It's time to put our impact data to work to get a better understanding of the value, use and re-use of research.

If published articles and research data are subject to open access and sharing mandates, why not also the data on impact-related activity of research outputs? **Liz Allen** *argues that the curation of an open 'impact genome project' could go a long way in remedying our limited understanding of impact. Of course there would be lots of variants in the type of impact 'sequenced', but the analysis of 'big data' on impact, could facilitate the development of meaningful indicators of the value, use and re-use of research.*

We know that research impact takes many forms, has many dimensions and is not static, as knowledge evolves and the opportunities to do something with that knowledge expand. Over the last decade, research institutions and funding agencies have got good at capturing, counting and describing the outputs emerging from research. A lot of time and money has been invested by funding agencies to implement grant reporting platforms to capture the myriad outputs and products of research (e.g. ResearchFish, Pure, Converis and ImpactTracker). The inclusion of the impact element in the UK's 2014 Research Excellence Framework (REF) alone, yielded nearly 7000 case studies of impact. Happy days indeed for those required to describe the impact of research, undertake funding programme evaluation and, create stories for research communication and engagement purposes.

Capturing and describing impact takes a lot of time and resource, not to mention the opportunity cost to science while researchers and research institutions enter information into platforms, and craft stories; drawing together the impact case studies for the UK's REF, is estimated to have cost around £55m (almost as much as the entire 2008 Research Assessment Exercise (RAE), even accounting for inflation). Yet, despite our efforts and considerable skill in *describing* impact, we have limited ability to *understand* whether and how we could make research more effective, productive and impactful. However, there is an opportunity to start to remedy this.
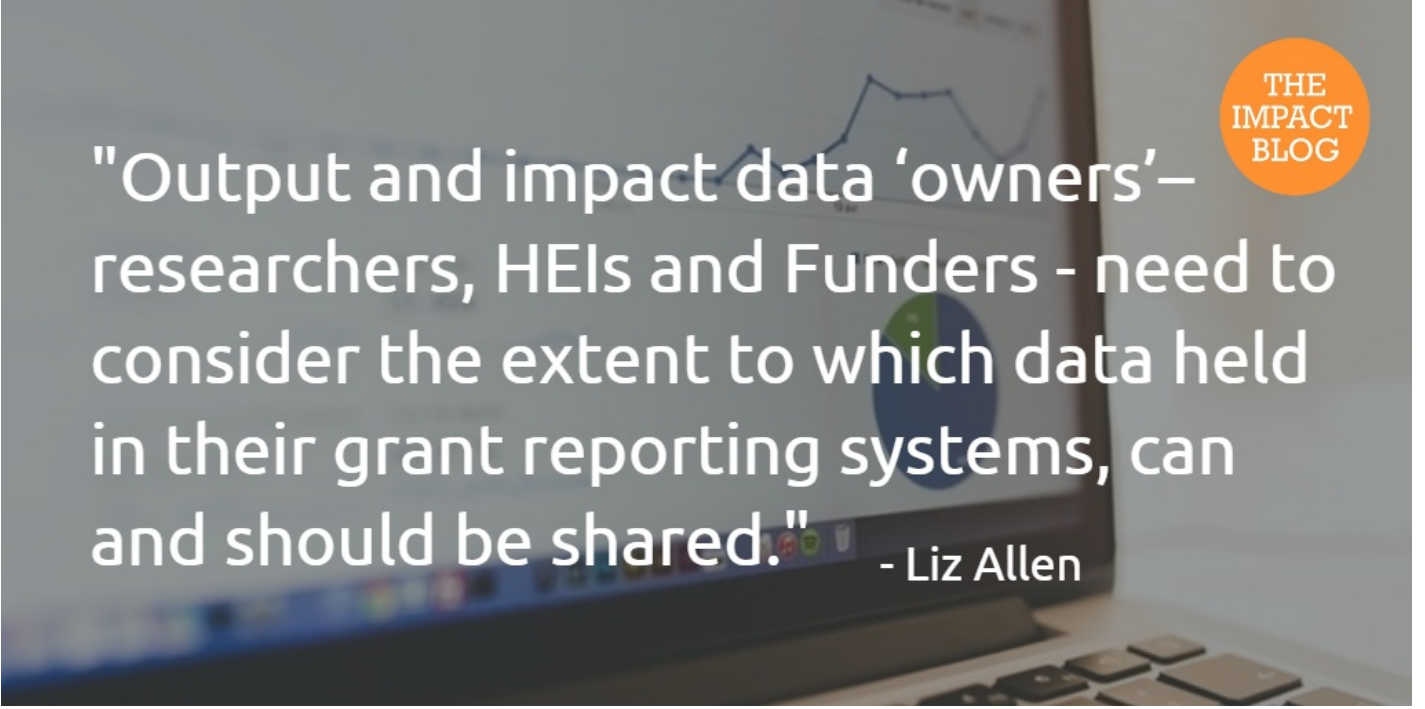
Image credit: Shaury Nash Alineando secuencias (CC BY-SA)

If it were possible to draw together our mass of impact-related data – currently stored across a range of platforms – policy and evaluation professionals, could be enticingly close to being able to go beyond *describing* impact, through perhaps (and I am not the first person to use this metaphor in this context) an impact genome project. Of course, it is unlikely that we will ever be able to precisely predict the outcomes of research, for a given funding investment or input; research moves mostly in incremental steps, serendipity has been a characteristic of some of the most important scientific breakthroughs, and it is tricky to define the precise contributions of individuals or teams associated with specific impacts. So, there would be lots of variants in the type of impact 'sequenced'; nevertheless, through such a large body of data we should at least spend time trying to identify patterns across research areas, and, for example, determine whether variation in how research is funded, executed, managed, delivered and shared, effects the products of research, and its impact.

The timing is good. Governments and funding agencies across the world want to make research assessment more efficient, quicker and cheaper. In the UK, there is much discussion about whether future national research assessment exercises could and should be more reliant upon research-related metrics; both BIS and the Stern Review want to make evidence-based decisions. It therefore seems obvious that, where potential evidence exists, it should be used. At the same time, we should be using technology to connect data – through infrastructure like ORCID – and reduce researcher burdens associated with reporting output-related information where possible. The Metric Tide included a call for connectivity across research information platforms, partly to bring efficiencies to research management and administration, but also to support the need for more research on research. Furthermore, through the drive towards open science, the basis on which the findings of research is made available, shared and used by others, is likely to change the way we perceive value impact and quality.

To allow proper scrutiny of the output and impact-related data now amassing, and serve a research on research imperative, a number of things need to happen. Key among these, output and impact data 'owners'– researchers, HEIs and Funders – need to consider the extent to which data held in their grant reporting systems, can and should be shared. If published articles and research data are subject to open access and sharing mandates, why can't other outputs and impacts related data (subject to considerations around any commercial or sensitive data), be also? Importantly for the future, if data are shared, appropriate consents and permissions need to be accommodated in reporting platforms.

"Output and impact data 'owners'— researchers, HEIs and Funders - need to consider the extent to which data held in their grant reporting systems, can and should be shared." - Liz Allen

THE IMPACT BLOG

A second, and not insubstantial challenge, concerns data compatibility. To enable robust analysis of data across systems, classification, taxonomies and descriptors of impact need to be agreed, curated, and managed. There are already relatively well developed taxonomies being used to describe impact contained within grant reporting tools which could be shared, agreed among stakeholders and evolved in collaboration over time. To permit analyses, data needs to be in formats that can be connected, aggregated and stored – perhaps through central repository or through system APIs (application programme interface); ensuring an ORCID id for researchers and assigning a Digital Object Identifier (DOI) to research objects would simplify connectivity.

A further consideration, is the inclusion of information on all aspects of the funding life cycle. Often grant reporting tools focus on the *post-award* period, and do not include data on aspects of the funding decision. Research evaluation has tended to focus on research outputs and impacts where there are some systematic data (e.g. scholarly output and numbers of doctoral awards) and largely ignore the inputs and broader context. Information on the latter, such as researcher demographics, career stage, or type of grant awarded, is currently less readily available for analysis alongside outputs. Perhaps most importantly, aspects of the grant selection process – and there could be lots of variables here – are also missing from much analysis. A recent study of National Institute of Health grants found that (at least in terms of publication outputs) peer review, beyond screening out poor applications early in the funding process, was not a good predictor for which grants would be most 'successful'; without more research on research, we don't really know what the implications of this could be for how we support science more broadly.

Finally, assuming access to research output and impact data, there needs to be funding for skilled researchers and data scientists to formulate the research questions and do the work. Given the potential benefits *to* research of research on *how* to *do* research more effectively, this seems like an obvious thing for funding agencies to get behind and support.

We can't have a sensible debate about the potential of metrics to support national research assessment exercises, unless we know more about which metrics are most appropriate to answer the questions being posed. Analysis of 'big data' on impact, could facilitate the development of meaningful indicators of the value, use and re-use of research. Indicators that are used to support research funding allocation and have such an effect on peoples' lives and careers, should be evidenced-based, and based on what matters rather than what is merely available. And, as the Leiden Manifesto recently emphasised, drawing on a combination of quantitative (e.g. Relative Citation Ratio; Normalized Citation Impact) and qualitative indicators (e.g. F1000Prime; Altmetric) in research evaluation is likely to

deliver balanced and robust conclusions.

It might be that impact (or aspects of) is not reducible to big data analysis and correlation. But if scrutiny of these data does not support *any* insights into research efficiency, productivity and impact, the research community would be justified in asking two questions: (1) are we collecting the right data? and perhaps more fundamentally, (2) if the best use of these data is to *describe* impact, is the effort spent in bringing it together worth it?

Note: This article gives the views of the author, and not the position of the LSE Impact 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

Liz Allen, PhD is a Visiting Senior Research Fellow at the Policy Institute, KCL & Director of Strategic Initiatives at F1000