Fourth RDA Europe Science Workshop:

The opportunities and challenges of open data for research

London, UK, 25-26 April, 2017

Brian Matthews and Juan Bicarregui

RDA Europe receives funding from the European Union Horizon 2020 Programme (H2020/2014-2020) under grant agreement n° 653194.

ABSTRACT

This report summarizes the outcomes of the RDA Europe 2017 Science Workshop, which was organised by the RDA Europe team in collaboration with the Wellcome Trust at the Wellcome Trust Headquarters in London, UK, 25-26 April 2017. The participants were selected from a range of disciplines, with a mix of experts involved in data sharing, scientific practitioners who were actively engaged with collecting and exploiting data, and researchers who were studying the field of research data sharing from the perspectives of their own particular disciplines.

The Workshop was organised as a series of presentations and discussions around three themes: (i) how do researchers share and work with other people's data within their discipline and across disciplines; (ii) what are the incentives and rewards to researchers for sharing their data; and (iii) what infrastructure tools and services are of particular use to researchers in their disciplines. These themes were inspired by the emerging European Open Science Cloud programme, which is drawing together research infrastructures and e-infrastructures across Europe. There were three sessions. Each session was organised around one of the above topics, seeded with a number of participants' statements, and included much open discussion.

This report describes the content of the workshop and summarises the views and suggestions expressed during the Workshop which support the global RDA strategy and reinforce the importance of many of the RDA activities. It also records some specific recommendations that were proposed for consideration by RDA Global and RDA Europe.

TABLE OF CONTENTS

1 Introduction

This report summarizes the outcomes of the RDA Europe 2017 Science Workshop, which was organised by the RDA Europe team at the Science and Technology Research Council (STFC), in collaboration with the Wellcome Trust. The workshop took place at the Wellcome Trust Headquarters in London, UK, 25-26 April 2017.

The successive RDA Europe projects have been organising yearly Science Workshops since 2014. The activity was launched by the RDA Europe 2 project, and taken up by RDA Europe 3 when it started in September 2015. The aim of these workshops is to gather input from high level scientists to influence the priorities and the directions taken by the RDA. These meetings are organised by invitation, with a relatively small number of participants to enable a high level of interaction.

The Science Workshops are co-organised between the RDA Europe projects and a research organisation. The 2014, 2015 and 2016 Science Workshops were co-organised with the Max Planck Gesellschaft (Germany), CERN, and CNRS (France) respectively, and held in Munich, Geneva and Paris. A call was opened during spring 2016 by the RDA Europe 3 project to identify the co-organising partner of the 2016 Workshop and the response by the *Science and Technology Facilities Council* (UK) was selected.

STFC proposed to hold the meeting in London at the Headquarters of the Wellcome Trust, one of the world's leading charities supporting medical research, which has been long standing advocate of open science, and an active member of the RDA. The Wellcome Trust offered the use of their excellent conference facilities at their headquarters, conveniently located in Central London; the organisers would like to express their thanks to the Wellcome Trust for their generous support to the workshop.

The lead organisers and contact persons for the Workshop were Dr. Brian Matthews and Dr. Juan Bicarregui for STFC and for RDA Europe, and Dr. David Carr for Wellcome Trust. Mrs. Jean Pearce from STFC assisted with organisational activities. The RDA-UK team at Jisc also supported the workshop, and was represented by Ms. Rachel Bruce and Mr. Christopher Brown. The meeting was chaired by Brian Matthews. Notes on the meeting were taken by Juan Bicarregui, Jean Pearce, Brian Matthews and Christopher Brown.

Section 2 of this report, summarises the theme, participation, preparatory activities and agenda of the meeting. An account of the findings are discussed in Section 3. The recommendations for RDA and RDA Europe are summarized in Section 4. The list of participants, agenda, abstracts from participants, acronyms used, and photographs taken during the meeting can be found in Annex I, II, III, IV and V respectively.

2 Meeting organisation

2.1 Participants

The participants were selected from across a range of disciplines, with a mix of experts, representing different perspectives on issues of data sharing. Participants had one or more of the following perspectives:

- Practitioners actively involved in the data infrastructure working in data repositories, data management, data sharing, and data reuse.

- Scientists who are actively engaged with collecting and exploiting data to further their research programme.
- Researchers who are studying the field of research data sharing from the perspectives of their own particular disciplines.

The aim was to allow a good representation of the local organisations and national strategic topics, to obtain a good balance of the disciplinary fields covered, and to involve participants from a range of European countries, although in practise, scientists from the UK was strongly represented for the sake of convenience[1]. The RDA Europe project partners and participants in the previous Workshops were polled for names of possible participants, as well as STFC, Wellcome Trust and Jisc. A first list was compiled, and potential participants were contacted in January 2017. Several iterations of the list were undertaken to establish the final list of 13 scientists. STFC, Jisc, Wellcome Trust and RDA Europe 3 staff also participated in the workshop.

The STFC brought a focus on physical science, large-scale facilities and materials, while the Wellcome Trust brought a complementary emphasis on medical and biological sciences. Other disciplines and topics represented included fusion energy, business studies, materials, linguistics, crystallography and chemistry, social science (including gender studies and media), agriculture, analytical facilities, and archaeology. Despite efforts to contact candidate participants, there was no representative from the environment and earth science; it was discovered late in the preparation process that the meeting date coincided with the European Geosciences Union General Assembly 2017 in Vienna, and most suitable candidate participants were engaged at that event. Participants work in France, Sweden, Spain and UK. As already mentioned, the UK was strongly represented at the meeting, nevertheless there was a strong international perspective: for example the nuclear fusion community has a strong emphasis on the international ITER collaboration to build an experimental fusion reactor, and the representative of the EBI is working in a cross-European organisation. Some of the participants had been involved in the RDA before the Workshop, but others had not and discovered RDA at this event.

The participation of the representatives from social science and gender studies, and from a business school gave interestingly different perspectives on the issues involved. In both cases, rather than considering the needs of their own community for data sharing, they were using the communities involved in research data management and sharing as the subjects of their research. Thus they brought the viewpoint of studying respectively the sociology of data management practice, and the economic and business environment of research data sharing, both of which provided useful professional input into the discussion.

A list of participants in the workshop can be found in Annex I, and some photographs of the event in Annex V.

2.2 Workshop Theme

In the two first Workshops, a list of questions or possible topics for discussion had been proposed to the participants before the meeting. During the 2016 Workshop, participants were asked to prepare to prepare a 15-minutes presentation describing her/his point of view on scientific data with minimal guidelines and the wide general topic of *Big and Smaller Data*, for a wide-ranging

[1] In several cases, when a non-UK candidate participant was approached, if they could not attend they recommended a UK based collaborator; the nature of research means that the work described was international.

brainstorm. For the 2017 workshop, it was decided again to frame discussion by providing a number of issues related to a theme.

The theme and issues of this workshop were inspired by the emerging European Open Science Cloud (EOSC), which is drawing together research infrastructures and e-infrastructures across Europe. This initiative is developing a framework for cross-European infrastructure for Open Data Science, in recognition that there are key barriers in current infrastructure and research practice[2]. In particular, there is:

- a lack of interoperability in the data and services, within and across disciplines;
- a culture of disincentives to making data available and shared within research practice;
- incompatibilities in the tools and services provided by research infrastructures.

The EOSC is attempting to provide a cross-disciplinary framework to surmount these barriers, and the RDA anticipates playing a significant role in this by providing a forum for setting and sharing best practice and standards for data interoperability and sharing. As a consequence, the Workshop was organised as a series of presentations and discussions around the theme of:

The opportunities and challenges of open data for research.

Data from public funded research is not always made open and it is not always usable when it is. In the workshop, we wanted to explore issues which may affect how and why researchers share and use data, taking a research practitioner's point of view. In particular, the workshop proposed to consider three specific issues.

1. How do researchers share and work with other people's data within their discipline and across disciplines?
2. What are the incentives and rewards to researchers for sharing their data?
3. What infrastructure tools and services are of particular use to researchers in their disciplines?

These issues correspond to the three barriers to open science identified in the EOSC programme above. A session was organised around each issue, seeded with a number of participants' presentations whilst leaving room for much open discussion. Each participant was invited to submit a short abstract in advance of the workshop, the abstracts were reviewed and a plan for the session proposed where participants gave 15 minute presentations, within the limitations of the number of talks which could be scheduled in each session and any restrictions the participant might have in attending the workshop.

Some documents were suggested as preparatory reading to participants to indicate the context of the workshop and frame discussion, as follows:

- European Cloud Initiative - Building a competitive data and knowledge economy in Europe http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266
- Wellcome Trust reviews on Open Research https://wellcome.ac.uk/what-we-do/our-work/open-research

[2] European Cloud Initiatives - Building a competitive data and knowledge economy in Europe http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266

Links to the reports on previous RDA Science Workshops were also provided.

The abstracts provided by participants in advance of the workshop are given in Annex III.

2.3 Workshop agenda

The meeting was organised over two days from midday to midday. Participants who did not know each other before the meeting were also able to discuss informally over two lunches and dinner, and this rapidly built a sense of belonging to a group engaged in a common endeavour, with discussion over dinner (provided at the Wellcome Trust) being particularly friendly and stimulating. The Wellcome Trust and STFC covered the costs of the venue and the meals of all participants, RDA Europe 3 the travel and hotel costs of the invited participants.

Each day of the workshop started by a keynote presentation from a representative of one of the two main organisers, STFC and Wellcome Trust, which set the scene for the subsequent discussion. Keynotes talks were given by Prof. Robert McGreevy, Director of the ISIS Neutron and Muon Spallation Source, within the STFC, and by Dr. David Carr, Programme Manager – Open Research at the Wellcome Trust.

Professor McGreevy gave a talk on the changing way that science is undertaken at analytical facilities, with a radical change from emphasis on hardware and instrumentation, to data and software. His message was "Currently there are more cost effective efficiency/capability gains to be made in scientific research through improved software and data management than through improved hardware techniques and instrumentation." However, funding decisions do not always reflect this.

Dr. Carr introduced the Wellcome Trust, emphasising its advocacy of open access & data sharing to maximise health & societal benefit. He went on to describe the work of the Trust in developing open science policy, and overcoming barriers to open science. The Trust has recently expanded on this, taking on extensive work to map the landscape and take forward pilots in Open Research. It aims to scope: how to embed a culture that incentivises openness and work towards best practices in different fields; how to develop infrastructure needed to ensure outputs are findable, accessible, interoperable and re-useable (FAIR); and how the Trust can spark disruptive innovation in the way research outputs are communicated, shared and re-used. Additionally, Dr. Francoise Genova gave an introductory talk on the RDA.

Sessions on the first two themes took place on the first day, with the third session on the second morning. The discussion was very lively and rich, and took place largely during the course of the talks; STFC and Jisc had note-takers as reporters, and Juan Bicarregui gave a summary of main points of discussion at the end of each day. The main topics addressed during the meeting are summarized in the next section. A detailed agenda can be found in Annexe II.

3 Main discussion points identified during the meeting

The participants came from different backgrounds, but shared a common appreciation and understanding of the value of research data and the benefit to research which the wider use of data would bring.

Below, we give a summary of the main points of the meeting, arranged to reflect the three initial questions posed as issues to be considered. Note that while the three sessions were arranged loosely around these themes, the presentations and discussion which followed did not rigidly follow this separation, but ranged wide over the whole theme of the meeting. In the summary below, we have organised the points raised during the meeting into the three themes.

3.1 How do researchers share and work with other people's data within their discipline and across disciplines?

Uptake in sharing of FAIR, open data.

It is clear that the uptake, use and acceptance of data sharing is still widely variable. Crystallography and bio-informatics are strong exemplars – for example it was commented that a culture of data sharing exists in Biology, where "data sharing is now accepted as the norm", and the size and funding of the bio-medical research field has enabled an efficient ecosystem of data repositories to emerge at scale. Several presentations in this field described efforts to develop FAIR data principles showing how this principle is gaining traction in this area, with tools, standards, metadata standards and support being developed to further promote this point of view in the community.

Crystallography has a well-defined and organised process with a number of important central data repositories (e.g. PDB, CCDC) whose use is to some extent driven by publishers (e.g. IUCr). This area is also leading on linking through from the source experimental data to the final analysed results and publication, resulting in a transparent, reviewable process available for inspection and re-analysis. The issue in this area is encouraging the take up of these capabilities in the community. Archaeology in the UK also has a well-establish system of data deposit, with the Archaeology Data Service forming a centre of expertise over a long period of time; for work across borders, this needs to be replicated across different countries, and the ADS is collaborating on efforts to disseminate expertise and capability across Europe.

Some other fields are less fully developed. Materials science and chemistry, fusion science and agriculture have a mixed situation when it comes to data sharing, and need further support to make data FAIR; data is often restricted to a small set of experts, and it is debateable whether making it more widely reusable is worthwhile.

Some fields are inherently cross-disciplinary. Agriculture is an exemplar of a domain where there is a powerful need for interoperability and cross-domain working, with disciplines from geology, meteorology, chemistry, genetics, economics, sociology, medicine and others all being relevant to studies in the area. Data sources are distributed widely across labs and in farms, with different public and commercial ownership and hugely various data which is often poorly documented giving access and discovery problems. Practices in data sharing are at different levels of maturity, and there is a need for a common perspective and infrastructure to bring these together to gain the benefit of the new data science.

Analytical facilities are also multi-disciplinary, although with a different emphasis, as they interact and provide services to a wide number of disciplines. This means that they need to be compatible at some level with all of them. However, as they can provide central services with relatively stable resources, they have the opportunity to provide a focus for practical data sharing and influencing community practise in areas such as molecular engineering and nano-technology.

Metadata standards and semantics

A point closely related to the uptake of open data is establishment of common open standards for data. Again, bio-medicine has an advanced body of expertise in this area, while others such as fusion and agriculture have mixed provision. There is a need for unified data models and universal data APIs, supplemented by suitable metadata standards for sharing and reuse. These data standards need to support trust, experimental context, and provenance. These may take many years of community engagement to develop as it is difficult to gain consensus even in relatively narrow domains.

Even in areas which have established infrastructure and practices for data sharing, there are still many issues to be resolved. It was noted that there are many standards for metadata – especially in areas such as biological and medical science. As a consequence, new standards may not be the solution, but ways of navigating around existing standards and understanding their connections and cross-points may be preferable. This is also reflected in other areas, such as agriculture where there is a need for a common semantics to share meaning between the widely various disciplines, data sets and organisations involved.

3.2 What are the incentives and rewards to researchers for sharing their data?

Perceptions of data infrastructure

There were a number of issues raised around persuading funders to take seriously providing infrastructure to support the collection, sharing and analysis of data. It was noted that this was sometimes seen as "boring". This exemplified by popular science accounts of "big science" discoveries, in high-energy physics or analytic facilities for example, which barely mention the contribution of the computing infrastructure or the teams of people which manage data, even though computing resources required to make discoveries are considerable and vital to the process. There is an image problem here; a need to change public perception. Even in big science, such as large-scale analytic facilities and nuclear fusion, investment in innovative equipment and techniques is often not matched with investment in the data infrastructure to support them.

The need for data infrastructure is better recognised in bioinformatics, with organisations such as the EBI and the Wellcome Trust providing strong support, but even here funders need to recognise the need to maintain funding for data repositories and other infrastructure. However, recognition of the importance of data infrastructure needs to be across the whole of research. This also has a significant impact in areas which are highly inter-disciplinary, such as in agricultural science which require significant investment to support the connection of researchers across disciplines.

Business models for Data Infrastructure

Funders and public policy makers will respond better to the needs of the community if they can see the impact, to science and to the wider society, of their investment. Open access data

resources are seen to be cost effective and allow re-use of data, and institutional data management realizes economies of scale.

However it is up to the research community to provide the evidence of the value of data, both to establish the case for providing data infrastructure and to make data sustainable for the long term. More than one speaker mentioned the need to provide convincing business cases. The participation of a representative of a business school was instructive as it gave a structured and systematic approach to describing the value proposition and value chain of data infrastructure. This work highlighted the separation of the sources of the costs and benefits of the use of data, and provided a foundation for good business cases for sustainable data infrastructure. It was noted that while the costs of data management and curation are well studied, the question of how value is created through data sharing is much less well understood. Traditional business model depend on gaining a return from tangible assets, but data assets and the added value gained from their good management are difficult to assess as the benefits are quite diffuse, long term and hard to identify in particular cases. There is also value in the network of collaboration itself, the social connections formed by data sharing and interchange, but these are equally hard to quantify.

Recognition and reward

A number of presenters also raised the issue of recognition and reward. This is a well-recognised problem. Data publication is becoming gradually more wide-spread, being driven by publishers requiring access to data behind publications, and funders requiring data management plans and data deposit. This is being extended across the science workflow to source data. However, it is questionable whether the data is normally assessed during the article review process. More expertise is needed for this, as well as ways of making data more transparent and open.

A question was raised of whether data publication a separate thing from the paper publication. The wider benefit may not be what researcher is funded to do or rewarded for, and opening out the data for scrutiny does not always get a positive response if the data is criticised. We need to make sure that researchers take a good attitude**;** if others improve our analysis then we improve theirs and we can share credit.

Formulating data infrastructure tools such as data standards, metadata formats, and ontologies is a hard and complex task, and there are not necessarily rewards for this effort which might only pay dividends in the long term. Disciplines do not recognise this as a contribution to the discipline, and it can in the worst case damage careers. There is a need to recognise standardisation as a research activity in its own right, and allow contributors to gain academic credit from participating in such activities.

Sociology of the data management teams

The participation of a representative from social science and gender studies led to the consideration from a sociological perspective of how the structures and cultures of data-oriented science affect how research is practiced, funded and organized. Science and technology are not neutral or natural processes, they are designed, implemented and used by people within specific contexts.

There has been little consideration of how the way data infrastructure is provided and managed might affect the knowledge that is produced. For example, competing scientific paradigms, political agendas, budgetary constraints, organisational infrastructure, reward structures for different parts of the scientific process, commercial influence, and personal bias can all influence the choice of what research is undertaken and how it is presented. The type of data

infrastructure organisation can also influence the balance between collaboration and competition in research.

Providers of data infrastructure try to balance their understanding of the needs of an "assumed user" and the material limitations of the technology itself. However, their assumptions of the user perspective may not be correct, which may lead to poorly designed and under used solutions.

To generalise on Jeremy Frey's presentation, Research is a Social Machine, a technology-enabled social system, seen as computational entities governed by social processes that enable them to solve complex social and computational problems in a decentralized fashion, at large scale. The combination of people, instruments and data, now operating in a social media oriented environment, offer new models for interaction and collaboration which may be unpredictable. These social issues are an open area which need further consideration.

Reluctance to share data

It was commented that although funding agencies and other organisations have made numerous policy recommendations towards sharing of data for more than 30 years, and despite data management now being a condition of grants, empirical studies show that only a small percentage of researchers do routinely share data. Although researchers often say that they are willing to share data something is happening in practice which is stopping data being shared.

The need to protect private and sensitive data, especially in the case of medical and sociological disciplines, can lead to a conflict between privacy and openness. However, it was pointed out that subjects were often happy and willing for data to be made available for research as long it is anonymised with suitable safeguards, and well established procedures exist to enable this.

It was suggested that a closer look at user communities would help. Different user communities have very different perspectives, incentives and norms. A user-centric perspective could promise better adoption of technologies and techniques for data sharing.

Open Science Policy and Data Management Planning

It was noted that when a leading institution in a field, establishes a strong policy and support for open science and data publication, it can have galvanizing effect on the whole discipline. The examples of the Wellcome Trust in medical research and European Synchrotron Radiation Facility (ESRF) in the analytical facilities were given as examples of changes which have affected whole communities. It was felt that this could be reflected in other areas. An institute such ITER could take a leading role in the fusion community for example. Such institutes need to engage with developing e-infrastructures to develop their policy towards data and support for data management and distribution.

Data management plans can drive the move to open science and by making them more active and actionable, monitored and assessed, DMPs, can form the basis for incentives to researchers. Furthermore, data management plans should cover not only data and publications but a wide range of research outputs such as software, workflows and methodologies.

Training and Expertise

Technology for sharing of Big Data and for high throughput and automation using advanced virtual platform systems is not mature. There is a shortage of knowledge and expertise in these areas such as Docker and OpenStack, and advanced analytics platforms, such as Apache Spark and Storm. The notion of a "Research Software Engineer", which is being developed in the UK, was highlighted. Research Software Engineer is a professional career path **c**ombining a detailed

understanding of research with expertise in programming and software engineering. This career path needs to be more widely recognized. Further, there should be an increased engagement with the whole area of data science in the training of doctoral students across disciplines.

3.3 What infrastructure tools and services are of particular use to researchers in their disciplines?

Common integrated Infrastructure

Some contributors recognised the need for a common basis for integrated data and compute infrastructure. Areas such as agriculture research are hampered by "silos" of expertise which are incompatible but have to interact with each other and also with many other data infrastructures, such as Elixir. It would be advantageous to have a common conceptual framework to describe e-infrastructures. Such a framework would describe, for example:

- The technical layers, and common services within the infrastructure, with their functionality and interfaces.
- Data interoperability rules around FAIR principles and exchangeable metadata and data standards.
- A set of principles of engagement.

Work in this direction is being undertaken by in the context of the European Open Science Cloud.

Cloud computing and access to large-scale compute resources

A number of participants highlighted the need to access large-scale compute resources, including mass data storage and distribution, high-performance computing and high-performance data analytics. This access can be distributed across a cloud platform. The large data volumes generated in modern research require large repositories, possibly using community distributed storage and a shared cloud infrastructure, to support large scale compute capability. Workflows can lead to compute requirements that come in bursts, so even if local resources are significant, there can be a need to burst out capability, ideally shared with across communities who each have "bursty" workflow

Unlocking data via text mining

It was recognised that there is much data which could be more openly and readably available, but is locked up in the legacy of literature, where it is only partially accessible and hard to reuse. There is a need to aggregate a lot of this information into databases and collections which can be reused more widely. The processes in crystallography where deposit of data in established data bases is expected common practise, are a good exemplar. But there are many other areas where this is not the case. For example, molecular engineering considers many properties of materials which are less widely studied than crystal structures and thus there are few common databases of properties available, with properties scattered across the literature instead. Text mining of the literature in this field was presented as a way of generating and populating new databases of properties, and modern text processing and analytic techniques can lead to the successful construction of new open data resources. Although this may need the cooperation of the publishers to access and mine the literature, once such databases exist, there can be a change in community practice so that entry in the databases is expected in parallel with the

production of the literature. Again, there is a need for sustainable business models to support such databases.

4 Recommendations to the RDA

As a consequence of the extensive discussion at the workshop, the following recommendations were proposed as actions or activities that could be undertaken by RDA to help address the issues raised.

1. **Consensus and Standards**:

- Provide a forum for the EOSC to formulate common standards and best practice for interoperability to underpin an open research commons.

- Provide a forum for the elaboration and promotion of the concept of FAIR data

2. **Business models**:

- Engage directly with funders and policy makers to determine their needs and expectations and to design sustainable models for data infrastructures.

- Encourage activities to explore the value chains of the research data ecosystem, taking special note of the benefits and gains of data sharing. This would need to involve different communities as the value proposition of different communities can vary wildly.

- Although provisioning cloud computing and access to large scale compute services is out of scope for RDA, RDA can provide a community forum to discuss issues on the requirements for large-scale data infrastructure to most effectively support data sharing, at the levels needed for different communities.

3. **Recognition**:

- Provide a forum for discussion on the social and organisational drivers to data-driven research, with an emphasis on how the data infrastructure can influence the research process.

- Develop mechanisms to allow formal recognition and reward for contributors to data infrastructure standards and tools, including within the context of the RDA itself.

4. **Communities**:

- Assist specific communities and disciplines to propagate the use of data infrastructure via training and dissemination of best practice.

- Advocacy at the user level within particular communities, taking into account community perspectives, incentives and norms. Perhaps the use of champions in particular domains will encourage the change of culture required to promote data sharing.

- Seek to influence and advise leading institutes and programmes in science and research, such as ITER, or large scale facilities, so adequate regard and provision on data and data infrastructure can be catered for at an early stage, and appropriate data policies and support developed. These institutes can often steer a whole community and thereby change attitudes to open science.

Annex I: List of Participants

Name	Institute	Country
Invited Scientists		
Rob Akers	Culham Centre for Fusion Energy	UK
Alvaro Arenas	Department of Information Systems and Technology, Instituto de Empresa Business School	Spain
Ennio Capria	European Synchrotron Radiation Facility	France
David Carr	Wellcome Trust	UK
Jean-Baptiste Cazier	Director of the Centre for Computational Biology, University of Birmingham	UK
Jacqui Cole	University of Cambridge and STFC Rutherford Appleton Laboratory	UK
Chuck Cook	EMBL-European Bioinformatics Institute	UK
Jeremy Frey	Department of Chemistry, University of Southampton	UK
Holly Wright[3]	Archaeology Data Service, University of York	UK
Katherine Harrison	Department of Gender Studies, Lund University, Sweden & Department of Media, Cognition and Communication, Copenhagen University, Denmark	Sweden
John Helliwell	Department of Chemistry, University of Manchester	UK
Odile Hologne	Institut National de la Recherche Agronomique	France
Robert McGreevy	STFC	UK
Susanna Sansone	University of Oxford and ELIXIR-UK	UK
STFC		
Juan Bicarregui	STFC	UK

[3] Holly Wright substituted for Katie Green at short notice

Brian Matthews	STFC	UK
Jean Pearce	STFC	UK
RDA		
Françoise Genova	CNRS – Université de Strasbourg UMR 7550 Observatoire Astronomique de Strasbourg	France
Leif Laaksonen	Finnish IT Center for Science, WP2 Lead RDA Europe 3	Finland
Chris Brown	Jisc	UK
Rachel Bruce	Jisc	UK

Annex II: Agenda of the meeting

4th RDA Europe Science Workshop

25-26 April, 2017

Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE

Outline Programme

Day 1

12:00 Lunch available at the Wellcome Trust

12:45 Welcome and introductions

13:00 Introduction to RDA - Françoise Genova

13:30 Keynote talk 1: Robert McGreevy, STFC

14:00 Theme 1 : How do researchers share and work with other people's data within their discipline and across disciplines?

- Chuck Cook - Enabling open data in the life sciences
- John Helliwell - Experience of data sharing within Crystallographic research
- Holly Wright - The Archaeology Data Service: 20 years of data sharing
- Susanna-Assunta Sansone - Going FAIR: highlights from the life sciences

15:15 Discussion on theme 1

15:45 Break

16:15 Theme 2 : What are the incentives and rewards to researchers for sharing their data?

- Katherine Harrison - Big Data in Big Science: a social sciences perspective
- Alvaro Arenas - A Value-Based Approach to Unveiling the Business Model Dimensions of Data Sharing

- Jeremy Frey - Chemistry as a Social Machine

17:30 Discussion on theme 2

18:00 General discussion on topics arising and collection of key points, issues, and recommendations

18:15 Close

20:00 Dinner - at the Wellcome Trust

Day 2

9:00 Recap on day 1.

9:05 Keynote Talk 2 : David Carr, Wellcome Trust

9:35 Theme 3 : What infrastructure tools and services are of particular use to researchers in their disciplines?

- Ennio Capria - Experiment is just the beginning: the importance of data management and sharing in the experience of the NFFA consortium
- Jacqui Cole - Chemical Database Auto-Generation Tools for Large-Scale Data-Mining
- Odile Hologne - E-infrastructure tools and services for open agricultural sciences : first approach
- Rob Akers - Towards the ITER era of Big Data and Extreme Scale Computing
- Jean-Baptiste Cazier - Bioinformatics Vagaries and its Challenges

11:00 Break

11:30 Discussion on theme 3

12:00 General discussion on topics arising and collection of key points, issues, and recommendations

12:30 Close, thanks and goodbyes

Lunch available at the Wellcome Trust

Annex III: Abstracts of talks for the meeting

Chuck Cook - Enabling open data in the life sciences

Biologists have long recognized the value in sharing data, with the first open access databases created 40 years ago. Molecular data, such as DNA sequences, have more value in aggregate than as individual experimental results, and data sharing is now an accepted part of the research culture for molecular data in the biological sciences. Technological advances have resulted in such large data volumes that centralized repositories are the only viable means of managing and keeping track of the data flows. The current challenges lie in searching the public repositories for useful data and in analysing very large data sets. EMBL-EBI manages many large public databases and provides analytical resources as well, but ever-increasing data volumes, and data types, mean that long term management will require distributed systems, such as ELIXIR, Europe's distributed infrastructure for life science information.

John Helliwell - Experience of data sharing within Crystallographic research

I work as a research crystallographer of biological macromolecules. Over my forty years career the Protein Data Bank has grown from its first depositions in 1971 to more than 120,000 today of which ~90% are from X-ray crystallography. As well as a depositor myself (of over 100 protein crystal structures) I very frequently use the depositions of others and always calculate the electron density map of the accompanying processed diffraction data to the atomic coordinates. Why do I routinely calculate the map? Understanding other researchers' crystallographic results should be through one's own eyes and thereby include the opportunity to work with a publication's underpinning data so as to check the various decisions made in the analyses by the original authors. As well as wrong steps that might be taken there are at present no community agreed rules on interpreting such maps although this situation has improved greatly as the Protein Data Bank has introduced a Validation Report on each of its structures. Decisions by authors start from the raw data and these raw data are currently not visible to users but today preserving the much larger raw diffraction data sets is technically and organisationally viable at a growing number of data archives, both centralised and distributed. These archives are empowered to register data sets, raw or otherwise, and obtain their preservation descriptor namely a 'digital object identifier'. This possibility of archiving of raw diffraction images is a quite remarkable revolution in my science as it is now a widespread opportunity. For archiving my raw diffraction images I utilise both the University of Manchester data library and also, most recently, the EU's Zenodo. A vital aspect of ensuring quality archives of data should be to ensure that referees are allowed access to the data underpinning a submitted article; this is common practice in chemical crystallography at least with IUCr journals (Acta Crystallographica C for example, which is thereby exemplary in its refereeing protocols). Also data skills for referees is an important topic for researcher education and should include training in implementing validation assessment procedures. I should finally mention that in terms of other research data in a publication I rely on those being checked by the relevant specialists. Choice of referees across all the underpinning techniques within an article by an editor of a journal is therefore important.

Holly Wright - The Archaeology Data Service: 20 years of data sharing

Founded in 1996, the Archaeology Data Service (ADS) is a discipline-specific digital archive. After twenty years, the ADS curates over 320,000 digital objects comprising over 2 million files and amounting to over 10Tb of data. As the amount of data held continues to increase so too does its re-use potential. This paper will present a brief synopsis of the work of the ADS and explore the challenges and opposition to data sharing experienced over the last 20 year. This will review both the 'carrot' and the 'stick' approaches to facilitating data sharing and will highlight the opportunities for enhancing data sharing and re-use that the ADS provides.

Susanna-Assunta Sansone - Going FAIR: highlights from the life sciences

Successful share and reuse of data in the life sciences is still significantly laborious and it gets harder when done across-disciplines. For example, the integration of clinical and basic research data requires manual intervention to match up and identify all digital entities of interest, such as molecules, compounds, cells, observations, drugs etc. To ensure all digital research assets are Findable, Accessible, Interoperable and Reusable, according to the FAIR principles, work is in progress within large life sciences programmes, infrastructures and public-private-partnerships - such as the NIH Big Data to Knowledge initiative (https://datascience.nih.gov/bd2k), ELIXIR (https://www.elixir-europe.org) and Innovative Medicine Initiative (https://www.imi.europa.eu). One common thread is the need of mapping the landscape of interoperability standards, to understand their maturity and usability in enabling FAIR data; in the life sciences, for example, there are over one thousand domain-specific standards (source: BioSharing portal, https://biosharing.org).

Reference: Sansone, S-A, Rocca-Serra, P (2016): Review: Interoperability standards. Wellcome Trust. https://doi.org/10.6084/m9.figshare.4055496.v1

Katherine Harrison - Big Data in Big Science: a social sciences perspective

Too often in the histories of Big Science the role of "computing" is framed as a support function or a side project to the experiments. My current research project sets out to challenge this perception by exploring ways in which data management plays a key role in knowledge production at Big Science facilities such as the European Spallation Source. Approaching the topic from a social sciences perspective, I am interested in the following questions: How do software and hardware shape the collection, processing and distribution of the vast amounts of data generated in Big Science experiments? And what effect does this have on the knowledge produced? How do design and use of data management software affect the creation of national and international collaborations between researchers, institutions and industry?

Alvaro Arenas - A Value-Based Approach to Unveiling the Business Model Dimensions of Data Sharing

This talk presents on-going work analysing business models for data sharing. We have followed a qualitative-research approach to uncover the dynamics of value creation in data sharing. Using a case study method, we examined the operating model of three organizations covering data-sharing practices in three academic disciplines: material sciences, represented by the Cambridge Crystallographic Data Centre; environmental studies, represented by the British Atmospheric Data Centre; and health-population studies, represented by the Medical Research Council Unit for Lifelong Health and Ageing. We conducted semi-structured interviews with managers from the three sites, and review documents about the technical and managerial practices in order to determine main characteristics of their business models. In addition, we applied the e3-value modelling methodology, a methodology widely used for analysing business models using a value-based lens, to tease out the value flows within each site. The findings demonstrate the importance of the value network dimension of a business model, as data sharing relies on a set of actors creating and getting value in the process, and the significance of intangible assets.

Jeremy Frey - Chemistry as a Social Machine

Ennio Capria - Experiment is just the beginning: the importance of data management and sharing in the experience of the NFFA consortium

NFFA-EUROPE sets out a platform to carry out comprehensive projects for multidisciplinary research at the nanoscale extending from synthesis to nano-characterisation to theory and numerical simulation.

Advanced infrastructures specialised on growth, nano-lithography, nano-characterisation, theory and fine-analysis with Synchrotron, FEL and Neutron radiation sources are integrated for a coordinated access to develop frontier research on methods for reproducible nanoscience research and to enable European and international researchers from diverse disciplines to carry out advanced proposals impacting science and innovation.

Moreover, in the context of the Joint Research Activities (JRAs) the own research activity of NFFA-EUROPE addresses key bottlenecks of nanoscience research: nanostructure traceability, protocol reproducibility, in-operando nano-manipulation and analysis and open data. In particular, with respect to the latter, a data management dedicated programme is contributing to creating the "Information and Data management Repository Platform" (IDRP). This is a platform-wide data model to allow efficient recording and sharing of results as well as a strong contribution to building a successful open data policy for European technological platforms.

Jacqui Cole - Chemical Database Auto-Generation Tools for Large-Scale Data-Mining

Large-scale data-mining workflows are increasingly able to predict successfully new chemicals that possess a targeted functionality. The success of such materials discovery approaches is nonetheless contingent upon having the right database source to mine. This presentation shows how to tailor-make databases to search for functional materials to meet the needs of a given device application.

Odile Hologne - E-infrastructure tools and services for open agricultural sciences : first approach

To tackle the societal challenges linked to food, health and environment, Agricultural sciences are facing important issues linked to integration of data from different disciplines, at different scales and geo-localisation. The landscape of the existing research infrastructures and e-infrastructures or data sources and data services can be seen as silos which prevent the development of cutting edge research in the agricultural domain. Based on 2 H2020 e-infra projects eROSA and Agrinfra+, we will give a first approach of what could be an e-infrastructure for open agricultural sciences.

Rob Akers - Towards the ITER era of Big Data and Extreme Scale Computing

Eurofusion was established in 1999 to preside over the European Fusion Programme. Currently Eurofusion overseas 14 national fusion devices and the Joint European Torus (JET) in Oxfordshire UK, used by over 40 research facilities around Europe and the world. Although over recent decades collaboration has become increasingly prevalent, sharing of data between relevant disciplines (plasma physics, neutronics, materials science, robotics) remains rather limited, involving large amounts of manual intervention. Sharing of data between research centres is even more difficult and even less commonplace, due largely to having no uniform access method and multiple data models spanning the many experiments. Researchers do make use of the many HPC facilities available both nationally and internationally, but data sharing, data re-use, provenance tracking and therefore traceability are often ad hoc and limited. JET is coming to the end of its operational 4 decade long lifetime. 35 nations are now collaborating to build its replacement, arguably the most ambitious, most complex piece of advanced engineering mankind has ever embarked upon – the ITER tokamak. ITER is designed to prove the feasibility of fusion as a large-scale and carbon free source of energy based upon the same principle that powers the Sun and the stars. It represents a step change for the community; it will produce more data in a single day than JET has produced in its 35 year history, truly bringing the community into the 'Big Data' era. To derive useful information from the plethora of heterogeneous objects both from the experiment, and from modelling, will require an increased use of both high throughput and exa-scale class high performance computing facilities

distributed across national borders. Significantly more technical and political coordination than has been the case up until now will be required. Grid/cloud computing, HPC as a Service, scalable, distributed and high availability storage systems will all be necessary in order to properly capitalise on the investment made in this device and the current generation of experiments which can be seen as 'feeder' experiments into future ITER operations. Over the next few years, as ITER data systems are designed and prototyped, an increased engagement with the communities developing next generation cloud e-Infrastructure and High Performance computing/Big Data technologies will be paramount for learning how to tame the thermonuclear plasma and make a 'star on earth' - ten times hotter than the core of the Sun and a source of sustainable, near infinite clean energy for all mankind.

Jean-Baptiste Cazier - Bioinformatics Vagaries and its Challenges

Annex IV: Acronyms used in this report

ADS	Archaeology Data Service
CCDC	Cambridge Crystallographic Data Centre
DMP	Data Management Plan
EBI	European Bioinformatics Institute
EOSC	European Open Science Cloud
ESRF	European Synchrotron Radiation Facility
FAIR	Findable, Accessible, Interoperable, Re-useable – principles for open data.
IUCr	International Union of Crystallography
PDB	Protein Data Bank
RDA	Research Data Alliance
STFC	Science and Technology Facilities Council

Annex V: Photographs taken during the meeting

Figure 1: Participants in the 4th RDA Science Workshop, Wellcome Trust, London, 25-26 April 2017

Figure 2: Dinner at the Wellcome Trust