

Document Description	
Project Name:	Shared Storage for Active Data
Acronym (if applicable):	IRLDAT
Targeted Action:	Work Project 3
Title of Document:	The Design and Building of a Pilot National Data Storage Solution for Active Research Data
Dissemination level:	Open
Nature of Document:	Report
Author(s) (Name and institution):	Paul Skelton and Fred Clarke University College Dublin
Contributor(s):	Helen Marie McNamara HEAnet, Michael Kugel South East Technological University, Brian Clayton University College Cork, Niall Wilson ICHEC
Delivery Date:	31.10.2023
Due Date:	31.10.2023
Version:	v4

Internal Reviewer Details	
Reviewer(s):	Roberto Sabatino, HEAnet, Barry O'Sullivan University College Cork
Approved by:	IRLDAT Project Board
Date:	30.11.2023

Lay Abstract
The IRLDAT project aims to design a national infrastructure to support the active part of the research data lifecycle. This report summarises the design decisions that were taken for the technical setup of the pilot, including technology choices (Nextcloud vs Owncloud,

AAI) and architectural setup (central, distributed, federated). The decisions were made following workshops with experts from EUDAT and SURF in the first few months of the project. The choices made are solely for the purposes of the pilot and to gain experience, and the final recommendation for a production quality service may differ based on the experience gained.

Revision History

Issue Date	Version	Comments
15.09.2023	V1	Table of content defined
02.10.2023	V2	First version of the document
16.10.2023	V3	Second Version
25.10.2023	V4	Final version

Table of Contents

<u>REVISION HISTORY.....</u>	<u>2</u>
Introduction	4
Executive Summary.....	4
Requirements, Initiation & Planning.....	5
Selection of Research Groups for the Pilot	5
Distributed v Centralised.....	5
Federated Access and ID Authentication	6
System Architecture	6
ICHEC Implementation	6
University College Cork Implementation	7
South East Technological University Implementation	7
NextCloud v OwnCloud	7
University College Cork IRLDAT NextCloud Instance	7
B2ACCESS and Identity Access Management	8
Glossary & Acronyms	11

Introduction

The IRLDAT project aims to design a national data storage infrastructure to support the active part of the research data lifecycle. The Work Package (WP3) phase is to design and create a pilot system to support active research datasets. The pilot includes a sample of research projects that are representative of the research community.

This document is the Work Program 3 deliverable of the IRLDAT project and is part of a series of reports outlining each work package. We recommend you read this as part of the series.

The report summarises the decisive steps taken and the suggested architecture solutions with their rationales.

Executive Summary

The IRLDAT project aims to design a national data storage infrastructure to support the active part of the research data lifecycle. IRLDAT members along with EUDAT and SURF collaborated to define a system design that would be most appropriate for this aim.

The key end user requirements were initially established:

- Store active research data in various formats
- Ease of use and ability to access from anywhere on any device
- Simplify collaboration with partners across multiple institutions and industries
- Fully featured solution providing modern cloud style functionality

The group considered a number of key questions, e.g., system architecture, hardware/software required, storage demands, local or federated access and identity access management. Each institution decided on the most appropriate implementation for their research groups.

Three pilot systems were developed and tested by ICHEC, SETU and the UCC. Research groups data are assigned to one of the three systems, the position of the data was selected based on the locations of the host institutions and for the diversity of demands the researchers would make on the system during the pilot. Their decisions have enabled a hybrid environment between physical, local virtual servers and public cloud.

UCD provided project management, coordination and oversight for this Work package. HEAnet provided identity access management expertise and sourced European expertise as part of EUDAT, from Jülich Supercomputing Centre in Germany, Surf in the Netherlands and Cineca in Italy.

Requirements, Initiation & Planning

Initially, the project groups engaged in an extensive information-gathering process. Both EUDAT and SURF, members of the IRLDAT working group, were heavily involved. The technical expertise from both European groups was invaluable in moving the project forward, especially relating to system architecture and its implementation.

Several key questions were considered:

1. What type of hardware would be most appropriate for the systems?
2. What type of software would allow researchers to collaborate more effectively?
3. How much storage capacity would be needed for each of the systems?
4. Should the systems be connected with a bespoke network and then connect each of the systems directly?
5. Which would be better, federated access over separate local institutional access?

Initial discussions centred on whether each system should be set-up purely for the use and benefit of their own research institute users. This would make installation and implementation simple and straightforward.

However, the big disadvantage was that it didn't enable a straightforward national connection. Therefore, a Federated solution, using the EUDAT authentication method B2ACCESS, was chosen.

Selection of Research Groups for the Pilot

A test instance was developed and used by the host universities' research groups. The research groups originally selected by EUDAT members were granted access to the national RDM storage infrastructure along with a select group of additional research groups from each of the host institutes.

These research groups were based on the location of the host institutions and the diversity of demands the researchers would make on the system during the pilot. Some might only have limited storage requirements whereas others would have larger storage needs. Others would be collaborating with several other researchers nationally and/or internationally.

Distributed v Centralised

Early discussions focused on whether it would be best to have a large, centralised instance, such as that currently used by EUDAT, or to go with a distributed system architecture, ideally among the IRLDAT institutes.

It was decided to go with distributed instances in three locations using federated access as it would enable each of the institutes to gain experience in working with both physical and virtual systems and how they combine with the storage software required.

Federated Access and ID Authentication

In line with EUDAT recommendations, and the distributed architecture of the pilot system, federated access was decided upon. It would use both eduGAIN and Edugate for ID authentication for the pilot. It is the authentication method advocated and used by EUDAT and enables IRLDAT instances to be easier to introduce as key nodes for EUDAT in the future.

System Architecture

Each of the research institutes decided on the system they would use during the pilot:

- UCC: Local on-premise storage with virtual servers on local compute hosts
- SETU: virtual servers
- ICHEC: AWS implementation

These decisions enabled a hybrid environment between physical, local virtual servers and public cloud.

ICHEC Implementation

Nextcloud Host: 1 x AWS r6i.large instance (2 vCPU + 16GB RAM)

OS: Debian 11

Nextcloud version 26

Nextcloud Primary Storage: AWS S3 Object Store

University College Cork Implementation

NextCloud hosts: 2 ESX virtual machines, running CentOS 7. One front end vm, one Database VM.

NextCloud version 25.

Storage is a Compellent SAN

South East Technological University Implementation

Nextcloud host: 1 VM on Openstack (8 vCPU, 8 GB RAM, 40 GB boot drive, 100 GB primary data storage, 100 GB backups storage, 600 GB external project storage.

OS: Ubuntu 22.04 LTS

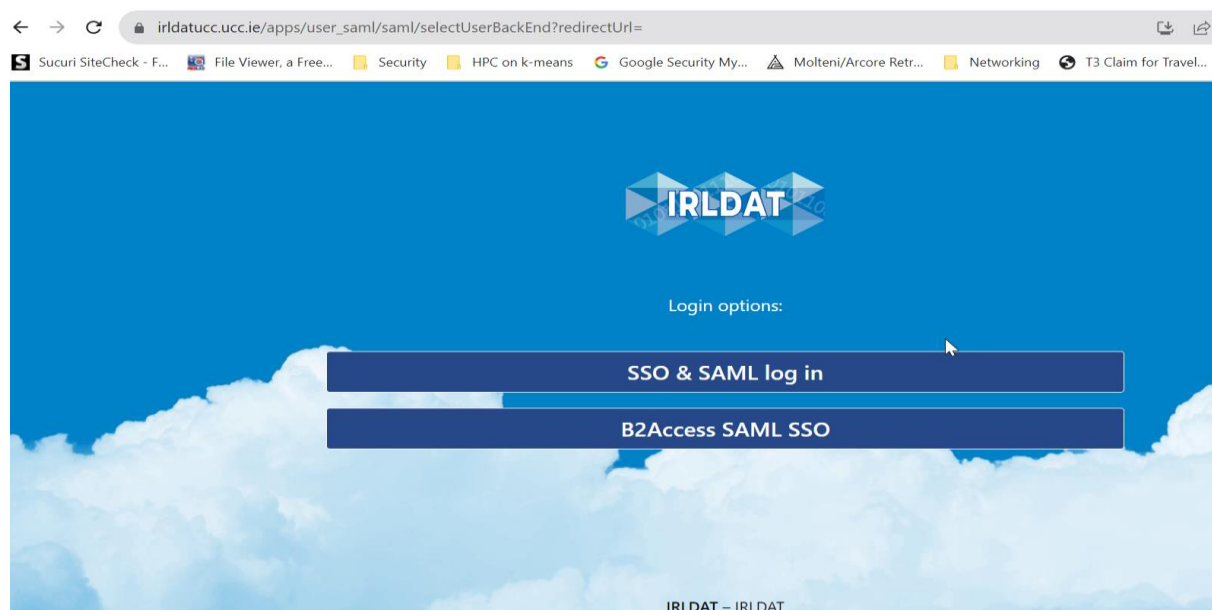
Nextcloud version 26.0.5

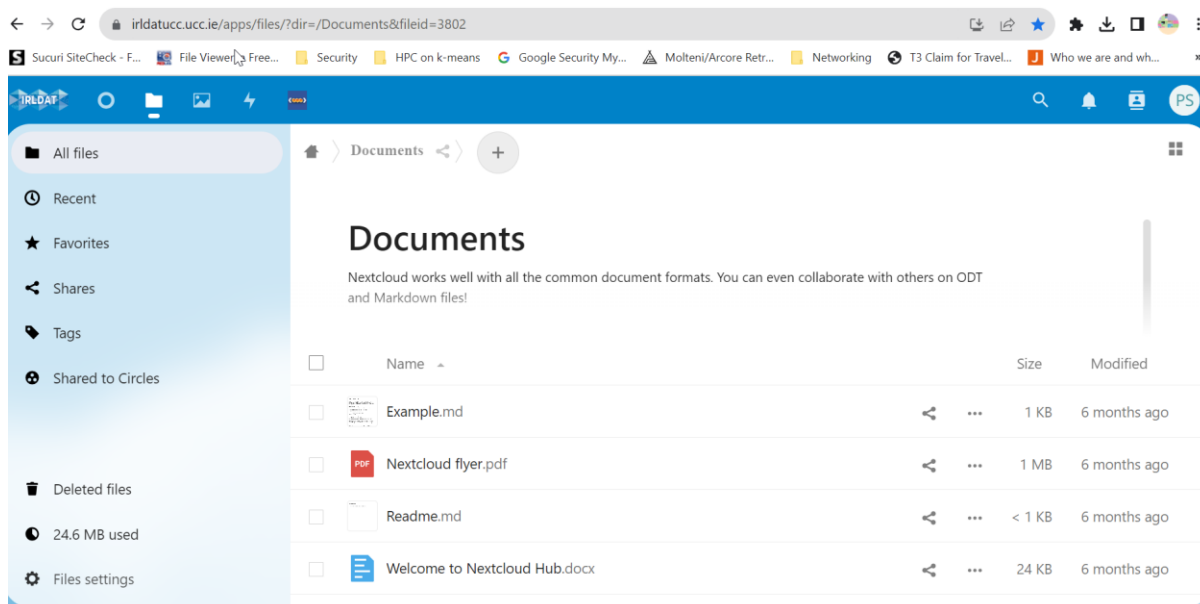
NextCloud v OwnCloud

After discussions with both EUDAT and SURF, NextCloud was chosen as the collaborative software. Despite OwnCloud being a more mature product and currently in use by SURF, it was felt that NextCloud was better aligned with EUDAT members who were already using it.

University College Cork IRLDAT NextCloud Instance

Below are screenshots from the UCC IRLDAT implementation as an example of what the implementation looks like.





B2ACCESS and Identity Access Management

B2ACCESS (<https://b2access.eudat.eu/>) B2ACCESS is a federated cross-infrastructure authorisation and authentication proxy for user identification and community-defined access control enforcement. It allows users to authenticate themselves using a variety of credentials providing federated access and single-sign-on to services and service providers in a trusted way. It was chosen as there is good collaboration with the EUDAT technical engineers and will be useful in relation to future discussions about a national instance.

B2ACCESS is integrated with the European Open Science Cloud (EOSC) and allows seamless access to services from other providers in Europe.

Edugate and eduGAIN

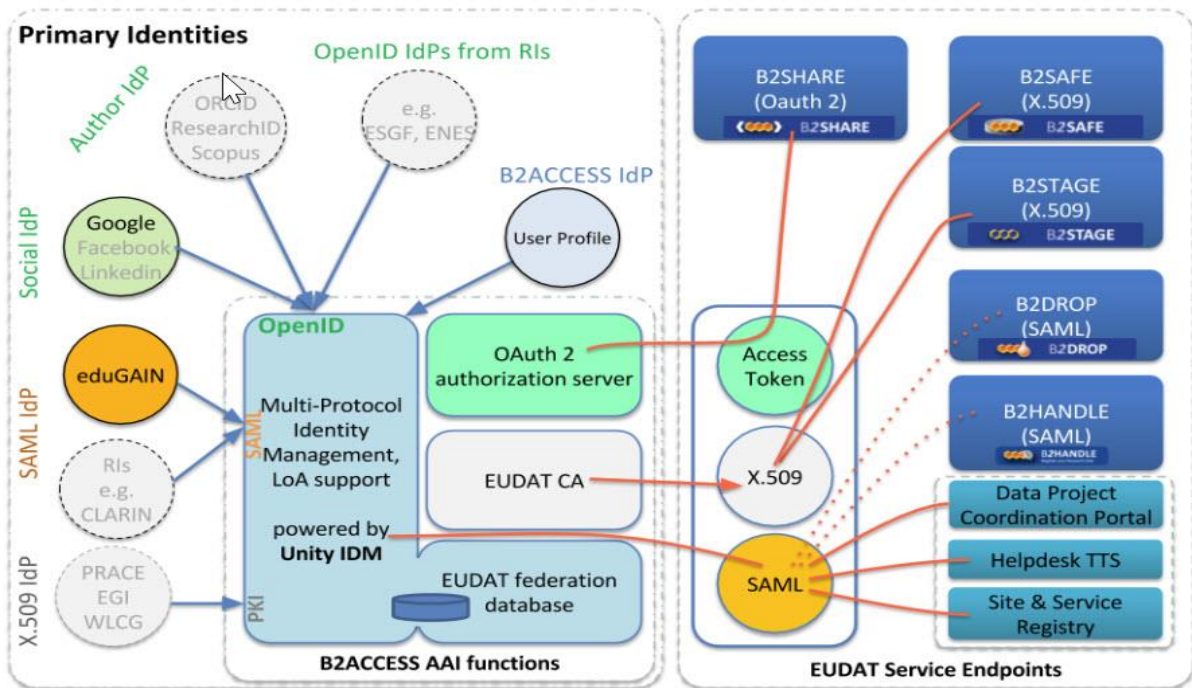
The pilot group advocated using the Heanet-provided service Edugate as the identity access management system for the IRLDAT project.

Edugate focuses on enabling secure access to various online resources, services, and applications within the academic and research community. It uses SAML as the Single Sign-On format, an XML-based framework for authentication and authorisation between two entities: a Service Provider and an Identity Provider.

eduGAIN

B2ACCESS uses eduGAIN which acts as a global inter federation service for the research and education community. As well as inter federation, eduGAIN also provides the following;

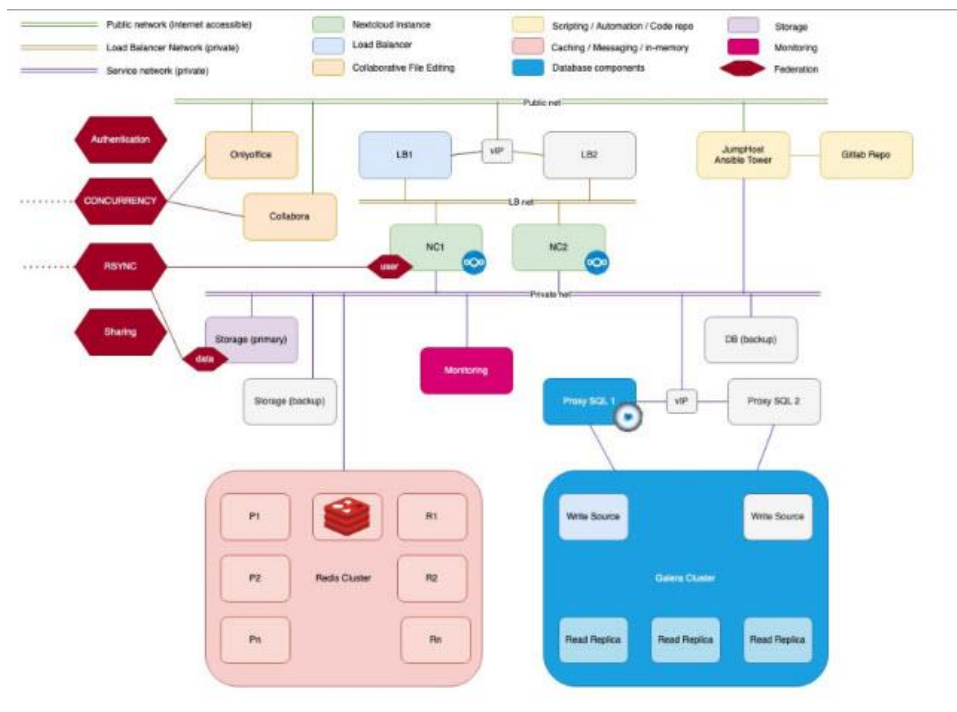
- Access to Global Services: eduGAIN enables users from one institution or federation to access resources and services provided by organisations and institutions in other federations without the need for separate logins or accounts.
- Metadata Exchange: eduGAIN maintains a metadata exchange service that provides information about identity providers and service providers, ensuring secure and standardised communication between federations.



Recommended Open Architecture

One of the benefits of the EUDAT /IRLDAT distributed installations with federated authentication is many of its layers are open to site specific preferences. Hardware for networking, server and storage (physical, virtual or IaaS cloud) operating system are all open to site specific preferences.

The group decided to choose the NextCloud platform for file storage and access via Edugate and eduGAIN. There were a number of reasons behind the decision. IRLDAT had bought the relevant support licensing and so were able to access excellent documentation on setting up both authentication and authorisation. NextCloud also has a good development community and is very pro-active with its security patching. However, the most significant reason for installing NextCloud as our collaborative software was that it was already being used by EUDAT and made using EUDAT B2DROP software more straightforward in terms of implementation. The following diagram gives some indication as to what a more advanced research storage solution might look like.



Glossary & Acronyms

B2ACCESS

This is a federated cross-infrastructure authorisation and authentication proxy for user identification and community-defined access control enforcement. It allows users to authenticate themselves using a variety of credentials providing federated access and single-sign-on to services and service providers in a trusted way.

Cineca Consortium Italy: made up of 117 members: the Italian Ministry of Education and Merit, the Italian Ministry of Universities and Research, 70 Italian universities and 44 Italian National Institutions and agencies.

eduGAIN: Education Global Interoperability and Authentication Infrastructure

This is a service that facilitates the exchange of identity and authentication information between different identity federations and institutions worldwide. eduGAIN acts as a global inter federation service for the research and education community.

Edugate: Education Global Authentication and Trust Environment

Edugate is an Irish educational institutional identity federation aimed at providing a secure and trusted authentication and authorization infrastructure for the global research and education community.

EUDAT: A collaborative data infrastructure service that shares data across borders and disciplines

HEAnet: Ireland's National Education and Research Network.

ICHEC: Irish Centre for High-End Computing

IRLDAT: The EUDAT pilot implementation for Ireland.

Jülich Supercomputing Centre

SETU: South East Technological University

SURF: A cooperative association of Dutch educational and research institutions.

UCD: University College Dublin

UCC: University College Cork