Trustworthy Repositories as Stable Pillars in the Evolving Eco-System of Data Infrastructures

24.7.2017

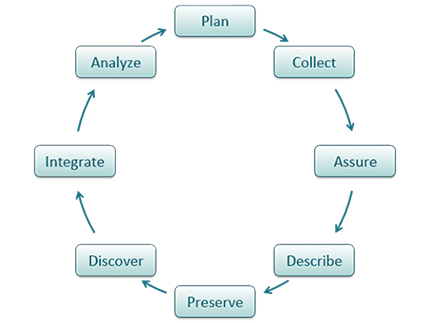
"Trustworthy Repositories" are now widely accepted as essential and stable pillars in the evolving eco-system of data infrastructures. Before elaborating on the term repository we should give a rough definition. We can cite re3data for example which states "*re3data.org has reached a milestone of identifying and listing 1,500 research data repositories, making it the largest and most comprehensive registry of data repositories available on the web. It has grown steadily since its launch four years ago to cover a wide range of disciplines from around the world*." Techopedia, for example, states in similar ways: "*data repository is a somewhat general term used to refer to a destination designated for data storage.*"

At first approximation we can state that a "repository" is a container where we can store bit sequences. But this is not sufficient since we want that repositories also inform stakeholders about which data they are storing, that repositories take care of data management and curation and that they give access to the bit sequences. Some see "repositories" just from a functional IT point of view, others also include the organisation taking responsibility. As we will see there are many different views on what "repositories" are, however they seem to all share the same basic notion of storing, managing and giving access to stored bit sequences and metadata about these bit sequences. In the following we will present a number of different views about "repositories".

It should be noted here that these bit sequences can contain data, software code, workflows, configurations, and many other types. In the following we will use the term "Digital Object" (DO) as defined in RDA's Data Foundation and Terminology Working Group to point to the fact that for many management operations it does not make a differences what type the DO has.

### Repositories in Data Lifecycles

Here we refer to the Lifecycle model as being described by the DataONE project as an example. According to DataONE the data life cycle provides a high level overview of the stages involved in successful management and preservation of data for use and reuse.

The DataONE data lifecycle has eight components:

- **Plan**: description of the data that will be compiled, and how the data will be managed and made accessible throughout its lifetime

- **Collect**: observations are made either by hand or with sensors or other instruments and the data are placed a into digital form

- **Assure**: the quality of the data are assured through checks and inspections

- **Describe**: data are accurately and thoroughly described using the appropriate metadata standards

- **Preserve**: data are submitted to an appropriate long-term archive (i.e. data center)

- **Discover**: potentially useful data are located and obtained, along with the relevant information about the data (metadata)

- **Integrate**: data from disparate sources are combined to form one homogeneous set of data that can be readily analyzed

- **Analyze**: data are analyzed

At least in most steps in this lifecycle repositories play a fundamental role, since they need to store the collected data and metadata, do need to take care of preservation, discovery and enable access.

### Digital Object Architecture

When Kahn & Wilensky wrote the first paper about the Digital Object Architecture in 1995 they shifted the focus from the internet as a set of network devices exchanging messages to a higher level of hosts that help discovering and delivering information in the form of digital objects. In this paper they introduced a few concepts amongst which is the term "repository": "*An originator, i.e., a user with digital material to be made available in the System, makes the material into a* ***digital object****. A digital object is a data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a handle (and, perhaps, other material). To get a handle, the user requests one from an authorized handle generator. A user may then deposit the digital object in one or more* ***repositories****, from which it may be made available to others (subject, of course, to the particular item's terms and conditions, etc.)"*. They also introduced the **repository access protocol** which would be used to access a digital object in a repository.

It is obvious that Kahn & Wilensky refer to a repository as an abstract entity that supports a specific protocol, independent of how the repository is being realised such as a cloud, a file system, a database etc. The software implementing the protocol would translate a request where a persistent identifier would be central into requests typical for the local implementation. These notions inspired various people to elaborate about object stores such as the Fedora community, the Cloud software designers and later the RDA community.

### Abstract IT View

In the RDA Data Foundation and Terminology Working Group a basic model was derived which is close to what Kahn & Wilensky have defined. In its core there is the definition of a DO. It states that a DO has a bit sequence stored at some repositories, assigned metadata (which is also a DO) and assigned a Persistent Identifier (PID) which uniquely points to a DO. It also defines that a "collection" is an aggregation of DOs and is itself a DO, i.e. it is assigned metadata and a PID. This definition of "collection" is recursive, i.e. it allows creating large and complex collections that consists of other collections etc. In an abstract IT term a "**repository**" is nothing else than a large collection, that consists of collections and perhaps individual DOs.

**DO**

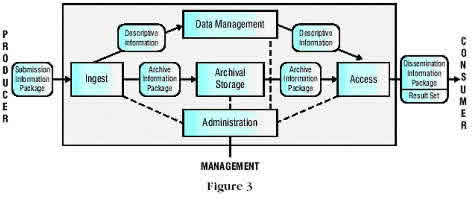
**collection**

**repository**

This view is in line with the view expressed in the Global Digital Object Cloud Concept being discussed in RDA's Data Fabric IG. It distinguishes 3 layers: 1) Users are just interested in abstract representations of Digital Objects (persistent identifiers and metadata), to form collections and ways to manipulate them with the help of workflows. It is the workflow machinery executing some function on the content that will also require access to the bit sequences of the digital objects. 2) Layer 2 is made up of **repositories** storing and giving access to data and metadata and by registries that aggregate different sorts of metadata information and giving services on them. 3) Layer 3 is made up of large storage systems to store the bit sequences of digital objects, powerful computers to carry out calculations and networks. The organisation of this layer is widely transparent to the user.

### OAIS Model and OAI

The very well-known OAIS model developed in the space data domain to address in particular data archiving has inspired many designers of repository systems. It speaks about data packages described by metadata as fundamental unit of archiving, but it does not say what a package typically should contain - this is left to communities and archives to be defined. It presents a slightly traditional model in so far as it does not support the view of flexible collections enabling to reuse digital objects in various contexts. It suggests that "packages" need to be retrieved if only a single digital object from that package is going to be reused.



The Open Archives Initiative (OAI) made very important contributions to the understanding of repositories in so far as they specified a model for metadata exchanging where a metadata provider (typically a repository hosting data) provides metadata for service providers (typically registries that support a search portal). This enabled them to design the OAI-PMH (protocol for metadata harvesting) which was used by many repositories to inform service providers of the data they are storing. Another relevant contribution was the specification of Object Reuse and Exchange (OAI-ORE) as standard for the description and exchange of aggregations of web resources.

### Scope of Repositories

Repositories can be organised according to different dimensions.

* **libraries** in institutions started to extend their holding to digital publications and also to data of limited size
* institutions built **institutional repositories** to capture all kind of publications and data of limited size
* data driven institutions extended their computing equipment by large storage systems and added a software to structure and maintain its data which increased in size and complexity rapidly; one could call these repositories that are often also set up close to computing facilities **Big Data Repositories**
* scientific organisations found that they need to maintain control of the data being created by their employees and built **organisational repositories**
* some countries developed national plans and set up (a network) of **national repositories**
* in some disciplines regionally or globally active community centers built **domain repositories**
* companies started to build ready-made software (Cloud software) to deal with vast amounts of data and now offer a) software for others to build a **cloud solutions** (software as a Service), b) **cloud storage capacity** (storage as a service), or c) **complete software solutions** as services (platform as a service)
* we should here mention "**workspaces**" of large high performance computers which store temporarily data to feed the HPC pipes and also the data being generated, we will not call these temporary storage containers repositories
* last not least we see a set of repositories that are focusing on long-term archiving of special data which are normally called **digital archives**

Often mixed forms can be seen and the scope of repositories is changing over time due to new requirements, insights and the use of new technology. All of these forms have their rational with different priorities in mind. Repositories being initiated by libraries traditionally look at data similar as to a book, i.e. as **static** data sets: a collection is stored and it can be extracted. Repositories that evolved in the scientific environment are seen as a rather **dynamic** entity that is part of complex scientific workflows where the stored collections change continuously, are being annotated etc. Of course, one can expect that the highest priority for domain repositories is its closeness to the research topic and the availability of special expertise, while the highest priority for organisational repositories can be found in the need to define clear service responsibilities for the organisation and its sustainability.

### Trustworthy and Registered Repositories

We call repositories "**trustworthy**" if they are certified according to standards such as World Data Systems / Data Seal of Approval who recently merged their rule set under the umbrella of RDA. It is important to know for various stakeholders such as depositors, publishers, politicians, and users whether a repository is trustworthiness which implies that certain quality criteria such as sustainability and suitability of practical policies are met. It should be common practice for every repository to regularly carry out a quality assessment.

We call repositories "**registered**" if they are listed in open registries such as re3data. A registration of repositories is increasingly important to inform stakeholders about their existence, their focus, and their offers in terms of available collections and services.

### Repositories in CLARIN

Already at an early stage in 2009 the CLARIN research infrastructure defined "centres[[1]](#footnote-1)" of different types. Repositories can be centres of type B if they offer data resources based on a regular assessment according to the DSA and WDS rule sets. CLARIN identified the crucial role of repositories as pillars of its infrastructure at an early phase an defined criteria for certain centre types to increase trust of users in the services offered.

Also other ESFRI research infrastructure initiatives realised the crucial role of repositories at an early stage. As another example we can refer to the CESSDA structure for example.

### FAIR Principles

The FAIR principles indirectly assume the existence of trustworthy digital repositories by stressing the accessibility of data and metadata via an open and obviously standardised protocol. Here is the corresponding FAIR principle:

***FAIR-A1****: (meta) data are retrievable by their identifier using a standardized communications protocol.*

### RDA

The Research Data Alliance has a number of activities that stress the importance of (trustworthy) repositories. The **Data Foundation and Terminology** WG and the **Practical Policy** WG came out with a few important definitions and statements:

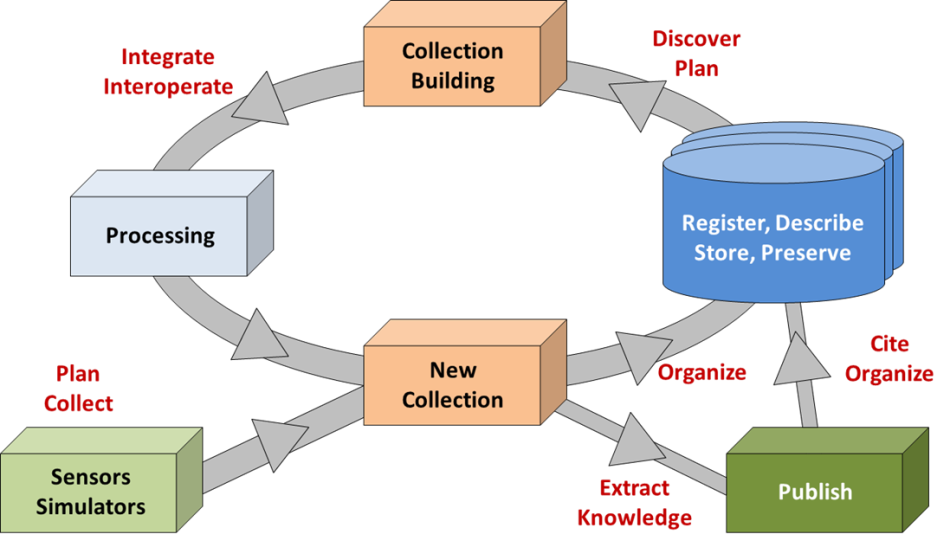
**RDA DFT1.9**: A digital repository is an infrastructure component that is able to store, manage and curate Digital Objects and return their bit-streams when a request is being issued.

**RDA DFT1.13**: A digital metadata repository is a digital repository that is able to store, manage and curate metadata.

**RDA DFT:** Digital repositories should have a repository software system that supports the data organisation as defined in DFT.

**RDA-DFT:** Data copies will reside in several trustworthy digital repositories. It is recommended to indicate in the PID record which repository is the original one and thus has authority about setting access permissions and original metadata descriptions.

**RDA-PP:** A trustworthy repository must specify auditable practical policies for its various tasks, turn them into executable procedures and workflows, and systematically apply them in all cases to document provenance of all its digital objects.

Further, the discussions in the **Data Fabric** IG where several infrastructural RDA groups joined forces referred also to the central role of repositories in their basic discussions which are captured in the following process diagram. The blue boxes include repositories and registries where repositories are storing, managing and preserving the collections that are being created.

**RDA**: Trustworthy repositories are digital repositories that untertake regularly quality assessments successfully such as Data Seal of Approval / World Data Systems.

**RDA:** Digital objects need to be stored in trustworthy digital repositories.

**RDA**: digital repositories should expose their characteristics and services in widely recognized schemas to enable service providers to create useful services for human and machine processing.

**RDA**: One of the services of a digital repository to be indicated in the schema is the metadata harvesting port supporting a standard protocol such as OAI-PMH.

**RDA:** The global Internet of Data is domain of registered digital objects, at registration with a digital repository a PID is associated and metadata are created.

A BoF at the 3rd RDA plenary in Dublin led to the fusion of two initiatives who wanted to offer a registry for repositories and this effort resulted in the creation of the **re3data** registry. This is an utterly successful registry for repositories from all kinds of disciplines and initiatives.

Further RDA groups discussing the issue of repositories:

**Research Data Repository Interoperability WG**

The Research Data Repository Interoperability Working Group is busy to establish standards for interoperability between different research data repository platforms. These standards may include (but are not limited to) a generic API and import/export formats.

**Storage Services Definition WG**

This WG is busy to create well-defined key terms that describe storage services which are part of the offering of repositories that will help users to compare different storage offers. Agreements should be made such that the offers are understandable; statements need to be made how lifecycle challenges are being treated, etc. Only a common language (vocabulary) will increase the quality of offers in the long term and will help to prevent misunderstandings.

**Domain Repositories Interest Group**

This group brings together experts from repositories that serve a variety of different scientific disciplines with services tailored to the needs of the corresponding discipline and acting as mediators between scientists and the rapid changes in information and data science. These repositories in general take care of data curation, dissemination, preservation, and institutional sustainability and thus have an important role.

**RDA/WDS Certification of Digital Repositories IG**

Repositories fulfill a crucial role in offering meaningful and usable resources (data) over time, i.e. researchers need to be able to rely on the quality of managing, curating and archiving the stored data and on the sustainability of the offers. One way to increase the level of trust in such offers is to do regular certifications. The group will build on the existing DSA/WDS certification standards and work out further recommendations. (see below)

**Repository Platforms for Research Data IG**

The group is basing its work on the matrix that gathers and analyses research data use cases in the context of repository platforms and relating these use cases with functional requirements for repository platforms. Potential gaps between needs and solutions need to be identified to help users to choose a platform (software) that may help them to setup a proper repository.

### Data Seal of Approval / World data System

A few initiatives such as DSA, WDS, ISO and DIN developed rule sets to assess data repositories. DSA and WDS turned out to find a brought acceptance in science and joined forces under the umbrella of RDA, i.e. they merged and harmonised their rule sets so that there is now one worldwide accepted and broadly applied standard. For reasons of simplicity we cite here those rules that can be applied to repositories.

**DSA4**: The data repository has an explicit mission in the area of digital archiving and promulgates it.

**DSA5:** The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

**DSA6:** The data repository applies documented processes and procedures for managing data storage.

**DSA7:** The data repository has a plan for long-term preservation of its digital assets.

**DSA8:** Archiving takes place according to explicit work flows across the data life cycle.

**DSA9**: The data repository assumes responsibility from the data producers for access and availability of the digital objects.

**DSA10:** The data repository enables the users to discover and use the data and refer to them in a persistent way.

**DSA11:** The data repository ensures the integrity of the digital objects and the metadata.

**DSA12**: The data repository ensures the authenticity of the digital objects and the metadata.

**DSA13:** The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

1. The term centres was used since in addition to repositories other types of centres were identified such as knowledge centres or centres offering registry services. [↑](#footnote-ref-1)