EC Expert Group on FAIR Data:
Some thoughts on costs and benefits and business models

Simon Hodson, Executive Director, CODATA
www.codata.org
Costs of doing FAIR...
Costs of not doing FAIR...
Some thoughts...

- If we agree that FAIR is essential for reproducibility / replicability then FAIR is part of the cost of doing science...
- The costs of good data management and stewardship can be quantified from case studies. What is hard to quantify are the hidden costs of current bad practices (the lost data, lost time, invisible labour and costs of ad hoc systems).
- There is considerable evidence of the economic benefits of Open Data from certain research areas.
- The counter factual is important: what is the cost of losing data, of no longer having access to data, of data not being interoperable and reusable?
- Important to consider not just how much we pay, but how we pay (importance of business models).
- FAIR depends on a system, not just repositories, and the costs of this (identifiers, creation and maintenance of metadata standards etc) needs to be considered.
Emerging Policy Consensus? FAIR Data

- **FAIR Data** (see original guiding principles at [https://www.force11.org/node/6062](https://www.force11.org/node/6062))
  - **Findable**: have sufficiently rich metadata and a unique and persistent identifier.
  - **Accessible**: retrievable by humans and machines through a standard protocol; open and free by default; authentication and authorization where necessary.
  - **Interoperable**: metadata use a ‘formal, accessible, shared, and broadly applicable language for knowledge representation’.
  - **Reusable**: metadata provide rich and accurate information; clear usage license; detailed provenance.
European Commission Expert Group on FAIR Data

Core Deliverables

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. Actively **support the creation of the FAIR Data Action Plan**, by proposing a list of concrete actions as part of its Final Report.

Related Deliverables

1. To contribute to the **evaluation of the Horizon 2020 Data Management Plan (DMP) template** and development of associated sector / discipline-specific guidance
2. To provide input on the issue of **costing and financing data management activities**
European Commission Expert Group on FAIR Data

Report Structure
1. Concepts: Why FAIR?
2. Making FAIR data a reality: technical perspective
3. Creating a culture of FAIR data
4. Skills and capacities for FAIR data
5. Measuring Change
6. Facilitating Change: a FAIR Data Action Plan
Attributes that give value to research data

• Builds on previous definitions...
• OECD Statement of Principles and Guidelines for Access Research Data: include a number of principles including accessibility, interoperability, quality, legal transparency, sustainability...
• G8 Science Ministers’ Statement, 2013, ‘Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.’
• FAIR Data now at the heart of H2020 policy, European Open Science Cloud etc.
  • Under the revised version of the 2017 work programme, the Open Research Data pilot has been extended to cover all the thematic areas of Horizon 2020.
FAIR Data – Additional Concepts

- As Open as Possible, as Closed as Necessary
- FREE-FAIRER: Findable, Rapidly Available, Ethical, Equitable; Forever, Accessible, Interoperable, Reliable.
- FAIRer: adding Reproducible / Replicable...
- FAIREST: adding Stewardship and Trust...
- Timely release.
- Long term preservation in a trusted (sustainable) digital repository.
- Quality markers from community and repository.
- Responsibilities of users.
- ...

- Action Plan retain FAIR, retain the set of priorities it lays out, while discussing how FAIR can be achieved in a broader ecosystem.
Open Science and FAIR Data: Implications and Directions

- Clarify the **boundaries of Open** for research data.
- Clarify and propagate **criteria and guidelines** for appraisal and selection.
- Recognise that data stewardship must be regarded as part of the total cost of doing research.
  - Invest in **sustainable data infrastructure** (including repositories, stewardship, standards), and develop appropriate **business models** for sustainability.
- **Incorporate research data in the process of scholarly communication** and ensure that researchers, research groups and institutions receive adequate reward and recognition for their efforts.
- Address the **skills** requirements for data scientists, data stewards, data liaisons and researchers themselves.
- **Refine and improve understanding of FAIR data, particularly I and R**...
- Work with and across disciplines on standards and vocabularies, to help address I and R...
Interoperability and Reusability

- The two most challenging areas of FAIR...
- Major challenges to clarify and unpack **interoperable** and **reusable**.
- Need to build on knowledge from archival community: what makes a resource usable?
- **OAIS Reference Model:**
  - ‘**Preservation Description Information** is divided into five types of preserving information called Provenance, Context, Reference, Fixity and Access Rights’
  - **Independently Understandable:** A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.
  - **Designated Community:** An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.
The Value of Open Data Sharing

- Report by CODATA for GEO, the Group on Earth Observation.
- Provides a concise, accessible, high level synthesis of key arguments and evidence of the benefits and value of open data sharing.
- Particular, but not exclusive, reference to Earth Observation data.
- Benefits in the areas of:
  - Economic Benefits
  - Social Welfare Benefits
  - Research and Innovation Opportunities
  - Education
  - Governance
- Available at [http://dx.doi.org/10.5281/zenodo.33830](http://dx.doi.org/10.5281/zenodo.33830)
Economic Benefits of Data Sharing: LandSat

- **2006 Study** estimated the loss in case of a data gap as equivalent to US$935 M.
- **2011 Study** estimated benefits of landsat-sourced information for agriculture as US$858 M just for the state of Iowa.
- **2015 Study** estimated worldwide economic benefit of US$2.19 BN.
- Estimated benefit in US of US$1.8 BN.
- Valuing Geospatial Information: Using the Contingent Valuation Method to Estimate the Economic Benefits of Landsat Satellite Imagery: [http://dx.doi.org/10.14358/PERS.81.8.647](http://dx.doi.org/10.14358/PERS.81.8.647) (Paywall... Irony...)
- Open data and open data infrastructure has a significant economic benefit.
The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres
(Neil Beagrie and John Houghton) http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

Figure 3: The value and impacts of the three UK data centres

<table>
<thead>
<tr>
<th>Investment &amp; Use Value (Direct)</th>
<th>Contingent Value (Stated)</th>
<th>Efficiency Impact (Estimated)</th>
<th>Return on Investment (Estimated)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ESDS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on Depositor &amp; User Counts</td>
<td>Based on User Count</td>
<td>Based on User Count</td>
<td>Based on Data Spend</td>
</tr>
<tr>
<td>Investment Value £23m per annum</td>
<td>Willingness to Pay £25m per annum</td>
<td>Willingness to Accept £83m per annum (£111m per annum)</td>
<td>Increased Rol £58m - £233m</td>
</tr>
<tr>
<td>Use Value £24m per annum</td>
<td></td>
<td></td>
<td>2.5 to 10-fold Rol</td>
</tr>
<tr>
<td><strong>ADS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on Deposit &amp; Download Counts</td>
<td>Based on Download Count</td>
<td>Based on User Count</td>
<td>Based on Data Spend</td>
</tr>
<tr>
<td>Investment Value £1.2m per annum</td>
<td>Willingness to Pay £1.1m per annum (Constrained by capacity to pay)</td>
<td>User Community £13m per annum</td>
<td>Increased Rol £2.4m - £9.7m</td>
</tr>
<tr>
<td>Use Value £1.4m per annum</td>
<td>Willingness to Accept £7.4m per annum</td>
<td>User Community £58m per annum</td>
<td>2.1 to 8.3-fold Rol</td>
</tr>
<tr>
<td><strong>BADC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investment Value £2.8m per annum</td>
<td>Willingness to Pay £5.2m per annum</td>
<td>User Community £10m per annum</td>
<td>Increased Rol £11m - £34m</td>
</tr>
<tr>
<td>Use Value £2.3m per annum</td>
<td>Willingness to Accept £16m per annum</td>
<td>User Community £58m per annum</td>
<td>4 to 12-fold Rol</td>
</tr>
</tbody>
</table>
80% of ecology data irretrievable after 20 years (516 studies)

Vines TH et al. (2013) Current Biology DOI: 10.1016/j.cub.2013.11.014
Benefits of Open Data: some examples from GEO

- Barbara Ryan, Director of Secretariat GEO, TED-X Talk Barcelona
- In 2008 US Government was convinced to make Landsat Data openly available, for free.
- Under charging, the highest number of downloads was 53 scenes per day.
- Now over 5700 scenes per day are downloaded.
- Spanish deforestation research: under the charging regime data access alone would have cost €260M

https://www.youtube.com/watch?v=9umWTFgFIVs
The Challenge: Business Models for Sustainable Data Repositories

- Research increasingly relies on digital repositories, curated databases and services over data.
- Research funder policies increasingly mandate data stewardship of data produced by funded projects.
- Increasing need for data repositories and data stewardship.
  - Increasing volume presents a challenge.
  - Requirements for stewardship present a greater challenge.
- **Sustaining digital data infrastructure is a major issue for science policy!**
- Genuine concern that current funding models will prove inelastic and not meet the growing requirements – concern on the part of repositories and funders.
  1. Essential that data stewardship and FAIR data is considered an essential part of the cost of doing research.
  2. Essential to demonstrate the value proposition of data repositories / data services.
  3. Sustainability is not just about whether something is funded, but how it is funded: what are the most effective and sustainable mechanisms for funding?
Final Report: Business Models for Sustainable Data Repositories

- Report summarizing findings and analysis, key features: http://dx.doi.org/10.1787/302b12bb-en
  - Improved understanding of mechanisms for funding research data repositories.
  - Criteria to assess the sustainability of business models for research data repositories in different circumstances.
  - Summary of the pros and cons of different business models.
  - Insights into the development of business models for research data repositories.
  - Recommendations that primarily target policymakers and funders in OECD member states, as well as repository operators and managers.
Final Report: Business Models for Sustainable Data Repositories

- **Recommendations:**
  
  [http://dx.doi.org/10.1787/302b12bb-en](http://dx.doi.org/10.1787/302b12bb-en)

  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.

  3. Policy makers, research funders, and other stakeholders need to consider the ways in which data repositories are funded, and the advantages and disadvantages of various business models 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 in order to be able to effectively manage digital assets over time.
Thank you for your attention!

Slide Credits: Geoffrey Boulton, Jane Hunter, John Broome, Ernie Boyko, Devika Madalli, LI Jianhui

Simon Hodson
Executive Director CODATA

www.codata.org
http://lists.codata.org/mailman/listinfo/codata-international_lists.codata.org

Email: simon@codata.org
Twitter: @simonhodson99
Tel (Office): +33 1 45 25 04 96 | Tel (Cell): +33 6 86 30 42 59

CODATA (ICSU Committee on Data for Science and Technology), 5 rue Auguste Vacquerie, 75016 Paris, FRANCE