RUHR-UNIVERSITÄT BOCHUM

CREATING TRUST IN RESEARCH DATA REPOSITORIES

Case study of an institutional research data repository at Ruhr University Bochum

Talk Outline

Part I: Background

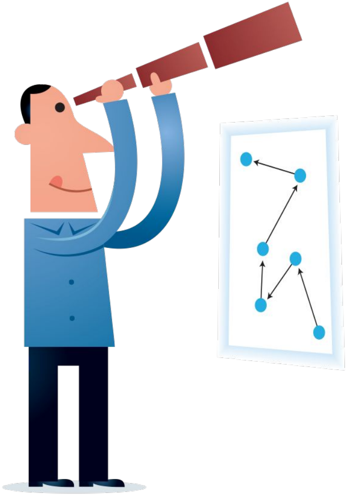
- Research Data Management in Germany

Part II: Who we are

- Ruhr University Bochum
- Research Use Case from Neuroscience

Part III: Our Approach

- From requirements to implementation
- Collaboration with Use Case
- Current status and future aims



Digitalbevaring.dk

Part I

Background

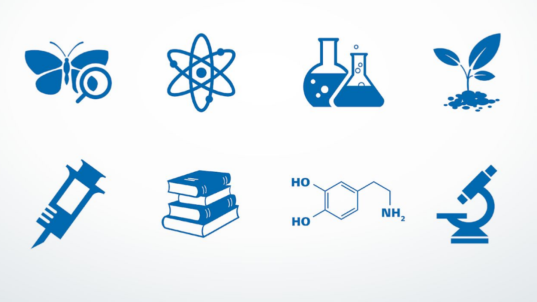
Research Data...

- are all data produced or used by research
- exist in (almost) all disciplines
- are created and processed by a variety of (subject-specific) tools, methods and devices
- usually lack a uniform description
- are very often produced in large quantities (file sizes and number of files)
- are unique and of high value in many cases (astronomy, medicine, history, …)

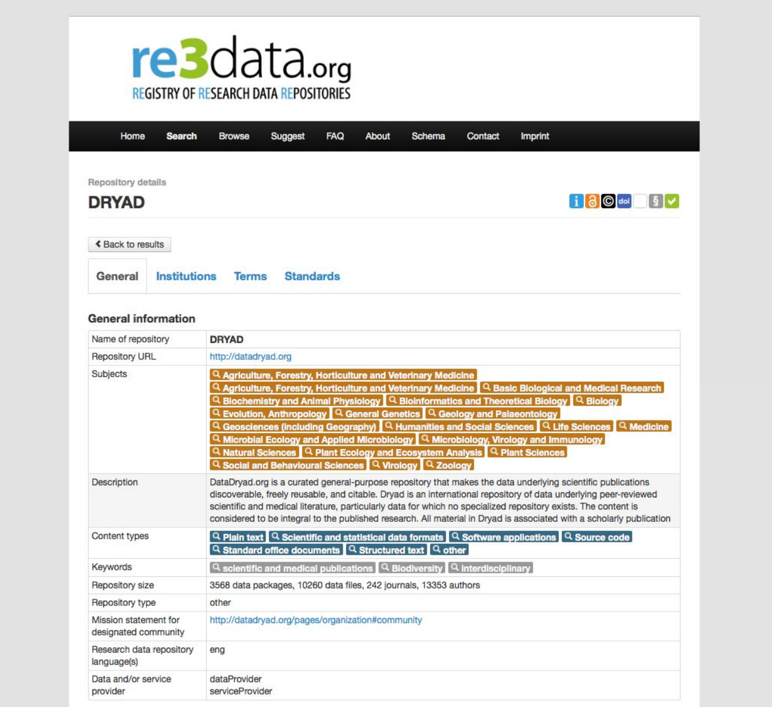
44056-005
AK C-swab
AI VTM 2006
DEUTSCHLAND
BENEZER PROUT.
Bear me on, my faithful steed, Bear-
me to my goal with speed,
Ebenezer Prout.
From "The Red-Cross Knight")
56

```
. Sep 15:53 ..
0. Sep 2015  bin -> usr/bin
9. Sep 09:31 boot
21. Sep 15:50 dev
19. Sep 09:32 etc
21. Sep 15:52 home
30. Sep 2015  lib -> usr/lib
30. Sep 2015  lib64 -> usr/lib
23. Jul 10:01 lost+found
1. Aug 22:45 mnt
30. Sep 2015  opt
21. Sep 15:52 private -> /home/e
21. Sep 08:15 proc
12. Aug 15:37 root
21. Sep 15:50 run
30. Sep 2015  sbin -> usr/bin
30. Sep 2015  srv
21. Sep 15:51 sys
21. Sep 15:45 tmp
12. Aug 15:39 usr
23. Jul 10:25 var
```

Research Data Repositories



Discipline-specific repositories are available, **however the are not accepted as expected**



Repository details

DRYAD

General information

Name of repository	**DRYAD**
Repository URL	http://datadryad.org
Subjects	Agriculture, Forestry, Horticulture and Veterinary Medicine Agriculture, Forestry, Horticulture and Veterinary Medicine Basic Biological and Medical Research Biochemistry and Animal Physiology Bioinformatics and Theoretical Biology Biology Evolution, Anthropology General Genetics Geology and Palaeontology Geosciences (including Geography) Humanities and Social Sciences Life Sciences Medicine Microbial Ecology and Applied Microbiology Microbiology, Virology and Immunology Natural Sciences Plant Ecology and Ecosystem Analysis Plant Sciences Social and Behavioural Sciences Virology Zoology
Description	DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad is an international repository of data underlying peer-reviewed scientific and medical literature, particularly data for which no specialized repository exists. The content is considered to be integral to the published research. All material in Dryad is associated with a scholarly publication
Content types	Plain text Scientific and statistical data formats Software applications Source code Standard office documents Structured text other
Keywords	scientific and medical publications Biodiversity interdisciplinary
Repository size	3568 data packages, 10260 data files, 242 journals, 13353 authors
Repository type	other
Mission statement for designated community	http://datadryad.org/pages/organization#community
Research data repository language(s)	eng
Data and/or service provider	dataProvider serviceProvider

re3data.org lists over 3000 research data repositories from around the world

RDM: Challenges

Legal

- Sensitive data (psychology, medicine, …)
- Intellectual property (esp. when project partners from industry are involved)

Practical

- Lack of standards for data handling
- Unpropriate workflows in existing repositories (complex data models, peer-review, large data sets)
- Lack of time (→ high pressure in scientific careers / „up or out“)

Cultural:

- Fear of sharing or publishing data
- Lack of trust in service providers

Digitalbevaring.dk

However: Research data management is key to sustainable research, reducing effort of future research and enabling innovative data analysis

Funding for RDM

DFG Deutsche Forschungsgemeinschaft

Main funder: German Research Foundation (DFG)

- Revision of „Guidelines for Safeguarding Good Scientific Practice“ in 2019: **RDM is mandatory**
- **RDM infrastructure must be provided by host institution**

However, **incentives for standardization are given**, especially in Collaborative Research Centers (CRCs)

- CRCs are one of the larges funding lines of DFG (~ 24% of the total DFG budget (€848 million) for 268 CRCs)
- Information Infrastructure projects (INF) within CRCs fund staff capacities for training, consulting, policy development regarding research data

Part II

Who we are

Ruhr University Bochum (RUB)

„ Deep in the west, where the sun gathers dust. It's better. Much better than you think.”
Herbert Grönemeyer

Location:

- Near the river Ruhr: Former industrial center → most dense populated area in Germany (5 million people in the area)
- Near (within a 30-min ride) to two other full universities collaborating in University Alliance Ruhr
- In the German state of North-Rhine Westphalia: One of the two German states with the largest higher education sector in German (about 40 universities, thereof 12 full universities)

Hamburg
Berlin
Bochum
Munich

Facts and Figures:

- One of the 10 largest universities in Germany (40,000 students, 6000 employees, 500 professors, 21 faculties)
- Strong research:
  - €188 million annual third-party funding (25% of total budget)

Research data management at RUB

Staff capacities:

- Research data services (library and IT)
- Started with 5 full-time equivalents of staff capacities
- Today: about 10 full-time equivalents of staff capacities

Storage for research data:

Two S3 object storage infrastructures in operation:

1. within a consortium of 5 universities (2 PB)
2. within the 3 universities of University Alliance Ruhr (300 TB)

Repository development

- 2020: €360,000 for repository implementation
- 2023: Money for 1 developer position for 6 years

CRC 1280: Research Use Case

extinction learning | SFB 1280

- Topic: “Extinction Learning” (Neuroscience)
- 81 researchers in 17 projects at 4 institutions
- Scientific disciplines: biology, psychology, medicine, and computational neuroscience
- Techniques: microscopy, single cell recording, magnetic resonance imaging, questionnaires
- Human and animal subjects → sensitive data
- Large existing data sets (32 TB, 24 million files in 2 million folders) → ingest strategy

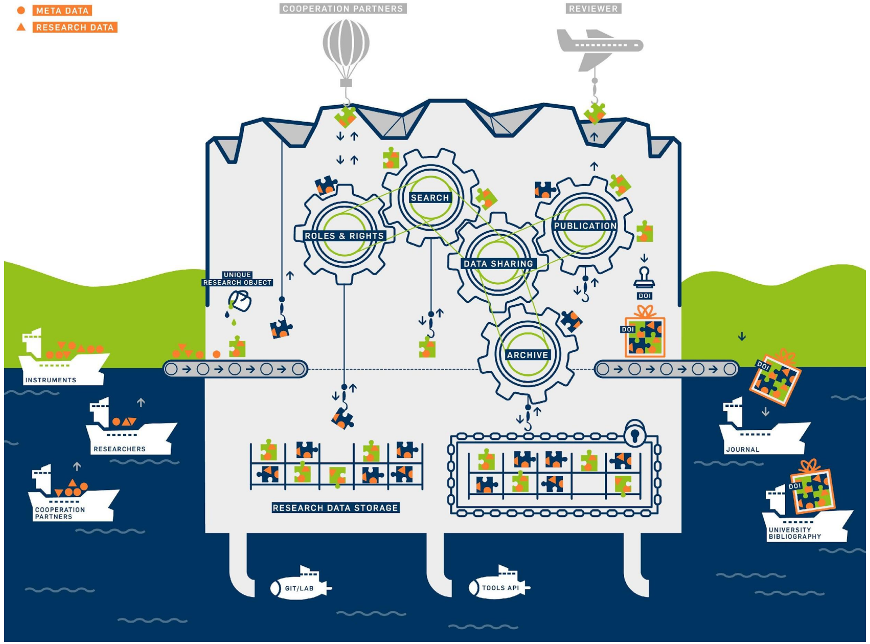
Common data model applying an inheritance strategy across folder structures

Part III

Our Approach

Requirements

- 2 data models: generic (university wide), neuroscience CRC (discipline-specific)
- Differentiated visibility of data
  → roles & permissions
- Complex (three-step) review workflows
- Data curation steps: Draft, archiving, publication, tombstoning
- Automated data import of hierarchical metadata
- Use of local S3 storage for data and metadata
- Login for project partners via ORCID



Market analysis

Starting point: Feature analysis of available open source repository platforms
→ modifications always necessary

Aim: Flexible toolbox for implementation of innovative workflows

Challenge: No own expertise to implement modifications

Solution: Tendering process for external service provider supporting

- Specification of technical requirements
- Choice of platform
- Implementation
- Roll-out, testing and training concepts

Digitalbeva

Participation of Use Case

Before ReSeeD implementation:

- Cooperative development of data model and mapping to Datacite and Dublin core
- Definition of requirements for ReSeeD

During ReSeeD implementation:

- Participation of use case in the project team (selection of and communication with service provider, coordination of beta tests)
- Preparation of data ingest into ReSeeD by storage of (meta)data in a prescribed way
- Provision of actual (not sensitive) data for testing during development

Digitalbevaring.dk

Implementation

ReSeeD
RESEARCH DATA MANAGEMENT SYSTEM

Current status:

- **Generic data model (university-wide):** Implementation finished, user acceptance tests running
- **CRC data model:** Beta tests of bulk ingest

Lessons learned:

- Implementation of CRC data model and workflows (→ roles and rights) more challenging than expected
- Building up expertise on the operation of ReSeeD is a challenge
- Technical capabilities still not fully meet expectations of researchers (regarding flexibility and performance)

Digitalbevaring.dk

ReSeed: Next steps

ReSeeD
RESEARCH DATA MANAGEMENT SYSTEM

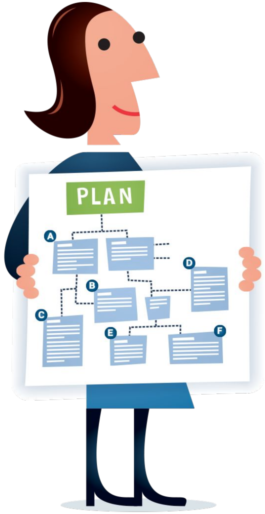
- Launch publication service
- Establish supporting measures: Training, user tutorials, user survey

Ongoing implementation and testing

- Ingest of data records with multiple TB using bulkrax
- Usability testing

Future aims

- Upgrade Hyrax 3.5 → 5
- UI re-design with input from usability testing
- Specific data models for further use cases
- Data ingest API
- Flexible metadata
- German language support



Digitalbevaring.dk

Conclusion

Challenge for acceptance of RDM infrastructures

1. (sensitive) data are not (yet) intended for publication or sharing among all repository users
2. (innovative) requirements from (interdisciplinary) projects are not represented by repository workflows
3. researchers are used to a fast evolution of RDM tools and do not trust in institutional processes

Digitalbevaring.dk

ReSeeD fosters acceptance by

1. connecting to local storage infrastructure (fulfilling requirements for sensitive data storage) + roles and permissions
2. providing individual workflows for interdisciplinary research projects
3. generating trust via a participation of researchers in the development

Thank you very much for your attention!

Nina Winter, PhD
Research Data Services @ RUB

extinction
learning | SFB 1280

UNIVERSITÄTSBIBLIOTHEK
BOCHUM

IT.SERVICES