BRIEF

Using open and FAIR data to increase research efficiency

April 2019

SPARC Europe

Using open and FAIR data to increase research efficiency

Briefing paper for policy makers, research funders and senior managers

1. Introduction

There are many reasons to pursue an agenda of open research, and more specifically open data in research. While other briefing papers in this series have considered the value of increased research visibility and related citations; our briefing examines [ST1] the implications for research integrity. A strong moral and policy argument exists for openness of products which are the result of public funding. Moral arguments aside, there is also clear evidence of economic efficiencies associated with data being made open as well as FAIR. In this briefing paper we present the findings of various models, which are intended to assist research decision makers, such as funders and senior management, when making evidence-based decisions regarding data management, especially open and FAIR data. While it is likely that they have a clear picture of the costs involved, their understanding of how open and FAIR data contributes to greater research efficiency may be more ambiguous. Further to presenting a summary of key findings, we also give examples of useful indicators for measuring efficiency gains, and the reports we reference present both qualitative and quantitative approaches for measuring the value of openness and FAIRness of research data.

2. Efficiency savings associated with data

One possible way to view this is informed by the calculations in the business case made to UK government in 2010 for significant investment in research data infrastructure and associated activities as an outcome from the UK Research Data Service project (UKRDS.) The final business case is not public, unfortunately, but earlier versions appear in the initial feasibility report.[1] The figures presented in this briefing differ in some respects from those presented at the time. In particular, we disregard efficiencies gained by centralising some aspects of the service, but the broad argument is the same. The UK business case indicates a clear benefit to the overall research budget from data reuse, even if much data remains unused. Governments might realise these benefits in one of two ways: they may reduce research budgets whilst still aiming to achieve the same outcomes, or they may choose to keep budgets constant knowing that they will be able to fund more research overall from the savings realised.

We must make a number of assumptions to construct our model. The figures presented here were based in part on figures from UK research funders at the time (which are unlikely to have changed in character) and information on actual costs of data reuse in research. Those relating to storage of data have changed markedly in the intervening years and have been updated to reflect current costs; the change improves the business case.

First, let us assume that the average research project costs €2m and that 10% of this cost relates to original data collection. Thus, an equivalent project that is able to reuse data already collected could conceivably run 10% below the budget of the former - which amounts to a savings of €200,000. But there is a penalty associated with using other people's data. It takes time to locate, and even more time to understand and ensure its suitable for the intended analysis.[2] The search costs are relatively modest – a few days of staff time; we will set them at €2,500. The reuse costs are somewhat higher, and best expressed as a fraction of the cost of original data collection. 10% is a realistic figure and so we end up with a total figure of €22,500 to reuse this dataset as opposed to €200,000 to generate it in the first place. Our total saving for each reuse is thus almost 9% of the total cost of the research project ((200 000-22 500)/2 000 000). But not every dataset will be reused, and there are costs associated with retaining the data and making it available for others to use. These costs need to be set against the savings. In addition, there are initial costs in setting up suitable storage and discovery services, whilst data reuse – and the benefits that arise from it – may take some years to realise.

The UKRDS feasibility study[1] reported an expected return of £5.84m after 5 years for an investment of £18.58m, or a return of 31%. Returns in later years would be higher since the first 5 years included an initial period of significant investment with no return. This figure also assumed savings which arose from centralisation of some services.

Let us instead consider our simplified model of the €2m research project, and assume that we fund 1,000 datasets per year and wish to retain this data for at least 5 years. Thus, at the end of the five-year-period, we must make available 5,000 datasets for reuse at a national average cost of €1,000/year/dataset, or an annual cost of €5m against our total research spend of €2,000m. If only 29 of our datasets are re-used each year, the savings generated are €5,147,500 – a few percent more than the cost of retaining all the datasets for reuse. Thus our open data reuse policy is financially effective even if less than 0.6% of the data retained is reused each year.

Expressed more formally, if our repositories hold K datasets, and the cost of doing so for each dataset is S euros/year, then the total cost of maintaining our discoverable, reusable data is K x S per year. Our savings from each reuse event were shown above to be €177,500. Thus, if R% of our datasets are reused each year, the total savings are (R x K x 177 500)/100 or R x K x 1775. If this number is greater than K x S then we are seeing efficiencies. This can be further simplified since K is a common factor. For retention and reuse to be worthwhile, R x 1775 must be greater than S.

[1] UKRDS (2008) *The UK research data service feasibility study: Report and Recommendations to HEFCE*.
https://web.archive.org/web/20120314155910/http://www.ukrds.ac.uk/resources/download/id/16

3. Value of data centres and data management

A synthesis study demonstrating the financial benefits of open research data (when stored centrally in research data centres), through the review of three studies of the value of three UK research data centres, was conducted by Beagrie and Houghton for Jisc in 2014[3]. The data centres in question were the Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS) and the British Atmospheric Data Centre (BADC), thus representing a range of different disciplines. The data services differ in maturity, operational budget, type and number of users. The notion of value was evaluated both qualitatively, i.e. in interviews researchers were asked what data repositories meant for their work; and quantitatively, e.g., measuring return on investment. In short, Beagrie and Houghton (2014) find that

"economic analysis indicates that data sharing and data curation via the centres studies has a substantial and measurable positive return on investment and, by facilitating additional use, increases the return on investment in the original/collection of the data hosted." (p.16)

The qualitative analysis revealed that many academic users value the centres very highly, many stating that if they were not able to access data and services through them, it would have a major or a severe impact on their work. Estimated efficiency gains, reported by users of data centres, ranged from 2 to more than 20 times the costs (operational, depositor and user costs)[2]. The authors state that although the three cases are different, and not comparable, they do illustrate a similar pattern across findings: data sharing via the centres has a significant and measurable impact on research efficiency.

A more recent impact evaluation of the European Bioinformatics Institutes, carried out by Beagrie and Houghton in 2016 reveals that the benefits of EMBL-EBI data and services can be estimated at a minimum of £1bn per year worldwide, which is 20 times the direct operational cost.[5] Over half of researchers surveyed in the evaluation claimed that not having access to EMBL-EBI resources would have a "major" or "severe" impact on their work[6]. Calculating the efficiency impact of the services and resources were estimated at as much as £26,000 on average per person per year and as little as £5,380, taking into account potential mis-reporting or misunderstanding of survey questions.[3] This is based on a calculation of researcher time spent on working with data, and how having access to well managed and curated data can reduce time spent on data related activities.

The cost of poor data management is also reported by the [U.S. Geological Survey website](#) to be approximately 15% - 25% of a project's budget. The website furthermore quotes Bill Michener, The Director of DataONE, who refers to the "80-20" rule, where "Eighty percent of a scientist's effort is spent discovering, acquiring,

[2] 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. (page 17)* Available at: http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

[3] Beagrie, C. and Houghton, J. (2016) *The Value and Impact of the European Bioinformatics Institute: Full Report*. Report for EMB-EBI. (p.26) https://beagrie.com/static/resource/EBI-impact-report.pdf

documenting, transforming and integrating data, whereas only 20 percent of the effort is devoted to (…) analysis, visualization, and making new discoveries."[4][7]
In addition to highlighting the value of openness, this also speaks to the value of good data management and the curation of data according to the FAIR principles (see section 4). Research data can be open, but at the same be difficult to find and non reusable due to a lack of metadata or incomplete metadata. As we present in the next section, the highest cost associated with research is time spent on searching for data as well as working with datasets that come with incomplete metadata. Putting this time to "better use", i.e. on analysis, visualisation and dissemination would allow for a more efficient R&I process overall.

4. The cost of non-FAIR data

As we have explored in a previous briefing paper, FAIR (Findable, Accessible, Interoperable, Reusable) data is an important concept when data curation is considered along with openness. That briefing paper also provides guidance on how to implement FAIR and open data, drawing from a report of the European Commission Expert Group on how to turn FAIR data into reality[5].
In March 2018, the European Commission published a report presenting findings from a cost-benefit analysis for FAIR research data[6]; the report reveals the cost of not making data FAIR. It uses seven indicators to calculate costs of not making data FAIR: time spent; cost of storage; licence costs; research retraction; double funding; interdisciplinarity and potential economic growth. Of these, 'time spent' and 'storage' are the most significant cost drivers (see table 1); noteworthy is the fact that the movement towards FAIR data will have significant impact on the way we use and store data in research. The report is rich in findings and presents in-depth analysis of the impact of non-FAIR data on research activities, collaborations and innovation. A summary and breakdown of cost savings is presented in in Table 1 below.

[4]USGS, *Value of Data Management* (no date). Available at: https://www.usgs.gov/products/data-and-tools/data-management/value-data-management#efficiency
[5] European Commission (2018) Turning FAIR Into Reality. Final report and action plan from the European Commission Expert Group on FAIR data. https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283
[6] European Commission (2018) Cost of not having FAIR research data: Cost-benefit analysis for FAIR data. https://publications.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en

Indicator	Explanation	Cost per year
'Time Spent'	Time spent on searching for data, often with incomplete metadata.	€4.5bn
'Cost of Storage'	Additional copies are made of inaccessible data. These would not be necessary if data was made FAIR.	€5.3bn
'Licence costs'	Cost of extra licences that researchers have to pay to access non FAIR data.	€360m
'Research retraction'	Non FAIR research would lead to less article retractions due to non-reproducibility, errors, fraud, plagiarism etc.	€4.4m
'Double funding'	Non-FAIR research leads to duplication of research effort.	€25m
'Interdisciplinarity'	Added value of interdisciplinary research made possible by FAIR data.	Cost impact could not be estimated reliably. Interdisciplinarity through the FAIR principles will increase rate of data re-use and allow for escaping disciplinary and data silos to perform better science.
'Potential economic growth'	GDP growth and number of jobs created if FAIR data was widely available.	Cost impact could not be estimated reliably. However report authors estimate that economic benefits of FAIR data to be parallel to open data, which is estimated[8] between €11.7bn and €22.1bn per year[7].

Table 1 Summary of costs associated with not making data FAIR in Europe[8]

[7] European Commission, (2017)*European Data Market: Final Report.* http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=44400

[8] The table presents a summary from European Commission (2018) Cost of not having FAIR research data: Cost-benefit analysis for FAIR data. https://publications.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en

The report estimates the annual cost of not making data FAIR in Europe is a minimum of €10.2bn per year. This does not take into account the impact of FAIR on innovation, due to a lack of data which hinders this link being explored fully. The authors of this EC report acknowledge that the cost of having data which is not FAIR will differ between disciplines, depending on their data-intensity and that FAIR readiness may differ between institutions. They however argue that an overall move towards making data FAIR will greatly contribute to a more efficient and thus a stronger science ecosystem in Europe.

5. Conclusions

This paper has presented calculations and evidence that support the argument for good research data management and a move towards open and FAIR data. We acknowledge that the exact cost figures will differ between disciplines and that calculations will need to be done on a case by case basis to account for different types of data as well as infrastructure investment needed. We have referred here to a number of reports which demonstrate, in detail, methodologies and approaches to estimating the impact and value of data management, data infrastructure and openness, as well as FAIR data, which will allow decision makers, at different levels, to calculate the value and impact of making their data open and FAIR. We recognise that in some instances there may be a barrier to realising such sizable gains as are referenced throughout as these only make sense at nation-state level or above; the investments made by individual organisations may or may not result in equivalent gains by those organisations but will result in a net gain for the science community as a whole as time spent on data activities can be reduced, making more room for discovery and dissemination. Making data available for re-use requires that institutions have robust data management processes and structures, which may be costly, specifically during the initiation phase. The benefits of having high quality open and FAIR data available for researchers are however clear. These benefits will result in a more efficient science system where greater focus is on discovery and production of knowledge rather than on searching for data, working with incomplete data or duplicating research effort.

April 2019

SPARC Europe would like to thank DCC for carrying out this work

DCC

www.dcc.ac.uk

Disclaimer

Please note that the views expressed in this paper are that of the author Kevin Ashley; SPARC Europe is not responsible for any resulting works derived from the information it contains