RESEARCH DATA ALLIANCE

European Scientists view on Research Data and corresponding challenges

RDA-MPG Science Workshop on Data Report - Compiled by Bernard Schutz, Leif Laaksonen, Raphael Ritz, Herman Stehouwer, Peter Wittenburg, Rob Baxter

The widespread adoption of the Internet in the 80s was met with scepticism by science as to whether it could truly foster scientific research. Within just a decade, science had fully adopted both the Internet and its various layered infrastructures such as the World-Wide Web, as science understood that the exchange of knowledge, information and data between the rapidly increasing number and types of computers could now be done within seconds, almost seamlessly. It relieved scientists from many time-consuming aspects of traditional communication and exchange channels. Agreement on a few basic principles (node numbering, protocols, registries) at a time where many competitive suggestions were brought forward allowed scientists to shift their attention back to new scientific questions, simply making use of the new facilities rather than trying to invent them.

Currently we seem to be in a comparable situation, where the number and complexity of data exceeds our abilities to deal with them manually or through traditional means such as file systems. Fragmentation within disciplines, across disciplines and often across organizational boundaries (projects, institutes, states) is increasing rather than decreasing, and in many scientific domains the amount of time needed to manage and manipulate data to make them re-usable has become intolerable without support from new, highly automated processes. These trends with respect to data in science and beyond require new approaches to our management of data in the coming decades. Hence the Research Data Alliance (RDA), an initiative inspired by the Internet Engineering Taskforce, started, like the IETF, as a grass-roots, bottom-up organization designed to come up with formal agreements, specifications, running code – by data practitioners, for data practitioners.

The RDA-MPG workshop on Scientific Data brought together a number of leading European scientists* and other experts* to discuss current points of concern in the context of research data 10-11 February 2014 at the Max Planck Society headquarters in Munich. The main

*See Participant list on page 7

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424

objectives were to discuss whether the participants see a role for the Research Data Alliance (RDA) in their daily activities, what the science community's expectations might be, and whether the RDA roadmap needs to adapt to meet those expectations. The context of the Workshop was set by a number of questions covering scientific concerns, data sharing & publishing, stakeholder aspects, data infrastructures, technological trends and aspects of data science education.

Recommendations

The outcomes of the Workshop discussions are classified under a number of headings – data sharing and re-use, publishing and citing data, infrastructures and repositories, and general observations on the nature of the research data challenge we face today. The discussions culminated in a number of recommendations for RDA to consider:

1. RDA can play an important role if it is able to come up with recommendations, API specifications, guidelines, etc. that help to overcome the many one-shot, point solutions currently being implemented and hence make infrastructure building more cost-effective.
2. RDA must indeed be a bottom-up organization, and needs to strike the right balance between a better balance between bottom-up and its current, rather top-heavy, state.
3. RDA must motivate a "middle layer" of data scientists and to get engaged, rather than hope for too much engagement from leading researchers.
4. RDA must be aware that it may find itself in a race towards specifications and solutions with big commercial players who may win with de facto standards, simply because they arrive first.

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424

5. Expectations RDA has to meet:

a) Invest in training younger generations of data scientists.

b) Push demo projects, act as a clearing house and should be able to give advice on data management, access and re-use to everyone in research.

c) Have data experts who can visit institutes and help them implement solutions.

d) Perform good quality assessment on the first working-group results due in September 2014, and take care to not fall into the trap of overselling.

Sharing and Re-use of Data

On the specific topics of data sharing and re-use, the Workshop made the following key observations:

1. Re-using and sharing data and information has only just begun for many reasons, such as the difficulty in understanding each other's data, lack of visibility and accessibility, lack of high quality metadata descriptions that facilitate re-use, a reluctance to invest time in proper documentation when the rewards are not obvious and other sociological factors (many noted in the previous section). Despite the general support for open access we need to accept that there are some serious limits to openness which mostly are of a sociological nature.

2. Despite enormous progress we still lack efficient, cross-disciplinary agreed methods to describe and process data semantically in a way which enables re-use. Too much handcrafting is required, leading to the creation of one-shot solutions which do not scale. On a stage where increasingly many players produce data, this cannot continue.

3. In some disciplines the mapping of data to agreed reference data is needed to create a common ground on which comparative analysis can take place. Establishing and maintaining such reference data is costly.

4. Re-using data can only be successful if we can trust its identity, integrity, authenticity and the seriousness of all actors that are involved in the production chain. However, the mechanisms to establish and prove trust in a seamless way are not in place.

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424

Publishing and Citing Data

On publishing and citing research data, the Workshop made the following key observations:

1. Publishing results and being able to cite them is at the core of the scientific process. Because of the increasing relevance of data we need to come to a data publication and citation machinery which is accepted worldwide, and which reflects the higher complexity of the data domain (volumes, dynamics, relations, etc.) compared to the domain of publications.

2. Referencing data (e.g. using some form of persistent identifier [PID] system) must be stable. In several fields PID systems have not been as stable over the years as is needed.

3. Being able to refer to accessible data has at least two different aspects: 1) to execute workflows in reproducible science we need to be able to refer to data objects and collections; 2) for referring to a record of knowledge we also need to have mechanisms to cite data that has been published in a catalogue or journal in association with a scientific paper, and thus has undergone some form of quality assessment.

4. It has not yet been clarified whether data publication can be as highly rated for career building as peer-reviewed scientific papers. Some researcher argue that there is also a difference with respect to career intentions in each case: scientists versus data scientists.

5. Being able to refer to or cite data requires an infrastructure to store identifiers persistently, along with attributes and the data themselves. This is costly and currently it is not obvious who will pay for such an infrastructure. The responsibility – national, regional, organizational – needs to be clarified soon to get this infrastructure in place. The currently available systems and approaches are not reliable enough.

Infrastructures and Repositories

On the nature and provision of data infrastructure and repositories, the Workshop made the following key observations:

1. There is no doubt that we need infrastructures to be able to deal with data in a much more seamless and efficient way. The components of such infrastructures are still not clearly identified, but trusted and persistent repositories are obviously a cornerstone. Repositories can be organized at discipline, organizational and/or regional level, and data and metadata flow between them should be as transparent as possible based on

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424

agreed interfacing and procedural standards. Repositories require continuous funding, clear responsibilities and participation in quality assessments.

2. Researchers need to be in the driving seat to ensure that infrastructure building and maintenance meet the needs of research, that trust can be established and that thin and cost-efficient layers are being implemented. Trust can best be established within regional boundaries and within disciplines, in both cases based on tradition and culture.

3. Open access as a general principle is to be supported but there are many reasons that some data need to be protected, be it for an incubation period to protect scientific advantage, be it because data contains sensitive information, or to meet the requirements of licences, etc.

4. Offering services on data rather than just data per se has a big advantage for some researchers. However, these services offer restricted views on data and thus can fail to meet the needs of all researchers. A combination of both ways makes sense but providing and maintaining services are costly.

5. Infrastructures need to encompass existing repositories which implies that lots of legacy systems need to be integrated. Only a focus on abstract interfacing layers can solve the integration, requiring adaptations and compromises at both sides. The costs must not be underestimated.

6. Commercial companies have realized that data, and the information enclosed in them, have a high potential value, and thus invest large amounts of money and effort to gain access to data and to sell services around them. The viability of this model in the science domain is not evident, since there is a clear lack of trust at various levels (restrictions on data with a potential economic value, persistency, protection, dependence, future costs, etc.). Companies have the advantage that they don't have to care about legacy data organization for their services. They define the rules of the game, making services more cost effective.

7. To find trusted repositories, useful services and interesting collections easily, infrastructures need to set up and maintain a variety of registries and catalogues.

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424

Spectra of Data

Regarding "Big" and "Small" Data, the Workshop noted that there are several axes or spectra that can be identified, with the poles marked in the two columns below. Every research discipline or project, and even individual researchers, find their place on these spectra. Obviously the various types of data require different strategies.

Some communities have very heterogeneous data (on many of these axes), which raises more issues.

Well-structured data	Heterogeneous data sets
Data with automatically generated metadata	Data with complex metadata issues
Static data	Dynamically changing data
Data acquired under controlled conditions	Crowd-sourced data
Centrally managed databases	Widely distributed data, no clear curation
Data that are computationally simple to handle	Data needing massive computing
Data that are used "raw"	Data that are understandable only after processing
Numerical data	Text data
Communities knowledgeable about data processing	Communities scared of data
Communities with trust	Communities with no tradition of sharing, even with distrust
Open data	Proprietary/embargoed data, data with copyright issues
Proprietary/embargoed data, data with copyright issues	Data with privacy issues
Privately generated data	Data with publicly funded stakeholders

A complete version of the Science Workshop report is available on RDA Europe web site:

https://europe.rd-alliance.org/Repository/document/Publications%20and%20Reports/RDA-Europe-Science-Workshop-Report_final_April2014.pdf

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424

PARTICIPANT LIST

Scientists

Bernard Schutz - Gravitational Physics - Cardiff University MPG
Bruce Allen - Gravitational Physics - MPI for Gravitation physics, Hannover
Bruno Leibundgut - Astronomy - ESO, Garching
Cécile Callou - Archaezoology/Biodiversity-Ecology-Environment - Museum d'Histoire Naturelle, Paris
Christine Gaspin - Bio-Informatics - INRA, Toulouse
Dick Dee - Meteorology - ECMWF
Jan Bjaalie - Neuroanatomy and Computer Science - University of Oslo
Francoise Genova - Astronomy - CNRS, Strasbourg
Jochem Marotzke - Climate Model - MPI for Meteorology, Hamburg
Manfred Laubichler - History of Science - New Mexico University
Marc Brysbaert - Psychology - Ghent University
Mark Hahnel - Biology - Figshare and Imperial College, London
Markku Kulmala - Atmospheric Sciences - University of Helsinki
Peter Coveney - Chemistry, biomedicine - UCL, London
Stefano Nativi - Earth System Science and Environmental Technologies CNR, Roma

Experts

Carlos Morais-Pires - e-Infrastructures - European Commission
Donatella Castelli - RDA/E Member/Computer Science & ISTI-CNR, Pisa
Frank Sander - MPDL Director - MPDL, Munich
Leif Laaksonen - RDA/E Coordinator & CSC-IT Center for Science, Helsinki
Peter Wittenburg - RDA-TAB Member/Linguistics - MPI for Psycholinguistics, Nijmegen
Ramin Yahyapour - GWDG Director - GWDG, Göttingen
Raphael Ritz - RDA/E Member/NeuroInformatics - MPG, Garching
Reinhard Budich - Data Scientist - MPI Meteorology, Hamburg
Riam Kanso - Data Policies/Cognitive Neuroscience - UCL, London
Stefan Heinzel - RZG Director - MPG, Garching
Herman Stehouwer - RDA Secretariat & MPI for Psycholinguistics, Nijmegen
Ari Asmi - Atmospheric Sciences - University of Helsinki

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424

ABOUT RDA

RDA Europe, the European plug-in to the global RDA, supports RDA global and brings the European voice to the table. RDA is a bottom-up alliance where different players (technology, community, policy, legal, etc.) self organise and collaborate. Their main co-operation mechanism being Working & Interest groups focusing on different domain and technological challenges working towards global interoperability.

RDA Working & Interest Groups as drivers

Scientists come together through focused Working Groups, exploratory Interest Groups to exchange knowledge, share discoveries, discuss barriers and potential solutions, explore and define policies and test as well as harmonise standards to enhance and facilitate global data sharing.

Some of the over 40 RDA Working & Interest groups are closely collaborating with initiatives such as, CODATA, WDS, W3C, etc and where possible RDA looks to interact with worldwide initiatives where collaboration is of mutual benefit and builds on what has been done already.

Join us on-line

What data sharing problem are you trying to solve? RDA boasts a broad, committed membership of over 1,600 individuals, from more than 70 countries. Get involved by becoming an RDA Member and signing up to the Working / Interest Group(s) on www.rd-alliance.org. Membership is free and open to all.

Join us at our Plenary Meetings

RDA members meet biannually at plenary working meetings, where they are the drivers behind the agenda and the discussions. The 4th Plenary Meeting 22-24 September 2014 Amsterdam, The Netherlands. Find out more on www.rd-alliance.org.

rdaeurope@rd-alliance.org | europe.rd-alliance.org

RDA Europe has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 312424