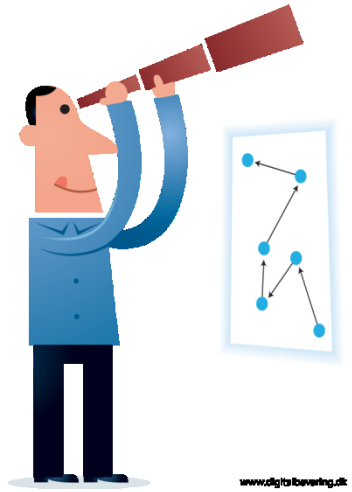


Collecting research data

Andrew Cox

Information School, University of Sheffield, United Kingdom

a.m.cox@Sheffield.ac.uk

A thought experiment

If you went to visit a senior researcher in their office to talk to them about their research data:

- How much data would they have?
- How do they store and back it up?
- Can they always refind it?
- Who do they believe owns it?
- Do they share it?

Drivers for Research Data Management

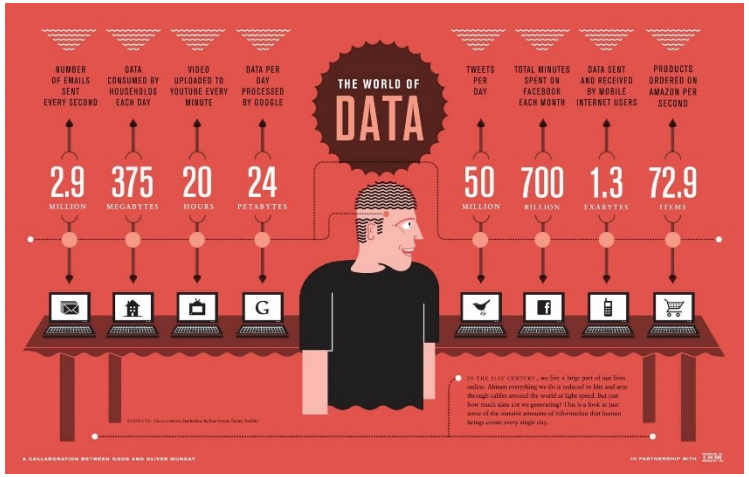
Drivers for RDM

- “Data deluge”… big data
- Government / Research Funder mandates
  - OECD *Principles and Guidelines for Access to Research Data from Public Funding* 2004/2007
    - To validate new research, to enable new research and collaboration, and for training
  - RCUK common principles
    - Concordat on open research data (2016)
  - EU Horizon 2020 Guidelines on data management (2016)
  - FAIR principles (2016)
- Crisis of replication/reproducibility
- Open Science movement
- Information security and confidentiality, including GDPR in Europe
- Journal mandates, publisher services

Big data examples

- Astronomy
- Particle physics
- Genomics
- Social media data

- “Every Six Hours, the NSA Gathers as Much Data as Is Stored in the Entire Library of Congress.”
- “In less than two years Instagram has already hosted more than 500 million images — more than 30 times greater than the entire photo archive of the Library of Congress.”
- “15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress”
- https://blogs.loc.gov/thesignal/2012/03/how-many-libraries-of-congress-does-it-take/



Duffy (2013) on scale of the data issue at University of Birmingham

- 3000 items in institutional repository
- 50,000 items in special collections
- 75,000 publications for REF
- 2,700,000 items in library
- 700,000,000 folders in top 100 accounts
- Perhaps 1,000,000,000 folders for the whole university

UK “Concordat on open research data”

1. Open access to research data is an enabler of high quality research, a facilitator of innovation and safeguards good research practice
2. There are sound reasons why the openness of research data may need to be restricted but any restrictions must be justified and justifiable
3. Open access to research data carries a significant cost, which should be respected by all parties
4. The right of the creators of research data to reasonable first use is recognised
5. Use of others’ data should always conform to the legal, ethical and regulatory frameworks including appropriate acknowledgement
6. Good data management is fundamental to all stages of the research process and should be established at the outset
7. Data curation is vital to make data useful for others and long-term preservation of data
8. Data supporting publications should be accessible by the publication data and should be in citeable form
9. Support for the development of appropriate data skills is recognised as a responsibility for all stakeholders
10. Regular reviews of progress towards open access to research data should be undertaken

FAIR principles

- Findable
- Accessible
- Interoperable
- Reusable

- https://www.force11.org/group/fairgroup/fairprinciples

“Crisis of replication (/reproducibility)” and trust in science

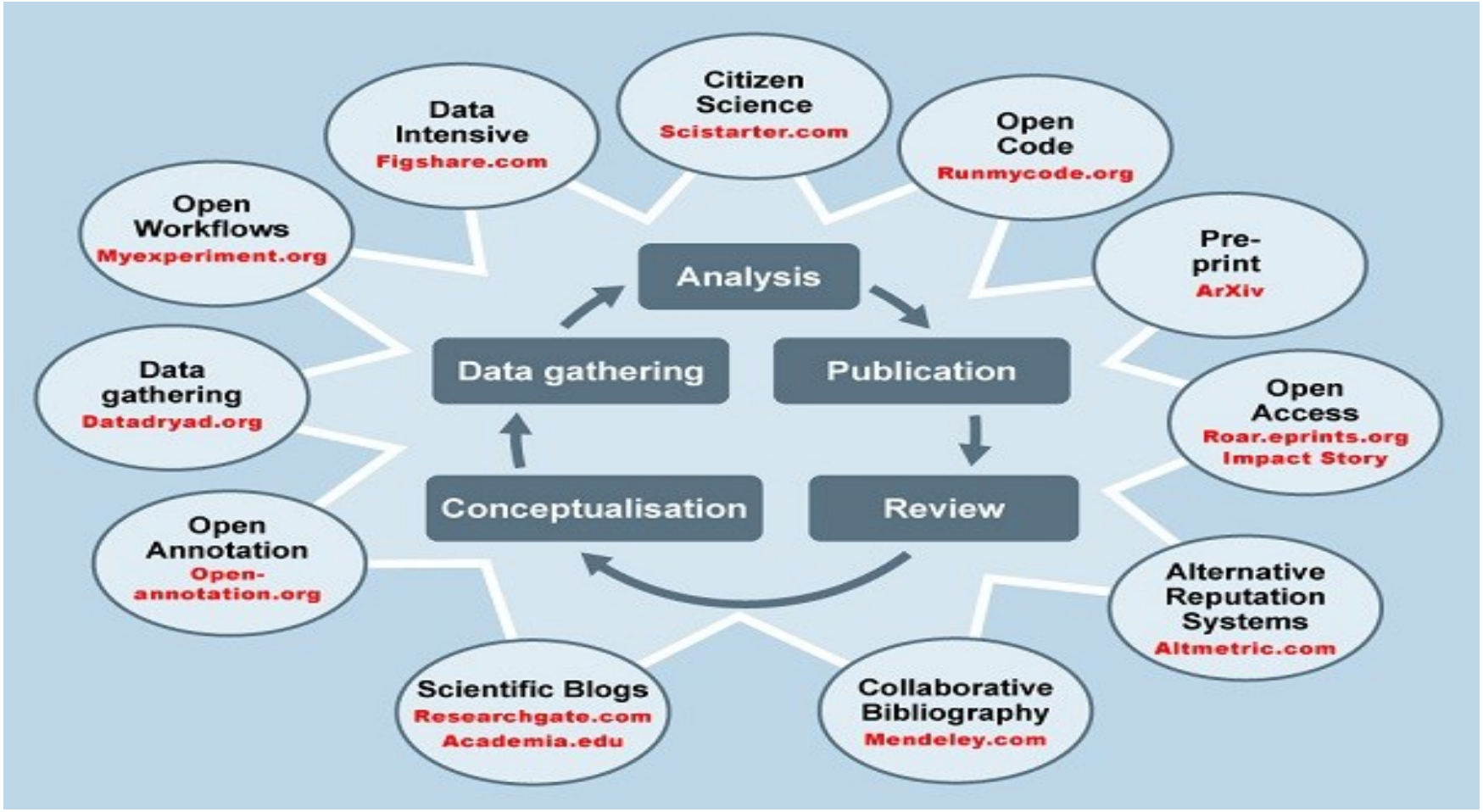
- Medicine, psychology, genetics and biology
  - Fraud
  - Failures of replication (eg 60-70% of biomedical findings could not be reproduced by Bayer (Ger), Amgen (USA))
  - Statistical fallacies (Ioanndis, 2005)
  - “Questionable research practices” under pressure to publish

    - 0.6% of psychologists surveyed admitted to falsifying data
    - 22% round off p-values (if you get a result of 0.054 you round to 0.05 to get significance),
    - 63.4% said that they did not report all dependant measure.
    - 55.9% said they decided to analyse and then decide if they were to gather more data
    - 38.2% said that they decided whether to exclude data after looking at the effect of doing so.

    (John et al. 2012)

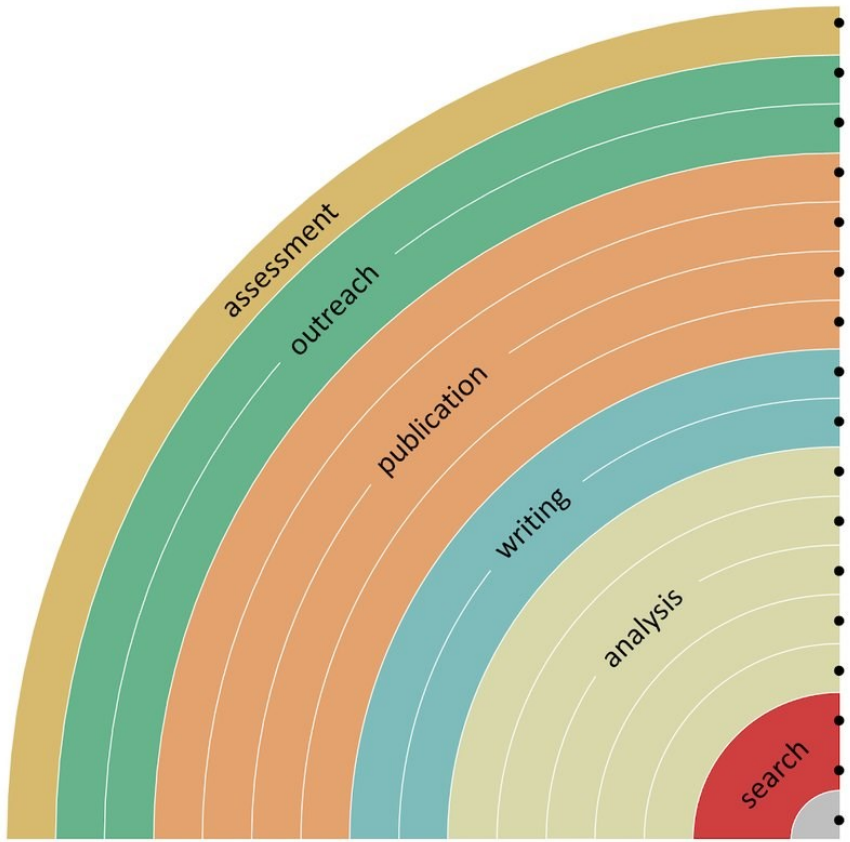
- Prompts increasing replication studies; easier to publish such studies
- Reproducibility Project in psychology and Reproducibility Initiative in biomedicine

Open Scholarship



https://www.libraryjournal.com/?detailStory=welcome-to-science-2-0-open-access-in-action

You can make your workflow more open by …



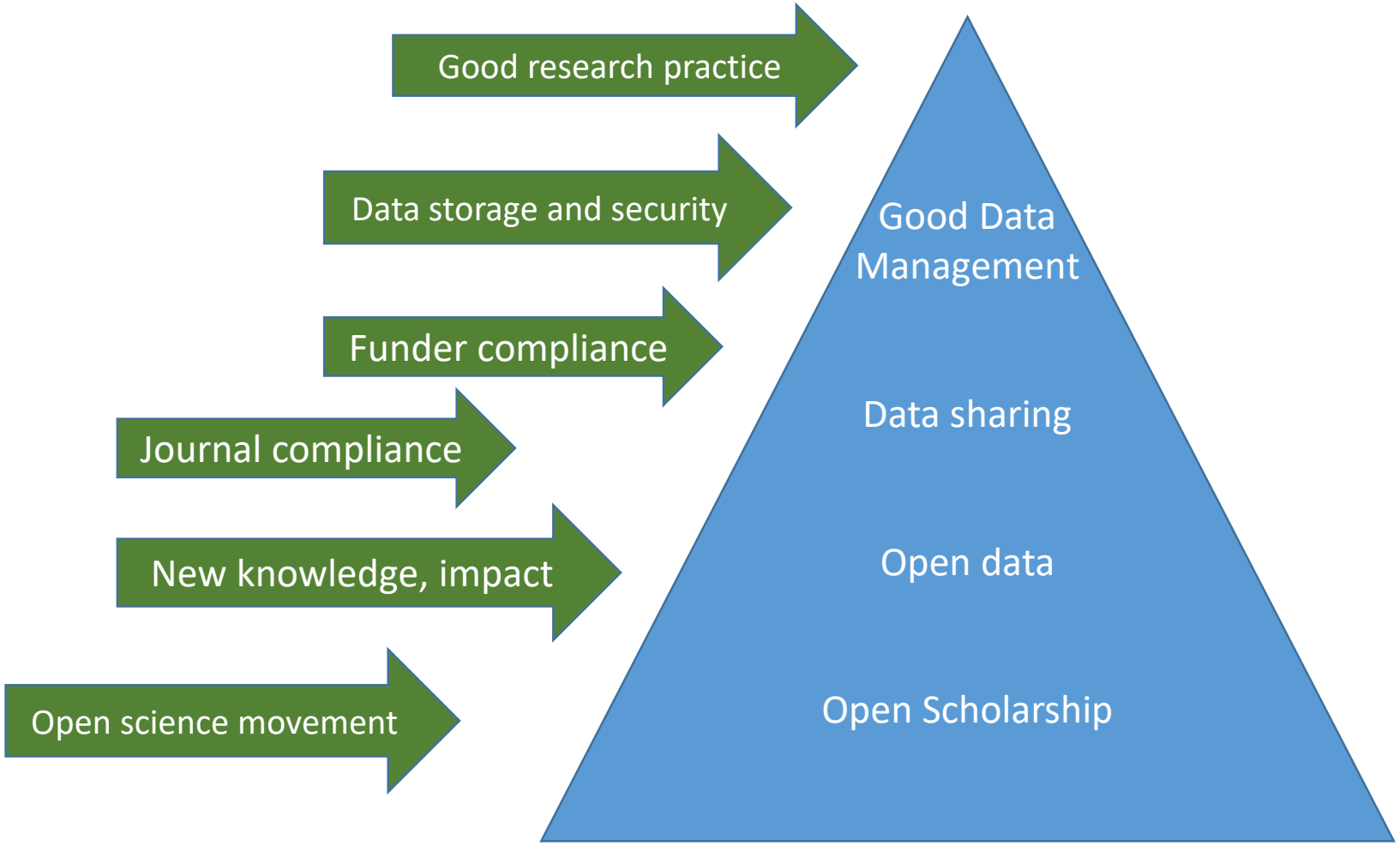
- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, 'green' or 'gold'
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO

DOI: 10.5281/zenodo.1147025

Masuzzo (2017) “What can you do?”

1. Use and cite existing public data
2. Share your research data and create relevant metadata
3. Release code
4. Post free copies of your research articles online
5. Post preprints of research manuscripts
6. Publish in open access journals

Force field analysis of RDM and open scholarship



Challenges of RDM

What is data like?

- Volume: scale of digital research data
- Variety of research data: print and electronic; many different file types and standards
- Some researchers use other terms, eg “sources” “primary resources”
- Ownership
- Fragility of digital data
- Complex: data can be produced from other data
- Exists or is reused for other purposes
- What is the data? The sound files of interviews, the transcripts, summaries of interviews, notes on interviews, nvivo files???

The variety of data

- Interviews and focus groups and questionnaires
- Weather measurements eg field or sensor data
- Results from experiments
- Government records
- GIS data
- Simulation data
- Log data
- Field notes
- Software
- Images (e.g. brain scans)
- Quantitative data (e.g. household survey data)
- Historical documents
- Moving images
- Physical objects: such as bones or blood samples
- Digitised photos / born digital photos
- Social media data: tweets
- Metadata

A definition of data

- “Data are facts, observations or experiences on which an argument or theory is constructed or tested. Data may be numerical, descriptive, aural or visual. Data may be raw, abstracted or analysed, experimental or observational. Data include but are not limited to: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artefacts; specimens; samples.” (University College London)

More definitions

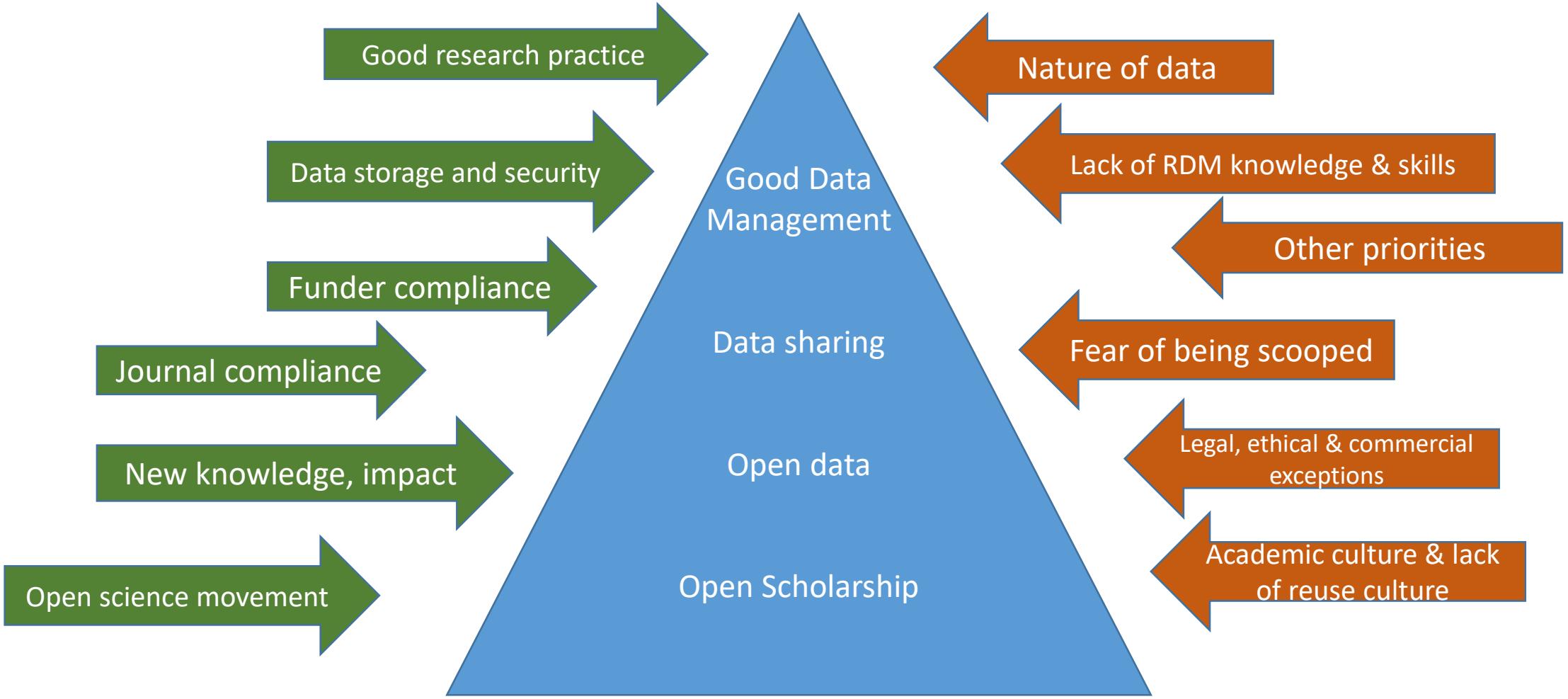
- “Research data is defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created.” (EPSRC)

- “Anything you perform analysis on” (Briney, 2015: 6)

Some issues for researchers around data sharing

- Desire to keep control over data after investment of time/ fear of being scooped
- Legal, ethical and commercial reasons for confidentiality
- Dislike of bureaucracy and a lack of time to process datasets
- Lack of know-how, skills and confidence (eg metdata, data selection, choosing a repository, licensing)
- Questions over the usefulness of data to a wider audience
- Issues with the feasibility of data reuse (methodological concerns)
- Lack of a reuse culture
- Fear of criticism and misuse
- Lack of direct incentives
  - “Does your institution have incentives, rewards or recognition for faculty/academic staff in your institution who engage in research data management good practice?”
  - 8/209 said yes

Force field analysis of RDM and open scholarship



The strengths of these forces differ for different individuals, subject fields, institutions – what do you think are the key drivers in your context?

For your context

- Adjust the central triangle
- Resize the arrows representing drivers and barriers to reflect their strength

International Library Research Data Services (RDS) survey context

- Conducted Mar-Apr 2018
- One response invited per institution
- Low numbers by country v stat significance
- Non response bias

Response rate

- Australia 34/39
- Canada 24/74
- Germany 23/250
- Ireland 11/12
- Netherlands 6/16
- New Zealand 8/8
- UK 80/169
- USA 23/86

Survey results: Drivers and barriers to RDS

- Drivers
  1. Compliance
  2. Library capabilities
  3. Researcher need

- Barriers
  1. Resources
  2. Skills
  3. Engagement of academic staff

- “Now - integrity, reproducibility, trust in research, resource justification for funders
- Future - strategic and economic impetus for AI, smart data visualisation, data skills ”

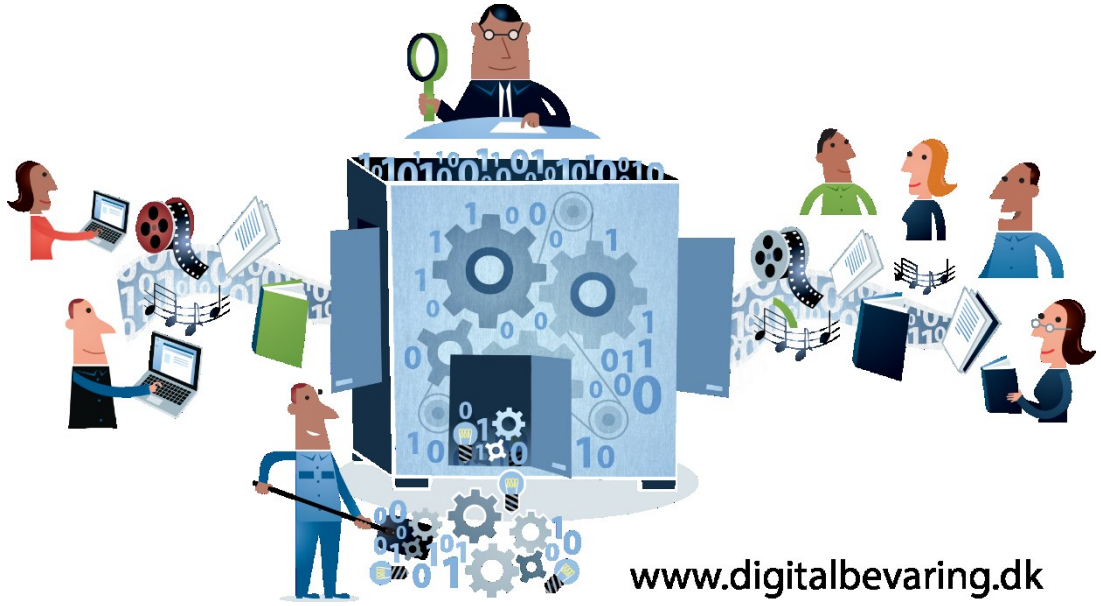
- “libraries can play an important part, since they are metadata experts; offering document repositories it makes sense to also engage in data repositories and also knowledge graphs / data integration (libraries offer knowledge - beyond books)”

- “We need to make the "library" profession something that people with the right aptitudes and skill-sets will come into. I do not have a single "librarian" in my Open Research team. We need to be better at recruiting people with an affinity to this type of work rather than "people who love books".

- “The chicken and egg scenario of RDM remains. You need to have a service in place to promote effective RDM practices, but it is hard to fund and develop a service without evidence of demand for that service, or to decide how to scope it. We are still in advance of academic demand for RDM” (UK)

- “A major challenge is doing this as well as everything else. Also, RDM is much more complex than most other things we do.” (UK)

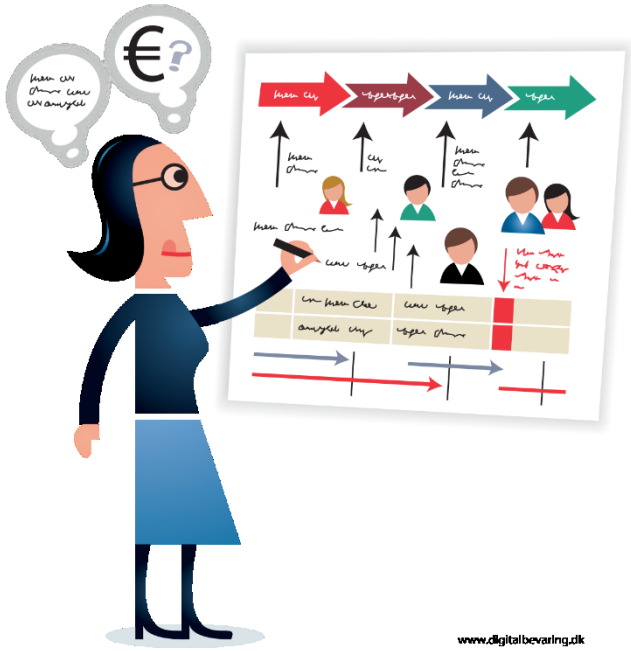
- “Awareness amongst library staff relating to services provided in RDM, also having staff who are reluctant to move beyond their traditional role of information literacy experts to embrace data literacy.”



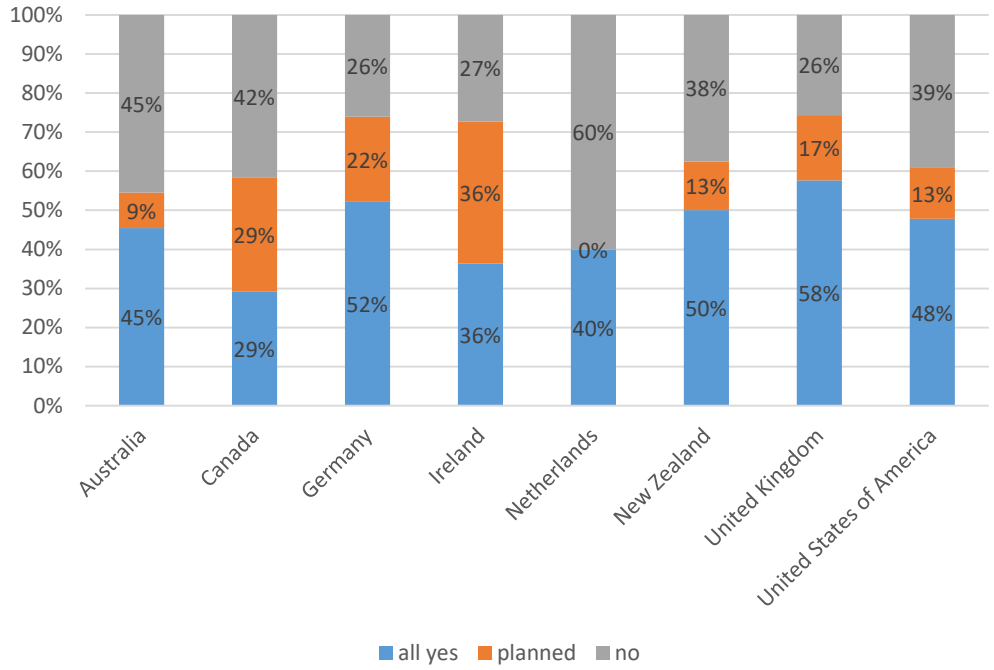
Research Data Policy and Services

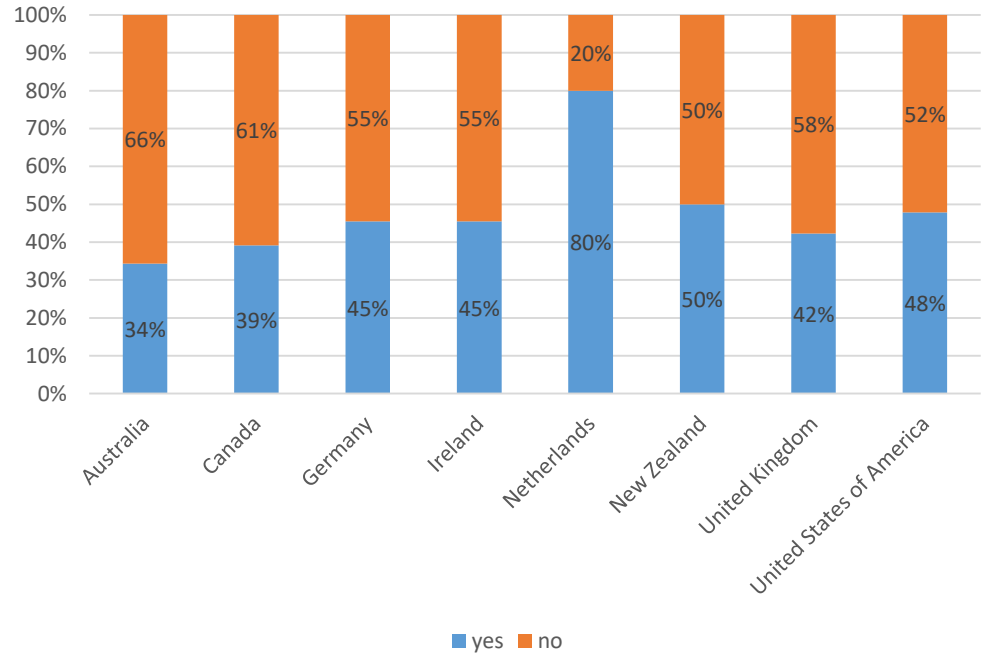
Supporting better RDM

1. User studies
2. Policy
3. Research data services (and infrastructure): from advice to repositories



1. Data audit and/or Staff attitude survey





Why do a survey in your institution?

- Gather vital information about what researchers think and are doing
  - How much data they have; how often they back things up
  - Awareness of policy
  - Attitudes to data sharing; desire for training
- Identify people and groups who are pathfinders or problem areas
- First step in raising awareness and making contacts
- Valuable material for advocacy at a strategic level: research leaders do not know about the practices in their areas
- Benchmark against comparable institutions

What you can ask

- Types of data being used
- Amount of data
- Ownership of data
- Awareness of policy
- Experience of Deposit/ willingness to share data
- Desire for training
- About the person completing the questionnaire

- DAF toolkit 2016
- https://www.researchgate.net/publication/318440738_Jisc_Data_Asset_Framework_Toolkit_2016

- 2014 Survey at Sheffield-
- http://www.ijdc.net/article/view/10.1.210

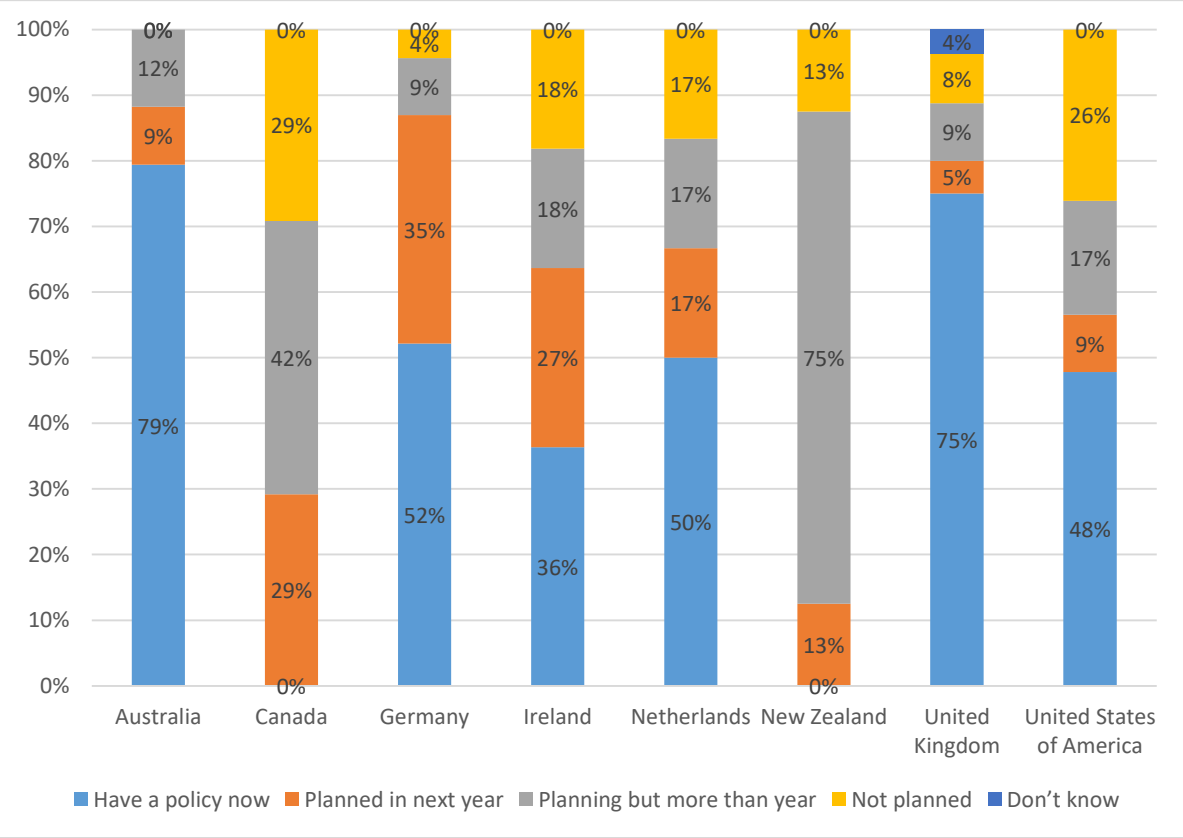
Attitudes to data sharing

	Agree strongly	Agree somewhat	All those agreeing	All those actively disagreeing	Tenopir et al. (2010) – strongly agree
I would use other researchers' datasets if their datasets were easily accessible.	23%	41%	64%	14%	43%
I would be willing to place at least some of my data into a central data repository with no restrictions.	19%	37%	56%	22%	42%
I would be willing to place all of my data into a central data repository with no restrictions.	4%	17%	21%	59%	15%
I would be willing to share data across a broad group of researchers.	19%	43%	62%	13%	37%
It is important that my data is cited when used by other researchers	64%	23%	87%	4%	69%

Training needs

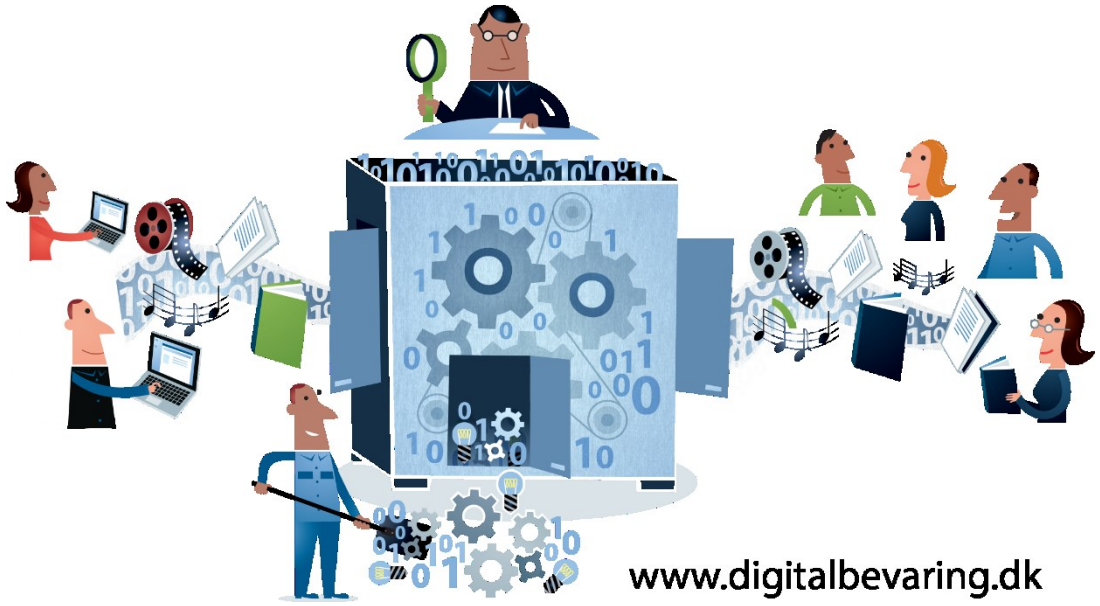
Training subject area	Might be interested	Definitely interested	Total with any interest
Storing your research data	36%	36%	72%
Developing a research data management plan	44%	30%	74%
Copyright and Intellectual Property	40%	30%	70%
Documenting your research	43%	29%	72%
Citing your research data	38%	28%	66%
Sharing your research data	46%	25%	71%
Funders requirements and RDM	46%	21%	68% (after rounding)
Creating metadata for research data	36%	21%	57%
Ethics and consent	35%	19%	54%

2. Institutional policy



- DCC Collection of examples: http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies

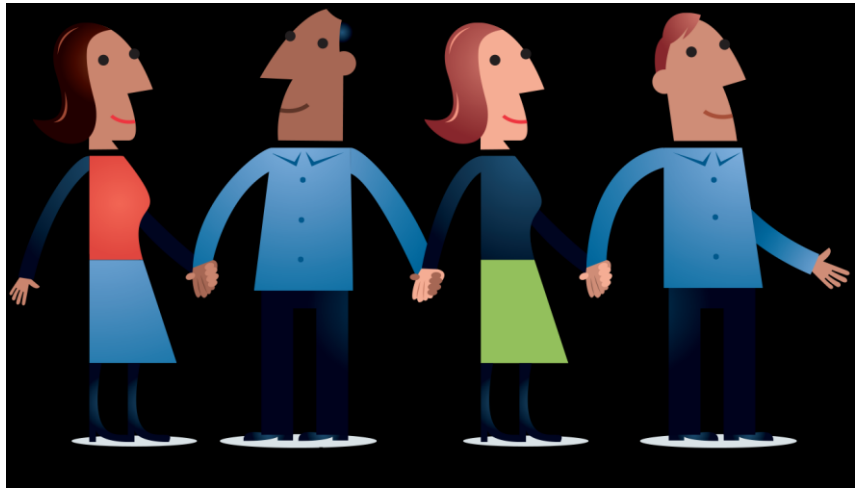
3. Research Data Services



The “who does what game”!

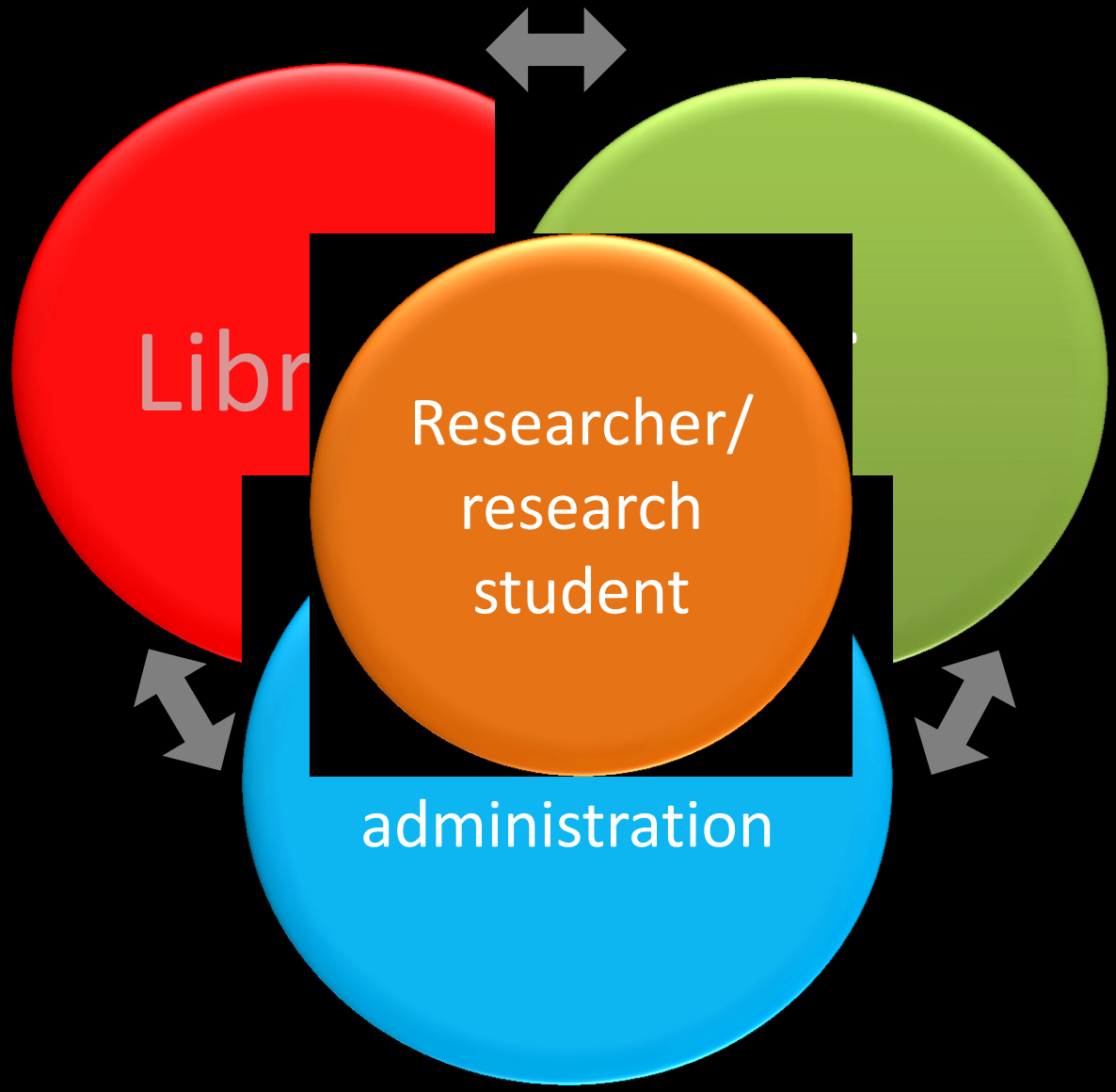
- Library
- Archives/ records managers
- IT services
- Departmental IT staff
- Research administration office, including Ethics / research integrity people (if separate from research administration office)
- Staff development office
- Researchers themselves

- What are the priorities?
- Who should do what?



Who should do what?

	The repository team	The library	Computing services	Research office	Researchers themselves
Survey current practices and attitudes					
Write a data sharing policy					
Advise on security of active data					
Train researchers in RDM					
Create a research data catalogue					
Build a data repository					
Provide overall leadership on RDM					



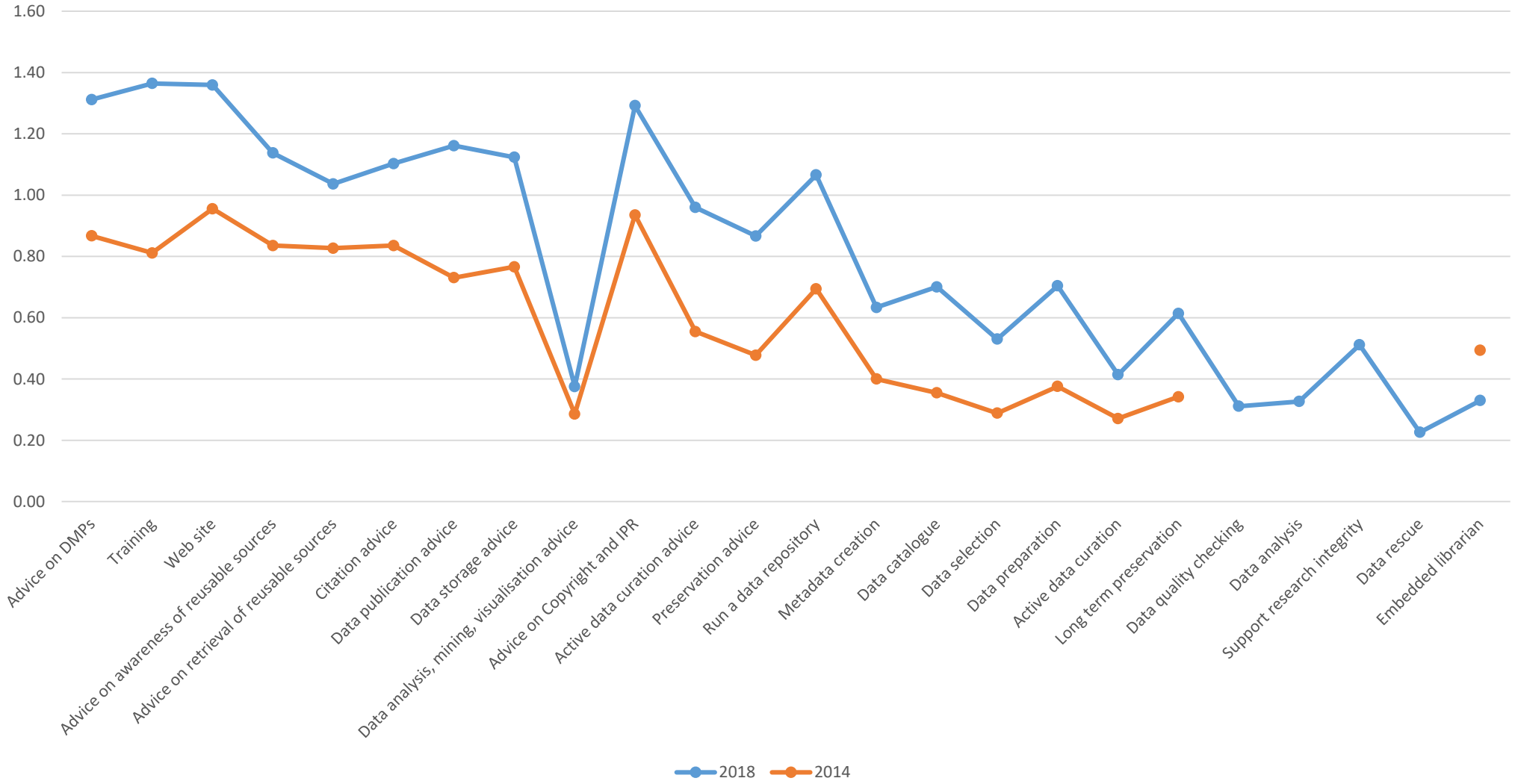
Slide by Martin Lewis,
Director of Library,
University of Sheffield

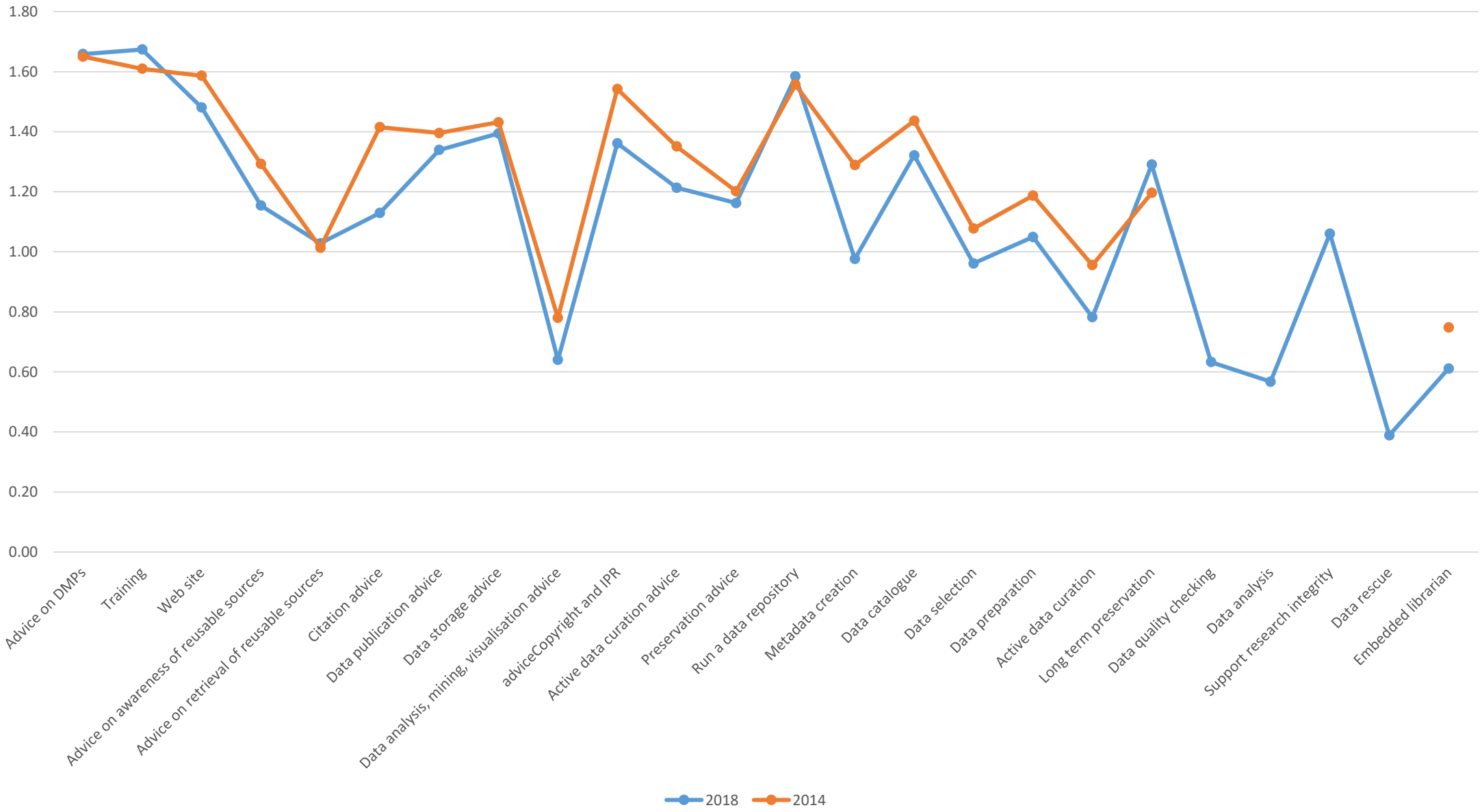
Survey results:
Services Currently Provided by libraries

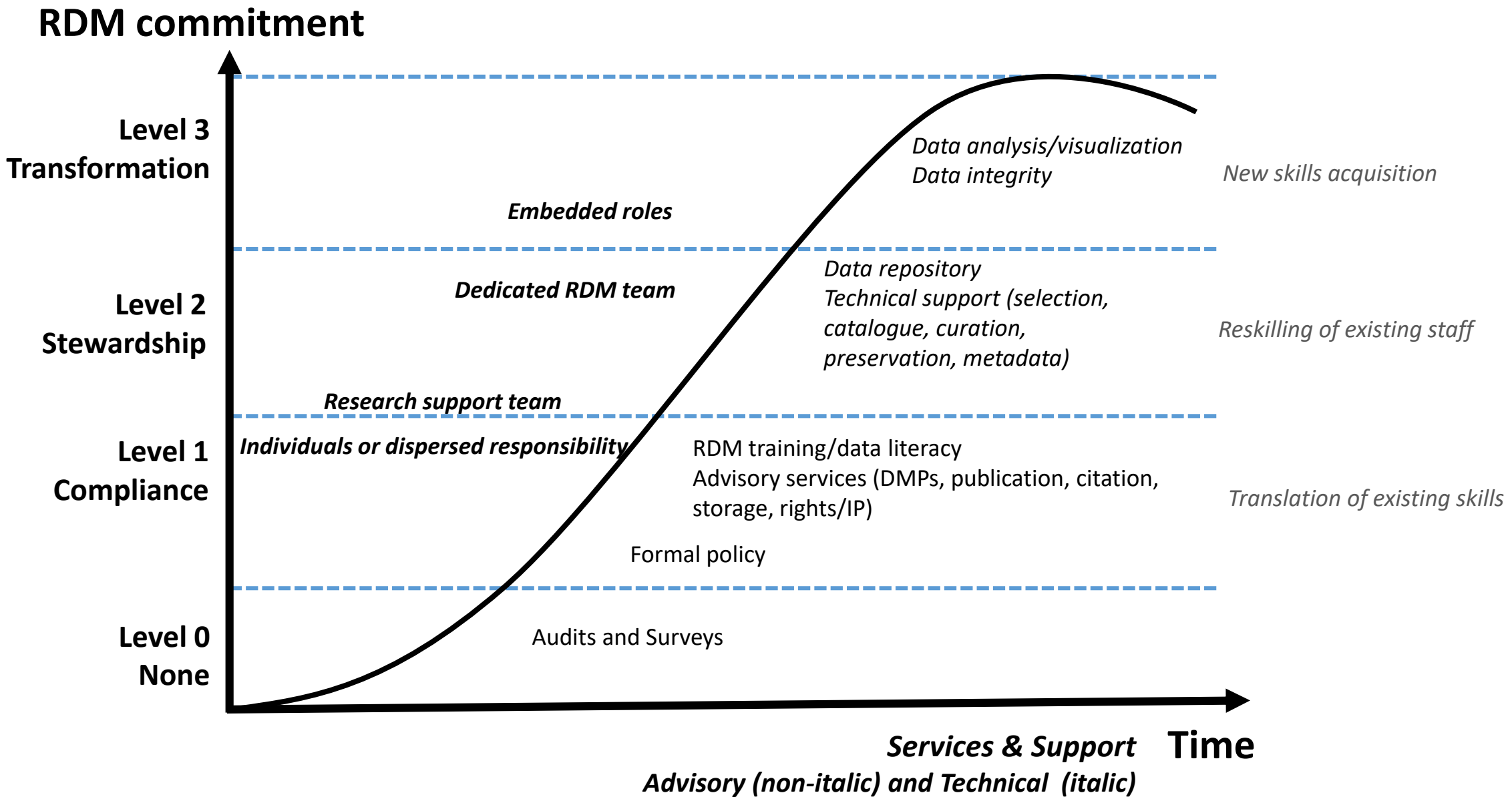
Ranking of services (providing any service – Basic, Well-developed, Extensive): Advisory rather than technical services predominate

1	Promote awareness of reusable data sources, such as data archives	83%
2	Offer advice on copyright and/or intellectual and/or licensing property rights relating to data and data management	81%
2	Data management training and/or data literacy instruction (e.g. to research students, early career researchers etc.)	81%
4	Maintaining a web resource/guide of local advice and useful resources for RDM	79%
5	Data Management Planning (DMP) advisory service	76%
5	Offer data citation advisory services	76%
7	Offer data publication advisory services	75%
8	Provide support for search and retrieval of external data sources	73%
9	Offer data storage advisory services	68%
10	Run a data repository/archive/store	67%
…		
24	Offer an advisory service on data mining	23%
25	Analyse and visualise datasets using Python scripts, SPSS, R and MS Excel software	21%
26	Rescue legacy data or perform data triage or forensic data recovery	16%

Library supported RDS 2014/2018




1.80
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Advice on DMPs
Training
Web site
Advice on awareness of reusable sources
Advice on retrieval of reusable sources
Citation advice
Data publication advice
Data storage advice
Data analysis, mining, visualisation advice
adviceCopyright and IPR
Active data curation advice
Preservation advice
Run a data repository
Metadata creation
Data catalogue
Data selection
Data preparation
Active data curation
Long term preservation
Data quality checking
Data analysis
Support research integrity
Data rescue
Embedded librarian
2018
2014


RDM commitment
Level 3
Transformation
Level 2
Stewardship
Level 1
Compliance
Level 0
None
Embedded roles
Dedicated RDM team
Research support team
Individuals or dispersed responsibility
Data analysis/visualization
Data integrity
New skills acquisition
Data repository
Technical support (selection, catalogue, curation, preservation, metadata)
Reskilling of existing staff
RDM training/data literacy
Advisory services (DMPs, publication, citation, storage, rights/IP)
Translation of existing skills
Formal policy
Audits and Surveys
Services & Support
Time
Advisory (non-italic) and Technical (italic)

V3

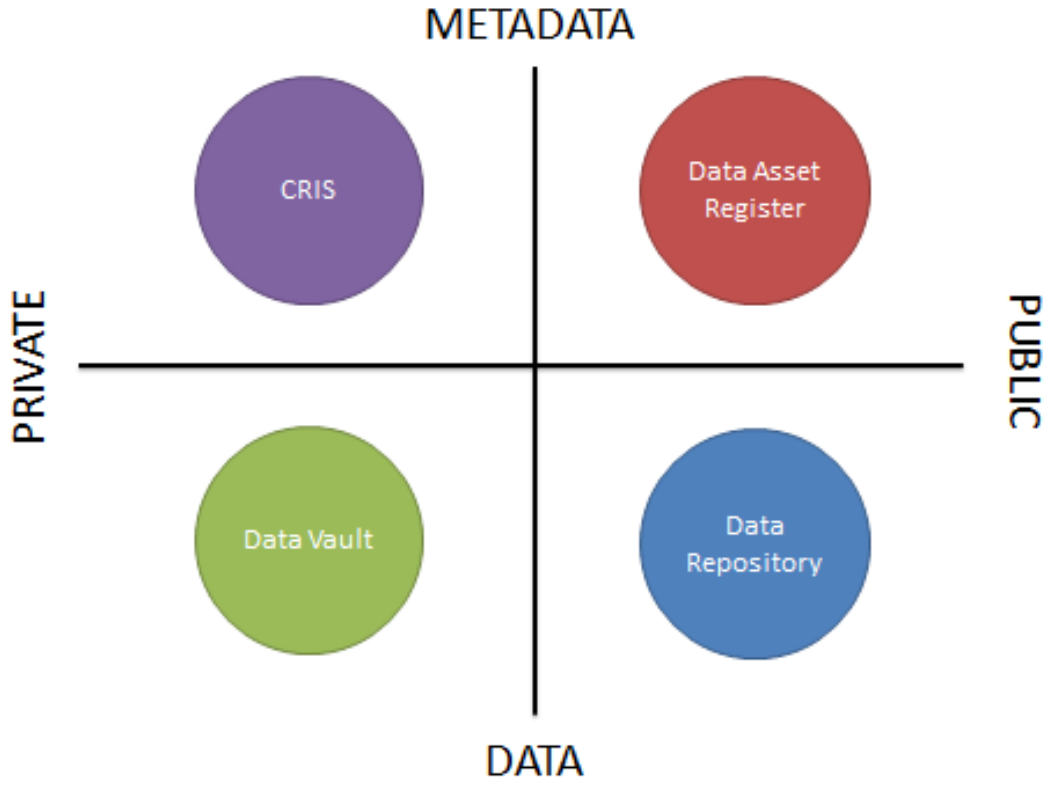
The data role spectrum

Familiar <										> Unfamiliar
Support for data search / access to data	Data literacy training and promoting awareness	Data collection management, including metadata	Gathering support requirements for services/tools	Data policy	Data Management planning advice	Data carpentry	Data curation	Data integrity	Embedded roles in a research team	Data analysis and visualisation

- Close to existing roles
- Resources required
- Demand

Issues for the repository

1. Engaging researchers
2. Developing staff competencies
3. Variety of data types
4. Influencing researchers early enough
5. Choosing technical solutions
6. Choosing external collaborators
7. Meeting competition from publishers



Source: University of Edinburgh http://libraryblogs.is.ed.ac.uk/blog/2013/12/06/the-four-quadrants-of-research-data-curation-systems/

- “The role of publishers, positive and negative, in this arena. They are marketing heavily to university faculty and administration and cutting out libraries from the discussion.”

Advocacy and culture change

What is data sharing? Who would be involved?

- With future self
- With collaborators
- With collaborators beyond the institution
- By request
- Linked to a publication
- Open data in a repository

- The data sharer
- The data repository and/or journal
- The secondary data user
- Support staff
- Research participants
- Research collaborators and external partners (e.g. government, commercial partners)
- Research funders and sponsors

Why share data?

- To enable research to be validated and so ensure integrity
- To enable new research to be done with data
- To increase the visibility of research. Sharing data may increase citation impact.
- To prompt further collaboration

Why share data?

- The public good argument: If public funds were used to create the research, so the results should be available to the project
- Because there is a mandate from funders
- Because there is a mandate from many journals

Some issues for researchers

- Desire to keep control over data after investment of time/ fear of being scooped
- Legal, ethical and commercial reasons for confidentiality
- Dislike of bureaucracy and a lack of time to process datasets
- Lack of know-how, skills and confidence (eg metdata, data selection, choosing a repository, licensing)
- Questions over the usefulness of data to a wider audience
- Issues with the feasibility of data reuse (methodological concerns)
- Lack of a reuse culture
- Fear of criticism and misuse
- Lack of direct incentives
  - “Does your institution have incentives, rewards or recognition for faculty/academic staff in your institution who engage in research data management good practice?”
  - 8/209 said yes

Competencies

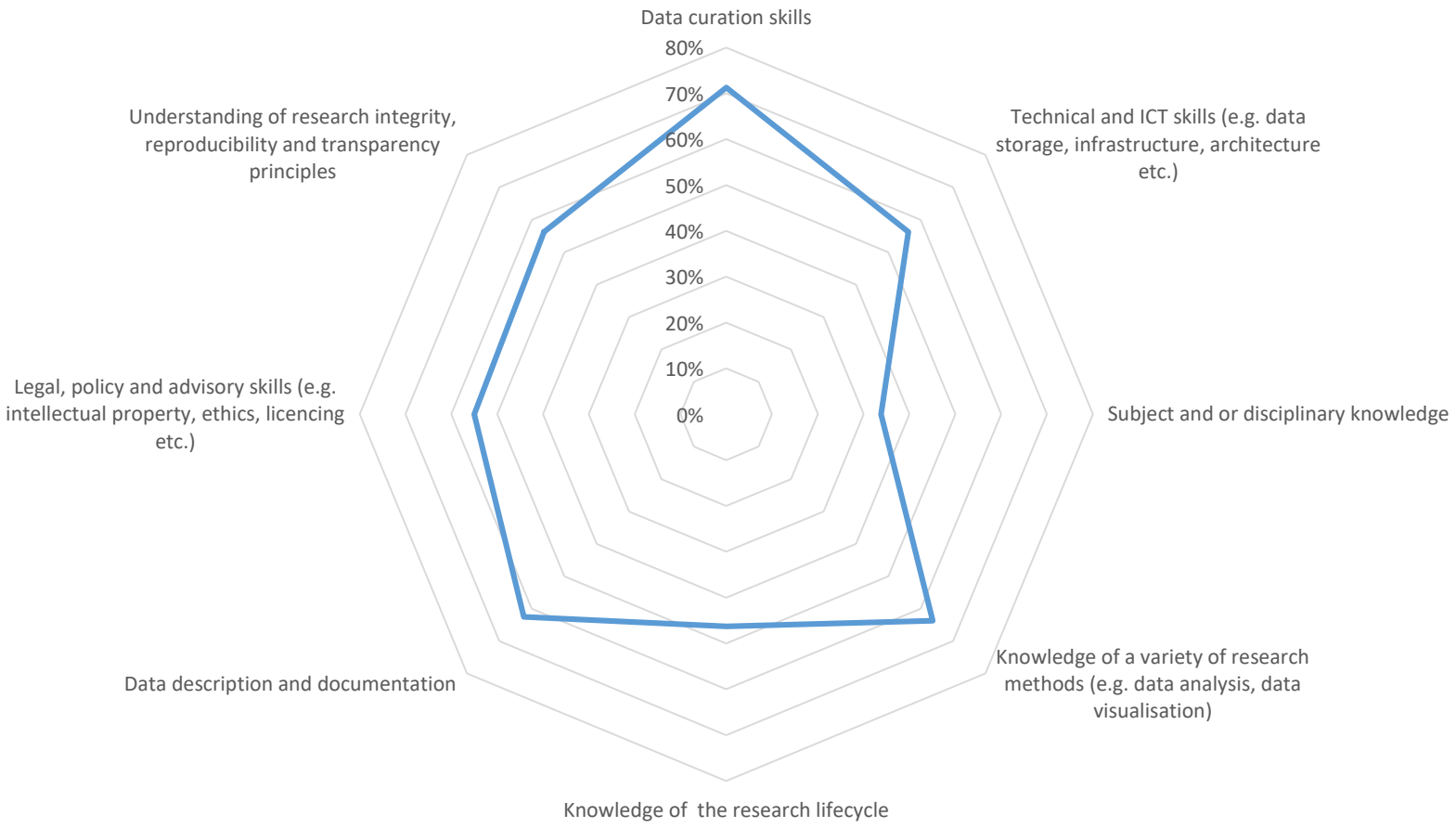
Why repository managers have a potential leadership role

- Their knowledge of and networks within disciplinary communities; their liaison and negotiation skills
- The strong professional network to copy best practice across institutions
- Their contact with many students and researchers in a way other support services do not
- Their generic knowledge of good information management practices
  - Understanding that research data management as a form of Information Literacy
- Their existing open access leadership role
- Relevance of collection development practices; their understanding of metadata

Where is the skills gap?

1. Data curation skills
2. Technical and ICT skills (e.g. data storage, infrastructure, architecture etc.)
3. Subject and or disciplinary knowledge
4. Knowledge of a variety of research methods (e.g. data analysis, data visualisation)
5. Knowledge of the research lifecycle
6. Data description and documentation
7. Legal, policy and advisory skills (e.g. intellectual property, ethics, licencing etc.)
8. Understanding of research integrity, reproducibility and transparency principles
9. Other?

The skills gap (according to UK librarians)



Thank you!

Acknowledgments

- Images are from Digital Preservation, https://digitalbevaring.dk/om-sitet/about-us/
- Co-authors of report on Research Data Services: Stephen Pinfield and Laura Sbaffi (Sheffield), Liz Lyon (Pittsburgh) and Mary Anne Kennan (Charles Sturt)