Project Acronym	RDA Europe
Project Title	Research Data Alliance - Europe 3
Project Full Title	RDA Europe 3 – the European plug-in to the global Research Data Alliance
Project Number	653194
Deliverable Title	Practitioner Engagement and Uptake Report 1
Deliverable No.	D4.1
Delivery Date	November 2015
Author	Raphael Ritz, Herman Stehouwer, Peter Wittenburg

ABSTRACT:

This WP4 Practitioner Engagement and Uptake Report describes the actions that are intended within WP4 in the coming 12 months to engage with the practitioners and to organize Uptake. It is the first report of this type and another report with an update will be written before the end of this first 12 month period.

RDA Europe receives funding from the European Union Horizon 2020 Programme (H2020/2014-2020) under grant agreement n° 653194.

DOCUMENT INFORMATION

PROJECT	
Project Acronym	RDA Europe
Project Title	Research Data Alliance – Europe 3
Project Full Title	RDA Europe 3 – the European plug-in to the global Research Data Alliance
Project Start	1st September 2015
Project Duration	30 months
Funding	H2020-EINFRA-2014-2
Grant Agreement No.	653194
DOCUMENT	
Deliverable No.	D4.1
Deliverable Title	Practitioner Engagement and Uptake Report
Contractual Delivery Date	November 2015
Actual Delivery Date	01 December 2015
Author(s)	Raphael Ritz, Herman Stehouwer, Peter Wittenburg (MPG)
Editor(s)	Raphael Ritz (MPG)
Reviewer(s)	Hilary Hanahoe (TRUST-IT), Sandra Collins (NLI)
Contributor(s)	all WP4 participants
Work Package No. & Title	WP 4: Practitioner Level Engagement and Uptake
Work Package Leader	Raphael Ritz (MPG)
Work Package Participants	MPG, CSC, TRUST, CNR, Athena, STFC, ACU, CNRS, RIA, NLI
Estimated Person Months	3
Distribution	OPEN
Nature	Report
Version / Revision	1.0
Draft / Final	Final
Total No. Pages (including cover)	23
Keywords	Quality, Management

DISCLAIMER

RDA Europe (653194) is a Research Infrastructures Coordination and Support Action (CSA) co-funded by the European Commission under the Research and Innovation Framework Programme, Framework Programme Horizon 2020 (H2020).

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

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

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

GLOSSARY

ABBREVIATION	DEFINITION
BoD	Board of Directors
CB	Consortium Board
DoA	Description of the Action
DFT	Data Foundation and Terminology Working Group
DFIG	Data Fabric Interest Group
DO	Digital Object
DOI	Digital Object Identifier
EC	European Commission
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
f2f	Face to face [meetings]
IG	Interest Group in the global RDA initiative
KB	Knowledge Base
PID	Persistent and unique Identifiers
PM	Person Months
PMO	Project Management Office
RDA	global initiative "Research Data Alliance"
RDA EU 3	Research Data Alliance - Europe 3
SJT	Senior Junior Team

SyA	Synchronisation Assembly of RDA Europe
TAB	Technical Advisory Board of RDA
TCP/IP	Transmission Control Protocol/Internet Protocol
WG	Working Group in the global RDA initiative
WP	Work Packages in the RDA Europe project
WPL	Work Package Leader

TABLE OF CONTENTS

Executive Summary

Engaging with European Practitioners is a fundamental activity in the project in order to present the RDA opportunities as well as accompany them in adopting and implementing solutions made available by RDA that will help them solve data challenges. On the other hand, interacting with practitioners and understanding how they can contribute to RDA and work on a global level is imperative. The activities planned in the **Practitioner Level Engagement and Uptake** work package (WP4) aim to do this and have the following specific objectives:

1. Set up and run an expertise centre that understands all RDA WG/IG activities, knows the results and their embedding in the standards landscape and can give advice to experts within RDA and RDA Europe and to interested people from outside.
2. Engage with the practitioners (scientists, data scientists, data managers/stewards, archivists, librarians) of European research infrastructures and e-Infrastructures (ESFRI + beyond) to organise bilateral meetings and collaboration projects to develop needs analysis and result uptake.
3. Organise concrete topic seminars, training workshops, summer schools, and hackathons/datathons using the results of RDA to engage in particular early career people.
4. Execute collaboration projects with European and developing countries that have been initiated within WP2 and WP3 at policy level. These collaborations are necessary to include new regions in RDA; however they need to be in balance with the effort for European practitioners.
5. Create a platform for intense and deep interactions about RDA activities including advanced technologists such as computer scientists, experts from relevant RDA working groups and interested industry to improve RDA and RDA Europe work. Interact closely with the Technical Advisory Board (TAB) of RDA, in particular the European members of TAB.

These objectives will be achieved through 4 distinct tasks, which all start at M1 and run throughout the lifetime of the project:

- Task 4.1 Activity Analysis and Clearing House
- Task 4.2 Practitioner Engagement
- Task 4.3 Advanced Technology Interaction[1]
- Task 4.4 Training Program

In just the first three months, the following goals have been achieved:

- The Senior Junior Team (SJT) has been established and first internal trainings have been completed.
- The Knowledge Base (KB) has been specified and is currently being filled.
- The first Working Group (WG) descriptions have been reviewed with respect to various aspects (groups, outputs, services, etc.) and inserted into the KB.
- The training plan for the first 6 months has been specified and the first courses will begin in December.
- The technology for webinars has been tested and is ready to go.

[1] Task 4.3 is covered by Deliverable D4.5.

So within the first months much has already been done and the plans for the coming months have been defined. Where necessary WP4 will react with agility depending on the feedback we will receive. Most of the core issues are new for the team (webinars, KB, support) so we will reflect on the utility and success regularly. A next face-to-face meeting of SJT is being scheduled for January where we will elaborate on the work so far and adapt where necessary.

1. Practitioner Engagement

Achieving the objectives, as outlined in the Executive Summary, set will be focal to all activities in this work package and the following sections describe in detail the plans, expected outputs and associated KPIs.

Furthermore, to achieve maximal transparency the plans will be presented to the RDA Europe Stakeholder Meetings and the Synchronisation Assembly (SyA) for comment and the Board of Directors (BoD) for final decision taking.

1.1 Senior/Junior Team

A group of seniors and juniors (SJT = Senior/Junior Team) has been constituted for this WP. It is intended to include external experts in the work in particular in tasks 4.2 and 4.4 since RDA Europe cannot cover all disciplines and all topics addressed by RDA WGs and IGs. Two external experts (Wolfram Horstmann from LIBER and Odile Hologne from INRA) have been added by subcontracting to the WP4 activities and here in particular to the SJT. Therefore the SJT includes the following persons:

Senior/Junior Team:

Leif Laaksonen, Anni Jacobsson (CSC); Herman Stehouwer, Kathrin Beck, Peter Wittenburg (MPG); Françoise Genova (CNRS); Sandra Collins (NLI); Natalie Harrower, Rebecca Grant (RIA); Dimitris Gavrilis, Panagiota Koltsida (Athena RC); Simon Lambert (STFC); Odile Hologne (INRA); Wolfram Horstmann (LIBER)

Another subcontract has been finished with Alan Blatecky who will help in particular in synchronizing testing and uptake between US and Europe.

In addition to what is described in the DoA the RDA Europe project decided to adapt the budget in so far that a number of short collaboration projects can be funded. The process to select projects, the guidance of the projects, the analysis of the results, and the aggregation of the essential skillsets will all be performed by SJT under task 4.1 (WP4).

1.2 Task 4.1/4.2 Activity Analysis and Clearing House / Practitioner Engagement

Task 4.1 will carry out analysis in specific data science domains combining all activities in the form of reports, hand them over to the interested RDA groups, in particular the TAB members and the WG/IG chairs and enter the essentials to an open knowledge base. These experts must also be involved in the editing of the final results to guarantee coherence in form and content and to ensure quality. It is important to understand what the discipline specific groups within and beyond RDA are discussing and to bridge the gap with the cross-disciplinary WG/IGs active in RDA. Given this variety we need young experts with high availability, engagement and flexibility in WP4 that carry out this work to have a deep coverage. Finally Task 4.1 participates in creating suitable documentation about the results from the WGs and in translating the key messages to different stakeholder groups (policy makers, industry, science, etc.).

Task 4.1 procedures cover the:

- Analysis of the work in all WGs and IGs and compare goals, notes, activities and results

- Analysis of the related work in related standardisation initiatives to look for synergies and collaborations
- Analysis of the data practices and identifying inefficiencies in scientific disciplines by regularly visiting domain meetings etc.
- Creation of expert memos to document knowledge improvements, give advice, answer questions and build up a common and open Knowledge Base
- Discussion of priorities, progress and quality of the work during monthly meetings with all task participants. Close collaboration with members of TAB, Group Chairs and Secretariat is required.

Task 4.2 focuses on interactions with data scientists, data managers, archivists etc. from all kinds of research disciplines as well as from industry through a variety of forms with the clear intention to deepen mutual understanding and to organise uptake. This will include:

- Organising joint meetings with practitioners of ESFRI project initiatives, with e-Infrastructures, archivists and librarian organisations, industry groups, etc. who are faced with sharing and interoperability hurdles.
- Participating in discipline/domain meetings.
- Organising bilateral meetings with data practitioners.

Discussing and implementing collaborations with the intention to put RDA results into place, to implement APIs to RDA registries, to adapt existing solutions to RDA results (APIs, protocols, etc.), to carry out tests with test implementations, support of national activities, etc.

Task 4.2 procedures cover the:

- Collection of all kinds of information in particular that from WP2 and that from other tasks in this WP to create and maintain this Engagement and Uptake Plan.
- Creation of short notes from all meetings to identify the state of interactions and possible uptake opportunities.
- Creation of a shortlist of concrete collaboration projects to guide the concrete uptake work.

Due to the close relationship between these two tasks we will do the planning and reporting in one subchapter.

1.2.1 Activity Report

In the first months the following activities were carried out:

- Planning the nature and structure of the Knowledge Base (KB)
- Deciding about the start-up technology for the KB
- Planning descriptors that could be used by people to easily find information in the KB
- Planning access rights on the KB
- Implement a first simple version of the KB
- Planning and inventing templates for different document types
- Start describing the first WG outputs by using the templates
- Launch & management of the Call for Collaboration projects

1.2.1.1 Knowledge Base Setup

The KB has been set up based on the wiki Confluence system running at CSC. First we will run a test KB being implemented in the SJT pages and when the structure has been settled we will start a new special KB site within the Confluence system. During the second phase the access permissions would be set as decided:

- Only SJT members should be able to write.
- All information is open worldwide.
- User questions and comments will be used to start moderated discussion threads to maintain a coherent knowledge base.

It has been decided to first use a simple wiki like structure and from practice and usage patterns evolve to a final, potentially more complex setup.

To start with, the structure will consist of a few major components that should help navigating:

- Description of RDA Recommendations & the groups that produce them, i.e. the Working Groups (WG) are the starting point
- Examples of adoption and the experiences of the adopters
- Services that are produced as a side effect of Working Groups (e.g. www.typeregistry.org or the Research Data Switchboard[2])
- Other Documents

These are presented in such a way as to indicate what problems can be solved and what an adoption of the result would roughly require in terms of local adaptions and resources. Links between the various pages will need to support easy navigation between these related pages.

The structure will be amended by the possibility to add keywords enabling easy discovery. As keywords we have chosen those that have been worked out by RDA TAB to describe all the various RDA groups. The vocabulary can be extended if necessary. At this moment this feature is not used in the current test phase.

The first phase KB is running and currently being filled with first information by most of the SJT members.

1.2.1.2 Document Types and Templates

We have already now identified a number of different document types that can be found in the KB, others may follow. The document types are:

- RDA Output descriptions
- RDA Output flyers
- Adoption examples
- Service description
- Other documents of relevance

For a number of these document types first templates have been developed and added to the KB test document. The intention is to have well-structured and thus comparable documents in the KB that make it easy for users to find the information they are looking for. The Data

[2] http://www.rd-switchboard.org/

Foundation & Terminology (DFT) example is added in Appendix A for the DFT group[3] following one of the templates. The first page gives an introduction to the DFT group, its goals, a graphic representation (not shown in this document), and a matrix which includes internal links to specific aspect of the group's results and external links to other related work. This semantic weaving could help users to see the relationships for example with the FAIR principles[4] etc.

1.2.1.3 RDA WG Outputs

The first types of documents to be entered into the KB are descriptions of various sorts about those RDA Working Groups that have undergone the full RDA review process (first 4) and those WGs that produced their final documents but did not yet finish the community review process. In detail the following documents are in preparation:

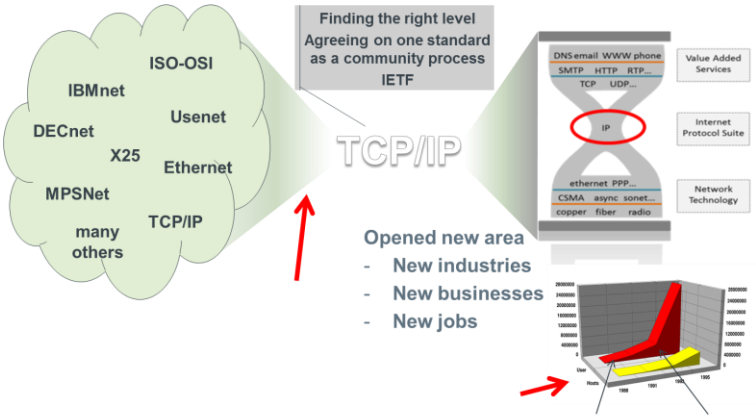
- Data Foundations & Terminology — Peter Wittenburg (MPG)
- Data Type Registry — Dimitris Gavrilis (Athena RC)
- PID Information Types — Herman Stehouwer (MPG)
- Practical Policy — Kathrin Beck / Peter Wittenburg (MPG)
- Dynamic Data Citation — Herman Stehouwer (MPG)
- Wheat Interoperability — Odile Hologne (INRA)
- DSA/WDS Certification — Françoise Genova (CNRS)
- RDA/WDS Data Publishing groups — Riina Salmivalli (CSC)
- Metadata Directory — Panagiota Koltsida (Athena RC)

1.2.1.4 Components & Recommendations

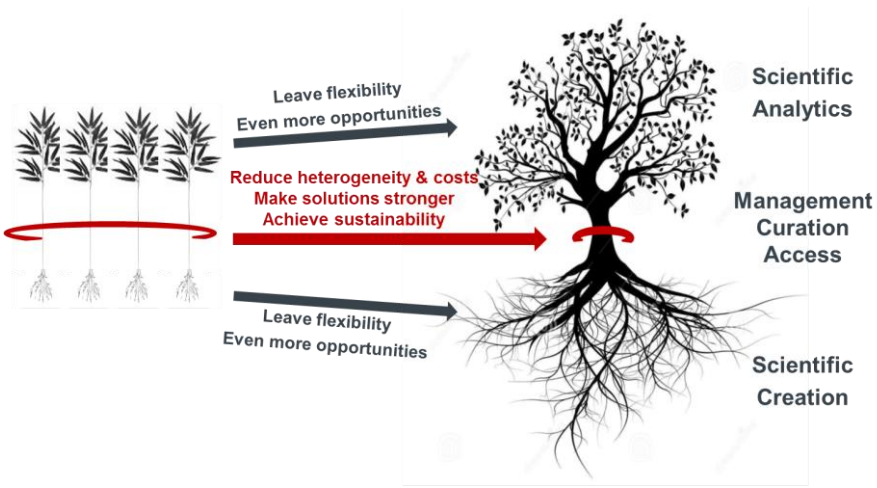
In the Data Fabric group [5], between the RDA WG chairs and in the cross-Atlantic synchronisation interactions (WP3) a discussion has taken place about the question of how to speed up the virtuous cycle from RDA outputs to global recommendations which finally will increase interoperability and reduce the costs of global infrastructure building. The so-called 'Paris document'[6] has been published and broadly discussed including several ESFRI projects. In the Data Fabric Interest Group (DFIG) sessions at the RDA plenary P6 in Paris it was agreed that the DFIG members and beyond will start a global discussion about recommendations. To make the global discussion process feasible it was agreed to start regional groups first and use the DFIG wiki[7] to aggregate and discuss these.

In Europe we started a discussion process by talking to various experts from ESFRI and other projects to form a core expert group for discussing recommendations. Also a first document has been written and discussed that describes a possible structure for different types of recommendations.

[3] https://rd-alliance.org/groups/data-foundations-and-terminology-ig.html and https://rd-alliance.org/groups/data-foundations-and-terminology-wg.html
[4] https://www.force11.org/group/fairgroup/fairprinciples
[5] https://rd-alliance.org/group/data-fabric-ig.html
[6] http://hdl.handle.net/11304/f638f422-f619-11e4-ac7e-860aa0063d1f
[7] https://rd-alliance.org/node/44520/all-wiki-index-by-group

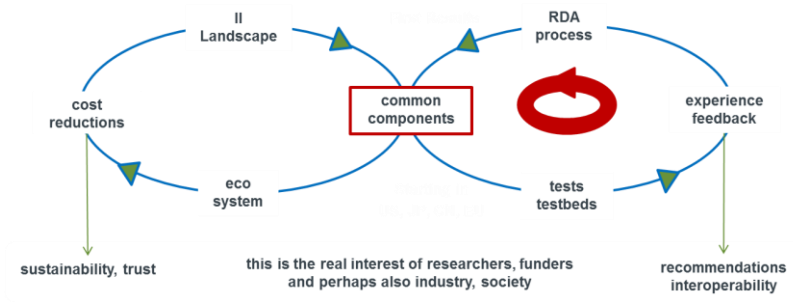


The following figures give an impression of the rationale behind the virtuous circle that RDA wants to accelerate and where RDA Europe (and here SJT) will play an important regional role in Europe. We can refer back to the success story of TCP/IP (Transmission Control Protocol / Internet Protocol) when we want to explain the rationale. During the early days of connecting computers there was also a large variety of suggestions and it was not clear how one could overcome the heterogeneity to create a momentum. Finally the bottom-up initiative from a number of experts led to the worldwide agreement that TCP/IP is a proper and tested solution. It is well-known that this global agreement had an enormous impact on innovation. As the diagram shows it took about 15 to 20 years between definition of TCP/IP and its wide acceptance. TCP/IP could be called a common component of the Internet.



The second figure indicates where experts expect common components to be needed. While there will always be heterogeneity in the methods that are used to create and analyse data, an analysis of a large variety of infrastructures showed that the components that are used to manage, curate and access data are fairly common despite all differences in detail. Common agreements need to be defined in this area and RDA groups are working in this direction.

The third figure indicates the virtuous circle the acceleration of which is now highest priority. RDA groups come up with results for common components, but a long and intensive period of testing is required to give feedback to the groups, improve the results and finally to create trust in them. This can be then the start of wide acceptance and thus increased interoperability. Also funders are interested in such common components since these will reduce complexity of solutions and costs of infrastructure building and maintenance.



1.2.1.5 Testing & Collaboration

RDA Europe has re-assigned some budget to allow the funding of short collaboration projects which are meant to do RDA result testing and therefore accelerate the virtuous circle. A call has been launched with a deadline of 30 October. In total 25 project proposals were submitted and are now in the evaluation & negotiation phase (Nov / Dec 2015) with a target to start the successful ones from January 2016 on.

Task 4.1 & Task 4.2 Outputs & KPIs:

- Planning documents and content memos about relevant discussion subjects need to be provided to the project and the RDA boards (such as TAB).
- After 6 months all early career experts participating in this task will understand the activities within the RDA WGs and IGs and have offered a first training seminar for all Europeans involved in RDA in some way to bring everyone on the same state.
- A first complete analysis will place all activities into a coherent data landscape and that can be used to give advice about further activities.
- Create and maintain an Engagement and Uptake Plan and each year RDA Europe will prove engagement with at least 30 different calls for help, advice and integration work from a variety of disciplines.
- Create short notes to achieve transparency about the interactions and to document the state of collaborations, test results (implementations) and documentations describing them.
- Add knowledge of various types extracted from the concrete work into the emerging Knowledge Base.
- The group of experts in this task must have shown their capabilities in almost all training courses being carried out by task 4.4.

1.2.2 Activity Planning

In the following we will indicate the planned activities for the coming 12 months.

1.2.2.1 Knowledge Base Setup

- The Knowledge Base (KB) will be moved to a separate space and opened for public consumption and commenting.
- An option for specifying and searching for keywords will be added.
- An option for commenting and raising questions will be added.
- The structure of KB will be subject of continuous optimization.

The goal is to establish the KB as a valuable information resource for data practitioners world-wide.

1.2.2.2 Document Types and Templates

- Depending on the upcoming needs new document templates will be defined.
- Existing document types will be optimized dependent on the experiences.

It is the intention to extend and optimize the content and include or point to a variety of external resources in the context of the RDA activities.

1.2.2.3 RDA WG Outputs

For each of the finished and finishing RDA WGs a complete and comprehensive document set must be made available.

1.2.2.4 Advice / Clearing House / Support

- More comparisons of different solutions will be carried out and the results will be uploaded in the KB.
- Internal training will be intensified to bring the SJT to a high level of knowledge.
- Areas of special expertise will be defined so that specialized knowledge will be available in SJT.
- A help desk will be announced and offered asap so that all ESFRI projects and in fact every interested person in Europe can make use of this knowledge.
- A ticketing system will be put in place to monitor responses to requests.
- Joint meetings with practitioners from ESFRI and other initiatives will be organised depending on interests.
- Requests from industry are important and eventual enterprise focused events will be organized together with WP2 and WP5.
- Where necessary SJT will also give support by sending members to sites for limited periods.

Given the broad spectrum of RDA activities which are a reflection of the complexity of the issues in the data area RDA Europe members cannot claim to have detailed expertise in all areas. Therefore we expected the need to include external experts in a number of activities such as workshops, training courses, support actions, etc. Therefore SJT will always check where the involvement of external experts is required and make suggestions to the WP4 leader who will send update plans to the BoD for decision taking on subcontracting.

1.2.2.5 Components & Recommendations

- During the plenary P7 in Tokyo in March 2016 the state of work will be presented and discussed in the Data Fabric IG.
- A first RDA Europe document with recommendation should be made ready by Easter 2016 through a discussion process between the European core experts – we plan many virtual group interactions to settle towards recommendations.
- The discussion process will be continued at global level also requiring many virtual interactions.
- At the plenary P8 in Washington in September 2016 a first draft global recommendation paper will be presented and discussed in the Data Fabric IG.

1.2.2.6 Testing & Collaboration

- All 25 proposals for Collaboration Projects will be analysed timely by SJT in a timely manner including interactions with the proposers where needed.
- External experts will be included to give advice on proposals.
- The summary of the analysis will be discussed in BoD and a ranked shortlist will be established.

- The selected projects should be started and for each project team a contact person in SJT will be appointed.
- Further actions (reporting, etc.) will be done dependent on the projects.

1.3 Task 4.4: Training Program

RDA Europe interaction with data practitioners from all kinds of research disciplines and also industry must be intensified using a variety of forms where organising training courses of different form is an important part. This includes:

- Training courses to present the results of RDA, which can be cross-disciplinary or within disciplines/domains and which can be stand-alone events or co-events to meetings
- Webinars that are easy to plan but also more limited with respect to their effect
- Organising summer and/or winter schools at least once per year to contact early career people and to foster community building
- Organise hackathons/datathons twice per year to prepare the plenaries and to do concrete code development and testing around RDA activities

In particular the last two action points require skilled people and considerable preparation time. This task itself needs a professional person to plan and organise all events, but it also needs to make use of the experts within RDA Europe as well as external experts from working groups or even experts from outside of Europe to achieve the level of quality required.

Participation at all events is mainly targeted to European early career people. However it should be possible to include a few non-European experts.

1.3.1 Activity Report

1.3.1.1 Internal Training

Already in September the SJT did its second internal Training Course in Paris in which SJT members participated. Furthermore, internal virtual training on webinar infrastructure has been arranged, i.e. how to use Adobe Connect Pro on giving webinars.

1.3.1.2 Training Concept (webinars/f2f meetings, tool, format, external experts, announcement)

The training concept consists of three major parts:

1. Webinars and webcasts
2. Workshops and other organized face to face (f2f) training, hackathons, etc.
3. Consultancy type training

For the first two types we aim to cover all RDA recommendations and related basic knowledge (e.g. metadata systems, PIDs in general). The third type is organized on an as-needed basis. Depending on the level of detail and the time commitment of the training we try to use the experts themselves to provide the training modules.

1.3.1.3 Training Plan

Month	Type	Date & time of training	Subject covered	RDA EU Interface	Speaker
Nov. 2015	Webinar	1 Dec. 2015 15:00	PID Information Types (PIT)	Herman Stehouwer	Tobias Weigel
Dec. 2015	Webinar	15 Dec. 2015	Data Type Registries (DTR)	Herman Stehouwer	Larry Lannom
Jan. 2016	Webinar	13 Jan. 2016 14:00	Practical Policy (PP)	Kathrin Beck	Reagan Moore
Feb. 2016	Webinar	16 Feb. 2016	Metadata Standards	Herman Stehouwer	tbd
Feb. 2016	Webinar	23 Feb. 2016	Data Citation	Kathrin Beck	tbd
March 2016	Webinar	15 March 2016	Data Organisation	Herman Stehouwer	Peter Wittenburg
April 2016	F2f training	tbd	tbd	Herman Stehouwer	tbd
April 2016	Webinar	10 April 2016	Data Description Registry Interoperability (DDRI)	Herman Stehouwer	tbd
May 2016	Webinar	17 May 2016	Wheat Data Interoperability	Herman Stehouwer	tbd
June 2016	Webinar	tbd	Repository Audit and Certification (Data Seal of Approval / Word Data System; DSA/WDS)	Kathrin Beck	tbd

It is widely known that webinars do not help in building a community; therefore it is intended to offer a physical training course approximately all 4 months. The details of the first one in April remain to be worked out.

1.3.1.4 JEEE Interaction

Discussions with the European projects OpenAIRE[8], EUDAT[9], EGI[10], EDISON[11], and LEARN[12] have been started to synchronize all activities wrt. training. An agreement has been achieved. Concrete actions need to follow after the RDA plans have been settled.

1.3.2 Activity Planning

1.3.2.1 Internal Training

Internal training courses need to be carried out regularly to achieve a high level of knowledge amongst all SJT members. It is agreed to spend time at the regular SJT meetings on thematic topics.

1.3.2.2 Training Concept (webinars/f2f meetings, tool, format, external experts, announcement)

We cannot make comments on the concept yet, since we need to wait on the feedback from the first events. At the end of January, after the 3rd webinar, we will evaluate the experience and adapt our strategies if necessary.

1.3.2.3 Training Plan

The table above in section 1.3.1.3 contains already a plan for the coming six months. Before making further plans we want to wait for the feedback for the webinars we are doing. Based on these experiences we will start making plans for the next six months. Also we need to discuss which experts we want to include for attracting participants.

1.3.2.4 Hackathons/Datathons/Summer schools

For the second half of 2016 a larger event must be prepared. It needs to be discussed what character the meeting should have. By the end of January 2016 a concrete plan will be worked out.

1.3.2.5 JEEE Interaction

- The concrete training agendas need to be exchanged and synchronized until end of December.
- Adaptations and expert exchange need to be discussed until end of January.
- Joint announcements need to be planned.

[8] https://www.openaire.eu/
[9] http://eudat.eu/
[10] http://www.egi.eu/
[11] http://cordis.europa.eu/project/rcn/198292_en.html
[12] http://cordis.europa.eu/project/rcn/194936_en.html

1.3.2.6 Training Targets

As indicated the project wants to have at least one training session event per month and f2f training events regularly. By including external experts on topics where this makes sense we will probably be able to extend the training schedule to a higher frequency addressing various stakeholders. Our experience with RDA related training courses in particular in Germany and France indicates that at this moment it is too early to tailor training content to very specific aspects such as explaining in detail how the programming interface of the RDA PIT WG needs to be integrated into own software stacks. Currently many of the training sessions are focussing on informing people about basics such as the existence and use of Persistent Identifier systems. We urgently need to address the existing information gaps that many of the people dealing with data have.

Currently we are addressing in particular those data practitioners that are active in ESFRI research infrastructures, e-Infrastructures and other large infrastructure building projects such as Human Brain Project. It is obvious from the discussions in WP2 that we need to extend our activities by addressing even more data practitioners from research institutes and projects. This most often until now is regulated by national activities. Here WP4 is dependent on impulses again from WP2.

After having evaluated our first experiences end of January we need to see how the training program needs to evolve to

- Address the data scientists, i.e. those people that want to go into details of interfaces and tools
- Address the data librarians who are more interested in generic metadata and curation questions
- Address the science policy makers taking into account that these people in general do not like a training format, but more of a workshop which is the task of WP2

It is also too early to speak about concrete content and formats for industry directed training offers. WP4 is ready to make offers but we will dependent on interactions in T2.2 with industry people to identify what kind of advice and information is urgently needed. Until now discussions with industry did not result in concrete ideas that can be turned into a plan.

Appendix A

Data Foundation and Terminology WG

The Data Foundation and Terminology Working Group (DFT) started as one of the first in RDA, had a first interactive preparation session at plenary 1 in Gothenburg in March 2013 and then formulated a Case Statement to start the real work after plenary 2 in Washington in September 2013. It was chaired by Peter Wittenburg, Gary Berg-Cross and Raphael Ritz and had about 65 members from very different backgrounds and disciplines. Since 2015, the group continues as an Interest Group to maintain and extend terminology and is chaired by Gary Berg-Cross and Raphael Ritz.

DFT Goals

- **harmonize data organizations to increase interoperability and reduce data integration costs**
- **harmonize terminology to reduce time to understand each other on basic data modelling aspects**

The work was mainly based on different data models contributed from various disciplines and use cases. The group produced six core documents[13] covering all aspects. Most important is the document 'DFT Core Terms and Model' which summarizes the core terms being defined and the model.

For citations we suggest using this Digital Object Identifier (DOI): http://dx.doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF [14]

References within Page	References to other Documents	References to other Documents
Problem Statement	Flyer about DFT	
Solution	Relation to other WGs	
Impact	Relation to FAIR Principles	
How to Adopt?	Kahn & Wilenski Paper	
Implementation Examples	Fedora Repository System	

Problem Statement

Open data sharing and re-use is one of the key requirements for the coming decade, but yet working with data is highly inefficient and thus costly and does not scale:

[13] https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html
[14] https://b2share.eudat.eu/record/247

- For industry a lack of 190,000 data professionals is estimated for the coming years to make data re-usable for business.
- Scientific practice shows that a data scientist spends about 75 to 80% of his/her time finding, curating and integrating data to make it usable for his/her research.
- We are still creating much data without proper registration and documentation which will cost a lot of curation effort in the coming years.

This all sounds like a paradox since we are used to re-purpose and integrate data coming from different directories paths and being stored in clouds identified by has tags. Where is the problem that needs to be solved? All research domains are increasingly faced with irreversible trends (volume, velocity, variety, complexity) that obviously require new ways to deal with data. Experimentalists who create large volumes of observational data often still can work with file systems since the "underlying logic of data" does not change. A new way to cope with the huge amounts of files efficiently is offered by clouds.

However the second grand problem is created by variety and complexity. We are creating so many different new data types and utterly complex relation structures between data that only the creation of metadata (context, provenance, etc.) and stable references through persistent and unique identifiers (PID) can help overcoming chaos. It is well-known that traditional ways such as using file/directory names, spreadsheets, etc. cannot help to prevent the dramatic loss of information about data as Bill Michener from DataONE[15] described it. It is often completely overlooked that cloud storage solutions do not address these aspects, since they are layers on top of the physical storage layer.

Many initiatives and communities started already defining metadata systems and using PID services. However, there is no harmonization of the way data yet, metadata (of different type) and PIDs are related to each other. Everyone does it differently and often relevant information is hidden even in programming code so that when integrating such data with the intention to preserve all relations one needs to do special programming efforts. In the EUDAT project it cost about three weeks of programmer's time to replicate an excellent database including all relevant metadata. However, there are thousands of such data bases out there and new ones are created daily. How to get them not only into a shared domain of data, but to make them practically re-usable?

We urgently need steps to harmonize data organisations so that integrating data from various sources including the relevant metadata and by making use of PIDs can be done efficiently and cost-effective.

Solution

Based on analysing a large number of practically used data models and use cases from different research communities the Data Foundation & Terminology Working Group (DFT) in RDA tackled two challenges:

- finding a common language what data practitioners are talking about when speaking about basic data organisations (terminology)

[15] https://www.dataone.org/

- grounding this language on a core model that mimics current practices and can serve to harmonize data organizations

Intensive DFT discussions and tests in some communities led to the model indicated in the diagram below which on purpose was kept simple to make it understandable by all data practitioners. The core of the model is indicated by the elements marked brown. Basically the model states that Digital Objects (DO) need to be described by metadata, that references by a persistent identifier and that its bit streams need to be stored in a digital repository. It also states that Digital Objects can be aggregated to Digital Collections. It is important to note that metadata descriptions and collections are also Digital Objects which have enormous implications. Digital Objects can thus be anything that can be referred to by a PID, i.e. it could be a file, a query that is stored, a software module, a configuration of a telescope, etc.

If this model will find wide acceptance and if software builders implement it, we can achieve a lot of improvements:

- It would allow us to rely on PIDs for all kinds of references enabling also identity and integrity checks at any moment.
- It would allow us to design a simple programming interface for repositories in order to store digital objects of different types and also for easy retrieval of the corresponding metadata and PID records.[16]
- Machines could be used to integrate data easily since they would systematically find the appropriate information for location, interpretation, etc.

Impact

By using the generated terminology we would understand each other much more easily across backgrounds and disciplines and be more efficient. By implementing this generic model in software that is used for data creation and management we would widely harmonize the way we organize data making data sharing and re-use much more interoperable and cost-efficient. This model is in full alignment with the data principles as they have been established by G8/FAIR and others.

How to Adopt?

Adoption of DFT results is straightforward:

- Use the DFT specification document, use its terminology and implement the basic data organization principles. ***Most important is here the document titled "DFT Core Terms and Model"[17] which summarizes the major results.***
- Use the Term Definition Tool (TeD-T)[18] that has been created by RDA to add comments to the current model and terminology which will lead to updates.[19]
- Contact the DFT IG co-chairs in case of questions and comments.

[16] In the general sense PIDs and attributes that are stored in the PID records are specific metadata, however they serve specific purposes.
[17] https://rd-alliance.org/sites/default/files/DFT%20Core%20Terms-and%20model-v1-6.pdf
[18] http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page
[19] The DFT Interest Group takes care of maintenance and updates.

The DFT group is aware of the fact that term definitions will change over time and that ongoing testing and adoption of the model will lead to new insights that may need to be reflected in the definitions.

Implementation Examples

As can be seen from the data models and use cases that were used as basis for the DFT results quite a number of initiatives worked in the direction of the DFT results, i.e. the DFT work led also to adaptations of original thoughts. Here we only want to refer to some cases:

- The World Data Climate Center[20] and its community adopted the model.[21]
- The CLARIN[22] community was part of the discussions from the beginning and follows the model.[23]
- The CLARIN community developed a collection builder tool which fully implements the model.[24]
- The EUDAT data infrastructure is strictly following this model by speaking about a domain of registered and documented Digital Objects (DO), i.e. each DO in the EUDAT sphere has a PID and a metadata description.[25]
- DFT Terminology and Data Model has been adopted for the description and structuring of atmospheric data.[26]
- The model is being used in experimental and collaborative ethnography.[27]

Many more applications can be found in the descriptions of underlying models and use cases being offered to the DFT discussions.

[20] https://www.dkrz.de/daten/wdcc
[21] https://rd-alliance.org/sites/default/files/attachment/RDA-2015-DKRZ-2.pdf#overlay-context=plenary-meetings/fifth-plenary/programme/adoption-day.html and https://rd-alliance.org/sites/default/files/attachment/6_GermanClimateComputingCenter_TobiasWeigel.pdf
[22] http://clarin.eu/
[23] https://rd-alliance.org/sites/default/files/attachment/dvu-clarin-dft-adoptation.pdf and https://rd-alliance.org/sites/default/files/attachment/07-CLARIN_ErhardHinrichs.pdf
[24] http://clarin.eu/content/virtual-collections
[25] https://rd-alliance.org/sites/default/files/attachment/RDA%205th%20Plenary%20Practical%20Policies%20in%20EUDAT%20CDI.pdf and https://rd-alliance.org/sites/default/files/attachment/5_EUDATCollaborativeDataInfrastructure_DaanBroeder.pdf
[26] https://rd-alliance.org/sites/default/files/attachment/RDA_Adoption_Day_Slides_Addison.pdf and https://rd-alliance.org/sites/default/files/attachment/08-DataFed_CynthiaHudsonVital.pdf
[27] https://rd-alliance.org/sites/default/files/attachment/4_PlatformExperimentalCollaborativeEthnography_LuisFelipeRosadoMurillo.pdf