Third RDA Europe Science Workshop "Big and Smaller Data"

Paris, 19-20 April 2016

Abstract

This report summarizes the outcomes of the RDA Europe 2016 Science Workshop, which was organised in collaboration with the Centre National de la Recherche Scientifique (CNRS) at the CNRS Headquarters in Paris, France, on 19-20 April 2016. The Workshop was organised as an open brainstorming, which provided a view of the disciplinary practices for data management and sharing in a variety of disciplines, and allowed to identify common needs and issues. The needs expressed during the Workshop support the global RDA strategy and reinforce the importance of many of the RDA activities, with specific recommendations to take into account by RDA Global and RDA Europe.

The Workshop was organised in collaboration by the RDA Europe project and the CNRS. The contact at CNRS was the *Direction de l'Information Scientifique et Technique* (IST Direction), led by R. Fabre. The participants were selected to obtain a good balance of the disciplinary fields covered, to get participants from a range of European countries, and to allow a good representation of the local Organisation and national strategic topics. The meeting was briefly introduced by R. Fabre for CNRS and P. Wittenburg for RDA Europe. Then the participants presented their contributions, with time for discussion after each of them. The rest of the meeting were devoted to general discussion, and at the end of the meeting the participants were invited to send a summary of their personal conclusions. The discussion was very lively and rich. The main topics addressed during the meeting are summarized in the next section. The conclusions with respect to RDA activities are described in Section 2. The list of participants is provided in Annex.

1. Main topics identified during the meeting

The participants were all interested in science data sharing, but they came from different disciplines which are not at the same level in that domain. They all started their presentation from the needs they identified, and some described the disciplinary international landscape or the RDA activities in which they are involved when relevant.

The presentations and discussions were a fascinating kaleidoscope of disciplinary practices for data management and sharing and ideas about future development. The participants noted at the end of the meeting that in spite of the high diversity of their backgrounds, they shared common topics and spoke the same language – one of them noted that this is not always the case between researchers from much closer disciplines who use different technical vocabularies. Also, the comment by one of the participants that for the practitioners "Riding the Wave" (by reference to the 2010 report of the EC High Level Group on Scientific Data) is rather like "climbing mountains" was certainly widely shared.

Several themes were present in most presentations, led to wide discussion, and were also recalled in the participants post-meeting messages. It was not surprising to see that the most critical high level points identified to enable data sharing are the need for standards and the reward system**.**

The **need for standards** is an obvious technical component of data sharing, discovery and reuse, which involves many sociological aspects to be properly set up. The discussion showed that in some cases standardization operates at the national level, whereas data sharing issues are global. In other cases there may be competing organisations or sub-communities with different objectives and agendas, which are difficult to coordinate and do not always satisfy the needs of scientists. RDA can certainly plays a role here.

The problems due to the **reward system** are on the sociological side. On that front, political backing, funder requests to define a Data Management Plan or to provide information on applicants' contributions besides publications, journal policies to require publication of supporting data, the establishment of data journals and the promotion of data papers in the existing journals, were cited among the **incentives to data sharing**. But the current reward system emerged forcefully, once again, as the worse obstacle to the sharing of scientific data when discussing how to initiate data sharing in the scientific community. The problem is now well identified, but it is difficult to propose an efficient way to improve the situation.

More generally, in many domains the **faculty culture** has to evolve with respect to the importance of data-based science. This shows once again that **sociological problems are more important that technical ones when dealing with science data** (as in many other contexts!). Also, the evolution of **curricula** to include data sharing and usage is of tremendous importance to enable evolution on the medium term. The **knowledge of digital tools and standards should be disseminated** in Schools, Universities and scientific communities. The scientific community should be accompanied to share data and use shared data, as well as rewarded for doing so. It was also noted that **facilities such as the ESFRIs can influence their communities**.

The presentations strongly indicated that a differential treatment of disciplines is required to accommodate for **data diversity**, with as a first step the specific requirements in each specific field. There are good examples of different ways to organise discussion on the specific **disciplinary aspects of interoperability standards for data discovery and reuse** (data formats, disciplinary metadata including a minimal set of metadata describing the reference ontology, the dataset and the data item, etc.) depending on the way the specific scientific community is organised. The Workshop demonstrated that sharing information, lessons learnt and best practices across disciplinary borders is important and can suggest possible solutions to the beginners. Diversity, which can include crowd-sourced data, also increases when **cross- and interdisciplinary** usage of data takes place, which is daily life for some topics, in particular societal challenges such as environment, biodiversity and health.

Despite diversity, there are **commonalities** with respect to standards, for instance on **identifiers** and **data citation**. **Trust** is also identified as an important keyword. The urgent need to establish **trustworthy repositories** was underlined. Certification should evaluate the technical procedures implemented by the repository, but also include criteria to evaluate scientific oversight to establish trust in the content and make sure that user needs are taken into account. An interesting additional point is how to identify trustworthy standards. What are the **health indicators of standards**? Also, **legal and ethical aspects** of data sharing are a key issue. **IPR** are at the core of the relationship between Open Data and industry, and **licences** can be an obstacle to data reuse inside the scientific community.

In the current trend towards creating a "multidisciplinary open or controlled data infrastructure", the **EOSC** or the **Commons** discussed by the e-IRG in the current European context, the usefulness of gathering all the necessary inputs to identify the open issues, the difficulties and the successes of the different communities was underlined. The need to **include also "long tail" data** is well identified, and it was noted that some "big facilities" produce "small science". It was noticed that the RDA already has several Groups working on different aspects, and that it would be interesting to see **how the different pieces can be put together** to build up a vision of this e-Infrastructure.

In more details, among the following common topics identified:

- The need to enable **data citation** and to establish **data bibliometrics**, which is linked in particular to the implementation of metrics to measure a scientist's contribution in the discussion of credits which could be used in career development. More technical and practical questions, for instance about the implementation and usage of **Digital Object Identifiers** (DOI), where also addressed. Some disciplines assign DOIs to experiments, not to data sets. The need to identify precisely text passages for quotation from data based corpora was also identified. Also, researchers should know how to get a PID hence the need for methods, dissemination and training.
- Commonly used **methodologies** were identified (statistics, GIS, text analysis, image processing, 3D models, data wharehousing). It is not clear how the usage of these common methodologies could be fostered and supported, but they are clearly key topics for education and training. Would it be useful to bring these topics in the "RDA arena"? This is the communities' decision, so it is important that they know that it is possible, and a powerful way of interacting with colleagues and gathering advice, if they need a neutral coordination place.
- **Open access to data** is a key point, from the domain of ancient texts to biology and medicine. It was also strongly stated that **the right to read is the right to mine**. This means understanding of and influence on copyright laws at the national, regional and international levels.
- The importance of **Data Management Plans** (DMP) was stressed once again, but it was noted that they are sometimes not popular among researchers. In this case, the landscape is evolving with the incentives from funding agencies, but it is also important to demonstrate that DMPs are a useful tool for data producers and providers.
- **Data quality** is another important element of trust. It has to be fully taken into account when setting up the data management system, and information about it should be provided. It is also important to assess ways of checking the quality of annotations in disciplines which use them.
- The fact that users often request **well documented data products** rather than the original data, **APIs** (Application Programming Interfaces), **data portals, workflows** and **Virtual Research Environments** was cited. **Documented** and **easy to use** tools and systems to share and exploit data are needed in addition to the data itself.
- **Software** is an integral part of data sharing, and the need for sustainable and adaptable code, and software management planning, preservation and citation was underlined. The specific need for well-defined software input and output for workflows was identified.
- It was noted that some facilities or disciplines have to store **huge data volumes**, but in many cases volume is not the problem but rather **diversity** and **complexity**. **Sensibility** should be added to the usual 4V, volume/variability /veracity/velocity, as critical factor for the Big Data endeavour. It is also critical for each discipline and each facility to assess what data to discard and when.
- The need to make data useful and usable should rely mostly on data producers and/or on **disciplinary data infrastructures** with dedicated and knowledgeable staff. **Local data managers** have to be trained and their role recognized.
- The evolution of data dissemination and data usage methods and practices has to be considered by **disciplines which have been sharing data for a long time**, and which built an international framework to do so. However for at least some of them, it can be difficult to adjust to the new

landscape, especially when the community is satisfied with the current approach. For instance, the data can be open and shared in international services but DOIs and data citation not be common practice. The storage size and complexity increases, more computing power as well as new ways of discovering and distributing data should be enabled. The trend to impose authentication to use open data (for instance to track users to feed Key Performance Indicators) can also be an issue.

- The interest of computing before data transmission ("**Bring computing to data**") to reduce data traffic was stressed for disciplines which produce huge data volumes. It was also noted that for instance Health data are getting more and more accessible "within the hospital", but that sharing them beyond remains an issue. This is another aspect of "**bring data analysis where the data are**". **Data Analysis** and **Data Analytics as a Service** were cited as an example of need due to increased data volumes.
- Individual standards were also commented, for instance the complexity of TEI was noted. The need for recommendations for corpora annotation was underlined.

Several possible strands of work for RDA Groups can be identified from the topics discussed. The following ones were specifically discussed during the meeting or in the post-meeting comments received from the participants:

- A RDA Group for **neuroscience data**, to broker activities in the area, in particular those undertook by the Human Brain Project (HBP) and the International Neuroinformatics Coordinating Facility (INCF). There has been an unsuccessful application for a COST action to bring together the two communities of data producing scientists and more data analytically engaged scientists. It is proposed that a RDA WG could stand at the start of a new COST action to bring together these two communities.
- Data is valuable and its re-use can boost scientific productivity, which is why there are many efforts to make data more accessible for re-use, integration and further analysis. Data is also valuable for companies, and there is a risk that discoveries made and products designed based on these data, limit further access and scientific use of these discoveries. This may be inappropriate because usually data acquisition and making it accessible has been publically funded. Furthermore, the recent drive of national governments and EU to have scientists pursue collaboration with industry, may limit the availability of data acquired by public funding through the use of restrictive IP policies. A working group is necessary to **review the consistency of data sharing policies** of funding agents, governments and industry and make recommendations to ensure industrial strength, scientific productivity and openness of data. Access to data produced in the framework of private companies was discussed, with for instance the example of Twitter data (access right allowed to only 1%), as well as the reuse of clinical data managed by hospitals for medical research. Another comment is that there can be IP issues for a single component of a workflow.
- Start a **WG to formulate recommendations on how to give credit to scientists for data sharing** for government, universities and other research organisations. One can cite here the work of the Bioresouce Research Impact Factor (BRIF) framework on the citation and acknowledgment of bioresources and the definition of indicators, in collaboration with the relevant journals. It would be very useful for them to share experience with similar initiatives. This type of practical work comes in complement of the policy aspects.
- Start an **IG on Industry engagement**.

- It would be very useful to **adapt the RDA PID type registry to environmental sciences**, which would also be an interesting exercise because of the data diversity in that domain.

2. Conclusions for the RDA

How do the RDA activities fit with the Workshop conclusions? The **diversity of the topics** discussed makes, once again, the point for the diversity of activities performed in the RDA. Some recommendations of the workshop can be used at the RDA Global level. Others fit with or could trigger activities managed by the RDA Europe project. This was the case for the discussion about persistent identifiers as explained below. As explained above, it was recommended that "RDA tackles" several new topics, but it was also well recognized that the RDA by itself does not create Interest and Working Groups. The RDA and RDA Europe can however work at motivating communities to join to deal with the topics of interest for them, keeping in mind the topics listed at the end of Section 1. For instance, a Bird of a Feather (BoF) session on how to give credit to scientists for data sharing was organised at the RDA Barcelona Plenary meeting (April 2017) with the support of one of the Science Workshop participant.

One of the **key links of the RDA with communities** is through people who participate in the activities, know and promote RDA culture, finding and outputs in their own communities, which meets the so-called "**Ambassador programme**" which is proposed in the RDA Future Direction Plan. The different level of maturity of different communities with respect to data management, sharing and reuse is one of the important factors to take into account. Another path is **participation in major scientific conferences**, which can be supported by RDA Europe (and the other Regional RDAs) like for the 2016 Digital Humanities conference. It was suggested to ensure RDA presence to the 2018 Conference of the EuroScience Open Forum (ESOF 2018), which will be held in Toulouse (France). These biennial conferences attract several thousands of participants and are a way to reach a wide range of communities. **National entry points** can also play a liaison role. All these liaisons should be done in a coordinated way. **Transfer to industry/business** has to be worked out, and a dedicated Interest Group may help by polling the knowledge of the diverse RDA membership, in addition to the Groups which already build up relationship with the private sector.

The RDA can play a key role for enabling communities to establish their **interoperability standards** in a neutral, international framework. Some disciplines already have their international forum for that purpose, but those who do not can use the RDA for that purpose. Agriculture is a striking example of how a discipline can use the RDA for its own aims: the agriculture community began to participate in the RDA at the first Plenary, through information given to INRA and shared by them with their international collaborators. They set up first an Interest Group on Agriculture Data, and then quickly established a Working Group on Wheat Data Interoperability, which produced the interoperability framework of the intergovernmental International Wheat Initiative (including exchange format, best practices, portal and prototype). The presentation during the workshop of the context of wheat data sharing and how the interoperability framework was progressively built was enlightening.

Organising the exchange of information and best practices on examples of **governance** for disciplinary interoperability frameworks would help disciplines which are not yet organised. The emerging proposal for

an *Interest Group on Disciplinary Interoperability Frameworks*, which was also the subject of a BoF session at the Barcelona Plenary meeting, would help to fill the need.

Several **technical/technological topics** addressed during the Workshop are already tackled in the RDA. In particular several Groups deal with data publication and citation, in liaison with publishers. On more **sociological aspects**, the Recommendations on Repository Certification prepared in the relevant DSA/WDS Working Group, which is being implemented by the DSA and the WDS, is for instance an important building block on a topic raised during the meeting, as well as the work on Active Data Management Plans.

RDA Europe organises **Workshops** on specific topics. The discussion about DOIs during the Workshop inspired the *Views about PID Systems. Training Course and Workshop*[1] organised in Garching 31 August- 2 September 2016. The Science Workshop participants were invited to attend.

Education and training are identified as key enablers. The RDA itself is an efficient machine to **disseminate good practices** in its wide membership, in particular towards plenary participants and the members who are involved in the activities of the Groups. The members then can inform and train their own communities. Several Groups focused on community needs also **assess and prototype training activities**, which can then be implemented by others, or produce guidelines (e.g. 23 Things: Libraries for Research Data). The RDA vision of its role for training is summarized in a recent paper produced by the RDA Council Strategy Subcommittee[2]. and submitted to a Request for Comments from the community.

RDA makes a very strong plea for instance for Persistent Identifiers, a key technical building block for scientific data sharing. But for the moment it has not been keen to produce **general statements** on what it can identify as essential building blocks of scientific data sharing on the sociological front. However sociological bridges to enable scientific data sharing are fully in its remit. The general understanding that the current **reward system** of scientists is a major obstacle and that there is for the moment no solution, suggests that the RDA should address the question. The RDA has no vocation to lead the implementation of the necessary evolution, but it can host assessments of possible evolutions if enough members are interested to participate. In addition, a clear statement building on its wide, international membership would come in support, in particular, to the strong recommendation of the EOSC High Level Expert Group on the same topic. The possibility for the RDA to produce these kinds of statement will be discussed in the RDA Council Strategy Subcommittee.

The discussion shows that the **RDA has a strong role to play in the definition of the EOSC**, because it gathers world-wide activities on many of its building blocks, including sociological ones.

In practice, the participants also underlined that the diverse activities of the RDA are sometimes difficult to follow, and that **support is needed to identify the activities of interest and their status**. The efforts done by RDA Europe on the Atlas of Knowledge (AoK), and the interest of the RDA Technical Advisory Board for assessing the possibility of mapping RDA activities in collaboration with the AoK, go in that direction.

[1] https://www.rd-alliance.org/views-about-pid-systems-training-course-and-workshop-31-august-2-september-2016-garchingmunich

[2] The RDA Coucil Strategy Subcommittee has begun to produce reference documents about RDA strategy and roles which are submitted for community comments.

In conclusion, the brainstorming exercise performed during the 2016 RDA Europe Science Workshop provided a good overview of disciplinary practices. The needs expressed during the Workshop support the global RDA strategy and reinforce the importance of many of the RDA activities, with specific recommendations to take into account by RDA Global and RDA Europe.

It was the first time that the Science Workshop was organised as a fully open brainstorming. This provided a good picture of the data landscape seen with scientists' point of view and of the general needs, which is useful for assessing and guiding RDA activities at the global and European level. It is likely not worth repeating the same exercise before a few years, to let enough time for evolution of the landscape. The next Workshop, which will be organised by STFC, will focus on a few topics.

Annex – List of participants

Invited Scientists		
Hans Bennis	Royal Netherlands Academy of Arts and Science	The Netherlands
Carine Bruyninx	Royal Observatory of Belgium and EUREF Permanent GNSS Network (EPN)	Belgium
Cécile Callou	UMS3468 BBEES (Bases de données sur la Biodiversité, Ecologie, Environnement et Sociétés – Databases on biodiversity, ecology, environment and societies), CNRS/INEE - MNHN	France
Anne Cambon-Thomsen	UMR1027 Epidémiologie et analyses en santé publique (Epidemiology and analyses in public health), INSERM-Université Toulouse III	France
Marc Cuggia	Inserm UMR 1099 LTSI - Equipe données massives en santé (Team Massive Data in Health) Département D'information médicale CHU de Rennes	France
Francisco Doblas Reyes	Institució Catalan de Recerca i Estudis Avançats (ICREA) & Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS) Barcelona	Spain
Marcello Maggi	Istituto Nazionale di Fisica Nucleare (INFN) of Bari	Italy
Marie-Angélique Laporte	Biodiversity International	International
Brian Matthews	Science and Technology Facilities Council	United Kingdom
Jean-Luc Minel	MoDyCO, Modèles DYnamiques Corpus, UMR 7114 Université Paris-Ouest Nanterre La Défense - CNRS	France
Stéphane Rondenay	Department of Earth Sciences, University of Bergen	Norway

Charlotte Schubert	Lehrstuhl für Alte Geschichte, Historisches Seminar, Universität Leipzig	Germany
Paul Tiesinga	Neuroinformatics Department, Donders Institute for Brain, Cognition and Behavior, Radboud University	The Netherlands
CNRS and INRA		
Renaud Fabre	Director DIST	CNRS
Laurence El Khouri	DIST	CNRS
Francis André	DIST	CNRS
Odile Hologne	DIST	INRA
RDA Europe 3/RDA		
Françoise Genova	CNRS – Université de Strasbourg UMR 7550 Observatoire Astronomique de Strasbourg	France
Wolfram Horstmann	State and University Library Goettingen	Germany
Leif Laaksonen	Finnish IT Center for Science, WP2 Lead RDA Europe 3	Finland
Raphael Ritz	The Max Planck Computing and Data Facility (MPCDF), WP4 Lead RDA Europe 3	Germany
Andrew Treloar	Australian National Data Service, co-chair of RDA TAB (partial attendence)	Australia
Peter Witenburg	The Max Planck Computing and Data Facility, Coordinator RDA Europe 3	Germany