2nd RDA Europe Science Workshop – Tentative Agenda and Topics

Peter Wittenburg, Raphael Ritz, Jamie Shiers, Leif Laaksonen, Herman Stehouwer

1. Introduction

The purpose of the RDA science workshops is to listen to ideas and feedback on the issues around the utilization and interoperability of research data from a selected group of leading scientists as one group of stakeholders as an input to the RDA Europe in general and the RDA global activities in particular. RDA is also interacting with other stakeholders such as data professionals of different favors, policy makers and research organizations to get input on urgent needs and priorities.

As for the first workshop of this sort the organizers will produce a summary document after the workshop with the main outcome to be taken further in Europe or on the global scale.

This note is about suggesting an agenda and topics to be addressed at the 2nd EU Science Workshop at CERN/Geneva at April, 8/9th 2015. It reflects the fact that RDA has some first concrete results and some new activities that a few new documents have been produced and that funders such as EC and NSF for example asked for recommendations for actions. In the appendix we briefly summarize the results. In particular we are referring to the following documents:

1. Report of the 1st RDA EU Science Workshop
   https://europe.rd-alliance.org/documents/publications-reports/rda-europe-science-workshop-report
2. Report of the 1st RDA US Science Workshop
   https://rd-alliance.org/groups/rdaus-science-meeting/wiki/final-draft-rdaus-science-workshop-report.html
3. Data Harvest Report
   https://europe.rd-alliance.org/documents/publications-reports/data-harvest-how-sharing-research-data-can-yield-knowledge-jobs-and
4. Data Practices Report
   https://europe.rd-alliance.org/documents/articles-interviews/rda-europe-data-practice-analysis
   http://media.icordi.eu/Repository/document/Interviews%20and%20articles/RDA%20Europe%20Survey_web.pdf
5. 2-page Flyers on RDA Results, etc. (see attachments)
   http://europe.rd-alliance.org/news/rda-working-groups-first-outputs
6. RDA group activities
   https://rd-alliance.org/

These documents will be in the focus of the suggested sessions, i.e. we suggest to organize the meeting so that we (1) briefly present the essentials of the first 4 reports and open the commenting and discussions on them, (2) present the current results and overarching activities of the first 2 years of RDA work and open the commenting and discussions on them, (3) leave session 3 for all sorts of statements on data issues in science the participants want to make and (4) finally discuss recommendations that may come out of the meeting. In this last session we will also briefly present the state of the EU-US consultations on impact of the RDA work which includes a number of concrete recommendations.

To minimize the effort we suggest splitting into two groups – one group reading and commenting on the papers 2 and 3 and the other group reading about the RDA results/activities (flyers and other material about the groups). All participants are kindly requested to look at the two workshop reports and are asked to make general statements on urgent data issues they would like to be taken up by RDA.

With respect to all sessions we will provide a number of reduced statements and questions which may help to structure the discussion. We will also ask you to come up with some statements before the meeting for the same purpose.

2. Tentative Agenda

11.00 Sergio Bertolucci Welcome

Session 1 – Report Summaries; Chair: Sergio Bertolucci

11.15	Bernard Schutz	Summary of 1st RDA EU Science WS
11.30	Rob Pennington	Summary of 1st RDA US Science WS
11.50	Leif Laaksonen	Summary of Data Harvest Report
12.10	Peter Wittenburg	Summary of Data Practices Report
12.30	Participants	QA & Discussion about Reports
13.00	*Lunch at CERN*	

Session 2 – RDA Activities; Chair: Bernard Schutz

14.00	Participants	Statements & Discussion – Policy Issues
15.30	*Coffee Break*	
15.45	Peter Wittenburg	Overarching RDA Activities and Landscape
16.00	Raphael Ritz	Concrete RDA Results and their impact
16.20	Participants	Statements & Discussion – RDA Issues
17.30	*Visit LHC at CERN*	
19.00	*Dinner*	

Session 3 – Wrap Up & General Comments; Chair: Bernard Schutz

9.00	Leif Laaksonen	Open Topics from Day 1
9.15	Participants	Statements & Discussion – Data Science
10.30	*Coffee Break*	

Session 4 – Recommendations; Chair: John Wood, Peter Wittenburg

10.45	Peter Wittenburg	EC-EU Recommendations
11.15	Participants	Q&A, – Discussion Policy Recommendations RDA Recommendations Recommendations for concrete actions
13.00	*End of Workshop* *Lunch at CERN* *Visit LHC at CERN*	

3. Participants & Guests

name	field	affiliation	country	Policy/ Practices	RDA Activities	Data Science
Cecile Callou	archezoology	Museum National d'Histoire Naturelle	FR	x		x
Massimo Cocco	earth science	INGV Rome	IT		x	x
Markku Kulmala	athmosphere	U Helsinki	FI	x		x
Ari Asmi	athmosphere	U Helsinki	FI		x	x
Gerd Grasshoff	hist. of sci, philosophy	Humboldt U, Berlin	DE	x		x
Nicolas Le Novere	systems biology	Babraham Institute Cambridge	UK	x		x
Barend Mons	bioinformatics	U Leiden	NL		x	x
Bernard Schutz	gravitation/ astronomy	U Cardiff	UK	x		x
Walter Thiel	theoretical chemist	MPG	DE	x		x
Gabriel Aeppli	physics/nano	ETH Zurich	CH		x	x
Jan Bjaalie	neuroinf	U Oslo	NO		x	x
Katrin Amunds	neuroinf	FZ Jülich	DE	x		x
John Nerbonne	humanities	U Groningen	NL		x	x
Thomas Lippert	physics/brain simulation	FZ Jülich	DE		x	x
Paul Olivier Dehaye	mathematics	U Zurich	CH	x		x
Tom Connor	immunology	U Cardiff	UK	x		x
Sergio Bertolucci	high energy physics	CERN	CH		x	x
				comment on existing papers and docs	comment on RDA activities	mention 5 main points for data science

Name	Field	affiliation
Doris Wedlich	RDA bioinformatics	KIT
Francoise Genova	RDA astronomy	CNRS
Peter Wittenburg	RDA	MPG
Leif Laaksonen	RDA	CSC
Jamie Shiers	RDA/CERN	CERN
Raphael RItz	neuroinformatics	MPG
John Wood	material physics	ACU
Cees de Laat	EU-US Liaison	U Amsterdam
Rob Pennington	EU-US Liaison	NCSA
Herman Stehouwer	RDA support	MPG

The "guests" are mentioned in the left table and they are participating for example due to their involvement in RDA. There is no difference with respect to the participation in the discussions during the workshop, however, guests will not be asked to make prepared statements on the various aspects of the RDA work.

Appendix - Report on recent RDA Activities

The first EU Science Workshop was organized in February 2014 in collaboration between RDA and the Max Planck Society. Quite some activities have been undertaken in the meantime. Here we want to refer to a few major activities.

1. Response to 1st EU Science Workshop[1]

The first EU Science Workshop came out with a number of recommendations which are listed here (RDA responses are in italics):

1. RDA can play an important role if it is able to come up with recommendations, API specifications, guidelines, etc. that help to overcome the many one-shot, point solutions currently being implemented and hence make infrastructure building more cost-effective.
   The first RDA results have been presented at the 4th plenary indicating a quick start and amongst others two big domains of activities have been crystallized much clearer and excellent experts are being engaged: (1) all activities around the daily scientific data creation and consumption machinery in the labs and making this work much more efficient. (2) all activities around data publication and citation. Therefore we believe that RDA is on a good way.

2. RDA must indeed be a bottom-up organization, and needs to strike the right balance between bottom-up and its current, rather top-heavy, state.
   The impression may be still that RDA is too much a top-down activity which partly has to do with the wish to have a quick start compared to the Internet history. However, all working and interest group activities are driven bottom-up by data professionals of different types who want to overcome barriers. The initially nominated Technical Advisory Board members have already been replaced by elected members and we are moving towards the next step, that members of the organization board and the council will be elected by the registered RDA activists. Most important, however, is that the process of creating concrete results is driven bottom-up. We need to take care that this line is being followed.

3. RDA must motivate a "middle layer" of data scientists and to get engaged, rather than hope for too much engagement from leading researchers.
   This message was well-understood and at least in Europe we are focussing indeed on engaging the "data scientists" which are mostly the middle layer people who do the data work in the labs. About 120 interactions with scientists and data experts and an increasing amount of national/regional meetings in the last two years show that we did a lot to engage and include experts.

4. RDA must be aware that it may find itself in a race towards specifications and solutions with big commercial players who may win with de facto standards, simply because they arrive first.
   This will always be a critical point we need to look at since industry will always try to achieve a competitive advantage and set de facto standards. For us it is of great importance to involve industry at a very early moment to include them in the specification work. In the last months two events have been used to engage industry: a workshop in Paris and a special session at the 4th plenary. This needs to be intensified and in the RDA Europe 2 (now running) and 3 proposals (to come after summer 2015) we reserved funds for activities involving industry led by two companies.

5. Expectations RDA has to meet:

[1] https://rd-alliance.org/rda-europe-science-workshop-report-european-scientists-view-research-data-and-corresponding.html

a) Invest in training younger generations of data scientists.
b) Push demo projects, act as a clearing house and should be able to give advice on data management, access and re-use to everyone in research.
c) Have data experts who can visit institutes and help them implement solutions.
d) Perform good quality assessment on the first working-group results due in September 2014, and take care to not fall into the trap of overselling

In the RDA Europe 3 proposal starting in September 2015 we suggested to invest considerable amounts of funds in exactly the recommended activities. (a) We reserved funds to train the next generation of experts by a variety of means (datathon, trips to plenaries, training courses, education, etc.) and in particular we reserved funds to engage a set of young people in RDA to a large fraction of their work time. (b, c) A group of senior experts and this team of young experts will be available to help, give guidance and advice, visit institutes, etc. to get the RDA results into operation and thus help changing current practices. (d) Ensuring quality of results and not overselling them will be an issue in the future. In particular the way how RDA results are transmitted to the researchers will be crucial, since different languages, styles and habits need to be taken into account. The first results have been produced but need some more work to make the useful for practice. The attached flyers give an impression, how we feel dissemination should be done.

In general we can say that we have taken the Science Workshop recommendations very serious and have put them on RDA's agenda.

2. Plenaries and Results

In March we had the 3rd plenary in Dublin and in September the 4th plenary in Amsterdam. For both plenaries we could identify that the number of data experts being engaged in RDA discussions increased. We now have about 1800 registered members and in particular in the Amsterdam meeting we could identify that even more data experts with deep knowledge from different scientific disciplines were participating leading to deep going discussions.

The first 4 Working Groups presented their results which are briefly summarized here:

- The **Data Foundation and Terminology** group worked out a basic model and basic terminology for the core of the data organization principles. Agreeing on a model of these core principles will improve efficiency when working with data. The group got some final comments and will finish their work until November.
  For a simple description see the attachment.
- The **Data Type Registry** group worked out a specification for registries that will help researchers to easily find tools to work with when they encounter a new unknown data type. This is a common scenario that one gets for example a file and does not know what to do with it. The first implementation will also become ready until the end of 2014.
  For a simple description see the attachment.
- The **PID Information Type** group worked on a unified programming interface (API) to register and resolve persistent identifiers (PID) that are associated with additional information. This is important since it is widely agreed now that PIDs are an ideal way to establish trust in the data in particular when one can associate identity and integrity information with it. However all PID service providers need to agree to register their information types in a Data Type Registry to establish interoperability. The core set of information types has been specified and the API has been developed.
  For a simple description see the attachment.
- The **Practical Policy** group collected areas where practical policies are being applied such as replication, preservation, etc. and in selected areas started collecting such practical policies

from a variety of repositories and projects. The goal is now to evaluate these examples and extract best practices that can be adopted by every repository to make data management much more trustful, allow certification and increase efficiency. Due to some unexpected event the work of the group was delayed so that they got an extension of 6 months to finish their first work on the selected policy domains.
For a simple description see the attachment.

Two other major outcomes of the Amsterdam meeting were

a) the intensification of the work of the experts dealing with question around **data publishing, citation** etc. 4 working groups have been setup and are now working on concrete results of how to streamline data publishing and how to make it available for everyone and
b) the start of a group dealing with the **data fabric**[2], i.e. the needs to make the daily data creation and consumption machinery in the labs much more efficient. What are the components and services that are needed to establish an efficient way of dealing with the huge amounts of data objects that we are creating in data driven science and to come to reproducible science which currently is not given in most cases as surveys have shown. This group was initiated by the core people in the early working groups, since they all understood that they are working on various edges of the same overall landscape of components. For more information on the Data Fabric ideas see the attachment.

Also at the Amsterdam plenary we had two sessions with the title "Interactions with Sciences". In the first session two ideas were being discussed: (1) **Trusted Open Service Agora** (TOSA) for data and services and (2) the nature of the **Data Fabric**. The **Trusted Open Service Agora** for data and services is highly required to allow researchers to easily find data and services/tools they can use and this across disciplines and countries. It is widely agreed that establishing such a **TOSA** is not trivial, but that we should start elaborating and piloting on this now based on existing initiatives and implementations. In the second session the results of the two Science Workshops (Europe, US) were presented and discussed.

3. 1st US Science WS[3]

The RDA US colleagues organized also a 1st Science Workshop in August 2014 with similar goals: comment on RDA work and give inspirations for priorities for RDA and also for US infrastructure needs. They restricted participation first on a few communities only and also invite some infrastructure providers. The major topics they addressed were (a) persistence, (b) sustainability, (c) tools, (d) discovery, (e) ease-of-use, (f) metadata, (g) data infrastructures, (h) education, (i) technology trends, (j) workflows and (k) provenance. For each of these topics the workshop formulated recommendations.
Compared with the EU Science Workshop that gave room for broad discussions and formulated concrete recommendations for the RDA process, our US colleagues made a number of concrete statements about urgently required measures to improve data practices.

4. The Data Harvest

In 2010 the High-Level Expert Group on Scientific Data handed over its report called "Riding the Wave"[4] to the commission and it had considerable effects on the European funding programs in so far as data infrastructure projects were started and as there was a request to foster global interaction on harmonization efforts in the area of data. This helped the EC to support the RDA initiative. After 4 years and in the view of the H2020 program it was obvious that a follow-up report is needed that describes the needs for the following phase. This follow-up report has the title "The

[2] https://rd-alliance.org/group/data-fabric-ig.html
[3] https://rd-alliance.org/groups/rdaus-science-meeting/wiki/final-draft-rdaus-science-workshop-report.html
[4] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Data Harvest"[5] to indicate that we now need to move to make use of the changes that have been initiated.

The claim that is being made that similar to the appearance of Internet we are at the start of a new wave of opportunities and that as a consequence nature of science will change towards a global data commons, a virtual and global science library. A number of recommendations are being extracted for policy makers at European and member states level such as asking for plans how to deal with data, promote data literacy at all levels across society, develop incentives for data sharing, develop tools and policies to establish trust as a key-point for increased data sharing and support global collaboration with respect to harmonization efforts.

5. Survey on Data Practices[6]

During the last 24 months two projects, RDA Europe and EUDAT, did a lot of effort to understand the practices with respect to data in the institutes, departments and projects across many disciplines. 40 interviews with data professionals were carried out and experts participated in more than 70 community meetings all devoted to a large extent on data issues. Here we want to mention a few of the major impressions from all interactions:

- The infrastructure projects (research infrastructures and e-infrastructures) had an enormous impact on the awareness about data issues in a number of disciplines.
- Open Access is widely supported, but there are a number of issues hampering open access which are often not mentioned such as bad state of data, legacy formats and unclear rights situation.
- Trust in its many facets is key for progress in data sharing and a chain of trust building mechanisms involving the various actors is needed.
- There is an enormous amount of legacy data around and due to not appropriate methods and tools we are still creating legacy data which will cost an enormous amount of effort to make it part of the sharable domain of registered data. Many senior domain experts are aware of this, but hesitate to invest due to a lack of widely accepted agreements, lack of experts to put better systems in place and lack of ready-made software.
- Many departments see the need to step into Big Data like scenarios and start using manual and ad hoc script based workflows. These are not appropriate, require an enormous data management effort and do not lead to reproducible data science. Automatic workflows are hardly applied due to a lack of experts and a doubt whether such workflows are flexible enough to handle all kinds of exceptions.
- Data Management costs a lot of time of highly qualified scientists and thus is very inefficient and cost-intensive.
- Practices with respect to metadata are still far from being satisfying. It requires additional efforts which are not taken and there is a lack of tools supporting easy MD creation at the very beginning already.
- There is a lack of explicitness of structural and semantic information hampering re-usage of files from other projects, disciplines, etc.
- Stable "centers" are crucial for the data landscape since they have the capability of offering persistent and reliable services to scientists.
- It is obvious that we lack data professionals of different facets (data scientists, data managers, etc.) and that this hampers progress in making data stewardship more professional.

---

[5] http://europe.rd-alliance.org/documents/publications-reports/data-harvest-how-sharing-research-data-can-yield-knowledge-jobs-and
[6] https://europe.rd-alliance.org/documents/articles-interviews/rda-europe-data-practice-analysis

- For normal researchers it is very difficult to get "trusted" information about all kinds of re-usable data and tools/services, since they do not have the time to try out all components offered via the web.

For a summary see the attachment.

Data Foundation and Terminology Working Group

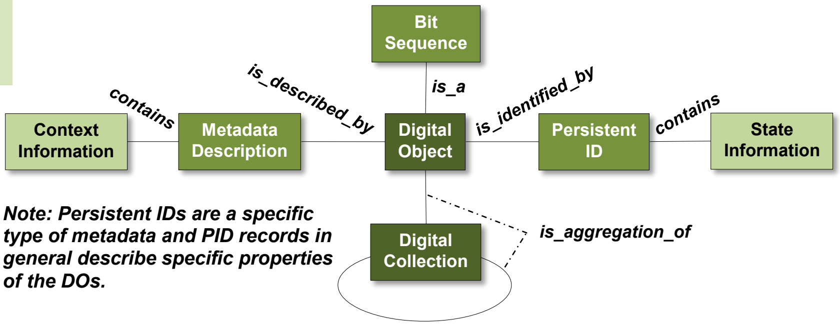
Responsible RDA Working Group Co-Chairs:
Gary Berg-Cross – Research Data Alliance Advisory Council, Washington D.C. USA
Raphael Ritz - Max Planck Institute for Plasma Physics, Germany
Peter Wittenburg – Max Planck Institute for Psycholinguistics, Germany

What is the Problem?

Unlike the domain of computer networks where the TCP/IP and ISO/OSI models serve as a common reference point for everyone, there is no common model for data organisation, which leads to the fragmentation we are currently seeing everywhere in the data domain. Not having a common language between data communities, means that working with data is very inefficient and costly, especially when integrating cross-disciplinary data. As Bob Kahn, one of the Fathers of the Internet, has said, "Before you can harmonise things, you first need to understand what you are talking about."

When talking about data or designing data systems, we speak different languages and follow different organization principles, which in the end, result in enormous inefficiencies and costs. We urgently need to overcome these barriers to reduce costs

For the physical layer of data organisations, there is a clear trend towards convergence to simpler interfaces (from file systems to SWIFT-like interfaces[7]). For the virtual layer information, which includes persistent identifiers, metadata of different types including provenance information, rights information, relations between digital objects, etc., there are endless solutions that create enormous hurdles when federating. To give an idea of the scale of the problem, almost every new data project designs yet more new data organisations and management solutions.



Note: Persistent IDs are a specific type of metadata and PID records in general describe specific properties of the DOs.

We are witnessing increasing awareness of the fact that at a certain level of abstraction, the organisation and management of data is independent of its content. Thus, we need to seriously change the way we are creating and dealing with data to increase efficiency and cost-effectiveness.

This diagram describes the essentials of the basic data model that the DFT group worked out in a simplified way. Agreeing on some basic principles and terms would already make a lot of difference in data practices.

What were the goals?

The goals of this Working Group (WG) were:

- Pushing the discussion in the data community towards an agreed basic core model and some bas ic principles that will harmonize the data organization solutions.
- Fostering an RDA community culture by agreeing on basic terminology arising from agreed upon reference models.

7 https://wiki.openstack.org/wiki/Swift

What is the solution?

Based on 21 data models presented by experts coming from different disciplines and about 120 interviews and interactions with different scientists and scientific departments, the DFT WG has defined a number of simple definitions for digital data in a registered[8] domain based on an agreed conceptualisation.

These definitions include for example:

- **Digital Object** is a sequence of bits that is identified by a persistent identifier and described by m etadata.
- **Persistent Identifier** is a long-lasting string that uniquely identifies a Digital Object and that can be persistently resolved to meaningful state information about the identified digital object (such as checksum, multiple access paths, references to contextual information etc.).
- A **Metadata description** contains contextual and provenance information about a Digital Object that is important to find, access and interpret it.
- A **Digital Collection** is an aggregation of digital objects that is identified by a persistent identifier and described by metadata. A Digital Collection is a (complex) Digital Object.

A number of such basic terms have been defined and put into relation with each other in a way that can be seen as spanning a reference model of the core of the data organisations.

What is the impact?

The following benefits will come from wide adoption of a harmonized terminology which will be expanded stepwise:

- Members of the data community from different disciplines can interact more easily with each ot her and come to a common understanding more rapidly.
- Developers can design data management and processing software systems enabling much easier exchange and integration of data from their colleagues in particular in a cross-disciplinary setting (full data replication for example could be efficiently done if we can agree on basic organization principles for data).
- It will be easier to specify simple and standard APIs to request useful and relevant information re lated to a specific Digital Object. Software developers would be motivated to integrate APIs from the beginning and thus facilitate data re-use, which currently is almost impossible without using information that is exchanged between people.
- It will bring us a step closer to automating data processing where we can all rely on self-docume nting data manipulation processes and thus on reproducible data science.

When can we use this?

The definitions have been discussed at RDA Plenary 4 meeting (Sept 2014) and will become available as a document and on a semantic wiki to invite comments and usage at January 2015. RDA and the group members will take care of proper maintenance of the definitions. For more information see https://rd-alliance.org/group/data-foundation-and-terminology-wg.html and http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

In the next phase of the work, more terms will be defined and interested individuals will have the opportunity to comment via the semantic wiki.

8 There will always exist data in private, temporary stores, which will not be made accessible in a standard way.

Data Type Registries Working Group

Responsible RDA Working Group Co-Chairs:
Larry Lannom - Corporation for National Research Initiatives, Virginia USA
Daan Broeder - Max Planck Institute for Psycholinguistics, Netherlands

What is the Problem?

Often researchers receive a file from colleagues, follow a link, or otherwise encounter data created elsewhere that they would like to make use of in their own work. However, they may not know how to work with it, interpret it or visualise its content, being unfamiliar with the specifics of the structure and/or meaning of the data, ranging from individual observations up to complex data sets. Frequently, researchers need to stop here since it requires too much work to look for explanations, tools, and where tools exist, install them.

When sharing data across disciplines, we often get files which we cannot process easily. Dragging such a file on the DTR would immediately yield results and reduce effort.

What was the goal?

The goal of the DTR WG was to allow data producers to record the implicit details of their data in the form of Data Types and to associate those Types, each uniquely identified, with different instances of datasets. Data consumers can then resolve the Type identifiers to Type information for gaining knowledge of the implicit assumptions in the data, finding available services that can be used for this kind of data, and any other useful information that can be used to understand and process the data, without additional support from data producers. DTRs are meant to provide machine-readable information, in addition to presenting human readable information.

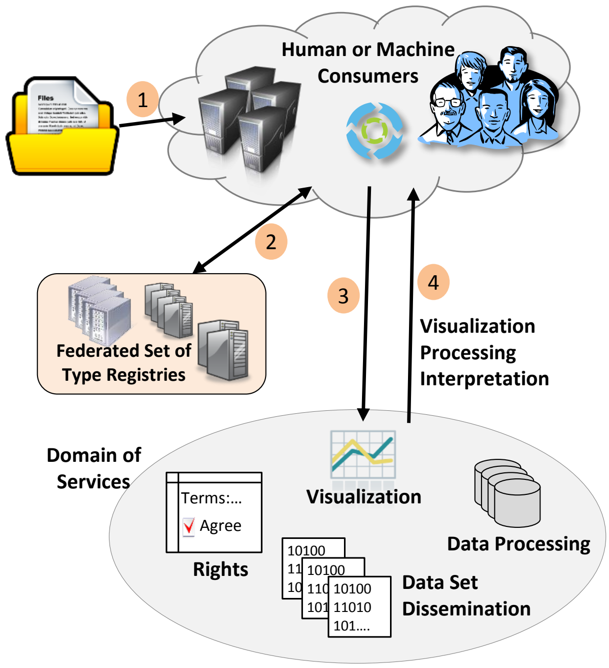
What is the solution?

DTRs offer developers or researchers the ability to add their type definitions in an open registry and, where useful, add references to tools that can operate on them. For example, a user who received an unknown file could query a DTR and receive back a pointer to a visualisation service able to display the data in a useful form. A fully automated system could use a DTR, much like the MIME type system enables the automatic start of a video player in the browser once a video file has been identified. We envision humans taking advantage of Data Types in DTRs through the type definitions that clarify the nuanced and contextual aspects of structured datasets.

Data Types in DTRs can be used to extend or expand existing types, e.g., MIME types, which provide only container-level parsing information. They can additionally describe experimental context, relationships between different portions of data, and so on. Data Types are deliberately intended to be quite open in terms of registration policies.

Two examples may illustrate the benefits of the DTR solution:

1. Researchers dealing with data (e.g. in a cross-disciplinary, cross-border context) find an unknow n data type and can immediately process and/or visualize its content by using the DTR service.
2. Machines that want to extract the checksum information of a data object from a PID record to c heck whether the content is still the same. Without knowing the details of the PID service provid er, the machine could ask for CKSM for example, since this is an information type which all PID s ervice providers agreed upon and registered in the DTR.



This diagram indicates how the Data Type Registry (DTR) is working. A user or machine receives an unknown type (1) which can be a file or a term for example. The DTR is contacted and returns information about an available service (2) that will allow the user or machine to continue processing the content (3, 4) such as visualizing an image without asking prior knowledge from the user. This will make cross-disciplinary and cross-border work much more efficient and enable data driven science even to those who are

What is the impact?

The potential impact on scientific practices is substantial. Unknown data types as described above can be exploited without any prior knowledge and thus an enormous gain in time and/or in interoperability can be achieved. In a similar way to the MIME types that allow browsers to automatically select visualization software plug-ins when confronted with a certain file type extension, scientific software can make use of the definitions and pointers stored in the DTR to continue processing without the user acquiring knowledge beforehand. DTRs pave the way to automatic processing in our data domain, which is becoming increasing complex, without putting additional load on the researchers.

Of course, a price needs to be paid in that type creators need to enter the required information into a DTR. We assume that there will be a federation of such DTRs setup to satisfy different needs.

When can we use this?

The first groups are building software to implement such a DTR concept and make the software available. The RDA PID Information Type (PIT) Working Group is already using the first DTR prototype version in its API. The latest version of a DTR prototype is made available here: http://typeregistry.org/. We expect software to become available for download around the end of 2014. Please check the information on the DTR WG's web page at https://www.rd-alliance.org/group/data-type-registries-wg.html for updates.

This simple model will be the start for designing DTRs, with the intention to extend the specifications according to priorities and usage.

PID Information Types Working Group

RDA Working Group Co-Chairs:
Tobias Weigel – DKRZ, Germany
Timothy Dilauro – John Hopkins University, Maryland, United States

What is the Problem?

Numerous systems and providers to register and resolve Persistent Identifiers (PIDs) for Digital Objects and other entities have been designed in the past and are used today. However, almost all of them differ in the way they allow researchers to associate additional information, such as for proving identity and integrity with the PID. For application developers this is an unacceptable situation, since for all providers a different Application Programming Interface (API) needs to be developed and maintained. Given that a researcher has found a useful file, but first wants to prove whether it is indeed the same stream of bits after some years, he should be able to request the checksum independent of the provider holding the PID. How should he do this not knowing whether the provider offers this information and if so, how to request it? We can overcome such extreme inefficiencies only if all providers agree on a common API, register their information types in a common data type registry and agree on some core types, such as the checksum.

Due to high demand, a variety of trusted PID service providers have been set up already, yet all of the different attributes associated with the registered PIDs make life of a software developer a nightmare. We need to harmonize the major information types and suggest a common API, so that if we request the checksum we simply have to program one piece of software

What were the goals?

The goals of this WG were:

- Coming to a core set of information types and register (and define) them in a commonly accessib le Data Type Registry
- Providing a common API and prototypical implementation to access PID records that employ registered types

What is the solution?

The PIT group accomplished the following:

- Defined and registered a number of core PID information types (such as checksum)
- Developed a model to structure these information types
- Provided an API, including a prototypical server implementation that offers services to request certain types associated with PID records by making use of registered types.

The set of core information types currently provided can help to illustrate cross-discipline usage scenarios. It can also act as an example for a community-driven governance process creating and governing more user-driven types. PID service providers and community experts need to come together regularly and add types to the data type registry to make full use of the possibilities of the results of the PIT group.

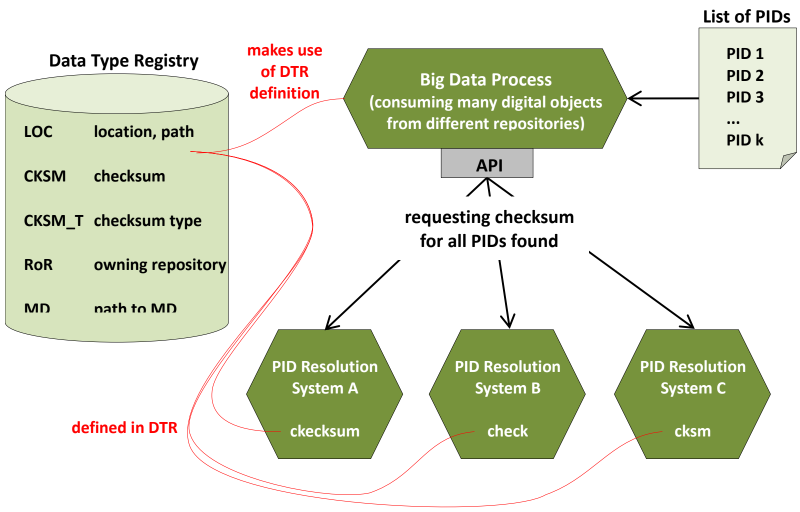
It is now essential to convince PID service providers such as those using the Handle System (DOI, EPIC, etc.) to adopt the API to unify access. In the diagram below, we give an example of the usage and potential of the suggested solution.

What is the impact?

We need to envisage the situation in a few years, when the amount and complexity of data has been increased in all sciences and there is a greater need to rely on automatic processes, as human intervention means loss of efficiency.

In such scenarios, communities can exploit the wealth of the data domain relying on semantic interoperability between all relevant actors for example for Big Data analytics. The above example is just one small usage scenario that would be enabled if the relevant PID service providers accept the results of the PIT WG and harmonize their approach. Application software writing would be reduced dramatically since only one API would be supported and one module would be sufficient for retrieving the checksum, for example, and checking identity and integrity.

Assume that you got a list of PIDs referring to data you want to use in a computation, that these PIDs are being registered at different providers and that you first want to check whether all data objects are still the same. You simply want to provide one module that reads a PID from the list and submits a request to the appropriate resolver to send the checksum. If all actors refer to the same entry in the DTR interoperability is given, i.e. one module would be sufficient to retrieve the checksums independent



The strengthening of PID information types could also move the existing identifier systems and the overall idea of identification into a more central and fundamental position as suggested by DFT's core model of a Digital Object, leading to an enormous increase in efficiency when dealing with data.

When can we use this?

First groups are building software to implement a first prototype based on the defined PIT API. This first prototype works together with the DTR prototype and both are publicly available, but not designed for production use. We expect another update of the prototypes to become available for download at the end of 2014.

Please check the information on the PIT group's web-page at https://www.rd-alliance.org/group/pid-information-types-wg.html.

It is now time to convince the PID service providers to adopt the solution.

Practical Policy Working Group

Responsible RDA Working Group Co-Chairs:
Reagan Moore, RENCI, North Carolina, USA
Rainer Stotzka, Karlsruhe Institute of Technology, Germany

What is the Problem?

Repositories' responsibilities of data stewardship and processing require a highly automated, safe and documented process. However, at this time, repositories design and implement these processes in a method that does not support this requirement.

Current practice in managing and processing data collections are determined by manual operations and ad-hoc scripts making verification of the results an almost impossible task. Establishing trust and a reproducible data science requires automatic procedures which are guided by practical policies. Collecting typical policies, evaluating them and providing best practice solutions will help all

With the increasing amount and complexity of data, repositories should not continue to use manual interventions and ad-hoc scripts any longer since they prevent us to establish trust.

All operations or chains of operations that have these capabilities and are enforced on collections of data objects should have "Practical Policies" (PP), which should be stated in simple languages and turned into robust and tested executable code. PPs are at the basis of reproducible science, an important element in the chain of building trust and one of the core elements in repository certification processes.

What were the goals?

The goals of this WG were:

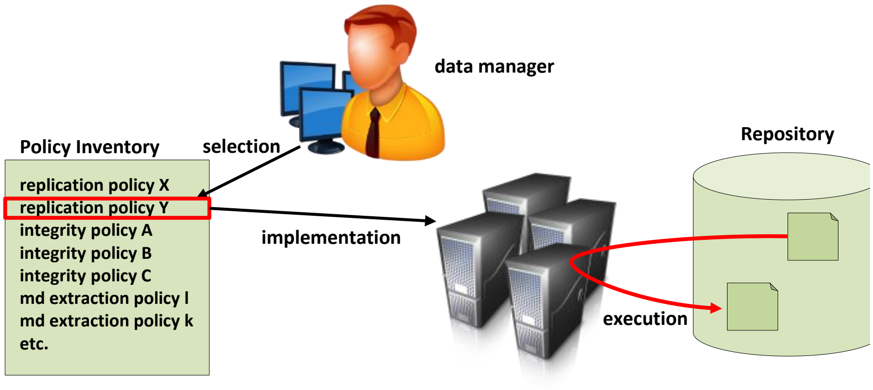
- Defining computer actionable PPs that enforce proper management and stewardship, automatin g administrative tasks, validating assessment criteria, and automating types of scientific data pro cessing
- Identifying typical application scenarios for practical policies such as replication, preservation, m etadata extraction, etc.
- Collecting, registering and comparing existing practical policies
- Enabling sharing, revising, adapting and re-using of such practical policies and thus harmonizing practices, learning from good examples and increasing trust

Since these goals were broad in scope, PP WG focused its efforts on a few application scenarios for the collection and registration process.

What is the solution?

In order to identify the most relevant areas of practice, the PP WG conducted a survey as a first step. The analysis of the survey resulted in 11 highly important policy areas which were tackled first by the WG: 1) contextual metadata extraction, 2) data access control, 3) data backup, 4) data formal control, 5) data retention, 6) disposition, 7) integrity (incl. replication), 8) notification, 9) restricted searching, 10) storage cost reports, and 11) use agreements.

Participants and interested experts were asked to describe their policy suggestions in simple semi-formal descriptions. With this information, the WG developed a 50-page document covering the



simple descriptions, the beginning of a conceptual analysis and a list of typical cases such as extract metadata from DICOM, FITS, netCDF or HDF files.

Due to unexpected circumstances, the WG will continue until Plenary 5 (March 2015). It will focus on further analysing, categorising and describing the offered policies. Currently, volunteers are reviewing the policies and different groups have started to implement some of these policies in environments such as iRODS and GPFS. The goal is to register prototypical policies with suitable metadata so that people can easily find what they are looking for and re-use what they found at abstract, declarative or even at code level. At this point, there is still much work to be done to reach a stage where the policies can be easily used.

What is the impact?

The impact is huge. In the ideal case, data managers or data scientists can simply plug-in useful code into their workflow chains to carry out operations at a qualitatively high level. This will improve the quality of all operations on data collections and thus increase trust and simplify quality assessments. Large data federation initiatives such as EUDAT and DATANET Federation Consortium (US) are very active in this group, since they also expect to share code development/maintenance, thus saving considerable effort by re-using tested software components. Research Infrastructure experts that need to maintain community repositories can simply re-use best practice suggestions, thus avoiding ending up in traps. In particular, when these best practice suggestions for practical policies are combined with proper data organisations, as suggested by the Data Foundation and Terminology Working Group, powerful mechanisms will be in place to simplify the data landscape and make federating data much more cost-effective.

The diagram indicates the final goal of the PP WG. A policy inventory will be made available with best practices examples. Data managers will have the ability to select and implement the procedures most relevant to them.

When can we use this?

The document mentioned above already provides a valuable resource to get inspiration and perhaps make use of suggested policies, thus improving people's own ideas or to even making profit from developed code.

Once evaluated, properly categorised and described, the real step ahead will be registering practical policies in suitable registries, so that data professionals can easily re-use them, if possible even at code level. The group intends to progress to this step by the end of March 2015 for a number of policy areas, making use of the policy registry developed by EUDAT.

For more details on the PP WG, see https://www.rd-alliance.org/group/practical-policy-wg.html

Revolutionising Data Practices

Gary Berg-Cross, Keith Jeffery, Rob Pennington, Peter Wittenburg

What is the Problem?

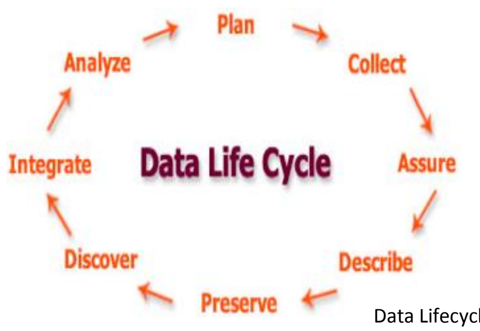
A large survey from mainly RDA Europe and EUDAT (including about 120 interviews and interactions with data professionals from various departments engaged in various research disciplines) demonstrated that the way we manage and process data is very inefficient and too expensive. In addition, data science generally is not reproducible as some reports have shown which is contrary to good practices and thus not acceptable.

The task of DFIG is to design a flexible and dynamic framework of essential components and services, identifying those that enable efficient, cost-effective and reproducible data science and making these known and available to researchers and data scientists. The goal is to make it possible for scientific users to easily integrate their scientific algorithms into such a data fabric without needing to master the underlying details.

Despite insights from computer science and excellent individual solutions from advanced infrastructure projects we lack a broad and systematic approach to understand the components, their services and their interfaces that are needed to change our data practices in a way that the deficits will be overcome and to make them available to every researcher. A number of RDA groups are working already on such components, yet doing it in a somewhat isolated way. There is a wide agreement that this needs to be changed urgently.

What are the Goals?

The Data Fabric Interest Group (DFIG) has been setup to address the design of such a framework as a whole, to locate the various activities on the landscape of components, to indicate gaps and to understand how the various groups need to interact to come to an interoperable flexible framework. The intention is thus not to design a relatively fixed architecture of a system that fulfills a particular set of functions, but a flexible framework that can be configured by changing components to meet varying needs, and thus is technology-independent. The framework identifies the minimal set of components required to let any system based on the framework function.



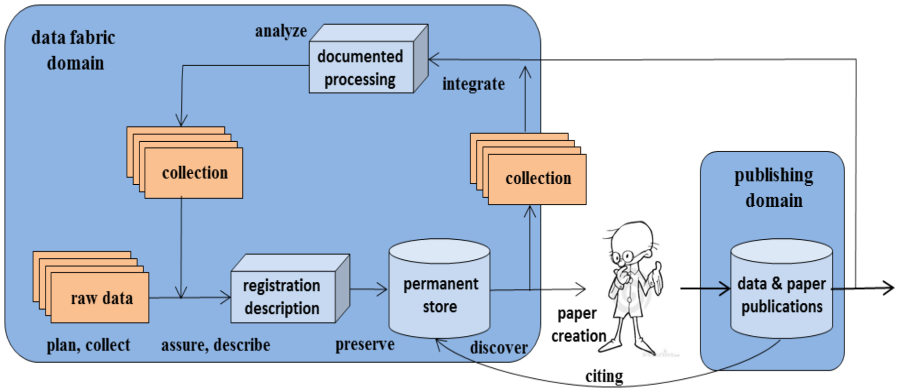
Data Lifecycle

To meet these goals we need to analyse large scale lighthouse infrastructure projects - which are mostly discipline-based developed exemplary solutions - and identify commonalities. DFIG does not start from scratch, but can build on the knowledge already gathered.

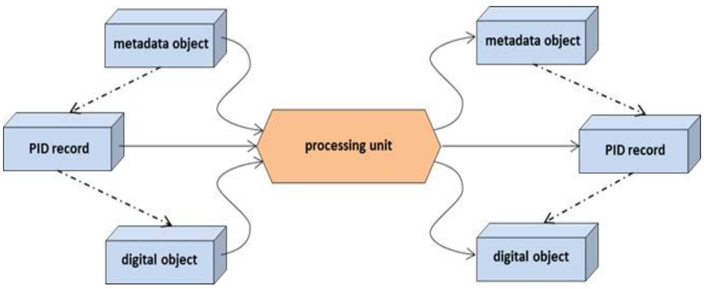
DFIG also needs to look at all phases of the lifecycle as schematically indicated by the diagram above.

What is the Solution?

DFIG needs to define a basic and flexible machinery framework that (when implemented as systems) makes data science reproducible, fulfils the G8+O5 recommendations and the need to carry out data management and processing much more efficiently. Recognizing that data intensive science is faced with increasingly large volumes and complexity of data we need to turn to processing which is guided by actionable and documented policies, in which all steps adhere to basic organizational principles are self-documenting, i.e. provide provenance metadata and are (as much as feasible) autonomic.



The diagram above indicates the data machinery which is being executed in some form in all data intensive scientific work. The relations to the phases in the previous diagram are indicated. Raw data (which can also be long tail data created on a notebook) will be brought into the accessible domain of data by registering it (assigning Persistent Identifiers), describing it by metadata and depositing it into a permanent and accessible repository which will be distributed. Using metadata scientists will now create new (virtual) collections by making selections which then will be subject to some kind of processing – be it management, curation or analytic. New collections are being created that which again are described, registered and deposited.



If all processing steps follow principles as schematically indicated above where new data and metadata is being generated extending the old objects, we will achieve the kind of self-documentation that is required. To unload the scientist DFIG needs to identify the components that are required to put such machinery in place and that allows researchers to simply plug-in their scientific algorithms so that they do not need to know about all the details of the machinery. We realize that achieving this, being compliant with the G8+O5 principles (searchable, accessible, interpretable, re-usable) and putting it in place so that everyone can take profit from it is a long road that requires a step-wise approach. But we need to start working on this today and convince software builders to follow these principles. RDA activities need to have this overall picture in mind where the act of publishing papers and data is an integrated phase requiring some explicit steps.

What is the Impact?

The impact of implementing such machinery based on a flexible framework is huge and will revolutionize data intensive science. It can be compared with optimizing the publication and citation machinery as we have seen over the past decades.

When can we use it?

Like with Internet where broad uptake happened about 15 to 20 years after the invention and optimization of the TCP/IP framework, RDA will stepwise optimize the way to deal with data in the various phases. Here the first working and interest groups in RDA take already now important steps and also large lighthouse infrastructure projects facing the inefficiencies daily have designed solutions which need to be analysed and considered carefully. Like with Internet we need to define the basic and essential components now that will allow us adding components and services

![RDA Research Data Alliance Europe]

RDA Europe Data Practice Analysis

Editors: Peter Wittenburg, Herman Stehouwer*

dependent on insights and technological advancements.

What did we do?

For the RDA Europe Data Practice Analysis Programme we held a large number of interviews with data scientists/practitioners from various communities. We interviewed these people about various aspects of their data environment including data acquisition, data processing, the computational environment, services and tools, and the data related policies being applied.

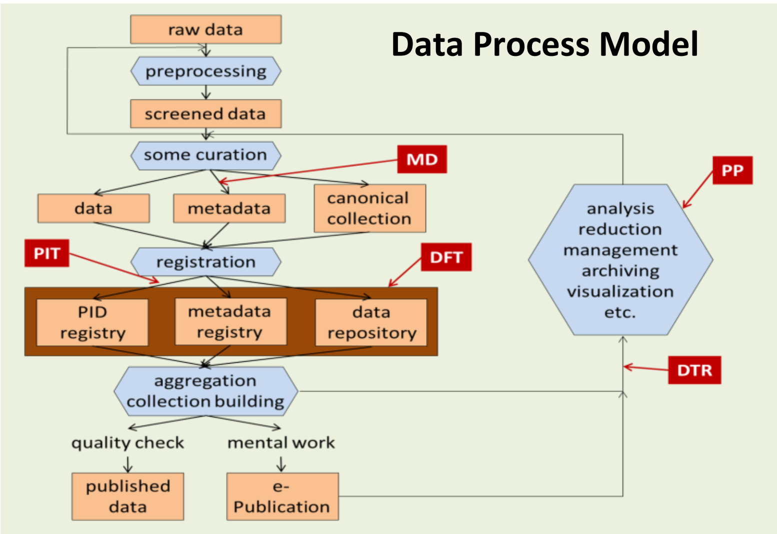
We interviewed 24 communities, and attended more than 70 community meetings. We combined these observations with the interviews and observations made in the EUDAT project, in the Radieschen project, and in the first RDA Europe Science Workshop. Based on these sources of information we came to a large number of observations, which are summarized here in form of the dominant underlying data process model, and 12 key observations.

Key Messages

Support Open Access
Ensure (Meta-) Data Quality
Explicit Structure & Semantics
Change to Documented Methods
Help Increasing Trust
Educate/Train Data Professionals

Data Process Model

The process model in the figure emerges as the dominant underlying process model that most data scientists/practitioners are implicitly using when processing data. In practice the methods used in the departments deviate slightly from this generic model in various ways, but it summarize what is being done at an abstract level very well. Furthermore, most often parts of the data processing are implicitly handcrafted with ad-hoc solutions rather than by following an explicit model.



The model helps us to clarify our observations and to identify specific steps as they relate to data, specifically: Data is scientifically meaningful and relevant after the pre-processing step; data is ready for upload to a repository after the curation step; data is ready for re-use after the registration step; and data is ready for citation after the publishing step. Currently most researchers do not distinguish between these steps explicitly. Explicitly separating these steps of the data process would increase efficiency and decrease cost.

This model shows similarities to existing models of data processing (such as the Kahn/Wilensky, CLARIN, EUDAT, ENVRI, EPOS, and DICE models[9]), and it can be used to place the observations made in the analysis program as well as to talk about a data management system. In the diagram we also placed where the topics of the first RDA Working Groups can be located.

12 Observations

1. **ESFRI** projects and the recent developments within **e-Infrastructure** have had a strong and positive influence on data management practices.
2. **Open Access** is supported everywhere as a basic recommendation. However in practise there are many barriers that still need to be lowered.
3. **Trustworthiness** is a key issue and new methods are urgently required to establish trust in the entire data processing chain.
4. **Legacy Data** is a problem in many communities, however even new data is often badly documented and organised, thus we are creating continuously new legacy data which will cost much effort to integrate them in the accessible data domain. There is 1) a lack of knowledge about principles of proper data organisation; 2) a lack of experts, time and money who could change practices; 3) a lack of off-the-shelf software methods for improved data management and access.
5. **Big Data** is driving many new scientific requirements that dictate the thorough adoption of this paradigm in increasing numbers of departments. However, big data only scales when data management and access methods are used that scale.
6. **Data Management** needs to move towards including the logical layer of information, i.e. metadata, PIDs, rights, relations to other data, etc. At the end the current file-system based methods are too inefficient and costly. A large amount of researchers' time is wasted in finding the right data objects, interpreting them and creating meaningful collections.

Establish Trust

Quality and Integrity of data
Availability of high-quality metadata
Sustainable services and PIDs
Clear Responsibilities and Funding

7. **Metadata** practise needs to be improved in order to help discovery and reuse (especially after some time). Guidance and ready-to-use packages and software are required to improve the situation.
8. **Lack of Explicitness** is an issue in relation to data, which hinders efficient machine-based processing of data. This lack ranges from non-registered digital objects (i.e. lacking PIDs), data integrity information (such as checksums), collection descriptions, 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.
9. **Centres** for managing data across communities are a clear trend. Such centres and repositories need to be established to provide a long-term reliable service to all researchers. Creating virtual collections or carrying out distributed processing jobs is still un unsolved issue. Some aspects of distributed authentication and authorization are still not in place at European level and

[9] references

distributed computing, although mentioned increasingly often, is not a well-understood scenario.

10. **Education & Training** is a clear need in order to address the lack of data professionals. This lack hampers changes and progress everywhere.
11. **Lack of Knowledge** and trusted information on services that are being offered (registries, data, storage, curation, analytics, etc.) is an issue. We have a large number of possibilities, but many can't cope with the information flood and have a hard time making selections. A more structured and trusted approach of offering information would have great impact.
12. **RDA** needs to ensure it is a true grass-roots organisation. It needs to provide demonstration cases, and give help and support to research communities.