Project Acronym	RDA Europe
Project Title	Research Data Alliance Europe
Project Number	312424
Deliverable Title	First year report on Roadmap to achieve full interoperability of EU-US data infrastructures
Deliverable No.	D2.3
Delivery Date	1. November 2013
Author	Peter Wittenburg

ABSTRACT (a brief 2 line summarizing statement of the document content):

This deliverable describes the work undertaken during the first year of the project towards a global and interoperable data infrastructure.

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

DOCUMENT INFORMATION

PROJECT	
Project Acronym	RDA Europe
Project Title	Research Data Alliance Europe
Project Start	1st September 2012
Project Duration	24 months
Funding	FP7-INFRASTRUCTURES-2012-1
Grant Agreement No.	312424
DOCUMENT	
Deliverable No.	D2.1
Deliverable Title	First year report on Roadmap to achieve full interoperability of EU-US data infrastructures
Contractual Delivery Date	September 2013
Actual Delivery Date	October 2013
Author(s)	Peter Wittenburg, Herman Stehouwer (MPI-NL), Rob Baxter (EPCC), Fotis Karayannis (Athena), Donatella Castelli, Costantino Thanos (CNR), Francoise Genova (CNRS), Leif Laaksonen (CSC)
Editor(s)	Peter Wittenburg, MPI-NL
Reviewer(s)	Donatella Catelli, Fotis Karayannis
Contributor(s)	EPCC, CNRS, MPI, CSC, Athena, CNR, STFC, ACU
Work Package No. & Title	WP 2
Work Package Leader	Peter Wittenburg
Work Package Participants	CSC, CINECA, EPCC, MPG, CNRS, STFC, UM, ACU, Athena, CNR
Estimated Person Months	42
Distribution	Public
Nature	Report
Version / Revision	V2.0
Draft / Final	Draft
Total No. Pages (including cover)	26

Keywords	Research Data Alliance, Building Blocks

DISCLAIMER

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

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

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

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
RDA Europe	Research Data Alliance Europe

TABLE OF CONTENTS

Executive Summary

In this report we examine the state of current data infrastructures and repositories and conclude that the European landscape is very heterogeneous and that there is no reason to assume that this will change soon. There are stakeholders at European level, at national level and at research organisation level, some organised along scientific disciplines, some cross-disciplinary. There are an increasing number of commercial services for which the question of trust will be very important for researchers. Improving interoperability and data sharing requires that we find ways of harmonisation which have the potential to overcome the barriers introduced by this diversity.

We introduce the notion of the RDA Agora (after the Agora, the main market and meeting place of ancient Athens), as the total of the Plenary interactions, the WG/IG interactions, the online interaction forum and the various contributions in RDA/Europe (WP2 Analysis, WP3 collaborations, RDA/E Forum, Science Workshops). It is this AGORA which has the potential to overcome the many barriers by defining and maintaining a roadmap of building blocks as basis of an overall architecture to be specified and implemented. Many initiatives and actors will contribute to this Agora, and some have already been addressed in RDA/Europe and RDA/global.

We review the first lists of building blocks that have emerged from these Agora discussions and have already been taken up in form of working and interest groups at RDA: it is an impressive start. Nevertheless, RDA needs to work on identifying concrete outcomes and tackle a number of gaps that have already been identified. This rapid progress convinced RDA/Europe that the Agora approach is the right one to take to meet our goals of improved sharing and interoperability.

Our conclusions on the state of the current RDA roadmap lead us to five recommendations:

- There needs to be an intensive discussion at political level in Europe to cross-fertilize with relevant scientific organisations as is already done by the RDA/E Forum.
- We need to find ways to engage with leading researchers in Europe, who have little time to go to RDA Plenary meetings but who could (and should) take profit from RDA's outcomes. RDA/Europe is currently planning its first RDA/Europe Science Workshop with 17 high level researchers as a vehicle to address this.
- We need to find more ways to motivate young data scientists to work on data solutions and implement RDA outcomes. The first steps have been made by RDA/Europe, but this needs to be intensified.
- We need to be more attractive to data archivists and curators and to develop special programmes for them. They are actually the experts dealing with many questions RDA is discussing.
- We need to include industry in the RDA process, to understand the feasibility of wide-scale implementations of the results in major software products. A major investment in efforts to foster innovation needs to be made in 2014.

Preface

This deliverable was due in month 12 (August 2013), but at the time the Description of Work (DoW) was agreed it was not apparent how fast RDA would grow and how fast relevant agreements could be reached at all levels of the RDA membership. Thus it seemed sensible to wait until after the Second RDA Plenary meeting in Washington the better to see what state we have achieved and which new perspectives we now have. The Washington plenary was a major step forward, since it:

- stabilized the organisation (Council, Technical Advisory Board [TAB], Organisational Advisory Board [OAB], Secretariat, Funders Group);
- had much deeper and more qualified discussions in most of the working and interest groups;
- showed that there is a growing global core group of experts (currently mostly from the US, EU and Australia) that believes in the future of RDA and that is participating in creating an RDA culture and terminology – all important ingredients of RDA's identity as a grass-roots based organisation.

We could not foresee these dynamics and therefore a few assumptions made in the DoW have been overtaken by events. This report reflects this new reality and describes the current state of RDA and RDA/Europe.

1 Introduction

The data domain is currently characterized by many actors and initiatives working in diverse, uncoordinated ways, and because of the pace of scientific and technological innovation we can assume that divergence will continue to increase. We can compare with the early days of computer networking and its explosion of approaches and solutions: only when agreement was reached on a few simple principles for a common basic layer (IP numbering, TCP/IP protocol stack, various registries) did worldwide communication truly profit.

In the domain of data we need to tread the same path: to identify a suitable harmonisation layer which must be determined by a simple layer of protocols, numbering systems and registries that has the potential to boost data sharing just as the introduction of TCP/IP and its associated specifications and organisational aspects boosted connectivity.

Such a harmonisation layer also will have the potential to allow for stable investments from all stakeholders in new sets of technologies and in new forms of harmonised data organisations. As stated in the *Riding the Wave* report[1] there will be no one technology that can cover all aspects that need to be addressed; rather we foresee the emergence of a variety of essential building blocks that may lead to improved interoperability at the various levels.

In D2.1 we describe the organisational structure of RDA that has already been implemented surprisingly quickly, with few tasks remaining outstanding. We know, of course, from the Internet community that governance structures change over time, and depend upon inherent dynamics. We already see a problem emerging which we will need to solve: through urgency and the lack of a fully-functioning TAB we began a series of regular interactions between the chairpersons of the five early working groups as a mechanism for early information exchange; now we have 23 active working and interest groups, we foresee that in a year this interaction of chairs will no longer work. It makes little sense to have video conferences with 20×2 experts (each group in general has two chairs), so for RDA to keep its momentum we will need to define *clusters of thematic topics* to group the WGs and IGs that will also be reflected in the TAB's way to follow group activities. This may lead to meetings between the chairs of the clusters which will have a limited scope and participation again. RDA's core leadership are aware of the need to be able to react flexibly in this way. RDA/Europe in its first year has already demonstrated its ability to react in an agile way.

Because of the advanced state of RDA's organisation we will focus on the non-organisational aspects of RDA's collaborations in this report. Right now, particularly in the working and interest groups, these are strongly driven by US and EU experts from cutting edge projects such as Datanet[2] (US), EUDAT[3] (EU) and OpenAire[4] (EU) (amongst others). This advanced state is also apparent at the political level: RDA has strong support from high political authorities in the US, Australia and the EU. The RDA Funders Group has been established as a funders' interaction body, and here also the US, Australia and EU actors at the political level agree very strongly on basic terms. In Europe some major actions have been initiated to complement the RDA process:

[1] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[2] http://www.cni.org/topics/ci/nsf-datanet-sustainable-environments-actionable-data-project/

[3] http://www.eudat.eu

[4] http://www.openaire.eu

- the RDA/Europe Forum[5] has been set up as a group of representatives of major European stakeholders;
- the first RDA/Europe Science workshop is currently being prepared to create close links with leading scientists;
- ideas to include industry in these interactions have been discussed and will be put into practice in 2014;
- ideas to reach out to young researchers have also been discussed and will also be put into practice in 2014 (already we have provided financial support for young data scientists to travel to the Washington Plenary).

It should be noted that any and all statements made in a roadmap for interoperability must anticipate the core developments in the data domain that will happen in the coming decades. Just as the *Riding the Wave* report did, we must ahead and try to tackle the significant problems which we can predict occurring. The major factors are "volume", "complexity", "diversity" and "dynamics", complemented of course by take-up by the community of data providers, and relevance to users (particularly science users).

[5] former called High Level Strategic Forum (HLSF)

2 Infrastructure Dimension

2.1 Roadmap and Infrastructure Principles

The data domain is currently driven by immediate research needs and brilliant ideas. Because of the tendency of research to apply bleeding-edge technology to yield new insights and shift knowledge borders, there is a sense of an increasingly divergent data domain making data sharing and re-use increasingly problematic.

The early days of computer networking were similar. To begin with it was almost only the physical layer of cables that was shared by everyone; it was the rapid convergence on fairly simple principles or protocol at a higher level that helped boost the connectivity of nodes worldwide and created the Internet. Despite the limitations of some of the solutions, the agreement on IP numbers, a few basic protocols and the setup of proper registries of nodes (in the DNS system) led us to the enormous explosion of global networked solutions. As R. Kahn[6] states, it is the ability to clearly identify the nodes, and the specification of simple protocol mechanisms between these nodes, that is at the core of the Internet and all its immensely valuable applications.

By analogy we can conclude that there is once again an urgent need to define a suitable harmonisation layer, firstly to boost data sharing and then to develop an increasingly rich domain of interacting data processing components. To learn from the Internet: we need to focus on the protocols and registries and leave to the endpoints the details of how to organise data. As noted in *Riding the Wave* there is no one technology that can solve everything, and no chance that a holistic solution can be designed top-down. We must begin from community-based data organisation solutions that have emerged over many years and that can only be adapted, if necessary, through a slow incremental process. The basic principles of data harmonisation can only be achieved by an approach that:

- is primarily grass-roots based but has complementary governance mechanisms to guarantee balance and coherence (i.e. bottom-up, supported also top-down);
- includes a variety of global stakeholders covering the most active regions of the world;
- analyses the state of the data organisations within the communities and regions and engages the different experts;
- focuses on identifying the necessary abstract building blocks of this protocol-centric layer.

Discussions in the *Data Foundation and Terminology* group highlight the need for this discipline and border crossing grass-roots approach. The collection of a large number of data models – static organisation descriptions as well as dynamic descriptions of workflows – is at the heart of defining a common language, identifying commonalities as well as differences and abstracting towards building blocks. This enterprise is currently complemented by the analysis work in RDA/Europe Task 2.3, by cross-disciplinary working initiatives (see D2.2) and by the cross-Atlantic collaboration projects in RDA/E WP3, all of which are focused on the better understanding of approaches and solutions and their abstractions towards suitable generic components.

[6] http://eudat.eu/system/files/B.%20Kahn.pdf

2.2 State of Data Infrastructures

In this short section we will limit our discussions to the US and EU regions, and even then will not be able to cover all projects and initiatives that are active in building data infrastructures.

2.2.1 European Level

At the European and national level we have many initiatives, some competing, to break down barriers in building state-of-the-art data infrastructures. First off we should mention the 48 ESFRI research infrastructures[7], most of which seek to make interoperable access to data and technology in a distributed environment much easier for researchers within specific disciplines or clusters of disciplines. At the core of many of these attempts is the setup of a network of trusted centers which store and preserve relevant data and permit further operations on them. Because of the intentions of these research infrastructures (RIs) to become long-term funded legal entities, the centers of the ESFRI RIs are at the root of European data infrastructures. However, all research infrastructures work to their own agenda and solutions, although so-called *cluster projects* have been started to make a step towards harmonisation.

Another cornerstone is the so-called *eInfrastructures* that have a cross-disciplinary approach. Here we should mention the European Grid Initiative[8] (EGI) that offers technology and some distributed services. EGI claim to have shifted from a purely computation-based approach to include a data approach, although it is still not their aim to preserve large amounts of data. EUDAT is a new initiative with the aim of building a collaborative data infrastructure as defined by the High Level Expert Group in *Riding the Wave*. Basically, it adds a layer of *common data services* to the heterogeneous community-specific data solutions. Much of the inspiration for RDA working groups in Europe has emerged from the many integration and interoperability problems that need to be solved for such a multi-site infrastructure, crossing national and discipline boundaries.

Another major initiative with a data aspect is the Helix-Nebula cloud[9] supported by CERN, EBI, ESA and a number of European companies. The intention is to develop a cloud solution that can be adopted by different service providers and thus offer cloud storage space for researchers in Europe and beyond. Persistence is being guaranteed by institutions such as the three mentioned. One service already making use of the Helix-Nebula is ZENODO[10] from OpenAIRE, offering researchers a place to drop their data. Another relevant initiative is Europeana[11] which provides a common portal for all digital objects stored in museums, archives and libraries. While Europeana only harvests metadata, its joint search portal brings all these holdings together virtually.

Two other important initiatives are DataCite[12] and EPIC[13], offering PID registration and resolution systems for everyone in research. While DataCite focuses on registering published, and thus citable, quality-controlled data, the EPIC federation focuses on efficient APIs to allow for the registration of the millions of data objects we are creating, to allow them to be referred to in workflows, for example.

---

[7] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

[8] http://www.egi.eu/

[9] http://www.helix-nebula.eu/

[10] http://www.zenodo.org/

[11] http://www.europeana.eu/

[12] http://www.datacite.org/

[13] http://www.pidconsortium.eu/

European efforts to come to an increasingly harmonised data infrastructure covering multiple levels are thus substantial.

2.2.2 National Level in Europe

At national levels in Europe we see, unsurprisingly, different strategies to establish data infrastructures that are broadly correlated with the size of country. In smaller countries such as the Netherlands and Finland a national strategy has been developed with strong national data centres offering core services. We see trends to split the *logical* and *physical* layers across different centers: one stores and provides access to the (physical) data, while another focuses more on (logical) metadata gathering and curation.

In larger countries such as UK and Germany the landscape is less clearly structured. Depending on national political organisation, different centres compete to offer national level services. Also strong research organisations such as the Max Planck and Helmholtz Society in Germany have their own data strategy where dedicated data centres have an organisation-wide remit with guarantees of persistence. In both Germany and the UK this is another emerging trend: until recently universities had no plans to build their own persistent digital repositories for data. We now see a change in opinion which has to do with visibility, image building and the research potential attached to the ownership of data – and, in the UK at least, the requirements of government funding councils.

We must of course mention the big libraries, archives and museums that are also seeking solutions to allow them to offer their digital content to readers and researchers. These institutions already have guarantees of long-term persistence, making them natural candidates to store and preserve data and offer access to them. In some cases these institutions focus on the logical layer leaving the physical layer to some big data centers. And we should also to refer to community driven centers and portals that are funded by national funds, but which nevertheless offer cross-national services.

There are many factors that determine the strategies and activities at national level. However there are a few strong centres emerging in each of the various countries to create the European data landscape.

2.2.3 US Level

In the US there is a wide variety of initiatives advanced by the different funding organisations. The NSF's Datanet program in particular funds a number of infrastructure and innovation initiatives. While the DataONE project[14] intends to build a persistent data infrastructure based on a federation of centres, and is investing in complementary activities such as community building, other projects such as SEAD [15] and DFC [16] are more technology oriented, i.e. developing and implementing new methods for shifting specific knowledge borders. Another important programme is DIBBs[17], supporting as it does projects that look at common building blocks for an emerging data infrastructure. EarthCube[18] is a well-known project funded in this programme.

[14] The DataConservancy project had similar intentions as DataONE, but was more technology bound

[15] http://sead-data.net/

[16] http://datafed.org/

[17] http://cyberling.org/node/823

[18] http://earthcube.ning.com/

Equally important are the big national labs that offer persistent data infrastructure solutions for specific communities, mostly coupled with data analytic offerings, along with a number of strong university strategies. However there is little interaction between the different actors except for an active exchange at meetings such as Datanet or ASIST[19].

2.2.4 Community Level

Some communities have data solutions at international level: the Protein Database[20], for example, the Genome Database[21] or the Astronomical Virtual Observatory[22] and there are many more of them. These are largely stable data collectors and providers, and will likely survive for a number of years due to significant user communities and interest from researchers. We can also categorise the big community data centers in Europe – CERN, EBI, ESA, etc. – in this way.

2.2.5 Commercial Level

An increasing number of commercial actors offer interesting data solutions, mostly positioned in a way that reduces the effort required for the logical layer of data, since this is expensive to maintain. DropBox[23] widely ignores the logical layer by offering a low level but very efficient data synchronization solution with a cloud-based system the data store. YouTube[24], Flickr[25], Figshare[26] etc. reduce their service to a specific standard data type, but also offer them in efficient ways, mostly backed by a cloud-based data store (often Amazon S3[27]). Some big players such as Google have been digitizing huge numbers of books, storing the digital content and offering related services.

Yet for all commercial offers there are huge trust problems for serious scientific data, since many, if not most, of these companies make no statements about where data are stored and who has access to them, and offer no long-term guarantees[28]. This raises concerns, particularly in Europe, and is why scientific organisations, national organisations and the EU see a need for Europe's "own" solutions.

[19] http://www.asis.org/

[20] http://www.rcsb.org/pdb/home/home.do

[21] http://www.ncbi.nlm.nih.gov/genome

[22] http://de.wikipedia.org/wiki/Virtuelles_Observatorium

[23] https://www.dropbox.com/

[24] http://www.youtube.com/

[25] http://www.flickr.com/

[26] http://figshare.com/

[27] http://aws.amazon.com/s3/

[28] For a comparison of the terms and conditions offered by 16 cloud storage providers in 2012, see Annex A of the *EUDAT Sustainability Plan* (available at http://www.eudat.eu/deliverables/d211-sustainability-plan).

3 Building Blocks

As we have noted there is no one technology that will solve all data-related problems, and imposing a solution top-down seldom succeeds. Therefore in this chapter we discuss a number of building blocks that we see emerging, and compare them with ongoing activities in RDA and other global initiatives. There is some overlap between some of the building blocks taken up within RDA, which is natural for a grass-roots initiative. In this document we will not try to cluster the various topics RDA groups are working on.

Aside from finding common language and terminology to better understand each other and to harmonise conversations in the data domain, recent discussions – in particular at the Washington RDA plenary – point to a number of building blocks that can already be identified very clearly as being essential. More building blocks will emerge from ongoing discussions, in particular those facilitated by the RDA; therefore this list is not meant to be complete or exclusive. For the building blocks listed below we indicate which regions are currently leading in each area, with the main drivers listed first.

3.1 Models and Terminology

Much work is currently being done to understand the various approaches and solutions taken to organise data. RDA/Europe and RDA has two tasks working on this:

- Task 2.3 in RDA/E is carrying out interviews with quite a number of communities and departments. WP3 in RDA/E also carries out joint collaboration projects between institutes at EU and US side to understand differences and commonalities based on concrete collaborations.
- The Data Foundation and Terminology[29] (DFT) WG in RDA also asked for submitting models (now 22 sketches) to base its analysis work on data organisations and to determine proper terminology based on harmonised models

State/Perspectives: The RDA DFT WG has collected 22 models from US and EU participants and has started analysis work on data organisations. Task 2.4 in RDA/E provided its first preliminary analysis as deliverable D2.4 in October 2013, leading to a workshop of all interested parties to discuss the results. WP3 will also organise a workshop with all collaborations soon to discuss the findings so far and compare them with their collaboration results. The results will then be feedback again to the RDA DFT which is aware of the need to extend its overview and analysis documents continuously.

Leading Regions: EU, US

3.2 Persistent Identifiers

It is widely agreed that we need a worldwide system for registering and resolving, with a recognized authority, persistent identifiers (PIDs) for all data objects that we create. PIDs created with "digital fingerprint" information, for example, will allow us to verify at any point in time the identity and integrity of digital objects, to refer to them in workflows and notes and to access them where there are instances of the object. This system must be open to every registered repository, and thus must be highly performant, robust and reliable and must

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

support APIs to register the millions of objects that are now being created every day in research.

State/Perspectives: The DONA[30] (Digital Object Naming Authority) will be established in October 2013 as a Swiss Foundation covering all the major regions of the world and it will take over steering of the Handle System[31] from CNRI. The initial set of major registration authorities will then start testing a worldwide redundant system of digital object naming. RDA is currently agnostic with respect to the PID system used, although it is widely agreed that only a global system based on a unified registry will help data harmonisation in the same way that DNS (for example) did for the Internet.

Leading Regions: US, EU, Australia, China.

3.3 Data Citation and Publishing

Increasingly, experts see the need to be able to cite data in publications. It is not yet clear whether citing and referring to data is the same and whether data citation includes the step of publishing data. *Publishing data* suggests that a data object or a collection has been quality controlled in some way and is ready to be used by everyone; *data citation* then means referring to such published data in a uniform way in publications so that the data being used can be accessed. DataCite for example specifies that registering a DOI[32] for an object or a collection means that certain attributes such as specific metadata fields must be provided, and some statement regarding quality and persistence is required.

State/Perspectives: Some global initiatives have made statements about data citation. In RDA the Data Citation WG[33] is working on this issue to come to widely accepted agreements. Within RDA there is also an IG on Publishing Data[34] that will work on the specific aspects.

Leading Regions: EU, US

3.4 PID API

From discussions in the RDA working group on PID Information Types[35] it is obvious that repositories want to store different types of information together with the PIDs and that there will be a large number of PID registration authorities for various reasons. We must ensure that, independent of the registration/resolution service, these all support the same core API. For software developers it must be possible to rely on the validity of an information request independent of the service provider: a request for a checksum and other core attributes must yield the same information everywhere.

State/Perspectives: The RDA WG on PID Information Types (PIT) is looking for information that should be associated with PIDs based on case studies. This group has genuine momentum and we expect to see suggestions for a core set of agreed information types at the RDA

[30] http://isocbg.files.wordpress.com/2011/09/dona-flyer2011.pdf
[31] http://www.handle.net/
[32] http://www.doi.org/
[33] https://rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html
[34] https://rd-alliance.org/internal-groups/publishing-data-ig.html
[35] https://rd-alliance.org/working-groups/pid-information-types-wg.html

Plenary in Amsterdam in September 2014. This will allow major players such as DataCite, Crossref[36], EPIC etc. to adapt their APIs towards a "standard" API.

Leading Regions: EU, US.

3.5 Metadata

There is no doubt that metadata describing the context and history (provenance) of a digital object is of enormous relevance for its interpretation and re-use, in particular in future workflow scenarios, yet the fragmentation in solutions and approaches is equally enormous. This is largely because metadata systems were defined reactively within disciplines to support contemporaneous research questions and to manage the data deluge. This is still an area of high flux, making it a non-trivial task to achieve a suitable level of harmonisation. The first work item, a registry of all major metadata schemas, can be built on what the Digital Curation Centre[37] (DCC) has already started.

State/Perspectives: The RDA Metadata information group and a working group[38] are active in discussing harmonisation steps, but it will take time to achieve major break-throughs. The registry of metadata schemas is an important first step. A Birds-of-a-Feather session on contextual metadata has been organised, but this work now seems to be included in the Metadata IG.

Leading Regions: EU, US.

3.6 Provenance Metadata

For the proper interpretation and re-use of research data, provenance metadata is as important as contextual metadata. Provenance metadata describes the history of processing steps that have been taken to generate a specific data object, such as a series of automatic transformations which may limit the number of useful or even valid subsequent operations. Provenance work in W3C resulted in the PROV[39] standard, and several research and data infrastructures are working on workflow implementations (web service orchestration and execution) that need provenance information. However, there is still no agreement on how to store provenance information which is "close to" the metadata work described above.

State/Perspectives: The RDA Research Data Provenance IG[40] has taken up this important aspect.

Leading Regions: EU, US.

3.7 Aggregations

Data aggregations are very common for storing, managing and accessing data which makes it necessary to have agreed mechanisms to build, store and manipulate them. Yet the terminology is not fully clear, since while some use the term *collection*, which normally associates semantics with an aggregation, others speak about *data sets*. The open archives

[36] http://www.crossref.org/

[37] http://www.dcc.ac.uk/

[38] https://rd-alliance.org/working-groups/metadata-standards-directory-masdir-working-group.html

[39] http://www.w3.org/TR/prov-overview/

[40] https://rd-alliance.org/internal-groups/research-data-provenance.html

community have created the ORE[41] concept in which a resource map with RDF assertions defines the aggregation. Some communities use metadata and some may use the PID record to do the binding.

State/Perspectives: The ORE definitions are available and can be used. Within RDA there is yet no initiative taking this aspect up.

Leading Regions: OAI, EU, US.

3.8 Trusted Repositories

N. Kroes, VP of the European Commission, stated that "data is the currency of modern research". It is thus widely agreed that we need a worldwide system of trusted and registered repositories that take care of this valuable commodity, i.e. that store, preserve and give access to data and metadata objects. Registering PIDs only makes sense if the objects they are pointing to are also stored persistently. Trusted repositories must be assessed regularly based on accepted criteria to assure users that their claims about policies being applied meet reality – and some idea of the *quality* of data stored must be considered (see below).

State/Perspectives: There is an increasing belief that trusted repositories need to be set up at various levels (community specific, organisation-wide, cross-disciplinary, cross-border). In Europe several ESFRI communities have begun to ask their repositories to participate in regular quality assessments. Other initiatives such as OpenAIRE, EUDAT and Helix Nebula also work on offers to the researchers. The availability of agreed assessment procedures, the setup of a registry for trusted repositories and pressure from funders will all help drive the emergence of trusted repositories. The EU is pushing this issue much more intensively than other regions. There are two RDA IGs working on these issues: Certification of Digital Repositories[42] (IG) and Preservation eInfrastructure[43] (IG).

Leading Regions: EU, WDS, US.

3.9 Certification of Digital Repositories

It is clear that coming to a network of trusted repositories requires the establishment of widely agreed assessment procedures and certification authorities that supervise the assessment. A few such assessment procedures have been defined by EU and US initiatives. They should be updated to meet emerging needs and require worldwide acceptance.

State/Perspectives: An RDA group[44] has been setup to stabilise and align the different suggestions: DSA[45]and WDS[46] and to disseminate the procedures across the world.

Leading Regions: EU, WDS, US

[41] http://www.openarchives.org/ore/

[42] https://rd-alliance.org/groups/repository-audit-and-certification/wiki/rda-candidate-working-group-certification-digital

[43] https://rd-alliance.org/internal-groups/preservation-e-infrastructure-ig.html

[44] https://rd-alliance.org/groups/repository-audit-and-certification/wiki/rda-candidate-working-group-certification-digital

[45] http://datasealofapproval.org/en/

[46] http://icsu-wds.org/our-members/membership-application/criteria-membership-certification

3.10 Clouds

Currently we see a differentiation between systems that, on the one hand, offer large storage capacity based on simple APIs and very fast access, and on the other include all kinds of organisational, contextual and referential information. We call the first the *physical layer*, and cloud solutions fill just this slot. We call the second layer the *logical layer*, a layer that will consist of a number of independent components such as PID systems, metadata, brokering etc. A "proper" repository should provide both layers for its own data, although we expect to see a rich landscape of services providers at the logical level. The relevance of trust in this context is widely underestimated and is viewed differently in the US and Europe.

State/Perspectives: Commercial offers are mainly driven by US companies, but initiatives such as Helix-Nebula, EGI, EUDAT and national ones in Europe also offer cloud services. Many of them can be used already, others will be launched in the coming months; many of the non-US initiatives will focus on the trust issue and thus arrive at different solutions. The relevant interfaces have been defined by, among others, Amazon (the *de facto* standard S3), by SNIA (CDMI[47]) and by the open cloud initiative (OpenStack[48]), yet there is no RDA activity on this issue.

Leading Regions: US on commercial offers, many countries, some communities.

3.11 Registry of Trusted Repositories

At the Washington Plenary (September 2013) a first BoF session was organised around this topic and there was agreement among a number of key players from widely known initiatives such as WDS[49], EUDAT, DataCite, DataONE[50], etc. who need to deal with an increasing number of repositories, that such a registry, including human (but especially machine) readable information, is urgently needed at a global level. Repositories are the nodes between which and from which data will be transferred, and requests for various kinds of information about a repository, such as the network port for harvesting metadata, should be found in a registry. At the moment too much information is implicit, or hard-wired in some software, restricting flexibility.

State/Perspectives: At the Washington Plenary a group was formed under the lead of WDS and including some of the major multi-site projects running in the US and EU. We expect this group to be inaugurated as an RDA working group at the Dublin Plenary. Based on the work that has already been done in initiatives such EUDAT, DataONE and past investment in Grid activities, we would hope for rapid convergence towards an agreement.

Leading Regions: EU, US, WDS.

3.12 Policy Framework

The execution of explicit policies should be the basis of all data management and processing since it will allow managers and users to trace what happened to a data object over time. Policy execution will be the basis for all assessments of trust and capability. Currently, many management policies are hidden in special software making it difficult if not impossible to

[47] http://www.snia.org/tech_activities/standards/curr_standards/cdmi

[48] http://www.openstack.org/

[49] http://www.icsu-wds.org/

[50] http://www.dataone.org/

check correctness. There is value in having a registry of such policies so that repositories can begin to harmonise their approaches. Since policies are translated into procedures which are then processed by workflow engines, it makes sense to have a reusable library of such procedures and certified workflow engines; only this will guarantee the level of trust needed to speak about trusted repositories.

State/Perspectives: Major US and EU initiatives support the idea of having "explicit policy rules" as the basis for proper data management and access. US colleagues lead this activity since they have running code (iRODS[51]) to implement and test policy rules. EUDAT has established a policy registry that is independent of a specific tool. At the Dublin Plenary we expect a first set of solid policy rules to be published for tasks such as data replication; others will follow in due course. Efforts to come to certified procedures and execution engines need to be taken. In RDA the Practical Policy WG[52] is working on these issues based on use cases and areas of policies.

Leading Regions: US, EU.

3.13 Data Synchronization

Policies are being used to determine the actions that need to be carried out for proper and verifiable data management. One action that is part of proper data management is the replication (or synchronisation) of web-accessible content. Here we mention the ResourceSync[53] activity of W3C which is an emerging standard that describes the data structures (resource list, dump list, etc.) that should be offered by a source repository so that other repositories can start replicating the resources indicated by the lists. Such lists can easily be embedded in a policy based data management.

State/Perspectives: As indicated W3C is working on the ResourceSync protocol which should be included by RDA.

Leading Regions: US, EU

3.14 Type Registry

A well-known problem for many data types created in science is that it is difficult to find, for example, viewers that can visualize the content of a digital object. Increasingly researchers want a quick appraisal of DOs from other communities, and this is hampered by an inability to access (or even find) the right software. The MIME[54] type system is limited to simple objects and is currently unsuitable to register an arbitrary number of different types. Thus we need type registries which allow scientists to register types and, for example, point to a piece of software that knows how to deal with this type – be it simple or complex.

State/Perspectives: The requirements of the type registry are currently being collected in the RDA WG on Type Registries[55]. In parallel, CNRI has a grant from the Sloan Foundation to implement a prototype. In Europe, EUDAT is committing itself to participate in the requirements specification process, to add to the software development efforts towards a robust tool and to set up such type registries to be open for researchers.

---

[51] https://www.irods.org

[52] https://rd-alliance.org/practical-policy-working-group.html

[53] http://www.niso.org/workrooms/resourcesync/

[54] http://www.iana.org/assignments/media-types

[55] https://rd-alliance.org/working-groups/data-type-registries-wg.html

Leading Regions: US, EU.

3.15 Syntax and Semantic Registries

Interpretation and re-use of data and metadata objects requires detailed knowledge about their structure and the meaning of the included symbols. For objects including "numbers" in a specific format, an increasingly used generic format description, NetCDF[56], has been defined. However in many other cases the underlying schemas describing the structure are not obvious. Even worse is the situation with respect to the meaning of the elements that are included. They are mostly not defined explicitly in a registry, making interpretation a tedious if not impossible enterprise.

A change in culture and tool support is required in the coming years to allow creators to specify their schemas and concepts easily. To support this, we need a much better way to make use of existing schemas and concept sets. This can only be achieved by having better mechanisms to register schemas, concepts and their relations. This is a field which still needs some careful thinking to find useful mechanisms.

State/Perspectives: This will take quite a while to formulate clear goals on top of current discussions in the existing RDA groups. Founding an interest group in this area was discussed explicitly between US and EU colleagues at the Washington Plenary.

Leading Regions: US, EU.

3.16 Data Categories and Codes

A topic very much related to semantic interoperability is the attempt to offer registries for concepts (data categories) and terms. The idea is simply that, if every community were to register and define their concepts and the labels they use to refer to these concepts, everyone would be able to make use of them in their own way, i.e. create an open market place for cross-semantic services. Currently too many concept definitions are not made explicit, are hidden in large ontologies that are not easy to use, are included in Wikis or prose descriptions etc. which makes it impossible to make significant progress.

State/Perspectives: Within RDA a WG[57] has been started to standardize language terms and codes with the help of a web-accessible registry supporting an ISO compliant data model.

Leading Regions: AU, EU

3.17 Mapping/Brokering

It is obvious that we have numerous protocols, schemas, APIs and semantic domains already in research and there is all reason to assume that research dynamics will lead to even more innovation. Tackling this diversity, with change everywhere, is a big challenge. Even if RDA were in a position to define a new layer of commonality which everyone could agree on, there will be higher layers where we need to do mapping between different solutions. These developments convinced some of us to think about new ways to perform mapping and thus achieve interoperability: (semi-) automatic brokering services seem to be one way forward, although ideas are very much in the early stages.

---

[56] http://de.wikipedia.org/wiki/NetCDF

[57] https://rd-alliance.org/standardisation-data-wg.html

State/Perspectives: There is an active group within EarthCube working on this issue and an Interest Group on Brokering[58] has been formed at RDA. However it will take time to establish proper mechanisms.

Leading Regions: US, EU.

3.18 Authentication and Authorisation (AA)

A functioning AA system at a global level is urgently needed, but it is well-known that there is no functioning cross-country solution for distributed authentication and authorization. These solutions include different aspects, from technology to politics and culture. This multi-dimensional nature of AA makes it difficult to achieve agreement even at the EU level, and certainly not across the Atlantic or including other regions. The Grid community have established a system based on user certificates, but this system cannot be transferred to other usage scenarios. With Shibboleth[59] we have some promising technology, and with SAML[60] we have a standard to exchange assertions in a safe way, but many aspects are still missing to make it a robust cross-border solution. With respect to technology we are missing, for example, a simple solution that allows us to delegate rights when starting a workflow chain that will invoke web services at different locations, access files at other locations, etc. With respect to politics we miss harmonised roles, a wide agreement to offer useful user credentials, and so on.

State/Perspectives: The EC wants to spend significant funds on a call to improve the situation. However, it is not clear whether this will indeed bring us to a robust and usable system. Currently tools like Contrail[61] and Moonshot[62] seem to offer progress, but we lack a global undertaking to overcome the limitations for a truly single-sign-on system based on single identities granted by trusted home institutions. Initiatives such as FIM[63] (Federated Identity Management) are helping us to understand the complexity of the requirements and indicate the solution space. Yet there is no activity in RDA although a research community based approach is urgently needed. Perhaps it would make sense to establish FIM as an IG within RDA to give the work more visibility and an embedding.

Leading Regions: No obvious champion yet to break down the barriers.

3.19 Legal Interoperability

An issue related to AA has to do with differences in legal systems, licensing agreements and ethical considerations in the various regions of the world. These differences are potentially the largest factor hampering progress in data sharing. Any improvement at global level to simplify legal and ethical issues across country-borders will be a big step ahead. Open Access is an important movement, but for many reasons (personality rights, commercial interests, incubation period, etc.) we will have to live with restrictions.

---

[58] https://rd-alliance.org/internal-groups/brokering-ig.html

[59] http://de.wikipedia.org/wiki/Shibboleth_%28Internet%29

[60] http://de.wikipedia.org/wiki/SAML

[61] http://contrail-project.eu/

[62] http://myterena.wordpress.com/2011/05/17/moonshot-grids-middleware-and-attribute-aggregation/

[63] https://tnc2012.terena.org/getfile/1333

State/Perspectives: RDA has initiated an IG group on legal interoperability[64] which is currently studying existing hurdles and possible improvements for different communities as case studies. The communities were selected to bring people together from various areas.

Leading Regions: US, EC.

3.20 Community Activities

RDA is a good umbrella for communities which do not yet have international forums to discuss data sharing and interoperability, or which have several such initiatives and want to move towards convergence. There are a number of community specific harmonisation activities within RDA upon which we will not comment here. These are, of course, very important within the RDA realm: Agriculture, Urban Data Exchange, Data Practices in History and Ethnography, Marine Data Harmonisation, Structural Biology and Toxicogenomics Interoperability.

State/Perspectives: We assume that there will be more community specific actions to use the RDA umbrella to gain from cross-fertilization aspects. These groups are also important for RDA itself since they bring data producers and users in the RDA community. Other communities have already developed their own forums and solutions for data sharing and interoperability and it is critical for RDA that they participate to share their findings and assess applicability of the solutions proposed by the RDA.

RDA Activity: EU, US

3.21 Other Activities in RDA

In addition to what has been described above we can identify a few other activities in RDA targeting obvious gaps on the way to an interoperable global data domain. Two groups are addressing aspects of scientific culture:

- The Community Capability Model WG[65] addresses the issue of collecting, validating and publishing a range of data-centric "capability profiles" to enhance inter- and intra-domain interoperability and catalyse RDA data-sharing goals.
- The Engagement IG[66] addresses the question of how to get the researchers dealing with data every day involved in the RDA interactions, knowing that all are under an enormous pressure to publish.

Two other groups do not have yet a clear focus but are analysing two areas of high interest:

- The Big Data Analytic[67] IG is addressing the question whether this new field will need proper agreements on data.
- In a similar way the Long Tail of Research Data[68] IG is addressing whether for dealing with long tail data will require additional considerations.

[64] https://rd-alliance.org/internal-groups/legal-interoperability-ig.html
[65] https://www.rd-alliance.org/working-groups/community-capability-model-wg.html
[66] https://www.rd-alliance.org/internal-groups/engagement-group-ig.html
[67] https://www.rd-alliance.org/internal-groups/big-data-analytics-ig.html
[68] https://rd-alliance.org/internal-groups/long-tail-research-data-ig.html

4 Roadmap

The ultimate goal of this deliverable, as it was originally formulated, is to characterize a *roadmap* towards full interoperability of EU-US data infrastructures. This formulation needs to be extended since Australia is already participating in these global interactions on improved sharing and interoperability, and we can expect that soon other regions and countries will join. We must therefore speak about a *global roadmap* for which finally RDA TAB has the final responsibility. From what has been discussed above it is obvious that there are many actors and initiatives that are working towards improved sharing and interoperability.

Primarily, it is the work in the RDA WGs and IGs, with their own dynamics, which will foster sharing and interoperability. It is the WGs and IGs where "data practitioners" from various research disciplines and regions will bring in their knowledge and experience to plan steps ahead.

Secondly, it is the work in the existing global standardization and best practice initiatives such as W3C[69], IETF[70], WDS/ICSU, CODATA/ICSU[71] and ISO[72] (to mention but a few highly important ones).

Thirdly, it is the analysis work on data models and organisations as carried out within RDA/Europe Task 2.3 which is contributing already now to the work in RDA WGs such as Data Foundation and Terminology and which will get a new impulse when the first analysis summary will be published.

Fourthly, it is the concrete collaboration work in global or cross-Atlantic collaboration work as it is being carried out in RDA/Europe WP3. Here we expect concrete contributions from the four collaborations (Data Intensive Astronomy, Earth Sciences, Open Data and Interoperability, Chemical Safety) in 2014.

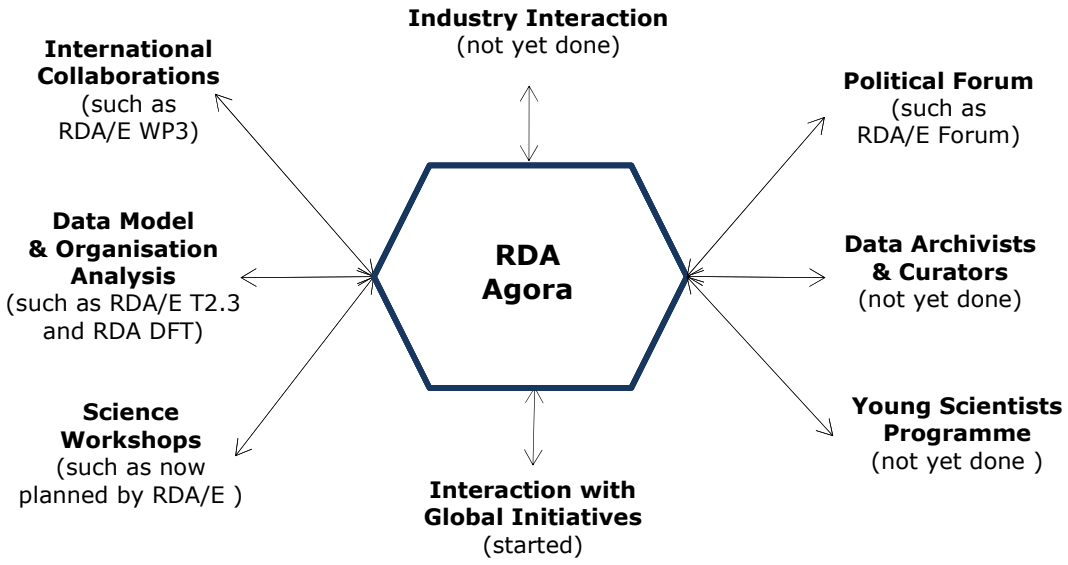
Our analysis of the state of RDA and the broader data harmonisation landscape leads us to a number of concluding recommendations:

- There needs to be an intensive discussion at political level in Europe to cross-fertilize with relevant scientific organisations as is already done by the RDA/E Forum.
- We need to find ways to engage with leading researchers in Europe, who have little time to go to RDA Plenary meetings but who could (and should) take profit from RDA's outcomes. RDA/Europe is currently planning its first RDA/Europe Science Workshop with 17 high level researchers as a vehicle to address this.
- We need to find more ways to motivate young data scientists to work on data solutions and implement RDA outcomes. The first steps have been made by RDA/Europe, but this needs to be intensified.
- We need to be more attractive to data archivists and curators and to develop special programmes for them. They are actually the experts dealing with many questions RDA is discussing.

[69] http://www.w3.org/
[70] http://www.ietf.org/
[71] http://www.codata.org/
[72] http://www.iso.org/iso/home.html

- We need to include industry in the RDA process, to understand the feasibility of wide-scale implementations of the results in major software products. A major investment in efforts to foster innovation needs to be made in 2014.

RDA, with its interactions with international standardization organisations, needs to become the anchor and meeting point for all contributions. It has already been established as the place for the global and cross-discipline interactions driven by the data practitioners from the various disciplines. It is the broader RDA Agora[73], consisting of the plenary interactions, the WG/IG interactions and the online interaction forum, that will help to define the roadmap driven by the most urgent needs of practitioners on the one hand and balanced by the interactions at political and scientific level on the other hand.



In addition, RDA has its own internal mechanisms to balance between the grass-roots-driven activities and the top-down guidance. With respect to the latter, it is important to mention the Council, with "statesmen" taking an overview and managing political balance; the Technical Advisory Board, with its senior technical people to identify overlap and gaps; and the Organisational Advisory Board, managing organisational balance within RDA. We see this document, indeed, as an input to the RDA task of discussing and continuously updating an RDA roadmap.

We conclude with the observation that describing the roadmap for RDA is not a static one-time action, but a result that will emerge from the RDA Agora, where principles such as *rough consensus forming* are essential to reach extended plans. It has already been astonishing how fast the various experts have synchronized on a number of urgent action points, and how much coherence could be achieved in the "chairs meeting" of the existing working groups. The progress at the RDA Washington Plenary – clearly visible – gives us confidence that we are on the right track towards a balanced, open and grass-roots-driven landscape of harmonised building blocks and protocols.

[73] http://de.wikipedia.org/wiki/Agora