2nd RDA Europe Science Workshop – Participant Statements

This note contains statements that we received from various participants in a structured form, i.e. we tried to group the statements according to the session contents. We start each session part with repeating our initial questions. These are followed by questions (italics) and statements from participants.

Session 1: Policy Level Aspects

- *Do you think that the take-up of Science WS recommendations by RDA EU is satisfactory? How can it be improved?*
- *Do you think that RDA as it works now is sufficiently owned by researchers?*
- *Are Data Sharing, Re-use and Interoperability issues of relevance for you and if so how?*
- *Are mechanisms to improve trust with increasing data worldwide?*
- *How can personal data be protected, or is that impossible? Do we need to simplify?*
- *To what extent should the lead be undertaken by computer scientists or by users from other disciplines? How can the maximum benefit be obtained by balancing push and pull?*
- *Do you agree with the major statements of the Data Harvest Report (summary see attachment)? If so, how can the recommendations be fully realised?*
- *Do you agree with the major statements of the Data Practices Report (summary see attachment)? If so, how can current practice being changed?*
- *Which are the main obstacles for you and your colleagues when working with research data and are they met in the reports?*
- *Should initiatives such as RDA be core funded by the EC and member states working together?*

- I am very excited about what you guys are doing with the RDA and think how we **manage, share, and reuse** Research Data is one of the biggest challenges within Data Intensive Biology today - and it is one that is currently, to my mind, not being really strategically tackled. I think that there are a large amount of things to potentially discuss.
- There is still little benefit anticipated by most mathematicians in **sharing** research data, which means that they tend not to document properly what their data is. As a consequence it becomes very hard to reuse.
- *Do you think that RDA as it works now is sufficiently owned by researchers?*
  I am afraid I do not think so. I am new to RDA, but it seems to me that it is **largely unknown from researchers**, including researchers interested in data management and sharing. For instance I never heard about RDA before receiving the invitation to come at the 2nd workshop.
- **Data Sharing, Re-use and Interoperability** issues are central to my research activity. I spent the best part of a decade and a half developing tools aimed at sharing and re-using computational models, and in the last half of that, solutions to make those models and experimental data interoperable. In addition, I am now acting chief data officer in a large cancer centre, with clinical and research activity. The circulation and re-use of clinico-biological data is at the centre of my mission.
- *How can personal data be protected, or is that impossible? Do we need to simplify?*
  This is a good question. Personally, I think people worry a bit too much about **protecting** their data. Before the advent of computing, they would not have cared so much about **privacy**. It is now getting too far. Not everyone is interested in other people's data, and I think people gain in general much more by sharing data than protecting it. Nevertheless, it is clear that most people

want to protect their data, and it is useless to fight it. Now, I wonder if **data protection** is actually part of the mission of the RDA. RDA is about data sharing, and more importantly research data sharing. Once a data is deemed of a "research" nature, the privacy is irrelevant. For instance, **patient data** is not research data. Data from a **clinical trial** is research data. But in this case, there was a patient consent for re-using their data.

- *To what extent should the lead be undertaken by computer scientists or by users from other disciplines? How can the maximum benefit be obtained by balancing push and pull?*
  I do not think the lead should be taken by **technicians** of any domain. **Computer scientists** have the expertise to help with implementation. However, many decisions must come from **data generators** and **data users**. RDA should be the "maître d'ouvrage" while computer scientists would be the "maîtres d'oeuvre".
- *Do you agree with the major statements of the Data Harvest Report (summary see attachment)?* If so, how can the recommendations be fully realised? I do agree with the **data harvest report**. I think the recommendations contain the path needed to implement them. But more importantly, we need **political will** and **significant funding**. The rest will follow.
- *Do you agree with the major statements of the Data Practices Report (summary see attachment)? If so, how can current practice being changed?*
  I agree with most (but not all) of them. The different statements are not independent. Some of the problems are consequences of the others. Therefore an identification of the bottleneck is needed, in order to maximise the efficiency of the measures taken to address the problems. At the end, it all boils down to: **No incentive, no tools**.
- *Which are the main obstacles for you and your colleagues when working with research data and are they met in the reports?*
  The **lack of structure and the metadata**. Most of the data is unstructured and cannot be automatically processed. The lack of metadata reduces their usefulness. Both lacks combined **preclude quality control and validation**.
- *Should initiatives such as RDA be core funded by the EC and member states working together?*
  I think that the EU must **support RDA**, but ultimately the conclusions must be endorsed and supported by the national governments because laws have to be passed. And also some data are within the national remit (e.g. medical records)
- Firstly, I would like to specify that after reading the **Data harvest** (and « A surfboard for riding the wave ») I'm very surprised (good surprise) about the conclusion and the statement of this report. Many of my questions can find answers on this document. But now what can we do to convince stakeholders of EU?
- **Questions and statements :**
  - **<u>Describe data</u>**
    How can we motivate researchers to describe by **metadata**? Actually, everybody knows the importance of metadata but nobody describes their data. There are several reasons: no time, no technical support, no human support, no direct interest, etc. In fact "The **Data Harvest**" document is a good argument and an interesting first step to persuade researchers (chapter 3. Why bother?)
  - **<u>Access to data</u>**
    The main problem to access data is getting **metadata** (context & methodology). Researchers don't have time and possibilities to give the free access to their data, not for a question of legal issue but mainly for a **technical problem**.
    For this, we need to have data bank or data center with metadata descriptions and the take care of accessibility of these data. But, in fact, many institutions/stakeholders don't (or can't) give us this possibilities.
- **A data management plan**
  Presently, we don't know how to **preserve and archive** our data. We need to use a common data management plan or, in fact, to be aware of the future of raw data! We publish a part of

data (not raw data but elaborated data) and we don't give the open access to the raw data. But without **metadata**, these raw data are lost for scientific community!

- My view may be parochial, so I'll begin by saying that I'm a linguist working in the Netherlands. **Archiving** data is something we hear a lot about -- the Dutch universities have all initiated data committees and archiving requirements in the wake of **scandals** such as that surrounding Diderich Stapel, who confessed to fabricated data throughout a good part of a twenty-year career. But so far these seem to be motivated by the wish to avoid more scandals, so they emphasize archiving requirements, but not making it **easy to find data**, nor the positive side for science of getting more research done based on the same data. Could RDA see this imbalance as an opportunity?
- *The need for data exchange (and thus the need for proper data management) is yet difficult to convey. The role of funders seems to change: build infrastructures to make data visible and accessible [Tsunematsu]*
  To what extent does this already happen? Are **funders** in general interested in data **sharing** and **sustainable access** to data for all projects? If not, why?
- How can we more effectively work with **stakeholders** including government, publishers, funders, charities and researchers to facilitate and enable research data sharing?
- *Within Biology data is often shared, but rarely reused. Why is this?*
  It probably relates to several reasons:
  1. **Accessibility** of data
     a. For large biological projects with multiple data types, and multiple files, **downloading/accessing** data can be extremely time consuming
     b. Data **searchability** is often poor
     c. Data is often **unlinked** (eg metadata is in a table in a PDF in a paper, sequence data in the European Nucleotide Archive)
  2. **Infrastructure limitations**
     a. Often biological IT **infrastructure** has been built piecemeal and so is not well designed for the task
     b. Often Biologists don't understand the **systems** that they procure, and don't focus on key components
     c. Biologists have, compared to other Data Intensive fields, a generally **lower skill** level in computing than would be ideal (most non-bioinformaticians can't use UNIX, and can't install software on a UNIX machine)
     d. Biological infrastructure is generally **local** (single groups often buy/own single servers for example)
  3. **System incompatibility** / inability to share software
     a. Most papers within biology will include work done on **local systems** with bespoke environments; it is not easy to **replicate** these to reproduce the results
     b. Software is often shared as a **git repository**, with no **incentive** for the author to maintain the code, or ensure it is portable
  4. For many groups there may not be an **incentive to share** data as widely as possible
     a. Some researchers have strong feelings of data **ownership**
     b. Mechanisms for **citing datasets** are currently still developing (therefore there is a potential impact problem)
     c. Researchers may want to exploit their data set for other research questions, which they don't want other people to be able to attempt to answer

- This is compounded by the fact that in my experience some Biologists do not want to **share data**, or believe sharing the minimum amount of data is acceptable. There are a growing number who believe that data should be free, and available to all, and that we should try to make this possible by developing systems and approaches to support this – however, this

view is not universal, and there is a **cynical view** that even if there are resources available, people won't use it.

- *What lessons can we in Biology learn from other fields in this area? How can we build cross-discipline collaborations? How can we develop best practice? And how can we build best practice across Data Intensive Science?*
- *Is it possible to develop views about data sharing and reuse within our students? Providing them with the tools to share data, in order to bring about a generational change in the way in which data sharing is viewed?*
- I suspect that what we need is to begin coalescing a young, active, engaged group of **Data Intensive Biologists**, who can develop the systems and approaches to share data in a way that is relevant to the practice of modern data intensive biological research. I think this is an area where the RDA can help, on several levels. Firstly, to **aggregate researchers** across continents. Secondly to **connect researchers** across disciplines, to spread and develop best practice across data intensive science. Thirdly, to facilitate the creation of resources to **enable data and method sharing**. Fourthly, when systems/approaches become available, to help champion these, and to help researchers to work with other stakeholders such as publishers to establish **standard approaches** for sharing data, based on what we know about data production, data richness and data quantity today.
- *How can RDA and other initiatives help funders take up an interest in and most efficiently build infrastructures for all domains of science (assuming that no domain would like to be behind the others by not having such infrastructures)? To what extent has OECD Global Science Forum dealt with this (beyond specific areas of research)?*
- The balance between generic and domain specific infrastructures for data sharing and sustainable access to data needs to be found.
- *How can we gradually expand, building on success stories?*

Session 2: RDA Results and Impact

- *It took up to 15-20 years between the definition of TCP/IP and its worldwide take-up by scientists, industry and societies. Can this huge time gap be avoided when talking about simplifying data re-use?*
- *Do the 4 concrete results achieved within 18 months mean something to you and do they satisfy your expectations? If not what would you have expected as results? (see attached flyers)*
- *Do the areas of activity meet your expectations or what is missing or not useful? (see attachment for a grouping of activities)*
- *Does the start of the Data Fabric Discussion as an umbrella for discussing optimizations of the scientific data production and consumption machinery make sense to you? (see attached flyer)*
- *Which technology trends do you see and how should they influence RDA activities?*
- *What kind of infrastructure components would you see as so essential that RDA should try harmonization as early as possible to prevent fragmentation?*
- *What kind of training and support actions would you like RDA Europe do to train more young data professionals and/or to foster the work in your discipline?*
- *Would you like to participation in adoption projects which are funded to take up results in your environment?*

- Data Foundation & Terminology **Model Diagram**
  A **metadata** description is usually itself something very **complex**, a mathematical object itself whose description might just be as complicated as the original object itself. The same mathematical object can have several different reasonable instantiations as digital object, which can be resolved only as the different metadata descriptions are worked out to be equivalent, which requires additional logic, heavier than equating ontology types or relationships, and most of the time a computationally hard problem.
- Emphasis on **sharing code and data in reusable** way. Avoid at all costs PhD thesis with locked data and code.
- Two main trends can be observed:
  - one particular strand of the **NoSQL push**: network/graph datasets, which obviously have importance in science (relevant because the "aggregation of digital objects" concept in the model of the Data Foundation Group breaks down a bit). With Neo4j for instance a graph is really several different aggregations of digital objects all at once, the metadata is associated to the whole collection of objects, and a lot of metadata is included as data in the graph itself)
  - **merging of data**: sharing of datasets between industries/companies, reuse of existing datasets (open data) requires careful attention to metadata.
- Another upcoming trend is:
  - more and more careful consideration of the **metadata** in the dataset, for privacy and efficacy reasons. For instance, location data is useful to sell ads, but the more so if it is GPS-level precision than if it is city-wide precision. It is also more of a privacy risk, of course.
- *What we are waiting for RDA?*
  Improve, convince and help the stakeholders of each country to create a **scientific platform of metadata/da**ta with a public administration, on the model of the https://www.data.gouv.fr/fr/ but for scientific data or something like this: http://www.coriolis.eu.org/ (Is it a dream or can we hope a short-term solution?)
- *Which Choice of metadata?*
  There are many existing standards of **metadata**:
  - policies metadata
  - thematic metadata
  - technical metadata
  - and we always surf between these constraints

- *What we are waiting for RDA?*
  A guideline to help the **choice** and also give **equivalence** between all these standards!
- *What we are waiting for RDA?*
  A guideline for **data management plan** should be very interesting but also some concrete solutions for the implementation of this plan (tools, technical solution).
- *Interoperability – how?*
  In my opinion, interoperability is not a problem if we respect 3 conditions:
  - we **describe** our data,
  - we choose the appropriate **metadata format**,
  - we give an **open access and sustainable access** to this data

  But these 3 conditions are clearly difficult to respect!

- While **data types**, and what they describe may be different, fundamentally **data is all the same**; we use the same technologies to store it, and the same hardware to analyse it. There are a lot of people who like to believe that their data is somehow special; however, in reality there are very large areas of **commonality between fields**, often without those fields realising. How can we start to get people to see the similarities, rather than the differences between data and how we analyse and store it?
- With an increasing number of large infrastructure projects developing large scale research storage and compute, can we move towards a situation where data is universally available; an '**eduroam' for data storage and software**? Can we develop systems that enable researchers to **share, and access data** wherever they are?
- And can we develop **universal researcher identifiers**, so that research data is tagged to individual(s) as they move institutions?
- The **cloud and virtualisation technologies** offer enormous promise to enhance the **portability** of software and enable more **effective sharing** of datasets. Imagine a situation where research is performed on a Virtual Machine (VM), and at the point of publication a researcher snapshots that VM, creates a **DOI** to it, and includes this in their paper. This would then enable any other researcher to literally pick up where the paper ends. On this basis, how can we make use of **cloud technologies** to enhance **data and software reuse and sharing**?
- Are attempts to **build on existing tools/systems** (many of which were designed in the 1980's or earlier) detrimental to the long term needs for data reuse and making data available, and do we need, instead, to attempt to develop **wholly new systems/environments/software tools** to enable data sharing? This might be a risky venture, but could support for a number of well-targeted small scale pilot projects that are independent of current systems, all with the potential to scale up, provide us with the tools that we need going forward.
- Are the systems for data sharing we have currently suitable for their task? How would **Google do data sharing**? And what does this tell us about where current systems are good, or poor.

Session 3: Data Science in General

- *Strong institutions seem to have enough resources (human, machines) to undertake data-intensive work, but most researchers do not have a chance to participate? Is that something we can and should change and if so how?*
- *Where are the biggest costs for doing data-intensive science and where is most of expensive specialist time lost?*
- *How to efficiently cross discipline boundaries when looking for resources? Would an Open Data and Service Agora help to quickly find useful components or would you continue to rely on discussions with colleagues?*
- *Which kind of cultural aspects (for example of sharing) need to be changed most and how can a change in culture be encouraged? Are there some fields of research where this could be easiest and act as examples? Does culture change come purely from scientists or are measures by funders required?*

- **Data in mathematics** takes three forms:
  - numerical data
  - computer software (implementation of algorithms, sometimes of constructive proofs)
  - theorem statements and proofs
- **Open Data and Service Agora** is a priority. It would allow for instance more atomic and relevant training around the use of data (can imagine for instance a machine learning/statistics course for scientific data based around the re-exploitation of public scientific datasets)
- **Sharing** of data brings little **recognition**. Even if someone does something useful with released data, at best the original creator will get a citation. Encouragement for funders/agencies etc. to include a "Dataset" **section in CVs**, just as there is one for "Awards", "Grants", etc.
- *Strong institutions seem to have enough resources (human, machines) to undertake data-intensive work, but most researchers do not have a chance to participate? Is that something we can and should change and if so how?*
  First of all, I believe that is not because institutions are strong that they have **enough resources** to undertake **data-intensive work**. Large institutions are often synonymous of very rigid workforce and strong stiffness that preclude the development of new areas of activities. Regarding the participation of researchers, I am not sure that they do not have the **chance to participate**, rather than the **will of participate**. While in the 90s, researchers were at the **forefront of computer usage** (regardless of their field of research), they are now lagging behind. My two kids (18 and 9) are more proficient in computer usage than most of my colleagues. So if we do something, I suggest the urgency is to convince researchers that indeed they could benefit from **data sharing, re-use and interoperability**.
- *Where are the biggest costs for doing data-intensive science and where is most of expensive specialist time lost?*
  I think the **big costs** are currently in the salaries and the storage. It is **hard to find specialists** of large dataset analysis. And storage is much more expensive than compute. As far as specialist time lost, I would say that it splits between **quality control/fixing things** and **lack of structure/format conversion**.
- *How to efficiently cross discipline boundaries when looking for resources? Would an Open Data and Service Agora help to quickly find useful components or would you continue to rely on discussions with colleagues?*
  I think the big data field is still young, so there is not a lot of **generic training and documentation**. Each small field re-invents the wheel, for instance re-invent statistical procedures or algorithm to structure and compress datasets. I think we should develop large **dataset analysis as a discipline in university**, in its own right, not as part of computing science. There should be more textbooks on the subject. etc.

- *Which kind of cultural aspects (for example of sharing) need to be changed most and how can a change in culture be encouraged? Are there some fields of research where this could be easiest and act as examples? Does culture change come purely from scientists or are measures by funders required?*
  Lack of sharing is definitively a **cultural problem**. Equally important in research I think is **re-using**. By that I mean **trusting data** coming from others. It is perhaps not a problem in physics, but in life science definitively. Another cultural problem to overcome is the walls that we all build around our subfields. There is a completely **arbitrary restriction** of the world of information to the close circle of our colleagues. We would not go to a different domain of research to find a product or a solution. We will first look under our nose, and then rather than looking up, we **re-invent** (often in an unsatisfactory manner) what already exists in a nearby field. A perfect example of that in my subject is the complete disconnection between systems biology and pharmacometrics. The two domains ignore each other, and would rather share a toothbrush than re-use an algorithm or a software tool. Despite the fact that for anyone not specialist of mathematical modeling of biomolecular processes, these two populations are barely different.
- I do see one "grass roots" initiative, namely the **Mind Repository**, which I and other colleagues have contributed to: http://read.psych.uni-potsdam.de/. Is there a way for RDA to benefit from this and other like-minded efforts?