DATA REPOSITORIES

UKB RDM Theme Session

April 26th 2022

Today’s program

- Welcome to the UKB RDM Data Repository Theme Session
  - *Time: 13.00 – 15.00 hours*

- Three topics will be presented:
  - *Data Repository Landscape: introduction, survey & discussion*
  - *Data Repository Finder: pitch & discussion*
  - *Software Curation & Sensitive Data Sharing @ 4TU: challenges & best practices*

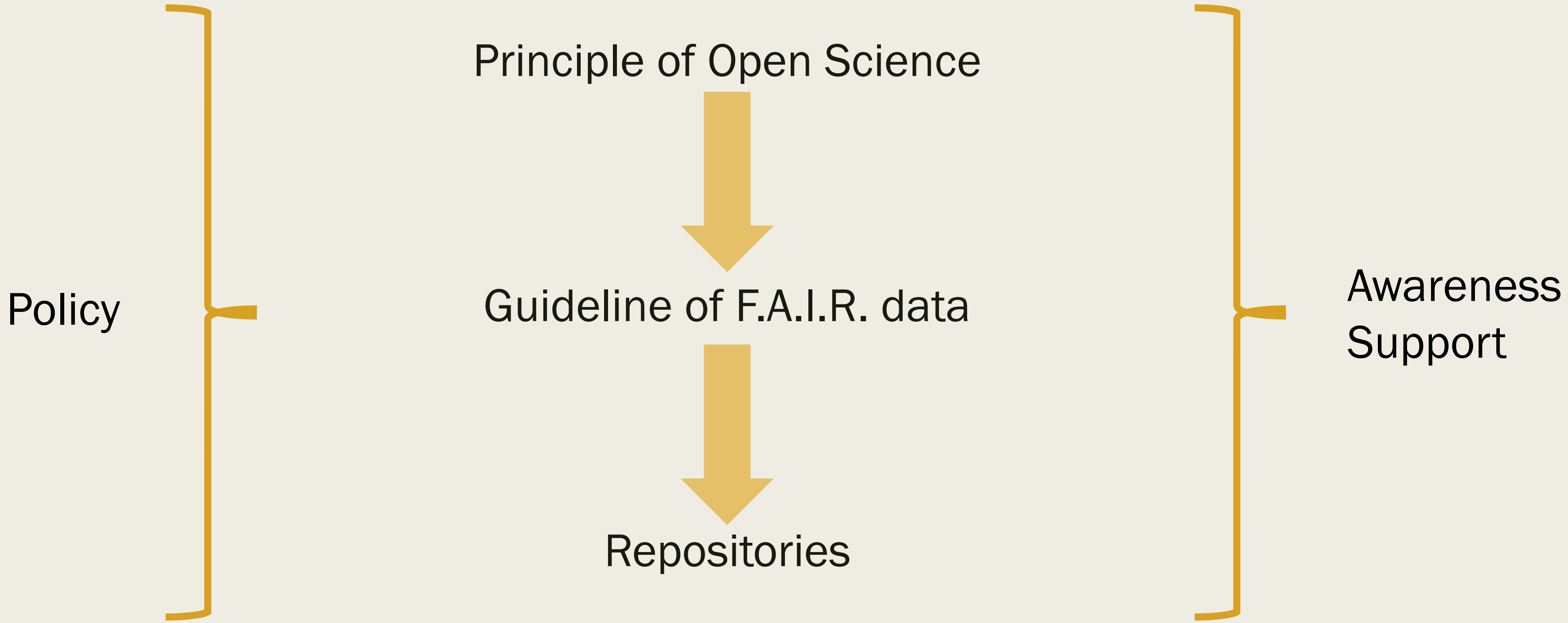
- We encourage discussion: please share your thoughts, opinions & experiences
  - *Raise your digital hand*
  - *Introduce yourself: name / function / where do you work*

DATA REPOSITORY LANDSCAPE

UKB RDM Theme Session

April 26th 2022

Data Repository Landscape: Introduction I



Policy

Principle of Open Science

Guideline of F.A.I.R. data

Repositories

Awareness
Support

Data Repository Landscape: Introduction II

Repositories: seemingly endless variety

- Large number
  - *Data Monitor harvests metadata of >2.000 repositories*
- Multiple types
  - *Institutional vs. Domain vs. General*
  - *Free vs. Paid*
- Different uses
  - *Researcher: publishing data & locating data for reuse*
  - *Institution: reporting, monitoring, and governance*

OpenNEURO

zenodo

DRYAD

figshare

Data Repository Landscape: Introduction III

If the landscape is organized, then ..

- How do researchers find their way in this complex landscape?
- What is the sentiment amongst researchers to share data?
- How many datasets are actually being reused?
- How do institutions benefit from data repositories?
- …
- …
- In sum: what is the utilization of the landscape?

Data Repository Landscape: Survey I

- Qualitative and exploratory design
- Three sections:
  - *Researcher: Publishing Data*
  - *Researcher: Locating Data for Reuse*
  - *Institution: Reporting, Monitoring, and Governance*
- Response:
  - *N=6 filled out the survey*
  - *N=1 responded in dialog*

    Total N=7
  - *All N=7 represented supportive, central, services*
  - *The respondents represent 6 universities*

Data Repository Landscape: Survey II

Digital Science Report

The State of Open Data 2021

The longest-running longitudinal survey and analysis on open data

Foreword by Natasha Simons, Australian Research Data Commons (ARDC)

November 2021

Researchers: Publishing Data

- The sentiment on publishing data differs per research domain / faculty
  - *It is mainly driven by funder- and journal requirements and personal interest in Open Science*
  - *Other factors are e.g. confidence in skills, experience[s], and whether or not it is common practice*
- The proportion of researchers actively using a repository to publish data varies from 2% to 100%
  - *Differs according to faculty [strict policy]: on average ±20%*
- Barriers in using a repository are a lack of:
  - *Time, awareness, confidence, fears [privacy, being scooped], know-how, incentives*
- Recommendations are mostly to use the institutional- or a general repository [e.g. DANS]
  - *Only one responded commented on data sensitivity*
- Support varies from providing info only on repositories, to data curation, and helping with the actual uploading of data to the repository

“I SUSPECT THAT CONFIDENCE IN STATISTICS AND PROGRAMMING SKILLS ALSO MAKES RESEARCHERS MORE LIKELY TO SHARE THEIR DATA”

Is this (confidence & skills) something that we,
as data support staff, can help with?

“I FEEL LIKE THERE IS A GREAT IMBALANCE BETWEEN URGING DATA TO BE OPEN AND REUSABLE, BUT NOT ENOUGH SUPPORT (AT THE UNIVERSITY-LEVEL) AND INCENTIVES (FROM FUNDERS, THE SCIENTIFIC COMMUNITY, JOURNALS ETC.) TO ACTUALLY REUSE DATA”

Besides offering support, what incentives should be offered?

Is there a role for data support staff there?

Data Repository Landscape: Survey III

Researchers: Locating Data for Reuse

- The sentiment for reusing data differs per domain, but common across domains is that:
  - *There is a big push for and support in sharing data but not [yet] for locating data for reuse*
  - *Researchers who are not inclined to share data are less likely to look for data for reuse*
  - *If sharing data is common researchers have trust in data repositories and the data that is available for reuse*
- The proportion of researchers looking for data to reuse is estimated to be small [10% to 15%]
- Researchers find data through various ways:
  - *Publications in their field*
  - *Domain specific repositories [general repositories mentioned once]*
  - *Network [e.g. conference, personal network]*
  - *Trusted sources [e.g. government website]*

They find their own way

Data Repository Landscape: Survey IV

Researchers: Locating Data for Reuse

- Barriers in looking for data to reuse are:
  - *A lack of prestige, control over generated data, trustworthiness, data quality [assessment]*
  - *Datasets not suitable to be reused [specific research needs specific data]*
  - *Datasets do not have enough metadata [lack of descriptive information for reuse]*
  - *Researchers do not know where and how to locate data [no systematic approach / options]*
- If provided, support in looking for data to reuse is minimal
  - *Support on case by case basis*
  - *Keeping a list of repositories or giving directions to a repository finder [re3data.org]*
  - *Support for commercial databases only*
- Support in [systematic] data searchers is much needed
  - *Offer this in addition to or part of a course in [systematic] literature searches*
  - *Keep a list of trusted / much used repositories and their requirements to help researchers prep data*

“VERY SPECIFIC RESEARCH QUESTIONS NEED SPECIFIC DATA COLLECTION … DATA MIGHT NOT BE SUITABLE (DESIGN) TO BE REUSED”

Is data reuse only suited for large-scale studies?

“… I DON'T KNOW HOW MANY RESEARCHERS REUSE DATA, BUT I WOULD BE CURIOUS TO FIND OUT”

What can we do to gain better insight?

Data Repository Landscape: Survey V

Institutions: Reporting, Monitoring, and Governance

- Institutions actively want to monitor data output via e.g. PURE or their Institutional Repository
- The number of registered / reported datasets is thought to be inaccurate
- Numbers used for reporting only, not [yet] for directing or developing new services

“THE INFORMATION WILL BE USED TO MONITOR OUR POLICY”

What are the biggest challenges in monitoring research output?

And how can / should these be addressed?

What should be reported on?

DATA REPOSITORY FINDER

UKB RDM Theme Session

April 26th 2022

SOFTWARE CURATION & SENSITIVE DATA SHARING

UKB RDM Theme Session

April 26th 2022

Final Thoughts

- Lessons learned
  - *Landscape has improved, challenges remain the same*
  - *Open Science still seems very much a one-way street*
  - *The growing amount of information requires good organization / structuring*
  - *Collaboration is key*
  - *Software curation is equally important as data curation*

- Next theme session June 28th on Data Registration and Data Monitor