# The future of research outputs: Summary of roundtable discussion co-hosted by the British Library and RAND Europe

## Introduction and provocation

On the 17th February 2020, British Library and RAND Europe convened a roundtable discussion on the future of research outputs. Since the workshop we have all been challenged by the profound changes to our lives and research caused by COVID-19, a crisis which re-focused a worldwide research effort in search of a vaccine and effective drugs, and which, among other things, brought into sharp relief the importance of rapid sharing of research data, but also importance of communicating research to the widest possible public. This makes this discussion even more relevant when considering the future of research outputs.

The aim of the workshop was to explore the issues and implications for different actors in the sector e.g. researchers, universities, libraries and funders presenting recent research and facilitating a discussion with the group.

RAND Europe presented research commissioned by Research England to survey researchers in universities across England and understand their views of what the future may hold.[1] Some of the key findings of the report include:

- Researchers currently produce a diversity of output forms, and expect this to increase
- Researchers expect to continue to produce journal articles and conference contributions
- Many researchers expect to start to produce more diverse forms of output aimed at a wider audience
- Many researchers want to diversify their output types
- Reasons researcher expect to diversity outputs vary
- There are differences between disciplines, and these are expected to persist.

From this, we understand that journal articles and books will remain important forms of outputs and these are well understood and well served by existing systems for cataloguing, archiving and assessment. There are also existing contributions that researchers frequently make which are not well captured or served, e.g. peer reviews, conference contributions, datasets. In addition, our research shows that researchers want to and expect to diversify their outputs. Key growth areas are likely to include public facing outputs such as social media, blogs, and websites. We know that research assessment processes, institutional expectations and system change are drivers of changing outputs. These actors also need to be ready to respond to and acknowledge the increasing diversity in academic outputs.

The British Library presented their current work and experience in collecting, preserving and making accessible research data, web and social media, as well as a range of new and evolving research output formats. The Library's presentation pointed out the following:

- Success and growth of interest in PhD theses as research output with the Library's EThOS platform bringing together 544,000 PhD records, of which 330,000 in total are fully available for download from either Ethos or HE repository (60% of all UK records). 60,000 theses are viewed via EThOS every month. www.ethos.bl.uk

[1] Parks, Sarah, Daniela Rodriguez-Rincon, Sarah Parkinson, and Catriona Manville, The changing research landscape and reflections on national research assessment in the future. Research England, 2019. https://www.rand.org/pubs/research_reports/RR3200.html.

- UK Web service, which ingests and preserves the UK web content, including web research outputs, growing at a rate of approximately 80 terabytes per annum. www.webarchive.org.uk
- Selective archiving of social media, including a reflection on technological and legal difficulties in trying to make this more useful for researchers.
- Open Data – how we continue to address the challenges arising in enabling better discovery in the complex data landscape, including the initiatives such as DataCite working with over 100 of UK research organisations, assigning DOIs to datasets enabling researchers to locate, identify and cite research datasets with confidence. www.bl.uk/datacite
- A fast growing and expansive range of emerging formats used by scientists such as collaborative and sharing platforms and tools enabling geolocation, visualisation and crowdsourcing.

## Key themes identified in discussion

The discussion bringing together representatives from research funders, publishers, research institutes, government and universities. The group discussed these findings and the implications. The following sections set out some of the key themes that emerged and the questions that resulted from the discussion.

### *How do we define and identify a research output?*

There are many types of outputs from research – spanning from the traditional outputs such as journal articles and books, to more diverse types of output such as code, artworks, blogs, datasets and peer review contributions. Figure 1 shows the top 10 output types most frequently produced now, and those expected in the future. The group discussed that formats are changing and expanding. One of the challenges is to identify which products from research are outputs, and which represent stages in the development of research on the pathway to producing those outputs. This reflects the fact that different materials produced from research may cover different aspects of the same information, some may not be finalised or complete, and many not be fixed in time. One example discussed here is a Github repository which may be fluid and changing on an ongoing basis – but in this case at least that history and emergence can be traced over time. Other products – for example social media exchanges – are a fixed point but may not represent a researcher's final perspective on a topic, rather the emergence and discussion of views and ideas. This fluid and dynamic mix of different media emerging over time makes it challenging to understand what is a research output as might be traditionally defined but can be important for historians and researchers to understand the creative process that researchers have undertaken. The group also discussed some of the advantages of this dynamic and fluid type of research output, reflecting on the difficulty of retractions in the current system of research publishing, and the fact that people reading research publications may not be aware of a subsequent retraction of that work. The importance of delineating process from output varies depending on the reasons for capturing research. Reproducibility purposes may benefit from access to the process information. However, access and archiving purposes may be better served by a more streamlined and well-defined set of outputs.

*Figure 1 Top 10 output types reported by researchers, now and in the future. Source: Parks et al, 2019*[2]

| Ranking | Most frequently reported output forms now | Most frequently reported output forms in 5 to 10 years' time |
|---|---|---|
| 1 | Journal article | Journal article |
| 2 | Conference contribution | Conference contribution |
| 3 | Chapter in book | Chapter in book |
| 4 | Research datasets and databases | Authored book |
| 5 | Working paper | Research datasets and databases |
| 6 | Social media content and blogs | Website content |
| 7 | Website content | Openly published peer review |
| 8 | Openly published peer review | Social media content and blogs |
| 9 | Authored book | Research report for external body (non-confidential) |
| 10 | Code | Edited book |

*What is the purpose – discoverability, preservation, reproducibility, reuse, assessment – what is the most important problem to solve now?*

It is important to consider how the outputs may be used in the future and what purposes capturing research outputs serves. We identify five key purposes in our discussions:

- Discoverability: To enable knowledge to be found easily
- Preservation: To preserve a record of the research endeavour for future generations
- Reproducibility and reuse: To enable researchers to use and build on existing research, and ensure it can be reproduced by others
- Assessment: To assess the performance of research systems and researchers, and learn how to improve that performance

These purposes speak to different approaches and priorities. For example, a full record of the Twitter conversations of prominent scientists may have value for preservation purposes and may – to some extent – be useful for the purposes of reproducibility, in understanding the process through which approaches came about. However, this type of information is likely to be of limited utility in terms of discoverability and assessment.

*Where does responsibility sit and for what?*

We discussed the fact that research is increasingly global and that research outputs may span national borders – hence there is limitation on the extent to which UK actors can, should, or need to act to provide a full record of all research endeavour – but also that drawing lines between what is and what is not 'UK research' is not straightforward. The role of different levels of actors within the UK was also discussed -for example, what is the role of a national library versus institutional library facilities, and where should different records be held. We also reflected on the role of the UK-wide electronic legal deposit and its important role for long-term preservation, but also its limitations in terms of remote access. The role of different actors in being either proactive or reactive was also discussed. Different actors – libraries, funders, institutions, publishers – can either look to shape and drive desirable changes in behaviour or respond to changes as they emerge 'bottom up'. Funders in particular

[2] Parks, Sarah, Daniela Rodriguez-Rincon, Sarah Parkinson, and Catriona Manville, The changing research landscape and reflections on national research assessment in the future. Research England, 2019. https://www.rand.org/pubs/research_reports/RR3200.html.

are aware that their rewards and incentives can drive researcher actions so there needs to be consideration of what behaviours are desirable and whether the current system supports or hinders those in terms of those incentive structures.

*How do we manage quality control?*

As the range and nature of output types broadens, there are questions emerging around how we can assess the quality of those outputs and make decisions about what is part of the scientific record. One point acknowledged was the weaknesses and limitations of our current approaches to this, in terms of peer review. It was also noted that the ultimate test of the quality and rigour of research was the extent of uptake and use by the academic community over time, and in that sense the change in types of outputs makes little difference to our ultimate assessment of their quality. However, it was also noted that as the volume of research increases, and with increasing concerns around issues of reproducibility, fake news and the reliability of different evidence sources, being able to point to legitimate and reliable sources that can be considered to be rigorous may be of increasing value.

*Do we have infrastructure for now and the future?*

The growing diversity of research outputs creates new challenges in relation to infrastructures needed at all levels – from capability to enable secure digital platforms to share, collaborate and enable computational analysis of data, to increased complexity of digital archiving, discovery and access. The increasing diversity of private sector tech solutions, publishing platforms, institutional and other public digital environments, means that the infrastructure in which research outputs are shared, exchanged, preserved, accessed and assessed is becoming more complex and difficult to navigate. The international nature of research adds to the overall complexity. There is emerging evidence that researchers are finding it difficult to navigate a growing range of repositories and platforms, and keep in touch with developments in their discipline. The infrastructure required for text and data mining remains insufficient and expensive, as well as difficult to navigate in legal terms, even with the copyright exception in place. Our overall understanding of the total infrastructure supporting research outputs is insufficient, especially how it functions across different players in the field – e.g. libraries, funders, publishers, tech etc.

*Are persistent identifiers and/or AI part of the solution – and how do we promote their use?*

The scope of technology and other solutions to address some of these challenges were explored. Persistent identifiers, such as DOIs, acting as unique IDs for outputs to enable their consistent identification and referencing were identified as a key part of the solution. Ensuring their consistent use was raised as a challenge and an important route forward to help make this more problem more tractable across many of the potential aims of output capture and analysis. We looked at the successful example of DataCite in establishing an international DOI solution. The role of AI was also discussed as a part of the solution, though the potential risks associated with this, such as biases, were also raised. Lack of transparency of some of the algorithms underpinning the way information is curated and presented among the research community, the public and other actors was also raised as a challenge. Engaging with and interpreting material when there is a lack of understanding regarding how and why it has been identified as relevant can present challenges – both in confidence where this is understood and acknowledged, or in the appropriateness and trust in that information where that is not clear to users.

*What skills do people need and what culture do we want to incentivise in the future?*

Linked to the challenges around new technological solutions is the need for data literacy, within and beyond the research community. Examples of different actors working in this space have started to emerge, such as journals now providing courses on how to provide datasets as the quality of data

provided to them has been low. Digital skills will increasingly become central to the skills set required by researchers and there is increasing need to work across the sector to ensure these skills are valued, sought and provided, and are mainstreamed across disciplines and contexts. To do this, support will be needed from funders and institutions. Data skills and quality data curation also need to be embedded across the research life cycle. Data planning should be part of research projects from the outset, rather than coming at the end of the project when it is time to publish. If data plans aren't in place from the outset, data may not be generated and stored appropriately creating challenges later. Creating a culture of openness, planning and data literacy will go beyond the provision of skills to their embedding and adoption across the sector.

## Reflections

Reflecting across these discussion points, it is apparent that the changing nature of research outputs has the potential to impact on a wide range of actors in the sector, and that joined up thinking and action is needed. As we see the explosion in diversity of research outputs, we can either be reactive, responding to needs and challenges as they emerge, or proactive, to help shape and guide the nature and effective preservation of research outputs. A more proactive stance could help drive research towards better practice in information storage, sharing and communication, but requires early action and shared goals at a sector level. Since the workshop, the Covid-19 situation has provided us with a stark example of the importance of the effective and rapid sharing of research outputs. Continued dialogue and sharing of views on this topic will be important to make sure these issues are appropriately and adequately addressed.