NTW2019 - ScienceData - from local to global

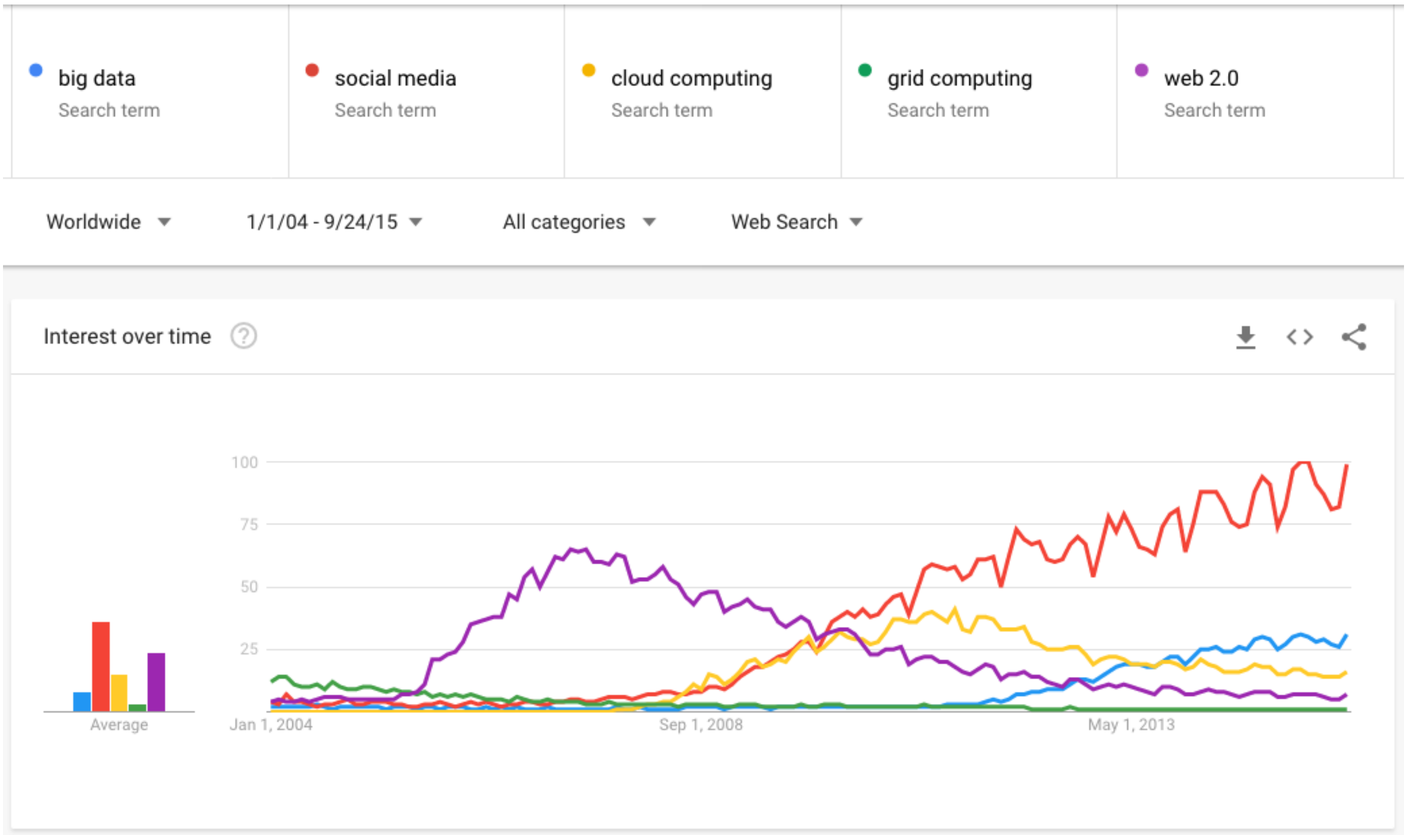
Background/boundary conditions

- "*DeiC - Danish e-Infrastructure Cooperation was established on April 19, 2012, with the purpose to support Denmark as an e-Science nation through delivery of e-infrastructures (computing, storage and network) to research and researchbased teaching*" (from deic.dk)
- The same year I was given the task of building national services for computing and storage
- We called the first incarnation of our services compute.deic.dk and data.deic.dk
- I presented data.deic.dk here 4 years ago
- After that, I immediately started the development of a distributed storage solution which I will return to later

Outline of talk:

- What has changed - IT landscape, challenges
- ScienceData - why and how
- CS3MESH4EOSC - going global - EOSC, Australia, national services, how it all fits together

What has changed



● big data
Search term

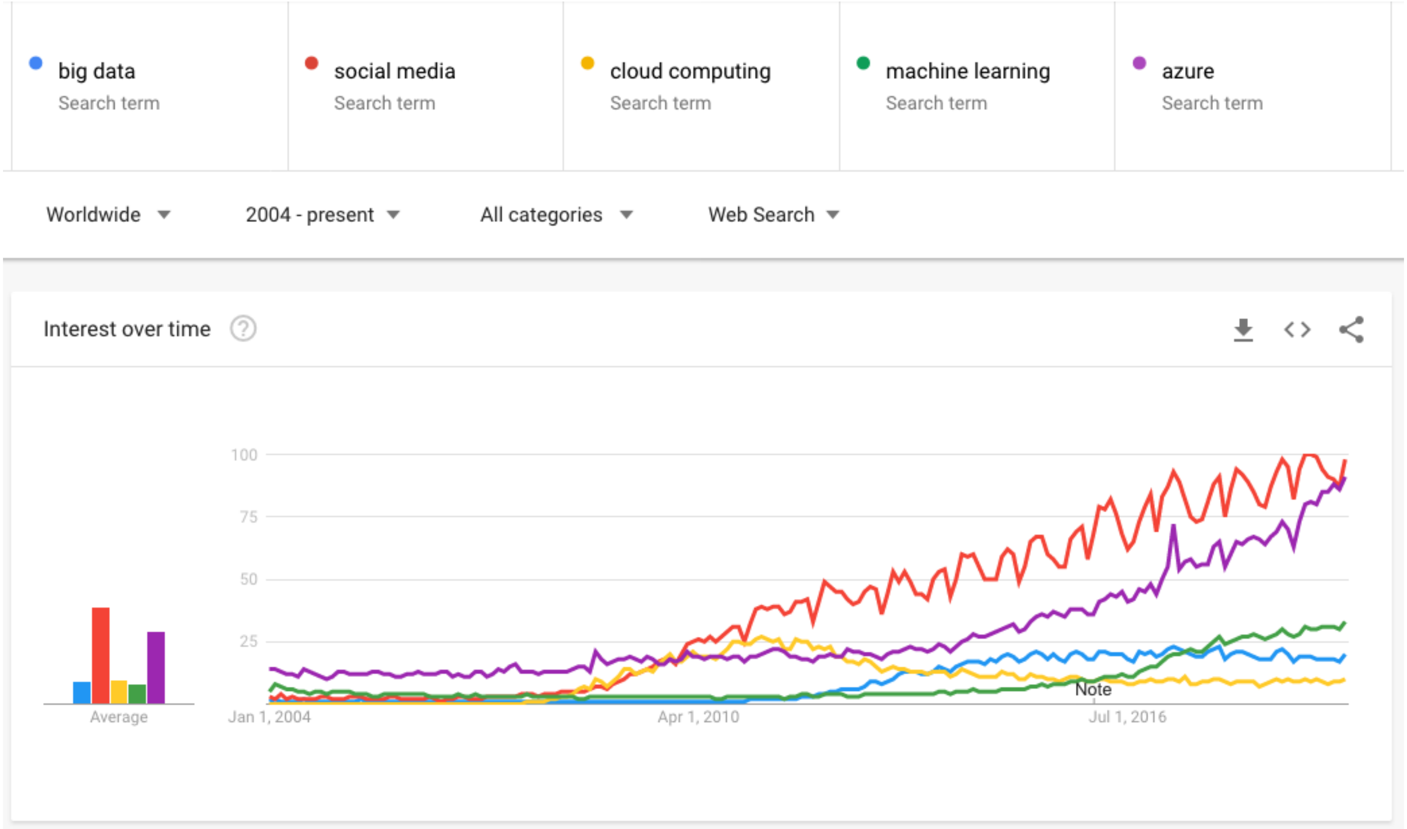
● social media
Search term

● cloud computing
Search term

● grid computing
Search term

● web 2.0
Search term

Worldwide ▾ 1/1/04 - 9/24/15 ▾ All categories ▾ Web Search ▾



● big data
Search term

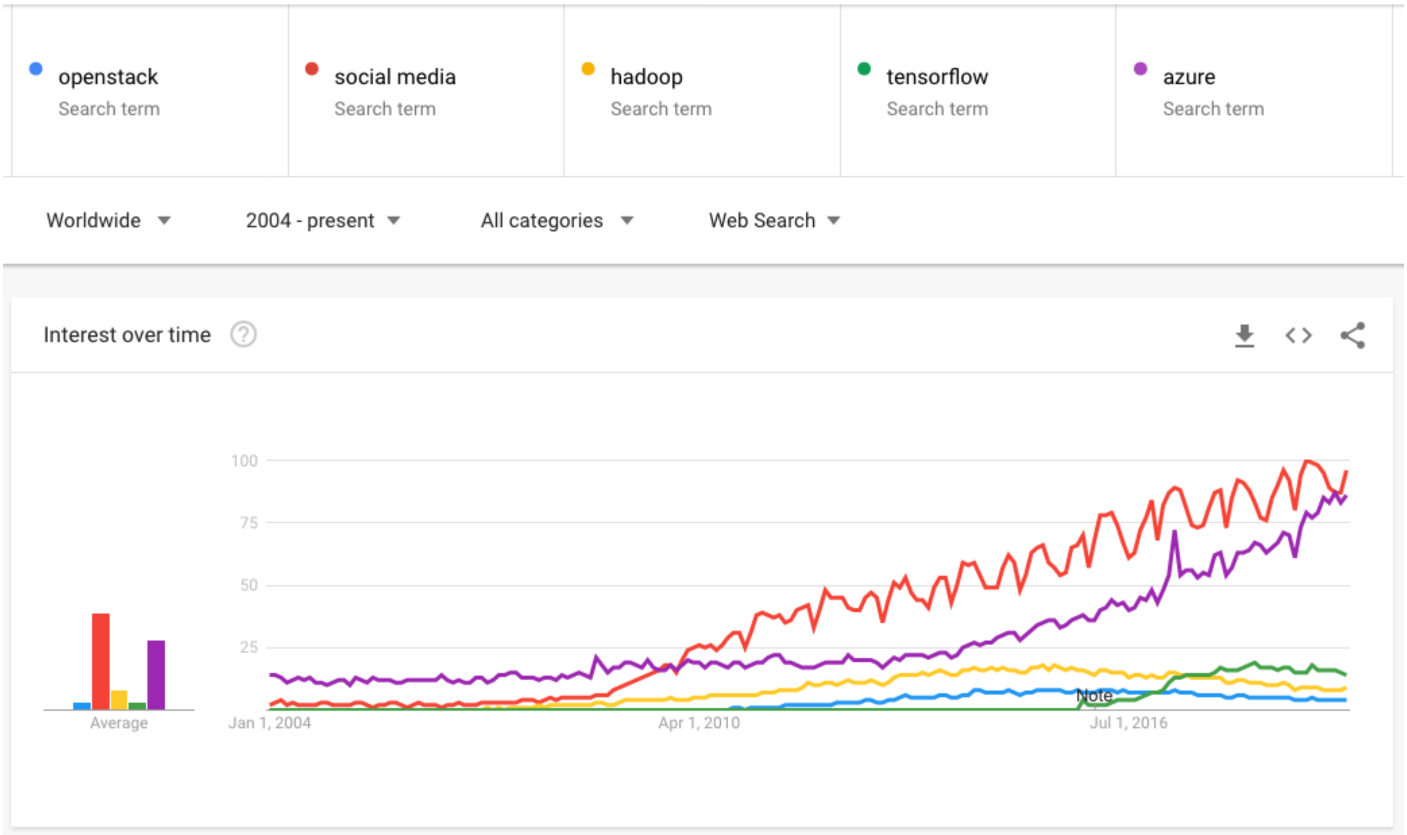
● social media
Search term

● cloud computing
Search term

● machine learning
Search term

● azure
Search term

Worldwide ▾ 2004 - present ▾ All categories ▾ Web Search ▾



● openstack
Search term

● social media
Search term

● hadoop
Search term

● tensorflow
Search term

● azure
Search term

Worldwide ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Interest over time

- In the open-source world my impression is that the OpenStack and Hadoop ecosystems have peaked
- In general IT infrastructure building is still loosing clout e.g. in university IT depts
- Public clouds are still growing and more and more targeting the public sector

In short:

††† ***Private clouds are dead*** †††

On the other hand:

- Privacy concerns have increased both in general and in academia: GDPR
- In academia, reproducibility and provenance concerns have increased: FAIR
- The EU is determinedly supporting FAIR/GDPR and European technology
- Rising demand from researchers for long-term, trusted, on-premise data-infrastructure
- Not clear who should address this

ScienceData - a data store for scientists

Researcher-oriented functionality

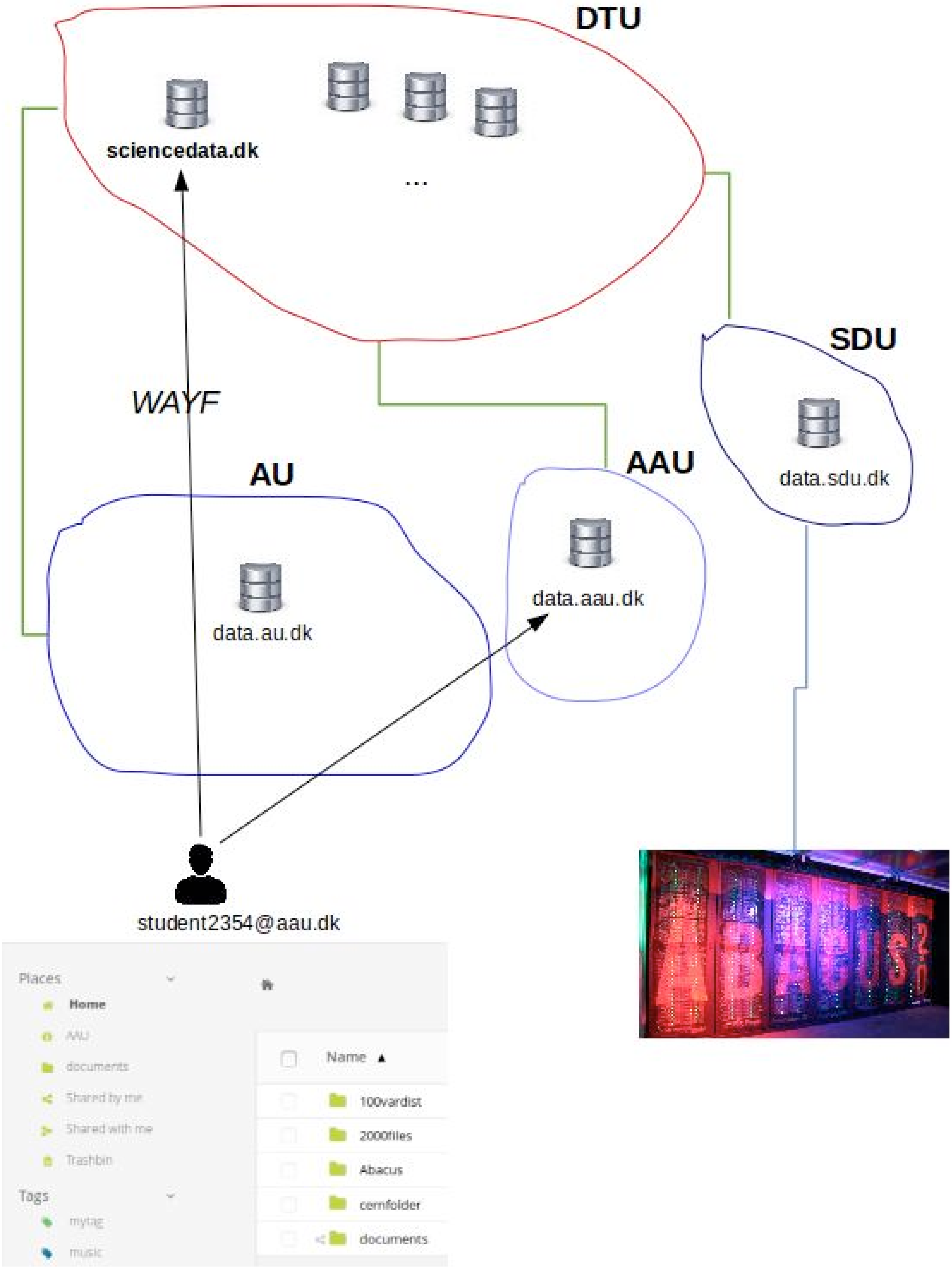
- **Take control:** Store, share, publish
- **Plan, share, reach out:** Websites, blogs, documentation, lab notes
- **Organize:** Tag your folders, mark up your data
- **Make your data live:** Process, analyze: Connect with computing resources
- **A safe home**: Finished university, got a postdoc position abroad? No worries, your data stays put
- **Never run out of space**: Research data valuable to you and your peers will always find a home here

Research-oriented architecture

- Must be configurable/extensible by standard FOSS hacking
- Must support ten-thousands of users
- Must support unlimited scaling of storage
- Must be open source and free of licenses
- Must support end-to-end encryption
- Must support distributed deployment
- Each participating site must be able to function as a stand-alone service, independently of central services
- Participating sites must be able to control where the data of their users is stored (e.g. only on-premise at their home site)
- It must be possible to migrate users from one site to another

Technical choices

- HTTP first and only
- Horizontal scaling by HTTP redirects and proxying
- FreeBSD, ZFS, ownCloud/Nextcloud
- Automated provisioning
- Sync-based user backup/replication


DTU
sciencedata.dk
...
SDU
data.sdu.dk
WAYF
AU
AAU
data.aau.dk
data.au.dk
student2354@aau.dk
Places
Home
AAU
documents
Shared by me
Shared with me
Trashbin
Tags
mytag
music
Name ▲
100vardist
2000files
Abacus
cernfolder
documents

User-contact matters - case story 1

- Feature request **number one** from researchers:

> I want to be able to share data with my local research group

- Feature request **number two** from researchers:

> I want to be able to share data with my research group - which is spread across institutions

- Feature request **number three** from researchers:

> I want to be able to share data with my research group - which includes members from industry

- Feature request **number four** from researchers:

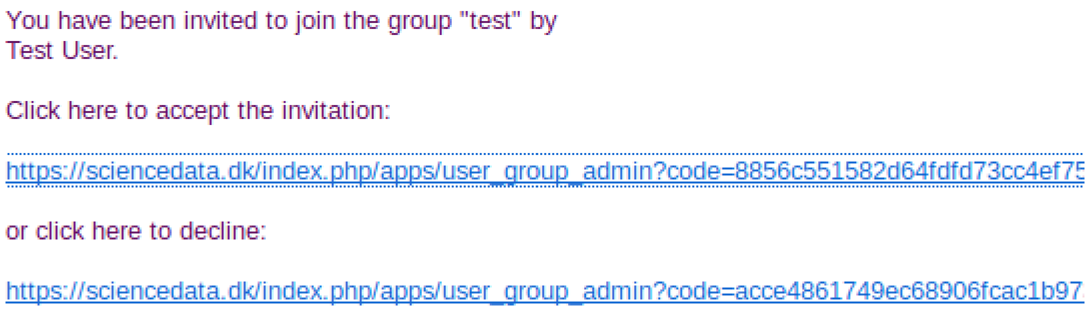
> I want to be able to share data and collaborate with anyone I choose - including colleagues in Malaysia

Implementation

Groups used as central concept - group owners legally responsible for external group members.

Sharing via email

When clicking on a group, the appearing dialog features an "Invite via email" button. To invite a collaborator to join the group, click this button, type in the email address of the collaborator and click "Send". This triggers the sending of an email containing an invitation link. You may also type in a comma-separated list of email addresses.



You have been invited to join the group "test" by
Test User.

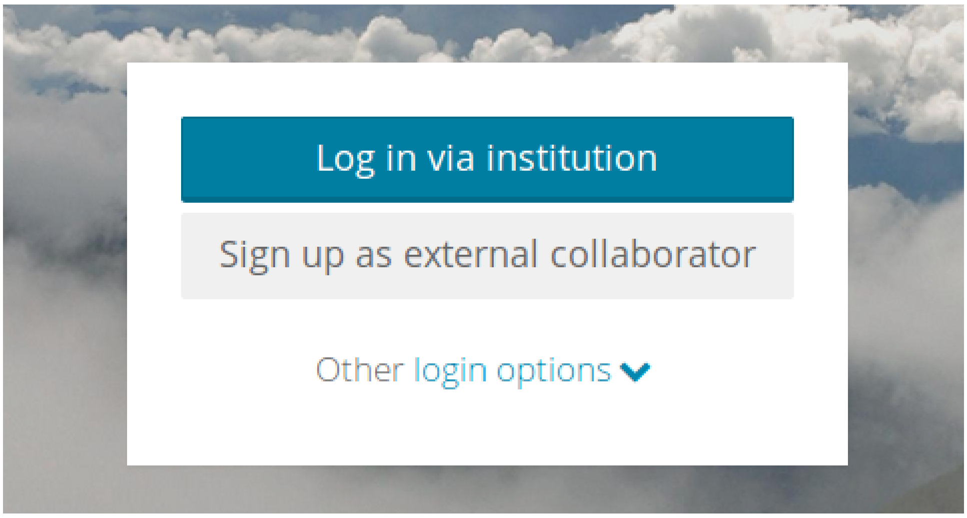
Click here to accept the invitation:

https://sciencedata.dk/index.php/apps/user_group_admin?code=8856c551582d64fdfd73cc4ef75

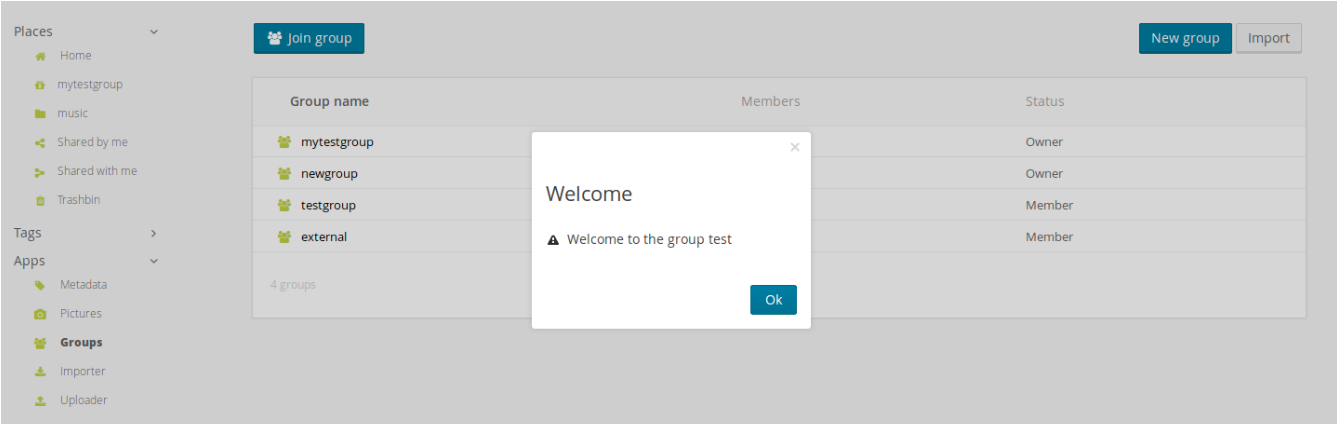
or click here to decline:

https://sciencedata.dk/index.php/apps/user_group_admin?code=acce4861749ec68906fcac1b97

When the recipient visits the link URL, he is asked to either log in or sign up for an external collaborator account.



After signing in or signing up, he is added to the group.



Group name	Members	Status
mytestgroup		Owner
newgroup		Owner
testgroup		Member
external		Member

4 groups

Welcome

Welcome to the group test

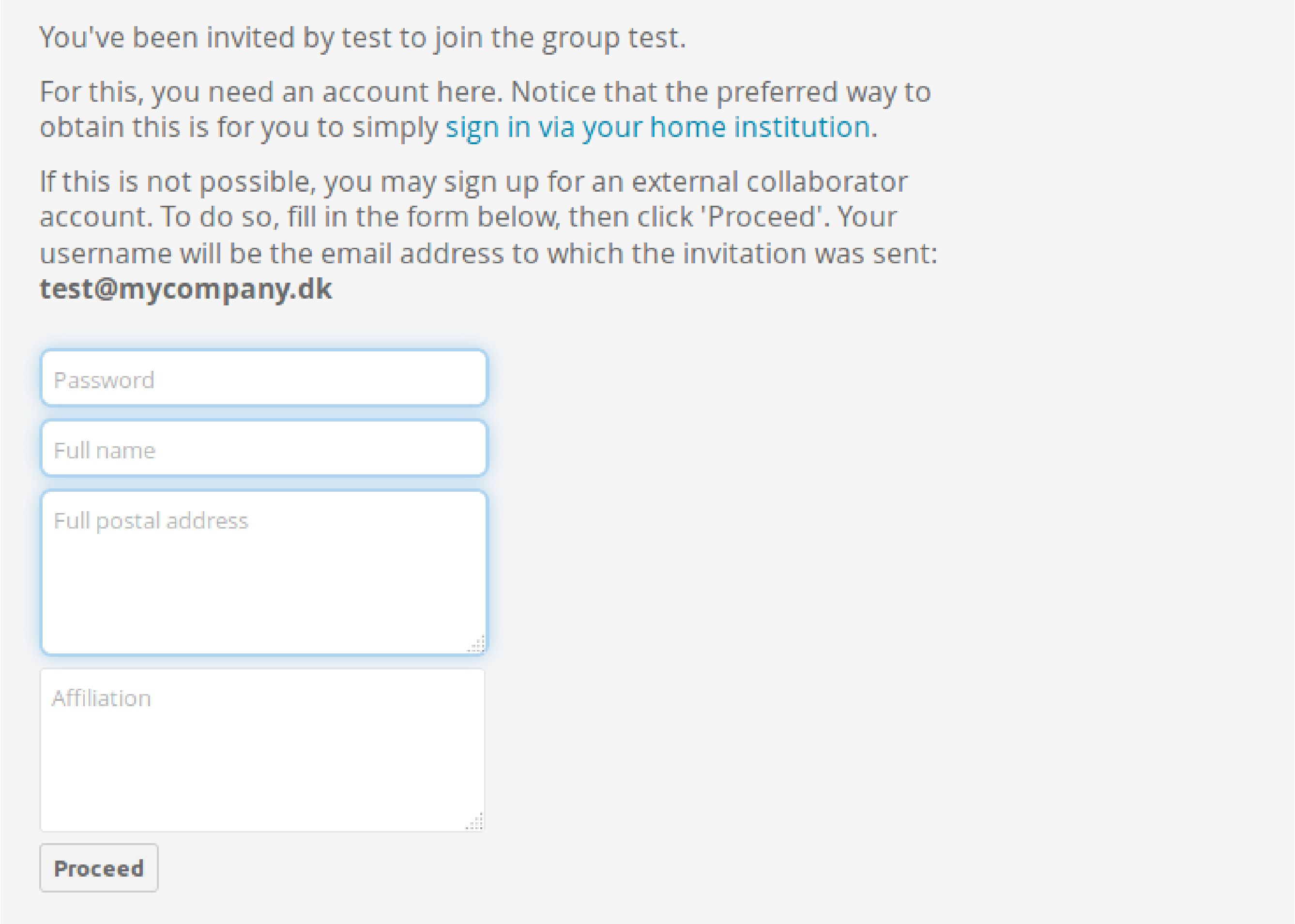
Ok

Sharing with external collaborators

First, notice that employees and students at European research/education institutions can typically log in via eduGAIN and are automatically granted an account.

Collaborators are considered external if they do not have a research/education institution affiliation or their institution is not participating in eduGAIN.

For more information, see our user agreement.



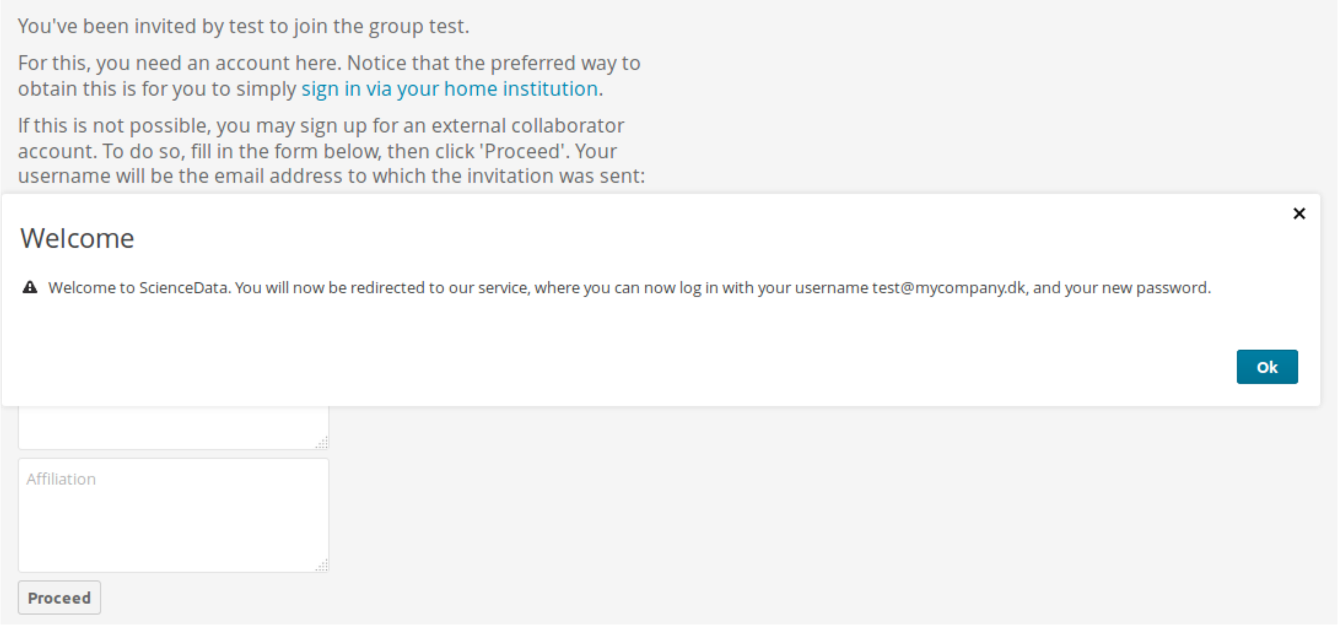
You've been invited by test to join the group test.

For this, you need an account here. Notice that the preferred way to obtain this is for you to simply sign in via your home institution.

If this is not possible, you may sign up for an external collaborator account. To do so, fill in the form below, then click 'Proceed'. Your username will be the email address to which the invitation was sent: **test@mycompany.dk**

When clicking on an emailed group invitation link, such a collaborator can click "Sign up as external collaborator". This will take him to a form for choosing a password and entering contact details.

Once completed, he must click "Proceed". This will cause an account to be created and the account holder to be added to the group in question.



You've been invited by test to join the group test.

For this, you need an account here. Notice that the preferred way to obtain this is for you to simply sign in via your home institution.

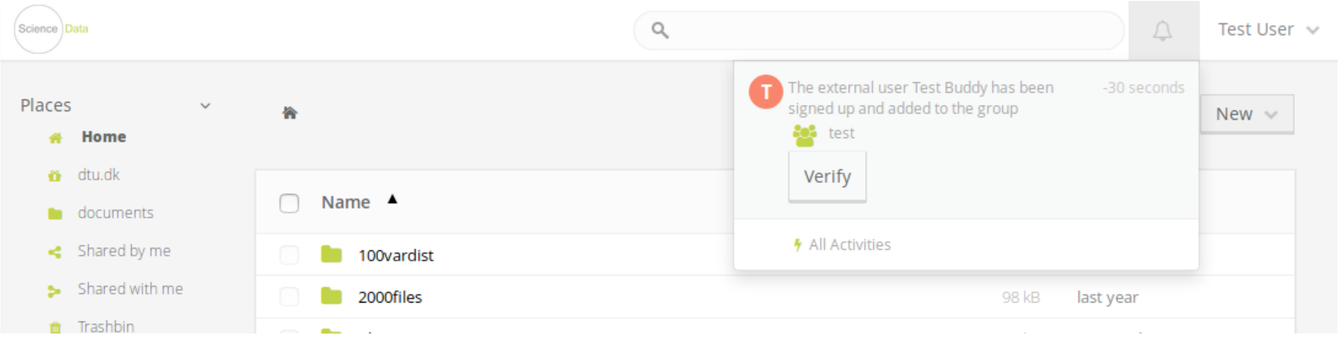
If this is not possible, you may sign up for an external collaborator account. To do so, fill in the form below, then click 'Proceed'. Your username will be the email address to which the invitation was sent:

Welcome

Welcome to ScienceData. You will now be redirected to our service, where you can now log in with your username test@mycompany.dk, and your new password.

After this, the new account holder can log in with his chosen password.

The group owner will receive a notification, asking her to validate the contact details

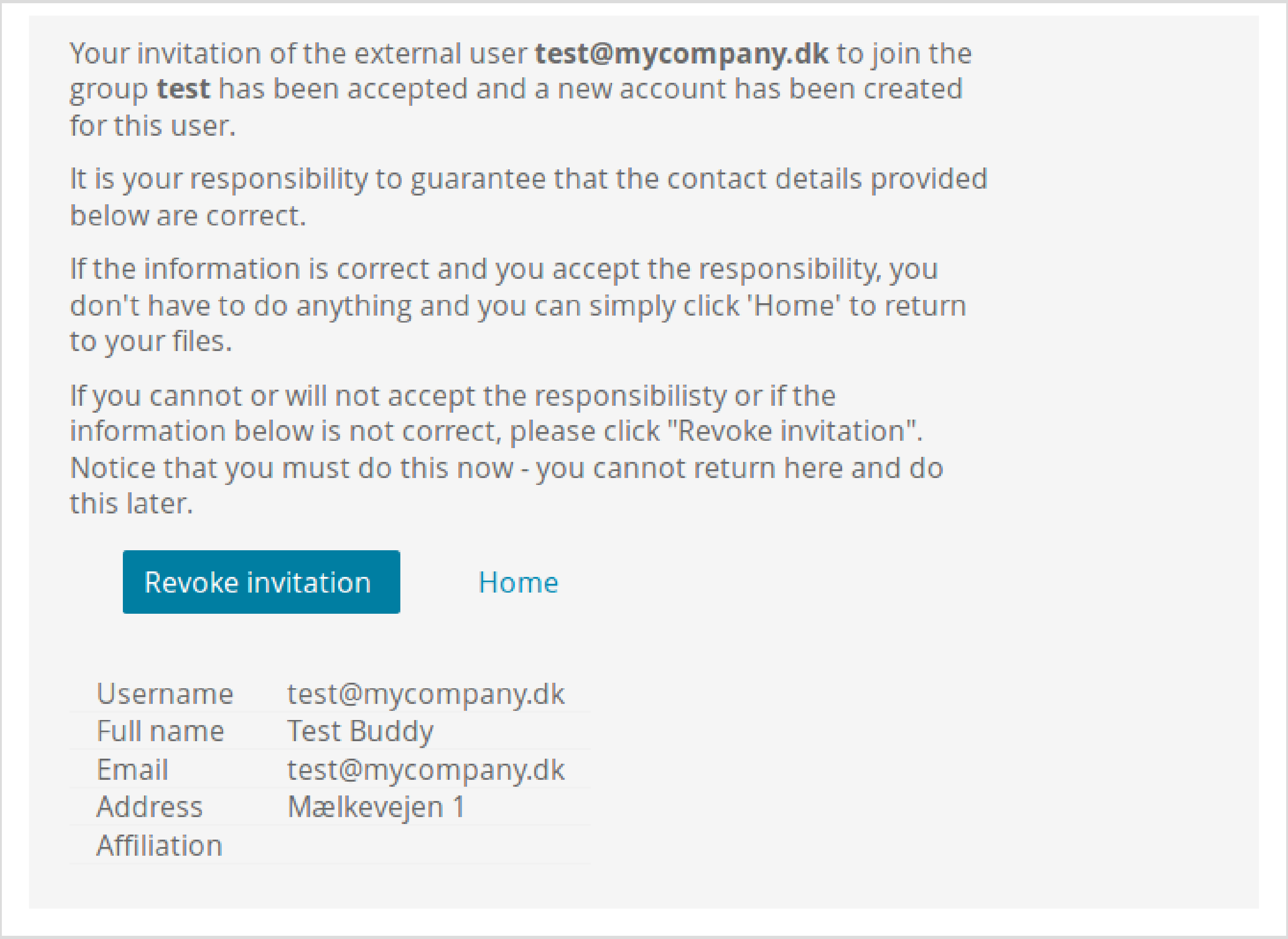


The external user Test Buddy has been signed up and added to the group

-30 seconds

test

and allowing her to revoke the created account if the contact details are not correct.



Your invitation of the external user **test@mycompany.dk** to join the group **test** has been accepted and a new account has been created for this user.

It is your responsibility to guarantee that the contact details provided below are correct.

If the information is correct and you accept the responsibility, you don't have to do anything and you can simply click 'Home' to return to your files.

If you cannot or will not accept the responsibilisty or if the information below is not correct, please click "Revoke invitation". Notice that you must do this now - you cannot return here and do this later.

Username	test@mycompany.dk
Full name	Test Buddy
Email	test@mycompany.dk
Address	Mælkevejen 1
Affiliation	

User-contacts matters - case story 2

- Feature request

> I want to be able to manage the storage space of my research group and keep data when group members leave

Implementation

Groups again used as central concept - with group owners controlling and accounted for the storage used by group members

The first time a user visits https://sciencedata.dk/, he is presented with a setup dialog, where he will choose site, backup policy and initial group memberships. The possibilities offered will depend on the participation-level of his home institution.

After logging in, the user will be redirected to his home site, where he has access to

- his home storage
  - this will be x GB free of charge (with x depending on his home institution)
  - if he uses above this, he will be personally charged (via PayPal)
  - if he does not pay, his access will be read-only until he is again below his free quota
- a personal group folder for each of the groups he is a member of and which is providing group storage (decided by the group owner)
- any number of shared folders, shared with him as an individual or group member

Data in group folders belongs to and is billed to the group owner, e.g. a university (service account)

Going global - CS3MESH4EOSC

Scope

We're not alone!

- 7 sister services across Europe (and in Australia) operated by people sharing our vision and serving hundreds of thousands of academic users
- CS3 concerences since 2014

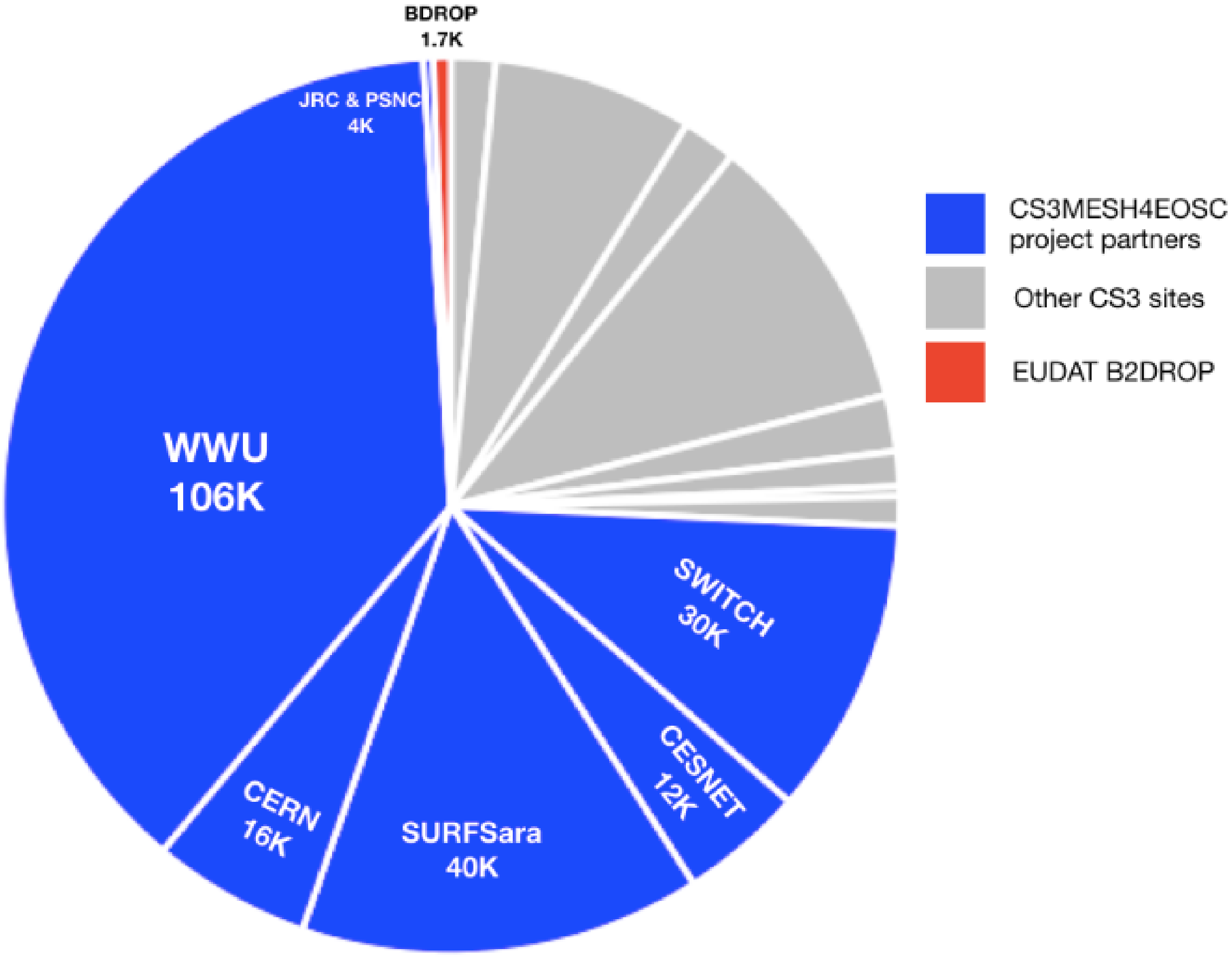
CS3Mesh4EOSC will combine our services into an

> **interoperable, pan-European mesh of data and higher-level services, which will allow friction-free collaboration between all European researchers, without requiring these researches to relocate their data.**

CS3Mesh4EOSC elevates the ideas of ScienceData to the European level:

> - the ability to form collaborative groups composed of domestic and remote users;
> - the ability to use toolsets available on a remote EFSS installation as if they were available locally;
> - access of locally and remotely stored data on the sites in the same collaborative workflow, without requiring, as a prerequisite, to export data to remote systems to reach functionality;
> - extension of local group definitions to natively include remote users;
> - maximal redeployability of relevant apps and services;
> - full metadata awareness in the research workflows.

CS3 Sites (number of users)



DeIC's contributions

DeIC will lead the **"Open Data Systems"** task (T4.2).

This will effectively turn ScienceData into a publishing/archiving platform, featuring:

- organization of research data via tags and metadata
- persistent identifiers for datasets (DOIs)
- persistent identifiers for researchers (ORCID)
- integration with established open data registries via OAI-PMH
- easy expunging to other open data repositories such as Zenodo

CS3MESH4EOSC is an EOSC project

Results will be shared across the consortium and made available via the EOSC Hub and in the Nordics via EOSC Nordic.

In particular, users can look forward to:

- integration of data science environments and compute resources (T4.1)
    - Jupyter notebooks, HPC
- collaborative document editing (T4.3)
    - Office applications

Conclusion

We have an interesting challenge ahead of us:

> "Keep European research data under European control"

It is a huge task and not one that can be addressed at institutional, or even national level.

The rationale for NRENs and others to engage in building infrastructure for research/academia still exists, but we don't stand a chance against Azure, AWS and Google if we don't find those hands, *build* something and build it together.

I encourage everyone interested to join and use the CS3MESH once it becomes operational.

In the Nordics, already now, you can join ScienceData with a local node and with time automatically join CS3MESH and EOSC.