



# FAIRsFAIR

Fostering Fair Data Practices in Europe

## Planning to meet the costs of managing research data to be FAIR

Angus Whyte, Joy Davidson, Ryan O'Connor (Digital Curation Centre, UK)  
Annalisa Montesanti (Health Research Board, Ireland)



# Aims of our session

Offer resources you can adapt to help your researchers

1. Understand why they should budget for the costs of making data FAIR, and keeping it FAIR, and include them in grant applications
2. Appreciate the benefits that services may provide to justify their costs
3. Know about the different kinds of data management costs, including costs that funding bodies may allow to be charged to projects
4. Apply a template to help budget for the costs that may arise in preparing data to be FAIR

Share experiences and expectations about costing the preparation of FAIR data - invite you to join us in co-creating new guidance

# About us

Angus Whyte, Joy Davidson, Ryan O'Connor

- Research data specialists at Digital Curation Centre
- Applied research, support e.g. DMPonline, guidance resources, training, consultancy, events e.g. IDCC since 2004
- FAIRsFAIR project partner

Annalisa Montesanti

- Programme manager – Health Research Careers
- Health Research Board – FAIR Funding pilot



# About FAIRsFAIR

Objective is to supply **practical solutions** for the use of the FAIR data principles throughout the research data life cycle. Emphasis is on **fostering FAIR data culture and the uptake of good practices** in making data FAIR.

Call: H2020-INFRAEOSC-5c

Budget: 10 million euro

Length: 36 months

Starting date: March 1 2019

6 core partners/WP leads





**(5a)**

**EOSC**  
secretariat.eu

Setup and management of the EOSC Secretariat  
supporting the EOSC Governance

**Regional Nodes /  
Thematic Projects (5b)**

**NIXOS**  
National Initiative for Open Science in Europe

**EOSC SYNERGY**

**EOSC-Pillar**  
Coordination and Implementation of National Research Initiatives to support EOSC

**EOSC NORDIC**

**ExPaNDs**  
European Open Science Cloud Platform and Research Data Services

**FAIR (5c)**

**Synchronisation  
Force**

**FAIRSFAR**  
Fostering Fair Data Practices in Europe

**EGFC**  
European Group of FAIR Champions

**EOSC  
Governance Board**



**EOSC  
Executive Board**



**EOSC  
Working Groups**

Landscape WG Sustainability WG

Rules of Participation WG

Architecture WG FAIR WG

Skills and Training WG

**INFRAEOSC-5 Cross  
Project Collaboration  
Board (CPCB)**

**INFRAEOSC-5 Task Forces:**

- Landscaping
- FAIR data and Infrastructures
- Services onboarding
- National policies and governance
- Training and skills
- Dissemination and events

**FAIR WG Task Groups**

- FAIR practice
- Interoperability
- PIDs
- Metrics and certification

**EOSC Interest  
Groups**

- Researcher engagement and use cases
- Service and research product catalogue
- Federating core
- Glossary
- Cluster Collaboration

**"Horizontals"**

**ARCHIVOR**  
ARCHIVING AND PRESERVATION FOR RESEARCH EXPERIMENTAL DATA

**eInfra Central**

**RDA**  
RESEARCH DATA ALLIANCE

**OpenAIRE**

**OCRE**  
Open Cloud for Research Environments

**EOSC-hub**

**ESFRI Clusters**

**ENVRI**  
FAIR

**panosc**  
photon and nuclear open science cloud

**SSHOC**  
SOCIAL SCIENCE & HUMANITIES OPEN CLOUD

**ESCAPE**  
European Science Cloud for Astronomy & Particle Physics Open Science Environment

**EOSC-Life**

**Other FAIR Initiatives**

**FAIR4Health**

**STM**  
Research Data

**GO FAIR**

**FAIRsharing.org**  
standards, databases, policies

**FAIRplus**

# Putting recommendations into practice



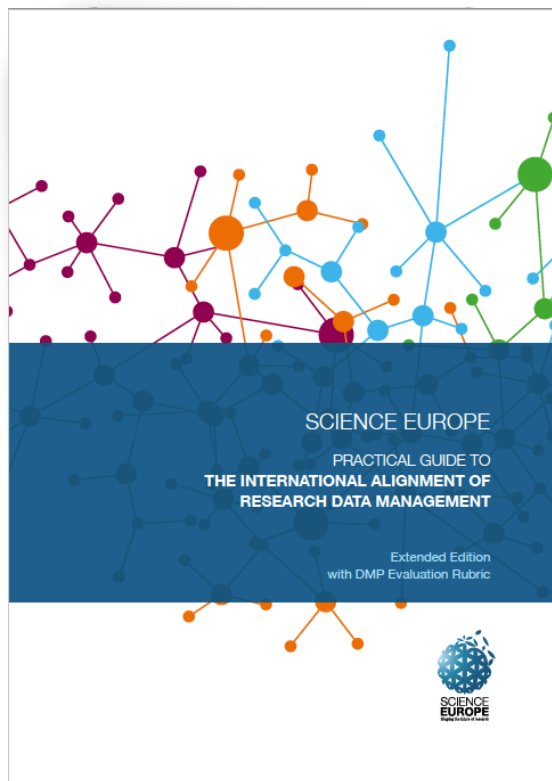
## Actions for funders, institutions, data stewards and communities

### “Rec. 18: Cost data management

Research funders should require data management costs and other relevant costs to be considered and included in grant applications where relevant. To support this, **detailed guidelines** and worked examples of eligible costs for FAIR data should be provided. ”

Action 18.1: **Questions about the costs of data management, curation and publication should be included in all DMP templates**. Information from existing and completed projects should be used to retrospectively identify costs and **develop examples and guidelines** based on these. Funders, institutions and data services should collaborate on retrospective analysis, including the cost of long-term curation.

# Funders expect DMPs that identify costs



## Science Europe DMP Evaluation Rubric

Section 6 Data Management Responsibilities and Resources  
“What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?”

Sufficiently addressed the DMP... Provides clear estimates of the resources and costs (for example storage costs, hardware, staff time, costs of preparing data for deposit, and repository charges) that will be dedicated to data management and ensuring that data will be FAIR and describes how these costs will be covered. Alternatively, there is a statement that no additional resources are needed.

# Making the case for costing FAIR data



## €10.2bn annual costs of **not** having FAIR data

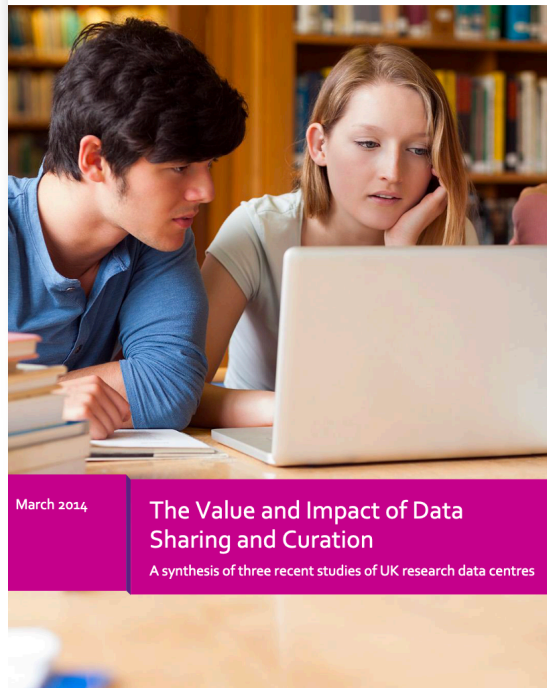
1. **'Time Spent'** - Time spent on searching for data, often with incomplete metadata **€4.5bn**
2. **'Cost of Storage'** - Additional copies are made that would not be necessary if data was made FAIR **€5.3bn**
3. 'Licence costs' - Cost of extra licences that researchers have to pay to access non FAIR data.
4. 'Research retraction' - Non FAIR research would lead to less article retractions due to non-reproducibility, errors, fraud, plagiarism etc.
5. 'Double funding' - Non-FAIR research leads to duplication of research effort.
6. 'Interdisciplinarity' - Added value of interdisciplinary research made possible by FAIR data.
7. 'Potential economic growth' - GDP growth and number of jobs created if FAIR data was widely available.

DG Research & Innovation/ PwC 2018 Cost-benefit analysis for FAIR research data

<https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1>

# Making the case for costing FAIR data

Jisc



## Returns on the investment in FAIR data

Qualitative and quantitative estimate of financial benefits from data centres- Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS) and the British Atmospheric Data Centre (BADC)

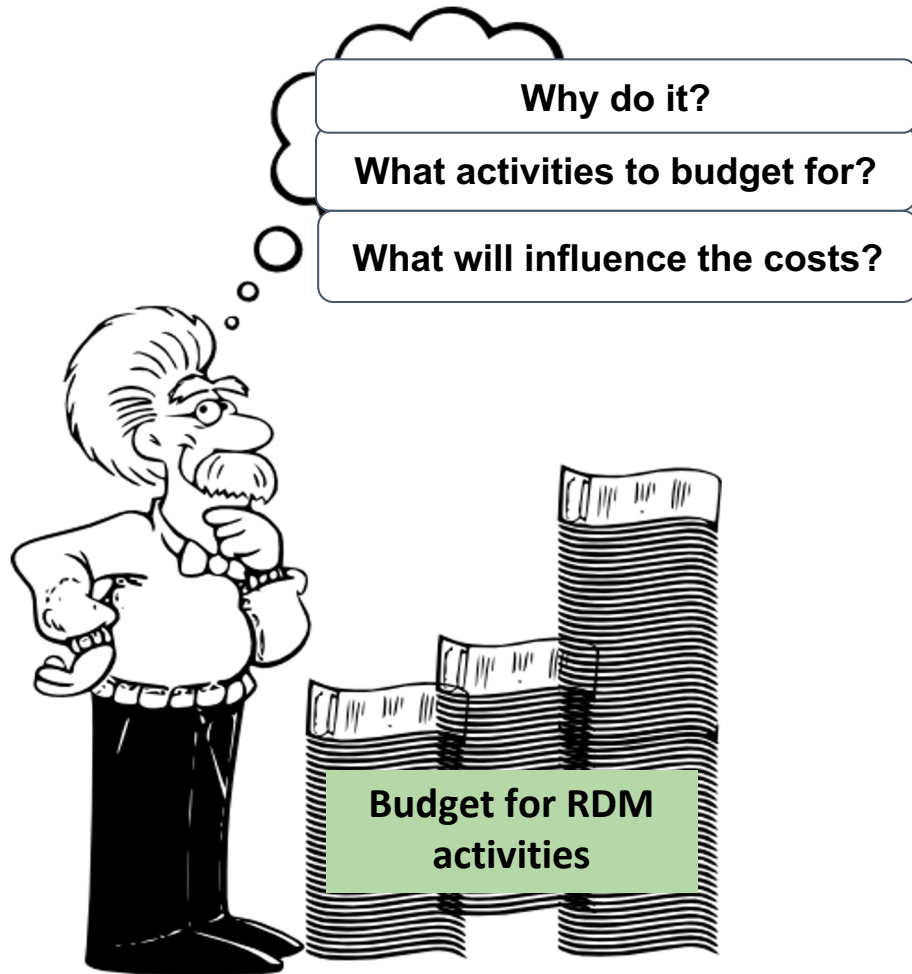
“ESDS facilitates additional use which realises additional returns that could be worth some £58 million to £230 million over 30 years (net present value) from one year’s investment expenditure – **effectively, a 2.5-to 10-fold return on investment.**”

Beagrie, N. and Houghton, J. (2014) for Jisc. *The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres.*

[http://repository.jisc.ac.uk/5568/1/iDF308\\_-\\_Digital\\_Infrastructure\\_Directions\\_Report%2C\\_Jan14\\_v1-04.pdf](http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf)



# Challenges for the individual PI



## Why

Incentivization for data management is still work-in-progress  
Low awareness of how to select data that should be FAIR

## What

Low awareness of what can be included in budget

## How

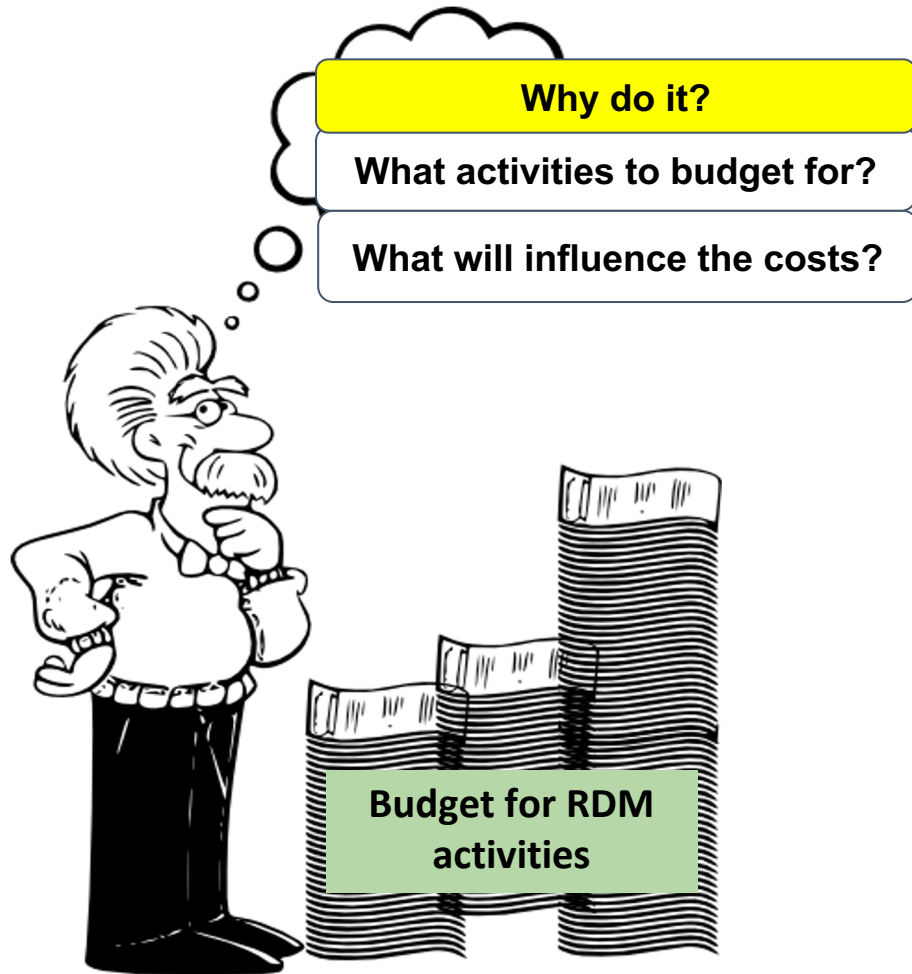
Time costs of preparing data to be FAIR tend to be underestimated

- Data cleaning/ wrangling
- Anonymisation
- Transcription
- Applying metadata standards

Examples help but are rarely shared, and are context specific

Guidance lacking on **cost drivers** - factors influencing relative costs

# Challenges



## Why

Incentivization for data management is still work-in-progress

Low awareness of how to select data that should be FAIR

**DCC Guide:** [Five steps to decide what data to keep](#)

## What

Low awareness of what can be included in budget

## How

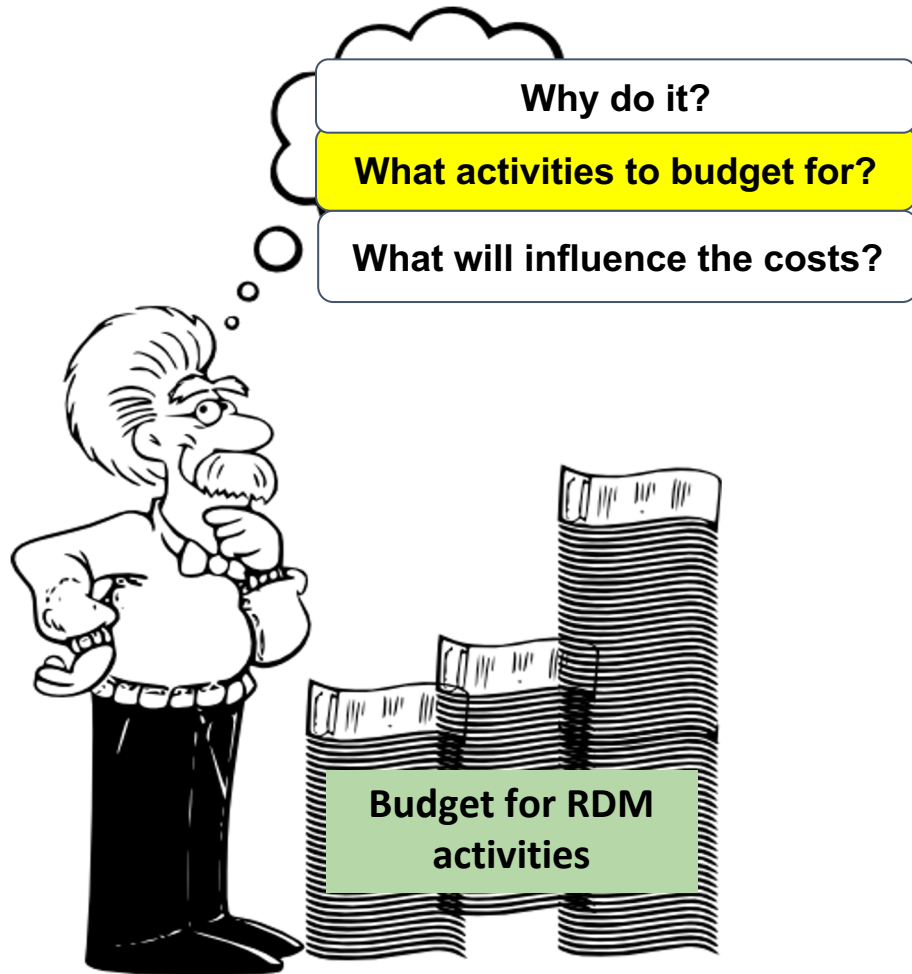
Time costs of preparing data to be FAIR tend to be underestimated

- Data cleaning/ wrangling
- Anonymisation
- Transcription
- Applying metadata standards

Examples help but are rarely shared, and are context specific

Guidance lacking on cost drivers - factors influencing relative costs

# Challenges



## Why

Incentivization for data management is still work-in-progress  
Low awareness of how to select data that should be FAIR

## What

Low awareness of what can be included in budget

*Current guidance identify activities but funders differ on which costs are eligible*

## How

Time costs of preparing data to be FAIR tend to be underestimated

- Data cleaning/ wrangling
- Anonymisation
- Transcription
- Applying metadata standards

Examples help but are rarely shared, and are context specific

Guidance lacking on cost drivers - factors influencing relative costs

# Current sources for researchers

- STFC: Horizon 2020 Costing Guide  
[https://twiki.pp.rl.ac.uk/twiki/pub/Main/PpdBids/EC\\_Horizon\\_2020\\_General\\_costing\\_guide.pdf](https://twiki.pp.rl.ac.uk/twiki/pub/Main/PpdBids/EC_Horizon_2020_General_costing_guide.pdf)
- UK Data Service: Data management costing tool and checklist  
<https://www.ukdataservice.ac.uk/manage-data/plan/costing>
- LCRDM: Guide Research Data Management and Costs  
<https://tinyurl.com/y9g2dr3q>
- OpenAIRE: How to identify and assess Research Data Management (RDM) costs  
<https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs>



## ESTIMATING COSTS RDM TOOL

INFOGRAPHIC: WHAT WILL IT COST TO MANAGE AND SHARE MY DATA?

### *Preparing*

#### **Make a Data Management Plan**

Make a DMP before you start creating data; make decisions about managing your data. You can find the template for H2020 DMPs [here](#).

2 hrs to 2 days, depending on the complexity of your project

Check if there is a department within your organization to support data management planning.

### *1. Data Collection*

#### **Acquiring External datasets**

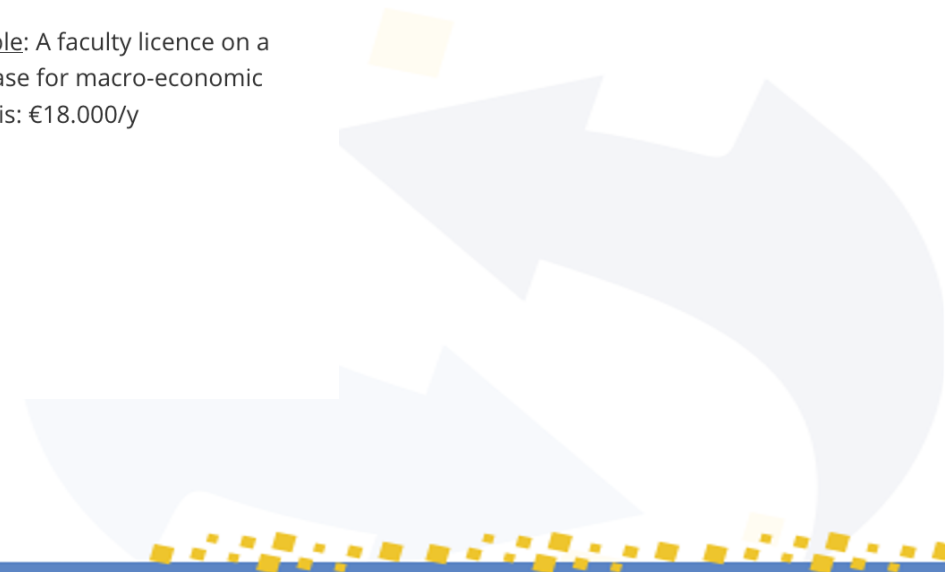
Do you plan to use existing data, and is the data available at a commercial partner?

Your library may be able to help you acquire a license to a crucial database

In research data repositories, data can be available at no or low costs

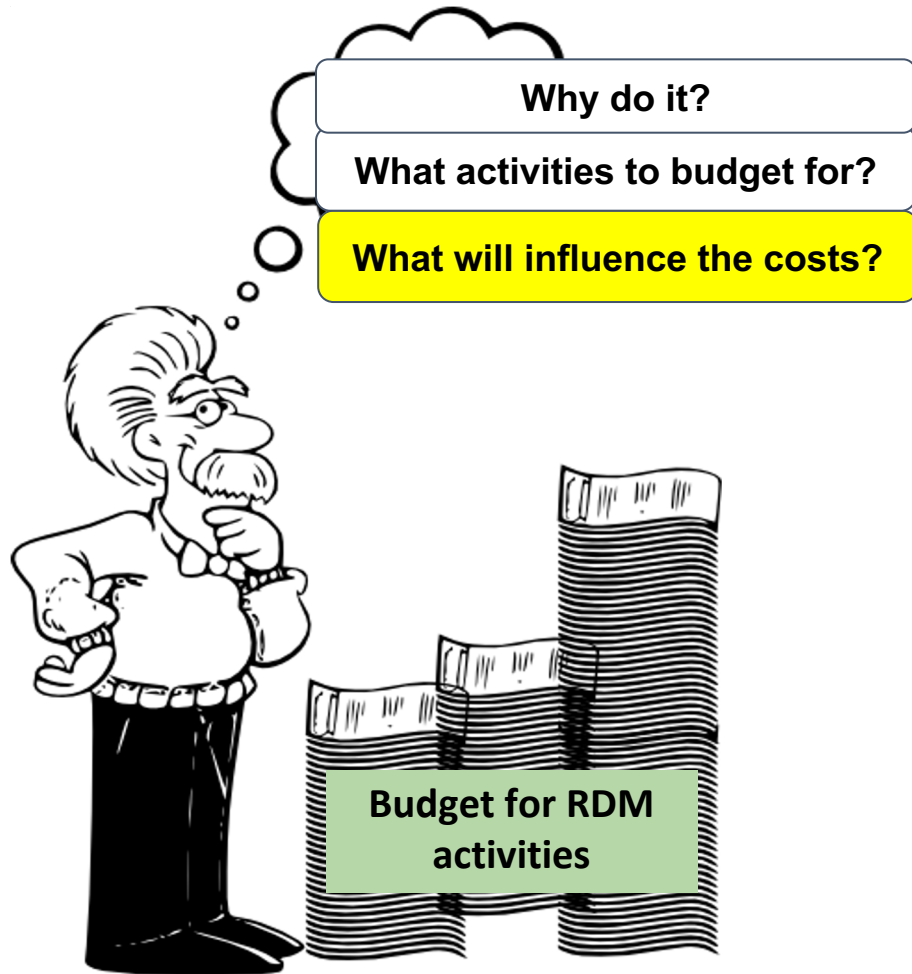
Example: A faculty licence on a database for macro-economic analysis: €18.000/y

<https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs>





# Challenges



## Why

Incentivization for data management is still work-in-progress  
Low awareness of how to select data that should be FAIR

## What

Low awareness of what can be included in budget

## How

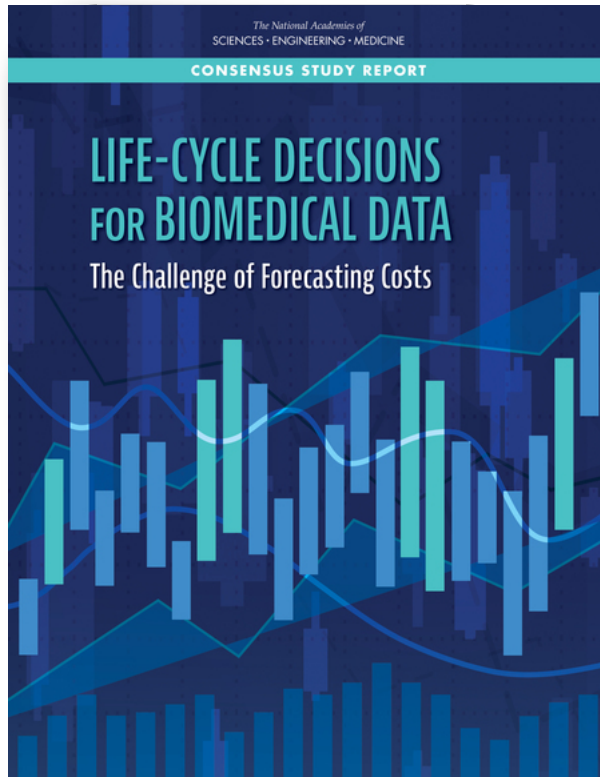
Time costs of preparing data to be FAIR tend to be underestimated

- Data cleaning/ wrangling
- Anonymisation
- Transcription
- Applying metadata standards

Examples help but are rarely shared, and are context specific

Guidance lacking on **cost drivers** - factors influencing relative costs

# Activity costing in context



## 5 Steps to cost estimation

1. Scope - why data needs to be FAIR
2. **Cost drivers** relating to the data, its expected lifecycle, contributors and users
3. Reuse value arising from the data, its production, its contributors and users
4. Activities, staffing, infrastructure, roles and resources needed
5. **Other major cost drivers** > relative costs, quantifiable short-term costs

Adapted from: National Academies of Sciences, Engineering, and Medicine. 2020. *Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*. Washington, DC: The National Academies Press.

<https://doi.org/10.17226/25639>

# Questions to guide estimation of relative cost

## Report provides a template based on 10 main cost drivers

1. **Content** (e.g., data size, complexity, and diversity; metadata requirements, depth versus breadth, processing level and fidelity; and replaceability of the data);
2. **Capabilities** (e.g., user annotation, persistent identifiers, citation, search, data linking and merging, use tracking, and data analysis and visualization);
3. **Control** (e.g., content, quality, access, and platform);
4. **External context** (e.g., replication, external information dependencies, and distinctiveness);
5. **Data life cycle** (e.g., anticipated growth, updates and versions, useful lifetime, and offline and deep storage);
6. **Contributors and users** (e.g., contributor base, user base and usage scenarios, training and support requirements, and outreach);
7. **Availability** (e.g., tolerance for outages, currency, response time, and local versus remote access);
8. **Confidentiality, ownership, and security** (e.g., data privacy issues and licensing);
9. **Maintenance and operations** (e.g., periodic integrity checking, data transfer capacity, risk management, and system reporting requirements); and
10. **Standards and regulatory compliance** and other governance concerns.

# Questions to guide estimation of relative cost

## Report provides a template based on 10 main cost drivers

1. **Content** (e.g., data size, complexity, and diversity; metadata requirements, depth versus breadth, processing level and fidelity; and replaceability of the data);
2. **Capabilities** (e.g., user annotation, persistent use tracking, and data analysis and visualization);
3. **Control** (e.g., content, quality, access, and security);
4. **External context** (e.g., replication, external dependencies, and legal requirements);
5. **Data life cycle** (e.g., anticipated growth, update frequency, and storage);
6. **Contributors and users** (e.g., contributor buy-in, user requirements, and outreach);
7. **Availability** (e.g., tolerance for outages, current and future needs);
8. **Confidentiality, ownership, and security** (e.g., data privacy issues and licensing);
9. **Maintenance and operations** (e.g., periodic integrity checking, data transfer capacity, risk management, and system reporting requirements); and
10. **Standards and regulatory compliance** and other governance concerns.

**Q. for breakout: could this help identify relevant cost drivers and DMP questions for Social Science and Humanities?**

# Breakout

- Feedback on the template- do cost drivers and questions deal with the ‘known and unknowns’ in your experience? (Angus Whyte)
- HRB Case study – budgeting for FAIR data (Annalisa Montesanti)
- Overview of OpenAIRE Costs Infographic (Ryan O’Connor)
- Exercise - overview of activity and instructions for Thursday feedback session (Joy Davidson)

