ENSURING TRUSTED FAIR DATA FOR RESEARCH:

INFRASTRUCTURE AND POLICY ISSUES

Carthage Smith, OECD Global Science Forum

EC eIRG, webinar, 1 December, 2020

OECD
BETTER POLICIES FOR BETTER LIVES

Menu

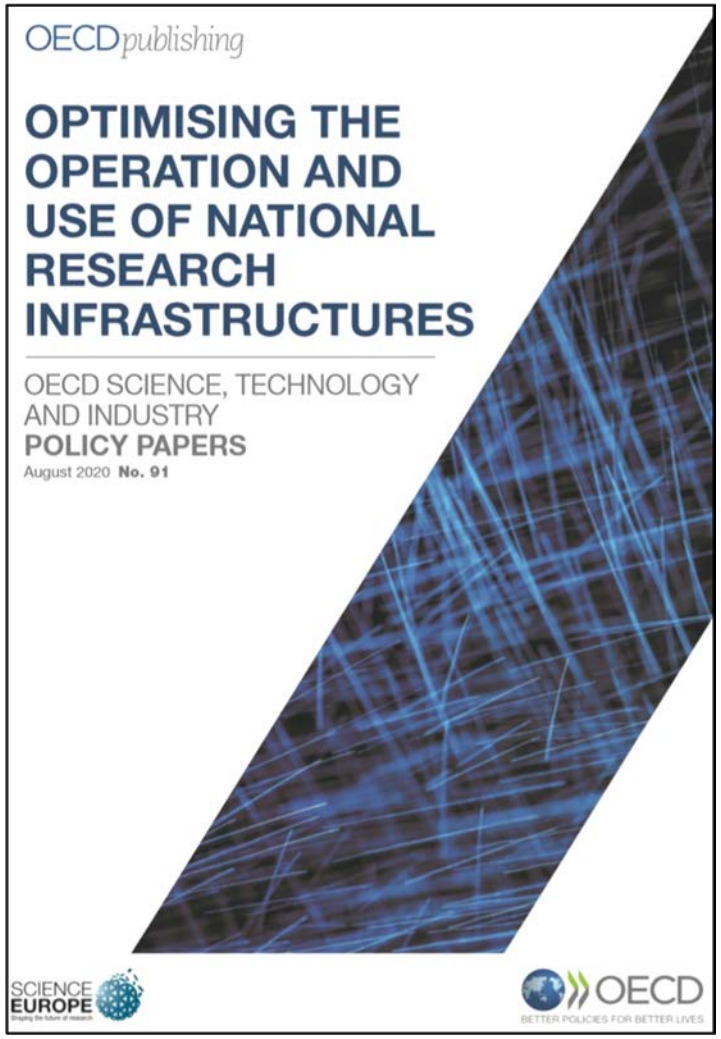
I. Outcomes from the ICRI 2018 session on Research Infrastructures, data and trust

II. Business models for sustainable data repositories

III. International cooperation and data networks

IV. Digital wokforce capacity and skills for data intensive science

I. Research Infrastructures and trusted data



RI ecosystem and data

- Trusted data and trusted cyber-infrastructure and trusted connections – need open PIDs
- Think in terms of the research data life-cycle
- Must design data services/cyber-infrastructure around specific user needs
- Build on best practices: different domains at different levels of preparedness
- Support development of open tools and software for connecting

➢ Develop a researcher-centric data ecosystem

Challenges

- Plethora of competing standards
- Standards for interoperability (disciplines – domains – cross-domain)
- Gaps between HPC infrastructures and traditional repositories
- Ownership of data
- Data changes in real time – requires snapshots, versioning
- Incentives and measurement - consequences of metrics

➢ Social, behavioural and technical barriers

Human infrastructure

- Need to build a workforce of data scientists and stewards
- Need new cohorts of data specialists and stewards
- Need critical skills and education to deal with the data rich World of the future
- Technology (AI) can help but not replace human analysis.
- Generational latency because of mentorship. Need cultural change

Google & Co.

- Google solns for storage and access to computing power meet a need
- Google are solving discovery challenge and driving semantic web, AI etc.

But

- Google pays more and is more attractive vs academia
- Google de facto sets the standards
- Google threaten to drive out (public) competitors
- ?long-term sustainability and customer-dependency/lock-in

Funder (policy) perspectives

- On-line Poll: - ~70% of audience saw incentives and mandates as a critical function
- Data sharing policies are not the same as ‘open data’.
- Balance national benefits and internat. cooperation?
- Communities need to triage and decide what must be kept and what is disposable
- Data life-cycle approach and importance of data management plans

II. Who pays for FAIR data and how?

Objectives

1. Identify and describe existing revenue sources and business models
2. Test potential business models with various stakeholders, including funders
3. **Make policy recommendations to promote sustainable business models** for research data repositories

Elements of a successful business model

Funding Source	Pros	Cons
Structural funding	• Compatible with open data principles. • Longer-term stability. • Larger-scale and efficiencies. • Flexible as to allocation.	• Fixed, multi-year may not scale easily. • Competes with research funding. • Too many eggs in few baskets.
Host or institutional funding	• Compatible with open data principles. • Longer-term stability. • Efficiencies through sharing services. • Close to researchers (customers).	• Limited purview, with focus on local community. • May lead to fragmentation of domain data and lower interoperability. • Limited incentive to add value to data and develop related services.
Data deposit fees	• Compatible with open data principles. • Demand oriented and scales with demand (data ingest). • Researchers price sensitivity ensures cost constraint. • Open data is part of research and its funding.	• Cost disincentive to depositing, so depends on strong mandates. • May lead to low level of curation to contain costs (price). • May be difficult for repository to compete for deposits with comparable repositories that do not charge.
Data access charges (subscriptions or use fees)	• Users pay for what they want, so funding reflects value. • More market-oriented approach may provide incentive for cost constraint.	• Not compatible with open data principles and many funder mandates, limiting the potential market size. • Charges limit use and will reduce the value of data. • Revenue scales with use and not ingest or curation costs. • Vulnerable to funding cuts.
Diversification of revenue sources	• No single source of failure. • Can maintain compatibility with open data principles. • Flexible and enables experimentation with new services.	• May lead to higher transaction costs (managing multiple funding sources). • May lead to Mission drift.

Recommendations

1. All stakeholders should recognise that **research data repositories are an essential part of the infrastructure for open science**.
2. All research data repositories should have **a clearly articulated business model** and value proposition(s).
3. Sponsors need to **consider the ways in which data repositories are funded**, and the pros and cons of various funding mechanisms in different circumstances.
4. Research data repository business models are **constrained by, and need to be aligned with, policy regulation** (mandates) **and incentives** (including funding).
5. In the context of financial sustainability, opportunities for **cost optimisation should be explored**.

III. International networking

Objectives

Establish principles and policy actions that can support open and sustainable international research data networks:

1. When is a data network needed?
2. How can governments use networks to maximize research data openness and reuse?
3. What is the best governance model for a particular network?
4. What interoperability arrangements are necessary for the effective operation of the network?
5. What business models can sustain a network over time?

Findings

- The most successful networks have **engaged and supportive users** who clearly understand and value the services of the network.
- The top issue faced by data networks in open sharing of data is **the varying attitudes and policies across countries.**
- Different research communities require **different data networks** because **the cultures** of data sharing vary.
- The most difficult aspects of **interoperability** are rooted in **human relationships and trust.**
- Developing a **coherent and sustainable business model** is a central challenge for virtually all data networks.

Recommendations

1. work toward **common definitions of, and agreements on, open data**. What is open data in different domains?
2. work toward commonly agreed and enforced **legal and ethical frameworks for the sharing of different types of public research data**.
3. Funders and host institutions should view internationally coordinated data networks as a **long-term strategic investment**
4. Networks should have clear business models, including value propositions and **measures of success** that are relevant to their different stakeholders and these measures **should be monitored**.

IV. Human infrastructure

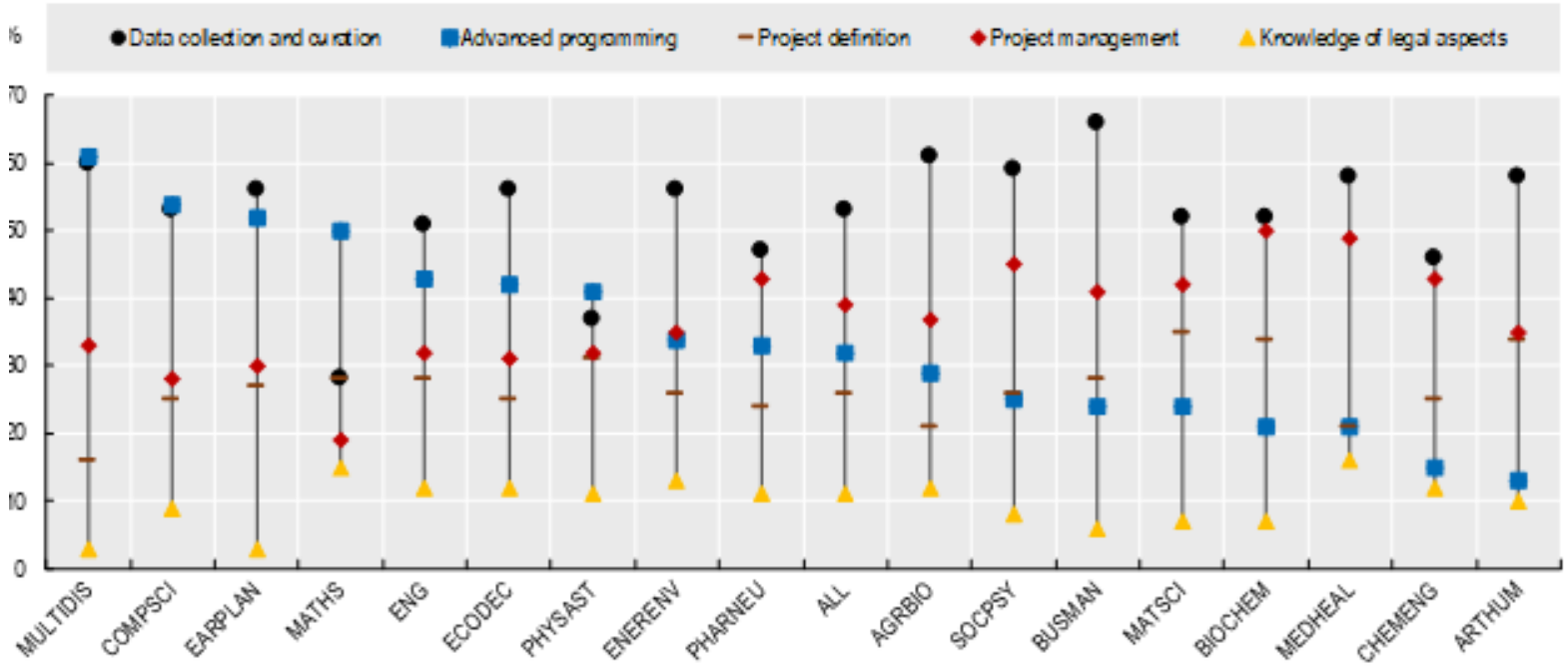
Objectives

- Identify skill and capacity needs and gaps for data-intensive science in academia
- Identify policy actions to address these gaps
- Promote mutual learning and exchange of good practices

Different needs in different research domains

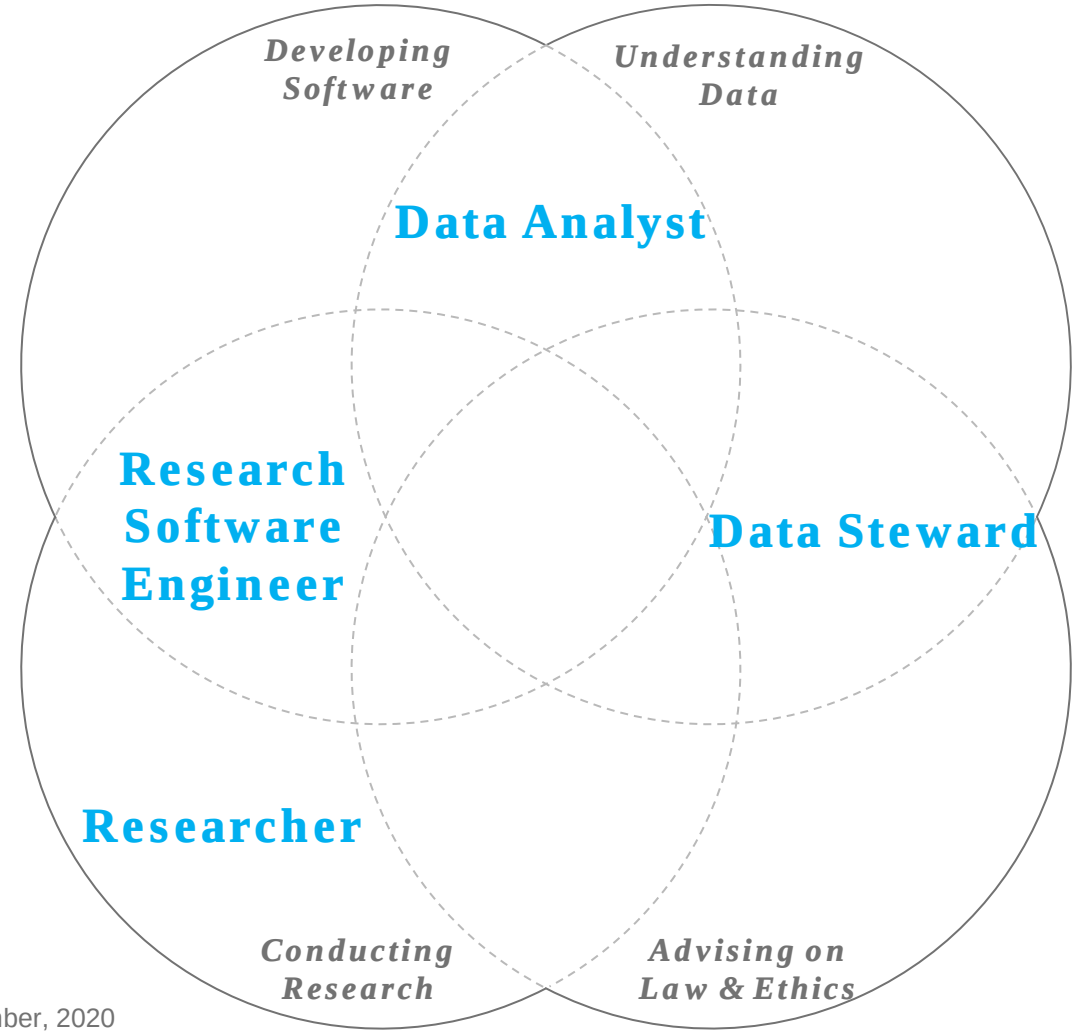
Figure 5.3. Most important skills for scientific authors' research work

Percentage of authors who deem each type of skill as important



Source: OECD, 2020, *Charting the Digital Transformation of Science*, in preparation

1. Digital skills and stereotype roles



5 areas for (policy) action

Integrate digital workforce capacity development into broader science policy frameworks and actions, e.g. for open science and research integrity.

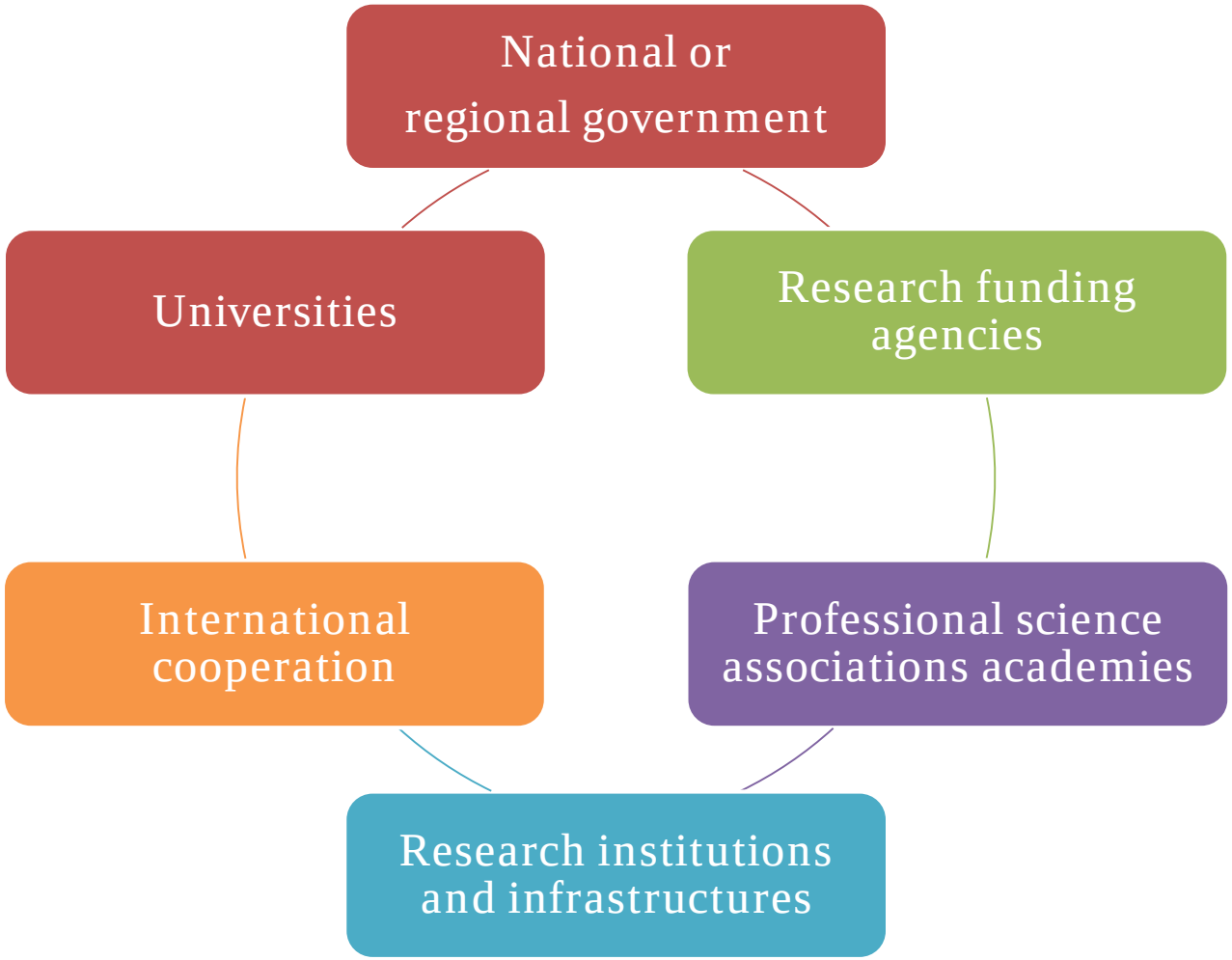
Identify the key competencies, skills and roles required for data-intensive science in different contexts.

Implement changes in academic evaluation and reward systems in order to attract and retain diverse digitally skilled staff.

Support training in foundational digital skills and more specialized skills for scientists and research support professionals.

Support development of communities for new professional roles, learners and trainers.

Requires complementary action at different scales and from different actors



OECD policy reports

2015: Making Open Science a reality

2016: Research Ethics and New Forms of Data for social and economic research

2017: Business models for sustainable research data repositories

2017: Co-ordination and support of international research data networks

2020: Digital workforce capacity and skills for data intensive science

2021: revised OECD Recommendation on Access to Research Data from Public Funding.