



European Research Council
Executive Agency

Established by the European Commission

Workshop on Research Data Management and Sharing

18-19 September 2014, ERC Executive Agency, Brussels

Abstracts of the presentations

SESSION 2: Setting the scene

Professor Sabina Leonelli, University of Exeter:

Data Sharing in Research: Four Key Concerns

Extensive data sharing and Open Data policies have a potentially transformative impact on scientific research. I discuss four aspects that I view as crucial for effective data sharing, which however raise serious concerns about how this is currently done and who should be involved in those efforts: (1) re-use; (2) sustainability; (3) size and (4) openness. I point out that effective data sharing requires shifts in research ethos and institutional structures, as well as large investments on long-term data infrastructures across the globe, including venues to coordinate and continuously update common standards. In the absence of such conditions, big data collections are destined to remain extremely partial and difficult to re-use by outsiders. I conclude by suggesting that caution in the use of the term “sharing”, which suggests, but does not necessarily entail, reciprocity and common ground among stakeholders. My analysis is grounded on ongoing empirical research on the conditions under which researchers share data in the UK, Europe, USA and South Africa, and the scientific and social implications of data handling practices around the globe. This research is currently carried out through an ERC Starting Grant on ‘The Epistemology of Data-Intensive Science’ and a Leverhulme Trust award exploring the digital divide. It has also been funded by the UK Economic and Social Research Council, the Max Plank Institute for the History of Science and the British Academy; and is closely aligned with the Global Young Academy Position Statement on Open Science, which I coordinated and co-wrote

http://www.globalyoungacademy.net/gya-publications/OpenAccess_GYAStatement.pdf.

Dr Hans Pfeiffenberger, Helmholtz Open Science working group (and Science Europe WG on Research Data):

The Open Science Imperative - Opportunities, Challenges and Limits

Over the last decade, for a variety of reasons, open access to research data has been found to be a guiding principle of handling research data, if not of good scientific practise.

Summoning first principles, the Royal Society observed in 2012, that “Open inquiry is at the heart of the scientific enterprise”. And there are quite “practical” reasons as well: It has been shown in 2014 that in a vast field of research up to one half of publications are not reproducible - which situation could only be remedied by providing openly all manner of supporting evidence, including primary data and software codes. On a more positive note, there is also evidence that sharing data for reuse could double the number of publications based on it.

But there is a number of real and perceived barriers to openness and a lack of drivers. Many of the barriers derive from tightly interwoven reasons and some may even be tracked to underdeveloped common understanding of terms and concepts or a hidden overload of meanings.

Older recommendations on good scientific practises and even new requirements of funders leave open a few questions, which are currently being addressed, such as: Is the author of an article required to release data on request or even to make it available at the time of publication of conclusions based on it? Which kind of data? When? To whom? This is now rapidly being resolved not only by funders but even more importantly by editorial boards of journals, which formulate, in quite unambiguous terms, how data are to be provided by authors.

Dr Bernd Pulverer, EMBO Scientific Publishing:

Publishing Source Data: Finding and Accessing the Data Behind Figures

In the biosciences research data are stored on a wide variety of platforms, but these can be constrained by a lack of standards and stable global infrastructures. In view of the rapid growth and diversification of data, efficient research progress depends on the development of such standards, validation processes and long-term scalable and accessible infrastructures.

The peer reviewed research paper remains the most widely used mode of sharing and archiving validated research findings in the biosciences. In papers data is published as figures, often representing a conglomerate of disparate experiments. Figures are little more than illustrations to support the textual description of the research - the reader cannot extract and reanalyze the data and it cannot be mined. A number of scientific journals now publish figures in association with the underlying source data, licenced for open access and reuse. Computer-readable metadata that describe the content of such datasets will connect the peer-reviewed scientific literature to biomedical databases and data repositories. This development may enable new search strategies, rendering data and experiments directly discoverable through the figure panels in a paper, facilitating data integration across the literature.

The publication of validated source data with associated reagents and methods in journals and repositories will ensure transparency, reproducibility and accountability for the reporting of scientific research.

Dr Roar Skålin, Research Council of Norway:

Norwegian Researchers Want to Share Data but Fear Jeopardising Their Career

As a part of developing a policy on sharing and archiving of research data, the Research Council of Norway (RCN) has taken a closer look at how researchers in Norway store their data, the extent to which they share research data and what kinds of obstacles get in the way of data sharing.

The findings of the study show that most Norwegian researchers are in favour of making research data accessible to other scientists. Nearly three of four researchers are willing to share their data. Eight of ten believe that open access to research data enhances research.

Although most researchers are in favour of using data generated by others, some remain sceptical about sharing their own data. The study finds that researchers would prefer to retain control over who gets access to their data and how the data are used. One factor behind this concern is that it

takes time to make data accessible. In addition, some scientists are worried that data sharing could compromise their possibilities for scientific publication in the future. The lack of a technical infrastructure is another obstacle.

The study, which is documented in the report "[Sharing and archiving of publicly funded research data](#)" (pdf-1 342.7 kB, in English), was carried out by [DAMVAD Norway](#), a research-based analysis consultancy company. A [news article](#) on the study can be found at the RCN webpages.

SESSION 3a: Domain specific perspectives: Life Sciences

Dr Iiris Hovatta, Department of Biosciences, University of Helsinki:

Towards improved data management and sharing in behavioral neuroscience and genomics

My lab is investigating the genetic and neurobiological basis of anxiety disorders. Anxiety and fear are normal emotional responses to threatening situations. In anxiety disorders these responses are excessive, prolonged, and disturb daily life. Anxiety disorders, which include panic disorder, obsessive-compulsive disorder, post-traumatic stress disorder, social phobia, specific phobias, and generalized anxiety disorder, were the most common mental disorders in EU in 2010 with prevalence of 14%. The major challenges in the field are to identify the molecular events that initiate and maintain pathological anxiety, and determine how to normalize this pathology. Accordingly, there is a need to find novel, well-defined and clinically relevant drug targets. We have used mouse and human genetic approaches to identify genes that regulate anxiety. We have setup several mouse models to investigate both innate anxiety and stress-induced anxiety. To identify gene networks involved in the regulation of anxiety, we carry out RNAseq and small RNAseq of known brain anxiety circuits, followed by bioinformatic analysis. We then carry out functional studies using pharmacological and genetic approaches to investigate the mechanisms underlying the demonstrated transcriptome changes. Importantly, to translate our findings from mouse to human, we carry out genetic association analyses in anxiety disorder patients and controls derived from the Finnish Health 2000 epidemiological cohort. In the presentation, I will discuss data management and sharing practices related to these projects. These practices are generally well-established in the genomics field, where several databases exist to store and share gene expression and genotype data. However, practices are considerably less-developed in behavioral neuroscience, although some databases, such as The Mouse Phenome Database (phenome.jax.org), maintained by the Jackson Laboratory exists. The main issues are the difficulty to standardize behavioral protocols across laboratories and the lack of publicly available databases to store the raw data from behavioral experiments.

Dr Bouke de Jong, Institute of Tropical Medicine, Antwerp:

INTERRUPTB – Data management & -challenges

In a cluster randomized trial on the effectiveness of enhanced case finding (ECF) of tuberculosis, the main outcome is the effect on TB case notifications. In a nested study, we assess the impact of ECF on the interruption of TB transmission, by testing whether the intervention is associated with lower genotypic clustering rates. The study generates complex and large datasets, with patient information (demographic, clinical, treatment outcome) linked with bacterial information (microbiology,

genomes). This data is 'clustered' on four levels, i.e. in space, time, by genotype, and at the level of belonging to an intervention versus a control cluster. Moreover, this large dataset will be analyzed by bioinformatics approaches to arrive at a clustering rate, and by a mathematical model that is yet to be developed to arrive at an effective case contact rate. Challenges consisting of data handling and storage, public access, and analysis of massive datasets will be addressed, as well as ethical issues surrounding public access to pathogens isolated from clinical samples.

Dr Sebastiaan Luyssaert, Laboratoire des Sciences du Climat et de l'Environnement (LSCE), Gif-sur-Yvette:

User and provider perspectives on data sharing at the interface between Life and Earth sciences

Earth system models are the most advanced tools to predict future climate. These models represent the interactions between the atmosphere and the surface beneath. Further developing the representation of the land surface in Earth system models requires process understanding, process parameterization and process validation - three tasks that heavily rely on data made available through field experiments, economic inventories, monitoring networks and remote sensing programs. These different data streams come from different communities which all have their own take on data sharing. Experiences with respect to obtaining the data and creating a more collaborative environment will be discussed. Following their release, the model is used by other researchers. Experiences with respect to software sharing will be discussed.

SESSION 3b: Domain specific perspectives: Physical Sciences and Engineering

Professor Debra Laefer, School of Civil, Structural and Environmental Engineering, University College Dublin:

The euros and cents of open access data

While many researchers perceive data sharing as a burden and commonly consider it as a loss of intellectual property, this is not necessarily so. This talk will explore the explicit financial and professional advantages to open data. In addition, some of the mechanisms for both storage and garnering the advantages will be discussed.

Professor Barbara Romanowicz, Institut de Physique du Globe de Paris and Collège de France:

Data access and other related issues in seismic imaging of the Earth's deep interior

The objective of my Advanced ERC grant, WAVETOMO, started in June 2011, is the development of three-dimensional models of structure in the earth's mantle and core at the global and continental scale using advanced seismic imaging approaches, in order to provide constraints on the dynamics and evolution of the earth through geological times.

Earthquakes generate elastic waves that travel through the earth's deep interior, illuminating its structure. These wavetrains are presently recorded digitally all over the earth's surface by hundreds of very sensitive instruments - seismometers. The corresponding waveforms as well as earthquake catalogs constitute the basic data that we exploit for imaging purposes.

Whether it is for the purpose of studying earthquakes themselves (determining their location, origin time, and rupture characteristics) or the earth's structure, it is always necessary to combine data from many seismic recording stations that are operated by different institutions and different countries. Because of this fundamental requirement, the international seismological research community has been ahead of many others by developing, since the early 1980's, global and regional data archives, most of which are open and freely accessible to any user.

So, for most practical purposes, our WAVETOMO group has not encountered any problems accessing the data needed for our research. There are just a few examples of regional data in Europe that are, for various reasons, still not freely shared.

I will describe how the seismological community is organized for data archiving and sharing, what efforts are underway for the construction and expansion of data access nodes in Europe, as well as issues and difficulties encountered in furthering the needed infrastructure and the access to data, one of which is proper attribution of the sources of data used.

Because understanding earth's dynamics and evolution necessitates a multi-disciplinary approach, I will also touch upon issues related to extending the open data archive efforts in other earth-science disciplines, such as geodesy, volcanology or geomagnetism and magneto-tellurics.

Dr Cordelia Schmid, INRIA Grenoble:

Data access to images and videos in the context of visual recognition

In this presentation we will first present the goals of the ERC advanced grant ALLEGRO and then given an overview over the associated data access problems. The ERC ALLEGRO addresses the issue of mining visual information and models from the massive and ever growing amount of digital image and video content available today, on sites such as Flickr and YouTube, in audio-visual archives such as those of BBC and INA, and in personal collections. In most cases, it comes with additional information, such as text, audio or other metadata, that forms a rather sparse and noisy, yet rich and diverse source of annotation, ideally suited to emerging weakly supervised and active machine learning technology. The ALLEGRO project will take visual recognition to the next level by using this largely untapped source of data to automatically learn visual models. The main research objective of our project is the development of new algorithms and computer software capable of autonomously exploring evolving data collections, selecting the relevant information, and determining the visual models most appropriate for different object, scene, and activity categories. An emphasis will be put on learning visual models from video, a particularly rich source of information, and on the representation of human activities, one of today's most challenging problems in computer vision. Although this project addresses fundamental research issues, it is expected to result in significant advances in high-impact applications that range from visual mining of the Web and automated annotation and organization of family photo and video albums to large-scale information retrieval in television archives.

SESSION 3c: Domain specific perspectives: Social Sciences and Humanities

Dr Peter Stokes, Department of Digital Humanities, King's College London:

Data (Re)Use in the Humanities: The Example of DigiPal

As the Digital Humanities has grown, so have issues surrounding the use and reuse of data in Humanities research. One example of this is the *Digital Database and Resource for Palaeography, Manuscripts and Diplomatic* (DigiPal), an ERC-funded project bringing digital methods to the study of historical handwriting. DigiPal includes a generalized framework for the exploration and communication of evidence in historical handwriting, and the application of this framework to a test-case of eleventh-century writing in the vernacular, central to which is about 850 digital images of medieval script (<http://www.digipal.eu/>). The project team has added structured annotations to this, allowing users to bring up images of specific letters in different contexts according to queries such as 'show me images of the letter **a** written in royal *versus* episcopal charters'. Users can save their results and share them with others for scholarly, pedagogical, or any other purpose. The framework also allows the data to be harvested automatically by anyone and this can in turn be used for aggregated searching, training computer vision systems, embedding content into online learning, and so on.

The potential here is enormous, and indeed the DigiPal framework and its implementation have already received substantial use by scholars and interested public. However, the challenges are also significant. The website includes previously published catalogue descriptions donated by other scholars, and high-resolution images of manuscripts provided by libraries and archives. Although the project team have permission to use this material, each typically comes with a different set of conditions, with different licences and copyright status. Attribution of IP can also be unclear. A library may retain ownership of the image, but what of a single letter which the team has cropped from it? Does this cropping constitute fair use or produce a derivative work? Does the DigiPal team own the coordinates of the cropping while the library retains ownership of the cropped image (as one licence has stipulated)? How can or should authorship be attributed in the case of automatic harvesting of data? Is a fixed-term licence acceptable given the limits this places on the site's sustainability, and, if so, what is a reasonable minimum period? These questions and others have had to be negotiated on a case-by-case basis, one institution at a time. There has been considerable success with this, and goodwill on all sides. Nevertheless challenges still remain, as the constraints and requirements of the various parties has meant the undeniable need to 'expend significant financial and human resources on negotiating and paying for reproduction rights' (*Dagstuhl Manifestos 2:1* (2013), p. 15), with the result still being the 'confusing patchwork of policies' that the Max Planck Institute identified in its *Best Practice For Access to Images* back in 2009.

Dr Daniela Stockman, Department of Political Science, Leiden University:

Authoritarianism2.0: Keeping Big Data Safe in China

In authoritarian regimes social media have often been described as "liberation technology" that may potentially destabilize authoritarian rule. What is it about social media that mobilizes people to act, and what kind of social media, such as Facebook or Twitter, are more influential than others? Funded by a Starting Grant of the European Research Council, Dr. Daniela Stockmann investigates

these questions in the context of China. Relying on multiple research methods (qualitative interviews, observation of internet users, online surveys, and a large nationally representative public opinion survey), the project brings together a team of one postdoctoral fellow and two PhD students. China's political system and its status as a developing country, together with the involvement of digital methods and big data, pose unique challenges to data management and data sharing. Such challenges include little differentiation between a public and private sphere among collaborators and research participants in China, Chinese law serving as an instrument of the state as opposed to protecting citizens from state interference, and differences between China, the Netherlands and the US in regulating the internet, among others. As the host institution, Leiden University has provided helpful advice by LURIS (Anna Groeninx) and the University library (Fieke Schoots). Support can be improved by adjusting the ethical review procedures, providing funding for projects that have to meet sudden or unanticipated challenges, building resources, investment into support staff, training of support staff on the specifics of conducting research involving authoritarian states, developing countries, digital methods and big data. For more information feel free to visit the project website at www.authoritarianism.net.

Professor Jo Van Biesebroeck, Department of Economics, University of Leuven:

Using confidential & proprietary administrative data in research

The use of administrative data in economic research, as well as in other fields in the social sciences, has greatly increased over the years. These data sources have clear advantages. Samples tend to be orders of magnitudes larger than what researchers could collect themselves using survey instruments. Reporting the information is often compulsory by law, leading to very high response rates. When the data is taken directly from administrative registries, no individuals need to answer additional questions for the purpose of data collection, avoiding problems of imperfect recall. In many circumstances, information on the universe of the population is available and where information from different data sources (government agencies) can be connected, it can provide a very rich picture of study subjects.

Unfortunately, using administrative sources of information in one's research comes with some drawbacks, in particular regarding data management and sharing. Due to the confidential and proprietary nature of these data, they can invariably not be shared freely with other researchers. This is a problem for the scientific community as it constrains verification of research methods and replication of results. It also makes it harder for a field of study to build upon prior research and collectively move the discipline forward. Access is often cumbersome, which hampers collaboration, slows research down, and tends to limit the sophistication of the research methods used.

On balance, there is certainly a place for administrative data sources in research, but their special nature needs to be acknowledged and accommodated in scientific journals' data policies and granting agencies' open access requirements.

SESSION 5a: Cross-cutting issues: Data management and data sharing

Sarah Jones, Digital Curation Centre based in HATII at the University of Glasgow:

Managing and sharing data

This presentation will give an introduction to managing and sharing data. Some of the justifications and benefits of data sharing will be outlined, as well as barriers that researchers face and how these could be overcome. The different activities involved in digital curation will be outlined and pointers given to existing tools and services that could be used. Gaps in infrastructure will be noted, drawing to a conclusion on a number of open questions for discussion, namely:

1. How do you foster open science?
2. Who is responsible for providing infrastructure and support?
3. Who should pay?

Dr Jacob Kornbeck, Office of the European Data Protection Supervisor:

The EU Data Protection Reform: what researchers should know

The presentation will provide an overview of the existing EU legal framework for data protection and the on-going reform. The main principles of data protection as defined in the current directive and the future regulation, in its current draft version, will be discussed and the relevant references to research, statistics and archives will be highlighted.

Short contribution:

Mahsa Shabani, Centre for Biomedical Ethics and Law, KU Leuven:

Ethical and Legal Challenges of Data Sharing in Genomic Research

In recent years, researchers have received an impetus from funding organizations to share data generated in the course of public-funded genomic studies. Data sharing and open access policies have been developed in order to unlock the research potentials of databases and to increase their statistical power. Concurrently, introducing data sharing practices into the genomic research arena has challenged the current mechanisms established to protect rights of research participants and triggered policy considerations. In the course of the current project, firstly, we will present the results of a systematic review that investigated the attitudes of research participants and the public towards the sharing of genomic data through public databases. In these studies participants' attitudes towards genomic data sharing revealed the influence of a constellation of interrelated factors including the personal perceptions of controllability and sensitivity of data, potential risks and benefits of data sharing and also governance level considerations. This analysis indicates that future policy responses and recruitment practices should be attentive to a wide variety of concerns, particularly protecting privacy of research participants, obtaining an adequate consent and maintaining robust oversight mechanisms over the downstream uses of data. Secondly, we will discuss the potential role of Data Access Committees in response to the pertinent concerns voiced by individuals.

SESSION 5b: Cross-cutting issues: Data storage and curation – quality, sustainability and interoperability

Dr Ingrid Dillo, DANS – Data Archiving and Networked Services:

European Digital Repository Certification: the way forward

Sharing research data is important. Data availability enables replication of research and thus enhances scientific transparency. Reusability of data also contributes to scientific progress and the advancement of knowledge. Research funders acknowledge the importance of data sharing and require a high degree of open access.

But there are all sorts of snags to sharing data. The data should be searchable and accessible to others than those who collected them, and they must be stored in a consistent and sustainable way. Sometimes the privacy of individuals needs protecting. Also, many researchers do not want others to publish about the data collected by them before they themselves have had the opportunity to do so.

Good, timely data management and secure data storage, both during and after completion of a research project, are essential prerequisites for sharing those data. In my presentation I would like to focus on the importance of certification as a means to establish trust in digital repositories.

Dr Mark Hahnel, figshare:

Future-proofing digital research

Responding to the need for open access to data, a variety of new models for sustainability of repositories have arisen in response to different contexts. From subject-specific repositories, to funder, institutional and commercial platforms, it could be suggested that research is becoming more and more silo-ed. However, by adhering to a set of basic principles defined here, we propose that the bucket the files are stored in is not the critical factor here, rather the access and availability for re-use by humans and machines alike.

It is no longer efficient or sustainable for humans to be the gatekeepers of academic content. Principles must be put in place so that content can still be reused long after the data creators are dead. Content must adhere to ethical and commercial sensitivities where necessary, but with the dawn of research data management plans the funders of research ultimately decide what level of access society has. Provided that new platforms for dissemination of academic content abide by these principles, the academic research community should succeed in the same way that other heterogeneous yet interoperable infrastructures have done previously. The obligations of academic institutions and governments mean that solutions will be put in place, but will the incentive structure for academics be high enough for researchers to use them in the most efficient manner?

[Short contribution:](#)

Dr Andreas-Alexander Maul, Federal Institute for Geosciences and Natural Resources (BGR), Germany:

Geodata Management at BGR and its Challenges

The Federal Institute for Geosciences and Natural Resources (BGR) is the central geoscientific authority providing advice to the German Federal Government in all geo-relevant questions. It is subordinate to the Federal Ministry for Economic Affairs and Energy. Numerous geologists, geophysicists and soil scientists are employed at the institute producing spatial data, maps and reports on geology, hydrogeology, soil and natural resources. Currently the main focus is on the chain from the working databases with map data at the data producing units to external data bases in the demilitarized zone (DMZ) which serve quality controlled map data for the public as required by the INSPIRE directive. BGR is thereby obliged to give access to metadata by discovery services, to view services for the maps and to download services for the data.

One big challenge is to integrate very heterogeneous data like various geophysical measurements taken and processed by different methods. Common storage systems for the data and processing parameters as well as visualization tools must be developed.

The challenges for an institution to store research data have in general an organisational component to build teams to set-up and to operate central data systems and to convince colleagues to use these systems. There are technical challenges to develop long-term storage systems and methods for visualization of the data. Standardization of data structures and set-up of exchange procedures are necessary for external users to re-use the data. Quality control during the data production process is essential.

SESSION 5c: Cross-cutting issues: Data discoverability, data access and data (re-)use

Dr Johanna McEntyre, European Bioinformatics Institute (EMBL-EBI):

Enabling Open Science: Data Discoverability, Access and (Re)use

In my presentation I will discuss what making data discoverable really means. I will briefly describe some of the currently available solutions, and explore the many challenges and benefits of making data discoverable using examples from the life sciences. Articles published in peer-reviewed journals is currently one of the most important measures of scientific output in our community, so I will also review some of the recent activities of journals in data publishing. Reuse of data is where the benefits of making data discoverable really become apparent: is making data accessible enough?

Dr Stavroula Karapapa, School of Law, University of Reading:

Copyright and the Big Data: Regulating the power of knowledge

Although the emergence of the big data promises unprecedented opportunities for innovation though the accumulation of information *en masse* and the possible generation of new knowledge, it also challenges existing copyright norms. Legal issues arise as a result of copying protected works at a massive scale, transforming them into data, licensing access to them and authorizing their re-use through the application of new research tools, such as text mining and data analysis. The paper discusses the copyright issues involved in mass digitization in light of the current legislative framework and the recent changes in EU (and UK) legislation. It offers a critical account and specific recommendations on how EU copyright law can evolve to embrace the mass digital opportunity.

Short contribution:

Dr Hylke Koers, Elsevier:

Supporting research data sharing, discovery, and usage at Elsevier

In this short contribution, I will talk about how Elsevier is supporting researchers to share, discover and use research data. This spans a broad range of initiatives, tailored to the needs of individual research communities and to different roles that researchers can play in the scientific process. I will highlight setting up links between articles and data repositories, which help to make data more easily discoverable and place them in the right context; data-integration and visualization tools that make it easier for researchers to explore and re-use data while reading articles; and data management and data capture tools that support researchers with data-intensive tasks in the lab. Finally, I will briefly touch upon the work carried out in international working groups on data citation and other aspects of research data management.

SESSION 5d: Cross-cutting issues: Rewards and incentives for good data management, data sharing and re-use

Dr Paul Ayris, Chief Information Officers Community, LERU - League of European Research Universities:

Implementing the Future: the LERU Roadmap for Research Data

This presentation will look at the [LERU Roadmap for Research Data](#), which was issued by the League of European Research Universities in January 2014. In a series of seven chapters, the [Roadmap](#) discusses a series of issues related to research data management and open data: Policy and Leadership; Advocacy; Selection and Collection, Curation, Description, Citation, Legal Issues; Research Data Infrastructure; Costs; Roles, Responsibilities and Skills; Recommendations. The [Roadmap](#) is the first analysis of its kind explicitly to the needs of universities and their researchers.

Following a description of the main Recommendations in the Roadmap, the presentation will look at levels of take-up of the [Roadmap](#) and its Recommendations based on a recent survey of LERU members. Comparisons with the level of take-up for the [LERU Roadmap Towards Open Access](#) will provide instructive insights into how European research-led universities are tackling issues around research data management.

Finally, the presentation will draw a number of conclusions about research data management in European research-led universities.

Dr Sünje Dallmeier-Tiessen, CERN / ODIN - Orcid and DataCite Interoperability Network:

Incentives for Open Science: Attribution, Recognition, Collaboration

With the evolution from Science to e-Science, the research community in general, and each individual researcher is confronted to the new opportunities and challenges. The success of Open Science depends on a delicate balance of incentives and disruption. Most researchers start to enjoy the unparalleled opportunities of (big-)data-driven research, increasingly discuss online or cooperate

across countries and time zones, and tweet to comment about interesting research results. At the same time, funding agencies, and society at large, are increasingly requiring transparency and accountability to the research community in many forms, among which Open Access to publications and forms of access to scientific data. The ODE (FP7 Project Opportunities for Data Exchange, 1) underlined that even when suitable e-infrastructures to share data exist, researchers often hesitate to embrace Open Data paradigms. This puts at risk the overall emergence of Open Science. This can be overcome by building appropriate incentives for researchers to get “credit” for the sharing of research results. This requires mechanisms must be put in place which enable researchers to be correctly attributed their research results, these scientific artefacts to be uniquely identifiable, and citable. The ODIN FP7 project (ORCID and DataCite Interoperability Network, 2) recently explored the global and disciplinary landscape for such an identification and attribution infrastructure. Many successful pieces are emerging and generating uptake and enthusiasm although local fragmentation is still evident. With the appropriate engagement of policy-makers, and leverage of the opportunities for researchers, it is possible to now build a technical and human infrastructure to support Open Science through rewards and incentives, powered by attribution and citation.

This conversation will briefly describe the relevant results from ODE in terms of incentives, drivers and barriers, and the gaps and roadmaps identified by ODIN, while giving some concrete example of successful emerging applications of identifiers for contributors and their scientific artefacts.

1: www.ode-project.eu

2: www.odin-project.eu

Short contribution:

Dr Veerle Van den Eynden, UK Data Archive, University of Essex:

Incentives for sharing research data: evidence from five European case studies

A recent study, commissioned by Knowledge Exchange, gathered evidence, examples and opinions on current and future incentives for research data sharing from the researchers’ point of view. This evidence is used to provide recommendations for policy and practice development on how best to incentivize further data sharing. Whilst most researchers appreciate the benefits of sharing research data, on an individual basis they may be reluctant to share their own data. The study is based on interviews with 22 selected researchers taken from five research teams with established data sharing practices, in Finland, Denmark, Germany, United Kingdom and the Netherlands. The case studies span various academic disciplines: arts and humanities, social sciences, biomedicine, chemistry and biology. The study found that different forms of data sharing can be distinguished, ranging from controlled sharing within a research group, consortium or peer network, to public sharing. Main motivations for researchers to share their research data are (1) when direct research or career benefits drive the need for data sharing; (2) the norms that researchers are exposed to within their research circle or discipline; and (3) a framework of external drivers such as funder and publisher expectations, policies, infrastructure, and data services. Recommendations are made for research funders, journal publishers, learned societies and research managers.

Further details: <http://www.data-archive.ac.uk/about/projects/incentive>