RDA Europe
Data Practices Analysis

Edited by Herman Stehouwer & Peter Wittenburg

Contributions from Rob Baxter, Diana Hendrix, Eleni Petra, Fotis Karagiannis, Daan Broeder, Marina Boulanov, Leif Laaksonen, Francoise Genova, Gavin Pringle, Rob Baxter, Franco Zoppi, Constantino Thanos, Herman Stehouwer, Peter Wittenburg, Giuseppe Fiameni, Natalia Manola

RDA Europe (312424) is a Research Infrastructures Coordination and Support Action (CSA) co-funded by the European Commission under the Capacities Programme, Framework Programme Seven (FP7).

1 Executive Summary

Scientific research has always been about observations transforming into data and from data to new insights. With the increasing possibilities enabled through modern technology it is possible to make observations increasingly fine-grained in time and space, paralleled with a gigantic increase of observation diversity and volume – be it by sensors or by humans. Utilizing the possibilities to simulate aspects of reality on powerful computers and combined with our increased capability to apply smart algorithms on this data we are also extending the amount of data and their complexity. While the extraction of scientific knowledge from all observations in former times was mainly an intellectual effort, we understand that due to the trends we need to use new computationally supported methods to extract scientific knowledge. Also, combining data from different sources in various ways will bring us new insights about natural and societal/cultural phenomena. We realize that "big data" is not just about higher quantities, but that "big data" is also a new quality in itself. J. Gray called this the "Fourth Paradigm: Data-Intensive Scientific Discovery[1]".

Quite a number of publications such as "Riding the Wave[2]" from 2010 have been written to demonstrate the challenges and opportunities coming with the data deluge, on the potential societal and economical value of the data and on the need of building bridges. Of course computer science has been working on advanced concepts to address some of these challenges. Europe is currently funding a broad range of research infrastructures to address data integration and interoperability issues with high priority. Europe is also funding e-Infrastructures that address and promote cross-disciplinary data related aspects and at an international level the Research Data Alliance (RDA) has been established to construct bridges of many different types to overcome the data interoperability barriers hampering data diffusion and reuse. Yet, however, there was no broad cross-disciplinary study of what the various problems are the research communities need to deal with to implement a seamlessly accessible and interoperable sphere of data that not only scientists, but also citizens can use. The overview documented in this paper based on about 50 interviews, more than 70 interactions at various community meetings and the results of a first RDA Science Workshop wants to fill this gap.

All relevant aspects brought forward in the interviews and interactions have been compared and classified into a number of “observations” described along a process model that seems to underlie the data operations being carried out mostly implicitly or explicitly by the researchers in the departments. These observations can be seen as describing "data practices" and are aggregated conclusions based on (1) the relevance of ESFRI and e-Infrastructure projects, (2) the consequences of the Open Access initiative, (3) trustworthiness of data as a key in the anonymized data domain, (4) the huge problems with some legacy data which are still being repeated with some new data due to inappropriate methods, (5) the challenges of Big Data and data management asking for new highly automated methods, (6) the challenges with creating and aggregating proper metadata and the lack of explicitness hampering progress, (7) the trend towards centers with well-established certified repositories with a long-term perspective, (8) the need to educate a young generation of data professionals, (9) the general lack of trusted information on services and (10) the need for a grass-roots organization such as RDA to address the challenges.

[1] http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf
[2] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

From these observations a number of concrete recommendations are derived that will help to speed up the process of changing our way of dealing with data in the various research communities. This increased speed will be determinant for the competitiveness of European science and in a second step the economy in general.

Therefore we need action in the following six areas:

(1) strengthening the education and training efforts;
(2) establishing a trusted open market place for data and services;
(3) carry out design and implementation studies on the data fabric to demonstrate usefulness and potential;
(4) fund a set of curation projects to integrate legacy in a convincing way;
(5) push structuring the landscape by establishing trusted and reliable repositories with sustainable funding; and
(6) make federation technology mature so that everyone can easily create integrative platforms.

TABLE OF CONTENTS

2 Introduction

One of the major action lines within the European iCORDI project (now called RDA/EU) was the analysis of the current data landscape in the various research communities and disciplines. This was seen as one of the core sources of both motivation and opportunity to kick off concrete activities within the RDA context. We feel that this process was indeed clearly helpful and some urgent and fundamental issues that stemmed from data analysis are consequently being addressed within RDA groups. The first deliverable from this activity was written at an early stage and was therefore based on a limited number of interviews within iCORDI[3]. This follow-up final deliverable is built on:

- 24 Interviews done in iCORDI;
- 16 Interviews obtained from the EUDAT project on understanding communities' data organization;
- 9 Interviews obtained from the Radieschen[4] project (a German-funded project);
- Interactions at more than 70 community meetings, many attended by the editors[5];
- The results of the first Science Workshop Organized by RDA/EU in collaboration with the Max Planck Society (see Appendix A).

The combination of these five sources of information gives us access to a large amount of information on data practices in many different scientific disciplines, in different organizational contexts, in different initiatives, and at different maturity levels of the data lifecycle. The results of this analysis potentially have a substantial impact on the work of the RDA, and also on the design and funding of research infrastructures. It should be noted that this report is meant to give insight to data practices as they are currently used within the research communities and that it is not meant to indicate possible new concepts and ideas from emerging technology research[6].

However, even though we have achieved broad coverage we cannot claim to be comprehensive in our description of data landscapes and organizations. There are two major limiting factors: 1) there was only a limited amount of time available for each interview and interaction, i.e. not all aspects could be covered in great detail; and 2) the conversion from interview to interview transcript and from interview transcript to extracted observations had to be done manually, i.e. it is influenced by the interviewers' and editors' biases.

Before continuing, let us briefly outline the method chosen to come to what we call "observations" in Chapter 5:

- A group of people (contributors, editors) interviewed community experts guided by an underlying questionnaire.

---

[3] Please see Appendix C for a list of interviews.
[4] Radieschen: http://www.forschungsdaten.org/index.php/Radieschen; Radieschen final report: http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:124265:3/component/escidoc:124264/ProjectRadieschen_Synthesis_EN.pdf ; the authors also had access to the interview texts which however cannot be published.
[5] Please see appendix B for a list.
[6] We often received the comment that a certain topic has theoretically been sorted out. While this may be the case, that is not the scope of this report.

    - The reports of these interviews are openly available for further reading.
- Additional interview reports where collected from the EUDAT and Radieschen projects.
- The authors and contributors extracted key points from across the interviews and from notes from additional interactions with community experts at various community meetings.
- The key points were aggregated, classified and combined, resulting in the observations.

One interesting point to note is that interviews and interactions with industry/companies did not prove to be very useful. We conjecture that the main reason is that companies tend to argue that they can do everything, have the know-how about all knowledge and possess ready-made platforms. What is often ignored is the fact that software technology and expertise can solve many problems, but there is a price that has to be paid. Pre-existing contracts for example for infrastructural services hinders progress in terms of adaption of new worldwide standards. It is evident that dependence on commercial solutions has significant consequences, namely the solutions cannot easily be changed once adopted.

The structure of this document is as follows. In Chapter 2 we present the conclusions that came out of the Science Workshop that took place in February 2014 in Munich. In Chapter 3 we summarize the results from our first deliverable which was in many ways preliminary. In Chapter 4 we present all observations that we could extract from the available material at this stage. We summarize them to a number of key observations in Section 4.3 and compare the observations with the current activities in RDA working groups and interest groups in Section 4.4. In Chapter 5 we conclude with recommendations.

3 Science Workshop Recommendations

RDA Europe, together with the Max Planck Society, organized a workshop involving leading European scientists with a broad interest in data. The goal of this workshop was to understand which opportunities, challenges and concerns researchers have in relation to research data while conducting their research, both currently and in the future. For a detailed report, and a list of all participants, please see Appendix A.

The two-day workshop fostered exchange and interaction on a wide range of topics that included Sharing and Re-use of Data, Publishing and Citing Data, and Infrastructures and Repositories. These discussions enabled the identification of a number of issues viewed as essential in helping to achieve the RDA vision of researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.

Below we go through each of these areas in order and briefly summarize the outcomes for each. Major recommendations follow at the end of this section.

3.1 General Observations

It is very clear that the many new possibilities in data generation are at the source of a number of major challenges. We need smarter algorithms, processes and automated workflows in order to keep on top of the generated data. At this point our ability to generate data far outstrips our ability to process data.

When dealing with larger volumes of data, we need more systematic solutions to process the data in order to have reproducible science. By that we mean, systems that need to cater for the disciplinary and multi-disciplinary approaches inherent in modern scientific practice. We note that when taking in multiple types of data with many differing properties, processing leads to a complex adaptive system where sociological hurdles play an important role. Currently a considerable effort is spent on reusing and combining different data sources.

It is clear that in order to deal with the increasing complexity and cost of combining data we need automated workflows that can cope with increasing demands for sharing, combining, staging, and processing data. Currently many solutions used are very situation specific. We need to stop relying on such one-use solutions for data exchange and interoperability for this to work.

3.2 Sharing and Reuse of Data

Reuse and sharing of data are problematic for a number of reasons. One reason has to do with our inability to explore/find/collect the data, i.e. lack of visibility due to insufficient descriptive metadata, or lack of inclusion in catalogues that are used by search engines. Other reasons arise from a lack of cross-discipline methods that scale, and data mapping difficulties – in other words, the lack of common ontologies and vocabularies. A further complicating issue is that of trust: can you trust the identity, integrity, authenticity and seriousness of all actors involved in the production chain?

3.3 Publishing and Citing Data

Publishing results and their citation is at the core of the scientific process. Due to the increasing relevance of data, data needs to become a first class citizen, i.e. data publications need to be impactful.

In order for data citation to work, the appropriate mechanisms need to be stable and in place, i.e. using worldwide accessible and interoperable PID systems. In order to be able to retrieve data at any point in time a stable infrastructure *must* be in place that makes not only the identifiers, but also the data and attributes of the data (metadata), available. This requires a considerable cost.

3.4 Infrastructure and Repositories

We need infrastructures that interconnect more seamlessly and more efficiently; however it is not clear how to get there. In order to encourage use the components of such infrastructures need to be trusted, reliable, findable, accessible, and (to an extent) interoperable.

We support open access as a general principle.

A large advantage is to offer services on the data, and not the data as such. However, these services need to provide alternative views on the data and not restrict usage of the data.

It is hard to find which infrastructures are available; we will need registries and catalogues in order to find the services that we need. Existing repositories and infrastructures will need to be integrated in such registries and catalogues. Infrastructure needs to be integrated to interoperate.

3.5 Conclusions

Overall, the Science Workshop drew the following conclusions. These conclusions reinforced the key ideas behind the RDA and have also been taken up by several new working and interest groups.

- Infrastructures must work and be reliable, persistent, and sustainable, i.e. the infrastructure must still function in the same manner after a number of years.
- Scientific work must be reproducible.
- Credit must be given for work on data and this credit must be valued in researcher career advancement.
- Data must be citable.
- *Provenance*, *validation* and *trustworthiness* of resources must be assured.
- Infrastructures must be trustworthy and this trust must be earned.
- Effort must be dedicated to training the new generation of data scientists.
- There is a leading role for RDA to play in defining and supporting recommendations that facilitate the creation and use of data infrastructures.
- Demonstration projects using the first results emerging from RDA must be set up.
- RDA should give advice on data management, access, re-use, guidelines, APIs, etc.
- Help from RDA in matters of Metadata (such as metadata standards) is appreciated.
- A forum, such as this one, bringing different disciplinary representatives together is essential.

For a full description of the workshop and the detailed recommendations please see the official RDA Europe document “Report on the RDA-MPG Science Workshop on Data”. A summarised version is included in Appendix A.

4 Analysis Programme Recommendations

The RDA/Europe Deliverable D2.4: "First year report on RDA Europe analysis programme" of 2013 elaborated on a number of recommendations and observations based on a set of initial interviews with research communities regarding their data management practices. The analysis of the interview reports showed that we still had a long way to go before good data stewardship is commonplace. Furthermore, it underlined the importance of having good metadata. Good metadata enables discoverability and reuse of the data.

Based on the analysis in the report we made a number of concrete recommendations, reproduced here in short form:

- Use basic data management practises as outlined in the e-IRG whitepaper[7]
- Consider and plan for data management *before* data collection takes place;
- Document data with high-quality metadata;
- Use persistent identifiers;
- Ensure discoverability of the data;
- Employ/implement data sharing policies;
- Educate researchers on the available services on cloud computing, grid computing, HPC, etc;
- Use computer enforced data management policies.

[7] http://www.e-irg.eu/images/stories/dissemination/white-paper_2013.pdf

5 Observations

5.1 Process Model

The process model in Figure 1 emerges as the dominant underlying process model that data practitioners are implicitly using when processing data, and it is used as a reference model in this report. This model is based on existing models of data[8] and the observations made in the RDA/EUROPE analysis program.

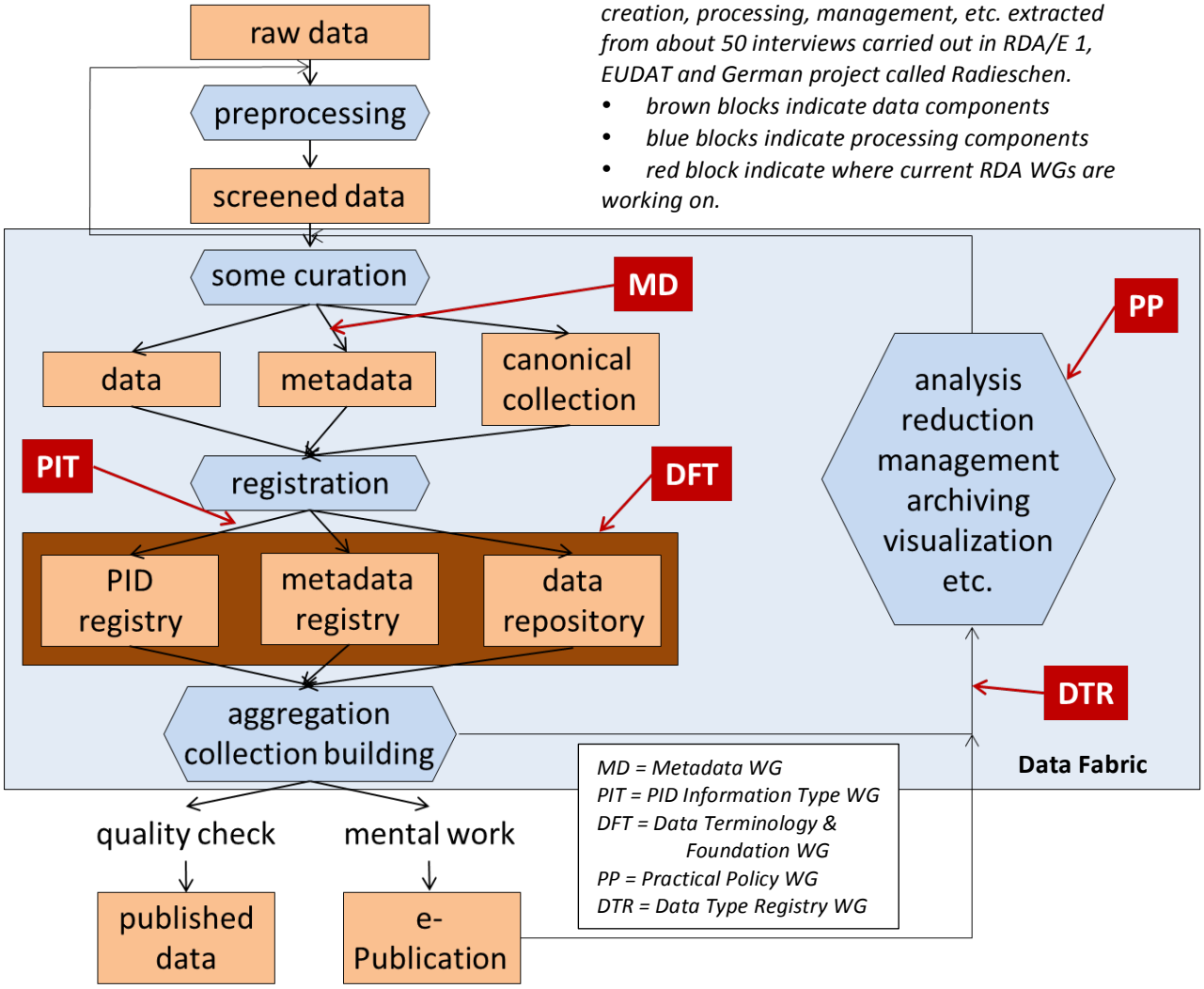

Figure1: Reference Process Model for data creation, processing, management, etc. extracted from about 50 interviews carried out in RDA/E 1, EUDAT and German project called Radieschen.

- *brown blocks indicate data components*
- *blue blocks indicate processing components*
- *red block indicate where current RDA WGs are working on.*

[8] We base ourselves here on the following data models: Kahn/Wilensky 2006, ResourceSync, CLARIN, EPOS, ENES, ENVRI, EUDAT core model, ORE, Europeana, OAIS, Datacite/EPIC, and DICE (as used by iRods).

Even though this model constitutes/describes a generic process we must note that:

- Some variation occurs in practice.
- Implicit hand-crafting with ad-hoc solutions is common.
- There are some specific process models that only describe part of the generation cycle[9].

The main purpose of this reference model in this document is to help in grouping the observations, and it helps to identify certain steps as they are applied to data. Data is:

- Scientifically meaningful and relevant after the pre-processing step.
- Ready for upload to a repository after the curation step.
- Ready for re-use after the registration step.
- Ready for citation after the publishing step[10].

Currently most researchers do not distinguish between these steps explicitly, which is one of the reasons for the inefficiencies and increased costs when dealing with data, in particular when data needs to be used long after its creation – researchers tend to forget details about what the actual processes that led to the particular data where. The Data Entropy diagram (by B. Michener of DataOne and University of New Mexico indicates the usual decrease of knowledge over time. Dependent on the particular purpose, other models are being used such as the LifeWatch model resulting from the ODIP work which separates in Acquisition, Curation, Access, Processing and Support and which can be mapped easily to what has been extracted from the interviews[11].

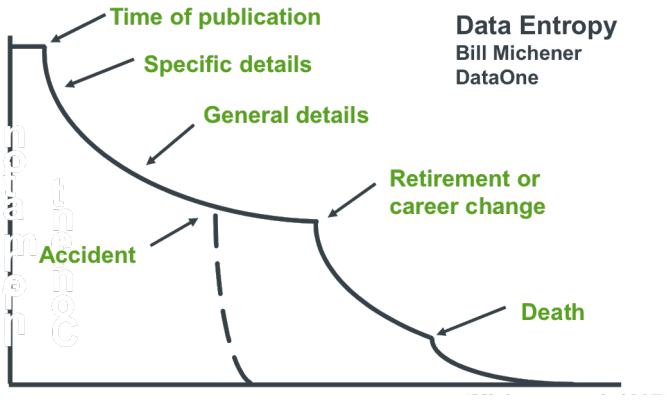

Figure 2: Data Entropy as described by Bill Michener

As can be seen from the model, management and analytic-type operations on data (collections) are included in one block. The rationale is that whatever operation is being selected it will change information about collections, will create replicas for different purposes (new instances) or create completely new collections with derived content. Both types of operations are part of what is now starting to be known as the "data fabric".

In the figure we also indicate the aspects being addressed by the five start-up RDA working groups, namely 1) Metadata Standards Directory, 2) PID Information Types, 3) Data Foundation and Terminology, 4)

[9] The HEP Tier concept for example describes how the huge amount of primary data sets from LHC experiments are pre-processed and then distributed to subsequent tier nodes to allow researchers to process the data.

[10] The publishing step in general requires validation and curationcuration, which is why we distinguish it as a separate step.

[11] See http://www.lifewatch.eu/web/alien-species-showcase/architecture for some details.

Practical Policy, and 5) Data Type Registry. A more detailed analysis of how RDA working group work fits in the data fabric can be found in Section 5.4.

5.2 Observations

5.2.1 General Observations

Many fields of research are changing rapidly, driven by the availability of data and thus by being able to make use of computational paradigms. In this process of fast changes scientists are often stuck with methods invented in a moment of need, that now hamper the search for and adoption of better solutions. One of the main reasons for this is that scientists like to stick with a solution that they are used to, since the pressure to publish results is enormous. In some cases, excellent data organization schemes have been worked out, but are not put into broad operation. Often there is a dearth of knowledgeable experts and of time/funds to change habits. Huge stock is placed on legacy data in almost all scientific domains, and inadequate methods and software tools are adding even more legacy data.

Often in daily practice the steps indicated in Figure 1 are carried on without any clear transition from one to another; this creates a lot of inefficiencies in dealing with the data at later steps. Data stored in local stores is being changed, re-generated, used to create other data etc., but there is no track of what has been done and what the relationships are.

Still people rely on their minds as metadata and relation stores, which does not work, especially when researchers leave a team. There is a clear need for *reproducible data science* and an abstract willingness to change habits and attitudes, but a general lack of experts and time prevents changing this.

Scientists often take an egocentric view and make an implicit estimate as to whether the creation of proper data documentation will be profitable compared with the overhead required to manage and access data as is. Given the general time pressure one cannot expect that curation is being done for others if there are no external motivations or pressure.

Only in a few communities can one find well-defined and widely accepted data organization concepts with a clear notion of what Digital Objects, their management, and processing are. Often when data is being created scientists continue to invent their own data management solutions, resulting in a large variety, which ultimately hampers easy interoperability. The difficulty is in part not a technical one, but can be found in testimonials such as “people who created solutions leave”, “software cannot be changed when problems occur”, “software cannot easily be extended to fit the state-of-the-art methods”, etc. Often communities invent their own schemes. One of the driving reasons for change will be the pressure to come to a method of reproducible data science.

Researchers and data practitioners in scientific communities see a clear need to reduce the heterogeneity of data management solutions. This trend is depicted in Figure 3, which illustrates the natural variety in data creation and analysis methods on the one hand, and the possibility of reducing variety in management methods given that proper data organization methods are being applied. The current heterogeneity is increasingly hard to justify funding and since we now understand the basic, common mechanisms of data management that apply to the majority of data infrastructures, a reduction in the number and complexity

of solutions will not hamper scientific progress. This is also obvious from the statements released by different fora w.r.t. data management, including the G8 Science Ministers[12] and e-IRG[13].

A clear trend can be seen towards *federations* of trusted centres and repositories within disciplines and/or scientific domains. The persistence of these centres is not yet ensured although communities are coming to rely on the availability of their services. The exact topology and task sharing within these federations depends on the organization of the communities served and the nature of the data and processes used. Obviously these networks of centres are drivers for structuring the data and service landscapes in the communities, partly on a global scale. The establishment of trust is being done by several means: some centres simply offer a special highly appreciated service, while others are undergoing formal certification steps; ideally the future will see a combination of both approaches. Trust in data centres when storing core data (not just published results) is often very much related to cultural and legal proximity between community and centre (for organizational or national reasons, for example). The lack of persistent funding for such centres is, in many cases, a serious barrier to acceptance. Some centres focus on analysis technology instead of data where it is more important to maintain the knowledge in the team.

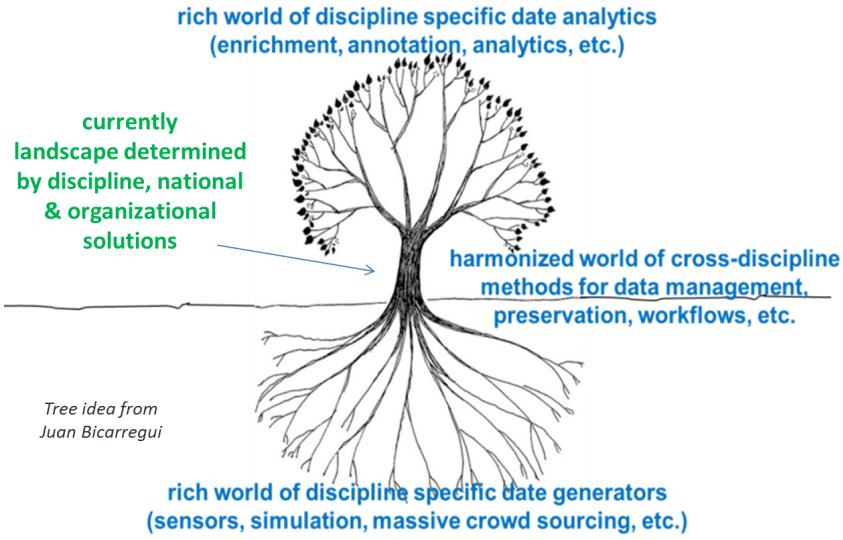

Figure 3 indicates schematically the assumption that despite all discipline-specific differences of data creation and data analysis methods the data management methods are widely discipline independent and can be reduced to a number of solutions that reflect special characteristics of data sets.

Large projects, which create huge amounts of data, have developed relatively fixed preprocessing and distribution workflows (pipelines) for the raw data. In most cases no responsibility is taken for derived data that is created in a distributed way, i.e. only parts of the data processing steps are covered and documented, and thus the methods cannot be transferred.

Big data[14] is new for many of the science departments and they are, in general, ill-prepared for the challenges. The main reasons are limited data management methods, and lack of knowledge and

[12] See the statement here: https://www.gov.uk/government/news/g8-science-ministers-statement.
[13] http://www.e-irg.eu/images/stories/dissemination/white-paper_2013.pdf
[14] In this document Big Data is not meant in terms of the volumes of data being used in computations, but is referring to the method of integrating data from various sources to relate for example phenomena observed with patterns to be found in data collections.

experience that are challenged even further when cross-disciplinary data needs to be integrated and processed.

Several departments have started to work on data processing workflows, even though further flexibility in processing steps and parameter control is required. In a paradigm[15] that uses a variety of learning and analytical algorithms to relate phenomena with patterns in data, flexible workflow systems have proved to be an advantage in speed and efficiency that motivates researchers to invest considerable effort in developing them. If this is paralleled by proper data organization concepts, reproducible science comes for free.

5.2.2 Raw Data

"Raw" data in the science domains is generated from sensors, simulations, observations, massive crowd sourcing, sequencers and other types of experiments. Sensors and simulations are currently creating increasingly large volumes of data as are crowd sourcing and user generated data (Web, social media, etc.). However, one of the problems is that it often seems that, after some preprocessing and reduction, the huge volumes of "raw" data are not further touched. They are often stored on external discs or tapes and disappear into cellar cupboards; a good example is image data in genomics after transformation into textual data.

Sensor and simulation data: Sensor and simulation data are largely generated in well-defined formats; for storing them, simple structures such as file systems are used. A major issue is the accompanying descriptive metadata, such as the instrument or model configuration, and how this is linked to the raw data. Often this metadata, is part of the file headers (for example, DICOM formats), while in other cases, e.g. when the sensor equipment or simulation environment is more complex, separate metadata files about the configurations are generated and linked to the data in an ad-hoc manner. It is mostly up to the researcher to maintain the relationships – in most cases no explicit infrastructural processes are used, but rather separate spreadsheets, databases etc. For simply-structured data, file systems seem to be adequate, since relevant information is represented in file and path names and data structuring is done through directory hierarchies. This often leads to severe problems over time (e.g. when people leave a research group), since contextual information has not been made explicit. Therefore, we see a clear trend towards thinking about alternatives to pure file systems.

Another big challenge in data management and access is the sheer transmission of huge amounts of data from sensors to data centres, where transmission packets from sensors arrive at unpredictable times at data centres, while the data analysis needs to begin upon data arrival.

User generated data: Much of the volume of raw data is generated by a large number of individuals, such as in the case of clinical data where doctors create valuable data on patients, or in research data crowd sourcing environments – an increasingly used paradigm. Such data can also amount to "big data" and its main characteristic is that it is less well structured, resulting in a data fragmented landscape, an increase in

[15] Such workflow paradigms are in use in several places such as at CERN and at the group of Barent Mons at Leiden University.

data types, and also in the creation over time of orphan data[16]. In the area of crowd sourcing many different technical solutions are being developed, most currently with ad-hoc approaches to data management. Much of the data created this way is unstructured (CSV, XLS, PDF, etc.).

5.2.3 Data Preprocessing and Curation

While preprocessing is a necessary phase to be able to work with generated (as opposed to raw) data, curation is often required to make data available for use by others. Researchers are in general reluctant to invest in curation of their data, since it costs time and often does not add to their own scientific career building. In this report we separate the curation phase from the preprocessing phase, probably in an artificial way, since some of the preprocessing operations (such as quality enrichment) could also be seen as part of the "curation" activity. The "curation" subsection mainly refers to metadata improvement.

5.2.3.1 Preprocessing

There is a wide heterogeneity of preprocessing operations being carried out, dependent on the data that has been generated. Most often preprocessing is guided by some ad-hoc scripts or, as in big installations, by code pipelines. Still in many disciplines preprocessing is done manually due to i) some special treatment required for specific data collections, and ii) a lack of expert time to turn practices into flexible, parameter-controlled workflows.

Key preprocessing tasks include:

- Documentation of the kind of operations on data that have been carried out (provenance); relations between data objects and preprocessing software components should be stored. In general it seems that communities are unprepared to deal with these two requirements (storing of provenance and relations) and use ad-hoc or widely unstructured methods.
- Quality control and error "treatment" of data in a variety of ways (which seems to be becoming more and more important), including: normalizations with the help of reference data sets; reduction by transforming patterns in one domain (images) to patterns in a highly reduced domain (texts); specific grouping of parts of large data sets to accommodate specific views, and in doing so also reducing the amount of data; anonymization or pseudo-anonymization of data sets; format transformations based on rich transformation libraries.
- Processes that improve quality and clarify IPRs, as found in traditional institutions such as museums, archives and libraries, but also encountered in clinical applications. These quality aspects, as well as the harmonization of quality norms throughout Europe in cross-border projects are a prerequisite for carrying out joint operations.

In the case that data from different sources and even disciplines needs to be integrated to allow "big data" type of computations, for example, much transformation work typically needs to be carried out. We summarize this work later under "collection building".

Naturally, there is some preprocessing that is so compute-intensive that data needs to be transferred to the input workspaces of HPC machines.

[16] With orphan data we refer to data that is in existence but has no longer any clear origin, project, or owner.

In many cases the preprocessing step results in “canonical collections” of data such as all files that belong to a certain experiment, all files that are created by one specific simulation, all files that belong to a specific observation (same day, same place, etc.) etc. In principle, these canonical collections are most often the basis for data management (see Sections 5.2.5/6).

In some cases where huge data amounts need to be handled the pre-processing and the dissemination steps are part of standard pipelines. As an example we can refer to the tier concept for the data treatment in the LHC experiment.

5.2.3.2 Metadata Curation

As has been shown by the DataONE overview[17] for Earth observation laboratories the situation of metadata is not at all satisfying: most laboratories don’t use an explicit structured metadata approach; often Excel files or similar widely unstructured formats are used for metadata, which in general can’t be interpreted by machines, and are difficult to interpret by others, or after some period of time. With few exceptions provenance metadata that capture the creation history of derived data is not being used in the communities.

For many the metadata concept is new since until recently researchers have relied on the “self-description” contained in file and directory names. The awareness that these traditional methods are not sufficient is growing, however the step to change practice is hard, since it requires additional software and additional efforts that many do not want to invest in. It should be mentioned that in particular research infrastructure initiatives, such as those started in ESFRI, have contributed a lot to raise the awareness around these concepts. The notion that metadata needs to be open for everyone has been spread and is widely accepted.

Only very few communities have a comprehensive metadata solution in place, but even then many of the community members do not have the capacity or the will to adhere to the norms. In case they are using a schema, it is seldom that the vocabularies which they are using are explicitly defined in registries, i.e. machine based semantic interoperability or easy checks by human users are not possible.

In case of complex sensors such as telescopes or accelerators where parts of the instrument (even some filtering software) are changed very frequently, people create metadata records that describe the sensor configuration at that specific measurement times. These descriptions need to be separated from metadata descriptions that describe the overall experiment or observation, however: the latter metadata should point to the sensor metadata, which is still not realized in an explicit way.

Metadata creation is still a largely manual task that raises the barrier for researchers to create quality metadata. Companies building sensor equipment do provide facilities to create metadata from the beginning that can be included in headers: information in photos or in the DICOM scanner format for example. This metadata information can be extracted and inserted into the metadata descriptions of the created data. We currently lack tools and setups that allow researchers to enter all required and useful metadata as defined by the communities from the start.

[17] http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021101

In communities that work with commercial software vendors or where we see a high turnover of their software development experts we can see enormous barriers to adapt their way of working. Software vendors will not invest in changes if there is no market or the cost is too high. Internal software solutions (often relational databases) created by local experts that have left an institute become static source since the lack of knowledge about the software limits its future evolution and maintenance. The importance of these hurdles is always underestimated.

Often metadata, and in particular relational information, is not stored explicitly but is hidden in software code. Such methods form additional, severe barriers to changing practice.

In some communities the experts have worked out standards for metadata that show a high degree of flexibility, and thus representational capacity, with the unfortunate effect that it requires experts to make use of them. Often departments do not have the funds to hire separate experts, which prevents them from making use of these complex metadata solutions.

In communities that work with textual data (e.g. the humanities) the need for descriptive type of metadata is questioned. These communities are focusing much more on linked semantic domains that can be exploited at a semantic level by users[18]. These discussions often ignore the management aspects, since the texts are expected to be available on the web. This includes the assumption that the data providers will solve data management in some form, but the web by itself is no guarantee of persistence.

Some communities that possess sufficient technological expertise are creating a variety of different metadata formats, using them for different purposes, and using standard protocols to exchange metadata (such as OAI-PMH).

5.2.4 Registration, repositories & access

"Registration" in this report denotes the step at which a data object receives a persistent object identifier (PID) from a trusted registration authority and thus is referable. This should be accompanied by the step to upload it to a persistent repository. Registration thus means that the data is a PID can be resolved into a digital (data) objects and its state. This state information includes access information (in the form of permissions, conditions, and licencing) and information on how to obtain the object itself (usually an URL, with AAI, to a repository which will than respond with the data itself.

5.2.4.1 Digital Objects and PIDs

It is a troubling fact that there is little knowledge in the research communities about proper Digital Object modelling and its implications for data treatment.

Also the knowledge of PID systems and its relevance for data management and access is very limited, although awareness of such systems is increasing. The message has been spread that it makes sense to register digital objects and assign DOIs (a specific type of PID) for collections that are referred to by publications, but only in some communities has the message arrived that it makes sense to assign PIDs to

[18] Text analysis offers the possibility to create relations between fragments of texts and in doing so dense semantic weaving is possible which can be used to find useful texts, fragments and to support navigation.

data objects that are being created by workflows (or manually), so as to be able to refer to the object, check its integrity, etc. Consequently the knowledge about existing PID registration services is limited.

Some communities use identifiers internally, for example to refer to certain pieces of information such as a specific protein, and they do not see the essential difference in referring to such a domain entity compared to openly registered PIDs for data management and access.

Other communities use other typed of identifiers (e.g., URIs) to refer to Digital Objects. However "identifying" is more than referring to "something". A PID resolution needs to provide fingerprint, location and other information to prove identity, integrity, access paths, citation information, etc. A URI can also take you to a location where this information is being stored, but it doesn't provide the direct inclusion of information types that PID systems do. This issue is still being debated, especially with respect to the coupling of metadata.

5.2.4.2 Repositories

Despite the fact that almost all research infrastructures speak about a restructuring of the data and service landscape towards a network of trusted, long-term centres or repositories, there is little debate in many communities about essential issues such as trust-building by certification[19], repository guarantees of access to the same data object over many years, proper management procedures based on explicit, auditable policies, etc.

Due to a lack of commonly accepted data organization models, and thanks to tradition, there is large heterogeneity in data organizations and data management principles. This creates major hurdles and expensive solutions when it comes to data federation building.

It is now well understood that one aspect of trust building is the capability to guarantee long-term access, thus implying long-term preservation strategies. This is hampered for many of the nodes in the emerging landscape by a lack of long-term funding. Large libraries and archives in general have a long term funding assurance, but most do not yet fully understand these new requirements for offering data services. Compute centres that move into the role of data service providers often lack the sensitivity towards the data communities. Universities in several countries have understood the important role of data for enabling advanced science. We do see an emerging interest in investing in new types of data centres, partly by merging data centres and libraries to make use of complementary knowledge.

Trust is a very important criterion for researchers looking to deposit their data in a data centres[20]. From the many factors influencing trust cultural and legal closeness is of greatest importance, i.e. researchers from a specific country would first choose to store their data with a repository in the same country. Commercial offers are viewed critically, despite the fact that they are widely used because of their compelling service offers.

[19] Currently there are three approaches for certification: Data Seal of Approval, DIN Nestor and MOIMS-RAC. None of these approaches is widely used yet.
[20] Note that some funding agencies mandate depositing at specific data centers.

An increasing number of repositories – although still not many[21] – have formal methods in place to upload or replicate data sets, easy for users to use which lead to proper, auditable data management and accessibility.

It is understood that cross-disciplinary science with data from various disciplines increases the effort of data management at a given repository, since they will be confronted with data that is often differently organized, is described by different metadata standards of varying quality and utility, and that requires different management methods etc. Additional effort, and thus cost, is required.

5.2.4.3 Open Access

There is a clear trend to a higher degree of open access to data, although in practice many barriers need still to be overcome. There are reasons such as the granting of exclusive access to the data creators, at least for a limited time period (the time period varies greatly, however a period of 1 year is fairly common); legal reasons that prohibit the exchange of data, very common in the medical area; privacy protection and ethical reasons that need to be respected.

Unfavourable licensing often restricts access to data. A practical barrier stemming from the existence of many different licences in use, all need to be studied and signed, making cross-border or cross-discipline access an enormous administrative task that many researchers do not want to deal with. Combining data released under differing licences is also difficult, if not impossible in some cases. A reduction in the number of commonly used licences and computer-readable licences would reduce this overhead. Often institutes share their data internally but don't give it to the outside world, simply for the reason that they don't want to invest the time in sorting out the complex rights situation and thus minimizing the risks is the preferred choice.

Despite all good examples the general view seems to be that "data re-using" has actually only just started.

5.2.5 Aggregation & Collection Building

5.2.5.1 Aggregations

We can observe a general trend to aggregate data, software components and metadata, either centrally or in some distributed fashion. This aggregation stems from community efforts to structure the data and service domain and to improve visibility, accessibility and maintenance. Often only metadata about data and software components is aggregated, and increasingly often metadata harvesters are requesting that a path to the actual digital object in question is also be included.

Metadata is being harvested within and from many projects and initiatives and integrated into searchable or navigable portals to enable users worldwide to find useful data and/or software services. Since metadata is generally agreed to be open, everyone can harvest it; this in turn leads to more portals with specialized services. For aggregating metadata into unified catalogues, different strategies are used to map the semantics: (a) using a "golden set" such as Dublin Core, with the effect of semantic blurring and a reduction

[21] For example let us look at the CLARIN infrastructure. They have a requirement for certification and up to this point some 14 CLARIN centers got certified. For certification you have to show that you have fixed, documented, steps for dealing with data.

of information; (b) using full-fledged ontologies, which are often seen as inflexible when changes occur; (c) using exhaustive lists, the result of a semantic comparison of all categories, which often leaves many unfilled cells; (d) using a component based approach with manually specified mappings; and (e) mixtures of these approaches.

For all aggregation and mapping activities the biggest problems are a lack of quality and content completeness. Researchers restrict the amount of time invested in adding information, leading to overall lower quality; maintaining semantic accuracy if there are no experts around is a known difficult problem; tools offering accepted controlled vocabularies at creation time are not widely used, etc. Thus creating meaningful, consistent catalogues is still an effort that costs a lot of time and money. Automatic enrichment methods are still in the IT labs, but not widely used in practice.

Aggregating of software components alongside data is less widely practiced, since the benefit of aggregating them is not yet obvious. Nevertheless the increasing frequency of requests for computation on both created and aggregated data makes it necessary to rethink strategies.

An increasing number of centres in various communities are taking the role of managing, curating and preserving data that requires an aggregation step. Some centres have even built global competencies here. Increasingly often, many research centres do this kind of aggregation across disciplines and countries in order to carry out "big data" type calculations with the help of software components often borrowed from other disciplines.

In general there is a wide agreement on using OAI-PMH[22] for metadata harvesting, although a few other protocols, such as those from OGC, are being used.

As already indicated above, some researchers and data providers see metadata aggregation as not useful, and rely more on "semantic weaving" and storing content relations in form of RDF assertions etc.[23].

5.2.5.2 Collection Building

As indicated above data comes mostly in "canonical collections", i.e. they are grouped according to some meaningful principles underlying the creation process. These "canonical collections" are usually the basis for dissemination (distribution of packages in the sense of the OAIS model), rights management and data management (replication etc.).

Current practices, however, indicate that scientists may well want to group data in different ways, either goal or context driven, to create new virtual and/or physical collections to run their calculations on. This type of collection building often includes and combines data from various centres and/or disciplines. In doing this, many small digital objects can quickly amount to "big data" (as web crawling shows).

Collection building is still largely a manual activity involving the actual transfer of files; only a few communities have begun testing virtual collection building, where the metadata is grouped in new ways, but where the actual data is accessed *in situ* at the original data centres. There are a number of reasons

[22] http://www.openarchives.org/pmh/

[23] Theoretically RDF assertions on data are metadata in the general sense, but we need to make the distinction between different types of metadata.

why physical collection building is still preferred: (a) there is no good virtual collection infrastructure; (b) researchers want to have all data on their machines to be flexible and efficient in processing them; (c) in the case of heavy-duty computations it is important to have the data close to where the pipelines are being executed; (d) the number of data types is increasing enormously for cross-disciplinary work and file transformations can still best be done in iterative steps, including tests and inspections, and thus can best be done under one's own control. The increasing number of transformations required for this kind of work is also one reason why this work is often limited to specialist institutes with "data professionals" that have deep knowledge about data formats etc. Often there is a lack of explicitness in syntactic and semantic descriptions, requiring experts to dig into the file details to find out how to carry out the necessary transformations or to interact with the creators.

A very common situation for almost all communities is that large, aggregated data sets need to be included in a computation but the sub-collections are stored at different places. Currently researchers are transferring the sub-collections manually to one centre to work on the whole set. Copying is a time consuming effort, both in terms of network bandwidth and copying to/from physical media, thus all these communities are striving to understand how they can overcome these hurdles. There is as yet no proper solution. One proposed solution, of course, is to deploy processing algorithms dynamically at the remote sites – compute to data – but a production-ready distributed computing infrastructure with virtualised hot deployment and federated access is still in the development and testing phase. Grid, or cloud, computing is only used by very few infrastructures, but they do not yet play a role in scientific debate and/or practice.

A few communities offer services to their users to create and save collections, with the appropriate metadata and to register them with a PID, thus making them referable and citable.

Often collections are used not just by individuals but by groups of researchers which requires more investment in the design of suitable access rights systems, in particular when distributed authentication and authorisation methods are going to be used.

5.2.6 Management Operations

It seems to be widely agreed that, given the huge volumes of data and the increasingly complex inter-relations between data objects, current data management practices need to change fundamentally.

Yet the forces that inhibit change are overwhelming, even though requirements for up-to-date data management have been converging onto a few core principles (see RDA). The reasons have already been noted: (a) change requires expensive experts who are scarce and often can't be paid; (b) the time pressure for generating scientific results is enormous, giving little room for risky changes; (c) software components are being used that can rarely be adapted easily; (d) disasters have not yet been so great that a change was mandatory. There may be more reasons, but it is obvious that the ESFRI process, with the many resulting research infrastructures, has had an enormous influence on awareness raising, and on kick starting the search for new solutions. It will take a while, though, until new methods will reach a critical mass of institutions. There are some exceptions: the LHC experiment, for example, where management of experimental data has been sorted out as a must to enable the planned research programme to proceed.

A common oversight still is that data management for linear experimental data[24] is comparatively simple compared to the management of more complex data. Solutions found for experimental data cannot be used for complex data, although solutions found for complex data can also be used for linear data.

Typical management operations such as data replication are comparatively easy if they are just based on the physical structure (file system, cloud objects). Simple replication like this, however, risks losing logical layer[25] information, including references. Management operations including logical layer information are expensive at the moment due to the heterogeneity of the solutions. Only harmonization will reduce the costs and thus facilitate exchange and re-use.

Institutes highly focussed on data-driven research report that data management costs an enormous amount of researchers' time (typically beyond 50%), at the expense of truly scientific work. The reason for this is as we have noted: increasingly out-dated data management methods. "Big data" definitively requires new, highly automated methods of data management.

Proper data preservation by systematic replication to other institutions is not widely used. Cheap, easy-to-use service offers are missing, or responsibilities have not been clarified. Data preservation in many cases is still done by copying data within or across file systems, and onto hard drives or tapes, which more often than not are then stored in cupboards.

5.2.6.1 Policy Based Management

Currently most data management work is based on manual operation or in commands embedded in scripts or software. Results tend not to be traceable, they lack a systematic approach and are widely undocumented.

Researchers understand that a great deal of time is lost from highly qualified persons because of these practices, tracking down the data objects one wants to use. Also, little quality control can be carried out – an issue as it is increasingly required.

With some exceptions, the adoption of explicit methods such as writing declarative policy rules in an easy to understand language is very distant from what people are either doing or even aware of. In some areas (such as clinical data) there is a tradition of managing data through explicit policy statements, which are often only available as documents and have not yet been transformed to executable form. However, until such explicit policy rules become the basis of data management, no effective certification of repositories can be done.

Some researchers argue that these kinds of declarative policy rules cannot be created easily because there are too many exceptions and special cases that cannot be transformed into simple workflow chains. While this may be true in some cases, there is also a lack of knowledge about these technologies and their opportunities.

---

[24] To give an example of linear data think of a bunch of sensors continuously making measurements. I.e. there is a clear time series to organize the measurements around.

[25] We use the term "logical layer" here to summarize all kinds of meta-information associated with a digital object such as metadata, PIDs, rights, relations, etc.

Developing expertise in this field of workflow chains would require building up additional knowledge, and many research groups simply lack the necessary manpower and skills. This situation is unlikely to change, so we need to come up with more off the shelf solutions like: “if you want to replicate your data in the following scenario, please take this policy rule and it will do”. Such solutions would also introduce greater harmony on how management is carried out, and might help improve quality.

As indicated above, most data communities do not have any good strategy in place to include the logical layer information in their management activities.

5.2.6.2 Federations

More and more groups of researchers want to access data, which is distributed across several repositories in a seamless way (single sign on, single identity), or a group of repositories want to create a platform that allows for the easier movement of data between centres. Therefore, the creation of federations based on formal agreements, for example about data management and access, and on some technical agreements, is increasingly popular.

To make federations work easily for users, distributed AAI methods that implement SSO and SI principles are necessary. However, the possibilities offered by the current platforms provided by most of the European NRENs, in collaboration with eduGain and GEANT, and their actual limitations, are still widely unknown. Again, it seems that most communities lack the experts to put such mechanisms in place and to manage broad uptake. Only when a critical mass participates at a suitable level can investments in establishing a federation be justified.

Given this lack of both expertise and experts, some communities consider “offering federation technologies as a service” to their members.

5.2.7 Analysis Operations

An increasing number of researchers want to work on data sets that are getting bigger and more complex (different types, relations, etc.). In general there is a growing awareness that computational methods need to change to maintain, or achieve, reproducible science: “big data” cannot be done in the same good old fashioned way, except perhaps for canonical operations on linear data sets.

Researchers want flexible access to data including legacy data. Access to newly created data is much more frequent, but access to old data is still common practise and a requirement. It is getting difficult to predict which data will be accessed to work on. This poses problems for storage and transmission technology, since often staging of some sort is necessary to be able to work efficiently. Researchers do not like IT-induced time delays since it slows down their work.

Alongside the increase in computational needs comes awareness that aggregating data and metadata alone does not solve the problem; an aggregation of metadata about tools and services, and a bundling of such tools and services in centres of competence, is just as relevant. Some researchers speak about the need for an “open market place” for software, where users can add comments about their experiences. The rationale is simply that researchers have no time to try out a range of components; community comments can guide them in a suitable selection from the beginning.

Research communities have adopted different approaches to bringing software tools together: some work on open toolkits which can easily be integrated into scripts; others aggregate web services and adapt them by using unified data structures for data exchange. Other methods may be in use as well.

Parallelism within some larger data set and the increasing number of computational requests also implies that combining the functions of "data centre" and "compute centre" into one institution makes a lot of sense, and reduces overhead. It would also solve the problem that several communities currently have of where large intermediate data sets should be stored for limited periods of time.

Developing the idea of combining data and compute capacity, it makes sense to offer "services on data" which can then be used by many groups. This is a growing paradigm, since interested researchers can, in principle, just "press a button" to get results. However, the success of this approach is under debate in research, since once "typical functions" are used by masses of researchers, the results naturally become less innovative. Indeed, most researchers want to do their own creative computations on their own collections, since this is what distinguishes them from others, and therein lies the potential to achieve breakthroughs.

Data processing, particularly on complex collections, leads to an enormous increase of formats and relationships, including provenance, between different components. Information about processing steps is still stored very often in unstructured, manually created file types such as Excel[26]. Researchers are increasingly coming to realize that this does not work anymore, but they hesitate to move towards self-documenting workflows, since this requires expertise that they do not have and hiring expensive experts is often out of (budgetary) scope.

Community experts are certainly discussing workflows, however, and some research infrastructures have introduced this kind of service to members of the communities they serve. For "canonical process chains" that show little variation, easy-to-use workflow environments have been developed that require no programming skills from researchers. Some researchers suffering from heavy data management requirements would like to move to workflow-based approaches, but they need to carry out a lot of testing and cope with many exceptions, and often their processes need manual interventions and parameter control, so that they don't yet see workflows as realistic alternatives. Thus, although much effort has been put into the development of workflow frameworks, they are still underused. Usage of workflow technology is not at a point yet where support for the "data fabric" principles is built in by default.

Annotating digital objects by automatic or manual methods is a common paradigm to enrich data and increase its value for deep analysis. Annotations can be made on whole objects or on fragments of an object. In the first case they can be compared with additional user-created metadata, and in the second they can extend to sequences of comments, each of them being associated with some information contained in the annotated object (time period in a time series, pixels in a picture, etc.). The way annotations and their relations are stored is heterogeneous, and yet knowledge about available open frameworks, such as the open annotation framework[27], is very limited.

[26] We note that Excel can be very structured (it is a spreadsheet after all), but that people tend to use it without specifying a schema.
[27] See also http://www.openannotation.org/spec/core/

Combining data with textual information from different disciplines, or even different creators within a discipline, requires some form of ontology or controlled vocabulary to support the "crosswalks". Despite the fact that many ontologies have been created, in practice they are heavily underused: still the norm is handcrafted, easy-to-manipulate mapping tables. Crosswalks are still seen as difficult and ontology-enabled crosswalks have the reputation that tight semantic control is being lost.

A new paradigm has recently appeared on the scientific scene: the automatic extraction of assertions and their formulation in (augmented) RDF triples. "Nano publications", for example, storing highly reduced statements on causal relations, allow researchers to generate statistics on them and thus manage the information flood. Metadata transcoded into RDF triples allows researchers to look for hidden information patterns. These technologies are still in their infancy, and only a few researchers are using them.

5.2.8 Data Publishing

Data publishing was not raised much as an issue in the various interviews and interactions, but the relevance of this topic was underlined by the Science Workshop (Section 3.3). In general we can state that it is a topic discussed broadly in the disciplines, that some communities already publish quality controlled data sets by registering them with DOIs and associated citation metadata and that some communities use highly respected community portals to publish their data or part of them.

5.3 Overall Conclusions

In this section we summarise our findings from all the observations made in the analysis programme, also factoring in the results from the Science Workshop reported in Section 0[28]. At the end of each summarized finding we refer back to the paragraphs that are most relevant for that finding.

1. **ESFRI and e-Infra**: Both the ESFRI discussion process and its project initiatives, as well as recent developments in e-Infrastructures, have strongly influenced the mind-sets, the practices and the interaction processes around data management crossing discipline boundaries. (§3.4, §5.2.3.2, §5.2.6)
2. **Open Access**: OA is supported everywhere as a basic recommendation. However, in practice there are many hurdles to make data really available. Three aspects in particular that are not often discussed are that: a) data is often in a badly organized, badly documented state and people hesitate to invest additional time if it is not explicitly requested; b) there is a lot of legacy data requiring much effort to make it accessible; and c) often the ethical and legal situation has not been clarified and people hesitate to invest time in addressing these complex issues. (§3.3, §5.2.4.3, §5.2.8)
3. **Trustworthiness**: In an era where data sharing and access across disciplines, countries and centres is becoming more of a default situation and consequently where the direct relation between data provider and data consumer is broken, new methods are required to establish trust on all levels. Therefore "trustworthy" in all its many facets is a recurring concern. A few key issues are: (a) researchers need to be sure that the quality and integrity of data is guaranteed; (b) researchers need to have access to good quality metadata to be able to interpret and make use of data without having the need to contact the producers (a process which simply does not scale anymore); (c) repositories need to offer robust, sustainable services of value, and they also need to guarantee stability of access to

[28] The main messages from the first report are also captured, although they may be presented in different form and context.

specific data objects by, for instance, registering them with the help of persistent identifiers; (d) there is an increasing urgency to clarify responsibilities at policy level and ensure funding streams for repositories, since without long-term guarantees no one will invest the time to deposit data. (§3.2, §5.2.1, §5.2.2, §5.2.4, §5.2.6)

4. **Legacy Data**: It is obvious that many communities have not only a huge problem with badly documented, badly organised legacy data, but also that large amounts of new, badly documented and badly organised data is still being created. This is largely due to unchanged practices in data creation and management caused not by lack of willingness to change, but by: a) lack of knowledge of up-to-date data organizations (in particular the maintenance of relationships between different data objects); b) the cutting off of logical layer information in management steps; c) lack of experts, time and money to modify existing software; and d) lack of clear, well-defined software methods at all relevant data management steps (registering encoding schemes, structures/formats, semantic concepts, etc.). The transformation of this still-growing corpus of legacy data to data that can be used in data-driven science is a long, incremental process, the difficulties of which are frequently underestimated. (§3.1, §4, §5.2.1)
5. **Big Data**: The requirements of "big data[29]" are new for many communities, but scientific competition makes it necessary that institutes and departments need to adopt this paradigm. It is understood that big data work will only scale efficiently when data management principles change. There seems to be consensus that it would be very good to move away from manually executed or ad-hoc-script-driven computations to automated workflows, but there is a reluctance to take this step. It is increasingly understood that use of ad-hoc methods cannot continue, and the need is present to move towards automatic procedures based on practical policies captured in executable workflows which are both documented and self-documenting and adhere to basic data organization principles (PIDs, metadata, provenance, relationships). Despite the principal agreement on current inefficiencies and the need for change, there is still a reluctance to implement changes in practice. The main reasons seem to be: a) again, the lack of resources and expertise that can transform current practices into the necessary flexible and parameter-controlled workflows; and (b) doubt as to whether such automatic workflows can cope with the inevitable exceptions, special testing requirements and the various parameter settings that control processing[30]. (§3.1, §5.2.2, §5.2.5, §5.2.6)
6. **Data Management**: For processing data for management purposes the same situation seems to apply. There is an urgent need to change current methods of dealing with data since the processes are too inefficient and too costly and do not lead to a reproducible science. Moreover, in data management file system-based operations are still dominant, now augmented with cloud-based approaches. Neither approach supports operations that include logical information, i.e. relations to metadata, PIDs, the relationships between files etc. (in cloud terms, typically this occurs when you leave the realm of the cloud application). Again, finding the right data objects and creating meaningful collections wastes a large amount of researchers' time. Further, the differences in software solutions used for data management are yet another source of unwelcome heterogeneity (§5.2.1, §5.2.2, §5.2.3, §5.2.3.2, §5.2.4, §5.2.6)
7. **Metadata:** Despite many years of discussion about metadata and its relevance metadata practice is still far from being satisfying which is not only hampering discovery, but in particular re-usage after some

[29] In appendix D a more detailed elaboration on Big Data Analytics requirements provided by C. Thanos can be found.

[30] It should be noted that in the US there are the first institutes such as RENCI who invested considerable funds to generate such workflows and thus get a competitive advantage.

time since people lack context and provenance information usually embedded in metadata. Obviously more guidance, ready to use packages (instead of complete schemas) and supporting software is very much required to improve the situation. (§4, §5.2.1, §5.2.3.2, §5.2.6)

8. **Lack of Explicitness**: There is lack of explicitness in the kinds of information important for efficient machine-based processing of data, be it for management or analytics. This ranges from non-registered digital objects (i.e. lacking PIDs), data integrity information (such as checksums) collections, encoding systems, format/syntax, and semantics up to the level of software components. Appropriate registration authorities and mechanisms do exist, but often they are unknown or not used. (§5.2.1, §5.2.2, §5.2.6)
9. **Centres**: A clear trend towards the use of trusted centres within and across communities is visible from the perspective of structuring the data landscape, and is now mentioned more often in relation to structuring the tool landscape, for transforming digital objects into a well-managed and well-maintained state. We urgently need to motivate research infrastructures and organizations to establish such centres in particular as repositories with a long term preservation mission and to offer reliable services to all researchers. Frequently, to make it easier to build virtual collections and carry out distributed processing jobs, these centres need to create federations of domains of single identities and single sign-on, but we do not yet have common ground for doing this. Some aspects of distributed authentication and authorization are still not in place at European level and distributed computing, although mentioned increasingly often, is not a well-understood scenario. (§5.2.1, §5.2.4, §5.2.6)
10. **Education & Training**: A recurrent observation is that we lack data professionals in their different facets and that this hampers changes and progress. This stresses the need to intensify education and training efforts, although if the increasing number of data professionals were to lead to an increasing number of solutions, we would even get more proliferation of incompatible solutions. Thus we also need to reverse the proliferation of data organization and management solutions, accepting that they are widely discipline-independent and thus can be standardised. (§4, §5.2.1, §5.2.7)
11. **Lack of Knowledge**: We can also conclude that there is a lack on trusted information about all kinds of services that are being offered (registries, data, storage, curation, analytics, etc.). The web offers a wide spectrum of possibilities, but many researchers can't cope with this information flood and have a hard time to make a selection. A more structured and trusted approach of offering information would have great impact. (§5.2.1, §5.2.6, §5.2.7)
12. **RDA**: An imperative on RDA is to ensure it can be a true grass-roots organization, and to provide demonstration cases and give help and support to research communities, while respecting that researchers are under heavy time pressure and are reluctant to spend time to test out new methods. (§3.5, §5.2.6)

5.4 Concurrence of RDA Activities

Many of the activities undertaken by the groups of RDA can be shown to address one or more of these observations or recommendations. Below we will go through all RDA working groups active as of July 2014 and detail how they relate to the recommendations of the science workshop, process model, and the observations from the interviews.

5.4.1 PID Information Types WG

The PID Information Types WG is developing a protocol that supports the registration and query of PIDs together with a harmonized set of useful information types. The crux is independence of any underlying PID systems: if a certain type like a checksum is requested, all providers should be able to understand this type.

This group addresses issues related to the registration and referral part of the process model. Furthermore, the efforts aid in building what we term the data fabric, and workflow systems could take profit from creating a uniform interface to deal with digital objects referenced by PIDs.

5.4.2 Data Type Registries

The goal of the Data Type Registries (DTR) group is to enable data producers and data managers to describe relevant data structures, assumptions, and usage conventions in order to enable humans and automated systems to process and understand data. Included are data types at multiple levels of granularity, from the types in the type/value pairs returned from identifier resolution up to and including single types defining complex datasets and collections of such datasets.

It is clear that this group addresses some core issues raised at the Science Workshop, i.e. data stored in an infrastructure must be easily usable and understandable. In particular, the process model addresses the reusability of digital objects by providing visualisations and conversions by making use of its types. Furthermore, it directly addresses some concerns raised in the observations about re-usability and interpretability of existing data, assuming that the formal definition of data types is available in such a federated registry.

5.4.3 Metadata Standards Directory WG

The Metadata Standards Directory group is working with the Digital Curation Centre (DCC) to further develop DCC's existing metadata directory. Developments will be made to DCC's website to ready it for "community maintenance". The metadata standards will be citable and usable using different modalities, e.g. using GIT[31]. Furthermore, efforts are being undertaken by the group to assess which metadata standards are in use or needed by members of the RDA community. These metadata standards are added to the directory.

The metadata group addresses one aspect of the metadata part of curation in the process model, i.e. it tries to promote the creation of good metadata (using a metadata standard, rather than say an Excel file) by listing all relevant metadata standards in an understandable manner, which is in particular relevant for curation steps. In terms of the Science Workshop recommendations, metadata is an essential component in discovering data, examining the provenance of data, and reproducing work. Proper metadata is also needed in many areas such as the workflow and data fabric paradigms.

5.4.4 Data Foundation and Terminology

The goals of the Data Foundation and Terminology (DFT) WG are twofold: 1) to foster a shared basic core model of data organizations, which will help harmonize research data management across data communities; and 2) to facilitate the adoption of shared basic terminology that is based on this model and principles.

The DFT group addresses issues raised at the Science Workshop and observations made in the interviews in so far as its clarifications may lead to more efficient and cost-effective data management procedures. The

[31] GIT is a distributed version control system that can work fully peer-to-peer. However the Metadata group means to use it more centralized (i.e. with a centralized authoritative repository).

work of this group is essential to cross-domain working as it seeks to generate a common frame-of-reference terminology that can be used to discuss data issues in a much easier way.

5.4.5 Practical Policy

The goals of the Practical Policy group are threefold, to: 1) identify a number of typical application scenarios for policies such as data replication, and preservation; 2) collect exemplary practical policies for a first number of such application scenarios, register them, allow people to compare and re-use them and to extract options for commonalities and optimizations; and 3) create awareness about ways to arrive at reproducible science, to achieve trusted repositories and to allow proper certification.

The Practical Policy group addresses the Science Workshop recommendation that "Provenance, Validation, Trustworthiness of resources must be assured", through the promotion of explicitly defined practical policies in archives which help them earn the trust of the users. It is essential is that computer actionable policies are easily readable and not hidden away in software and configuration layers. Practical policies are understandable to the user, and are thus also auditable. Clarity and an audit process help create and maintain trust. Having such policies in place can also ensure a provenance trail, assuming appropriate mechanisms are included in the policy rules. Having computer actionable policies makes it clear what happens to the data in the infrastructure, which allows for various actions and checks to be made automatically. Automatic policies also help preventing human errors which are one of the most common causes of data loss.

As an example policy on data preservation as used in the EPOS project in the EUDAT infrastructure, all files that come into an EPOS centre are to be replicated across two other data centres. Furthermore each of these centres will do both full and incremental backups, calculate MD5 checksums to check replication success, perform regular MD5 checksum checks in order to verify integrity, update and use PIDs for all digital objects. This policy is then ensured by a human readable but computer-actionable policy description.

Within the process model, practical policies can be used to automate many of the underlying processes.

5.4.6 Data Citation: Making Data Citable.

The Data Citation working group is working on making dynamically updated data citable. To give a sketch of the reason for this, data from sensors does not necessarily arrive in order. In reality the stream received from a sensor is non-linear in time: some parts of the data can come in later due to delays, or perhaps not at all. It is not a logical linear stream of measurements.

The problem here is that even though the data is not complete the expert (e.g. a volcanologist) often has to make immediate decisions on how to act. That means that statements (potentially health- or life-critical) are sometimes made when the stream of measurements is incomplete. If called to justify a given statement, how can one refer to the state of a data resource as it was at the time the statement was made? It might well be that by the time the statement is re-evaluated the underlying data stream has been changed.

This working group addresses the repeatability concerns raised at the Science Workshop and in the observations. It is important to be able to reproduce the data as it was at the moment of a computation; it is also important that new computations are run with the most up-to-date data available. In the process diagram this working group addresses the published data step, with the caveat that the *version* of the data used in a process should be published if that process leads to a publication or decision.

5.4.7 Wheat Data Interoperability

The Wheat Data Interoperability working group is gathering as much information about data on wheat as possible, and seeks to combine these in order to gain new insights on better exchanging these data. The ultimate goal in furthering this data exchange is to be better able to grow wheat worldwide. This is a pilot study in order to be able to do the same for data on other crops. The wheat group does not address any of the observations or recommendations globally except, indirectly, the data type aspects. Their goal is rather to try to organize data better within a single field.

5.4.8 Data Description Registry Interoperability

The Data Description Registry Interoperability group aims to address the problem of cross-platform discovery through a series of bi-lateral information exchange projects, and to work towards open, extensible, and flexible cross-platform research data discovery software solutions. Developing such solutions requires answers to problems including author disambiguation, persistent identifiers (required for identity resolution and disambiguation), authentication (e.g. commercial publishers), access rights management, search optimisation (search ranking), metadata exchange (crosswalks), and creating a connected graph of research datasets, authors, publications and grants.

The group is targeting specific interoperability cases between different registries or archives. This addresses observations raised on aggregation and interoperability. In the process model this fits within the aggregation/collection building part of the process.

5.4.9 DSA-WDS working group on Certification

The working group on certification is comparing and contrasting two repository certification standards in the hope of coming to a common understanding and integrated set of requirements, thus hopefully improving levels of trust in certified repositories. To quote the group themselves: “Certification is fundamental in guaranteeing the trustworthiness of digital repositories, and thus in sustaining the opportunities for long-term data sharing and corresponding services.”

The group addresses the trust issues raised in both the Science Workshop and interviews as one of the key issues towards improving data sharing practice. The group is addressing issues that often arise when communicating with a repository, be it during data deposition or access.

5.4.10 RDA/WDS Publishing Data (quadruple of groups)

The RDA/WDS publishing data working and interest groups deal with four different aspects of publishing data, namely: 1) publication workflows; 2) use of bibliometrics; 3) creation of data publication services; and 4) cost recovery for data centres.

As such the groups address several of the concerns from the Science Workshop, such as the longevity of data centres (via cost recovery), and the issue of getting credits for your work (via bibliometrics). Similarly they address a number of observations with respect to data publishing by attempting to streamline the data publication process through services and workflows.

Within the data model the activities of this group fall in the publishing data and publishing papers fields.

5.4.11 Infrastructure WGs

For the “infrastructure” related WGs that started early in the RDA process, the mapping was comparatively easy. Here we explain the inter-relations between the Metadata (MD), PID Information Types (PIT), Data Type Registry (DTR), Practical Policy (PP), and Data Foundation and Terminology (DFT) groups.

The **MD** group on metadata registries is working on a registry for metadata schemas which are used in the communities that may help others find useful schemas for re-use and correctness checking. The **PIT** group is working on defining an API to allow the registration of PIDs together with useful information types associated with them and retrieved from a Data Type Registry. The **DTR** group is working on a generic schema for formally describing types, one that can be used register all kinds of simple and complex types and that can thus be used to interpret data, in particular at the interface to the management/processing step. The **PP** group is busy creating a register of practical policies that are being used in a variety of processing steps (management, analytics) and that can thus serve to harmonize practice. Finally, the **DFT** group has analysed the core model of a proper data organization and has defined terms such as Digital Object, PID, Metadata, Repository and Aggregation to improve understanding. They have now agreed to build the “Data Fabric Interest Group” to discuss the complete set of components that could help to transform our current data management and analysis practices into automatic, self-documenting processes.

6 Recommendations

In this chapter we present a number of recommendations to funders for concrete actions that follow from the many observations presented in the document. Of course many issues are directed to the research community such as

- acceptance of Open Access principles where possible,
- increase quality of all data products incl. metadata,
- make structure and semantics explicit and
- be open to change practices and invest funds.

We also see the urgent need to facilitate transition by pushing boundaries with the help of key investments. Some of the issues raised have already been addressed in some form, but mentioning them here again stresses the importance.

Education & Training
A recurring point is the lack of a new generation of well-trained data professionals that will finally put the required changes into practice. European funders need to strengthen their efforts if Europe wants to maintain its leading role. These professionals are not just for supporting data work in research and industry, but should also be motivated to join SMEs or act as entrepreneurs. Given the availability of well-educated data professionals we cannot imagine, however, that more funds will be available in the many departments to add professional staff, which is effectively required for more efficient methods of dealing with data.

Open Data and Service Market Place
To foster exchange and competition we need to come to an open market place of data and services which helps researchers to easily find their way efficiently and to evaluate components they are using. We suggest funding two mid-size projects at first that are driven by domain researchers that first come up with a design and work out the design in a prototype within a one year project. A subsequent evaluation should result in a choice of funding the project with the most promising concept. The design must (1) specify how initiatives from different disciplines can be covered, (2) include policy, sociological, organizational and technological factors, (3) foster wide participation, (4) include industrial offers and (5) come to a self-supporting state after a few years.

Data Fabric Success Stories
We urgently need success stories for implementing the data fabric concept in the basic scientific domains that have the potential to indicate the essential components of such a concept and their interfaces in collaboration with RDA, to implement and integrate them by making extensive use of existing technology and to convince researchers to change practices to a reproducible data science. There must be a strong coordination and evaluation between the selected projects to achieve maximal coherence with the domain scientists having a leading role. The selected projects need (1) to implement a processing and management strategy based on practical policies to make the step towards automatic, systematic and self-documenting procedures, (2) to implement chains of trust in a convincing way, (3) to establish and integrate registries of various sort building on existing services where possible (such as PID registries), (4) to include convincing provenance and context metadata creation and data preserving solutions and (5) base their procedures on proper data organization principles as discussed in RDA.
Also here it seems to be most promising to fund design studies for a year and then select those projects that promise to have a high transitional potential.

Overcome Legacy

The problem with legacy data has widely been underestimated and we are still creating unsuitable “legacy data” by not adhering to proper data organization principles. We need to support communities in curating their data and metadata in a way that they become part of our open and accessible data domain based on proper data organization principles, that its structures and semantics are registered in accepted registries so that machines can work with them without human intervention.

Here we suggest funding quite a number of small and larger curation projects that have a potential to also be show cases for other data sets. One aspect of these curation projects is that independence of data from technologies needs to be ensured and that data is integrated in existing and certified repositories.

Repositories

Since it is widely accepted that trusted and strong repositories will be the key to store, preserve and offer access to data and metadata in particular when they are built on sustainable funding models we need to convince research organizations and funders to invest in such repositories that are smart enough to support the research workflows and not act as passive archives. It must be a high priority task of research infrastructures and e-Infrastructures to make proper suggestions how to structure their domain taking into account the existing offers and to evaluate them regularly according to the emerging standards. Some are specialized in metadata aggregation which needs to be much better supported by better quality of metadata adhering to standards.

Federation Solutions

Increasingly more departments want to participate in changing trust federations, yet the technology is not ready to facilitate this. A project needs to be carried out that is focusing on making federation technology mature. This includes authentication, authorization, gatekeeper technology, education and much more. Also here it would make sense to fund two short-term design studies determining essential components and assessing the current state of developments. As a result a goal-oriented project could follow that implements or improves components.

Appendix A. RDA/Europe and Max Planck Society Science Workshop on Data

A.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.

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 A.8 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.

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.

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.

A.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:

- 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.
- 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.
- 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.
- 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 (explain) establishing a “data fabric”.
- 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.

- 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.

A.3 Sharing and Re-use of Data

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

- 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.
- 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.
- 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.
- 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.

A.4 Publishing and Citing Data

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

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

A.5 Infrastructures and Repositories

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

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

A.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

A.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:

- 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.
- 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.

- 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.
- 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.
- There are a few expectations RDA has to meet:
    - RDA should certainly invest in training younger generations of data scientists.
    - 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.
    - RDA should have data experts who can visit institutes and help them implement solutions.
    - 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.

A.8 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
Bernard Schutz	Gravitational Physics	Cardiff U / MPG
Bruce Allen	Gravitational Physics	MPI for Gravitationphysics, 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, RDA TAB	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
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, 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 Secretariate	MPI for Psycholinguistics, Nijmegen
Ari Asmi	Atmospheric Sciences	University of Helsinki

Appendix B. List of Attended Community events

IVOA meeting, October 2014, Banff	Astronomy
Avoin Suomi, September 2014, Helsinki	Open Data
INNET Meeting, September 2014, Budapest	Linguistics
OSGIS, September 2014, Nottingham	Open Source Geographic Information Systems
INRIA, September 2014, Paris	Computation
Humanities Regional Center Meeting, August 2014, Berlin	Humanities
Neuroinformatics Meeting, August 2014, Leiden	Neuroinformatics
DataCite annual meeting, August 2014, Nancy	Data Citation
Nordforsk Open Data meeting, August 2014	Open Data
Developing countries workshop, August 2014, Nairobi	Data in developing countries
ISC, June 2014, Leipzig	Supercomputing
e-IRG, June 2014, Vilnius	e-Infrastructures, various communities
Datajalostamo-Keskusteluseminaari, June 2014, Helsinki	Data Refinery
Infrastructure Meeting, June 2014, Brussels	Burning Issues in Infrastructures
Terena, May 2014, Dublin	Network Infrastructure
ECRIN Data Meeting, May 2014, Düsseldorf	Medical/BioInformatics community
LREC, May 2014, Reykjavik	Language resources and tools
ERF Meeting, May 2014, Brussels	Physical Facilities
ELIXIR NL Interaction, May 2014, Utrecht	BioInformatics Experts
EGI user forum, May 2014, Helsinki	various community experts
IVOA Meeting, May 2014, Madrid	Astronomy
ISO Study Group on Big Data, May 2014, Amsterdam	Big Data
Big Data and Open Data, May 2014, Brussels	Research Infrastructures, Open Data
Digital Cultural Heritage Roadmap for Preservation, April 2014, Tallinn	Cultural Heritage
MENESR, April 2014, Paris	Research

UCL Big Data Symposium, March 2014, London	Big Data
RDA WG Chairs Meeting, March 2014, Munich	various community experts
Linking Geospatial Data, March 2014, London	Geospatial Data, Linked Data
Social Science Congress, February 2014, Berlin	Social Sciences
RDA-MPS Science Workshop, February 2014, Munich	various community experts
MPS Data Science Meeting, February 2014, Ringberg	various community experts
Linked Data Workshop, February 2014, Amsterdam	Linked Data
Symposium with Marco Baroni, Nijmegen	Computational Linguistics
Big data at Britisch Geological Survey, January 2014, London	Big Data, Geology
RNA-seq data analysis workshop, January 2014, Espoo	Genomics
MPD – CAS (China) Interaction on Data, January 2014, Munich	MPS/CAS Data Experts
DataCite Meeting, January 2014, Hamburg	various data professionals/librarians
Museum für Naturkunde Meeting, January 2014, Berlin	Biology
Copori, December 2013, Brussels	Research Infrastructure
e-IRG, November 2013, Vilnius	Research Infrastructure
EUDAT, October 2013, Rome	Infrastructure
SSH Tackle the big data challenge, October 2013, Rome	Social Sciences & Humanities
IEEE Conference on e-Science, October 2013, Beijing	e-science
RDA-CAS (China) Interactions, October 2013, Beijing	various community experts
MedOANet Conference, October 2013, Athens	Open Access
DASISH workshop for SSH project, October 2013, Gothenburg	Social Science & Humanities
The Challenge of Big Data in Science, September 2013, Karlsruhe	Big Data
EUDAT Working Group Meetings, September 2013, Barcelona	Various community experts
IVOA Meeting, September 2013, Waikoloa	Astronomy
Atila, September 2013, Corsendonck	Computational Linguistics
EPOS Meeting, August 2013, Erice	earth observation community
HPSC 2013, July 2013, Helsinki	Computing

SSH Meeting, July 2013, Berlin	Social scientists & Humanities experts
COOPEUS Data Meeting, June 2013, Bremen	Environmental Data Experts
Humanities Annotation Meeting, June 2013, Berlin	Humanities Experts
DOBES Conference, June 2013, Hannover	Linguists
Terena, June 2013, Maastricht	Network Infrastructure
ALLEA Meeting, June 2013, Rome	various community experts
Cluster Interaction Meeting, June 2013, Hinxton	various community experts
Meeting BioInformatics Experts, June 2013, Berlin	BioInformatics, Genomics
e-IRG, May 2013, Dublin	Research Infrastructure
CCGrid, May 2013, Delft	Grid
Nordic e-Infrastructure conference, May 2013, Trondheim	e-infrastructure
IVOA Meeting, May 2013, Heidelberg	Astronomy
EGI Community Forum, April 2013, Manchester	Grid
EDF, April 2013, Athens	Big Data
EGU, April 2013, Vienna	Geospatial
Interaction at various US research labs, April 2013, US	various community experts
HPC2013, April 2013, San Diego	HPC
HPCC, March 2013, Goat Island	HPC
GigaOM's Structure:Data, Put Data to Work, March 2013, New York	Big Data
ISGC, March 2013, Taiwan	Grid
EUDAT User Forum, March 2013, London	various community expertise
e-IRG, December 2012, Amsterdam	Research Infrastructure
Interaction with astronomers, March 2013, Strasbourg	astronomers
Atila, September 2012, Corsendonck	Computational Linguists
iRODS User forum, March 2013, Munich	various community experts
IDCC Conference, January 2013, Amsterdam	various community experts

Appendix C. List of Interviews

Interviews done in the RDA Europe project:

• EMBL-EBI	Bioinformatics
• Genedata	Bioinformatics
• TNO: Toxicogenomics	Toxicogenomics
• ENVRI	Environmental Research
• Svali	Arctic land ice
• Eiscat 3D	Atmospheric research
• Engage	Public sector information
• ESPAS	Upper atmosphere research
• Math Community	Math
• Huma-Num	Social sciences
• INAF	Astronomy
• ML-Group	Machine Learning
• CL-group	Computational Linguistics
• CLST	Natural Language Processing (text-to-speech)
• Donders	Brain research
• Meertens	Dialect research
• MPI-DevBio	Developmental Biology
• NIOZ	Arctic
• UNESCO-IHE	Hydrology
• Global Geothermal Energy DB	Geothermal Energy
• iMARINE	Marine life
• CDS-Strasbourg	Astronomy
• INCF	Neuroinformatics
• MPI for Molecular Genetics	Molecular Genetics

Interviews done in the EUDAT project:

• EUCLID	Space
• Virgo	Cosmology simulations
• CDS-Strassbourg	Astronomy
• Cessda	Social Science Data
• CLARIN	Linguistics Data
• Clarin Metadata	Linguistics Metadata
• Dariah	Social Science Data
• Dixa	Toxicology on Microbiology
• Ecrin	Medical Trials Data
• Emso	Sea Floor Data
• Enes	Climate modelling
• INCF	Neuroinformatics
• Museum für Naturkunde Berlin	Physics
• PANdata	Photon and Neutron Data
• VPH	Virtual Human Data

Appendix D. Big Data Analytics

Provided by Costantino Thanos

(many of the concepts presented in this report are drawn from a white paper developed by leading researchers across the United States and a report published by BI research)

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the data analysis pipeline that can create value from data. The problems start right away during the data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured forma; transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value.

Data analysis is another foundational challenge. It constitutes a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms, and due to the complexity of the data that needs to be analyzed. During the last 30 years, data management principles have enabled the first round of business intelligence applications and laid the foundation for managing and analyzing Big Data today. However, the many novel challenges and opportunities associated with Big Data necessitate rethinking many aspects of existing data management platforms.

It is necessary to develop a new wave of fundamental technological advances to be embodied in the next generations of Big Data management and analysis platforms, products, and systems. However, achieving these advances is also hard and requires a rethinking of data analysis systems. The analysis of Big Data involves multiple distinct phases, each of which introduces challenges:

Acquisition/Recording → Extraction/Cleaning/Annotation → Integration/Aggregation/Representation → Analysis/Modeling → Interpretation

Many people unfortunately focus just on the analysis/modeling phase; while that phase is crucial, it is of little use without the other phases of the data analysis pipeline.

Data acquisition and Recording: Much of the data produced/collected must be filtered and compressed by orders of magnitude. One challenge is to define filters that do not discard useful information. A second challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured.

Data Extraction/Cleaning/Annotation: Frequently, the data produced/collected will not be in a format ready for analysis. The data cannot be leaved in this form if we want to effectively analyze it. A data extraction process is necessary that pulls out the required data from the underlying sources and expresses it in a structured form suitable for analysis.

Data Integration/Aggregation/Representation: For effective large-scale analysis it is required that differences in data structure and semantics should be expressed in forms that are computer understandable and then robotically resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

Query Processing/Data Modeling/Data Analysis: Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy Big Data are more valuable than the tiny samples. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, etc.

Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms and big-data computing environments. Big Data is also enabling the next generation of interactive data analysis with real-time answers. Scaling complex query processing techniques for terabytes while enabling interactive response times is a major open research problem. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses.

Interpretation: Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Usually, the interpretation of the analysis results involves examining all the assumptions made and retracing the analysis. It is rarely enough to provide just the analysis results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data. Systems with a rich palette of visualizations become important in conveying to the users the results of the queries in a way that is best understood in the particular domain.

Challenges in Big Data Analysis

Heterogeneity and Incompleteness: Machine data analysis algorithms expect homogeneous data and cannot understand nuance. Consequently, data must be carefully structured as a first step in data analysis. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge. Recent work on managing probabilistic data suggests one way to make progress.

Scale: Of course, the first thing anyone thinks of with Big Data is its size. Managing large and rapidly increasing volumes of data when these volumes are scaling faster than the compute resources, and CPU speeds are static is a challenging issue. First, over the last five years the processor technology has made a dramatic shift: processors are built with increasing numbers of cores. In the past, large data processing systems had to worry about parallelism across nodes in a cluster; now, one has to deal with parallelism within a single node. Unfortunately, parallel processing techniques that were applied in the past for processing data across nodes don't directly apply for intra-node parallelism, since the architecture looks very different. Second, currently there is a move towards cloud computing, which aggregates multiple disparate workloads with varying performance goals into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures.

Timeliness: There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding

qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

Privacy: The privacy of data is another huge concern, and one that increases in the context of Big Data. Managing privacy is both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data. There are many additional challenging research problems. For, example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. Yet another important direction is to rethink security for information sharing in Big Data use cases.

Human Collaboration: In spite of the tremendous advance made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-fields of visual analytics is attempting to do this, at least with respect to the analysis pipeline.

Data Analysis Platforms: A data analysis platform should support the complete data analysis pipeline. Therefore, such a platform should gather data from a variety of sources, blend it together, and then analyze it. Data may come from structured data sources such as a data warehouse and business transaction systems, from multi-structured data sources like document management systems and web-based platforms, or from sensors on intelligent hardware devices. When large volumes of raw multi-structured data are involved the source data are pre-processed by a data refinery prior to it being used by the data analysis platform. Many organizations are beginning to build data refineries running on systems such as Hadoop as a cost-effective way of managing and transforming large volumes of raw data.

One of the main differences in a data analysis platform, as compared with a traditional data warehouse workflow, is that the information worker can blend, explore, analyze and visualize data in different ways without the need for rigid pre-defined data schemas and data integration workflows. This flexibility is provided by a data analysis workbench, which includes a set of tools that allow information workers to dynamically build data schemas and blend together as it read from various source systems. The data analysis workbench also provides a variety of different tools for analyzing and visualizing the blended data. It may, for example, include tools for OLAP, statistical and text analysis, forecasting, predictive modeling and analysis, and/or optimization.

Conclusion: Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines. However, many technical challenges must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation.