RADAR
A Repository for Long Tail Data

Angelina Kraft, Janna Neumann
German National Library of Science and Technology TIB

36th Annual IATUL Conference
Hannover, July 6th, 2015
IN A NUTSHELL

RADAR = Research Data Repository

- Goal: Establish a interdisciplinary research data repository
- Online: http://www.radar-projekt.org
- What kind of data? → DIGITAL! Including:
  - Raw (primary, machine output)
  - Secondary (working data)
  - Negative
  - Analyzed in scientific articles
- Funded by DFG
SCIENCE – NOW & THEN

• One thousand years ago, science was empirical: described natural phenomena

• Over the last one hundred years, a theoretical branch developed: building on models, generalisations

• In recent decades, an IT branch: Simulation of complex phenomena

• Today, science is data-based (eScience): Combination of theory, experiment & simulation
“The majority of datasets produced through research are part of the ‘Long Tail of Research Data’”


Science Survey 2011:
- 48% of respondents were working with datasets that were <1GB in size
- 50% stored data exclusively! in labs


Source: Ferguson et al. (2014): Big data from small data: data-sharing in the ‘long tail’ of neuroscience. DOI: 10.1038/nn.3838
Current situation of data publishing:

- Articles
- Supplements
- Data centres / repositories
- Data on private hard disks / institutional servers

- Potential for 'data dumping’ → overburdened!
- Few
- Lack of archives in many subject areas!
- ~ 75 % of RD is never published

Modified based on STM / Smit, E: Avoiding a Digital Dark Age for Data: why data and publications belong together ICSTI workshop Delivering Data in Science PARIS, 5 March 2012

RESEARCH DATA REPOSITORY
http://www.radar-projekt.org

Hannover, July 6th, 2015
Ideal case of data publishing:

RD in articles

Supplements

Data on private hard disks / institutional servers

- Linking text & data → ‘enhanced publications’
- If no other data integration is possible
- Journals request and check data deposition
- Generic & discipline-specific; interfaces for good connection!
- Support ‘enhanced publications’; Persistent Identifiers

Modified based on
STM / Smit, E: Avoiding a Digital Dark Age for Data: why data and publications belong together
ICSTI workshop Delivering Data in Science
PARIS, 5 March 2012

RESEARCH DATA REPOSITORY
http://www.radar-projekt.org

Hannover, July 6th, 2015
RADAR – The Domain Model

1. Private domain
   - Researcher’s workplace

2. Collaborative domain
   - Institutional infrastructure

3. Public domain
   - Archive
   - RADAR – 2 Services:
     1. Archival
     2. Publication
   - Business model
     - Infrastructure
     - Software
     - Metadata standards
     - Persistent Identifiers
     - Contracts
     - Interfaces
   - DataCite, publishers

4. Dissemination domain
   - Portals, researchers

Data selection
Data documentation
Data types / Data formats

Reuse


RESEARCH DATA REPOSITORY
http://www.radar-projekt.org

Hannover, July 6th, 2015
Software, framework & business model
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Data management & preservation services
KIT SCC
Karlsruhe Institute of Technology

Scientific specification & evaluation

Data publication, metadata & contact to publishers
TIB GERMAN NATIONAL LIBRARY OF SCIENCE AND TECHNOLOGY
CUS OF RADAR

Archival of research data as a generic service
Trustworthy preservation & traceable publication
Long tail“ of research data

Services
• Basic service: interdisciplinary data preservation
• Extended service: data publication
**TARGET AUDIENCE**

**Researchers**
- Archive and publish project-based research data

**Libraries and Research Institutions**
- Integration with existing institutional portals

**Cultural Heritage Organizations**
- Long-term preservation & web access of digitized materials

**Publishers**
- Infrastructure for providing access to research data
- Linked to publications
SERVICE: Archival Storage

- Aim: Trustworthy data preservation

- For whom?
  - Completed research projects
  - Internal resources, not to be publically available (yet)

- Properties:
  - Minimum metadata set (9 parameters)
  - Handle
  - Variable storage period: up to 15 years + extension
  - Bitstream Preservation for storage period
  - Regular reports on data integrity
  - Access rights for selected groups/users
SERVICE: Data Publication

• **Aim:** Trustworthy preservation & traceable publication

• **For whom?**
  - Projects: Data basis for scientific papers
  - Independent data publications (e.g. negative data)
  - Digital representations

• **Properties:**
  - Expanded metadata set for discipline-specific data
  - DOI
  - Unlimited storage period
  - Regular reports on downstream use to data provider
  - Access management (embargo & publisher services)
<table>
<thead>
<tr>
<th>Workspace</th>
<th>Date of creation</th>
<th>Last update</th>
<th>Size</th>
</tr>
</thead>
</table>
Edit metadata

Drag & Drop your metadata XML file
or browse your filesystem

Don't have one? Just fill out or edit the form below.

Metadata fields

Identifier

- e.g. 10.1371/journal.pbio.0020449

DOI

Additional available fields: Alternate identifiers, Related identifiers

Creators

- Apache Software Foundation (ASF)
- Community

Additional available fields: Contributors
Dataset: apache-maven-3.3.1-src.zip

Details

Description:
Apache Maven is a software project management and comprehension tool. Based on the concept of a project object model (POM), Maven can manage a project’s build, reporting and documentation from a central piece of information. If you think that Maven could help your project, you can find out more information about it in the “About Maven” section of the navigation. This includes an in-depth description of what Maven is, a list of some of its main features, and a set of frequently asked questions about what Maven is.

Production year:
2006

Subject areas:
Information Technology;
Software Technology

Date of creation:
May 5, 2015 1:24:22 PM

Total size:
9 MB

Identifier:
-

Creators:
Apache Software Foundation (ASF); ;
Community:

Title:
Apache Maven

Publishers:
Apache Software Foundation (ASF)

Status:
PENDING

License:
CC_BY_4.0_ATTRIBUTION

Retention time:
-

Creator:
admin

Last update by:
admin

Language:
English

Resource:
Apache Maven is a software project management and comprehension tool: SOFTWARE

Rights holders:
Apache Software Foundation (ASF)

Data citation:
Creator (Publication Year): Title of the data set. Publisher. Resource Type. Identifier
WORKFLOW

RESEARCHER / INSTITUTION

RADAR

WEB PORTAL

COMMUNITY

1. Registration / Sign in
2. Data selection
3. Service types:
   A. Preservation
   B. Data publication with integrated data preservation

4. Data ingest
5. License
6. Data transfer & checksumming
7. Data preservation
8. Allocation of persistent identifiers
9. Reporting for the data provider

RESEARCH DATA REPOSITORY
http://www.radar-projekt.org

Hannover, July 6th, 2015
RESEARCH DATA MANAGEMENT

• Guidelines
  → How-To’s, recommendations on formats, citations, licenses ...

• General & discipline-specific glossary
  → Step-by-step addition of examples

• Business model & quotes
  → Indicative price, e.g. for funding applications

• Integration services for publishers/journals
  → Data for peer review
RADAR – Reliable Storage Space

- Management of **storage quotas**
- **Bitstream Preservation**
- **Regular fixity checks**
- **PID Service (DOI & Handle)** on data set or file level
- **Generic** metadata schema
- Managing **license metadata** & **access rights**
- Access may be restricted to the institution providing the data (resp. another authorized party) and service operator

- But: No functional long-term preservation!
RADAR Roadmap

- Software – further development of services
  - 1. Middleware infrastructure → realized
  - 2. Archival service → realized
  - 3. Publication service → in progress

- DSA certification → in progress

- Roll-out to further disciplines → in progress

- Workflows & interfaces to data providers → in progress
Thank you for your attention!

Questions?

RADAR Test Account – Contact:
angelina.kraft@tib.uni-hannover.de