

DATA REPOSITORIES

2nd Data Analysis Methods (DAMe) Workshop

14.03.2019 | Sander Apweiler | JSC



Member of the Helmholtz Association

1. Data repository

2. EUDAT

- 2.1 overview
- 2.2 service suite

3. Hands-on



Data repository



WHAT IS A DATA REPOSITORY?

A data repository is a software suite to store and aggregate data sets from multiple sources. The data can be stored unstructured with additional metadata, describing the data itself. The access to the data can be limited, if needed, before the data is reused by other researchers.

 \Rightarrow A data repository is a collection of (domain specific) data sets, and additional metadata, for secondary use.



WHAT IS A DATA REPOSITORY NOT?

- An extension for your hard drive.
- A storage for volatile data.
- A storage for sensitive/personal (raw) data.



POSSIBLE CONSTRAINTS

- The data repository is public/private.
- The access to the data must be restricted.
- The data has an embargo date.
- Data protection issues in data.
- The storage is subject to legal restrictions, like storage for medical data.
- The availability is not guaranteed to the retention period of the data.



REASONS FOR USING

- A lot of initiatives/projects/guidelines/... which promote "open data" and shared data like:
 - (GO)FAIR
 - RDA
 - "Gute wissenschaftliche Praxis"
 - Funding agencies like BMBF, EU Commission
 - Publications
- You want to share your results and data with other researchers.
- They support you to hold the retention period.



- Findable
- Accessible
- Interoperable
- Reuseable



FAIR¹ principles:

- Findable
 - F1. (meta)data are assigned a globally unique and persistent identifier
 - F2. data are described with rich metadata
 - F3. metadata clearly and explicitly include the identifier of the data it describes
 - F4. (meta)data are registered or indexed in a searchable resource



¹https://www.doi.org/10.1038/sdata.2016.18

FAIR¹ principles:

- Findable
 - F1. (meta)data are assigned a globally unique and persistent identifier
 - F2. data are described with rich metadata
 - F3. metadata clearly and explicitly include the identifier of the data it describes
 - F4. (meta)data are registered or indexed in a searchable resource

Accessible

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1. the protocol is open, free, and universally implementable
- A1.2. the protocol allows for an authentication and authorization procedure, where necessary
 - A2. metadata are accessible, even when the data are no longer available

¹https://www.doi.org/10.1038/sdata.2016.18



FAIR² principles:

- Interoperable
 - 11. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
 - 12. (meta)data use vocabularies that follow FAIR principles
 - 13. (meta)data include qualified references to other (meta)data



²https://www.doi.org/10.1038/sdata.2016.18

FAIR² principles:

- Interoperable
 - 11. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
 - 12. (meta)data use vocabularies that follow FAIR principles
 - 13. (meta)data include qualified references to other (meta)data
- Reuseable
 - R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards



²https://www.doi.org/10.1038/sdata.2016.18

DATA REPOSITORIES

- Arctic Data Center For NSF³ Polar Programs
- NOMAD CoE Materials science
- PANGAEA Earth & environmental science
- Publisher
 - Springer Nature
 - dataplanet
- Community based solutions
 - LTER Network Data Portal
- Research center based solutions
 - Zenodo

³National Science Foundation



LOOKUP SERVICES

- CESSDA Consortium of European Social Science Data Archives
- DataCite
- DataONE Data Observation Network for Earth
- re3data Registry of research data repositories



EUDAT: Overview



EUDAT

EUDAT is a pan-European e-infrastructure solution for pan-European research infrastructure data challenges.

- All research infrastructures are facing data challenges
 - Where to store the growing amount of data?
 - How to find the data?
 - How to make the most of the data?
- \Rightarrow Solutions are needed at pan-European level.
 - We need to promote synergies
 - Some services are common to many communities
 - Costs and investments can be optimised
- ⇒ Better integration of e-infrastructures and research infrastructures can be achieved.



EUDAT(2020) PARTNERS



14.03.2019

EUDAT CDI

- Started October 2016
- Agreement among 26 partners
- Major research organisations, data and computing centres
- Agreement: Sustain EUDAT for the next 10 years







EUDAT: Service Suite



EUDAT SERVICE SUITE - OVERVIEW





EUDAT SERVICE SUITE - B2ACCESS





B2ACCESS

- EUDAT Authentication and Authorization Infrastructure
- Based on Unity 2.4.2
- CA for short lived X.509 (testing)
- Acts as a proxy IdP
- Users: 2002



user registration 21.06.16 - 12.03.19



B2ACCESS - ACCOUNTS

- eduGain (SAML)
- Community IdPs (SAML)
 - ARIA
 - Clarin
 - EGI
 - ELIXIR
- Social IdPs (OAuth2)
 - Facebook
 - GitHub
 - Google
 - Microsoft Live
 - ORCID
- Local B2ACCESS accounts
 - Username
 - Certificate with optional username





B2ACCESS - LOGIN SCREEN



B2ACCESS - ATTRIBUTE RELEASE SCREEN



B2ACCESS SAML web authentication

A remote service has requested your authentication

https://fsd-cloud48.zam.kfa-juelich.de/index.php/apps/user_saml/saml/metadata Address: https://fsd-cloud48.zam.kfa-juelich.de/index.php/apps/user_saml/saml/acs

The following information will be sent to the requesting service

Your identity			
Fully anonymous ide	ntifier		
0			
Details of exposed	information		
٥			
Remember the sett	ings for this service and do not show this	dialog again	
	Confirm Decline L	ogin as another user	
			JÜLICH
Helmholtz Association	14.03.2019	Slide 22	Forschungszentrum

14 03 2019

B2ACCESS - ATTRIBUTE RELEASE SCREEN

Details of exposed information							
Note that in any case your credential (password, private key,) will NOT be exposed.							
urn:oid:2.5.4.49: /C=EU/O=EUDAT/OU=B2ACCESS/CN=							
unity:persistent:							
distinguishedName: /C=EU/O=EUDAT/OU=B2ACCESS/CN=							
firstname surname: Sander Apweiler							
email: sa.apweiler@fz-juelich.de							
Remember the settings for this service and do not show this dialog again							
Confirm Decline Login as another user							



EUDAT SERVICE SUITE - B2DROP







For whom?

Researchers (including citizen scientists) who want to share their research with a small set of collaborators prior to making it available to a wider audience.

What can you do with it?

- A service to store and exchange data
- Automated desktop synchronization on many platforms



B2DROP

For whom?

Researchers (including citizen scientists) who want to share their research with a small set of collaborators prior to making it available to a wider audience.

What can you do with it?

- A service to store and exchange data
- Automated desktop synchronization on many platforms
- Move data to other EUDAT services



B2DROP - SOFTWARE

- Based on Nextcloud
 - Currently 13.0.5
- New theme for EUDAT CI
- New app called B2SHAREBRIDGE
 - Integration with B2SHARE
- user_saml module
 - Integration with B2ACCESS
 - Some own modifications



B2DROP - USAGE STATISTICS

- EUDAT catch all instance
- Users: 1813
 - 1340 in 10/2017
 - 600 in 05/2016
 - 330 in 09/2015
- Volume: 3.5 TB (15 TB available)
- Objects: 2.759.179 #
- Requests:
 - 358.000 #/day
 - 4,1 #/sec
- Daily backup to tape





B2DROP - USAGE STATISTICS

Shares

- With other B2DROP users
- By link
- Federated to other services
- Federated from other services

		Ki	nd of shares	
-				
-				
-				
-				
0	400	800	1200 1600 2000 2400 280 Shares [#]	C



• 71 uploads with B2SHAREBRIDGE

B2DROP - LOGIN SCREEN



Forschungszentrum

B2DROP - FILE LIST

G	O TO EUDAT WEBSITE	4 🖬	(005)								
	B2DROP	Т	WHAT IS B2DROP	USER GUIDE	FAQs	CONTACT					
	Alle Dateien										
٩	Aktuelle		Name 🔺						Größe	Geände	rt
*	Favoriten	\star . Int	test.txt				<	***	< 1 KB	vor 3 St	unden
<	Mit Ihnen geteilt								5 B		
<	Von Ihnen geteilt										
8	Getellt über einen Link										
۹	Tags										
Î	Gelöschte Dateien										
5 E	3 verwendet										
¢	Einstellungen									JÜL	ICI



B2DROP - PERSONAL SETTINGS

G	o to eudat website 🛛 🖌	•			÷	0
	B2DROP EUDAT	WHAT IS B20	DROP USER GUIDE FAQs CO	DNTACT		
4	Persönliche Informationen	-				
С	Sync-Clients	Sie verwenden 5 B der ve	erfügbaren Unbegrenzt			
E	Sicherheit					_
4	Aktivität	Profilbild 😃	Vollständiger Name 😃	E-Mail 🚢		
	EUDAT B2SHARE Bridge		20495dd7-1bcf-4661-a421-79298c09	sa.apweiler@fz-juelich.de		
<	Federated Cloud	S	Telefonnummer 🔒	Adresse 🔒		
			Ihre Telefonnummer	Ihre Postadresse		
		± 🖿	Webseite 🖴	Twitter 🔒		
		png oder jpg, max. 20 MB	Link https://	Twitter-Handle @		
		Gruppen				
		Sie sind Mitglied folgender	Gruppen:			
		Sprache			üı	

Forschungszentrum
B2DROP - FILE SHARING

GO TO EUDAT WEBSITE	4 🖂 🔤		९ ६ छ
B2DROP EUDAT	WHAT IS B2DROP USER	R GUIDE FAQS CONTACT	
Alle Dateien	* +	::	×
() Aktuelle	🗌 Name 🔺	Größe Geändert	test
🚖 Favoriten	★ test.txt	<pre>% *** < 1 KB vor 3 Stund</pre>	
< Mit Ihnen geteilt			
🗳 🛛 Von Ihnen geteilt		5 B	
🔗 🛛 Geteilt über einen Link			
🗞 Tags			test.txt ∰ ★ <1 KB, vor 3 Stunden %Tags Aktivitäten B2SHARE Kommentare Tellen Versionen
Gelöschte Datelen S B verwendet			sander i S sander S sander S Sander 7ech1132.bd/r .92.bfbe-12./f.4r/92.2ef0
Member of the Helmholtz Associati	ion 14.03.2019	Slide 32	Forschungszentru

EUDAT SERVICE SUITE - B2SHARE





B2SHARE

For whom?

Researchers (including citizen scientists) who want to publish their research data to a wider audience and get a unique identifier as a reference to the data.

What can you do with it?

- A service to publish and exchange data
- Store additional (community specific) metadata
- Reference the data with a PID and/or DOI



B2SHARE

For whom?

Researchers (including citizen scientists) who want to publish their research data to a wider audience and get a unique identifier as a reference to the data.

What can you do with it?

- A service to publish and exchange data
- Store additional (community specific) metadata
- Reference the data with a PID and/or DOI
- (Move data to other EUDAT services)



B2SHARE - SOFTWARE

- Based on INVENIO 3
- Own Web UI
- Additional PID/DOI handling
- Deployed with Docker
- Uses OAuth for B2ACCESS integration
- Interacts with B2DROP and B2FIND





Metadata Service

B2SHARE

- Build checksum on files and validates it periodically
 - in case of mismatches the administrator is informed
- File changes by the user create a new version
 - with own PID
 - user can switch between versions
- Records can be deleted by administrators only



B2SHARE - LOGIN SCREEN

GO T	FO EUDAT WEBSITE				
	ezshare EUDAT	arch records for	UPLOAD	Q SEARCH	*J Login
	Store	and publish	your res	search data	
	Search in public dat	asets or register a	s a user to u	oload and publish your da	ata!
		Login o	or Register		
	Create Record			Create a new record	
	Latest Records				
	Wind throw monitoring valleys 1 Aug 2017 by , ; Vegetation relieves from the the mon	itoring of wind thr	ows.		
	Wind throw monitoring uplands 1 Aug 2017 by , ; Vegetation relieves from the the mon	itoring of wind thr	ows.		K,
Aember of the Helmh	Wind throw monitoring slopes 1 Aug 2017 by .; Vegetation relieves from the the mon oltz Association	itoring of wind thr	ows.	Slide 37	

B2SHARE - PERSONAL SETTINGS



B2SHARE - LIST DRAFTS/SEARCH

publicatio	n_state:draft	Q Search
Show records from:		
All communities		•
Sort by:		
Most Recent		-
Page size:		
10		-
1 - 10 of 14 results		« 1 2 »
test3 21 Jun 2017		
<mark>test</mark> 23 Jun 2017		
<mark>license_test</mark> 22 Jun 2017		
<mark>test</mark> 3 Aug 2017		
<mark>test</mark> 27 Sep 2017		



Editing draft		
Add files	,	
	Drop files here, or click to select files	
	Add B2DROP files	
Basic fields		K
Community *	EUDAT	1 - be
Titles	title_test	
Member of the Helmholtz Associati Descriptions	14.03.2019 Slide 41	Forschungszentrum

Open Access *	True		
Embargo Date			#
License		© Se	lect License
	URL		
Disciplines			•
			O Add
Keywords			
	[O Add
Contact Email	email@example.com		
Publication Date			
	Show more details >		
	Submit draft for public	cation	
	When the draft is published it a published record's files can r	will be assigned a PID, making it publicly o longer be modified by its owner.	/ citable. But
		The draft is up to date	
er of the Helmholtz Association	14.03.2019	Slide 42	

Open Access *	True			
Embargo Date			m	
License		©	Select License	
	URL			
Disciplines			-	
l			O Add	
Keywords			0.444	
l			VAdd	
Contact Email	email@example.com			
Publication Date				
	Show more details >			- I
				I
	R ed W + 64 - 10			I
	Submit draft for publication	ation ill be assigned a RID, making it public	dy citable. Put	
	a published record's files can no	longer be modified by its owner.	ciy citable. But	
		Save and Publish		
mbolin Accession	14.02.2010	Slide 42		JULIC

Member of the Helmholtz Association

14.03.2019

B2SHARE - API

Upload with web frontend has some cons:

- Limitation in file size
- Inefficient if you create multiple uploads
- Can not be automated

REST-API solve this issue:

- Use favoured tool, e.g. curl
- Can be automated in scripts
- Is not limited to browser timeouts



B2SHARE - API COMMANDS

Commands for

- Listing all
 - communities
 - records
 - records of a community
 - files from a record
- Fetching
 - a community schema
 - the metadata of a specific record
- Searching for
 - records
 - your drafts



B2SHARE - API COMMANDS

Commands for

- Uploading
 - a file into your draft
 - metadata into your draft
- Creating a new record
- Deleting a file from your draft
- Submitting a draft for publication
- Reporting abuse
- Sending a request for access to restricted data



B2SHARE - API WORKFLOW

Common API workflow

- Identify a targeted community by listing the available ones
- Use the community identifier to fetch the community metadata schema
- Create a draft
- Upload files to your draft; one command per file
- Set the metadata to your record
- Publish your draft



EUDAT SERVICE SUITE - B2SAFE







For whom?

Researchers or communities whose repositories lack the capacity and / or funding to offer reliable storage and access services over longer periods of time or without adequate compute capacity.

What can you do with it?

- Guard against data loss in long-term archiving and preservation
- Optimize access for users from different regions
- Bring data closer to powerful computers for compute-intensive analysis



B2SAFE

- Based iRODS
- Own rules for safe replication and PID management
- Use EPIC Handle service



- Each EPIC record knows "parents" and "children"
- Starts with replication from community to first EUDAT partner
- Additional replications between EUDAT partners are possible



EUDAT SERVICE SUITE - B2STAGE





14.03.2019

B2STAGE

For whom?

Researchers or communities who need to access both large-scale data storage and high-performance computing systems and transfer their data easily between the EUDAT storage resources and remote HPC facilities.

What can you do with it?

- Transfer large data collections from EUDAT storage facilities to external HPC facilities for processing
- In conjunction with B2SAFE replicate community data sets, ingesting them onto EUDAT storage resources for long-term preservation
- Ingest computation results into the EUDAT infrastructure



EUDAT SERVICE SUITE - B2FIND





14.03.2019

B2FIND

For whom?

Researchers (including citizen scientists) who are searching for specific research data and publications.



B2FIND

For whom?

Researchers (including citizen scientists) who are searching for specific research data and publications.

Research communities who want to provide their metadata to a wider audience.

What can you do with it?

- Find collections of scientific data quickly and easily, irrespective of their origin, discipline or community
- Get quick overviews of available data
- Browse through collections using standardized facets



EUDAT SERVICE SUITE - B2HANDLE





B2HANDLE

For whom?

Research communities with own data services or data models that are not compliant with the data models in EUDAT services and wish to add PIDs to (data) objects.

What can you do with it?

- Assign PIDs to various kinds of managed objects
- Shielding end-users from the complexity of infrastructure details
- Retrieve and refert to PIDs to objects



EUDAT SERVICE SUITE - B2DROP





B2DROP - B2SHAREBRIDGE PLUGIN

Use Case:

A researcher worked on a publication. The volatile digital objects were stored, synchronized and shared using B2DR0P. Now, after finalizing it, he wants to publish the final document using B2SHARE.



B2DROP - B2SHAREBRIDGE PLUGIN

Use Case:

A researcher worked on a publication. The volatile digital objects were stored, synchronized and shared using B2DR0P. Now, after finalizing it, he wants to publish the final document using B2SHARE.

- Extend B2DR0P WebUI
 - Create deposit on a per-file base
 - Make community (metadata) selectable
- Implement third-party transfers



B2SHAREBRIDGE - FILE LIST

einem Jahr
einem Jahr



B2SHAREBRIDGE - COMMUNITY SELECTION

G	TO EUDAT WEBSITE	4 🛎 🗂 🚥	۹ (5
	B2DROP	WHAT IS B2DROP USER GUIDE FAQS CONTACT	
	Alle Dateien	★	×
(1)	Aktuelle	□ ▶ Nextcloud.mp4 < ••• 452 KB vor einem Jahr	
*	Favoriten	Nextcloud Manual.pdf < 4.4 MB vor einem Jahr	Nextcloud.mp4
<	Mit Ihnen geteilt	2 Ordess and 2 Datalog 7.2 MB	★ 452 KB, vor einem Jahr Stags
<	Von Ihnen geteilt		Aktivitäten B2SHARE Kommentare Teilen
8	Geteilt über einen Link		Versionen
•	Tags		Title: Nextcloud.mp4
ŵ	Gelöschte Dateien		Community: EUDAT -
0	373.9 MB von 20 GB verwendet		Open access:
¢	Einstellungen		deposit



B2SHAREBRIDGE - DEPOSIT LIST

GO TO	O EUDAT WEBSITE	4 😃	曲	()))				S
B.	SZDROP EUDAT	Г	v	VHAT IS B2DROP	USER GUIDE	FAQs CO	NTACT	
All	II Deposits	Transfer ID	#Files	Title	Status	Deposit URL	Triggered At	Last Update
1.00		63	1	b2sharebridgetest	deposited	Deposit URL	Thu, 17 Jan 2019 07:09:38	Thu, 17 Jan 2019 07:10:03
Pe	ending Deposits	58	1	Test Upload	deposited	Deposit URL	Wed, 19 Sep 2018 14:53:08	Wed, 19 Sep 2018 15:00:04
Pu	ublished Deposits							
Fa	ailed Deposits							



B2SHAREBRIDGE - DEPOSIT IN B2SHARE

GO TO EUDAT WEBSITE						
<u> </u>	Search r	ecords for			Q SEARCH	
EUDAT	HELP COMM	UNITIES UPLOAD CONTAC	ст	L b.von.s	Lvieth@fz-juelich.de +	
# > RECORDS > 1EDC6g7	FFD154C36A4580F38384	1790F3 - EDIT				
Editing 201	60620_B2	DROP-Technical-	-Meeting_Barc	elona.pptx		
Add files		<				
			Drop files here, o	or click to select f	files	
						/
			eet) Add	B2DROP files		
Uploaded files		Name		Date	Size	
		> a 20160620_B2DROP-Technical-	I-Meeting_Barcelona.pptx	Nov 23, 2016	2.90MB	×
Basic fields						
	Community '	EUDAT				
		EUDAT				
	Title '	20160620_B2DROP-Technical-M	feeting_Barcelona.pptx			
	B ernsteller					
	Description					6
		1100 0010		0114-00		
Association		14.03.2019		Slide 62		

B2SHAREBRIDGE - PERSONAL SETTINGS

G	D TO EUDAT WEBSITE 📄 🖌	- 🛎 🗇 🚍	S
	B2DROP EUDAT	WHAT IS B2DROP USER GUIDE FAQS CONTACT	
	Persönlich	✓ Ihre eigenen Aktivitäten im Stream auflisten ☐ Über Ihre eigenen Aktivitäten via E-Mail benachrichtigen	
i	Persönliche Informationen	E-Mails senden: stündlich •	
	Sicherheit		
С	Sync-Clients		
<	Teilen	EUDAT BZSHARE Bridge	
٥	Zusätzliche Einstellungen	https://b2share.eudat.eu/443 External publishing endpoint	
		9wTZVOqKJcVbNC56qRF1rtNF7qw9MfWryJOcLavZ8Zfw1cur1GpQ B2Share API token	
		Save B2SHARE API Token Delete B2SHARE API Token	
		CLARIN LRSWITCHBOARD Bridge	



B2SHAREBRIDGE - ADMIN SETTINGS

G	D TO EUDAT WEBSITE	4 😃 🛗 🚥	
	B2DROP EUDA	WHAT IS B2DROP USER GUID	DE FAQS CONTACT
	Verschlüsselung	EUDAT B2SHARE Bridge	
4	Aktivität	https://b2share.eudat.eu:443	External B2SHARE API endpoint
•	Workflow	5	# of uploads per user at the same time
	EUDAT	512	MB maximum filesize per upload
1	SSO & SAML-Autorisierung	Check valid secure (https) connections to B2SHARE	
	Umfrage zur Benutzung		
4	Gruppen-Ordner		
E	Protokollierung		
¢	Zusätzliche Einstellungen		
i	Tipps & Tricks		


Hands-on session



PREPARATION

- Go to b2drop-devel.zam.kfa-juelich.de/index.php
- Select B2ACCESS
- Login with your existing account or create a new one
- Create a file with your user name from paper, e.g. haf13.txt
- Copy your cloud id from /settings/user/sharing in the created file
- Share the file with user sapweiler
- Create an app password at /settings/user/security and store it some where, you won't see it again



MOUNT YOUR B2DROP STORAGE

In some cases you want to mount your B2DROP storage locally:

- Use ssh to connect to zam10141.zam.kfa-juelich.de
- Create .davfs2 folder in your home directory
- Create ~/.davfs2/secrets file
- Write mountpoint, username and app password in the first line e.g.:/home/haf13/b2drop sapweiler MYPASSWORD
- Change permissions of secrets file to 600
- Run mount b2drop within your home directory

Prerequisite:

- Installation of davfs2
- User is member of davfs2 group
- Valid entry in /etc/fstab



USE YOUR B2DROP STORAGE

- Create a file with some content
- Move the file in the B2DROP folder
- Review if it is uploaded
- Share it with your neighbour
- Create a new txt file online
- Review if it is downloaded



PREPARE B2SHARE

- Go to trng-b2share.eudat.eu
- Login with your B2ACCESS account
- Create a new API token in your settings
- Store the API token in your personal B2DROP settings
- Store it in some file, it is needed later again



USE B2SHAREBRIDGE

Upload a file from B2DROP to B2SHARE using the B2SHAREBRIDGE

- Select the context menu of a file (···)
- Select B2SHARE
- Enter a deposit name and select the metadata schema
- Press publish

The upload is done in background, during the next cronjob.

- Review you deposit status in the B2SHAREBRIDGE app
- If the status is deposited, click on the URL
- Finish your deposit
- Make another upload with two files



USE THE B2SHARE API

Using the B2SHARE depositer

- Fetch the Depositer from B2DR0P
- View options and syntax python3 run.py -h
- Upload a file with the depositer, do not finalize it
- Review the draft online



USE THE B2SHARE API

Using the B2SHARE depositer

- Fetch the Depositer from B2DR0P
- View options and syntax python3 run.py -h
- Upload a file with the depositer, do not finalize it
- Review the draft online

Using curl

- Identify a targeted community by listing the available ones
- Use the community identifier to fetch the community metadata schema
- Create a draft
- Upload files to your draft; one command per file
- Set the metadata to your record
- Publish your draft

