've just created, you can click on 'Publish':



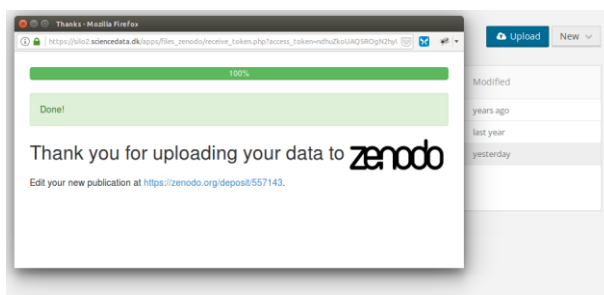
This will tag the file with a Zenodo tag and a metadata form will pop up:



What you type in is stored, allowing you to cancel the process and resume later without re-entering the metadata. You can always delete the tag and its associated metadata if your wish. Once you've entered the metadata you must choose the repository to which you wish to deposit - currently "scicerepository.dk" or "zenodo.org", then click "Next: Deposit file to". You'll be asked by the repository in question to allow access:



The file is then uploaded, but not actually published until you choose to do so at the repository in question (scicerepository.dk or zenodo.org):



At the repository, you can also fill in more extensive metadata:

zenodo Search Upload Communities frederik@orellana.dk

Delete Save Publish

New upload

Instructions: (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press 'Save' to save your upload for editing later. (iii) When ready, press 'Publish' to finalize and make your upload public.

Filename (1 files)	Size	Progress	Delete
MyData.zip md5:73b28c6627882edff5469b8017eeb035	1.9 Mb	✓	🗑️

Note: File addition, removal or modification are not allowed after you have published your upload. This is because a Digital Object Identifier (DOI) is registered with DataCite for each upload.
(minimum 1 file required, max 50 GB per dataset - [contact us](#) for larger datasets)

Upload type required

Basic information required

Digital Object Identifier e.g. 10.1234/foo.bar
Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.

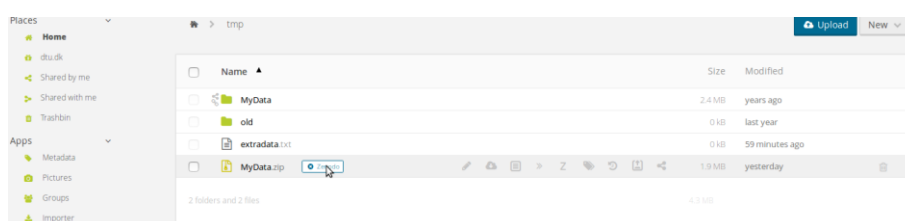
Pre-reserve DOI

Publication date * 2017-03-23
Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

Title * My Data.

Adding files

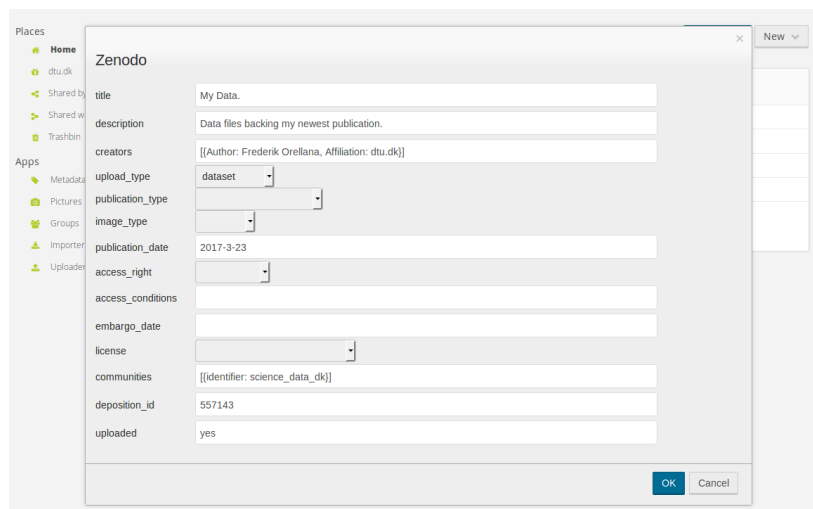
Should you wish to add additional files to your deposit:



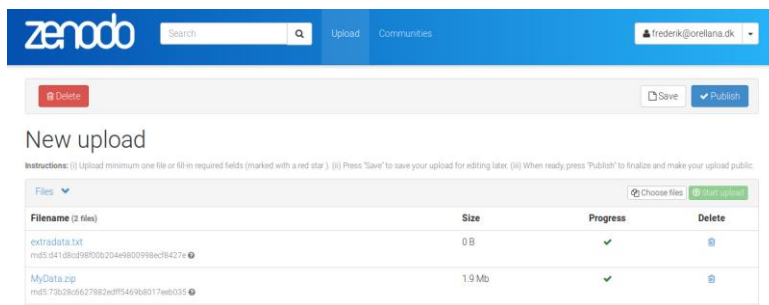
you can inspect the assigned deposition ID, `deposition_id`, by clicking on the Zenodo tag icon:



and then deposit further files to the same deposition by entering this deposition ID in the metadata popup:



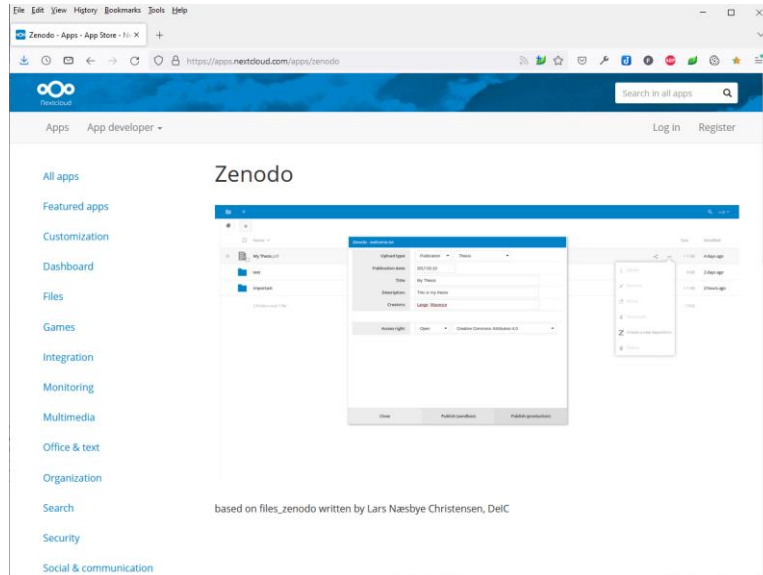
The file will simply be added to the uploaded files in your deposit:



5.2 Nextcloud

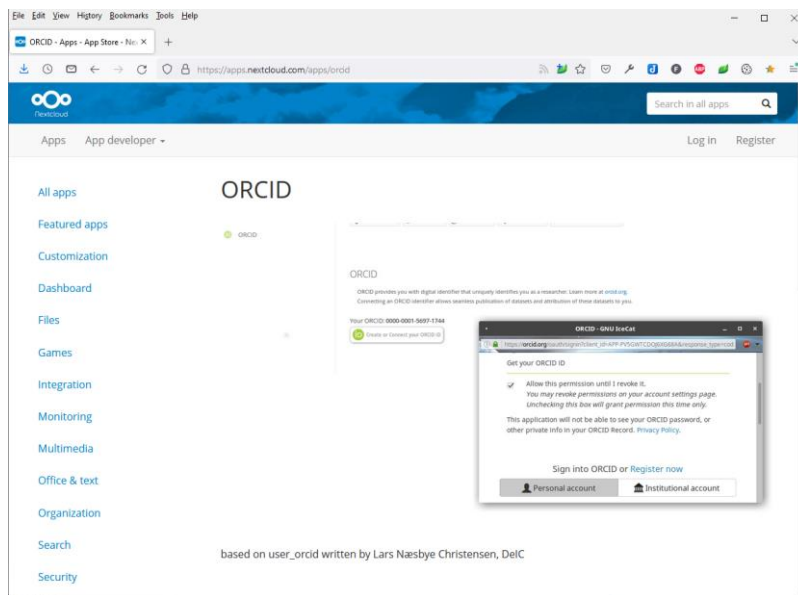
The DTU EFSS service is a customized version of ownCloud – the customizations include support for Nextcloud sync clients. In general, Nextcloud is an important player due to its prominence in the

comparatively small market segment of research IT. It is also on a possible future scenario for DTU to migrate fully to Nextcloud. One step in this direction would be to have the DTU customizations supported on Nextcloud. Therefore, after finishing a first prototype ownCloud Zenodo app, we (DTU) also commissioned a functionally equivalent Nextcloud app with Nextcloud GmbH. The app is available in the [Nextcloud app store](#):



6 ORCID integration

[ORCID](#) is an example of an open service, developed and operated by an international non-profit organization – and in widespread use in academia. It provides researchers with unique electronic ID to unequivocally distinguish each individual on research papers, grants etc. For expunging from EFSS systems to a service like Zenodo, the service is indispensable, since it allows carrying a unique ID from one system to another. Zenodo already supports ORCID. Supporting it on the DTU EFSS system and on Nextcloud was achieved via an app, in the same way as described above for the Zenodo app. Again, our code is available in a public [GitHub repository](#) and the Nextcloud app, commissioned from Nextcloud GmbH, is available in the [Nextcloud app store](#):



7 Discussion

The purpose of T4.7 is to provide feedback for WP4 and more generally on challenges and opportunities for our project in the European FAIR landscape - and by extension for a standards-based and open approach to building FAIR data services for European research. The method has been: 1) Gather infrastructure needs as expressed by researchers and as perceived by other stakeholders, 2) explore tangible FAIR service deployment. This and the discussion below supplement the data collection and analysis of [deliverable 4.3](#).

We stress that the considerations below, specifically relate to *generic large-scale infrastructure services*. Moreover, the described challenges are not equally pertinent across our consortium or across the broader CS3 community – with DTU perhaps situated at one extreme and CERN at the other.

One general takeaway is that we believe the observed trend of “enterprization” of university IT departments is a general one - and one that can complicate the deployment of open-source-, GDPR- and open-data-friendly solutions. This leaves a project like ours, and T4.2 in particular, with both opportunities and challenges:

IT entreprization

- University IT departments and libraries often don't have the impetus, staff or budget to do actual coding and software development
- The same can be said of on-premise service operation: In general there is a move to commercial clouds
- Generic data services (like OneDrive, figshare,...) offered to researchers, are generally licensed from commercial cloud vendors
- These are by mandate (by IT depts., administration) accompanied by service contracts, support offerings etc.

The heroic effort

- The individuals engaged in EU-funded infrastructure projects are often a small group of open-source advocates
- These may be challenged in providing service contracts etc.
- Deployment and hosting may end up requiring considerable effort from the same few persons

Research relevance

While open source, GDPR-friendly and open access technology may not be a perfect match with enterprise IT, such technology and practice *are* requested by researchers. This possible gap between what is offered by enterprise IT and what is requested, presents a project like ours with an opportunity.

For most of the consortium members, we believe this is the mechanism behind the successful base EFSS services. In theory, this should present a leverage for onboarding existing users to new functionality and new services. In practice, we have seen this still needs work.

Enter CS3

We believe that one prospect of the present project is to help participating organizations and institutions bypass the above predicament – by taking advantage of the fortunate situation described in our proposal: that a handful of IT groups in the research sector, scattered across Europe, have consolidated on software from the same open-source vendor and, combined, serve a number of users hitherto unseen in our sector.

Indeed, a strength of a community like ours is that together we can team up with vendors and open-source projects in long-term efforts on building credible alternatives to overseas cloud services: European data service alternatives, based on openness and standards. If successful, the European research and education IT sector – and by extension the European IT sector, stand to benefit.

8 Recommendations

One overall aim of our project is to get researchers to use our *new* functionalities and services. This can happen either because their administration, IT department or library mandates or encourages them to do so, or because they find our services useful, compared to alternatives. Achieving the former is generally contingent on processes preceding project start as described in section [3.1](#), but a positive difference can be made at any point.

Of the various stakeholders, we see university administrations, university libraries and IT departments as in many ways the most crucial. Although their persuasion is in principle a local matter for each consortium member, we see a potential for a project like ours to provide direct support for each member in interacting with their local administration. This could positively accompany more high-level political efforts, like EOSC symposia, workshops etc.

Although such political efforts contribute positively towards FAIR, openness and European self-reliance, we are not convinced they are enough to counter the tides of enterprization and commercial mega-clouds. Although, in this light, our project is certainly a small player, we believe it can make a difference. Concretely we recommend:

Incorporation

One ingredient of enterprise IT are license or service contracts agreed between two legal entities: The IT department (formally, the University) and the software or service vendor. For a new product to be deployed by an IT department, there needs to exist a legal entity in the other end.

Another aspect is that what we are building is *public infrastructure*: Services which may change, but which will be around for decades or more, which we expect researchers to rely on, keep their life's work on, use with complete trust and conviction that data is not used for dubious purposes.

In short, we recommend creating a non-profit foundation – or start e.g., as a programme under [The Commons Conservatory](#).

Product

This non-profit needs an asset – a *product* that can provide the content of license and service agreements. One possibility is to offer support and/or responsibility in setting up EFSS and related services.

To be credible as an enabler of *permanent* infrastructure services, however, we recommend shedding reliance on third-party, commercial vendors. This would be a potentially risk-prone move that should be undertaken only if an initial critical mass of skills can be amassed. If that is the case, we recommend forking one of the two leading European EFSS products and create an EFSS distribution, targeted at research and education – the direction and development of which would be governed by the above non-profit legal entity.

Operations

Besides offering support to IT departments, we envisage the non-profit organizing training workshops for IT personnel.

We moreover envisage partnerships with selected European cloud vendors, coupled with data privacy and easy migration guarantees.

Use cases, use cases, use cases

To lower the engagement threshold for researchers, we need tangible examples of collaboration, data processing, analysis and final deposition workflows. This could e.g., be copy-paste notebooks they can download from our website, execute on their own EFSS, and publish directly from EFSS to Zenodo.

What we are up against are commercial services with a free tier, like Binder, Azure Notebooks, Google Colaboratory and [many others](#). These services offer features like collaborative editing of Markdown, LaTeX and Jupyter notebooks, integration with GitHub and easy sharing and publishing. Commercial publishing services like figshare offer integration with Binder, ORCID, Overleaf (online LaTeX editor), RSpace (online lab notebook), GitHub etc. Googling e.g. Google colab returns a multitude of videos and howtos, showing you how to get started. On colab.research.google.com, anyone can log in and with a few mouse-clicks run an example notebook in their browser, processing data off their own Google storage, and finish by publishing their notebook on figshare.

These services were developed for researchers and are used by researchers. From our interactions, it seems,

not a large fraction - but very much the same fraction which our project is targeting. Luring researchers from the above platforms requires large and dedicated outreach efforts to and engagement with research environments. In T4.1 the data analysis part is already in progress. It needs to be complemented by the FAIR publishing part.