Data Sharing: A Global Movement

Kevin Ashley
Digital Curation Centre
www.dcc.ac.uk
@kevingashley
Kevin.ashley@ed.ac.uk

Thanks to Sarah Jones, DCC, for FAIR-related content

This work is licensed under the Creative Commons Attribution 2.5 UK: Scotland License.
My home – the DCC

because good research needs good data

capacity for research data services in research institutions

▶ Not just a UK problem – an international one

▶ Training, consultancy, shared services, research, guidance, policy, standards, futures
The source of our experience

- In 2011, DCC began working closely with 20 UK universities to develop research data management (RDM) services
- Putting guidance, learning into practice
- Since expanded to > 60 universities and other organisations around the world
- Participation in European & global projects (EOSCPilot, FOSTER+, OpenAire, FAIR data expert group, Research Data Alliance,...)
DCC ‘institutional engagement’

Assess needs
- Workflow assessment
- Supported assessments (DAF/RISE)
- Advocacy with senior management

Make the case
- Institutional data catalogues
- Pilot RDM tools
- Guidance and training
- RDM policy development
- Customised Data Management Plans

DCC support team

Develop support and services

...and support policy implementation
In 2017, I spoke about

- The case for open, reusable data
- Its value to society
- Its value to research funders
- Its value to individual researchers and their universities
- The skills researchers need
- The role of universities
- The elements of research data management
The Old weather project

Data for research, not from research
Often your data tells stories that your publications do not.

Not all data comes from other researchers.

Discipline-bounded data discovery doesn’t give us all we need or want.

One person’s noise is another person’s signal.
Since then...

▷ FAIR has become a core principle
▷ The idea that data should be:
  » Findable
  » Accessible
  » Interoperable
  » Reusable
▷ ... but not neces

Should all data be open?
▷ NO
▷ Many reasons – most to do with human subjects
▷ But data existence should always be open
▷ Allows discovery & negotiation on use
▷ Avoids pointless replication
ON FAIR
What is FAIR?

A set of principles to ensure that data are shared in a way that enables & enhances reuse, by humans and machines.

[Image: FAIR_data_principles.jpg]
The acronym come from a workshop in Leiden in 2014

Original focus was the life sciences & data – since extended to other domains and other type of research output

Echoes earlier principles relating to research data, such as...
OECD Principles and Guidelines for Access to Research Data from Public Funding (2007)

A. Openness
B. Flexibility
C. Transparency
D. Legal conformity
E. Protection of IP
F. Formal responsibility
G. Professionalism
H. Interoperability
I. Quality
J. Security
K. Efficiency
L. Accountability
M. Sustainability

Science as an Open Enterprise (2012) notion of ‘intelligent openness’ where data are accessible, intelligible, assessable and useable

“Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.”

G8 Science Ministers Statement (2013)
Is FAIR straightforward

▷ A popular notion – we all want to be ‘fair’ in how we work

▷ But.. a 2018 survey of European researchers shows:
  » 10% say they don’t understand the concept
  » 16% do not know if they understand

▷ Concepts are simple, terminology is hard

▷ ‘Interoperable’ presents particular difficulties for many

https://doi.org/10.5281/zenodo.1120245
Respondents mentioned 40 terms which were unclear to them.

“Researchers are not familiar with the following terms/phrases: Metadata, standards for metadata/data, ontologies, mapping with ontologies, interoperability, ... All the ICT jargon”

“With the help from Swedish National Data Service we could clarify many questions. Without this help we would not be able to finish the DMP.”

15 FAIR principles

**FINDABLE**
- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

**INTEROPERABLE**
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

**ACCESSIBLE**
- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

**REUSABLE**
- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.
Being FAIR

- FAIR is not a specific technology or standard to be conformed to.
- Its practice lies in a continuum – and different research communities must decide what FAIR means for them.
- Supporting it will require investment by research organisations.
FAIR ACTION
European investment – the FAIR data expert group remit

1. To develop recommendations on what needs to be done to turn each component of the FAIR data principles into reality
2. To propose indicators to measure progress on each of the FAIR components
3. To provide input to the proposed European Open Science Cloud (EOSC) action plan on how to make data FAIR
4. To contribute to the evaluation of the Horizon 2020 Data Management Plan (DMP) template and development of associated sector / discipline-specific guidance
5. To provide input on the issue of costing and financing data management activities
FAIR data expert group members

Simon Hodson, CODATA Chair of FAIR Data EG
Rūta Petrauskaitė, Vytautas Magnus University
Peter Wittenburg, Max Planck Computing & Data Facility
Sarah Jones, Digital Curation Centre (DCC), Rapporteur

Daniel Mietchen, Data Science Institute, University of Virginia
Françoise Genova, Observatoire Astronomique de Strasbourg
Leif Laaksonen, CSC-IT Centre for Science
Natalie Harrower, Digital Repository of Ireland – year 2 only
Sandra Collins, National Library of Ireland – year 1 only
FAIR data expert group report

- Launched November 2018
- 27 clear recommendations for stakeholders, structured around:

  1. Concepts for FAIR
  2. Creating a FAIR culture
  3. Creating a technical ecosystem for FAIR
  4. Skills and capacity building
  5. Incentives and metrics
  6. Investment and sustainability

Turning FAIR into Reality: Report and Action Plan [https://doi.org/10.2777/1524](https://doi.org/10.2777/1524)
FAIR data objects

- Digital objects can include data, software, and other research resources
- Universal use of PIDs
- Use of common formats
- Data accompanied by code
- Rich metadata
- Clear licensing
Project convened by the American Geophysical Union

- Develop standards to connect researchers, publishers and repositories
- Builds on the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS) Statement

- 50+ organisational signatories so far:
- http://www.copdess.org/enabling-fair-data-project
Skills – a critical component

▶ Two cohorts of professionals to support FAIR data:
  » - data scientists embedded in research projects
  » - data stewards who will ensure the curation of FAIR data

▶ Initiatives to coordinate, systematise and accelerate the pedagogy

▶ Support formal and informal learning

▶ Ensure researchers have foundational data skills
Concepts of FAIR and Open should not be conflated. Data can be FAIR or Open, both or neither.

The greatest potential reuse comes when data are both FAIR and Open.

Align and harmonise FAIR and Open data policy.
STUDIES ON INFRASTRUCTURE AND ATTITUDES
Dutch study on domain use cases

- Six detailed use cases from engineering, social science, climate science, physics, health care
- Highlights differences in culture and practice and need for domain specific guidelines
- Tension between domain approaches and interoperability cross-domain

https://doi.org/10.5281/zenodo.1246815
CLIMATE SCIENCE

▷ FAIR used implicitly. Sharing data is ingrained as climate is international.
▷ Established global standards and exchange platform of data centres.
▷ Transition from GRIB to netCDF to exchange more with other communities.

DUTCH REPOSITORIES

▷ Repository compliance is not high—
  ▷ 38% not compliant in terms of Findability
  ▷ 52% not compliant in terms of Interoperability
  ▷ 46% not compliant in terms of Reusability
▷ 49% of repositories did not assign a DOI, or URN
▷ None of the repositories had visible ontologies or controlled vocabularies
▷ Social science & climate science repositories fared worse

Based on study of 37 Dutch repositories:

https://doi.org/10.2218/ijdc.v12i2.567
Almost all provide open access to metadata, but majority (70%) do not provide open access to data.

60% of Nordic repositories did not assign a persistent identifier.

56% do not employ metadata standards.

80% not certified.

Based on study of 61 Nordic repositories in The State of Open Science in the Nordic Countries NEIC report.
.... And on researcher behaviour
How many researchers make data open?

79% of researchers have made data openly available

The State of Open Data 2017
Digital Science
2300 respondents worldwide

only 1 in 10 provides their research data as open data for the public

Researchers and their data (2015)
eInfrastructures Austria
3026 Austrian respondents

64% agree that they are willing to share their data

Open Data: the researcher perspective (2017), Elsevier
1162 respondents worldwide

68% of researchers already share data or expect to do so in future

Jisc DAF studies (2016)
1185 UK respondents
How do researchers share data?

Over half only allow access on request.
54% share data by using external storage devices or email.

“When asked where they have published data, most commonly respondents had done so as an appendix to an article (just over 30%) with a data repository close behind (just under 30%) and 20% having published in a data journal.”

Of 13 methods stated, top 4 options for currently sharing data were:

1. Emailing data files (65%)
2. Cloud service e.g. Dropbox, Googledrive (59%)
3. Portable storage (35%)
4. Supplementary data (20%)

Formal repository (public / institutional) c.12%

Less than 15% publish data in a repository.

Jisc DAF studies

Elsevier: Open Data - the researcher perspective

Slide: Sarah Jones, CC-BY
Why do researchers share data?

“For more than half of the researchers, the most attractive incentives for sharing their data were increased visibility and impact, new cooperation opportunities, recognition in professional circles, as well as their contributions being regarded as scientific output.”

-eInfra Austria: researchers and their data
Integrity – not without data

Cyril Burt
- Twin studies on intelligence.
- Questioned 1976; now discredited.
- Neubauer twin study – (‘Three identical strangers’ – no publications, data sealed, purpose unknown).

Dutch cases - psychology
- Stapel – 55 publications – “fictitious data”
- Poldermans – fabricated data or negligence?

Clinical trials
- Alltrials.net initiative – get them registered
- Weekly shaming by Goldacre et al in BMJ- trials that have finished without publication

“Lies, Damned Lies and Research Data: Can Data Sharing Prevent Data Fraud?” – Doorn, Dillo, van Horik, IJDC 8(1); doi:10.2218/ijdc.v8i1.256
What about the cost?

- A rough guide – 5% of total project cost on data curation
- May not all fall to original research group
- How to pay depends on funder and university costing models
- Benefits to society & industry are proven
  » E.g. PWC study in Europe showing cost over €10bn of NOT following FAIR data principles
- Automation and simplification of many processes is helping
My messages to researchers

- Sharing is difficult
- Reusing is difficult
- Both are key to advancing science, and advancing your own career
- Your data can live longer than your findings
- All this can be easier than you think