Response to Request for Information - FR Doc. 2011-28621

Dr Cameron Neylon – U.K. based research scientist writing in a personal capacity

Introduction

Thankyou for the opportunity to respond to this request for information and to the parallel RFI on access to scientific publications. Many of the higher level policy issues relating to data are covered in my response to the other RFI and I refer to that response where appropriate here. Specifically I re-iterate my point that a focus on IP in the publication is a non-productive approach. Rather it is more productive to identify the *outcomes* that are desired as a result of the federal investment in generating data and from those outcomes to identify the services that are required to convert the raw material of the research process into accessible outputs that can be used to support those outcomes.

Response

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Where the Federal government has funded the generation of digital data, either through generic research funding or through focussed programs that directly target data generation, the purpose of this investment is to generate outcomes. Some data has clearly defined applications, and much data is obtained to further very specific research goals. However while it is possible to identify likely applications it is not possible, indeed is foolhardy, to attempt to define and limit the full range of uses which data may find.

Thus to ensure that data created through federal investment is optimally exploited it is crucial that data be a) accessible, b) discoverable, c)interpretable and d) legally re-usable by any person for any purpose. To achieve this requires investment in infrastructure, markup, and curation. This investment is not currently seen as either a core activity for researchers themselves, or a desirable service for them to purchase. It is rare therefore for such services or resource need to be thoughtfully costed in grant applications.

The policy challenge is therefore to create incentives, both symbolic and contractual, but also directly meaningful to researchers with an impact on their career and progression, that encourage researchers to either undertake these necessary activities directly themselves or to purchase and appropriately cost third party services to have them carried out.

Policy intervention in this area will be complex and will need to be thoughtful. Three simple policy moves however are highly tractable and productive, without requiring significant process adjustments in the short term:

a) Require researchers to provide a data management or data accessibility plan within grant requests. The focus of these plans should be showing how the

project will enable third party groups to discover and re-use data outputs from the project.

b) As part of the project reporting, require measures of how data outputs have been used. These might include download counts, citations, comments, or new collaborations generated through the data. In the short term this assessment need to be directly used but it sends a message that agencies consider this important.

c) Explicitly measure performance on data re-use. Require as part of bio sketches and provide data on previous performance to grant panels. In the longer term it may be appropriate to provide guidance to panels on the assessment of previous performance on data re-use but in the first instance simply providing the information will affect behaviour and the general awareness of issues of data accessibility, discoverability, and usability.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

As noted in my response to the other RFI, the focus on intellectual property is note helpful. Private contributors of data such as commercial collaborators should be free to exploit their own contribution of IP to projects as they see fit. Federally funded research should seek to maximise the exploitation and re-use of data generated through public investment.

It has been consistently and repeatedly demonstrated in a wide range of domains that the most effective way of exploiting the outputs of research innovation, be they physical samples, or digital data, to support further research, to drive innovation, or to support economic activity globally is to make those outputs freely available with no restrictive terms. That is, the most effective way to use research data to drive economic activity and innovation at a national level is to give the data away.

The current IP environment means that in specific cases, such as where there is very strong evidence of a patentable result with demonstrated potential, that the optimisation of outcomes does require protection of the IP. There are also situations where privacy and other legal considerations mean that data cannot be released or not be fully released. These should however be seen as the exception rather than the rule.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

At the Federal level only very high-level policy decisions should be taken. These should provide direction and strategy but enable tactics and the details of implementation to be handled at agency or community levels. What both the Federal Agencies and coordination bodies such as OSTP can provide is an oversight and, where appropriate, funding support to maintain, develop, and

expand interoperability between developing standards in different communities. Federal agencies can also effectively provide an oversight function that supports activities that enhance interoperability.

Local custom, dialects, and community practice will always differ and it is generally unproductive to enforce standardisation on implementation details. The policy objectives should be to set the expectations and the frameworks within local implementation can be developed and approaches to developing criteria against which those local implementations can be assessed.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Prior to assessing differences in performance and return on investment it will be necessary to provide data gathering frameworks and to develop significant expertise in the detailed assessment of the data gathered. A general principle that should be considered is that the *administrative and performance data* related to accessibility and re-use of *research* data should provide an outstanding exemplar of best practice in terms of accessibility, curation, discoverability, and re-usability.

The first step in cost benefit analysis must be to develop an information and data base that supports that analysis. This will mean tracking and aggregating forms of data use that are available today (download counts, citations) as well as developing mechanisms for tracking the use and impact of data in ways that are either challenging or impossible today (data use in policy development, impact of data in clinical practice guidelines).

Only once this assessment data framework is in place can detailed process of cost benefit analysis be seriously considered. Differences will exist in the measurable and imponderable return on investment in data availability, and also in the timeframes over which these returns are realised. We have only a very limited understanding of these issues today.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

If stakeholders have serious incentives to optimise the use and re-use of data then all players will seek to gain competitive advantage through making the highest quality contributions. An appropriate incentives framework obviates the need to attempt to design in or pre-suppose how different stakeholders can, will, or should best contribute going forward.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

As with all research outputs there should be a clear obligation on researchers to plan on a best efforts basis to publish these (as in make public) in a form that most effectively support access and re-use tensioned against the resources available. Funding agencies should make clear that they expect communication

of research outputs to be a core activity for their funded research, that researchers and their institutions will be judged based on their performance in optimising the choices they make in selecting the appropriate modes of communication.

Further funding agencies should explicitly set guidance levels on the proportion of a research grant that is expected under normal circumstances to be used to support the communication of outputs. Based on calculations from the Wellcome Trust where projected expenditure on the publication of traditional research papers was around 1-1.5% of total grant costs, it would be reasonable to project total communication costs once data and other research communications are considered of 2-4% of total costs. This guidance and the details of best practice should clearly be adjusted as data is collected on both costs and performance.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Ideally compliance and performance will be trackable through automated systems that are triggered as a side effect of activities required for enabling data access. Thus references for new data should be registered with appropriate services to enable discovery by third parties – these services can also be used to support the tracking of these outputs automatically. Frameworks and infrastructure for sharing should be built with tracking mechanisms built in. Much of the aggregation of data at scale can build on the existing work in the STARMETRICS program and draw inspiration from that experience.

Overall it should be possible to reduce the burden of compliance from its *current* level while gathering vastly more data and information of much higher quality than is currently collected.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

There are a variety of proven methods for stimulating innovative use of data at both large and small scale. The first is *to make it available*. If data is made available at scale then it is highly likely that some of it will be used somewhere. The more direct encouragement of specific uses can be achieved through directed “hack events” that bring together data handling and data production expertise from specific domains. There is significant US expertise in successfully managing these events and generating exciting outcomes. These in turn lead to new startups and new innovation.

There is also a significant growth in the number of data-focussed entrepreneurs who are now veterans of the early development of the consumer web. Many of these have a significant interest in research as well as significant resources and there is great potential for leveraging their experience to stimulate further growth. However this interface does need to be carefully managed as the cultures involved in research data curation and web-scale data mining and exploitation are very different.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

The existing norms of the research community that recognise and attribute contributions to further work should be strengthened and supported. While it is tempting to use legal instruments to enforce a need for attribution there is growing evidence that this can lead to inflexible systems that cannot adapt to changing needs. Thus it is better to utilise social enforcement than legal enforcement.

The current good work on data citation and mechanisms for tracking the re-use of data should be supported and expanded. Funders should explicitly require that service providers add capacity for tracking data citation to the products that are purchased for assessment purposes. Where possible the culture of citation should be expanded into the wider world in the form of clinical guidelines, government reports, and policy development papers.

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

At the highest level there are a growing range of interoperable information transfer formats that can provide machine readable and integratable data transfer including RDF, XML, OWL, JSON and others. My own experience is that attempting to impose global interchange standards is an enterprise doomed to failure and it is more productive to support these standards within existing communities of practice.

Thus the appropriate policy action is to recommend that communities adopt and utilise the most widely used possible set of standards and to support the transitions of practice and infrastructure required to support this adoption. Selecting standards at the highest level is likely to counterproductive. Identifying and disseminating best practice in the development and adoption of standards is however something that is the appropriate remit of federal agencies.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

There is now a significant literature on community development and practice and this should be referred to. Many lessons can also be drawn from the development of effective and successful open source software projects.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

There are a range of global initiatives that communities should engage with. The most effective means of practical engagement will be to identify communities that have a desire to standardise or integrate systems and to support the technical and practical transitions to enable this. For instance there is a

widespread desire to support interoperable data formats from analytical instrumentation but few examples of bringing this to transition. Funding could be directed to supporting a specific analytical community and the vendors that support them to apply an existing standard to their work.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Development in this area is at an early stage. There is a need to reconsider the form of publication in its widest sense and this will have a significant impact on the forms and mechanisms of linking. This is a time for experimentation and exploration rather than standards development.