Report Knowledge Exchange workshop
Making data count - Research Data and Research Assessment
11-12 April 2013, Berlin

Changing the vicious circle into a virtuous circle

Executive summary
The workshop Making Data Count was attended by almost 80 participants from a wide variety of countries, organisations and disciplines. The keynotes showed a clear picture that data sharing is very important, in fact not sharing data should be considered scientific malpractice. The ‘Value of Research Data’ report argued for changing the vicious circle of researchers not being rewarded for sharing their data and there being insufficient metrics to prove the sharing. This should be changed into a virtuous circle in which researchers are rewarded for sharing, in which there are a sufficient number of cases which allows for metrics which show the sharing. Researchers can then be rewarded for their sharing. The researchers on the panel brought up clear arguments why data should be shared, irrespective of discipline. Although in some disciplines data cannot be made public due to privacy or security issues it is still worth making it available to other researchers. Breakouts then discussed the challenges in various areas and came up with recommendations1 in each area.

At present there is insufficient experience with alternative metrics and therefore it is hard to judge their value. Experiments are needed and as a first step DataCite showed an interest in sharing downloads and CWTS is eager to analyse these. Alternative metrics are not (at present) considered suitable as a measure of scientific quality. Peer review (both prior to publication or after through citations) still has an important role to play.

Funders should incorporate the sharing of research data sets more strongly when judging project proposals and should offer funds to make this possible. Institutions should also provide training, support and awareness raising not only among junior but also senior researchers. The underlying infrastructures (identifiers, etc.) must be in place and easy to use.

Welcome to the workshop
Angela Holzer from the German Research Foundation welcomed2 all participants and introduced the aim of the workshop. The purpose was to discuss how the publication of data can be integrated in formal research assessment and review procedures and explore where metrics can play a role in this.

Simon Hodson from Jisc then offered a historical perspective on the question at hand. This question is not new, back in 1680 John Flamsteed did not share his observations on the great comet. He was relentless in withholding his data and regarded these as his own private property. He argued that there was not sufficient quality in the data. It is interesting to note that such arguments are still heard today3. Several initiatives have responded to this challenge by giving data publishing a place in the scholarly process. This offers researchers better discoverability, their results are interpretable and this provides credit for their research. In the three keynotes this challenge was addressed from different perspectives.

1 The recommendations are available as annex to this report
2 All presentations, photographs and video recordings are available online at: http://www.knowledge-exchange.info/Default.aspx?ID=576
Assessment places demand on universities
Katrien Maes from the League of European Research Universities (LERU) presented how research universities are tackling open data and research assessment. LERU released a paper in May 2012 on research assessment which argued for the need to take stock and promote better understanding of research assessment. Assessment is valuable for universities to identify other research institutions to engage with, for strategic planning, to identify weaknesses and to track individual researchers’ competences. It also connects with the need for accountability for their public funding. Assessment does however require a lot of effort and can be unhelpful and lead to a short term focus. It can even lead to bean counting rather than identifying high quality research. Therefore a more sophisticated approach is required which allows for disciplinary differences and differences in time frames on which return on investment is expected.

Peer review is a wide spread method to assess research, although it does have drawbacks. It is time consuming, costly, difficult across disciplinary boundaries and can be biased. Bibliometrics as an alternative is complex: it raises challenges on how to address multiple authors, the labelling of researchers and institutions and how to prevent self citation. Impact is a new trend: in the UK the REF will assign 20% of its score based on impact and in the US Starmetrics is for example looking at employment outcomes of research. It is important to realise that there is a cost linked to collecting these measures. Research assessment is here to stay but good human expertise is required and central databases need to be in place that produce fine grained accurate data. The position of LERU is to take a sensible approach to assessment, be mindful that it does not lead to excessive costs.

LERU is also working on Open Knowledge, the LERU Roadmap to Open Access presents actions for universities to embark on. Policies are not yet in place and it is not yet clear how detailed these should be. These policies will also need to be implemented. Several questions need addressing: how to describe the data, can open be the default, how to deal with public-private data? Coordination on metadata standards is necessary and there has to be clarity on the provenance and trustworthiness of data. There is the question whether every university needs to provide their own solution. Advocacy would need to take place among researchers, to ensure that benefits are sufficiently clear. Data sharing could be embedded in researcher careers, both early on in their career and also when training the senior researchers. The new role of the data scientist also needs to be developed. It is important to build research data into review and recognition processes.

In discussion the question was raised who should pay for measuring assessment, should this be the institution or the funder? The returns are potentially important, however universities still need to travel a way to recognise that these tools are worthwhile towards running business intelligence.

Not sharing is scientific malpractice
Geoffrey Boulton, the chair of the working group that wrote the Royal Society report Science as an open enterprise then offered the perspective of the research funder. The aim of funding is to produce the best science and its useful application. This is a different motivation to universities and institutes which are in competition for funding and reputation.

Historically there was already the realisation that evidence should be published along with the argument. The current data wave which is encroaching us offers both problems and opportunities. It raises the question where we should store this large amount of data as only 11 percent of research can be replicated. For a large proportion of research data metadata is not available or not adequate. If we can integrate databases the potential for science will grow further. However a whole series of tools and trained experts are required to manipulate data effectively. Scientists will have to change the way they work, the current practice of not offering insight into data is embarrassing.

Funders should fund to get science moving and should focus on the outcomes and not the outputs. The current system of credits inhibit the process of sharing. Openness of itself has no value, we require accessible, intelligible, assessable and re-usable data. It is worth targeting the public interest in science as the distinction between paid and not paid scientists (citizen science) is diminishing. However there are boundaries to openness which we have to keep in mind: commercial interests, privacy, safety and security. However these boundaries are fuzzy.

Young scientists do matter in this process. They will need data which is easy to remix, credit, training and support. The next step would be to mandate an intelligent openness and accept diverse outputs. We do need to change the mind-set. Universities have to accept responsibility and offer incentives for example by incorporating data sharing in promotion criteria. Assessor will need to recognise a diversity in contributions.

**Reward the effort required to share**

Denis Huschka from the German Data Forum then offered the perspective of a research performing organization. He remarked that more data is available than ever before, but most of it is invisible, hidden away. Finding data should be as easy as finding a book on Amazon. However somebody has to do some work to make it visible.

There are four good reasons to share: first of all it is a precondition of scientific work that findings should be replicable. There is also the economic argument, making optimal use of public funded results. It will also have a positive impact on society, will boost competition between researchers and reduce the chances of wrong conclusions. Finally those who share data are cited much more. This is the most important argument as this actually touches on those that should do the work to make the data available. To support sharing a systematic arrangement on credit is required.

In social science there is real work involved: data may need to be anonymized, complex data sets need to be cleaned up, metadata needs to be supplied and code books need to be written. A suitable infrastructure is required which also offers restrictions on data use and offers formalised access.

In Germany they have taken the approach to not just leave the data in archives. They have set up 25 specialised data centres for social and economic data with contacts which can be approached by specialised users. This offers a direct link between producer and users. The providers now see what happens to their data and get free analyses which can be a further incentive to sharing. The work has been successful so far: many datasets have been made available and it has contributed to a high quality of data sets.

However the Data Forum still does not address individual researchers and changing the policies of research funders is not as easy as it sounds. Convincing researchers will require not only sticks but also carrots. They need to receive credit for their work on sharing data. As research is international an international solution is required to improve the sharing of research data.

**From the vicious to the virtuous circle**

Rodrigo Costas from the Centre for Science and Technology Studies (CWTS) at the University of Leiden presented the report ‘The Value of Research Data – Metrics for datasets from a technical and cultural point of view’ which was commissioned by Knowledge Exchange. The main aim of the project was to gain an overview of current use of data metrics and data citations and to recommend steps for the future. This was achieved through literature review, technical analysis of data repositories and interviews with diverse stakeholders.

---

7 The report is available at [www.knowledge-exchange.info/datametrics](http://www.knowledge-exchange.info/datametrics)
There appear to be two models for publishing data: data publication is simply putting data online and data publication is the process in which data is published in a journal after peer review, mirroring the scientific publishing model. The literature also showed that data metrics is an underdeveloped concept. There appear to be two types of indicators: citation based and Altmetrics\(^8\) based, (examples are downloads or references in tweets). There are still challenges in using data metrics at present. There is a very low involvement of researchers in data sharing, as it is not a criterion for tenure and it takes researchers’ time and money. This leads to insufficient critical mass to set up data metrics and properly assess the meaning of these metrics. We appear to be stuck in a data sharing vicious circle.

The solutions could be a reward system that considers data metrics. This will need a formalisation of data metrics and for example the tracking of data metrics to be part of research environments. To keep track we will need data citation standards and institutional commitment.

The report closes with a long list of recommendations. The central point is that data sharing should be promoted amongst researchers and reward systems should be developed that include data metrics. Technical and organisational problems will need addressing and standards for data publication and interoperability of systems are required. The future steps would be to develop a conceptual model for data metrics. Experience needs to be acquired in assessing the value of data citation. Policies are required for selection and efficient preservation of data.

**Data available to all disciplines**

Wolfram Horstmann from the Bodleian library at the University of Oxford led a panel session in which a variety of researchers offered their perspective on the recognition of data sharing in research assessment. Ortwin Dally from the German Archaeological Institute explained that archaeological research is often multidisciplinary and they are used to sharing data and metadata. Lost documentation cannot be repeated and data has a high reuse potential. As there are many points of contact with other disciplines there is a need to work on integration of data on a higher level in intelligent data repositories.

Ross Mounce at Bath University noticed in the meta analysis during his PhD research that in the field of morphology there is no tradition of sharing data. The data appears to be there, but this only partly true. Various barriers stand in the way: information is hidden in pdfs, papers are subject to copyright and there are text mining barriers. There truly is a crisis of reproducibility in research. His interest in getting people to share data, has led him to work for the Open Knowledge Foundation.

Riitta Mustonen is deputy director of the funding agency NordForsk and was heavily involved in research assessment in the past. In her view the researcher should have incentives to share information and should know how to write a data management plan when applying for funding. Proposals are reviewed by peers and in this review they should take data sharing into account. This does require research councils to follow up and check on the actual sharing and there is a need for indicators keeping track of what has been shared.

Joachim Wambgsans works at Heidelberg University in astronomy. In this discipline publishing research data has always been an important issue and was done in an appendix or as a supplement. Twenty years ago the Astrophysics Data System was set up to get access to the journals and the data. At present there are clear agreements on sharing. For example researchers get 12 months to use the data from the Hubble telescope and then it is made open. Standards have been agreed to describe the data collected from a telescope.

In the subsequent discussion several issues were raised. Should data be stored in few large repositories or in many small repositories? The dispersion of data over many repositories could make tracing data difficult and impair cross disciplinary use of data. However both solutions were considered to be valuable. If data is spread out there does need to be a good interface to discover the data and it has to have

\(^8\) More information on Altmetrics can be found at: [http://altmetrics.org/manifesto/](http://altmetrics.org/manifesto/)
proper metadata. Examples of large repositories mentioned were Dryad, which offers a place for long tail data and Research Data Australia for all datasets in Australia. As data is multifarious the data archive is more than just the place of storage. It is also the experts behind the data and we need hubs to make the expertise supporting the data available. The current data universe at present is very much an ecology. For massive data sets a large number of supporting staff are required. But do not forget the long tail of small data stored on desktops. To support this storage we need the experience of an established data centre. Such expertise needs to be spread. At present on average only .75 FTE is employed for this at each university library.

In the example of morphological data a database was set up using Treebase which was a success in the botanical community. However applying this successfully in another scientific community will not happen naturally and will require their involvement. Funding agencies should systematically ask researchers the question whether they have shared data in the past. It does sometimes appear in peer review but review panels do need to be reminded.

Karl Popper argued that researchers first should develop an idea, then collect the information and then establish whether the theory is correct. Therefore data cannot be subsidiary to the theory, but data are part of the same process and should be rewarded as such. The practice of making data available should be considered good science. Such good conduct of science should be rewarded in grant applications. There could be a tick box in application forms checking whether they shared research data in the past. This would require some system to see whether data has been shared. Perhaps funders should have a misconduct hotline, not sharing data being regarded as malpractice. It was agreed that required extra funding should be provided by the funder thereby resulting in less research budget.

Data as the new currency of science

The participants then split up into five break out groups to discuss the requirements from five different angles. They reported back during the closing session. The closing session was started with presentations by DataOne and DataCite. Bill Michener from DataOne had provided a video message in which he highlighted three key elements to enable data publication. First of all the researcher should receive credit, this requires unambiguous identification of authors (for example through ORCID) and materials. This will allow linking of data to publications which can lead to recognition in funding. Also alternative metrics (e.g. Impact Story) can be considered to measure other types of impact. Secondly there should be user friendly tools as researchers are already overloaded with commitments. DataUp is a tool which is part of the DataOne infrastructure which allows for easy upload and adding of metadata through user friendly interfaces. Finally peer pressure is critical to change researcher’s behaviour. This will require training and encouragement by research sponsors. Researchers should be made aware of new tools which fit in to their scientific workflow, for example lab notebooks. We will need to change the way we teach at universities and data should be regarded as the new currency of science.

Jan Brase was involved in the start of DataCite 10 years ago, originally as a concept study funded by DFG. It currently assigns Digital Object Identifiers (DOIs) to datasets and offers statistics which show how often DOIs are being resolved. The top 10 most popular datasets are now highlighted on the website. The top 10 scientists are now proudly tweeting they are in the list. DataCite would be happy to work with others to gain more experience on the meaning of these metrics. DataCite also offers other tools for example to search by filters and to create links to other datasets.

In the break out group on quality assurance of data publication the discussion had been focused on the meaning of quality and the different states of openness. Discussions on data publication are still dominated by the journal publication model. There was agreement that scientific quality should be assessed by the research community. To ensure the technical quality new expertise is required. The question was also raised how to work with long tail of research that does not clearly fit in an existing metadata format. One of the tasks ahead is connecting data review with data management planning and funders have a role to play demanding that research data be deposited. Publishers also have a role in ensuring that there are links from articles to data sets, checking these and reporting issues back to the institution. When connecting data review to article review permanent and unique identifiers are key.

The top priority was considered to set up templates for different disciplines showing how data could be published and shared. This could contain examples from other domains showing peer review processes and data quality checks. A working group has been proposed under the RDA\(^8\) to work on this and they can work further on these recommendations.

The second break out group had focused on linking data to other research information. These links provide context to the data and are crucial for interpreting research data. The group suggested a package of work dealing with data citations. Citation should be standardised and repository owners need to provide user friendly tools, training and support to make citation possible. A second package of work should

\(^{8}\) The Research Data Alliance http://www.rd-alliance.org
work on providing contextual information through the common metadata model CERIF. Regarding metadata standards, the challenge now seems to be to implement policy rather than just to develop more policy.

The breakout group on new metrics and citation systems saw a two pronged approach first seeing what we can get, grabbing as much information as possible and publishing this. This could then be analysed thereby providing more clarity on the value of metrics. An immediate action could be to collect the new metrics available in DataCite and experiment with these. This can result in small studies showing why some data sets are cited and others not. This in turn should provide a richer view on how data citations can be incorporated in scientific knowledge.

On the longer term it would be worth exploring which metrics can be measured. Based on experience it could be considered what the purpose is of using a new model with alternative metrics. It might be worth doing a sociological study into how researchers share research data, thereby targeting disciplines that do not have a history of sharing data. At present we do not know what metrics mean.

The break out group on codes of conduct had noticed that there are already codes of conduct on various levels, national and international, European, self-expressions of research communities, etc. Research data are always mentioned but in the context of scientific integrity and not in the context of data sharing. Also these codes have very general points of departure. To make this clearer such codes should be supported by implementation documents and should be reviewed regularly. Funders should coordinate these policies, all learned societies should have a code of conduct and publishers should have a data availability policy.

The final break out group to present had discussed research assessment procedures at universities and funding agencies. The good news is that there are procedures in place. However these are few, patchy, sector related and focus on other things than research data availability. Alternative metrics are not seen as a prerequisite for data sharing, peer review is considered a much more convincing measure of quality for researchers.

Changing the academic culture will incentivise the researchers to change. It will help if these responsibilities are incorporated in training modules, policies and job descriptions. There should be career options for those sharing. Funders can best use projects and initiatives to drive the change. Mandates have to be implemented and formulated in a sensible way, strict mandates can help and can be a means to change culture but can have an adverse effect. It can be better to provide funds for specific issues e.g. making negative results available. An action that funders could take up now is supporting projects that explore bibliometrical tools.

Participants then discussed the meaning of metrics. Peer review is seen as a measure for scientific quality and this can take place prior to or after publication, for example through citations. Other metrics like downloads are considered a measure of attention and not quality. In this view peer review is about establishing scientific quality and alternative metrics can show the value for others.

Closing considerations by stakeholders from keynotes

Maes offered her reflections on the recommendations. She stressed that it is important to make sure that the benefits to the researcher are clear and understood and accepted. For universities it is not about compliance, but looking ahead on the road we are travelling. Universities need to establish policies on research data and commitment at the top is required.

Implementation on the other hand is a bottom up process coming from research departments. It is worrying that they are silent in some disciplines. We must beware that the sectors that are driving progress should not define the rules and we should make sure to involve those that are less active. Social sciences and humanities (SSH) could make a real contribution to this debate.

Reward and recognition are powerful incentives for researchers. Universities have to make sure to develop the right structures and mechanisms, this includes policy development, training and offering support. It is a shared responsibility in which libraries and research management offices also have a role to play.

Boulton reminded the participants of the historical context we are in. This is a disruptive technology and universities will have to adapt and cannot carry on in the same way. Data is fundamental to science and researchers measure with a purpose. Therefore the data contains the science and citing the data is crucial. The cultural change should be driven by the realisation that there are benefits to sharing data. In some areas SSH are miles ahead. Researchers working on social statistics are exemplary how they order and share data.

Huschka added that simplicity is required. A simple quick, service must help to make data available and the standards should be easy to comply with. Researchers are not being paid for writing codebooks and metadata. However visibility is very important to researchers and data should be very visible, here metrics and citations have an important role to play.
In the final closing words Simon Hodson listed a few key findings raised at the workshop. Metrics do appear to have their place in combination with peer review and we are only just starting to grasp what data metrics can mean. We do live in a world in which accountability is very important. Peer judgment is central for researchers and metrics can support this, but metrics do not have to be the unique basis. A new approach to peer review could be considered. Transparency and replicability should be part of the social contract of research.
Annex: Recommendations

By topic

Quality assurance for data publication
- Set up a template showing how data review is done in a number of domains (to share with other disciplines)

Linking of scientific information
- Standardise data citation
- Work on multidirectional linking
- Standardise contextual metadata
- Show cases in Linked Open Data

Metrics:
- Immediately start collecting and collating available metrics.
- Initiate a study to consider what metrics are appropriate to data. (We do not simply want to use existing metrics). It is too soon to understand what we want do measure. Need to consider what can be collected and study what metrics may be taken to mean.

Codes of conduct
- Improve knowledge about community codes, analyse features of data sharing communities.
- Support codes with implementation documents (including criteria and standards, discipline awareness)
- Funders should coordinate policies
- All learned societies should have a code of conduct covering data sharing and preservation
- Journals should have adequate data availability policies; where appropriate should have data editors
- Publicity and advocacy for codes of conduct; training for researchers

Research Assessment
- Metrics should not be the sole basis of research assessment. Peer review is better suited to measure scientific quality.

For stakeholders

Universities
- Universities to promote culture of data sharing in the institution (and offer RDM support)
- Universities should ask at all points of academic life: What are you doing with your data?
- Universities to promote jobs related to data: data stewards, data scientists etc.

Funders
- Funder should set up policies, mandates should be implemented and well defined (linked to grants, encourage data reuse)
- Funders should observe development of practice and provide funding in key areas.