Data Management Trends, Principles and Components – What Needs to be Done Next?

Authors: Bridget Almas (Tufts), Juan Bicarregui (STFC), Alan Blatecky (RTI), Sean Hill (EPFL), Larry Lannom (CNRI), Rob Pennington (NCSA), Rainer Stotzka (KIT), Andrew Treloar (ANDS), Ross Wilkinson (ANDS), Peter Wittenburg (MPS), Zhu Yunqiang (CAS)

Acknowledgement: authors are thankful to RDA-Europe[1] and RDA US[2] for supporting this work.

This document is a ***request for comments*** from the authors with the purpose to seed discussions about components that need to be put in place to support data practices, to make them efficient and to meet the data challenges of the coming decade.

1. Introduction

RDA (Research Data Alliance) is a worldwide initiative to improve data sharing and re-use and to make data management and processing more efficient and cost-effective. After two years of intensive cross-disciplinary and cross-country interactions within and outside of RDA and after having produced concrete results of the first Working Groups, four factors have been identified:

- some inefficiencies of our current data practices[3]
- stabilized data principles from various funders and initiatives
- some widely accepted common trends
- on-going discussions about the consequences of principles and trends and the components which seem to be urgently needed

RDA aims to be a neutral place where experts from different scientific fields come together to determine common ground in a domain which is fragmented and, by agreeing on "common data solutions", liberate resources to focus on scientific aspects. RDA does not claim to be the first to comeup with ideas and concepts, since important contributions will often come from other discussion forums and research efforts.

In this document, we refer to current data management trends (chapter 2), discuss principles (chapter 3) and their possible consequences (chapter 4), and review some components that emerge as being required (chapter 5). While trends and principles seem to be widely agreed, the the type and nature of these components are still being debated. This paper aims to promote discussions that will lead to consensus across national and disciplinary boundaries about what is needed to meet the data challenges of the coming decades.

[1] RDA-Europe is an EC funded initiative, Grant 632756

[2] RDA/US receives primary support from the National Science Foundation. Larry Lannom's RDA/US efforts are supported by NSF ACI-1349002

[3] http://europe.rd-alliance.org/sites/default/files/RDA-Europe-D2.5-Second-Year-Report-RDA-Europe-Forum-Analysis-Programme.pdf

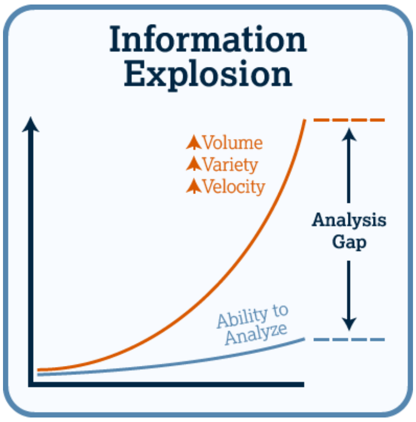
In appendix A, we define some terminology about roles and tasks in the data domain as we are using them in this document. In appendix B, we add some elaborations on the components mentioned in chapter 5.

2. Common Trends

Summarizing discussions in RDA and community meetings during the last months we can observe a few common trends that we will briefly explore in this chapter.

2.1 Changing Data Universe

Many documents, such as "Riding the Wave[4]", commented on the major developments in the data domain, such as increasing volumes, variety, velocity, and complexity of data, a need for a new basis for trust, increased re-usage of data even across borders (disciplines, countries, and creation contexts) and so forth. Figure 1 indicates the increasing gap between the amount of data that we produce and the amount of data that we are actually capable of analysing.



These developments are widely known and mean that we require new strategies for managing data if we are to keep up with and use what we generate. These matters are widely discussed and well known, yet not all consequences of the explosion are well understood.

Figure 1 the information explosion

2.2 Layers of Enabling Technologies

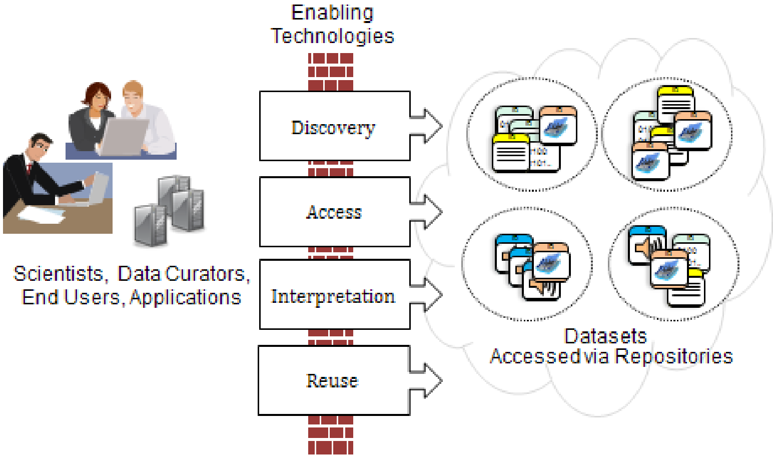

Enabling Technologies
Discovery
Access
Interpretation
Reuse
Scientists, Data Curators, End Users, Applications
Datasets Accessed via Repositories

Figure 2 shows layers of enabling technologies

Without going into detail, we can see that there is a wide agreement on the layers that can be distinguished when working with data and which were discussed at the DAITF[5] ICRI workshop in 2012[6]. These layers should be dealt with by different technology stacks as indicated in figure 2 since the properties of data which are being processed at the various layers are different. We find these layers mentioned back in documents from G8[7] and FAIR[8], amongst others, and they have guided our

[4] Riding the wave, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[5] DAITF = Data Access and Interoperability Task Force, which merged into RDA in 2012

[6] Larry Lannom, 2012, *DAITF: Enabling Technologies*, DAITF ICRI 2012 workshop

[7] G8 London Science Minister Statements, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

[8] FAIR Data Principles: https://www.force11.org/group/fairgroup

way of structuring and what it is needed in the data infrastructure of the future.

2.3 Data Management Commons

There is an increasing understanding that, to a certain extent, we can claim that all basic data elements (which we will call Data Objects) can, when ignoring their content, generally be treated as being discipline-independent, in much the same way as email systems are being used across disciplines.

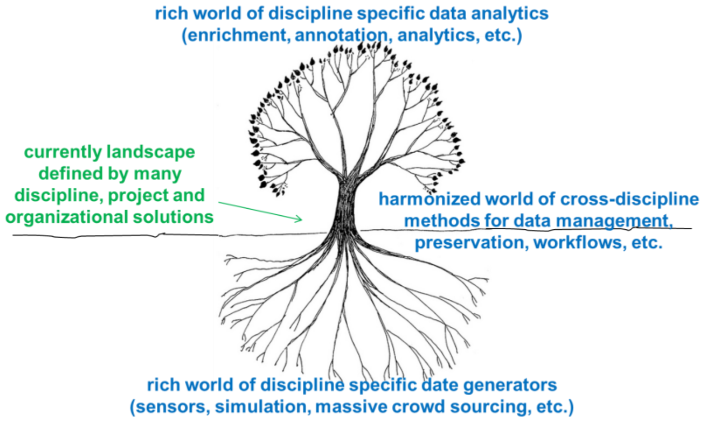

rich world of discipline specific data analytics
(enrichment, annotation, analytics, etc.)
currently landscape defined by many discipline, project and organizational solutions
harmonized world of cross-discipline methods for data management, preservation, workflows, etc.
rich world of discipline specific date generators
(sensors, simulation, massive crowd sourcing, etc.)

Figure 3 data management commons

However, we need to distinguish the external characteristics of data from the internal characteristics to ensure that we really can separate common data management tasks from discipline-specific heterogeneity in the processes of creating and analysing data as indicated in Figure 3[9]. It is not yet obvious where the borderline between external and internal properties of data actually lies, but we do know for sure that it is best to regard all the information describing the structure and semantics of the contents of a Digital Object as coming under the internal properties.

2.4 Central Role for PIDs

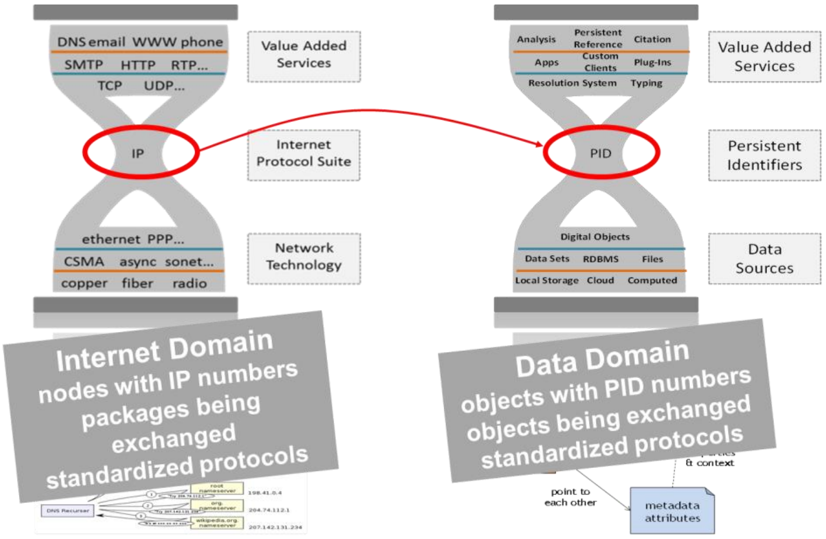


Figure 4 The PID - IP-number analogy

In many research communities dealing with data intensive research there is now widespread agreement on the central role of PIDs (that is, Persistent (and Unique) Identifiers). The idea of PIDs was introduced at an early moment[10] and is explained in figure 4 [11] where a parallelism is drawn between the use of IP addresses in the Internet and use of Persistent Identifiers in the domain of data that is to be shared and re-used. PIDs associated with some additional information – such as fingerprint data (checksums) – are strong mechanisms, not only to find, access and reference Digital Objects independently of their location, but to also check identity and integrity even after many years. Of course, PIDs on their own aren't enough – it is also necessary to have systems that allow the PIDs to be resolved, mechanisms to allow the resolution targets to be updated, and storage

[9] http://www.elettra.eu/Conferences/2014/BDOD/uploads/Main/Booklet_BDOD.pdf
[10] http://www.cnri.reston.va.us/k-w.html
[11] http://www.eudat.eu/events/conferences/eudat-1st-conference

solutions that allow the referenced objects to persist. PIDs can also be used to identify publications and services.

2.5 Registered Data and Trusted Repositories

We are seeing that is becoming more and more important that we distinguish between registered and unregistered data. There is a huge and growing amount of unregistered data stored on various computers and storage systems which is not described and not easily accessible, and may even be partly hidden for certain reasons. We cannot make useful statements about the management, curation, preservation, and citability of this data since it is not officially integrated into the domain of data that is subject to explicit rules. As a result, researchers making use of such data cannot rely on it complying with any data management mechanisms that are widely agreed upon. Most of the data that is currently being exchanged between researchers is unregistered data, which means that it is often necessary to copy the data to in-house storage systems before re-using it.

When we discuss "registered data", it is also important to be aware of the terms "Persistent Identifiers" (PIDs) and "Trusted Repositories". PIDs have been mentioned previously as anchors that uniquely identify Data Objects and thus make it possible to find and access data. Trusted Repositories are data repositories that follow certain explicit rules (such as the DSA[12] and WDS[13] standards) with respect to data treatment. Thus Trusted Repositories explicitly state what can be expected by people who deposit data in them and by other people using data from the repositories. We must remember that Persistent Identifiers and their extensive capacities are useless if the associated data ceases to be available, that is, if there is no persistence of the data itself.

2.6 Physical and Logical Store

Over the last few decades, there has been a gradual change in the methods used for storing different types of information that has mainly come about as a result of the trends described in section 2.1. Traditionally the bitstreams that carry the information content of the data were stored within structures inside files and, in most cases, the file type (as shown by the extension of the file name) indicated how to extract the information from the file. Organisational and relational information, such as the experimental context of the data, was indicated by the choice of directory structure setup and by the directory and file names. However, with the increasing volumes and complexity of data that are being produced, we have been seeing a growing need to describe more details of the properties of the data and of the relationships between different digital objects often being established long after their creation. In addition, traditional file systems were simply not designed to offer efficient access to millions of objects as that was not necessary when they were initially developed. Some disciplines, particularly those that deal with data being automatically captured by instruments and sensors, have moved away from a file-based approach to large, multi-table

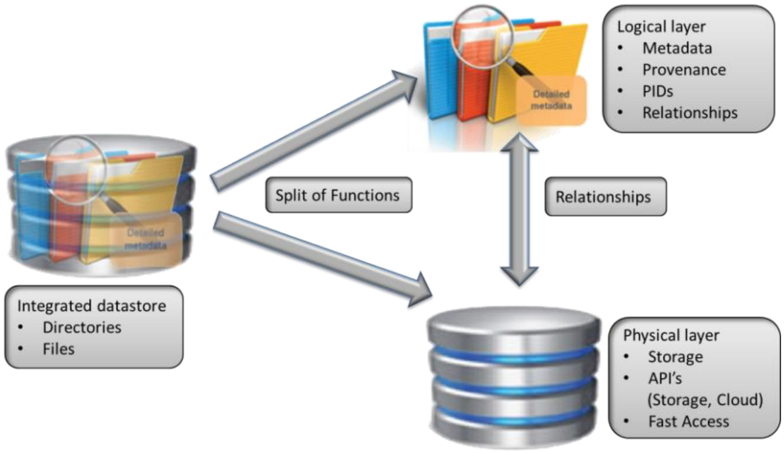


Figure 5: Split into Physical vs. Logical Storage as an evolution to better deal with volume and complexity

[12] Data Seal of Approval: http://datasealofapproval.org/en/

[13] World Data System: https://www.icsu-wds.org/

databases. With this approach, the distinction between the data and the metadata is less clear – both are columns in a table, and the relationships are stored as primary key <-> foreign key pairs.

Over time, two different trends in data storage emerged. On the one hand, new simple and fast technology (namely clouds) became widely available, and reduced the amount of descriptive information for each data item by basically using one internal hash tag per stored item. On the other hand, people started to build complex structures to store metadata, provenance information, PIDs, information about access rights and various sorts of relationships between digital objects. This split between a simplified physical layer and a complex logical layer is indicated in figure 5. Figure 6 indicates this split from a different perspective. The physical storage system can be optimized to access while the "logical" information is being extracted to a cloud of services making the different types of information accessible to the users, which can be humans or machines.

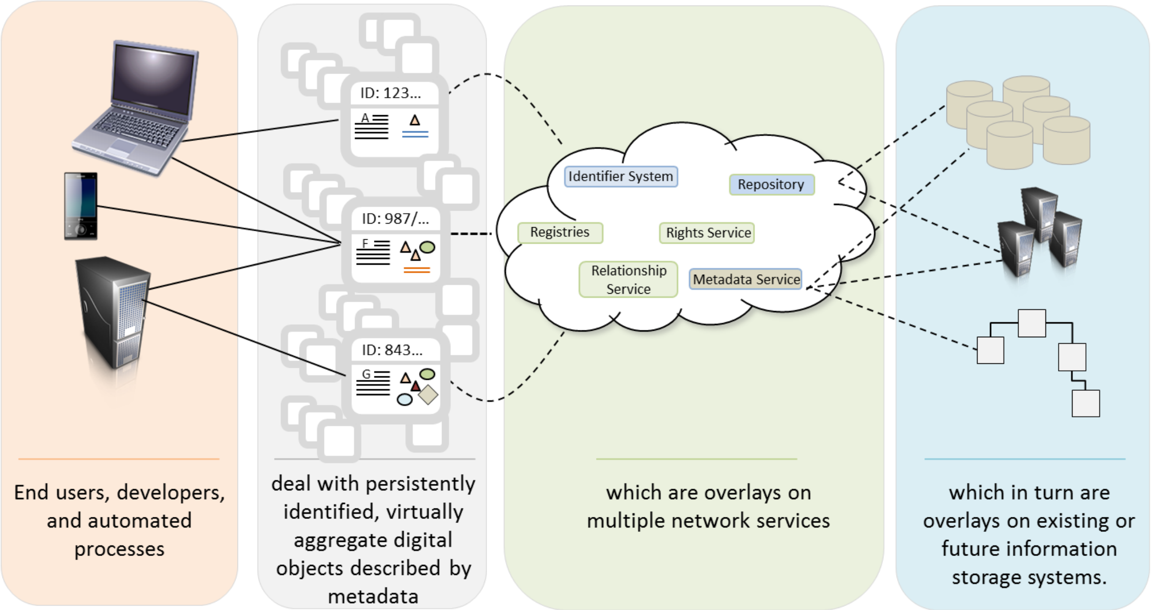

ID: 123...
ID: 987/...
ID: 843...
Identifier System
Repository
Registries
Rights Service
Relationship Service
Metadata Service
End users, developers, and automated processes
deal with persistently identified, virtually aggregate digital objects described by metadata
which are overlays on multiple network services
which in turn are overlays on existing or future information storage systems.

Figure 6: This diagram indicates virtualization layers. Users deal with Digital Object information stored in PID records and metadata descriptions. These are accessible via a variety of registries referring to bit sequences stored in storage systems.

As of now, there is no common agreement on guidelines about how to store, organize and manage all this information (which we call "logical information[14]") and how to maintain all relations, although some efforts are headed in the right direction. Assembling this information, virtually combining data and the essential metadata needed to understand it, is the goal of the Digital Object Architecture, from the Kahn/Wilensky framework referenced above, subsequent implementations such as Fedora Commons Objects, and the recently published ITU-T X.1255 digital entity data model. There are different implementations to store and organize this kind of logical information including structured database solutions (for example, using relational databases or XML) .

2.7 Automatic Workflows

There is an increasing conviction that (semi-)automated workflows that are documented – and that are themselves self-documenting (in terms of metadata, provenance, PIDs and so forth) – will be the only feasible method for coping with the data deluge we are experiencing and for keeping data intensive science reproducible. These automated workflows (which, in RDA contexts, are guided by practical policies) will take bitstreams of many digital objects as input. In some way they will then

[14] All the types of information that are included in the "logical layer" can be seen as metadata, however, they generally have different purposes, and hence it makes sense to differentiate between them.

read the PID record (to locate instances of the digital object, to check integrity etc.) and the metadata (to be able to interpret its content). As it is important that the process be reproducible, the workflows will then create one or more new digital objects that are associated with metadata, including rich provenance information, and new PID records with useful state information. This process, which is schematically indicated in diagram 6, is independent of the type of action that is being carried out by the processing unit – it could be a typical data management task such as data replication or a scientific analysis task.

While it is obvious that such self-documenting workflows offer many advantages, in daily practice, systems implementing self-documentation in the manner illustrated in figure 7 can rarely be observed.

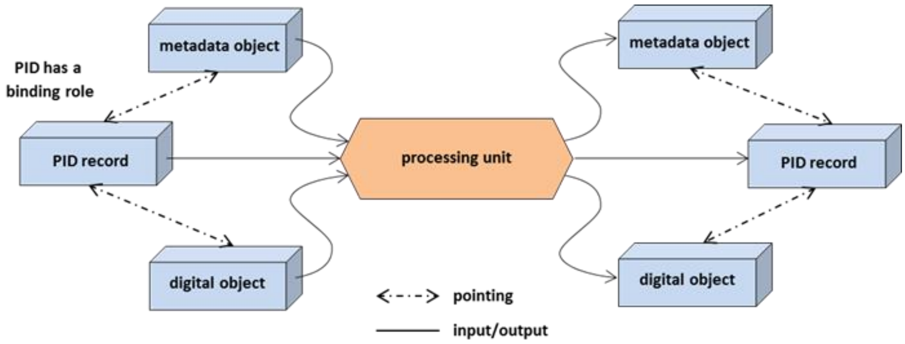


Figure 7 indicates the nucleus of a self-documenting workflow engine that has the potential to create reproducible data science. It makes use of PID and metadata information describing the digital objects being processed and about the knowledge about the processing details to create metadata descriptions and register a new PID for the new data being generated.

2.8 Federations

We are now seeing a broad trend towards working in data federations for various purposes. These federations are networks of data repositories and centres that offer processing frameworks and that act based on agreements about legal and ethical rules, interface and protocol specifications and a stack of common services for handling data. Increasingly often such centres are members of multiple federations: a climate modelling centre, for example, is a member of the climate modelling data provider federation, as well as being a member of the EUDAT data federation and also a member of the European AAI federation. This trend is likely to continue and will lead to even more federation arrangements.

Currently, all these data federations are being created without having the whole picture in perspective. This means that, for each federation, each centre creates and maintains its own form of description of its characteristics (both for humans and machines). This is very inefficient and urgently needs to be replaced by a coordinated approach where each centre creates a description of its characteristics based on a widely agreed upon set of properties, so that for each federation the same description can be (re)used to extract the information needed.

3. Principles

In various policy forums and initiatives a number of data principles have been established. The following documents are of relevance in this respect:

- G8 Principles for an Open Data Infrastructure[15],
- G8 Ministers Statement London[16],
- U.S. OSTP Memorandum on Increasing Access to the Results of Federally Funded Scientific Research[17]
- U.S. OMB Memorandum M-13-13: Open Data Policy – Managing Information as an Asset[18]
- HLEG Riding the Wave[19],
- RDA Europe Data Harvest Report[20],
- Research community results (such as FAIR[21] and FORCE11[22] Recommendations and the Nairobi Principles[23]),

A comparison of principles[24] shows that they all elaborate on a number of core principles that are relevant for data management/stewardship and show wide agreement:

- Make data discoverable to enable it to be used efficiently.
- Make data accessible with as few restrictions as possible to enable it to be used.
- Make data understandable to enable it to be re-used effectively and to make it possible to extract knowledge.
- Make data efficiently and effectively manageable to guarantee that it can be used in the long-term and to make sure that proper acknowledgment methods are used.
- Train people who are able to put efficient mechanisms into practice.
- Establish useful metrics to measure the use and the impact of investments in data.

4. Consequences of Principles

It is important to see interest supporting the convergence of these principles, but it is now as well important to face the consequences that follow from them and list actions that are required to make them reality in the daily data practices. As has been shown by the Report on Data Practices[25] we are currently far away from a satisfying situation. This chapter describes a number of courses of actions which seem to follow from the principles and that need to be discussed broadly. There will be questions about the intentions of the statements below, there will be disagreement and questions about feasibility, they will certainly not be complete etc. This document is therefore a ***request for comments*** and we create a place on the RDA Data Fabric Interest Group (DFIG) wiki[26] to open a broad discussion.

[15] G8 Principles for an Open Data Infrastructure, *http://purl.org/net/epubs/work/12236702* (*White Paper: 5 Principles for an Open Data Infrastructure*, editors on behalf of the Data Working Group of the G8+O6 Group of Senior Officials on Research Infrastructures, Dr A Blatecky (NSF) (Ed.), Dr J Bicarregui (STFC Rutherford Appleton Lab.) (Ed.), Dr C Morais Pires (EC) (Ed.) – May 2013

[16] G8 London Science Minister Statements, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

[17] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[18] https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

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

[20] https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html

[21] FAIR Data Principles: https://www.force11.org/group/fairgroup

[22] FORCE11 Data Citation Principles: https://www.force11.org/datacitation

[23] Nairobi Data Sharing Principles: http://www.jkuat.ac.ke/wp-content/uploads/2014/08/Nairobi-Data-Guidelines-20142.doc

[24] http://hdl.handle.net/11304/1aab3df4-f3ce-11e4-ac7e-860aa0063d1f

[25] http://europe.rd-alliance.org/documents/articles-interviews/rda-europe-data-practice-analysis

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

4.1 Change Data Culture

- Make research data "open" by default and help change the current research culture to promote data sharing.
- Convince researchers to adhere to a simple high-level data model with digital objects being registered and metadata described.
- Educate researchers to do proper data citation to acknowledge data-related work.
- Help change the existing culture to make data work be a recognized part of CVs and included in metrics for granting tenure.
- Help to define proper mechanisms to use data citations in impact metrics.
- Help to train a new generation of data professionals (see appendix A for definitions).

4.2 Discoverability

- Describe each digital object with adequate metadata to support data discovery.
- Register the digital objects and make the discovery metadata available via machine-readable interfaces, e.g. OAI-PMH.
- Register metadata schemas and their semantic categories in open registries to facilitate the process of metadata interpretation.
- Register metadata vocabularies that are being used in open registries.
- Associate suitable information with PIDs to make it possible to trace digital objects back.
- Create provenance records that make it possible to trace back digital objects history.

4.3 Accessibility

- Store digital objects in trusted repositories to make them accessible and referable.
- Have repositories adhere to certification rules.
- Assign a PID to each deposited digital object to register it and make it citable.
- Declare the legal, ethical, privacy and license rules each repository will apply.
- Define the protocols for access permission negotiation.
- Define the access protocol, for each repository, for accessing digital objects.
- Reserve sufficient funds to enable suitable data stewardship.
- Create a limited set of widely usable license models for data[27] (in analogy to Creative Common licenses).
- Create model agreements that make it possible to establish international data federations.

4.4 Interpretation and Re-use

- Associate information with the assigned PIDs that makes it possible to prove identity and integrity.
- Describe each digital object with metadata – including contextual information such as prose text describing the creation process and its manifold relationships) – that supports interpretation and re-use of the data.
- Register schemas and their semantic categories in open registries to make it possible to interpret content.
- Register vocabularies being used in open registries.
- Register data types in type registries and associate executable data processing functions with them.

4.5 Data Management/Stewardship

- Define policies that guarantee accessibility, preservation and re-usability of data over time.

[27] RDA's Legal Interoperability IG is working on this topic.

- Develop practical policies that turn these policies into executable and verifiable procedures and register them in open registries.
- Create policy frameworks that make it possible to easily integrate scientific algorithms into pre-fabricated workflows.
- Create registry systems for different purposes (such as storing metadata and information about schema, semantics, vocabularies, and practical policies) where necessary and maintain them.
- Agree on rules for data creation and handling that allow us to assess the quality of the data.

5. Technical Components

From the previous chapters, the discussions within RDA, in particular the Data Fabric Interest Group (DFIG), and discussions with various communities, a number of components can be identified that seem to be required to become available as professional services accessible to everyone. For many of the components mentioned below, some suggestions have already been made, thus we do not need to start from scratch. What we are often missing are cross-border systemic[28] solutions with professional support.

A broad discussion is also required about respect to the components and their services. Therefore we see this chapter as a ***request for comments*** to open this much needed discussion on the DFIG wiki[29]. In this chapter we briefly indicate components. In appendix B there are elaborations on these components.

5.1 PID System

A worldwide highly available and scalable PID system is needed that can be used immediately from the moment when a digital object (DO)[30] is created to uniquely reference that DO in scripts, practical policies and so forth, and thus guarantees reproducibility of the results of data intensive science at all steps.

5.2 ID System for Actors

A system is needed that is able to uniquely identify users who are involved in the data creation process. Currently ORCID[31] is well accepted by some communities for registering authors, however, in parallel we are using other systems for authentication.

5.3 Registry System for Trusted Repositories

A “Birds of a Feather” session on the topic of repository registries was organized at RDA's Fifth Plenary, where a decision was made to produce a position paper that details what a group within the RDA working on a Repository Registry could achieve. Some stakeholders need such a registry system.

[28] We are using the term "system" for all components to indicate the systemic approach without any particular implementation in mind.

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

[30] It should be noted here that practices include already associating PIDs with code versions, sensor configuration documents and others.

[31] http://orcid.org/

5.4 Metadata System

Metadata systems are an extremely complex topic since the term "metadata" can be used in different contexts to denote different things. Despite the many years of discussing metadata in its different forms we are still far away from widely and efficiently used metadata solutions.

5.5 Schema Registry System

For each digital object in the common data infrastructure, both the format and structure of the object is needed to allow users to parse the object and extract information from it. Also here we cannot speak about a systemic solution that can be used efficiently despite all the many approaches which we currently have.

5.6 Registry System for Semantic Categories, Vocabularies, etc.

When making use of data or metadata, a basis must be given to understand the meaning of all its elements. Therefore, shared semantic categories, vocabularies, taxonomies and so forth[32] that include concept definitions and partly concept relationships are very important. Much has been done in this area, but its usage in daily data practice is still poor.

5.7 Registry System for Data Types

RDA's Data Type Registry Working Group (DTR WG)[33] presented the concept of a Data Type Registry (DTR) which links data types of all sorts with the executable data processing functions that can be useful to work with a specific data type. We need to make it a usable service.

5.8 Registry System for Practical Policies

The RDA's Practical Policy (PP) WG is working on a list of best-practice practical policies (PPs) which are recognized as being the basis for self-documenting and reproducible data processing and managing. We need to have a registry system that is ready to be used by everyone to re-use policies.

5.9 Prefabricated PP Modules

Currently, when speaking about practical policies, we most often discuss PPs that fulfil a certain function, such as replicating a Digital Object. However, we can identify generic components for such PPs that, for example, implement standard functions such as required by RDA-compliant data management (register a PID, etc.). We need to provide a number of standard components to facilitate the integration of scientific algorithms into the complex policy landscape easily.

5.10 Distributed Authentication System

Obviously we need to have a professional and secure system to allow us to authenticate users across borders. Yet, even at national and regional levels, we do not have any such systems which meet the essential criteria and are also working smoothly.

5.11 Authorization Record Registry System

Up until now, distributed authentication (rather than authorization) has been the main focus in data access discussions. However, increasingly often, we see scenarios where data is for example being replicated for various reasons. To make all the instances of the data accessible, the data’s authorization records need to be accessible, for example, across different data repositories. Until now we do not have an efficient and secure solution that guarantees that authorization records are the same for all copies.

[32] In this paper we will use the term "semantics" to refer to all these entities. Some prefer the term "ontologies" which we would like to reserve for entities that combine concepts and their relationships.

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

5.12 OAI-PMH, ResourceSync, SRU/CQL

Metadata is an important part of any data management infrastructure. Special protocols exist for aggregating or harvesting metadata (OAI-PMH[34], ResourceSync[35]) as well as transferring metadata and data queries (SRU/CQL) to search engines. Yet they are not used broadly.

5.13 Workflow Engine & Environment

When it comes to executing practical policies, and any other kind of process chains orchestrated by a user, there are several essential components: a flexible workflow engine (such as Taverna or Kepler), and an environment which allows users to easily deploy and execute practical policies (see 5.8) and scientific algorithms at centres offering sufficient computing power (and often storing the relevant data in the neighbourhood of HPC computers[36]). The exact requirements for a suitable workflow engine and a sufficiently flexible solution for a whole environment have not yet been fully established, as there are many aspects that need to be considered and a lot of problems to be solved.

5.14 Conversion Tool Registry

Format conversion between differently structured files is a recurring issue and there are basically endless numbers of converters. There is no doubt that a registry of "major" converters would be a great help and save much time.

5.15 Analytics Component Registry

Big Data Analytics is a growing field, and since it needs to cope with large amounts of data it is very much related with efficient data management and faces many aspects of data interoperability. Yet it is not obvious how the field of BDA can be structured to make it easy for scientists to participate.

5.16 Repository API

In section 3.7 we explained that a split in data storage functionality occurred over time and stated that there are no standard ways of organizing the information in the logical layer, or for accessing it in a simple standard way. The Data Foundation and Terminology (DFT) Working Group in the RDA came up with a simple organizational model which is now being tested against practices in many other disciplines. Such a unified abstract data organization model could lead us to a standardized API for logical layer information.

5.17 Repository System

A large number of repository software packages (Fedora, D-Space, etc.) have been developed for different purposes over the years. They form a layer on top of storage systems making it possible to store, maintain and access the "logical information" associated with the data. RDA Repository System WG started to compare the systems against requirements. Help and harmonization is urgently required.

5.18 Certification & Trusted Repositories

One of the trends we are observing is that researchers not only use data from well-known colleagues, but also make use of data that they find somewhere on the web. New mechanisms need to be in place to certify repositories and thus help raising the trust level for users. Preliminary

[34] http://www.openarchives.org/OAI/openarchivesprotocol.html

[35] http://www.openarchives.org/rs/toc

[36] The rationale behind this statement is that we cannot afford to transfer large data sets for all the kinds of computation we want to carry out.

suggestions, such as those from DSA[37] and WDS[38], have been made that make it possible to assess the quality of repositories in delivering requested data and thus add to the basis of trust.

5.19 Training Modules

Besides of the technical components mentioned above we need to build-up a training and education infrastructure. Training modules such as tutorials, presentations, lectures, videos, webinars, etc. need to be composed and shared openly.

6. Organisational Approaches

In addition to the above technical components, it is necessary to think about the most appropriate organisational scaffolding to support their development, delivery and use. This scaffolding needs to exist at the Institutional, National/Regional and International levels.

6.1 Institutional

Infrastructure does not exist for its own sake – it is there to support the activities of researchers. Researchers occupy an interesting position within the scholarly ecosystem – they sit at the intersection of the discipline(s) to which they belong, and the institution at which they work. The former provides them with their professional identity, and the latter with many of the services on which they rely. For this reason, it is helpful to engage with institutions on infrastructure issues. This is for a number of reasons:

- they are the long-term custodians of the research data assets generated by their researchers
- they need to implement new infrastructure services in support of their researchers
- they can block externally-provided services if they wish to

6.2 National/Regional

Researchers are often funded at National (national funding councils/programs) or Regional (i.e. H2020) levels. The data requirements that funders build into funding calls have an impact on the kind of infrastructure that needs to be provided in order to meet those requirements. For this reason, it is highly desirable for infrastructure development to be coordinated with research funder requirements.

6.3 International

As indicated earlier, researchers belong to both institutions and to disciplines. All disciplines are international in nature, and so it is critical that there are coordinated international approaches to reducing barriers to data exchance and re-use. The RDA is one example of such a coordinated approach.

[37] Data Seal of Approval: http://datasealofapproval.org/en/
[38] World Data System: https://www.icsu-wds.org/

Appendix A: Roles and Tasks

A number of data-related roles have been mentioned in the foregoing discussions; often it is not obvious exactly how each of these roles are defined. The following table outlines what the different roles are currently seen as covering. Note that there is some overlap between the scope of various roles at present, however that may be reduced over time as clearer role definitions within the data field emerge. The second table describes the scope of some commonly used terms for various data-related tasks.

Role	Explanation
Data Professionals	all experts who deal in some form with (research) data
Data Practitioners	synonym for data professional
Data Scientists	all experts who carry out scientific processes on data (transformations, analytics, annotations, etc. based on scientific algorithms)
Data Managers	all experts who carry out typical management processes on data (transcoding, replication, preservation, curation, metadata creation & curation, PID assignment, etc.) and have a deep understanding about structures, metadata semantics, PID systems etc.
Data Custodians	this term is used by some to describe the group of experts that are responsible for the safe custody, transport, storage of the data and implementation of policy rules; this term quite often overlaps with the term data managers
Data Stewards	some research communities distinguish between data managers/custodians and data stewards – the former covers the daily tasks mentioned above while the latter encompasses more advanced tasks related to data content, its context and the definition of policy rules with the intention of keeping data re-usable persistently, which, for example, may require complex curation
Data Librarians	all experts who have a librarian background and often carry out curation and metadata related work; yet there is much overlap with data managers and data stewards
Data System Developers	all experts who create software that carries out some form of process on the data; experts that develop scientific algorithms are mostly called data scientists

Term	Explanation
Data Handling	this is a generic term describing all aspects of dealing with data in some form
Data Processing	this term is a generic term referring to all kinds of procedures being executed on data which can range from management to curation and analytics tasks
Data Management	this term refers to all tasks being carried out by data managers which are in general ignoring the content aspects of data.
Data Stewardship	the term refers to all tasks being carried out by data stewards which are related to data content, its context and policy aspects
Data Analytics	the term refers to all tasks that are directed towards extracting scientific knowledge from (combined) data

Apendix B: Elaborations on Components

In this appendix we elaborate on the components indicated in chapter 5 maintaining the same numbers. As indicated we see this chapter as a ***request for comments*** which will be discussed at the Data Fabric IG wiki[39].

1. PID System

A worldwide highly available and scalable PID system is needed that can be used immediately from the moment when a digital object (DO)[40] is created to uniquely reference that DO in scripts, practical policies and so forth, and thus guarantees reproducibility of the results of data intensive science at all steps[41]. At a minimum, next to the DO location, a checksum must be associated with each DO to allow identity and integrity checks to be performed at any moments in time. One of the initial RDA outputs is a recommendation for a conceptual model for PID types and an Application Programming Interface (API) specification for working with this model. We need to build on this work, and encourage PID systems to implement and improve the API. Any PID system that is used should support the PIT[42] application programming interface (API) independent of its underlying implementation to guarantee that all user software can function for all kinds of PID services. We must also remember that technology by itself is not sufficient to maintain a well-functioning PID system, so each component of the worldwide set of interoperable PID systems must have a stable organisation behind it which is committed to maintaining the system.

2. ID System for Actors

A system is needed that is able to uniquely identify users who are involved in the data creation process. In addition, the users' IDs should be included in provenance records (metadata) at all steps. ORCID, which is an organisation that provides unique and persistent digital identifiers for individual researchers, has received a lot of support in the publishing world. However, there are two things we need to take into consideration with the ORCID solution:

- There are many people who are active in the "fabric" of the world of digital data who are not *per se* authors of final data publications, and consequently they may not appear in the ORCID system of identifiers despite its openness.
- For authentication of anyone who is involved as actor different systems are being used with possibilities to assess identities from users.

From discussions within RDA, it has become obvious that we need one system to identify actors so that identities can be used in the federated solutions we are all building, and that this system must function in such a way that all of the potential actors (including users such as data managers) are part of the system.

3. Registry System for Trusted Repositories

A "Birds of a Feather" session on the topic of repository registries was organized at the RDA Fifth Plenary, where a decision was made to produce a position paper that details what a group within the RDA working on a Repository Registry could achieve. This paper is currently under discussion and

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

[40] It should be noted here that practices include already associating PIDs with code versions, sensor configuration documents and others.

[41] An agreement, such as the one between EPIC and DataCite that allows users to turn Handles with any prefix into Digital Object Identifiers (DOIs) which are Handles with prefix 10, at the moment of publication would, of course, be most welcome.

[42] https://www.rd-alliance.org/groups/pid-information-types-wg.html

it will soon be available via the DFIG list – for more details and final decision taking, we refer to the emerging documents of that RDA group.

Two concrete approaches were mentioned which emerged from different contexts:

- registries such as re3data[43] with particular human readable information emerged from the needs of funders, publishers and also users, to find "trusted repositories" that they could recommend
- large infrastructures such as EUDAT maintain a registry based on GOCDB[44] with particular machine readable information to do efficient operation

4. Metadata System

Metadata systems are an extremely complex topic since the term "metadata" can be used in different contexts to denote different things. Within the RDA, it was already identified that metadata about data occurs in different forms, such as attributes in PID records, state information records[45], access control lists, provenance records and the classical type of metadata to describe external and internal properties and contexts of digital objects. Since the various kinds of metadata descriptions fulfil the needs of different types of groups of data users, different groups within the RDA deal with the varieties of metadata. This document therefore refers to the work of these RDA groups.

Obviously we need a better classification of the different types of metadata and their uses and functions to come to a widely agreed specification of the different packages[46] that are required to provide an adequate and comprehensive metadata system.

5. Schema Registry System

For each digital object in the common data infrastructure, both the format and structure of the object is needed to allow users to parse the object and extract information from it. Therefore the RDA needs to ensure that there is a "system" allowing all users to register data schemas (which specify the format and structure of the components of a certain type of data). Quite a number of such schema registries are currently in use, such as "schema.org" or "ddc.org". In addition, some research communities maintain similar registries. These registries all emerged in a bottom-up fashion in response to urgent needs. As a result, the current situation cannot be described as a "system" for data schema registry because we do not have an integrated domain of such registries allowing everyone to easily find, upload, access and re-use schemas of all types.

When it comes to actually accessing a particular data item, there is usually an entry in the metadata associated with the item that points to the relevant data schema, and thus makes it possible to correctly interpret the contents of the data. This means that there is generally no problem with accessing the data itself, however we obviously need to maintain a reliable registry of the schemas that are being used – without the schema information, the data becomes useless.

6. Registry System for Semantic Categories, Vocabularies, etc.

When making use of data or metadata, a basis must be given to understand the meaning of all its elements. Therefore, shared semantic categories, vocabularies, taxonomies and so forth[47] that include concept definitions and relationships are very important. One of the basic principles guiding the domain of registered digital objects is that the semantics must be defined and therefore explicit

[43] http://www.re3data.org/

[44] https://rd-alliance.org/sites/default/files/EUDAT_Registry_overview_SAF.pdf

[45] Some people call this system metadata.

[46] The term "package" is being used in the RDA metadata groups to indicate re-usable components.

[47] In the following we will use the term "semantics" to refer to refer to all these entities. Some prefer the term "ontologies" which we would like to reserve for entities that combine concepts and their relationships.

as far as is possible. It is understood that this is probably the most complex requirement for achieving an interoperable common infrastructures since, in many scientific disciplines, making semantics explicit is a whole research topic in itself or may be barely possible.

Much excellent effort has already spent in the domain of "semantics" and whoever wants to become active needs to build on the knowledge and experience of the domain. But the discussions within RDA and in many communities show that a new pragmatic approach is required to overcome the hesitations with respect to using "semantics technologies" in daily practice. Since "semantics" to a large extent is discipline specific the question remains to be answered what cross-disciplinary initiatives such as RDA can offer in addition to all the efforts already been done. But there is no question that the current situation is not satisfactory.

7. Registry System for Data Types

RDA's Data Type Registry Working Group (DTR WG)[48] presented the concept of a Data Type Registry (DTR) which links data types of all sorts with the executable data processing functions that can be useful to work with a specific data type. Thus we can see the parallels that make DTRs complementary to schema registries and semantic ontologies. Data Types span the range from complex digital objects to simple categories that occur in digital objects. Functions that can be applied to data types include things such as performing transformations or mappings, and creating visualizations or interpretations.

The DTR WG understood that there could be different DTR instances, each serving a specific role in the data landscape (registering for example complex file types in biology or registering categories that appear in PID records to describe data properties), and therefore a registry of all DTRs conforming to the RDA DTR specifications is envisaged. Adhering to the same set of specifications would facilitate searching for specific types across instances.

Discussions with early adopters of the first DTR implementations show that some communities already think about using DTR specifications by automatic algorithms, i.e. a specific category appearing in specific file types could always be transformed in a specific way to make it comparable to categories from other contexts.

8. Registry System for Practical Policies

The RDA's Practical Policy (PP) WG is working on a list of best-practice practical policies (PPs) which are recognized as being the basis for self-documenting and reproducible data processing and managing. This work could be continued indefinitely since there will be more and more areas where automatic procedures will be applied in future. Infrastructure projects such as EUDAT realized that, in large federations, a registry is required that allows users to register practical policies and to share them with others. We need an initiative that will specify a registry system for practical policies which will then form the basis of open sharing, independent of current technologies such as iRODS[49].

We cannot yet imagine how machines will automatically look for the most suitable PPs, depending on the relevant tasks and frameworks, and then integrate those policies, but we must dare to look ahead. Broker technologies may help to find such solutions in a registry system where PPs are documented sufficiently well.

9. Prefabricated PP Modules

Currently, when speaking about practical policies, we most often discuss PPs that fulfil a certain function, such as replicating a Digital Object. However, we can identify generic components for such

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

[49] http://irods.org/

PPs that, for example, implement standard functions such as required by RDA-compliant data management. The following are several relevant examples that have been extracted from a drawing being used in the White Paper of the Data Fabric IG (see also 2.7):

- for all processing steps that create new digital objects, a PID record must be read, a new PID record must be created, and checks using PID information (such as the checksum) must be carried out and so forth,
- for such processing steps, existing metadata needs to be read and then, based on additional information describing the characteristics of the processing step, a new metadata object will be created which is extended by provenance information, and
- for all such steps when new digital objects are being created, a deposit into a trusted digital repository is required (for this part, ready-made components could be provided).

There will be increasingly more of these types of standard components that can be integrated into workflow chains so that scientific users developing code can focus on the arithmetic part of the code (for example, for performing simulations or analysis) and do not need to bother about the routine tasks related to data access and so forth, which are essentially a waste of time just re-inventing code for commonly performed tasks.

Such components could be registered in special practical policy registries or registry sections. It needs to be determined whether these kinds of components can be described by the same type of metadata as PPs in general.

10. Distributed Authentication System

Obviously we need to have a professional and secure system to allow us to authenticate users. Yet, even at national and regional levels, we do not have any such systems which meet the essential criteria and are also working smoothly. At present, we cannot envisage how this will work at global level either. The RDA has the Federated Identity Management (FIM) group which should take care of these aspects and interact with the major players such as Internet2, and GEANT/eduGain, amongst others. People have widely agreed on ORCID as a harmonized namespace for authors, yet this system cannot be used for authentication, since it does not cover all involved users nor (yet) includes any mechanisms to authenticate users (nor was that the intention when it was designed).

The FIM group in RDA needs to anticipate the user authentication challenges posed by the scientific automatic workflows that continuously create data in the labs, i.e. consider delegation of user credentials along those workflows.

11. Authorization Record Registry System

Up until now, distributed authentication (rather than authorization) has been the main focus in data management. However, increasingly often, we see scenarios where data is for example being replicated (which can be necessary for various reasons). To make all the instances of the data accessible, the data’s authorization records need to be accessible, for example, across different data repositories. Currently, we see the following two major types of solutions.

- A cloud service provider maintains a central database that includes all the types of logical information, such as access rights, that can be replicated as a whole as well. This solution will work in an isolated domain, such as that defined by a company, for example. This solution does not work easily in federations.
- In data federations, records of access rights need to be copied as well as the main part of the data to make the copies of the data accessible. However, this puts a burden on the mechanisms responsible for creating the copies, as they then need to synchronize all the copies properly since access rights are changing continuously.

- There are also complex issues to solve around if and how to apply access policies to data sets that are derivatives (rather than replicas) of other access protected data sets

It is therefore obvious that we urgently need to rethink our strategies. In the Finnish national data federation, a system was put in place that aggregates all access records that are offered by the member repositories so that this system can act as a central storage place for up-to-date access rights for all instances of any particular data item, wherever they are stored. Such a system must be highly secure, robust and scalable.

A new RDA group is needed that will work on this aspect anticipating DFIG requirements and come up with suitable suggestions for feasible solutions to this important issue.

12. OAI-PMH, ResourceSync, SRU/CQL

Metadata is an important part of any data management infrastructure. Special protocols exist for aggregating or harvesting metadata (OAI-PMH[50], ResourceSync[51]) as well as transferring metadata queries (SRU/CQL) to search engines.

OAI-PMH is already widely used by many metadata providers and service providers to harvest metadata compliant to different schemas. Some also use OAI-PMH to exchange other types of information.

The SRU/CQL standards[52] are also used to exchange information. SRU is a standard XML-based protocol for search queries, utilizing CQL - Contextual Query Language (a standard syntax for representing queries). Although SRU/CQL was initially developed for metadata querying, it is now also used for searching in non-metadata content.

13. Workflow Engine & Environment

When it comes to executing practical policies, and any other kind of process chains orchestrated by a user, there are several essential components: a flexible workflow engine (such as Taverna or Kepler), and an environment which allows users to easily deploy and execute practical policies (see 5.8) and scientific algorithms at centres offering sufficient computing power (and often storing the relevant data in the neighbourhood of HPC computers[53]).

The exact requirements for a suitable workflow engine and a sufficiently flexible solution for a whole environment have not yet been fully established, as there are many aspects that need to be considered and a lot of problems to be solved. Of the latter, the problem of rights (such as deployment, execution, and delegation) is one of the most difficult. The introduction of Virtual Machines may help make it possible to easily transfer applications between computers.

Another aspect that must be considered when looking at workflow engines is that there needs to be agreement on the languages that are used to formulate workflows. A number of languages have been defined.

It might also be beneficial explore iterative solutions which allow manual workflows to be gradually transitioned to automated engines and solutions, while introducing approaches and guidelines to enable workflow reproducibility even before full automation is achieved.

[50] http://www.openarchives.org/OAI/openarchivesprotocol.html
[51] http://www.openarchives.org/rs/toc
[52] http://www.loc.gov/standards/sru/
[53] The rationale behind this statement is that we cannot afford to transfer large data sets for all the kinds of computation we want to carry out.

14. Conversion Tool Registry

Format conversion between differently structured files is a recurring issue and there are basically endless numbers of converters. The main reason for this is that conversion between data formats is mostly dependent on the intended use of the converted data. Armies of students, researchers and software developers are busy creating such converters, which are often made just for one job. We certainly have a huge "waste" of energy happening in this area, however, it is yet not clear how we can improve the situation. Researchers creating a one-time ad-hoc script will not be willing to spend time describing it in such a way that others can easily find it and understand what it is doing. Conversion is also an essential part of the on-going curation process, and since not all data will get upgraded at the same time, a registry for finding the right upgrade converter will be most useful.

There is no doubt that a registry of "major" converters would be a great help and save much time. For popular formats such as for audio and video streams, many tools and libraries have already been created, thereby saving lots of effort. Whether the Data Type Registry could be used to register all kinds of conversion tools is not yet clear.

15. Analytics Component Registry

Big Data Analytics is a growing field and since it needs to cope with large amounts of data, it is very much related with efficient data management and faces many aspects of data interoperability, since in general, it will combine data of different types to relate for example phenomena with patterns found in data. An example can be the correlation between specific brain diseases and finding patterns causing these diseases in quite varying data streams such as brain-images of different types, genomics/proteomics data, patient behavioural data, etc.

Various groups (NIST, RDA Big Data Analytics IG, etc.) try to structure the field of Big Data Analytics to make it easier for newcomers to enter this field and carry out data intensive science and we know of large libraries of useful algorithms such as for supporting machine learning. Yet it is not obvious how this structuring can best be done.

16. Repository API

In section 3.7 we explained that a split in data storage functionality occurred over time, which resulted in a simplified physical storage layer with a very simple interface (in the case of cloud storage) and a "logical layer" where we put all the complex descriptive information associated with the basic data. We also said that there are no standard ways of organizing the information in the logical layer, or for accessing it in a simple standard way.

With respect to accessing the data elements in the physical storage within the domain of registered digital objects, we obviously just need the PID for the data and a PID system that resolves this PID into useful access path information. Once authorization has been granted and the optimal location of the data has been clarified, the PID is sufficient to access the stored bits and use them.

With respect to the logical layer, the Data Foundation and Terminology (DFT) Working Group in the RDA came up with a simple organisational model which is now being tested against practices in many other disciplines. Such a unified abstract data organisation model could lead us to a standardized API for logical layer information.

17. Repository System

A large number of repository software packages (Fedora, D-Space, etc.) have been developed for different purposes over the years. They form a layer on top of storage systems making it possible to

store, maintain and access the "logical information" associated with the data. All of these systems have limitations, and yet it is difficult for users to make judgements and proper decisions about which systems they should use. Too often it is the case that such systems are fraught with problems for the non-professionals. Such systems often use proprietary encoding, which makes it hard to extract the stored information. In all likelihood the software will not be maintained after a certain period of time, thus leaving users with a huge upgrade and data conversion problem. The chances are that the software will not be scalable when a new research paradigm is being added, and the software may not be able to handle larger amounts of data. More problem areas could be listed.
The RDA Repository System WG started working on this.

18. Certification & Trusted Repositories

One of the trends we are observing is that researchers not only use data from well-known colleagues but also make use of data that they find somewhere on the web. In the case of data from colleagues, they trust the quality of the data they receive as they trust their colleagues. However, the situation changes fundamentally when people use data found on the web, since they do not usually have any information about the quality of the work behind the data – here we have to consider not only the people who created the data, but also those who have been involved in all the other processes related to the data, such as managing and curating the data.

In addition, a further trend is that we find data creators increasingly often deposit their data in repositories for various reasons which have already been indicated (see 2.5). If the deposited metadata includes the names of the data creators, potential users can enquire about the quality of the creators' data work. If the data has a registered PID with associated fingerprint information, users could, for example, check whether the data is still the same. We can see that any trusted repository must have procedures in place that ensure that the data will be available over time, that document changes to data for example due to curation, and so forth.

This all indicates that users of trusted repositories must be able to rely on a number of mechanisms that are in place and being carried out systematically by the repository. Preliminary suggestions, such as those from DSA[54] and WDS[55], have been made that make it possible to assess the quality of repositories in delivering requested data and thus add to the basis of trust.

[54] Data Seal of Approval: http://datasealofapproval.org/en/
[55] World Data System: https://www.icsu-wds.org/