Agenda

- Short introduction to bwDataDiss

- Concept
  - General
  - Characterization
  - Archive

- Bibliographic metadata

- (Policy)
bwDataDiss – Short introduction

- **Purpose of bwDataDiss:**
  - On the one hand: Allow libraries to store and conserve research data, respectively enable libraries to offer such kind of services to their PhD candidates.
  - On the other hand: Provide PhD candidates the possibility to conserve their research data.

- **In terms of Open Science, the research community benefits from the possibility to**
  - re-use and to
  - review the archived research data
bwDataDiss – Short introduction & partners

- Long term preservation:
  - Before the actual archiving in tape libraries, bwDataDiss performs a so called ‘characterization‘ of the research data
  - The result of this characterization can be used by libraries to maintain a certain standard of quality and by the storage operator to develop adequate preservation strategies

- Project partners:
  - University of Freiburg
    - Library
    - Datacenter \(\rightarrow\) Characterization
  - KIT (Karlsruhe Institute of Technology)
    - Library \(\rightarrow\) Coordination
    - SCC (Steinbuch Centre for Computing) \(\rightarrow\) Archiving (LTS)
Concept - Basics

- Basics:
  1.) The interest of the clients (PhD candidates) is the top priority
     - bwDataDiss should be as simple as possible to use
  2.) Integrity of research data must be ensured
     - Checksum* calculation when research data is transferred
  3.) Flexible and easy integration in existing library systems
     - For libraries, it should be relatively easy to integrate the services provided by bwDataDiss. Also, different integration scenarios should be supported
Concept - separation of duties

bwDataDiss
- Archiving of research data in tape libraries
- Provides the research data stored in tape libraries
- Central characterization for long term archiving
- ...

Libraries
- Counseling of PhD candidates and researchers
- Captures and maintains bibliographic metadata
- THE place to go for PhD candidates

→ Upload of the research data is (usually) performed on library websites
Separation of duties
bwDataDiss as perceived by a PhD candidate

- Existing library system (simple form for example) to hand in the dissertation
  - Transferred to library:
    - Dissertation
    - (bibliographic) metadata

- Additional form to transfer research data + metadata
  - Additionally transferred to library:
    - (bibliographic) metadata for the research data
    - (Usually,) research data is directly transferred to bwDataDiss
Access to bwDataDiss

bwDataDiss uses bwIDM to authenticate users

- PhD candidates who want to store research data connect with their bwIDM account
- Library co-workers connect to bwDataDiss also using a bwIDM account
- Researchers and others who just want to re-use archived data don’t need an account to access the data – except if an embargo is in place

If no access restriction is in place, research data can be accessed without user authentication

- We expect this to be the default case
Reliable upload of big* files

bwDataDiss allows to upload files of arbitrary* size using a web browser (> 10GB)

bwDataDiss ensures the correct transfer of the data:
  - By calculating checksums (MD5, hash-function)
    - Before the actual upload in the browser (JS)
    - At reception of the data by bwDataDiss
  - And, if necessary re-transfers the data if checksums mismatch

bwDataDiss allows to resume uploads at a later point in time
  - This is especially useful for bigger files and / or low bandwidth
Reliable upload of big files

file: test.xml
Size: 378.3MB

MD5
Data

#1
#2
#3
#4
Hand in and release workflow

- Head for your library
  - Now (something like that):
    - Provide metadata for your dissertation
    - Upload of the dissertation
  - Additionally:
    - Authenticate with bwIDM (if not yet happened)
    - Provide metadata for your research data
    - Upload research data

- Once research data is transferred to bwDataDiss and metadata to the library:
  - Research data is locked for changes
  - The library checks the provided metadata
  - bwDataDiss performs characterization of the research data
Hand in and release workflow (2)

- It’s up to the library to decide whether the provided research- and meta-data is acceptable or not.

- If library does not release the data:
  - It contacts the Ph.D. candidate and sorts out the problem.
  - It allows the Ph.D. candidate to modify the research data:
    - Thus, to delete, replace, complement the data.
    - This can be performed by the Ph.D. candidate directly on the bwDataDiss portal.
  - Then, the characterization and release process starts over again.
Hand in and release workflow (3)

- If library releases the data:
  - Library notifies bwDataDiss about release
    - Per API or
    - Providing the corresponding metadata (using OAI-PMH, or metadata push)
  - bwDataDiss retrieves the corresponding metadata
    - bwDataDiss only expects a „minimal dataset“
  - bwDataDiss creates a DOI
  - bwDataDiss transferrs the research data to the archive
    - Whereby data integrity is ensured
  - bwDataDiss provides the link to the landing page to the library
    - per API
    - on the bwDataDiss portal page
Characterization

- Characterization is based on file-types

- Results of characterization are provided to the libraries
  - The interpretation of the results is library-specific
  - The visualization of this interpretation is very simple:
    - in form of a traffic light
Characterization

Library

Simplified result visualization

Library policy

Research data

bwDataDiss

Characterization results
Characterization – example (simplified)

Results: 30% pdf, 20% xef*, 40% csv, 10% xlsx

Pdf: super, Office-files: ok, etc.
Encrypted files: 😞 we don’t like those!

Based on the results and the library policy bwDataDiss provides a quality estimation → encrypted files are unacceptable

An easy to understand visualization: a traffic light indicator: RED (encryption 😞 (even if it might not always be a bad sign…))

*xef: WinAce encrypted file
Archive

Based on: High Performance Storage System (HPSS),
http://www.hpss-collaboration.org/
- Hierarchical storage system (disk arrays ↔ tape-arrays)
- Focus on long-term storage of huge amounts of data (PetaBytes)
- Horizontal scalability of all system components

KIT-Installation:
- Used by different projects (bwDataDiss bwDataArchiv, RADAR, …)
- Frontends for ‘end-users’:
  - SFTP: file-based access → used by bwDataDiss
  - REST-Interface: object-based access and metadata services
    (checksums, attribute-based search, etc.)
bwDataDiss – bibliographic metadata

- bwDataDiss only requires a small set of bibliographic metadata
  - A library is free to collect additional metadata

- The DFG (Deutsche Forschungsgemeinschaft) classification is used

- Open licenses CC-BY and CC-BY-SA
  - Further licenses will be added as requested by libraries
bwDataDiss – bibliographic metadata

- **Title**: hübscher titel
- **Publisher**: jemand ;)
- **Publication Year**: 01.01.2012
- **Classification**: 30901 - Kern- und Elementarteilchenphysik, Quantenmechanik, Relativitätstheorie, Felder
- **Resource Type**: Dataset
- **Keywords**: test, mechanik
- **Readme**: Liesmich ... LiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmichLiesmich
- **Creation Year**: 01.01.2014
Embargo, Rights, Policy

- Files that have reached the archive will never be deleted
  - Exceptions: Maximum archivation time reached, data damaged, etc.
  - No version support! Files can’t be overwritten! → new dataset

- All research data is linked to a publication, has a license and are published worldwide*

- Data that lay under (limited) embargo:
  - Readable only by the uploader (Ph.D. candidate)
  - Other users (bwIDM account required) may be given read rights:
    - By the library
    - By the Ph.D. candidate

*There might be exceptions
Thanks for your attention!

Questions?

https://bwdatadiss.kit.edu/
Backup / Detail slides
API (Application Programming Interface)

- **API-Authentication with API-Key**
  - API-Key is generated after first successful bwIDM connection
- **Responses in JSON or XML**
- **REST-Style**
- Allows a deeper integration of bwDataDiss with (existing) library systems

**Example:**
- **GET** http://<host>:<port>/api/v1/datasets/<name/id>/files[.xml/.json]
  - List files of a dataset
- **POST**
  http://<host>:<port>/api/v1/datasets/<name/id>/file/<name/id>
  - Writes data in the specified file on the server
Simplified: API and API-Key
Simplified design of bwDataDiss
Detail Architecture