Project Acronym	RDA Europe
Project Title	Research Data Alliance Europe
Project Number	312424
Document Title	Report on the RDA-MPG Science Workshop on Data
Date	April 2014
Editors	Bernard Schutz, Leif Laaksonen, Raphael Ritz, Herman Stehouwer, Peter Wittenburg, Rob Baxter

ABSTRACT

This report summarizes the main outcome of the RDA-MPG workshop on Scientific Data on February 10-11th 2014 at the Max Planck Society headquarters in Munich. It offers a view of current research data infrastructure and management issues from the perspective of a panel of distinguished researchers, drawn from fields as wide apart as gravitational physics and biodiversity. In conclusion, it offers a number of recommendations for the RDA to consider.

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

DISCLAIMER

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

This document contains information on RDA Europe (*Research Data Alliance Europe*) core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as RDA Europe Forum members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the RDA Europe Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 28 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

TABLE OF CONTENTS

Executive Summary

This report summarizes the main outcome of the RDA-MPG workshop on Scientific Data on February 10-11th 2014 at the Max Planck Society headquarters in Munich. The Workshop brought together a number of leading European scientists to discuss current points of concern in the context of research data. The main aims of the Workshop were to discuss whether the participants see a role for the Research Data Alliance (RDA), 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 circulated to participants beforehand, covering scientific concerns, data sharing & publishing, stakeholder aspects, data infrastructures, technological trends and aspects of data science education.

This report summarises the outcomes of the Workshop discussions 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 more first.
5. There are a few expectations RDA has to meet:
    a. RDA should invest in training younger generations of data scientists.
    b. RDA should 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. RDA should have data experts who can visit institutes and help them implement solutions.
    d. RDA should perform a good quality assessment on the first working-group results due in September 2014, and should take care to not fall into the trap of overselling.

1 Background and Aims of the Workshop

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, since 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 scientist to shift their attention back again 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**, 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.

1.1 Workshop Goals

The primary goal of the cross-disciplinary RDA Europe/MPG Science Workshop was to bring together a number of leading European scientists to discuss current points of concern in the context of research data (see Appendix A for a list of participants). The participants represented a broad range of scientific and research disciplines, including astronomy, biodiversity, bio-informatics, chemistry, Earth system science, ecology, environment, gravitational physics and meteorology and were joined by a number of guests representing RDA and the European Commission.

The main questions for the Workshop to discuss were: is there a role for the Research Data Alliance (RDA); what are the science community's expectations; and, how does the RDA roadmap need to adapt to meet those expectations. Since the RDA is not just focused on the here-and-now, all participants were asked to look ahead a little and describe the trends in their discipline.

1.2 Workshop Process

The scope for the discussions was set by a number of questions sent to all participants beforehand, plus the statements presented by the invited scientists. In addition, there were two dedicated presentations setting the context for the Workshop: one presentation introduced RDA and its possible benefits for science; and a second from the European Commission described the expectations and context from the funding policy angle.

For the core part of the Workshop the topics were grouped into two sessions. Each session was then:

- initiated by a few short statements from seven of the invited scientists;
- followed by an open discussion, structured and facilitated by the chair;
- concluded with a short summary.

The pre-Workshop questions, and summary statements drawn from them, are collected in Appendix B. For the Workshop itself, session 1 covered the questions of scientific concerns, data sharing and publishing & stakeholder aspects, while session 2 covered data infrastructures, technological trends and education aspects.

2 General Observations

Some of the concerns that were described by the scientists both beforehand and during the Workshop address topics that only the researchers themselves can solve – creating smart algorithms to reduce the amount of data needed/produced, for example, or negotiating with funders access to even bigger high performance computers. In this report we discuss only those aspects that have to do with the infrastructure that is required to be able to work efficiently with data. The borderline of what is science and what is infrastructure changes over the years.

Obviously scientists are interested in using operational and persistent infrastructures that add no additional overhead in working with them. For them the difference between the RDA, that *specifies* elements of an infrastructure, and others who *implement* infrastructure is of little relevance.

The main general observations arising from the Workshop sessions were:

1. It is evident that there are challenges which can only be solved by researchers themselves, by developing smarter algorithms and processes and by making use of cutting-edge technology. Our capabilities to compute and move data lag behind those of creating them; we require new methods and (obviously) a choice of optimization directions.
2. Leading-edge research is confronted with the challenges of larger volumes of data and the increasing need to introduce more sophisticated ways of organizing them. Only proper, systematic solutions will guarantee reproducible science in an era where data usage will largely be at distance, i.e. those re-using data will not know the details of each individual data object and will have to rely on software operating on collections defined by specific attributes.
3. For leading-edge science multidisciplinary research is a reality, requiring data from different disciplines and regions, different spatial and temporal resolutions, small and large collections, structured and unstructured types all to be combined. The need to combine data in such ways leads to a continuously evolving, complex adaptive system where sociological hurdles caused by traditions, culture, procedures, etc. need to be overcome to be successful. Currently re-using and combining data requires an enormous – and increasing – amount of effort.
4. Although many data are still being created by manual workflows, only automated workflows will have the power to cope with increasing data demands – not only for efficient data management, but in particular for smart data analysis. These will necessarily become part of new scientific application scenarios and thus need to be equipped with all modules establishing a “data fabric”.
5. The costs of dealing with data in all its different dimensions are currently too high; too much of the capacity of excellent researchers is occupied in managing, accessing and re-using data. Too many one-shot solutions dominate current practice, solutions which are obsolete within a short time. Bridging the gap between the acts of data creation and data consumption is too challenging because of the lack of appropriate metadata, little documentation of sufficient quality and too little information about structure and semantics.
6. To meet the challenges of seamless infrastructures, persistent and trusted repositories need to be built. In particular a new generation of data scientists needs to be trained, able to carry out all tasks at a high level.

3 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 hand-crafting 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.

4 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.

5 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 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.

6 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
Impersonal data	Data with privacy issues
Privately generated data	Data with publicly funded stakeholders

7 Conclusions and Recommendations for RDA

Two days of stimulating and engaging discussion were summarised and structured into a set of recommendations for the Research Data Alliance to consider. These were as follows:

1. Researchers are primarily interested in working, stable infrastructures that help solving challenging problems. RDA, as an organization working on specifications, is therefore far away from the researchers' main concerns, but it is nevertheless recognized that RDA can have an important role if it is able to come up with recommendations, API specifications, guidelines, etc. that help to overcome the many one-shot, restricted solutions and hence make infrastructure building more cost-effective. RDA can be a forum to bring together the good people working in these directions.
2. It is agreed that RDA must be a bottom-up organization if it wants to be successful. However, at this moment the impression is of an organization run too much from the top down. Since RDA is relatively young there are still quite some risks of failure; a better balance between bottom-up and top-down has the potential to reduce the risks.
3. RDA cannot expect leading researchers to engage in RDA activities; a middle layer of practitioners (data scientists and data librarians) needs to be motivated to get engaged. The critical question remains who has the time to spend the efforts.
4. The Workshop asks whether RDA will come up with specifications and solutions fast enough, compared to the big commercial players, and whether there is any chance for it to compete with commercial *de facto* standards.
5. There are a few expectations RDA has to meet:
    a. RDA should certainly invest in training younger generations of data scientists.
    b. RDA should 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. RDA should have data experts who can visit institutes and help them implement solutions.
    d. In September, when the first RDA results will become available, a good quality assessment should be done on the results, and RDA should take care to not fall into the trap of overselling.

Appendix A. Participants

15 leading scientists from different disciplines and countries were invited to this workshop. In addition we had a number of guests from different background who also participated in the discussions. Due to an emergency case Cécile Callou was unable to travel.

	Name	Field	Affiliation
R	Bernard Schutz	Gravitational Physics	Cardiff U / MPG
R	Bruce Allen	Gravitational Physics	MPI for Gravitationphysics, Hannover
R	Bruno Leibundgut	Astronomy	ESO, Garching
R	*Cécile Callou*	*Archaezoology/Biodiversity-Ecology-Environment*	*Museum d'Histoire Naturelle, Paris*
R	Christine Gaspin	Bio-Informatics	INRA, Toulouse
R	Dick Dee	Meteorology	ECMWF
R	Jan Bjaalie	Neuroanatomy and Computer Science	University of Oslo
R	Francoise Genova	Astronomy, RDA TAB	CNRS, Strasbourg
R	Jochem Marotzke	Climate Model	MPI for Meteorology, Hamburg
R	Manfred Laubichler	History of Science	New Mexico University
R	Marc Brysbaert	Psychology	Ghent University
R	Mark Hahnel	Biology	Figshare and Imperial College, London
R	Markku Kulmala	Atmospheric Sciences	University of Helsinki
R	Peter Coveney	Chemistry, biomedicine	UCL, London
R	Stefano Nativi	Earth System Science and Environmental Technologies	CNR, Roma
G	Carlos Morais-Pires	e-Infrastructures	European Commission
G	Donatella Castelli	RDA/E Member/Computer Science	ISTI-CNR, Pisa
G	Frank Sander	MPDL Director	MPDL, Munich
G	Leif Laaksonen	RDA/E Coordinator	CSC, Helsinki
G	Peter Wittenburg	RDA-TAB Member/Linguistics	MPI for Psycholinguistics, Nijmegen
G	Ramin Yahyapour	GWDG Director	GWDG, Göttingen
G	Raphael Ritz	RDA/E Member/NeuroInformatics	MPG, Garching
G	Reinhard Budich	Data Scientist	MPI Meteorology, Hamburg
G	Riam Kanso	Data Policies/Cognitive Neuroscience	UCL, London
G	Stefan Heinzel	RZG Director	MPG, Garching
G	Herman Stehouwer	RDA Secretariate	MPI for Psycholinguistics, Nijmegen
G	Ari Asmi	Atmospheric Sciences	University of Helsinki

Appendix B. Pre-Workshop Questions

The following questions were circulated to potential participants before the workshop to help scope the working sessions on the day.

B.1 Scientific Concerns

Questions asked:

- For what purpose and in which way do you generate which amount of data at what rate?
- How and where are those data stored - initially as well as in the long term?
- Are those data shared, with whom and how - initially as well as in the long term?
- How and where are those data (pre)processed and analysed? By whom?
- How do you keep track of this?
- Are there suitable tools, standards or best practices helping with the above?
    - If not: why not and should there be any?
- How do you expect the above to evolve in the near to mid-term future?
- What do you wish for?

Collected statements:

- Data Creation & Characteristics:
    - Large projects with cross-border and cross-discipline character are being formed. The data volumes these and other projects are creating are growing roughly with high speed reaching out to PB and EB.
    - We need to distinguish between for example sensor and simulation data which can amount to very large volumes, but where the structure and the semantics are comparatively simple and for example social science data where structural and semantic heterogeneity, vagueness and complexity play a much greater role. Different ways of treatment of such data is obvious.
    - Crowd Sourcing revolutionizes research in many disciplines in particular since millions can be equipped with sensors of all sorts. The result will be highly dynamic databases the content of which will emerge asynchronously leading to problems of citability etc. that need to be addressed. The citizen scientists included in these approaches will want to participate in the analysis of the data as well.
    - In many disciplines growth in all dimensions is not possible to the extent desired - this holds for data storage and management requirements, for computing cost requirements and for data transport capacities and costs. One driving factor is the much higher resolution of sensors, of simulation grids, etc. where an end of the dynamics cannot be seen.
    - The challenge in these sciences with high requirements is to compromise on some dimension. In simulation based science some will focus for example on ensembles, others on resolution or complexity or much longer simulation runs.
    - In many sciences the need for massive and smart reduction of incoming data is a challenge, this holds for humanities that are receiving multimedia streams from hundreds of subjects per day to climate modeling for example.
    - In many sciences temporal and spatial resolution scales cover tremendous ranges which require completely new techniques to store, use, combine and analyze the data.

  - Science is a source of new ideas and thus new formats and semantics, any hope to restrict these dynamics would be naive. This heterogeneity is challenging with respect to methodologies and harmonization. In some areas (quasi-) standards exist, but it remains a bottom-less pit.

- Re-Usage:
  - Data is going to be used by many unknown users, information about uncertainty must be conveyed to allow proper interpretations and re-usage.
  - Universal accessibility of data is an excellent goal, but we need to have mechanisms in place to ensure proper usage.
  - Science can often not just offer access to data, but needs to provide services on data that then include specific interpretations and usages. It is not easy to determine these services since one can invent many different ones.
  - The availability of metadata is essential to support interpretation and reusability.
  - Science is increasingly interested in combining different data sets, but yet we do not have generic ways and methods to support this in a flexible way.
  - In some sciences it is of big interest to see what kind of applications, queries etc. users are carrying out on the gathered data.
  - Multidisciplinary approaches are very much en vogue but extend the heterogeneity in all dimensions (scales, resolution, structures, semantics).
- Computation Aspects:
  - Scientific re-analyses is a very important task in some areas of research. Measures need to be taken and sophisticated tools need to be available to make it possible for numerous users of different background making use of a wide variety of applications.
  - Scientific data and its analysis are in very close proximity which requires optimized logistics and high communication speeds to meet researchers' wishes. Data is used in various scenarios such as analytics, modeling, simulation etc. To support special work in special areas (for example doctors and clinicians) one needs to be able to react quickly, i.e. research needs fast decision structures where possible
- Long Term Accessibility:
  - In some disciplines (such as genomics) it is well-agreed that large repositories store key data that partly cover the whole data cycle from raw data to derived data. For scientists this can have enormous advantages, since high throughput data creation is beyond what scientists can deal with in local infrastructures. However, in many cases it is not clear whether the repositories can give commitments for long term access which needs to be solved urgently.
  - In many sciences experiments or observations cannot be repeated making at least part of the data so valuable that it needs to be stored for a long time. In many cases the value of data can only be realized after many years by new generations. We have a responsibility, but are not ready for this.
  - Errors of all sort can slip in into chains of operations with data and can emerge to patterns that are misinterpreted. The documentation of such errors when observed has a high relevance to maintain high quality of results.

B.2 Sharing and Publication

Questions asked:

- Who else besides you has access to (subsets of) your data?
- How is that handled (access control, repositories, licensing, integrity, privacy, ...)?
- In which way can others discover your shared data?
- Do you register them somewhere?
- Is there a well-defined and stable way to refer to them?
- Do you aim for long-term availability of your data?
    - If so: how do you do that?
- Do you ever formally publish data? If so: where and how?
- Is there sufficient information available to 3rd parties to reproduce your analysis?
- If you do any of the above: do you get appropriate credit for that?
    - If not: what do you think has to change in order to alleviate this?

Statements collected:

- Funders' Statement:
    - Effective data management in science and its accessibility is indispensable to improve result quality and increase trust level.
    - Five principles science should adhere to:
        - data should be discoverable: requiring open metadata;
        - data should be accessible: requiring proper repositories and openness;
        - data should be understandable: metadata etc must support interpretation;
        - data should be manageable: requires data organization principles and proper workflow mechanisms;
        - we need educated experts: currently too much management work by scientists;
    - We need metrics for impact of data.
- Principal Issues:
    - There is still doubt about whether scientists like to share their data. There is much talk about a win-win situation when data is being shared, but there seem to be big differences between the disciplines in being convinced. Certainly there are areas where scientists see the competitive advantage when not sharing their data.
    - The funders increasingly demand openness of data which has been created by public money and there is often also a social aspect why data should be made open. In general the pressure on openness of data is increasing.
    - In many areas also industry has an interest in data with the goal of at least indirectly making money with it.
    - Data creation, gathering and curating and building tools to create and analyse it means often years of work. How to motivate people to invest all this time when they are requested to make data open after short time periods. What are appropriate mechanisms for giving credits, what could be successful carrots?
    - There is the fear that other scientists "profit scientifically" from open data at the expense of the creator who invested the time. Data thus is seen as the first step of creating scientific evidence. Do embargo periods help and if so, how long should they be?

- Publishing data only makes sense if people can read and interpret them. In some areas widely used formats have been agreed on that facilitate effective sharing, in many cases effective openness is not given due the lack of standards.
- In particular when sharing data across disciplines it is a question how effective sharing can be achieved. Do we need services on data etc.?
- Some communities have built a “fair” system of data exchange for a long time already which is working given the trust that everyone behaves in the same way.
- It is not always clear how to best achieve sharing with access options for a longer period. The roles (repositories, journals, etc.) are not yet clearly defined.
- Proper citing of data is not yet established as a general principle.
- Given that the scientists are willing to share, do we have appropriate mechanisms to “publish” data so that it can be appropriately cited?

- Special Aspects:
  - Often copyright etc. on data prevents sharing. In all areas where data is being created by companies (newspaper texts, etc.) the situation gets even worse and for science there are no special rules. Sometimes it is not obvious how sharing could be done without breaking laws. Only in some areas clearly identified repositories have been established, but it is not always clear what their funding basis is.
  - There is much talk about how to assign persistent identifiers to data objects to ensure long-term citability. Different practices have been established. Is the way towards a worldwide available and scalable system for registering PIDs (that is comparable with the IP number system to let computers talk to each other) a way out ?
  - In some disciplines ethical issues are much more severe to be considered than legal issues, since ethics is closely related with trust if the involved people.
  - In many areas (patients, human subjects, etc.) anonymity and privacy is an enormous challenge for data sharing - do we trust the available mechanisms. Do we trust all stakeholders in not misusing the data.

B.3 Infrastructure Demand

Questions asked:

- On which IT-infrastructure do you rely for data management and sharing (local, remote data centre, ...)?
- Do you consider the IT-infrastructure you use well prepared to keep up with your expected needs? If not: what's missing? where are bottlenecks?
- Is outsourcing some of your storage or computation needs an option for you?
- If so: what are the constraints (cost, privacy, location, legal, ...)?
- Who are potential providers for you?
- If not: why not?

Statements collected:

- Data Organization Aspects:
  - There are many aspects of data organization that are not solved and thus hamper easily integrating data:

        - granularity of what a dataset is and how it should be identified is not defined
        - do we fully understand the relevance of persistent identifiers
        - very often the nature of metadata is not defined
        - it is not clear which data constitutes an intellectual value in science
    - In science data will increasingly often be automatically generated by workflow scripts. It is not at all clear how we automatically create identities of data objects and describe them with provenance metadata tags so that we can trace histories of data. Also if data is being created manually these questions have not been addressed.
    - Often data annotation standards are not suited to the deep scientific applications
    - If we agree on some systematic approach it is not obvious who is responsible to implement these mechanisms to make them happen.
    - Databases in science will often be dynamic in the sense that the content of an object is changing frequently at asynchronous moments. How to identify and describe such objects, since they will be used to create scientific evidences and thus need to be citable.
- Repositories:
    - Repositories have an important role in making data accessible. However many questions need to be addressed as well.
    - In some disciplines national and international centers are respected repositories to store various types of data being generated in scientific workflows. It is optimal for science that there are agreements with respect to mirroring, access policies etc.
    - There are many places to get rich data and tools, however, existing repositories are not always suited to the nature of the data and sometimes stop with their services leaving big gaps for the scientific domain - is this good or do we need better portals or better supported repositories.
    - Repositories need to have a long term perspective to ensure researchers about accessibility of data.
    - Are repositories that have working data also archives of data, do we need to separate these two in digital era. Do we need to separate the functions of computation and data centers? If we separate then data shipment via slow networks can become a hurdle for science.
    - Who is responsible for data stewardship - is it a task of institutions/ disciplines, is it a private responsibility or is it the task of the repository experts?
    - In some countries there are serious efforts to come to national information/data infrastructure. These are often limited to specific disciplines dependent on their economic relevance. Also at the EC level there is funding for data infrastructures. How does this work together, how do we need to integrate the efforts?
    - Should long term storing/preservation being centralized per organization or discipline. There is clearly an economy of scale factor when data availability is being centralized. But what about support for data then, since knowledge about the content is often in the departments?
    - Some journals ask for depositing data into recognized and trusted repositories when a paper is being published. Do we have such repositories that are “persistent”.
    - Centralized repositories can issue strict policies which in the end are valuable for the community.
    - The advent of cloud systems may require rethinking strategies.

- Metadata/Portals:
  - many initiatives harvest metadata from many sources, build portals and allow to search for useful data - it is important that these metadata records support immediately click-through to the data objects and even through history of data by using the provenance information
  - does it make sense to harvest metadata within large disciplines or even across disciplines - looking for useful data can lead to looking for a needle in a haystack - where is the balance
- Services/ Visualization/Apps:
  - Software to perform analysis, visualization in distributed environments is getting increasingly complex. How can we manage this complexity in the long run - do we need experts who know how to use these tools and how adapt them where necessary?
  - It is widely agreed that on large data sets services on data is a good option towards users. However, who can pay for the maintenance of the software and which choices can we make given that in science there are so many different use cases? How to optimize this?
  - If we decide to create services to access data we are creating some inflexibility again - how to support the easy combination of different data sets?
  - High performance computer systems do not allow fast decision making in science, since CPU time needs to be requested early enough. This static scheme is against scientific dynamic.
  - Visualization methods are very important to understand the content of data. Yet visualization in science is very much behind the state of the art which is defined by the gaming industry for example. How can we close the gap?
  - Do we fully understand the role of apps and mobile devices for scientific applications with the possibility to virtually combine data etc?

B.4 Technology Trends

Questions asked:

- Where do you see your field heading with respect to data handling at large?
- Which technological trends that you see do you consider promising for your field?
  - What would be their expected impact?
- Is using cloud technology an approach that is applied or discussed in your field?
  - If so, what is the current state of affairs there?
- Do you think that you/your team/your collaborators have the capacity to follow current technological trends? Should you/they even?

Statements collected:

- In some scientific fields the agreement on data formats has proven to be extremely useful and improved data work enormously.
- Moore's Law for computing, and Kryder's Law for storage, are coming to an end (exponential growth cannot be sustained). Does this mean that the "golden age of computing" is over and that future progress will be much slower than in the past 50 years?
- In the past few years, a new generations of equipment such DNA sequencing platforms based on different methodologies has become a "disruptive technology". The availability of these new generation technologies raises many questions for the research community.

- What is the potential of the cloud - can we easily deploy new operations to be carried out on data? Is it a solution to the big new challenges - Big data, large computations.
- How can we optimally use clouds? From a data perspective they reduce the data model to a very simple one and by technical solutions such clouds can scale up even for huge amounts of files. From a computational point of view automatic parallelization (map reduce schemes) can be applied to also scale up operations.

B.5 Education/Efficiency/Cost/Education/Roles

Questions asked and statements collected:

- Whatever we do with data and software - who pays when it goes beyond the normal traditional research task?
- Working with (large) data and accessing it is still very inefficient in all respects.
- In general scientists will need a lot of guidance and support from experts to make working more efficient. Often support from centers is minimal, since it is cost-intensive.
- Can we transfer knowledge from one discipline to another? There are examples such as Grid that completely failed, since the HEP model was transported to other disciplines. Isn't the case that every discipline needs to go through a number of infrastructure steps to build up the required knowledge within the discipline?
- Should PhD students be trained in these, or should the "housekeeping" be left to professionals, so that the PhD students can get research and publication done?
- Is it satisfying to rely on user friendly interfaces and services on data instead of educating your own experts who can do programming etc.
- in many countries one can buy storage and computation facilities, but there is little money to train and keep experts to manipulate data and we lack trained people.

B.6 Stakeholder Aspects

Questions asked:

- Who else besides you has a say in how your data need to - or must not - be handled?
- What are the demands arising from this?
- Do you get sufficient support and guidance to meet these demands?

Statements collected:

- Funding agencies (with good intentions) might insist that data and software be made open. After all, it is typically taxpayers who cover the costs. But how can the best scientists be motivated to invest the time, if others "profit" scientifically "at their expense/ See my first question above.