DEUTSCHE INITIATIVE FÜR NETZWERKINFORMATION E.V.

Report **RDA-/DINI-Workshop**

Karlsruhe, Germany
May 28 – 29, 2015

Herman Stehouwer, Rainer Stotzka, Peter Wittenburg

This report summarizes the essential outcomes of the workshop held at May 28–29, 2015 in Karlsruhe that require actions. The workshop brought together 45 experts representing a variety of different institutions in particular also experts from small departments and libraries. In this workshop many experts participated for which RDA was new, i.e. the RDA presentations and the discussions included more aspects that were basic to RDA's functioning. Since the participants were mostly data practitioners from various background no time was spent on organizational/administrational matters.

We received quite a number of positive statements for the work so far and for what RDA is aiming at which we will not mention in this report, since this can be seen in the slides that can be found on the workshop site[1]. In this report we will mainly refer to points of critique or new ideas that should be taken up within RDA.

Goals & Agenda

The workshop had several goals:

- Inform about RDA and get feedback on the work of RDA so far.
- Start an interaction on what needs to be done next.
- What kind of components (as referred to in the Paris paper) need to be specified?
- Is RDA with its activities on the right way or do we need to adjust?
- Which collaboration projects should be started and what are the opportunities?

The agenda widely mirrored the goals of the workshop in so far as we had sessions that

- presented the RDA work so far (Working Group results, general overview, what is in the pipe)
- 10 experts presented views from their community or initiatives with respect to the points and questions mentioned above which are partly addressed by the Paris and two other documents[2]
- RDA EU and EUDAT experts presented possibilities of collaboration and support from September on[3]

[1] http://www.forschungsdaten.org/index.php/RDA-DE-DINI-Workshop-2015

[2] http://hdl.handle.net/11304/1aab3df4-f3ce-11e4-ac7e-860aa0063d1f
http://hdl.handle.net/11304/ea286e5a-f3d1-11e4-ac7e-860aa0063d1f
http://hdl.handle.net/11304/992fe6a0-fe34-11e4-8a18-f31aa6f4d448

[3] EGI Experts announced that they also are willing to engage in collaborations.

- allowed us to have discussions on many aspects.

General Points

- **RDA D and DINI**
  The collaboration between RDA D (D stands for Germany) and DINI to prepare this workshop was seen as positive and should be continued. The next RDA D meeting will be the RDA D plenary meeting in Potsdam in November (see below). It needs to be discussed with DINI whether they want to co-organize also that meeting.

- **Who is RDA**
  The difference between RDA Global, RDA Europe and RDA D was explained. While RDA Europe is a project that manages funds and human resources to mainly support the RDA by various activities, RDA Global is an initiative that is made up by all engaged people. Thus, RDA Global is all of us working in the research domain primarily and dealing with data in one or other ways. RDA D does currently not have any particular funds, however, funding support of meetings such as the one in November could be possible. RDA D participants can participate of course in opportunities offered by RDA Europe, however, balanced decisions will be required to serve several countries. Own German funds would of course be preferable given the broad interest in Germany.
- **RDA as Place for Exchange and Argument Collection**
  Some see RDA in particular as a platform to exchange ideas with others and use RDA outcomes to argue with institutional hierarchies and funders. Both are legitimate ways of acting.
- **Terminology**
  Also this meeting was again an example that terminology harmonization as DFT WG did it is very important, in particular if many participants were not involved in RDA discussions. It still is an issue to understand each other seamlessly when talking about data issues. It was suggested to add and discuss terminology and it was recommended to use DFT's term tool[4] for this purpose. An email to Thomas Zastrow (thomas.zastrow@rzg.mpg.de) is sufficient to be added as registered member.
- **Big vs. Small Initiatives**
  In contrast to the workshop in Amsterdam where many large and well-funded research infrastructure initiatives participated, this RDA D workshop made clear that there is a huge gap between the well-funded initiatives and institutions and those that are not so well-funded and thus do not have so many data professionals (or even none). These institutes need different types of help, support, clear guidelines and if possible ready-made solutions which they can easily integrate. Probably RDA E as support project needs to do special actions for these institutions and that also the well-funded initiatives help in offering their knowledge etc. There should be easy ways for the small groups to participate in some form in larger infrastructures. RDA can and should act here as mediator. In general it was obvious that service offers which already exist are not well-known.
- **Short vs. Long-term**
  This meeting with so many experts coming from small departments also made obvious that people are looking in particular for ready-made solutions and services now. We need to stress that RDA needs to address the challenges of the coming decade, however, it seems that we also need to act as a catalyzer for short-term help. The senior/junior team, which will be in place from September 2015 on, needs to react on these requests.
- **RDA as Clearing House**
  It is obvious that for many the "solution space" is simply overwhelming and that guidance is urgently necessary. People do not have time and do not want to test various solutions but want a neutral actor who can give advice on standards, software, etc. It was also mentioned that

[4] http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

currently there is too much of “Wildwuchs” (unguided growth) that adds to the “solution space” without convergence towards interoperability.

- **Expectations**
  The need for changes at various levels in the data domain is high, in particular since so many basic problems need to be addressed. RDA needs to be careful with its messages since it is and will remain a lean organization.

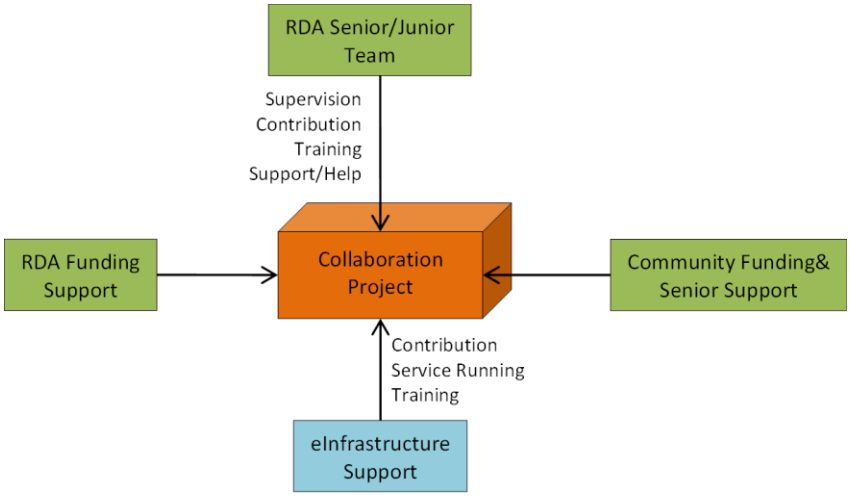
Collaborations

RDA

RDA presented its capabilities to give help and support and to start collaborations from September on. We need to carry out implementation projects that combine different components as partly been worked out by RDA WGs in close collaborations with interested partners from communities. RDA’s contribution should be at two levels to meet the needs: (a) give help, support and guidance in particular with the help of the senior/junior team and (b) use expert funds etc. to be able to contribute directly to collaborations.

Others

It was mentioned that EUDAT (a European data infrastructure project where 4 German data centers[5] are involved, www.eudat.eu) offers services and will probably also participate in collaborations. A call has been announced. It needs to be seen in how far the service offer needs to be extended. It was added that obviously also the Grid community may join in doing collaborations, however there is no German branch anymore. From OpenAIRE no signals have yet been given.



German Funding

It would of course be great to have German funding for collaborations on concrete implementation projects to guarantee that a new generation of experts is being trained and made fit for the markets.

Standard Bodies etc

It was argued that RDA should not re-invent the wheel and interact with other bodies like W3C, IETF, etc. RDA has been established official collaborations with many such global initiatives such as CODATA, WDS, W3C, IETF, etc. and indeed RDA’s policy is to refer to standards and best practices that have already been defined as long as they can be applied to the data domain. Many results from W3C for example (XML, RDF, ResourceSync, PROV, etc.) are being used within RDA. However, it is the primary task of the groups (and their chairs) within RDA to look around what is already existing when a new activity is being started and to include the experts from other initiatives.

General Topics

November RDA D Meeting

Given the funding situation we need to make optimal use of the coming November meeting in Potsdam. Plenary talks and parallel breakout sessions of different character (from training to

[5] FZ Jülich, RZ Garching, KIT Karlsruhe, DKRZ Hamburg.

advanced discussions should be planned dependent on the wishes of the participants). Possible topics may result from the list mentioned in this report and a questionnaire should find priorities. A PC has been established organizing the event: Jens Ludwig, Rainer Stotzka, Ralph Müller-Pfefferkorn, Hans Pfeiffenberger, Peter Wittenburg.

Policy and Principles Statement
RDA will take care that the German community will be involved in some form in the discussions about policy and principles documents that will be written at European level. It might be of interest to create German versions.
Principles being discussed are such as "all data to be shared should be assigned PIDs and metadata" and "all repositories for persistent data access should be certified". The policy paper would be a follow up of the "Data Harvest"[6] report probably aligned with the principles document. All documents need to focus on changing cultures and facilitating changing practices.

DMP Support
Also in Germany the creation of DMPs is a request from funders, but given the current practice these DMPs are more of bureaucratic acts rather than being productive. Help and guidance should be given. Some projects in D already give advice, but RDA D could have a more important role with specialized guidelines if there would be some funds. Currently we can think about organizing a workshop or session about this topic and take profit from what will happen at European level.

Synchronized Training
RDA Europe will organize training courses and seminars at different levels (incl. webinars) where German experts of course could participate. However specialized actions in Germany would make sense. Given the funding situation we can only manage to organize sessions at other meetings such as the one in November. We need to see how we can use the excellent knowledge that is available in German projects and in many cases the D language may remove barriers.

Usage Stories
For many participants the presentations on RDA results remained abstract. It would be of great importance to create Usage Stories that show what the various components do, how these components can be combined to concrete architectures, how they are being used in concrete examples. This may result in Primers that go beyond the 2-page descriptions. RDA is aware about this wish and has it high on the priorities list, however creating them given the limited personnel we have costs time which hopefully from September on is more easy due to the senior/junior team. Already now we can refer to the slides of the adoption day at the San Diego plenary[7] where a number of domain experts presented how they are applying the various results.

Offered Services
There is a clear lack of knowledge about available services. A workshop with practical hints how to use existing services could be very attractive. This can be extended to available technology that could be used across discipline borders. In this context the "Virtual Collection Builder" from CLARIN was mentioned for example. A flyer and web-site could emerge from a meeting. Again RDA EU needs to focus on the European level services, some special efforts would be relevant to indicate typical German services.

Library Series
A discussion was started about the possible roles of libraries and it was mentioned that librarians in general lack much of the knowledge to participate in discussing advanced data activities as being

[6] https://europe.rd-alliance.org/documents/publications-reports/data-harvest-how-sharing-research-data-can-yield-knowledge-jobs-and
[7] https://www.rd-alliance.org/plenary-meetings/fifth-plenary/programme/adoption-day.html

dealt with by RDA for example. A specific series of training courses was seen as very important for librarians. The liaisons at the research institutions (Fachreferenten) are seen as good interaction partners to discuss concrete measures. Also actions at European level via LIBER could be of interest. One of the participants will check opportunities with the liaison people at the universities.

Computer Scientists Involvement

A better inclusion of CS knowledge was mentioned as needed. At least two options are of relevance out: (1) The most direct way is that CS experts participate in RDA WG/IGs such as the Vienna group did it in the Dynamic Data Citation WG. (2) The other possibility is given by organizing special workshops bringing practitioners and CS together. The latter method however requires additional funds and time, and thus does not scale very well. There are some discussions going on where however special attention to inviting CS should be given to make use of their knowledge.

Specific Topics

Licenses

Licenses for data play often a very important role and it is not clear what the RDA group[8] on this topic is exactly doing and what their aims are. Some information should be spread. A CLARIN expert working in Mannheim (Pavel Kamocki) is highly active in that group and could be contacted.

Repository Setup and Trust

An interesting topic for many seems to be practical help in how to set up well-organized repositories that will be scalable and easy to federate. There is some experience around and some guidelines on data organisation principles could be given from RDA DFT[9] and on the choice of useful software could be given for example from CLARIN and others dependent on the requirements. There is also an RDA group now establishing requirements and evaluating solutions.

Furthermore certification of repositories or data collections is important to enable trust. Guidance how to build up trusted repositories as well as automated procedures, whenever possible, are required.

PIDs Usage

PID usage and services remain topics of high interest. Basic questions such as “what is the difference between Handles and DOIs[10], which functions PIDs should have and which kind of information should be associated with PIDs” should be explained and also advanced usages should be discussed to give an idea how professional services could be setup. Handles can be requested for example from EPIC (hosted by GWDG) while DOIs can be requested via DataCite. One issue that raised special attention is which information should be associated with PIDs. It is agreed that this information is metadata having special purpose such as the information associated with passports to prove identity and also information to check integrity. Yet there is no clear guidance about what to best do. Organizing workshops or sessions on basic and advanced aspects of PIDs could make much sense.

Metadata

Also metadata remains to be a wide field that still lacks agreed structuring to make it feasible and practicable. On the one hand standardization is required, on the other hand flexibility is necessary. It seems that only an approach based on re-usable components/packages anchored in semantic registries can overcome the diffuse discussions. Also flexible collection building was mentioned as a useful option in many areas. Some attempts in these directions have been made. A workshop would

[8] https://rd-alliance.org/groups/rdacodata-legal-interoperability-ig.html

[9] https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html

[10] Briefly it should be noted that technically DOIs are Handles with a prefix 10. Handles can have any other prefix and institutes could set up their own Handle server with an own prefix for performance reasons for example. Due to their additional requirements DOIs are good for published collections, while Handles in general are good for being used early in the production process to be able to refer to data.

help to discuss solutions and their capacity. Also the RDA metadata groups support now the idea of re-usable packages/components as it has been implemented for example by CLARIN. Also with respect to metadata a workshop could help clarifying.

Virtual Collection Building
Having a generic tool that allows building, registering and documenting virtual collections seems of interest to many. A workshop or session could explain such a tool in its practical use.

Workflows
Workflows were mentioned in some talks, however it was also obvious that building or using a WF environment is not trivial and too far going for many of the small departments. There is also the question in how far procedures in the labs are so normalized that working with WF environments make sense. The investments to turn to workflows cannot be neglected. This is one of the reasons that the idea of a library of pre-fabricated workflow components that can be re-used was seen as interesting to promote workflows. A workshop or session with practical hints of how to use WF technology with concrete examples seems to be useful.

Repository Registry
The issue of a repository registry with its different flavors as currently being discussed in the Data Fabric IG seems to be of interest. Currently three very closely related flavors are being discussed:

- a registry of trustful repositories mainly for human readability as it has been set up successfully by re3data for example
- a registry of "collections" that allows the huge number of small repositories to indicate the kind of usable data collections they store and offer – a concept that has been discussed also by DARIAH
- a registry that includes detailed machine readable data about services a repository offers as it has been designed by Grid projects (GOCDB) and as it is used by EUDAT

For certain reasons experts are discussing an approach where repositories offer such data according to defined schemas so that any trustful service provider can harvest the information and build useful end-user services.

Data Types Registry
For most of the participants the idea of a Data Type Registry and combining semantic types and functions was new. This paradigm seems to be of interest making a workshop or session with concrete usage scenarios attractive.

Privacy and Distributed Search
Colleagues from the medical field pointed out how difficult it is to release clinical and thus personal data so that it is available for research outside of the hospital. In this realm a method was mentioned where data does not leave an institution and where the operations can be controlled by tested software. The Human Brain Project and CLARIN worked for different reasons on a solution for distributed content search or data mining that is operating on the available data, extracting features and returning only those (summarizing and thus anonymized) features to the caller. It seems that also a workshop about such methods could be attractive.

Method

RDA D will continue to raise issues and obviously we need a separate space for interacting. All communication will be done including those that registered for this workshop and who participated in the last November meeting.

Data Fabric Use Cases and Comments on "Paris" Document

It is expected that in the coming two weeks the participants will start

- contributing with Use Cases to improve the abstraction process towards common components and
- making comments to the "Paris" document

Upload Use Cases here:

https://rd-alliance.org/group/data-fabric-ig/wiki/data-fabric-ig-use-cases.html

Comment on Paris document here:

https://rd-alliance.org/group/data-fabric-ig/wiki/data-fabric-ig-componentsservices.html

Paris Plenary

Please register for the Paris plenary which will take place from 23.-25. September. We expect interesting side meetings also at 21. and 22. September. At 22. September a so-called 5 stream meeting will be organized where relevant infrastructures will organize intensive discussions about similar questions as discussed at our workshop. The topics in focus are: Human Brain Project, Environmental Projects, Language Resources and Technology for Science, Data Infrastructure Services (EUDAT, EGI, etc.), Knowledge Infrastructure Services (OpenAIRE, etc.).