Practice Safe Science: Five reasons to protect your scientific data

blogs.lse.ac.uk/impactofsocialsciences/2014/01/21/five-reasons-to-protect-your-scientific-data/

1/21/2014

Research data management is quickly becoming one of the most pressing issues facing the scientific community, not just for university management teams, but for every individual researcher. The tech company Digital Science produced an infographic that captures five reasons why more attention is needed to attain a more secure system. **Nathan Westgarth** *elaborates on the points presented and on how the research process can be made more efficient through the better use of technology.*

We've been hearing a common theme from the scientific community – researchers are having difficulty managing and accessing their data. It seems to be an on-going problem for research scientists, at any stage of their careers. So we decided to do some investigation work and look at the stats. What we found was a rather concerning picture of the effect of poor scientific data management.

The amount of research data being generated is currently increasing by 30% every year. Worryingly, one study has found that a massive 80% of scientific data is then lost within two decades and the odds of sourcing datasets decline by 17% each year (Vines T.H. et al. 2013). As data output grows, effective data organisation is only going to get more difficult. If data continues to be managed poorly then science will ultimately suffer; experiments will be hard to replicate, findings called into question, papers retracted and careers will be impacted.

At Digital Science we engage with users of our products, and with researchers in the community in general through our UserLab program. Scientific researchers globally participate in surveys and focused interviews to discuss problems and concerns they have throughout their research workflows. One of the most common themes reported back to us is typical of Mark Hahnel's comment that "During my PhD I was never good at managing my research data. I had so many different file names for my data that I always struggled to find the correct file easily when it was requested." In a more extreme case of data management issues, Biologist Billy Hinchen told us "I lost 400GB of data and close to 4 years of work after my laptop was stolen from my lab. As a result I ended up getting an M.Phil rather than a PhD".

Our infographic tells the story of the impact of poor scientific data management and highlights 5 reasons to protect scientific data:

1. Data output is growing rapidly – 90% of all the data in the world has been generated over the last 2 years and scientific data output is currently increasing at an annual rate of 30%.

2. Despite significant investment, data is not being managed effectively – the current estimated total global spend on research and development is $1.5 trillion, which could be at risk. Much of the data generated is lost – in one study, the odds of sourcing datasets declined by 17% each year, with 80% of

Love your data – practise safe science

Data output is growing rapidly

90% of all the data in the world has been generated over the last 2 years[1]

2008 2009 2010 2011 2012 2013

30% JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

Scientific data output increases at an annual rate of 30%[2]

Despite significant investment, data is not being managed effectively

$1.5 TRILLION

80% lost

datasets over 20 years old not available.

3. Much of the data remains unverifiable – 54% of the resources used across 238 published studies could not be identified, making verification impossible.

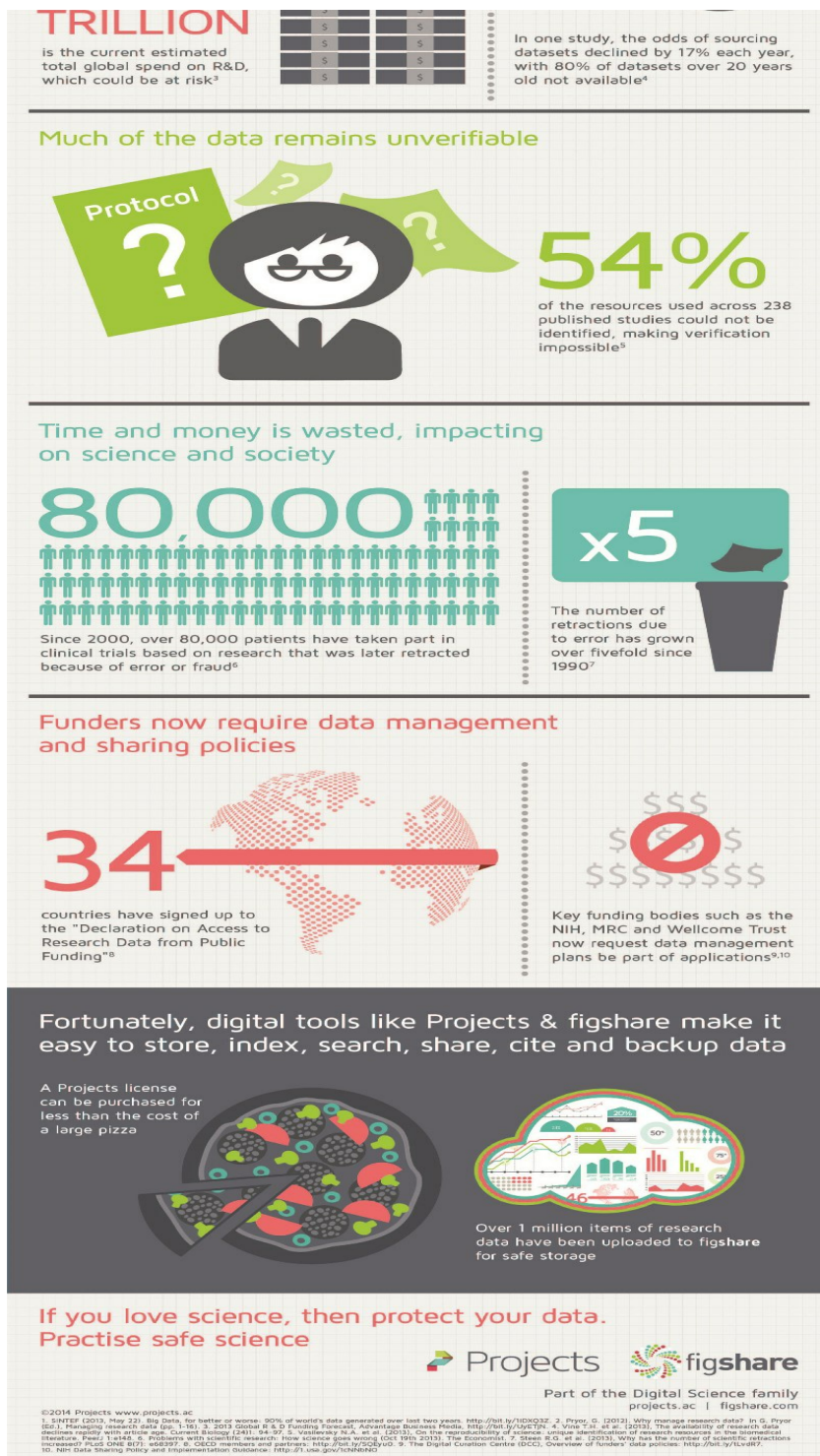**4. Time and money is wasted, impacting on science and society –** Since 2000, over 80,000 patients have taken part in clinical trials based on research that was later retracted because of error or fraud. The number of retractions due to error has grown over fivefold since 1990.

5. Funders now require data management and sharing policies – 34 countries have signed up to the "Declaration on Access to Research Data from Public Funding". Key funding bodies such as the NIH, MRC and Wellcome Trust now request data management plans be part of applications.

We think it's time to start practising safe science and protect your data! But how? Digital Science is part of a growing effort among start-ups to resolve this problem and was founded to make research more efficient through the better use of technology. We've started to develop some tools of our own to help researchers with their data. We've developed Projects, a simple desktop app that lets researchers safely manage their research data by providing a visual timeline to make finding files easy and backup functionality to recover previous versions of files.

Projects integrates with one of our other tools figshare, a cloud based repository where researchers can store their data, share it with colleagues, or make it publicly available and citable, with a permanent DOI. Here's a taster of the Projects interface:

TRILLION is the current estimated total global spend on R&D, which could be at risk[3]

In one study, the odds of sourcing datasets declined by 17% each year, with 80% of datasets over 20 years old not available[4]

Much of the data remains unverifiable

Protocol?

54% of the resources used across 238 published studies could not be identified, making verification impossible[5]

Time and money is wasted, impacting on science and society

80,000

x5

Since 2000, over 80,000 patients have taken part in clinical trials based on research that was later retracted because of error or fraud[6]

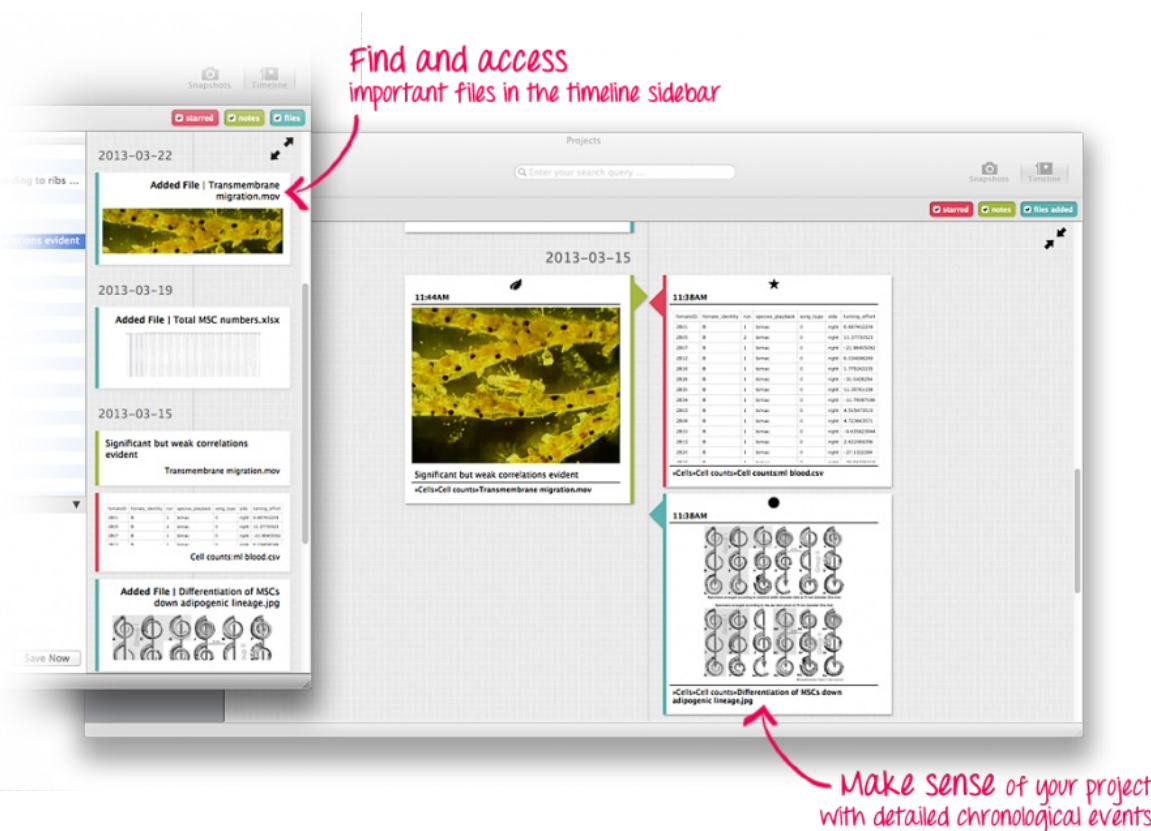The number of retractions due to error has grown over fivefold since 1990[7]

Funders now require data management and sharing policies

34 countries have signed up to the "Declaration on Access to Research Data from Public Funding"[8]

Key funding bodies such as the NIH, MRC and Wellcome Trust now request data management plans be part of applications[9,10]

Fortunately, digital tools like Projects & figshare make it easy to store, index, search, share, cite and backup data

A Projects license can be purchased for less than the cost of a large pizza

Over 1 million items of research data have been uploaded to figshare for safe storage

If you love science, then protect your data. Practise safe science

Projects figshare

Part of the Digital Science family
projects.ac | figshare.com

©2014 Projects www.projects.ac
1. SINTEF (2013, May 22). Big Data, for better or worse. 90% of world's data generated over last two years. http://bit.ly/1IDXQ3Z. 2. Pryor, G. (2012). Why manage research data? In G. Pryor (Ed.), Managing research data (pp. 1–16). 3. 2013 Global R & D Funding Forecast, Advantage Business Media. http://bit.ly/UyETJN. 4. Vine T.H. et al. (2013), The availability of research data declines rapidly with article age. Current Biology (24)1: 94–97. 5. Vasilevsky N.A. et al. (2013), On the reproducibility of science: unique identification of research resources in the biomedical literature. Peez J 1:e148. 6. Problems with scientific research: how science goes wrong (Oct 19th 2013). The Economist. 7. Steen R.G. et al. (2013), Why has the number of scientific retractions increased? PLoS ONE 8(7): e68397. 8. OECD members and partners: http://bit.ly/SQEyu0. 9. The Digital Curation Centre (DCC), Overview of funders' data policies: http://bit.ly/1LrdR7. 10. NIH Data Sharing Policy and Implementation Guidance: http://1.usa.gov/1cNNbND

colleagues, or make it publicly available and citable, with a permanent DOI. Here's a taster of the Projects interface:

options for tools that also help to resolve data management issues are to make use of the various electronic laboratory notebooks available to help researchers collect notes and metadata about their research and protocols. Alternatively researchers can try to make the host of generic tools fit into existing research workflows. Some such tools proving popular are Evernote, cloud storage services like Google Drive and Dropbox, and code hosting sites like GitHub. Mark Hahnel's post on the Scientific American blog, Research Management for Dummies, goes into more detail on the range of tools available.

What do you think about the issues of data availability and how do you manage your own data? Projects would like to hear your thoughts on Twitter @projects.

The 'Love Your Data' infographic was produced by Projects, a Digital Science product.

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Nathan Westgarth *joined Digital Science (Twitter) – a technology company that serves the needs of scientific research, as a Product Manager for Research Tools. Nathan manages the Projects (Twitter) product at Digital Science – a new research data management tool to help scientific researchers organise their data in a safe, simple and structured way.*