Using EUDAT services to replicate, store, share, and find cultural heritage data

...in PSNC... and beyond...

Maciej Brzeźniak, Norbert Meyer, PSNC
Damien Lecarpentier, CSC
Plan and purpose of presentation

- Summarize the work done and planned in EUDAT an DCH-RP project
  - How DCH sector can use e-Infrastructures?

- Understand how selected EUDAT services can be used in DCH domain?
  - How EUDAT CDI architecture can be integrated with domain-specific services?
  - Overview of current services (Simple Store, Safe Replication)
  - Presentation of integration work in DCH-RP
  - Discussion on possible future extensions / services
Some aspects of EUDAT

**EUDAT – European Data Infrastructure**

- **Vision:** to support a Collaborative Data Infrastructure
- **Aims:**
  - Provide a **sustainable platform of technologies, tools, services driven by user needs**
  - **Engage users in defining/shaping a platform for shared services**
  - Support **data-intensive, multi-disciplinary research:**
    - Humanities and Social Science: CLARIN
    - But also: earth science (ENES, Earth system modelling; EPOS: European Plane Observing System), ecology (LifeWatch), Virtual Physiological Human (VPH)
  - **Deliver common low-level services** that are required to provide the level of interoperation and trust of data
  - Ensure that the data **infrastructure is robust/scalable** (able to address 'data tsunami')
  - **Build community/domain-specific services** on top of the common services with participation of users
CDI layers
Commons vs community-specific services

The Collaborative Data Infrastructure:
A framework for the future

Data Generators | Users
---|---
User functionalities, data capture & transfer, virtual research environments

Community Support Services
- Data discovery & navigation, workflow generation, annotation, interpretability
- Persistent storage, identification, authenticity, workflow execution, mining

Common Data Services

Trust | Data Curation
---|---

Source: High Level Expert Group on Scientific Data, Riding the wave, 2010.
CDI layers vs tools and services in DCH Integration challenges...

- **High-level services:**
  - EUDAT simple store
  - eCulture Science Gateway
  - dLibra dArceo dLab
  - Invenio
  - Another community solution

- **Low-level services:**
  - Local LTS: long-term storage
  - EUDAT storage
    - Safe Replication
  - Cloud storage
    - S3
    - CDMI
  - Grid storage
    - FTS
    - GridFTP
CDI layers vs tools and services in DCH Integration challenges...

**High-level services:**
- **EUDAT**
  - simple store
- eCulture
  - Science Gateway
- dLibra
  - dArceo
dLab
- Invenio

**Another community solution**

---

**Low-level services:**
- **Local LTS:**
  - long-term storage
- **EUDAT storage**
  - Safe Replication
- **Cloud storage**
  - S3
  - CDMI
- **Grid storage**
  - FTS
  - GridFTP

---

**DCH-RP <-> EUDAT proof of concept: Do they go together?**
Services covered in the presentation:

- **Simple Store service:**
  - enable researchers and scientists to upload, store and share data
  - designated for the "long tail of data"

- **Safe Replication:**
  - Allow communities to replicate data to selected EUDAT data centers
  - Automated replication (iRODS), PID registration (EPIC)

- **Data Staging:**
  - Staging data from user community premises/systems (iRODS)
  - to computing systems, e.g. PRACE’s HPC centres (GridFTP, FTS)

- **Metadata service:**
  - Joint metadata domain for all EUDAT data centres
  - Searchable catalogue covering all data stored within EUDAT

- **AAI:**
  - Provide a solution for a working AAI system in a federated
EUDAT for DCH
Simple Storage Service

Simple Storage Service:
- Address the issues of small user groups and individual users
- Provides solution for “long tail data”: often stored on laptops and departmental servers

Functionality:
- allowing registered users to upload typical data objects into the EUDAT store
- enabling users to share such objects and collections with other researchers,
- lets utilising other EUDAT services
  - Safe Replication
  - PIDs
  - etc.
- May be integrated with AAI

More:
http://www.eudat.eu/simple-store
Simple Storage Service internals:

- Referred also as Researcher Data Store
- Based on Invenio:
  - Developed by CERN
  - [http://invenio-software.org/](http://invenio-software.org/)

- Storage backend:
  - Disk
  - iRODS (EUDAT safe replication)

- Front-end:
  - Developed a new submission portal to invenio
Durable Storage of Research Data

The EUDAT BE2Share (ex-SimpleStore) service provides a quick and easy mechanism for preservation of research data.

In future, the SimpleStore will work with other EUDAT services to ensure data is stored reliably and can be easily searched and retrieved.

The BE2Share is currently in an early testing phase. Users are encouraged to sign up and test the service, but please be aware that is not ready for production use. In particular, note that data submitted to the service in its current state is not stored reliably and will be removed at some stage. Also note that all submissions are made public; there is currently no support for restricting access to files.

Please contact livenson@kth.se with any questions or feedback.
1. Upload files (drag and drop or select files)

Deposits are done in 3 steps:

1. you select and upload one or several data resources
2. you then select the metadata set you want to use and describe the resource by metadata
3. you deposit the selected resource including its metadata elements. The deposit takes a moment to be processed. You will get a reference URL at once.

[Select files]
[Start upload]
1. Upload files (drag and drop or select files)

Deposits are done in 3 steps:

1. you select and upload one or several data resources
2. you then select the metadata set you want to use and describe the resource by metadata
3. you deposit the selected resource including its metadata elements. The deposit takes a moment to be processed. You will get a reference URL at once.
1. Upload files (drag and drop or select files)

Deposits are done in 3 steps:

1. you select and upload one or several data resources
2. you then select the metadata set you want to use and describe the resource by metadata
3. you deposit the selected resource including its metadata elements. The deposit takes a moment to be processed. You will get a reference URL at once.

[Select files] [Start upload] [Stop upload]

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>apples.jpg</td>
<td>784 KB</td>
<td></td>
</tr>
<tr>
<td>mountains.jpg</td>
<td>159 KB</td>
<td></td>
</tr>
</tbody>
</table>

2. Select a domain or project

Please select a metadata description set which best fits to your data. The EUDAT Generic set includes a common set of elements which are
# EUDAT for DCH

## Simple Storage Service

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>apples.jpg.JPG</td>
<td>784 KB</td>
<td></td>
</tr>
<tr>
<td>mountains.jpg.jpg</td>
<td>159 KB</td>
<td></td>
</tr>
</tbody>
</table>

## 2. Select a domain or project

Please select a metadata description set which best fits to your data. The EUDAT Generic set includes a common set of elements which are inspired by Dublin Core, MARC and DataCite and this set is included in all sets as basis. So if you just want the common set select Generic, if you want in addition a specific community set select that community. Soon there will be more communities. You can always revise your decisions and select another set, however you may lose what you have already entered.
### 3. Add metadata

**Generic**

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Maciej's photos</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Photos taken in 2013</td>
</tr>
<tr>
<td><strong>Creator</strong></td>
<td>Maciej</td>
</tr>
<tr>
<td><strong>Open Access</strong></td>
<td>ON</td>
</tr>
<tr>
<td><strong>Licence</strong></td>
<td>Public</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>Maciej Brzeziak</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2013</td>
</tr>
<tr>
<td><strong>Tags</strong></td>
<td>art, photography, landscape, still nature</td>
</tr>
</tbody>
</table>

Add more details?  

4. **Deposit**
Deposit Successful

Your submission will shortly be available at:
https://tmng-b2share.eudat.eu/record/2

Please note that it may take a few minutes to process your submission.

Deposit another item
# 2013 Photos by Maciej

**Maciej Brzeziak**

**Maciej Brzeziak**

Abstract: Still nature and landscape photographs

Keyword(s): art; photography; still nature; landscape

The record appears in these collections:

- eudat

## Files

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>mountains.jpg.jpg</td>
<td>30 Oct 2013</td>
<td>163.3 kB</td>
</tr>
<tr>
<td>apples.jpg.JPG</td>
<td>30 Oct 2013</td>
<td>803.2 kB</td>
</tr>
</tbody>
</table>

**Checksum:** 3f19aeac340de9465b69f3db978b338389870466a9bf223f6aa2e2dc4bee458

**PID:** ad99da14-4149-11e3-bc23-14feb57d12b9

**Publication:** Maciej Brzeziak

**Licence:** Public

**Uploaded by:** maciekb@man.poznan.pl

**Domain:** eudat
Safe Replication Service:
- Allows communities to replicate data to EUDAT data centers
- Can be integrated with portals and community tools

Functionality:
- Ingested file/data are:
  - automatically replicated to many data centres
  - get persistent identifiers registered (PID based on EPIC)
- Various interfaces supported
  - iRODS: icommand, API,
  - WebDAV, GridFTP
- Can replicate data on top of various different data stores:
  - Disks, tapes, HSMs
  - Clouds (e.g. S3)

More:
http://www.eudat.eu/safe-replication
Integration option 1: EUDAT Simple Store + EUDAT Safe Replication

Simple Store Service

Support for sharing

EUDAT Safe Replication

Easy deposit & access to the data

Transparent replication of data, persistent storage

Data Store 1 → Data Store 2
Integration option 2: Community portal/solution + EUDAT Safe Replication

Community service

Support for sharing

Storage & access typical for community

Safe Replication

Transparent replication of data

Data Store1 → Data Store2

Simple upload → Simple metadata → PID registration → Data center store → Simple store portal
Option 1 in practice:
EUDAT Simple store + EUDAT Safe Replication

**DCH-RP <-> EUDAT PoC 1**

- **Simple Store Service**
  - Support for sharing
  - Easy storage & access to the data

- **Safe Replication**
  - Transparent replication of data

**Memory institutions:**

- **EUDAT Simple Store**
  - Provided by PSNC

- **EUDAT Safe replication**
  - Provided by PSNC (Poznan) & EPCC (Edinburgh)
Option 2 in practice:
Community CMS solution + EUDAT Safe Replication

**DCH-RP <--> EUDAT PoC 2**

Memory institutions:

**EUDAT Simple Store**  
*Provided by PSNC or EUDAT partners*

**EUDAT Safe replication**  
*Provided by PSNC (Poznan) & EPCC (Edinburgh)*

---

**Community service**

- Feature-full domain-specific tool with large user base

**Safe Replication**

- Transparent replication of data

**Support for sharing**

- User
- User

**Data Store1** @PSNC

**Data Store2** @EPCC
Basic statistics:

- ± 100 digital libraries
- several hundreds memory institutions
- over 1,1 M of digital objects
- 98% content delivered via services based on dLibra software ([http://dlibra.psnc.pl/](http://dlibra.psnc.pl/)) which uses Solr for content and metadata indexing and searching
We investigate two possible ways to offering data preservation services for DCH

- Top-down solution for 'citizen scientists' / 'citizen DCH people'
- Well-established solution backed by EUDAT Safe Replication

We exploit the layered EUDAT CDI architecture

- In theory: It enables integration with existing solutions
- We try to understand how it works in practice
Planned extensions of simple store

Discussion trigger 1

- Development roadmap

  - Premium service:
    - Customisation for layout, metadata – for community needs
    - increased storage capacity
    - increased support
    - increased bandwidth

  - Premium vs regular service:
    - Providing premium service requires enrolling with EUDAT
    - Regular services to be offered to 'citizen scientists/users’ – no close relationship needed

  - AAI integration
    - On the roadmap
Possible extensions of simple store
Discussion trigger 2

- From yesterday's discussion about Simple Store:

  - Thousand of files?
  - Upload reliability / robustness?

    => Batch upload
    - API to be developed
      » Enables integration with existing systems
      » Tools can be offered by EUDAT to support batch upload

- Collections upload?

  => Support for meta-data extraction
  - Implemented client-side?
  - E.g. based on pre-prepared collections (e.g. DIPs)
Possible extensions of simple store
Implementation: user-side tool?

Data

Meta-Data

Client-side application

Data upload control

Meta-data extraction and upload

Meta-data exchange API

Data upload API

Simple Store API

Simple upload
Simple metadata

PID registration

EUDAT CDI Domain of registered data
Possible extensions of simple store
Implementation: user-side tool?

- **Highlights:**
  - Automation: ease of use, reliability, performance
  - Functionality: data upload, meta-data extraction

![Diagram of data and meta-data flow](image)
Possible extensions of simple store
Implementation: user-side tool?

- **Highlighted:**
  - Automation: ease of use, reliability, performance
  - Functionality: data upload, meta-data extraction

- **Challenges:**
  - Portability
  - Universality: standards need to be identified
  - Sustainability
Possible extensions of simple store

Implementation: user-side tool?

- **Highlights:**
  - Automation: ease of use, reliability, performance
  - Functionality: data upload, meta-data extraction

- **Challenges:**
  - Portability
  - Universality: standards need to be identified
  - Sustainability

- **Discussion needed!**

Data

Meta-Data

Client-side application

Data upload control

Meta-data extraction and upload

Meta-data exchange API

Data upload API

Simple Store API

Simple upload

Simple metadata

PID registration

EUDAT CDI Domain of registered data
Possible extensions of simple store
Implementation: user-side tool?
Extensions of EUDAT services discussion

Summary

Message:

- EUDAT infrastructure and services are layered, modular
  - This enables integration

- Further extensions possible
  - Users are welcome to influence them

- We want to make sure that we recognised and support necessary standards

- Technical details / organisation / etc. to be discussed
Using EUDAT services to replicate, store, share, and find cultural heritage data

...in PSNC... and beyond...

Maciej Brzeźniak, Norbert Meyer, PSNC
Damien Lecarpentier, CSC

THANK YOU!