Bilag til Notat om

Datamanagement i Danmark

Opdateret

Bilag 2 opdateret 18. september 2019

Udarbejdet af et udvalg bestående af:

AAU Prodekan Torben Larsen
AU Projektleder Birte Christensen-Dalsgaard (formand)
CBS Chefkonsulent Lars Nondal
DTU Sektionsleder Michael Rasmussen
KU Professor Kasper Møller Hansen
RUC Vicedirektør for Digitalisering Galina Ianchina
SDU Institutleder Peder Thusgaard Ruhoff

samt

DeiC COO for datamanagement Anders Sparre Conrad (sekretariatsbetjening)

Gruppen har undervejs fået assistance fra:
Ulrik Sørensen Rohde, KU

Inholdsfortegnelse

BILAG 1: KOMMISSORIUM FOR ARBEJDSGRUPPE VEDR. FORSLAG TIL FREMTIDIGT NATIONALT STORAGE OG DATAMANAGEMENT LANDSKAB 3

BILAG 2: CONCLUSIONS ON THE NEEDS FOR A NEW NATIONAL DATA MANAGEMENT INFRASTRUCTURE 5

BILAG 3: OPDATERING OG UDVIDELSE AF LANDEANALYSE FOR HOLLAND 15

1. AKTØRER 16
2. GOVERNANCE 16
3. SERVICES 17
4. DATALIVSCYKLUS 19
5. PUBLICERING 21
6. DATA STEWARDSHIP 21
6. OPSUMMERING 21

APPENDIKS A 23

REFERENCER 26

BILAG 4: REVIEW AF TRE EKSISTERENDE STORAGE LØSNINGER 28

ERDA/SIF 28

ERDA 28
SIF 28

CLAAUDIA CLOUD 29

CLOUD SERVICES 29
TEKNISKE VALG 30
SYSTEM EGENSKABER 30

SDUCLOUD 31

SOLUTIONS & SERVICES 32
TEKNOLOGI 33

DeiC 8. maj 2019

Bilag 1: Kommissorium for arbejdsgruppe vedr. forslag til fremtidigt nationalt Storage og Datamanagement landskab

Indledning

DeiCs bestyrelse har besluttet at nedsætte en arbejdsgruppe bestående af repræsentanter fra universiteterne til at komme med forslag til sammensætningen af et fremtidigt nationalt storage og datamanagement landskab til gavn for forskere i Danmark.

Formål

Udgangspunktet for arbejdet er visionen for digital forskningsinfrastruktur, som den fremgår af ”Strategi for Nationalt Samarbejde om Digital Forskningsinfrastruktur”:

Forskere ved de danske universiteter skal have adgang til en digital infrastruktur, der muliggør forskning og uddannelse i verdensklasse

De grundlæggende principper bag visionen er

- Forskere inden for alle hovedområder skal have adgang til digital infrastruktur på transparente vilkår
- Der skal være en klar og tydelig rolle- og arbejdsfordeling mellem aktiviteter på det nationale niveau og på institutionsniveau
- Investeringer i kostbar digital infrastruktur til forskning skal koordineres og udnyttes og drives effektivt
- Der skal etableres et langtidsholdbart og fleksibelt samarbejde med en stærk international forankring
- Der skal være stabilitet omkring økonomi for at sikre kontinuerlig modernisering af teknologi og løbende kompetenceudvikling

De danske universiteters rektorkollegie behandlede på deres møde den 21 maj DeiCs bestyrelses indstilling DeiCs bestyrelses indstilling vedr. national HPC, storage og datamanagement – proces, styring og økonomi. Principperne i denne indstilling er rammerne for det fremtidige nationale storage og datamanagement landskab. Indstillingen vedlægges som bilag.

”Storage” defineres som et system, der anvendes til opbevaring af data, og ”datamanagement” defineres om de lag, der ligger ovenpå som services, politikker og lignende

Som led i fastlæggelsen af storage og datamanagement landskabet gennemføres et analyse- og udredningsarbejde, der vil danne baggrund for beslutninger omkring og tidsplan for implementeringen.

Arbejdsgruppen

DeiCs bestyrelse ønsker at nedsætte en arbejdsgruppe med faglige og tekniske repræsentanter for storage og datamanagement området

Gruppen vil bestå af en repræsentant fra hvert af de otte universiteter. Gruppen skal konstituere sig ved valg af formand blandt gruppens medlemmer. Gruppen vil blive sekretariatsbetjent af DeiC.

Den udnævnte formand refererer til bestyrelsen.

Medlemmerne bør have teknisk indsigt på området.

Det er op til det enkelte medlem og universitet at sikre inddragelse af øvrige interessenter internt. Udgangspunktet for det enkelte medlem og universitet kan eventuelt tages i den vision for det nationale storage og datamanagement landskab, som bestyrelsesmedlemmerne fremlagde på DeiC bestyrelsesmøde den 26. april 2019.

Arbejdsform og interessentinddragelse

Arbejdsgruppen skal sikre høring og om ønsket inddragelse af

- DM Forum
- DM LedelsesCAB
- CIO gruppen
- Bevaringsinstitutionerne
- RA Sundhed

Opgaven

Arbejdsgruppen skal senest 1. september 2019 levere en rapport, der belyser følgende områder:

1. Hvilken storage infrastruktur skal være tilgængeligt på nationalt niveau for at understøtte FAIR data og open science, og dermed øget tværgående samarbejde?
2. Hvordan sikres sammenhæng mellem lokale, nationale og internationale løsninger, herunder aktiv dansk deltagelse i EOSC-initiativer.
3. Hvilke trin i datas livscyklus skal dækkes af nationale services? Herunder tanker om understøttende support og kompetenceudvikling
4. Hvilke udviklingsressourcer og hvilken organisering er der behov for, for at sikre en bæredygtig national udvikling i forhold til den internationale
5. Anbefalinger omkring det nationale samarbejde i DM Forum, herunder om det skal fortsætte og i givet fald med hvilke opgaver og budget.
6. Hvordan håndteres deltagelse i internationale medlemsskaber og samarbejde, og med hvilket engagement og investering?
7. Hvordan sikres national videndeling omkring internationale resultater og aktiviteter, med henblik på anvendelse såvel teknisk som forskningsmæssigt

Resultat

Der forventes en rapport, der behandler ovenstående emner, indeholdende anbefalinger for området for perioden 2020-2025.

Tidsplan

Drøftelser og analyser gennemføres af arbejdsgruppen i perioden 1. juni 2019 – 15. september 2019. Rapporten skal leveres til DeiCs bestyrelse senest den 15. september 2019. Rapporten sendes til sekretariat@deic.dk

Bestyrelsen vil drøfte rapporten på et ekstraordinært bestyrelsesmøde i slutningen af september, hvorefter den sendes i høring hos universiteterne.
Endelig beslutning og tidsplan for implementering baseret på rapporten vil foreligge 1. november 2019.

Bilag 2: Conclusions on the needs for a new national data management infrastructure.

The University of Copenhagen – August 2019

Interviewene er indsamlet af Susanne den Boer, Københavns Universitet

Six (Associate) Professors and two research support staff at 4 Faculties across UCPH were asked for their views on the needs for new national data management infrastructure. The text below summarizes their feedback given in response to a short survey and/or during a separate discussion.

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 <u>new</u> 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 analysing 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 <u>existing</u> 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.

Results of the questionnaire on the needs for national data management infrastructure.

A questionnaire was sent to 11 persons across UCPH on 16 August 2019. By 27 August, it had been answered by 8 persons, either in writing, during a short phone conversation, or in a meeting.

The 8 respondents can be characterised as follows:

- *6 VIP and 2 TAP*
- *VIP: 3 Professors, 1 Professor mso, 2 Associate Professors*
- *TAP: 1 Engineer and 1 Research IT coordinator*
- *1 person from the Faculty of Humanities, 3 persons from the Faculty of Science, 1 person from the Faculty of Health and Medical Sciences and 3 persons from the Faculty of Social Sciences.*
- *Two of the respondents (one at HUM and one at SCIENCE) answer on behalf of a larger group of researchers at their Faculties, while the others relate to a single research project or to their own typical way of managing data.*

All feedback is shown below, with the original questions in blue.

1. WHAT DATA DO YOU WORK WITH?

1.1 Do your datasets contain personal or confidential information?

HUM Humanities uses datasets of a very broad diversity:

- Digitized images
- Scanned and OCR'ed texts
- Digital texts
- Audio and video recordings
- Other datasets containing language data, e.g. transcriptions of language data, dictionaries, keyword collections; a lot of different datasets exists, some as tabular data, but in a diverse set of formats
- Survey data
- Log books and lab looks with observed data
- Databases
- Tabular data
- GIS-data
- Datasets based on binary data: Vector data, raster data, SVG data files
- Dataset that describes various types of models

SAMF Yes: Personal data in the form of photographs of persons. Confidential data, as these photographs are owned by professional photographers, who loan their photographs to the research project.

Yes: Survey data.

Yes: CPR numbers and information about children and their parents.

SUND No.

SCIENCE No.

No personal data, but sometimes confidential information, due to collaborations with companies.

1.2 What is the origin of your data?

HUM Data is collected in a long list of ways. A few examples of sources:

- Image data from an excavation
- Physical objects collected, scans of physical objects
- Manuscripts on paper or digital
- Audio and video recordings performed by the research group
- Interview data collected by the research group or collaborators (often sensitive data)
- Measurements done on persons e.g. eye movements and brain activity
- Closed databases, with license fees or only available via negotiation
- Open datasets and databases
- Data collected from publishers and companies that want to contribute to research but will not give away their data to others than the researcher, who they have an agreement with

SAMF Digital photographs taken by professional photographers.
Coding obtained from photos, to be used in algorithms.

The project reuses existing survey data, obtained from Statistics Denmark, the National Archives or from international surveys.

Personal data come from a large cohort study, from intervention studies and from surveys carried out across Denmark.

SUND The data are obtained from human and animal tissue samples and biopsies, which are stripped of all personally identifiable information.

SCIENCE Agricultural research in plants, plant breeding, plant production. Also other field relevant data such as soil maps and climate data.

We generate our own data in the lab.

1.3 What are the file formats you typically work with?

HUM A list of file formats cannot be listed here. From the above list of diverse examples of data sets and sources, it is clear that researchers in the humanities handles a very long list of data formats.

SAMF Photo formats that carry a lot of information.

R data formats.

Quantitative and qualitative data, in Stata, SPSS and R formats.

SUND Excel, R and special data formats generated by laboratory equipment.

SCIENCE

- Image formats both proprietary and open. From a few kB to several GB per mosaic. RGB, thermal, multi- and hyperspectral.
- Databases: SQL (PostGIS, MS, other), GeoPackage (SQLite for GIS)
- R, Python, C#, MatLab, GIS formats, bioinformatics formats.
- CSV, other clear text
- Most of the Microsoft products
- Homemade in- and output files in cleartext.
- Proprietary data formats for sensing hardware (e.g. climate data)

Laboratory equipment output files, such as spectrometer output files.

Almost all is in NetCDF format.

1.4 What approximate volume of data do you on average produce in a single project?

HUM Usually the data sets are smaller than the large data sets found at e.g. Faculty of Science. Volume is in some cases an issue for excavations, models trained, images, video and scans.
A few examples are a specific sound database of 1.2 TB, a collection of scanned documents of 7.5 TB, and a research groups collections of videos and other data of 18 TB. Other data set take up less than 1 MB.

SAMF Small to medium sized data sets.

Small to medium sized data sets. Only rarely big data sets.

Data sets ranging from records on 600 persons to 56.000 persons.

SUND Data set range in size from a few kb's to very large data sets containing sequencing and imaging data.

SCIENCE

- Most projects generate a few GB to some hundred GB.
- The image intensive projects are from several thousand images for a trial of a few weeks (circa 1 Terabyte), to projects spanning several years producing many terabytes of data.

Datasets of approximately 100GB on average, with which big calculations and simulations are carried out.

Typical projects generate a 1 TB. But we also have had multi-year projects bringing in 1 TB/day from a PRACE resource.

2. DO YOU TYPICALLY MAKE YOUR DATA AVAILABLE TO OTHERS AFTER PROJECT END?

HUM Many of the datasets that are used cannot be shared outside the specific research team, neither during the project nor after the project end.
For the majority of datasets the agreements and the rights are not allowing for re-sharing of data or sharing of the derived data. For some research groups and projects, data sharing is common practice, and more research groups are becoming aware of the benefits of open data, but Humanities will also in the future work with and perform research on sensitive and restricted data of many kinds. It is not typical to make data sets available.

SAMF Photos will not be made available as they contain personal information (images of humans), and are property of the professional photographers. Coding will be made openly available.

This will depend on where the data come from. Some data providers impose restrictions on data sharing. If those restrictions do not exist, data will typically be shared with peers.

Whether the data will be shared and how the data will be shared (Open Access, upon request etc.) depends on:

- Whether data can be anonymised.
- Whether the data provider allows it.
- Whether the persons involved have provided informed consent to do so.

SUND Data are shared openly when this is a requirement by the funder or publisher. Otherwise, data are typically made available upon request.

SCIENCE No. Not very often. Not yet. But, several of us are in the process of investigating sound ways of doing this for our phenotyping data, and we are currently looking into adopting the open-source Phenotyping Hybrid Information System (PHIS) created in France.

We are starting to do this more and more, either in connection with publication or a MSc/ PhD thesis. We aim to share the data openly, and ensure a DOI is attached to the data set.

3. WHAT TECHNICAL SOLUTIONS DO YOU CURRENTLY USE WHEN MANAGING RESEARCH DATA? INDICATE WHETHER THESE SOLUTIONS ARE OFFERED LOCALLY, NATIONALLY OR INTERNATIONALLY.

HUM Usually the research project frame their data collection and data processing to fit very specific needs in the current project. Some projects can handle their data with a small set of commonly used tools. Some projects depend on specialised software developed for exactly this type of research task.

Making datasets FAIR is a new thing for researchers, but some groups have for years worked with open data, used standards, documented and shared the data, which fulfils the most of the FAIR requirements. Others can only work on making the metadata FAIR, as data are restricted.

Researchers sharing data often use community specific repositories.
If the data cannot be shared in the community specific repositories like CLARIN, researchers are asking for a place to preserve data for 5 years, and in some cases also to preserve them longer. Repositories to do this for sensitive data and restricted data are currently not available at UCPH, but network drives H, N and S are used for these tasks.

SAMF Hard drives are used to transfer personal data from between international collaborators.
Coding will be shared with peers after project end. To do so, data will be deposited in data repositories recommended by the publisher of the associated research article, or by the funder.

Data are collected from Statistics Denmark, the National Archives and international survey databases in, among others, the UK and US. R is used to process the data.
Unless restrictions exist, the data are stored on a UCPH personal drive, or the secure (S) drive, which has the necessary security features to safeguard personal/confidential data.
To collaborate with others on the same research projects, data will be shared via Dropbox (only for anonymous data). There is no need for a sharing solution for personal data, as those data remain at Statistics Denmark etc., where others can find them too.
To share anonymous data with peers after project end, data repositories will be used that are recommended by the publisher of the associated research article, if the terms of reuse posed by the original data provider allow this. UCPHs GitHub is used for sharing with colleagues at UCPH.
The National Archives are frequently used to preserve data.

Secure FTP servers were used to retrieve the large data set collected by the municipality.
Data are processed in Stata, SPSS and R, and stored on UCPHs secure S drives.
If data can be shared with peers after project end, they will typically be shared through the National Archives or a repository recommended by the publisher.
To preserve data long term, The National Archives, Statistics Denmark, or one of the other cohort data bases are used.

SUND Computerome is used to process and store the RNAseq data.
To share data with internal collaborators, a university drive is used. The BlueWhale email plug-in is used to securely share bigger data sets with external collaborators. Small data sets are sent via regular university email.

To share and archive finalised datasets with peers, the UCPH/SCIENCE solution 'DATA DOI' is used, which is based on the local repository ERDA. Alternatively, a repository that is recommended by the journal is used (e.g. Figshare). But there is a strong preference for using the UCPH solution.

SCIENCE

To collect data created by others: UCPH servers, SCIENCE servers, ERDA, UCPH OneDrive, Email, Cloud solutions like DropBox and what else the project partners have in their institution/company, Physically transferring in person or by postal services sending a USB memory or an external hard disk.
To process data: Local instalments of commercial software (e.g. MatLab, Excel, VideometerLab, ArcGIS, Photoshop and the likes. Pix4D, Metashape), free software (e.g. R, Python, OpenDroneMap, GIMP, QGIS), homemade software (e.g. ThistleTool, PlotCut, Root-image analysis), Cloud-based solutions (e.g. Pix4D, Solvi (e.g. plant counting), misc. online bioinformatics tools).

To collaborate with others on an active data set: sending it back and forth by email, Having it on an UCPH/SCIENCE server, UCPH OneDrive, or Google Docs.

To share finalised datasets with peers: Not too much currently. A few times on GitHub for R packages.

To preserve data long term: UCPH SCIENCE's I-drives and SCIENCE's ERDA solution.

To share active data sets with external collaborators: Dropbox, email.
To share finalised data sets with peers: SCIENCE's ERDA / DATA DOI
To preserve data sets long term: ERDA

I use SCIENCE's ERDA for the storing and sharing of my research data with (external collaborators). I currently have 2 Pbyte on ERDA.

4. IS THERE A NEED FOR A TECHNICAL SOLUTION TO WORK WITH YOUR DATA THAT CURRENTLY IS NOT AVAILABLE TO YOU?

HUM

For data that cannot be shared in community specific repositories like CLARIN, or general repositories like Zenodo, researchers are asking for a place to preserve both sensitive and restricted data for at least 5 years.
Agreements about use of some data sets are complicated and need institutionalised solutions, because bringing yet another party/institution on board brings further complications into the handling of data.
Working on sensitive data is currently very restricted, and a solution to easily get a dedicated server or computer to work on sensitive data on the researchers own institution or at other institutions providing the data is needed.

SAMF

A solution to securely share sensitive/confidential data with collaborators. There currently are no solutions locally, nationally, internationally that can demonstrate to have the necessary security measures. This is why data exchange occurs in person via portable hard disk.

- There is a need for solutions to facilitate collaborating on active data sets. Dropbox is currently used, because there is nothing with the same ease of use on offer locally, nationally or internationally.
- There is also a need for project management / collaboration software, including a 'chatroom' for commenting and information exchange between collaborators.
- A secure repository for the long term preservation of personal data.

- There is a need for a solution to securely send and retrieve big data sets that contain personal information.

- There is also a need for an application that allows building and managing online surveys and databases (containing personal data), such as REDCap. There is no national instalment of REDCap, and no university wide instalment at UCPH yet at this point.

SUND

- A Dropbox equivalent to securely share active data sets with collaborators.
- A service that would make datasets findable and accessible, e.g. optimizing UCPHs existing solution DATA DOI by adding more metadata fields. There should also be more support for these services.

SCIENCE

- Data analysis on more powerful machines, including the necessary software instalments.
- For the meta part of the data acquisition and analysis we should have an UCPH solution for project management. Especially needed when we are many people, cross-department, cross-faculty even partners outside of UCPH. We need something like Microsoft Projects, with both UCPH and WAYF login (the latter to enable colleagues) and guest invitation possibility too.

We need a better Dropbox equivalent for the sharing of active data, that is free of charge to use and allows for a larger volume of data to be stored (than the current storage allowed for free in Dropbox).

5. WOULD YOU CONSIDER USING NATIONAL INFRASTRUCTURE TO MANAGE YOUR DATA, BOTH FOR THE TECHNICAL SOLUTIONS MENTIONED IN QUESTION 4, OR TO REPLACE THE SOLUTIONS YOU CURRENTLY USE AND MENTION UNDER QUESTION 3? IF SO, WHAT INFRASTRUCTURE?

HUM

Currently, national infrastructure storage solutions do not seem to be a usable data storage solution for data that cannot be shared, and for data with complicated rights restrictions.

A national safe and legally accepted management tool for non-sharable data for processing would be appreciated, if it were easy to use.

Computer calculation hours for analysing and modelling data are much needed, but the transfer of data to these server systems has to be easy both technical and legally.

SAMF

A national solution to safeguard / exchange sensitive data will only be considered if the security features are extremely convincing and on par with those at Statistics Denmark. Right now, there is no national solution that convinces, that can be trusted in this regard.

An interface that connects various already existing local, national and international solutions to retrieve and manage data is lacking. An example will be an interface that connects all the survey data bases in Denmark that could facilitate pulling data from multiple sources. Right now retrieving data on similar studies from multiple sources that contain CPR numbers (Statistics Denmark, National Archives) is time-consuming. Alignment between these sources through a 'CPR-link' would greatly facilitate registry-based research.

Rather than investing in a new national infrastructure, invest in improving existing infrastructures in Denmark and ensure that they are better connected to / aligned with each other, and to solutions locally at the Universities and internationally. There are already so many solutions out there. We do not need yet another similar solution. We need better alignment, better services, and better information on what is available.

A national solution would be considered, if a similar solution was not present locally. However, besides an instalment of REDCap accessible to us, there are currently sufficient data management solutions present locally and nationally to carry out our research. Some of these solutions should be improved, and this is where the effort should go.

SUND How useful is a “national data management infrastructure layer”, when you can access similar solutions at European level? Why add another layer, if the EU solutions are secure enough according to Danish legislation / standards?

SCIENCE For some tasks yes, most certainly.

We are happy to use other (national) solutions. It does not really matter which solution it is, as long as it works and does not require a lot of effort to use.
However, I prefer to use one solution that can do everything, instead of 10 different solutions for every element of the research data lifecycle.
Also, make existing solutions conform to a common standard. There are too many different standards for how to work with data, for example different journals demand different data uploads. Standardization across already existing solutions will make our lives easier.

I do not see the need to bring anything to a national level, ERDA provides a perfect service for all my needs. Perhaps ERDA can be expanded (inter)nationally? I can see that non-UCPH scientists would like to enjoy the benefits of ERDA, some have asked already.

6. COULD YOU SEE AN ADVANTAGE OF USING NATIONAL OR INTERNATIONAL INFRASTRUCTURE FOR DATA STORAGE / MANAGEMENT, OVER A LOCAL/INSTITUTIONAL SOLUTION. OR VICE VERSA? WHY?

HUM International research infrastructures and national infrastructure for data storage and data management are limited to sharable data or closed communities sharing data. Many data sets used at Humanities have complicated rights management, and that is best handle on the specific institution using the data.
Public data e.g. data from ministries and data from the Royal Library and the National Archives could be made available through national infrastructures.

SUND A national solution should be considered, if it can help the universities save money by reducing the need to make their own institutional solutions. The money saved should then be invested in local support in using these national and local solutions (or in research).

SCIENCE Yes, we could see an advantage of using a national infrastructure. Better annotation of data, better data sharing (especially long-term and for publication). Sure, in the beginning it will be more time-consuming, but I think the benefits can outweigh the extra effort.

7. WHAT WOULD BE YOUR INCENTIVE FOR USING A NEW NATIONAL DATA MANAGEMENT SOLUTION?

HUM The incentives will focus on stable and easy use, easy administration and low cost.
Usage of a new data management solution will depend on clear infrastructure commitment regarding continuous access to data and options to define restricted access to data. The solution should have a clear and accepted legal framework, access restrictions should be easy to specify, and it should be free or for a small cost.

SAMF If the solution is easy to use (easier than the current solutions) and has the necessary security features to handle personal data according to the GDPR.

Solutions that can be integrated with the Windows interface (file explorer) would be great. Most important is ease of use. Sharing solutions must be easy and intuitive to use, not just for us in Denmark, but also for our collaborators abroad. As long as solutions on offer are more time consuming or complicated to use than Dropbox, there will be no incentive to stop using Dropbox.

SCIENCE The solution needs to be reliable and accessible. The learning curve should be really quick. No need for a sexy solution. Reliability, ease of use and costs are the most important factors in deciding what system I will use. Click and go, no manuals, nothing too complicated. Complicated and expensive solutions will only ensure that I will go out and buy 10 hard disks and solve the problem myself.

8. WHAT SUPPORT, GUIDANCE AND KNOWLEDGE COULD YOU IMAGINE YOU NEED TO MAKE OPTIMAL USE OF NATIONAL DATA MANAGEMENT INFRASTRUCTURE? WHAT WOULD BE THE PREFERRED FORMAL FOR DATA MANAGEMENT SUPPORT?

HUM Data management support for specific data formats and data types would be preferred, combined with a general **legal support service** about rights to data and support to get data sharing agreements with data providers.

Knowledge bases of guidance and **experts that can advise** about the concrete circumstances in the specific project.

SAMF What is needed is a service that would provide an overview of the different data management solutions already available locally, nationally and internationally. This service should also include guidance on which solution is best for a given scenario, e.g. guidance in choosing the right cohort database for data preservation. There are already many solutions out there, but it is time-consuming and not always easy to find the most appropriate solution. Support in this would be welcomed.

SUND Regardless of the data management infrastructure that will be provided, support in using this infrastructure is crucial. This so that researchers can limit the time spent on tasks such as figuring out which solution to use and how it works. Support can be provided over email, or through FAQ etc. Ideal would be a local help desk, to get a quick answer to a data management related question. Whether that local support is located at the Faculty or Department depends on the needs. The closer to the research environment, the more likely it is that support persons can translate the researcher's needs into the best technical solution.

SCIENCE There should be phone/remote-screen-sharing/takeover user support from 0700-2000h at least the first year or two. Thereafter maybe just within business hours would be ok. The support is not just using the infrastructure but understanding the meta data concepts. Thus, it is not enough for the supporter to just be able to say, "have you re-started the software". They would need to have a deep knowledge of the platform and it intended usage/outcome. Please remember that email support only is not good enough. It can be very difficult for many people to sufficiently describe in writing their problem with a new technology. Phone and screen-sharing is essential for success.
-It should also be possible to book (maybe by small co-payments) a platform/infrastructure consultant to come out on the premises and help for a day or more with the data handling, the annotating, the uploads etc. needed to conform with the platforms design.

Some sort of Wiki website. However, in addition, people are essential, as not everything can be automated and described only on websites. I rather call a support person for 2 minutes, than cruise around on a website for 2 hours and fail to find what I was looking for. Websites are often written by people who are experts in the solution described on the website. This does not mean that others (non-experts) can understand this description.

9. DO YOU HAVE ANY OTHER COMMENTS OR SUGGESTIONS?

HUM Answers to some of the questions for the "DeiC arbejdsgruppe":

Sammenhæng mellem lokale, nationale og internationale løsninger:
In research projects e.g. EU-projects the handling of and rights to data has often been negotiated before handing in the application to EU. If Danish Researchers are requested to use specific data storage and data sharing solutions, it will be an obstacle for participating in international projects. Therefore, the selection of a potential national storage solution should be voluntary.

The most important aspect is that the storage is available without administrative burdens for the researcher and that the service is free, or handles without direct payments/economic administration between the research group and the storage provider.

National institutions that host interesting research data, could offer storing and sharing possibilities that make use of their data easy, e.g. Rigsarkivet, The Royal Library, the government and the Danish Regions.

National storage solutions could be build up to handle specific research areas. This would make it easier for the researchers to find relevant data (using more research specific metadata filtering) than in a general national data storage solution.

Concerning rights management and negotiation of agreements about data, local storage would be less complicated than national solutions. An institutional general storage solution would be able to benefit from Identity management and knowledge of the organization when structuring data. Where a general national solution could end up being a large mess of data sets with very different metadata elements, making it difficult to find the right data.

Hvilke trin i datas livscyklus skal dækkes af nationale services? Herunder tanker om understøttende support og kompetenceudvikling
Data collection: Data sharing faciliteter hvor man kan finde datasæt til brug i forskningen, fx af data fra KB, RA, offentligt støttede kulturprojekter, offentlige data kilder.
Data dokumentation: Fælles national juridisk enhed der rådgiver og forhandler om rettigheder og adgang til data.
Data analyse: Platforme hvor personfølsomme data kan analyseres og midlertidigt opbevares hvor databehandler aftaler er gjort nemt, eller er løftet til institutions niveau.

Det vil være et tigerspring fremad, hvis man i dansk lovgivning gjorde det lettere for forskere at få adgang til data.

Det ville også være et tigerspring fremad, hvis man havde en fælles juridisk enhed der kan forhandle forskeres adgang til data i forhold til de enkelte forskningsprojekters behov. Hvis der ikke kan være en forhandlingsenhed, så ville en national – evt. datatypeopdelt -juridisk rådgivningsenhed være guld værd for forskningen.

Hvilke udviklingsressourcer og hvilken organisering er der behov for, for at sikre en bæredygtig national udvikling i forhold til den internationale?
Der er brug for et stærkt national netværk mellem institutionerne. Netværk mellem institutionernes medarbejdere der arbejder med forskningsdata management – både med hensyn til tekniske løsninger og organisatoriske udfordringer – vil gøre arbejdet på de enkelte institutioner mere kvalificeret. Der vil være en langt bedre udnyttelse af ressourcerne af de samlede kræfter der bruges på datamanagement nationalt, hvilket kunne gøre det lettere for alle institutioner at være på et højt niveau vedrørende rådgivning.

Det kunne være hensigtsmæssigt hvis der er en fuldtids eller to deltidsmedarbejdere der holder en finger på pulsen internationalt, og videreformidler til de enkelte institutioner.

Der er særligt brug for at **opruste på det juridiske område nationalt**. Gerne ved at pålægge de institutioner der har store nationale dataresurser at tilbyde forskerne klare muligheder for at få adgang til data, og det ville være godt hvis de også kunne være progressive i at fjerne barrierer for adgang til data. NB: her er ønsket ikke at data skal kunne deles åbent, men blot at forskerne på en nationalt ensartet måde kan få adgang til data

Anbefalinger omkring det nationale samarbejde i DM Forum, herunder om det skal fortsætte og i givet fald med hvilke opgaver og budget.
DM Forum kan med fordel fortsætte som netværksenhed, der følger de internationale udviklinger og initiativer. Måske kan der tilføjes juridisk rådgivning og lobbyvirksomhed for adgang til forskningsdata.

Hvordan håndteres deltagelse i internationale medlemsskaber og samarbejde, og med hvilket engagement og investering?
Vigtigt at der også er institutionelle repræsentanter der har energi og tid til at indgå i de internationale samarbejder, men et overordnet samarbejde/koordinering af dette nationalt kunne være interessant.

Hvordan sikres national videndeling omkring internationale resultater og aktiviteter, med henblik på anvendelse såvel teknisk som forskningsmæssigt?
Vedrørende håndtering af forskningsdata kan der med fordel fastholdes et nationalt netværk hvor der helholdsvis deles informationer fx via en mailliste og ved at der afholdes 2 årlige møder hvor der videndeles omkring internationale resultater og aktiviteter.
Vedrørende teknisk håndtering af data storage på universiteterne kunne et netværk måske også være interessant.

Bilag 3: Opdatering og udvidelse af landeanalyse for Holland

DeiC's rapport *"Digital Infrastruktur til forskning i 2025"*[1] indeholder benchmarks i forhold til de nordiske land, UK og Holland. Fokus var på de nationale institutioner og deres understøttelse af forskningen. Gruppen blev hurtigt enige om, at specielt organiseringen og løsningerne i Holland var interessante og kunne danne model for en dansk tilgang til løsningen. Vi besluttede derfor at lave en opdatering – og udvide til også at have fokus på, hvordan institutionerne brugte de nationale løsninger.

Analysen i dette bilag er for en stor del lavet på baggrund af information fra internettet. Der må derfor tages forbehold for, at analysen ikke på alle punkter er udtømmende. Det har bl.a. været vanskeligt at finde specifik information om den bagvedliggende tekniske infrastruktur.

Analysen tager udgangspunkt i de overliggende organisatoriske strukturer og principper, der anses for at være styrende for det konkrete datamanagementlandskab. Dernæst beskrives de services, der udbydes, og endeligt i en vis udstrækning de anvendte teknologier og den grundlæggende hardware infrastruktur.

[1] https://www.deic.dk/da/analyserapport-digital-infrastruktur-til-forskning-i-verdensklasse-2025

1. Aktører

Den hollandske datamanagementinfrastruktur er distribueret og drives af mange aktører. RDNL, Surf og DANS er de væsentligste aktører i datamanagementlandskabet.

- **RDNL:** RDNL (Research Data Netherlands) er en alliance mellem 4TU.Centre for Research data, DANS og SURFsara med det formål at sikre langtidsarkivering af forskningsdata.
- **SURFsara:** SURFsara modsvarer DeiC.
- **DANS:** DANS (Data Archiving and Networked Services) er ddet hollandske institut for permanent adgang til digitale forskningsressourcer. En sammenslutning af 16 partnere med det formål at fremme FAIR-principperne. DANS er et institut under Koninklijke Nederlandse Akademie van Wetenschappen (KNAW).
- **4TU.Centre for Research Data:** Langtidsarkivering af forskningsdata inden for de tekniske videnskaber. Tjenesterne udbydes af TU Delft, der også udbyder specifikke tjenester til University of Twente og Technical University of Eindhoven.
- esciencecenter.nl**:** Kompetencecenter for udvikling og anvendelse af videnskabelig software med det formål at fremme akademisk forskning.

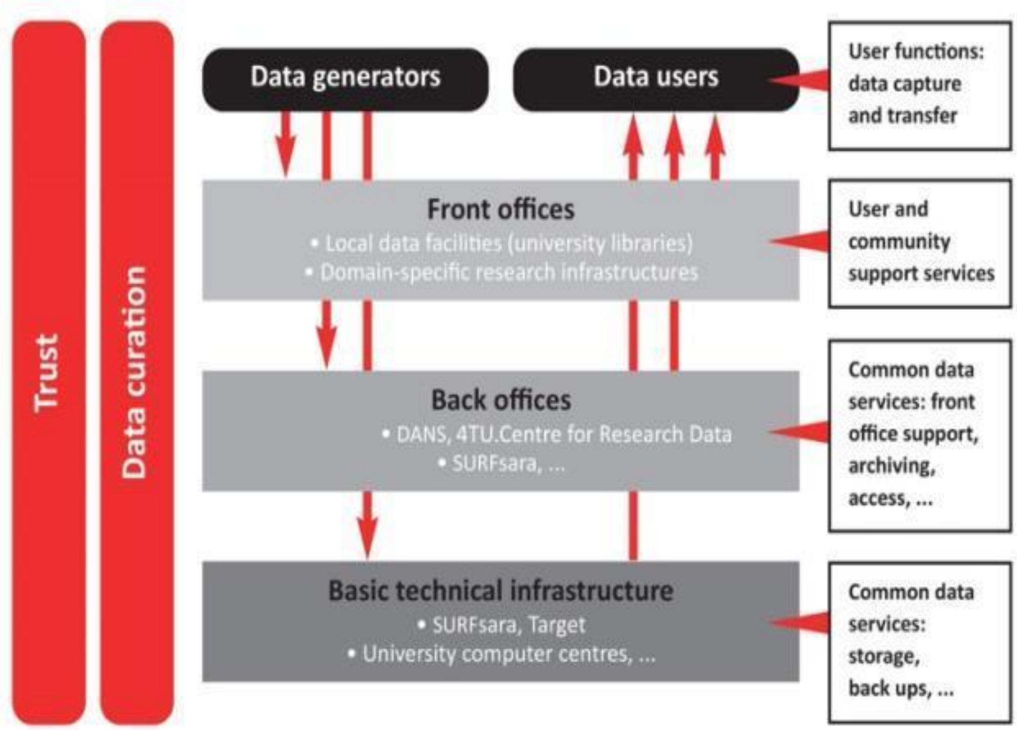
2. Governance

De nationale datarepositorier finansieres gennem et centralt nationalt Research Council (NWO). I Danmark finansieres datarepositorier dels gennem bevillinger fra ministerier, dels gennem medlemsbidrag, brugerbetaling, EU-bevillinger og andet.

Til styring benyttes den såkaldte FOBO-model, hvor FOBO står for Front Office Back Office. Front Office fungerer som mellemled mellem den ofte centralt leverede teknologi og forskerne.

Modellen nedenfor illustrerer den overordnede struktur i datamanagementlandskabet.

En række lokalt forankrede Front Offices varetager kontakten til brugerne, mens centralt placerede Back Offices tilvejebringer generelle og domænespecifikke dataservices, der hviler på en centralt koordineret, men distribueret og fødereret teknisk infrastruktur, der udbyder et antal Trusted Data Repositories (TDR) samt tjenester til håndtering af datamanagement.


Trust
Data curation
Data generators
Data users
User functions: data capture and transfer
Front offices
• Local data facilities (university libraries)
• Domain-specific research infrastructures
User and community support services
Back offices
• DANS, 4TU.Centre for Research Data
• SURFsara, ...
Common data services: front office support, archiving, access, ...
Basic technical infrastructure
• SURFsara, Target
• University computer centres, ...
Common data services: storage, back ups, ...

3. Services

3.1. RDNL

Rådgivning og træning i forhold til anvendelse af de repositorier, der drives af SURFsara, DANS og 4TU.Centre for Research Data.

3.2. SURFsara

Research Drive: En generel data storage service, der gør det muligt at lagre og dele store datasæt.

Data Archive: Langtidsarkivering af forskningsdata på bånd. Data gemmes i to kopier på to lokationer i Amsterdam. Desuden udbydes forskellige services, der muliggør effektiv overførsel af store datasæt i form af HPN-SSH eller GridFTP. Data Migration Facillity DMF. SURFfilesender.

B2SAFE: En generel distribueret, fødereret datamangementservice, der gør det muligt, at store projekter sikkert og robust kan replikere data til forskellige lokationer. B2SAFE tilvejebringer værktøjer, så datamanagementpolitikker kan implementeres på lokale B2SAFE-instanser med henblik på at sikre, at identiske datasæt håndteres ud fra samme politikker. B2SAFE datamanagementpolitikkerne er i overensstemmelse med EUDAT-politikkerne, men kan udvides med politikker, der gælder communities, projekter eller institutter. B2SAFE er baseret på iRODS (Integrated Rule-Oriented Data System). iRODS er et Open Source datamanagementværktøj, der muliggør at modellere regelbaserede politikker. iRODS kan binde distribuerede, heterogene datalagre sammen og samtidigt sikre en ensartet håndtering af identiske datasæt baseret på regelbaserede politikker.

Object store: Effektiv lagring af (ekstraordinært) store datasæt, der består af (overvejende) statiske filer. Velegnet til lagring af mediefiler (streaming) og filer, der har tilknyttet metadata. Kan desuden fungere som diskbaseret backup og arkiv. Er desuden storagefacillitet for f.eks. Research Drive samt figshare og Dataverse-baserede repositorier. Endelig anvendes Object store som data storage for iRoODS-instanser (B2SAFE). Obejct store er baseret på OpenStack SWIFT.

3.3. DANS

DataverseNL: Lagring og deling af forskningsdata. Alle formater er understøttet. Filstørrelse op til 10 GB. Understøtter versionering, hierarkisk strukturering og tilpassede metadata.

EASY: Certificeret langtidsarkivering af forskningsdata. Understøtter DOI og standard Dublin Core-metadata. Det er ikke muligt at strukturere data hierarkisk eller versionere.

NARCIS: Det nationale datakatalog for forskningsdata. NARCIS : National Academic Research and Collaboration Information System.

NARCIS er etableret med den målsætning at have en central søge facilitet for al forskningsinformation produceret af de Hollandske universiteter, the Royal Neterlands Academy of Arts and Sciences (KNAW) og the Netherlands Organisation for Scientific Research (NOW).

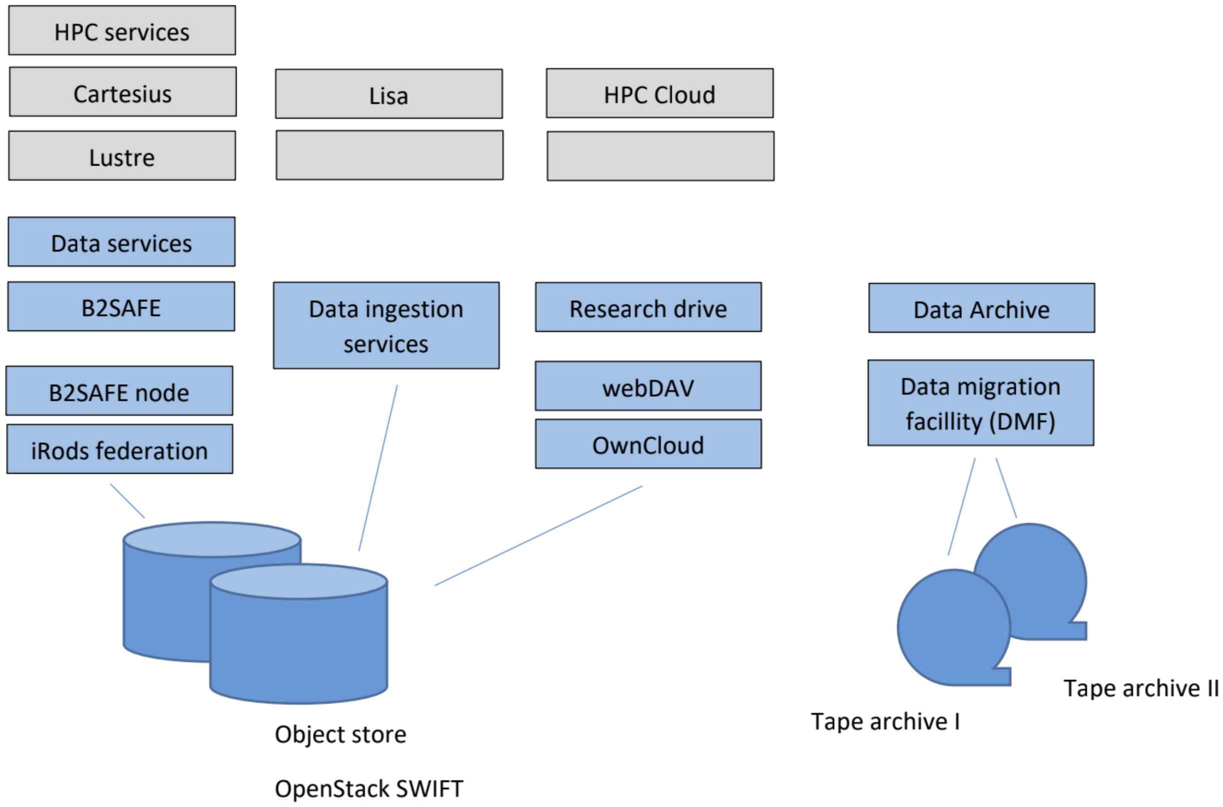
Portalen kombinerer struktureret forsknings information med information fra OAI-repositorier (Open Archive Initiative for Metadata Harvesting), websites og nyhedssider fra forskningsenheder.

Anvender webDAV-protokollen og er dermed tilgængelig fra alle systemer. Det er muligt at tilgå Research Drive fra lokale enheder.

3.4. Tekniske infrastrukturkomponenter:

Nedenstående figurer er baseret på informationer om den tekniske infrastruktur, der kan hentes ud af servicebeskrivelser fra SURFsara, DANS og 4TU.Centre for Research Data. Figurerne er, som det fremgår, ikke udtømmende.

SURF infrastructure :



DANS infrastructure :

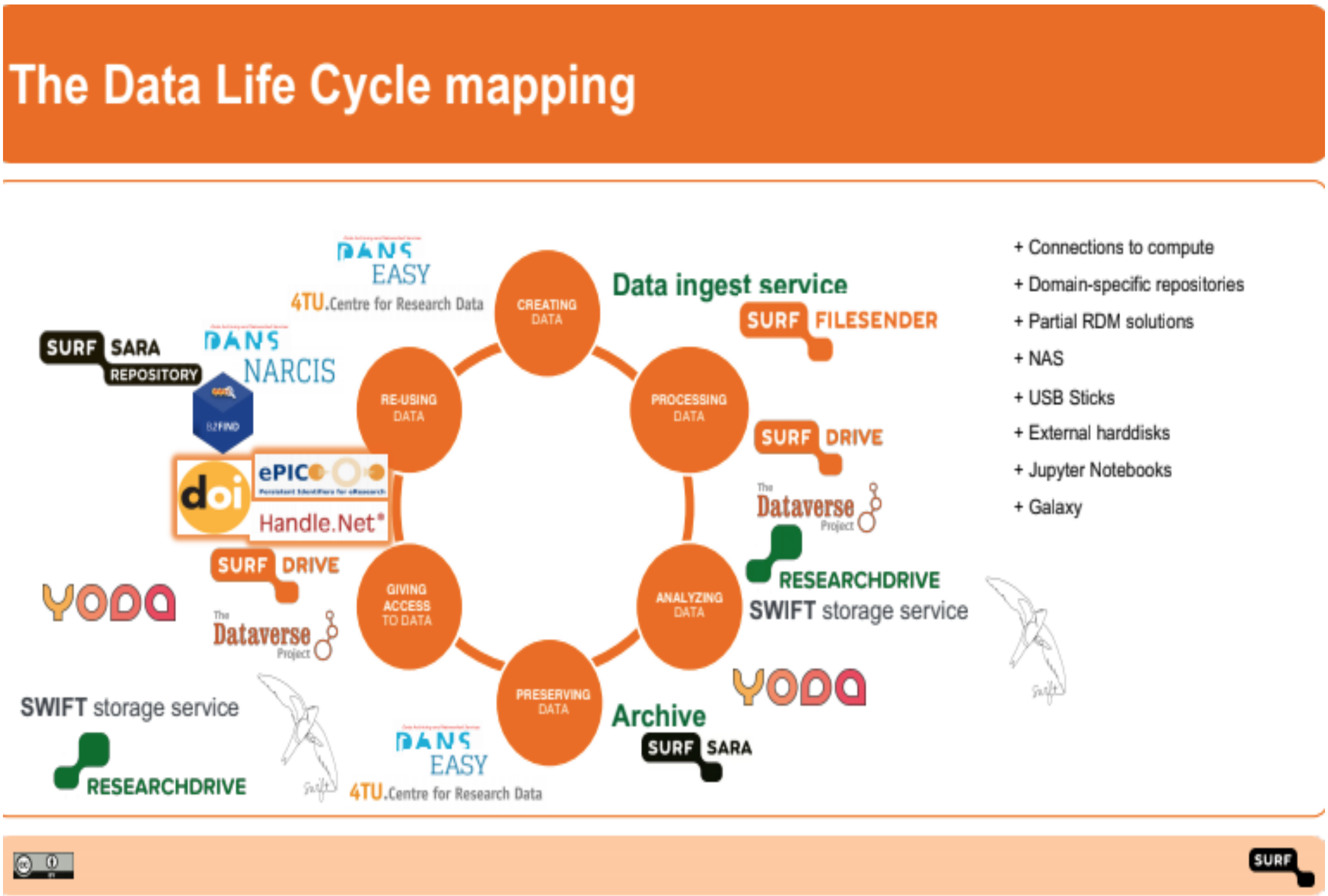
dataverse.nl	EASY	NARCIS
dataverse		

4TU. Researchdata Archive infrastructure :

4TU.Researchdata Archive

4. Datalivscyklus

Figuren ovenfor giver et overblik over de tjenester, som tilbydes i de forskellige faser af livscyklus. For at få et bedre indblik i anvendelsen blev det besluttet at se nærmere på to universiteters tilbud, nemlig Utrecht og Delft. Begge hører til i front, men med to meget forskellige tilgange til en løsning. Delft er en del af 4TU og står for at udbyde fælles tjenester. Utrecht har tradition for at være langt fremme – ofte med egne løsninger.



The Data Life Cycle mapping

- + Connections to compute
- + Domain-specific repositories
- + Partial RDM solutions
- + NAS
- + USB Sticks
- + External harddisks
- + Jupyter Notebooks
- + Galaxy

En detaljeret sammenligning er vist i appendiks A. Nedenfor er nogle af hovedobservationerne samlet.

4.1 Projektplanlægning, genbrug af data

Begge universiteter understøtter denne funktion ved at stille DMPonline til rådighed. Begge universiteter har udviklet et stort informationsmateriale, som kan bruges til at vurdere alle facetter af data – rettigheder, etiske overvejelser, etc.

Begge universiteter understøtter genfinding af eksisterende data. Begge peger på Re3data.org som et sted, hvor relevante repositorier kan findes. Delft peger på sit eget dataarkiv, 4TU.Centre for Research Data. Utrecht peger på den nationale løsning NARCIS samt på det kommercielle produkt Zanran.

Med en usikkerhed på uendelig tyder noget på, at har man sin egen løsning, er man mindre tilbøjelig til at understøtte de nationale tilbud.

4.2 Forskningsprocessen

4.2.1. Storage

IT-afdelingerne tilbyder lokalt lager som personlige drev og/eller gruppedrev. Begge peger på SURFdrive som en god, sikker løsning og et klart alternativ til Dropbox.

Utrecht understøtter hele forskningsprocessen, inkl. storage, med et egenudviklet system, YODA.

4.2.2. Styring

Utrecht peger, ud over sit eget system YODA, på eLABJournal og OneNote. Desuden har de oprettet deres eget GIT-repository. Endelig peger de på Open Science Framework som en mulig platform.

4.2.3. Samarbejde

Utrecht peger på SURFfilesender og på Boxcryptor til kryptering.

5. Publicering

Begge universiteter understøtter de meget forskellige behov i forhold til storage, som forskerne har. Dette afspejler sig i en myriade af forslag til, hvor data skal lagres, inkl. nationale løsninger som EASY (humanities og social science) og 4TU.Centre for Research Data (geo, engineering and technical science) og internationale frie og kommercielle løsninger som Zenodo, B2Share og figshare.

6. Data stewardship

TU Delft reklamerer for deres data stewardship-support på følgende måde[2]:

Your faculty Data Steward is the first contact point for any inquiries related to research data management and sharing. Your Data Steward can offer you a broad range of support, including:

- tailored consultations on your data management practice
- information about data storage and backup options available to you at TU Delft
- advice on the use of data management tools (including electronic lab notebooks and tools for software management)
- help with data management plans and funders' policies
- help with meeting journals' requirements for data availability
- information about data repositories, including the 4TU.Centre for Research Data
- advice on how to increase impact with data sharing
- dedicated workshops and information sessions about research data management
- information about working with confidential research data (e.g. commercially sensitive data or personally identifiable data)

6. Opsummering

Hollands strategiske opdeling i Front Office Back Office samt teknisk infrastruktur fungerer.

De to universiteter, som er undersøgt, bruger et miks af egne løsninger og centralt udbudte løsninger. De sidste kan være driftet af SURFsara eller DANS, men kan også udbydes og vedligeholdes af andre

[2] https://www.tudelft.nl/en/library/current-topics/research-data-management/r/data-stewardship/support/support-overview/

institutioner – enten i nationale samarbejder som 4TU.Centre for Research Data eller på basis af internationale infrastrukturer.

Utrecht har investeret i deres eget system, YODA, som er universitetets foretrukne platform til at understøtte alle faser af datalivscyklus. YODA er baseret på iRODS, som også er den foretrukne protokol i forhold til de nationale løsninger.

Begge universiteter har en differentieret tilgang til lagring af data, som afspejler krav til åbenhed og størrelse af datasæt. Det skal her bemærkes, at Holland har en tradition for at certificere repositorier. Der er 16 repositorier, som har fået … eller CoreTrustSeal. Til sammenligning er der i Danmark kun tre, hvoraf CLARIN-DK er et af dem.

Inspireret af løsningen i Holland foreslås, at følgende typer tjenester også bør stilles til rådighed for danske forskere:

Data Management Planning: DMP online

Datakatalog: Et nationalt datakatalog for forskningsdata med søgefacilitet (NARCIS). Suppleres med brug af R3Data.org.

Workflowstyring: Mulighed for at få systemet til at understøtte de forskellige faser (f.eks. YODA). Brug af iRODS synes at fungere godt.

Research Drive: En generlsk data storage service, der gør det muligt at lagre og dele store datasæt.

Deling af filer: Der udbydes forskellige services, der muliggør effektiv overførsel af store datasæt (HPN-SSH eller GridFTP, Data Migration Facillity DMF, SURFfilesender)

Lagring af forskningsdata: Lagring og deling af forskningsdata. Løsning afhænger af type og størrelse af data (f.eks. DataverseNL, Object Store, Data Archive, figshare, 4TU.Centre for Research Data, Zenodo).

Langtidsarkivering: Certificeret langtidsarkivering af forskningsdata. Understøtter DOI og standard Dublin Core-metadata. Det er ikke muligt at strukturere data hierarkisk eller versionere (f.eks. EASY).

Appendiks A

	Utrecht	Delft
Er der formuleret vision/strategi/politik – på nationalt plan og/eller på de enkelte universiteter?	At Utrecht University, it is important that all researchers honour scientific standards, including the meticulous and ethical treatment of research data. Utrecht University's requirements and expectations in this regard have been established by an administrative order in the University Policy Framework for Research Data which took effect as of 1 January 2016.	The TU Delft Research Data Framework Policy supports high quality research data management across each of the faculties at TU Delft.
Nationalt	**Research Organizations (VSNU, KNAW)** A number of organizations that promote the advancement of science in the Netherlands (such as KNAW and VSNU) have formulated policies or guidelines with regards to RDM. The **Code of Conduct** by the **VSNU** (the Association of Dutch Universities) contains guidelines on the adequate handling of raw data safeguarding the quality of data collections, and the storage of raw research data. The VSNU Code of Conduct stipulates that raw data should be stored at least 10 years. These rules are clearly based on common principles with regards to academic integrity. The VSNU, NWO and KNAW have also developed the standard evaluation protocol (SEP), which forms the basis for the evaluation of scientific research in the Netherlands. The SEP protocol includes a data management section in which faculties should describe their policies, activities and infrastructure with respect to research data management.	
Ejerskab til data	Officially Utrecht University, as your employer, is considered the rights holder to the research data you create. You, as a researcher, have the primary responsibility for taking care of the data.	If you have a working relationship with TU Delft it is likely that TU Delft owns your research data, but check whether other arrangements have been made with third parties. …… Remember that the TU Delft Research Data Framework Policy expects you to deposit your research data, code and any other materials needed to reproduce research findings in a research data repository in accordance with the FAIR principles (Findable, Accessible,Interoperable and Reusable), unless there are valid reasons not to do so.

Input til håndtering af livscyklus		
Research Planning		
Understøttes Project Proposal ja/nej	Ja. DMP online Tilbyder at lave review på udarbejdede DM-planer Værktøj til estimering af prisen	Ja. DMPonline Værktøj til estimering af prisen
Understøttes "Reuse of existing data" Ja/nej	Rådgivning. Peger på tre tjenester: ● Re3data, register over data repositories, ● NARCIS, national løsning driftet af DANS, som understøtter søgning. ● Zanran, kommecielt produkt, som understøtter søgning på nettet efter grafer og tabeller, som har været indlejret i dokumenter.	Rådgivning, peger bl.a. på ● Re3data.org. ● 4TU.ResearchDataArchive: Centre for Research Data (TU Delft is co-founder) maintains a data archive that offers access to research data from applied technical scientific research. Conduct a tailored search for data using the filters in the data archive.
Project start-up hvordan støttes project start-up, specielt hvor der er tale om personfølsomme data. Er der udviklet støtteværktøjer som "consent form generator", databehandleraftaler eller lignende?	Data protocols Ethics Decision Aid (DEDA for research) developed by Utrecht Data School	Copyright
Active State of Research		
Collect data		
Storage and backup	**Lokalt:** En blanding af egenudviklet, indkøbt og nationale løsninger. Indkøbte løsninger: IBM, lokaldrev samt onedrive Egenudviklet: YODA **Nationalt**: Surfdrive	Opererer på to niveauer: Lokalt (P-drev og lignende) eller nationalt. **Lokalt:** Personlige drev (gratis op til 8 GB), gruppedrev (gratis op til 50 GB), midlertidigt bulk data) eller SharePoint (mere til projekt information). **Nationalt:** SURFdrive

		Styring af processen, herunder også metadata og dokumentation	Peger på fire ”lokale” muligheder: ● YODA ● eLABJournal ● Microsoft OneNote ● GIT Andre: ● Open science Framework	Ikke identificeret
	Process/collaborate around Data			
		Understøttes “Information security, sharing large files” som en integreret del af datamanagement, eller er det en separat organisation?	Nationale løsninger: SURFfilesender Kryptering: Boxcryptor	Peger på en sikker FTP-server. The default size of files is 10 GB.
	Analyze data			
		Understøttes analyse som en integreret del af datamanagement, eller er det en separat organisation?	Anbefaling af værktøjer	Anbefaling af værktøjer
Sharing Results				
	Publish Results			
		Publication in data repositories	Med langtidsbevaring: **Lokalt**: YODA **Nationalt:** ● DataverseNL ● Surf Data Archive Har udviklet et beslutningsværktøj til at vælge mellem: ● (DANS) EASY (for humanities and social sciences) ● 4TU.ResearchData (for geo, engineering and technical sciences) ● DataverseNL (a system to preserve, share and publish data while keeping datasets together as a collection). ● B2Share (for European scientists and researchers to store and share small-scale research data)	*Peger på* ● **As supplementary materials attached to a journal** ● **As a description of your dataset in a data journal** **In a repository or data archive**, hvor relevante repositories kan finds via Re3data.org. ● DataverseNL is specifically designed to store, back-up, organise, annotate and share research data with colleagues all over the world. With this open source application you can grant multiple individuals controlled access to your data.

			• Zenodo (a generic data repository for EC funded research) • Dryad (a general-purpose home for a wide diversity of data types) • Figshare (to manage, publish and share research data). • Open Science Framework (OSF) (Open source environment to cooperate in projects) • YODA (Institutional research data storage platform from Utrecht University).	
		Persistent identifiers		
	End of Project			
		Long term preservation	Kalder nedenstående for long-term (preservation) storage: YODA DataverseNL SURF Data Archieve	4TU.Centre for Research Data

Referencer

A federated data infrastructure for the Netherlands: the Front Office Back Office model:

https://researchdata.nl/fileadmin/content/RDNL_algemeen/Documenten/RDNL_FOBOmodel-UK-web.pdf

DANS: http://www.dans.knaw.nl

Dataverse and EASY – wich service is the best solution for ICT: https://www.itc.nl/library/research/dans-itc20170421.pdf

EUDAT B2SAFE: https://www.eudat.eu/b2safe

National Nodes – Getting organized; how far are we?: http://e-irg.eu/documents/10920/238968/NationalNodesGettingorganisedhowfararewe.pdf

NARCIS : The Gateway to Dutch Scientific Information

Paula Martinez Lavanchy, Research Data Officer @ TU Delft RDNL: www.researchdata.nl

Research data management – An overview of recent developments in the Netherlands: https://pure.knaw.nl/portal/files/5752346/WhitepaperResearchdatamanagementAnoverviewDEF.pdf

SURFsara: http://www.surf.nl

Bilag 4: Review af tre eksisterende storage løsninger

ERDA/SIF

ERDA

University of Copenhagen - Electronic Research Data Archive (UCPH ERDA or just ERDA) is a storage, sharing and archiving facility provided by University of Copenhagen to employees and students. ERDA delivers centralized storage space for personal and shared files in addition to archiving of e.g. PhD theses and research data for safe-keeping and publishing. It also comes with a file synchronization service similar to Dropbox, but with the data stored locally at UCPH.

Although ERDA comes with a strong security focus, it is only approved for general scientific data and not for highly sensitive data. In particular, it is not for personal data classified as sensitive in the EU Regulation 2016/679 (General Data Protection Regulation).

Users working with sensitive data can instead use the ERDA sister facility SIF, which is intended and approved for exactly that purpose.

For now ERDA/SIF use is limited to the faculty of Science, but it is the intention to extend the coverage to UCPH as a whole, to support the general UCPH Data Management Guidelines. However, non-Science users and external users without a general UCPH account, can still get an ERDA account, provided that they are engaged in a project collaboration with one or more employees at UCPH. It is also possible for Science users to preserve access to their ERDA account beyond their employment at UCPH in case they need it to keep collaborating with one or more remaining employee(s).

Replication & Back-up

All data on ERDA is automatically replicated over multiple disks with RAID technology. Only files explicitly duplicated with the Archive feature are backed up to tape. This includes eventual remote tape backup for extra safety. ERDA also provides a Seafile file synchronization service for any small to medium sized data sets, for which automatic file versioning is wanted, so that deleted files can be recovered or 'rolled back' to earlier versions.

DOI

UCPH has a license to mint DOIs through DataCite and KU-IT has developed a service to help do that. This service is integrated with ERDA.

SIF

Popularly speaking SIF is the Sensitive Information Facility part of ERDA at University of Copenhagen (UCPH). It is set up for storing sensitive data, in particular personal data requiring special care under the EU General Data Protection Regulation (GDPR). SIF delivers secure centralized storage space for various projects involving sensitive data and integrates safe sharing with other project participants. Replication and back-up olicies are equal to those of ERDA.

When creating new projects storing data the Faculty Secretariat is automatically notified, any login attempt is logged, any access to data is logged along with ID, time of day and other relevant information, and any access pattern that is considered "suspicious" will result in a suspension of your account

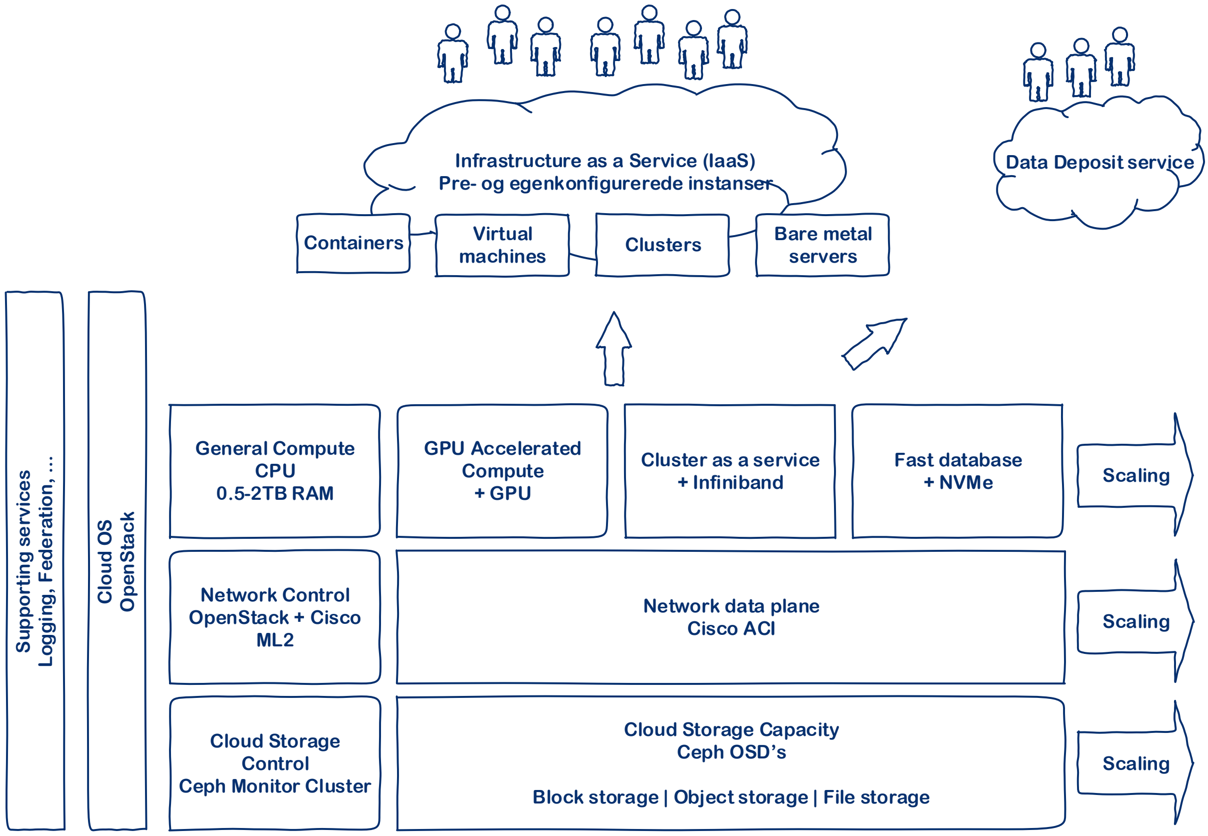
Examples of suspicious behavior includes:

- Login from different IP addresses within a short time frame
- Login from outside Denmark
- Structured download of data
- Attempts to circumvent security controls or access data to which user have no privilege.

CLAAUDIA Cloud

CLAAUDIA's primære infrastruktur er en "privat cloud" baseret på Ceph og OpenStack. Dette er efter model fra andre anerkendte forskningsinstitutioners private clouds. Her kan nævnes Cambridge University, DESY, NASA, Cern, SLAC og mange flere.

På figuren ses de væsentligste system komponenter i cloud'en.


Infrastructure as a Service (IaaS)
Pre- og egenkonfigurerede instanser
Data Deposit service
Containers
Virtual machines
Clusters
Bare metal servers
Supporting services
Logging, Federation, ...
Cloud OS
OpenStack
General Compute CPU 0.5-2TB RAM
GPU Accelerated Compute + GPU
Cluster as a service + Infiniband
Fast database + NVMe
Scaling
Network Control OpenStack + Cisco ML2
Network data plane Cisco ACI
Scaling
Cloud Storage Control Ceph Monitor Cluster
Cloud Storage Capacity Ceph OSD's
Block storage | Object storage | File storage
Scaling

Cloud'en består af 3 lag, som hver kan udbygges og skaleres efter behov. Bunden er et storage-lag, midten er et netværks-lag og toppen er et beregnings-lag. Det hele styres af et Cloud Operativ System (Cloud OS).

Cloud services

Det direkte produkt er "infrastruktur-as-a-service" (IaaS). Det betyder at en (avanceret) bruger kan bygge en applikation til dataprocessering ud af standard infrastrukturkomponenter, som virtuelle maskiner, netværk og containere. En sådan applikation kan udstilles som en cloud service til andre brugere på et højere abstraktionsniveau (pre-konfigurerede instanser). Altså noget der opleves som en out-the-box løsning for den enkelte bruger. Udviklingen af applikationer, som kan målrettes bestemte anvendelser, laves af support funktioner som CLAAUDIA i samarbejde med superbrugere indenfor de enkelte fagfælleskaber.

Et eksempel på en cloud applikation er AAU's Data Deposit Service, der sigter mod høj governance på lagring af forskningsdata. Den er udviklet som en selvstændig service, men anvender den samme underliggende cloud infrastruktur. Det muliggør tæt integration til de øvrige services, så eksempelvis unødig kopiering undgås. Som en cloud service kan den desuden integrere mod 3. parts services over Internettet, eksempelvis Zenodo.

Tekniske valg

Som Cloud OS er valg OpenStack fordi det globalt har langt den største udbredelse på private clouds til forskning og undervisning[3]. Det er vigtigt at den API som cloud'en udstiller til cloud applikationer er den samme som bruges i de miljøer hvor ny software til videnskabelig brug udvikles.

Som storage-lag er valgt Ceph. Ceph er en meget stabil og udbredt storage platform for OpenStack clouds. Ceph er et software defined storage system til cloud storage. Det er opbygget i 2 lag. Kapacitetslaget består af storage servere. Kapacitet og båndbredde øges samtidigt med antallet af storage servere. Kontrol laget består af almindelige servere. Replikering i både control og storage lag sikrer høj driftsstabilitet. Ceph har intet "single point of failure" og er designet til kontinuert drift under opgraderinger af både software og hardware.

Beregningslaget har 4 hovedtyper af noder:

- Almindelig dataprocessering (36-128 cores, 0.5 – 2 TB RAM)
- GPU acceleration til f.eks. AI og video processering (2 – 4 GPU'er per node)
- Lav latency lokal disk (NVMe) til databaser der behøver hurtige transaktioner
- Noder forbundet via InfiniBand til "cluster-as-a-service"

Som netværk er valgt Cisco's ACI, som er et Software Defined Network (SDN) der implementerer OpenStack's netværksmodel direkte på netværks hardwaren. Direkte hardware undertøttelse er nødvendigt for at garantere fuld segmentering når der anvendes både virtuelle- og bare-metal instanser. Netværket er af "spine-leaf" typen, som kan skalere med udvidelse af compute- og storage-lagene.

Generelt gælder at både kapacitet og båndbredde øges ved udvidelse af cloud'en.

System egenskaber

Følgende er nogle af de centrale egenskaber ved arkitekturen og de tekniske løsninger

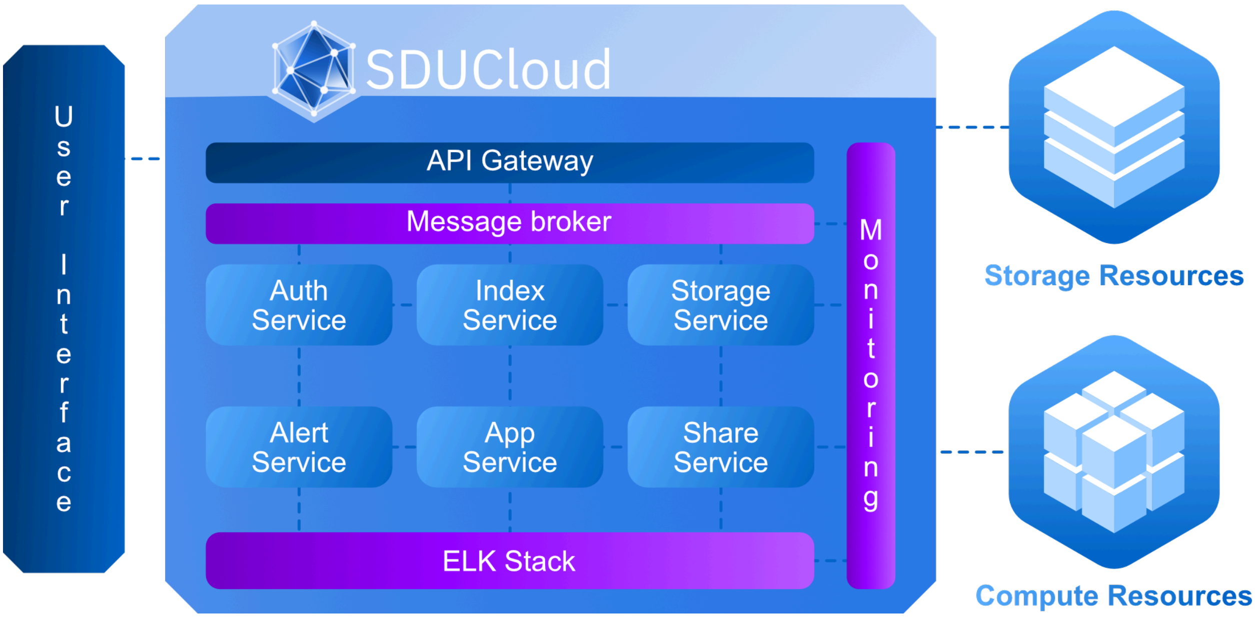
- Skalering af compute-lag, netværks-lag og storage-lag kan udføres sammen eller individuelt
- Skalering af control-lag for netværk, storage og Cloud OS kan udføres under drift
- Support for flere tenants/kunder med egen governace for brugere og roller
- Tenants er fuldt adskilte på data og netværks lag
- Fysiske noder kan knyttes til en tenant
- Distribueret virtuel routing gennem et SDN med en hardware optimeret datapath

[3] https://www.nasa.gov/offices/oct/40-years-of-nasa-spinoff/openstack-cloud-computing-platform

- Der bygges med standard rack-servere med fokus på pris/ydelse frem for redundans på hardwarelaget
- Hardware er bygget i moduler af 19 tommer rack enheder med storage, compute og netværk
- Høj driftssikkerhed er sikret gennem software laget med høj redundans i alle Cloud'ens kontrol-planer
- 5 separate fejl-zoner betyder at cloud applikationer selv kan etablere redundans
- Undertøtter instanser i form af containere, virtuelle maskiner og bare-metal servers
- Standard OpenStack REST API for integration med "cloud schedulers" som f.eks. Kubernetes

SDUCloud

SDUCloud er et digitalt forskningsmiljø som er bygget til at understøtte forskeres krav til computerkraft og data management. Det er designet til at være brugervenligt med en intuitiv grafisk brugerflade (GUI). Platformen er designet til at kunne udvides for at møde de evigt skiftende krav til forskning. Platformen er fyldt med features designet til research workflows. Missionen for SDUCloud er at gøre kompleks digital teknologi tilgængelig for alle forskere.



Solutions & Services

Datalagring

En datalagrings løsning designet til at være **holdbar, robust** og **sikker**. Løsningen er fleksibel for at kunne håndtere de **store mængder og variation i forskningsdata.** Forskellige datalagrings arkitekturer er understøttet, fra ustruktureret data lakes til high-performance filsystemer.

Data Management

SDUCloud integrerer en **skalerbar metadata service** som kan bruges til at **automatisere normale data management opgaver**, som f.eks. backups, data ingestion, præprocessering og arkivering. Metadata featuren er fleksibel og kan håndtere både standard og brugerdefinerede metadata formater.

Fildeling

Fildelingsfeaturen tillader brugere at **samarbejde på datasæts**, ved at gøre mapper tilgængelige til alle medlemmer i et hold. Forskellige brugere i et hold kan tildelt forskellige rettigheder.

Samarbejde

SDUCloud tillader brugere at lave **virtuelle workspaces dedikeret til hvert projekt**. Medlemmer af et team kan tilføjes til projektet med forskellige roller (f.eks. PI, ekstern kollaboratør, data steward osv.), hver rolle med deres eget sæt af rettigheder. Medlemmer af et projekt kan dele data, metadata, workflows og resultater fra applikationer.

Compute

En fleksibel feature til computerkraft som understøtter mange forskellige workflows, fra **interactive computing,** til mere traditionelt **high-performance computing** eller **data analytics og visualization**. Platformen understøtter også "cloud"-lignende services, dette kunne for eksempel være web-baseret dashboards eller websites til research grupper. Alt fra simple one-page applikationer til komplekse konfigurationer kan bygges. Compute featuren understøtter både **applikationer på "bare-metal" eller i virtuelle miljøer.**

Applikationer

SDUCloud tilbyder en "App store" hvor forskere kan finde deres yndlingsapplikationer. **Simple såvel som komplekse applikationer kan blive deployed ved hjælp at enkelte klik** via en intuitiv brugerflade, som ikke kræver nogen IT træning. Applikationer kan let tilføjes efter behov.

Indeksering

Metadata, fildata, applikationsdata og brugeraktivitet er indekseret i et skalerbart system som tillader **hurtig og nær real-time søgning** på tværs af filer og metadata i systemet. Indeksering af data og metadata tillader forskere at finde de datasæt de ønsker, når de har brug for dem.

Sikkerhed

SDUCloud er bygget med sikkerhed i tankerne fra begyndelsen. Sikkerhed er indbygget i apps, services og infrastruktur med autentificering, kryptering, netværkssikkerhed og monitoring. Brugere har mulighed for at tilføje multifactor authentication. En formel ISO27k certificeringsprocesse startede i foråret 2019 og vil blive fuldendt i 2020.

Auditering

Alle handlinger i SDUCloud producerer et detaljeret audit spor som gemmes i en central og sikker placering. Hver bruger kan også se og søge i deres egen fil- og applikationsaktivitet.

Adgangskontrol

SDUCloud **integrerer et avanceret access control system** for at styre adgang til ressourcer, såsom filer og applikationer. Forskellige roller kan defineres, hvert med et sæt rettigheder. Rettigheder kan tillade eller forbyde operationer i SDUCloud, for eksempel læs/redigering af filer, læs/redigering af metadata, adgang til applikationer osv.

Teknologi

Systemarkitektur

SDUCloud bruger en microservice arkitektur som gør det muligt for de enkelte services at være skalerbare og fleksible. Microservices er små og uafhængige enheder som, når de arbejder sammen, udbyder en række funktioner af SDUCloud. Microservices samarbejder ved hjælp af REST APIs. Microservices kan deployeres uafhængigt af hinanden, hvilket minimerer risiko og simplificerer operationelle procedurer.

SDUCloud er udviklet in-house ved SDU eScience center og bygget ovenpå veletablerede open source teknologier.

Backend

Kotlin: Et moderne programmeringssprog på JVMen. Det meste af SDUClouds backend er skrevet i Kotlin.

Ktor: Asynchronous web servers og clients, bliver brugt af SDUClouds microservices.

Postgresql: SQL database som gemmer det meste af SDUClouds service data (ikke bruger / research data)

Redis: In-memory datastruktur store som kan bruges som en database, cache og message broker. SDUCloud microservices bruger Redis til intern kommunikation.

Elasticsearch: Open source løsning til skalerbar og nær real-time indeksering. Den leverer metadata, søgning og auditerings services til SDUCloud.

Ceph: Skalerbar og fleksibel software defined object storage. Bruges af SDUCloud både som storage provider til den interne backend samt som en storage provider til research data.

Jenkins: Bliver brugt til at levere diverse interne funktioner, f.eks: continuous integration (compile og tests), continuous delivery (klargøring af containers).

Kubernetes: Skalerbar container orchestration system som gør det muligt at definerer en komplet software stak i kode. Blandt dens features er: service discovery, load balancing, self-healing, automated rollouts og rollbacks, secret- og konfigurationshåndtering, fleksibel skedulering af jobs.

Frontend

Typescript: En type-sikker udgave af JavaScript udviklet af Microsoft. Frontend koden er primært skrevet i Typescript.

React: Et JavaScript bibliotek til deklarative brugerflader, brugt af mange hjemmesider.