Neuroinform (2014) 12:361–363
DOI 10.1007/s12021-014-9239-0

EDITORIAL

Data Persistence Insurance

David N. Kennedy

Published online: 2 July 2014

As various journals begin to ratchet up the demands on their authors for including data sharing as part of their manuscript submission process,[1,2] one can see the angst increasing within the neuroscience community. Lots of discussion will continue to transpire about why neuroscience data should or shouldn't, can or can't be shared,[3,4] but perhaps we can consider a different approach. Lets say that a neuroscience funding agency would like to take out an insurance policy on the 10-year availability of the raw data acquisition funded by their programs. This is not data 'sharing'; this is just a guarantee that the specific data acquisitions of a specific funding award are accessible to the specific investigator for a specific period of time. What would such an insurance policy cost? And what would such an insurance policy be worth? As always, the answer to 'cost' and 'value' are exceedingly sub-domain specific and varies drastically from single-cell electrophysiology to phase 2 human treatment trials, etc.

Instead, let's consider how such an insurance policy could be implemented within a specific sub-domain. We would want to consider a sub-domain that is large enough to represent a substantial investment in research dollars, and mature enough to have some standards established for data representation and best practices in study design and execution. For this example, lets consider neuroimaging. Now, consider what would be the impact of a policy announced by some large neuroimaging funding agency that, as of some specific future date, all neuroimaging data acquired as part of that agencies funded research had to be *archived* in a 'certified' repository for a period of 10 years from date of acquisition. The cost of such archival would be covered by the funding agency as part of the original funding that supports the acquisition, and this additional cost would be capped at, say, 5% of the original acquisition cost. This could be perceived as a 5% 'tax' on data acquisition in order to support long-term data persistence. For grant programs that have specific total budgetary caps, a 5% tax for data persistence would equate to a 5% reduction in the number of subjects that could be acquired for the same amount of grant dollars.

Such a policy announcement, with a sufficient lead-time, could establish what the criteria for 'certifiable persistence' would be, and expose this substantial future market to the commercial sector, reducing the funding agencies need to develop and support their own data storage infrastructure. The lead-time can be set such that an evaluation of the suitability of available 'products' could be conducted in order to establish that viable solutions exist prior to proceeding to the implementation phase of the policy.

Can data persistence for neuroimaging be achieved at a 5% cost, and what would this 'market' look like? Establishing exactly how much research funding is spent in neuroimaging is quite challenging. But for the sake of discussion, we can consider the following lower bound. A search of the NIH Reporter[5] grant database indicates that there are currently 1,229 active R01 research grants that include 'MRI' and 'brain' in their description. At an average direct cost funding of $400K per grant, this represents $0.5 billion in grant support for just this small sector of the overall neuroimaging research portfolio.

[1] http://www.plosone.org/static/policies#sharing
[2] http://f1000research.com/author-guidelines#submission
[3] Kennedy DN. The benefits of preparing data for sharing even when you don't. Neuroinformatics. 2012 Jul;10(3):223–4.
[4] De Schutter E. Data publishing and scientific journals: the future of the scientific paper in a world of shared data. Neuroinformatics. 2010 Oct;8(3):151–3.

D. N. Kennedy (✉)
Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA 01605, USA
e-mail: David.Kennedy@umassmed.edu

[5] http://projectreporter.nih.gov/reporter.cfm

A neuroimaging R01 might acquire approximately 30 subjects per year, and a typical MRI exam (including structural, functional and diffusion scanning) (see [6] for example) might include approximately 400GB uncompressed raw data per 1-hour session (approximately 140MB after lossless compression). If we estimate that a typical 1-hour MRI session might cost $500 for the data acquisition alone, this represents about $15K per year per grant, for a total of approximately $20M in imaging costs for just this small sector of the overall neuroimaging research portfolio. The 5% insurance on this image acquisition would represent a $1M new market, and this is a gross underestimate of what the true neuroimaging investment is and what this resultant insurance market would be.

Using todays cloud data storage solutions (as provided by Amazon Web Services,[7] just as an example), 400MB of data can be stored in S3[8] at a cost of $0.14 per year, bringing the 10-year insurance policy cost to $1.40. This storage cost is well under the average $25 target price of insurance that the 5% persistence tax would support. Clearly there is room for building an improved 'product' that would better deal with billing (grantees would want to be able to pre-pay the 10-year storage at time of acquisition), simplicity of data transmission to the archival location (direct DICOM transmission from the scanner), security and privacy issues, data access costs, etc. Many of these issues, however, are already routinely solved in the clinical domain by the RSNA ImageShare program.[9]

Is this a large enough market to draw commercial interest? Will research institutions see an opportunity to retain this funding in-house, and provide a similar class of certified neuroimaging data persistence to their investigators? Within the commercial sector, one can expect competition for this market, and this competition should help to either lower the costs below the target, or to generate a higher level of service.

An important question is, then, if the effective 'tax' for data archival and persistence would generate a greater 'value' in the net scientific enterprise? If not, then the scheme could be considered unfounded, or the cost would need to be lowered further in order to at least match the actual value. But what is the 'value' of this persistence insurance? The immediate 'value' would be that this would enable virtually trivial compliance with increasing demand by publishers for data related to published papers to be available. The problem with the oft-used 'data available upon request' data availability position taken by many authors is that it is well documented that data is usually not available (or readily usable if it is) .[10] This itself would be a major step forward on the initiative to increase scientific reproducibility within the field. Estimates of the rate of (unintentional) errors in the scientific literature are startling,[11] and the archival process provides both a means of retrospective checking, and prospective generation, of more accurate data reporting details. While data sharing *per se* would not part of this initial mandate, this archival step will potentiate future data sharing and integration privately (within individual laboratories, between collaborating labs) and eventually publically. Supporting archival up front in the process (no data should be acquired that is not archived and tagged by funding source) will make data sharing, which typically happens years later in the overall scientific process, easier to facilitate.

The routine impact to the researchers operations will be relatively minimal. With the potential for transparent, secure scanner to archive facility data transfers, the researcher goes about their scientific life as usual: acquiring, processing, deriving, thinking, concluding, and publishing. The insurance policy, form the raw data point of view, is that there is a guarantee that for all published studies from a funded research project, an authorized individual can get back to the raw imaging data. This is, in effect, a guarantee that 'data available upon request' would have a chance of being true. Those investigators who want to, of course, can do so much more.[12,13] While initially envisioned specifically for raw data, where the access point between data acquisition and archive facility could be seamlessly facilitated, it is clear that an equivalent scheme can be envisioned for derived data. Conceptually, for a given derivation, there is a processing cost and an amount of subsequent data generated, and these results should also persist at a reasonable percentage of cost in order to insure availability. Although the value of the derived data availability to the pursuit of scientific integrity is high, the access points in the routine scientific process, a costing framework, and a standardized data representation are less clear at this point, and could be deferred until a future date.

One of the benefits of this scheme is, in a sense, that it will cost the funding agencies virtually nothing. In the end, the same grant funding limits apply, so exactly the same amount of grant funding can be supported. The funding agencies do not need to invest directly in the underwriting of the archival infrastructure. They pay for it out of current spending, and

[6] Brown TT, Kuperman JM, Chung Y, et al. Neuroanatomical assessment of biological maturity. Curr Biol. 2012 Sep 25;22(18):1,693–8.
[7] http://aws.amazon.com/s3/pricing/
[8] http://aws.amazon.com/s3/
[9] http://www.rsna.org/Image_Share.aspx
[10] Wicherts JM, Bakker M, Molenaar D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PLoS One.
[11] Casadevall, A., Steen, R. G., Fang, F. C. Sources of error in the retracted scientific literature. *FASEB J.* 28, 000–000 (2014).
[12] Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., … Milham, M. P. (2010). Toward discovery science of human brain function. Proceedings of the National Academy of Sciences of the United States of America, 107, 4,734–4,739. doi:10.1073/pnas.0911855107
[13] Mennes M, Biswal BB, Castellanos FX, Milham MP. Making data sharing work: the FCP/INDI experience. Neuroimage. 2013 Nov 15;82:683–91.

defer the actual implementation to the commercial sector. The cost to the agency will be in lower number of subject that might be scanned (minus the cost of improved scientific integrity), and the cost of issuing and evaluating the 'challenge' to develop the archival solutions. The commercial market forces drive the rest of the process.

In a sense, this entire policy would change nothing. It doesn't change the funding bottom line costs, it doesn't prohibit 'data available upon request' availability policy, it doesn't itself promote data sharing, and it doesn't change an investigators day-to-day operation. It is an insurance policy that would promote a new data archival habit, a proper cost apportionment system for the objective of data availability, a fiscally responsible way to approach data persistence, and the engagement of the commercial sector to work for the neuroimaging community. And in this sense, it could change everything.