Repository Topic Group Report

Role of Repositories in Research Infrastructure Building

November, 2019

Editor:

Peter Wittenburg (GEDE)

Contributors:

We like to thank the GEDE colleagues, in particular Damien Boulanger (IAGOS), Ingemar Häggström (EISCAT), Margareta Hellström (ICOS), Venkata Satagopam (ELIXIR) and Christophe Blanchi (DONA) for their excellent contributions.

The Research Data Alliance is supported by the European Commission, the National Science Foundation and other U.S. agencies, and the Australian Government.

Abstract

It is widely agreed that trustworthy repositories are essential pillars in the evolving digital data domain. Therefore, many initiatives such as working and interest groups in the Research Data Alliance, ICSU's World Data Systems, the EOSC process, software developers and commercial services are looking at repositories from various angles, define properties and functions of them and offer services related to them. When characterising repositories one can take different views. From a computer science point of view a repository is nothing else than a complex collection, thus a digital object having a content, a persistent identifier and some metadata description. Other views (societal, functional, phasing, organisational) lead to different sets of requirements.

Repositories are abstract functional entities which are finally employing some technology to store and manage digital objects such as for example cloud systems, file systems or database systems. It is forgotten that much more important is how information belonging to digital objects is being organised. Bit sequences are stored differently than the various kinds of metadata that are necessary to access and process them. Until recently we were lacking unified models to bind all necessary information together. With the concept of digital objects this can be changed.

Repositories offering rich digital assets ar part of trust federations to exchange digital objects for various purposes. Until now every federation requests specific information to be provided. In the evolving data domain for open science this should be replaced by a generic approach where repositories provide all relevant information about them and where service providers make selective use of this information dependent on the services being offered.

About GEDE

The aim of the Group of European Data Experts in RDA (GEDE-RDA) is to promote, foster and drive the discussions and consensus relating to the creation of guidelines, core components and concrete data fabric configurations, based on a bottom-up process. To achieve these goals GEDE-RDA is composed of a group European data professionals appointed by invitation from various research and e-Infrastructures and European co-chairs of Research Data Alliance (RDA) Groups. GEDE-RDA will operate within the global RDA framework, thereby guaranteeing that discussions are openly communicated and publicly accessible to the global community of experts – RDA members. For more information, see the group's web pages at https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda.

Contents

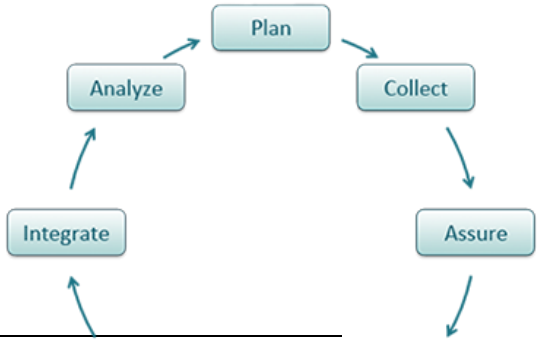
1. Nature and Role of Repositories

"Trustworthy Repositories" are now widely accepted as essential and stable pillars in the evolving eco-system of data infrastructures. The important role of repositories as they are for example maintained within the ESFRI initiatives was confirmed in the recent EOSC formation meeting [1]. Before elaborating on the term "repository" we should give a rough definition. The Turning FAIR into Reality Report [2] states: "*Repositories manage access to valuable data and metadata and offer services to support access and re-use.*" As first approximation we can state that a "repository" is an entity where bit sequences are being stored and maintained which includes proper management, stewardship and curation for the benefit of a user community. Bit sequences alone are not sufficient to enable reuse as is widely agreed. They need to be augmented by metadata descriptions that make assertions about their formal and content characteristics, their re-use conditions, their state and many others and they need to be identified by persistent identifiers. Compliant to the definition made by RDA DFT and various recent descriptions we call this complex entity a FAIR Digital Object[1] [2,3]. Therefore, we can state that a repository is an entity that takes care of Digital Objects[2]. RDA DFT [4] states: "*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*".

Repositories are service providers that allow users to deposit and access/re-use DOs, to see which DOs they are hosting, and to understand what their procedures, terms of services and guarantees are. The Trust Principles White Paper [5] states that in addition to supporting the FAIR Principles which define the properties of data and metadata, repositories need to be trustworthy which is about having transparent policies, organisational capabilities and skilled people to perform the tasks. As a recent paper [6] indicates, we are in a phase where new revolutionary infrastructures are being built and the investments required are huge as, for example, the plans for EOSC [7], the ESFRI projects and ERICs [8] and various national initiatives are demonstrating. Therefore, and to prevent a dark digital age [9] repositories will be key pillars of future data infrastructures and they need to offer their services for long time-periods to be accepted as being trustworthy.

It is not surprising that most large research infrastructure initiatives see repositories as core elements. This is in line with one of the key recommendations of the "Turning FAIR into Reality Report" which states: "*Research data should be made available by means of Trusted Digital Repositories*". It is widely agreed that data science needs to rely on the stable and trustworthy provisioning of Digital Objects and that this cannot be achieved by the individual researchers using their notebooks and servers. Repositories need to be entities that take responsibility for their holding and have the capability to guide Digital Objects through their whole life cycle.

According to DataONE [10], data is generated and processed in eight phases:
- **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 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

[1] The cont R) DOs can be da ata, software, semantic assertions, etc., i.e., the definition is agnostic about the content a DO's bit sequence is encoding.
[2] RDA DFT: A digital obj t sequence stored in trustworthy repositories, has a PID and has metadata of different types. Metadata descriptions are DOs and DOs can be aggregated to collections which are also DOs.

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

In most steps in this lifecycle data managers and stewards engaged in/by repositories play a fundamental role.

Also in RDA repositories play an important role as can be seen in appendix A. The activities in RDA with respect to repositories include topics such as

- the role of repositories in data fabrics, the core model for data organisation now called FAIR Digital Objects and ways to achieve interoperability
- the nature of digital collections being managed by repositories and the relationships to physical collections
- the nature of data repositories in research domains, managed by libraries and organised at national level
- the different software packages to allow setting up repositories
- requirements for archiving and preserving of data and
- certification of repositories to achieve trustworthiness which led to the development of the CoreTrustSeal rule set [11]

2. Views on and Organisational Forms of Repositories

Some see "repositories" just from a functional IT point of view, others also include organisational aspects. There are many different views on what digital "repositories" are, however they seem to all share the same basic notion of storing, managing, curating and giving access to stored bit sequences and metadata about these bit sequences over long time periods.

2.1 Cultural/Societal View

There is the overarching question what the essential pillars will be to manage our "digital cultural/societal" memory over centuries in similar ways as libraries and archives did it for physical objects. The preservation of digital objects follows completely different rules, since carrier the content are separated which is not the case for physical objects. Media for storing, formats for encoding content, and procedures and tools for curating digital objects are subjects of fast changes due to rapid innovation and aging. It is widely agreed that trustworthy digital repositories need to take over responsibility for these aspects and that trustworthiness needs to be documented by widely accepted certification processes.

2.2 Functional View

The tasks of repositories can roughly be classified by three overlapping areas coupled with specific skills: data scientists, data managers/stewards and computer/IT specialists. In the following we list a number of typical functions ranging from generating data products to taking care of the necessary hard/software basis. This diversity of the required functions and thus skills implies that often different organisational entities share a joint responsibility for "a given repository".

Task/Function	data scientists	data manager	IT specialist
data analytics			
data semantics			
data transformations			
metadata semantics			

data access			
user interfaces			
user interaction/support			
legal/ethical/licensing issues			
organisational embedding			
data formats/structures			
metadata formats/structures			
metadata exporting			
format transformations			
preservation			
software designs			
standards/protocols			
storage & server system			
software development			
data protection			
fast indexing			

2.3 Phasing View

As indicated above, data and their metadata is being created and processed in different phases which partly can be correlated with functions in the above table, i.e., different skills are required at different phases being another reason to split the tasks related to different phases between different partners.

2.4 Digital Object View

RDA DFT specifies that (FAIR) DOs can be aggregated to collections. In this sense a repository is nothing else than just another specific complex type of Digital Object, since it hosts a set of collections of DOs and thus itself is a DO which means that a repository needs to have a PID and to be described by metadata that characterises its essential properties. Even all its basic processes could be specified declaratively by a schema as part of its metadata enabling automatic auditing.

2.5 Organisational View

There are many different ways of organising a repository and splitting responsibility and functions. Yet we cannot speak about a clearly structured landscape and new organisational forms will emerge. Therefore, we can only refer to some typical forms which in reality often appear in mixed forms.

Project Related Repositories

Such repositories have been set up in very close collaboration between scientists, data managers and IT experts to meet the requirements of a specific project that in many cases then later serves a broader community. Good examples are the DOBES archive [12], the NOMAD archive [13] and the Protein DataBase [14]. These repositories are often devoted to specific types of material and take care of most of the above functions in one unit combining different skills. There are a number of attractive characteristics of these repositories such as closeness to the researchers and their needs, leading to excellent support for specific functional wishes. The disadvantages are often that these repositories are dependent on visionary individuals and lack persistency and that economy of scale factors do not apply.

Domain Related Repositories

Such repositories take over responsibility for a broader scientific domain and in many cases ESFRI infrastructure projects are based on such repositories. Often they are hosted by larger institutions taking responsibility for some primary functions and delegating others to collaborating organisational entities. These repositories in general need to provide a broader portfolio of services and thus cannot

address specific researcher interests. One of their goals is to have long-term funding applying economy of scale business models.
There is a wide spectrum of organisational solutions that can be found dependent on the disciplines these repositories are servicing.

Organisational Repositories

Such repositories are being set up by scientific organisations or governments to address the needs of their researchers, to offer ready-to-use services and to protect intellectual property. They can be established at different levels from research organisations (universities, research institutions) up to national or regional governments. Their establishment is mostly the result of long-term planning coupled with the intention to have continuous services for their researchers. Various models are being applied from a centralised service provider addressing different disciplines and offering a broad range of functions up to models where a few specialised institutions collaborate each taking over a special role dependent on the specific skills they can provide.
The frequently used term "Institutional Repositories" describes one type of organisational repositories. Large computer centres, often having key roles within their organisations, are also mutating to take over responsibility for the data they are storing. Traditionally they are not trained on aspects of persistency and services beyond pure storage provisioning, but there is an increased understanding that they need to adapt.

Generic Repositories

These repositories offer their services basically to all research disciplines such as ZENODO [15] and B2SHARE [16]. Any researcher can drop their documents and in most cases small data objects (long tail of data). Here on can also mention commercial services such as Dryad [17] and Figshare [18] where small companies emerged from science. Also cloud offers from big companies such as Google, Amazon, Microsoft, etc. can be classified as generic repositories. In most cases these services are restricted to those tasks that can be carried out by data managers and IT experts. Their services are based on a business model where costs need to be refunded in some way. Therefore, long-term persistency cannot be guaranteed. In the case of the big companies one aspect of the business model is to get control about as much data as possible to be able to offer data products that can be sold and to create long-term dependencies.

Libraries/Archives/Museums

These traditional institutions often get the task to move their holding into the digital domain and to extend their scope to digital data without getting the funds needed to develop the required skills and to take care of a broad spectrum of functions. Thus, collaborative models are often implemented. These institutions also have the challenge to adapt to the completely different interactive service model which is typical for digital repositories servicing data intensive research. On the other hand these institutions have a deep understanding about persistency and documentation aspects.

3. Data Organisation in Repositories

According to the main tasks of repositories and being compliant with the FAIR principles, repositories need to carefully design the way they organise their digital data and other digital information. The FAIR principles are explicit in requesting the assignment of PIDs and metadata to all digital entities that are being stored and managed. In accordance with the RDA DFT Core Model [4] we call this bundle a FAIR Digital Object. A FAIR DO consists of a bit sequence stored in some repositories, a PID and different types of metadata. Metadata and aggregations of DOs, called collections, are FAIR DOs as well, i.e., PIDs and metadata need to be assigned to them. Accessing and reusing such FAIR DOs means supporting access to all these different entities in stable ways. As Schultes and Wittenburg [3] worked out FAIR DOs are one way to implement the FAIR principles, since they offer a way to bind all these entities together by using persistent identifiers.

Repositories can use different ways to store the bit sequences such as file, cloud and/or database (relational, noSQL, etc.) systems. In file systems the bit sequences of a DO are typically stored in a particular file where the file hierarchy and file naming offer limited opportunities to organise the data, however, PIDs are not provided and there is no inherent way to link to the different types of metadata. In Cloud systems local digital objects are managed in so far as every bit sequence is linked via hash code mechanisms with an entry in a local database which typically stores system type of metadata. In databases mostly metadata of all sorts are contained in some tables with one entry pointing to the bit sequences stored somewhere. In some no-SQL databases the intention is to also store large bit sequences in the database. It is the logical structure of the database that gives information about the database's content and structure.

For all these storage concepts there is no defined mechanism to link all informational entities together as specified by the FAIR DO concept which leads to the fragmentation we are faced with, i.e., globally resolvable PIDs and the different types of metadata are spread dependent on individual solutions. Databases mostly include much of the metadata and partly the data and with the help of queries one can access the informational entities. However, in general there are no clearly identifiable queries that can be executed to extract the intended information. In addition, schemas define the structure of the content in files and cloud objects, but it is not specified where to find them in many cases of scientific data. Even worse is the findability and accessibility of vocabularies that enable interpretation and re-use.

The FAIR DO organisation concept of data clearly specifies how the stable linkage between all relevant informational entities can be achieved and repositories should adopt this model. The PID record and of course the metadata description, if all of this is made machine actionable, can be used to access the relevant links. Repositories should make use of attributes defined in data type registries [19] and create profiles for its sub-collections informing users how they can use the PID record and thus where to find the required information about the bit-sequences. All relevant information including schemas and vocabularies need to be made explicit by registering them in open registries and two protocols should therefore be supported by each repository: a) the protocol to interact with the PID resolver that will offer access to the PID record and b) the DO Interface protocol [20] to interact with digital objects independent whether they are stored in files, cloud objects, databases or any other technology which will emerge in the coming years.

4. Requirements for Repositories

The FAIR principles are directed towards data being compliant with a number of criteria including statements on PIDs, metadata, semantics, etc. The FAIR principles do not speak about repositories; however, since we believe that repositories will be essential pillars to come to a sustainable world of data we can derive requirements for repositories from the FAIR principles. On the other hand there is CoreTrustSeal as the result of broad agreement finding after years of learning and it provides a widely used set of certification rules that is directed to repositories and should be applied. We will mainly use these two aspects to describe the requirements for trustworthy repositories.

Core Trust Seal Rules

The CTS Rules were developed by combining the rule sets from Data Seal of Approval [21] and World Data System [22] under the umbrella of RDA. Both rule sets had different origins in different communities and due to this merger we can now speak about a world-wide accepted standard, i.e., repositories should aim at getting the CTS seal which currently is based on a self-assessment procedure and which will probably evolve towards automatic tests to cope with the many requests that can be expected in the future. Due to their relevance we list the rules here.

R0. Please provide context for your repository (repository type, designated community, level of curation performed, outsource partners)
R1. The repository has an explicit mission to provide access to and preserve data in its domain.
R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.
R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).
R7. The repository guarantees the integrity and authenticity of the data.
R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
R9. The repository applies documented processes and procedures in managing archival storage of the data.
R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.
R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
R12. Archiving takes place according to defined workflows from ingest to dissemination.
R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.
R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.
R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Additional Requirements

In addition to these requirements, we can make statements about FAIR supporting requirements for repositories. Some can be directly derived from the FAIR principles:

- F: Repositories need to ensure that its digital objects are assigned a PID and are described by "rich" metadata which also include the PID, and that metadata can be harvested.
- A: Repositories need to ensure that the PID can be used to retrieve the DOs bit sequence using standard protocols which are open, free and universal, that authentication and authorisation is being checked and that metadata exists even if the bit sequence is not accessible anymore.
- I: Repositories need to ensure that well-known languages are used to represent (structure and) semantics, that the vocabularies used in the DOs are FAIR and that relevant relationships are included in an explicit way.
- R: Repositories need to ensure that DOs are being described by accurate attributes that include clear usage licenses and provenance descriptions, and that domain-relevant community standards are being used.

Various actions are currently being taken to establish FAIR Maturity Indicators [23] and to enable automatic assessments [24].

Some repositories already include for example software libraries containing many routines to check format compliance during upload. We can assume that extensive type checking libraries will be applied more frequently by repositories to increase interoperability and re-use.

To come to a stable and accessible data landscape the requirements for quality assessment and to increase findability, accessibility, interoperability and re-usability there will be an increasing pressure on repositories as key pillars to meet the mentioned requirements and the ESFRI initiatives welcome this development. Steps towards automated assessments will be necessary to cope with the demand. These will also require repositories to take measures to formalise the descriptions of their characteristics and processes. There is no doubt that these requirements will impose an increasing load on repositories, i.e., efforts towards efficiency are urgently required. In addition to the automation of checks, there is the need to reduce the overhead where possible. There is overlap between the CTS rule set and the FAIR Maturity Indicator set which suggests efforts to synchronise these sets. The vague formulation of some requirements (the FAIR principles for example speak about "sufficiently rich metadata", CTS speaks about "appropriate metadata") indicates that there will be limitations for automatic assessment.

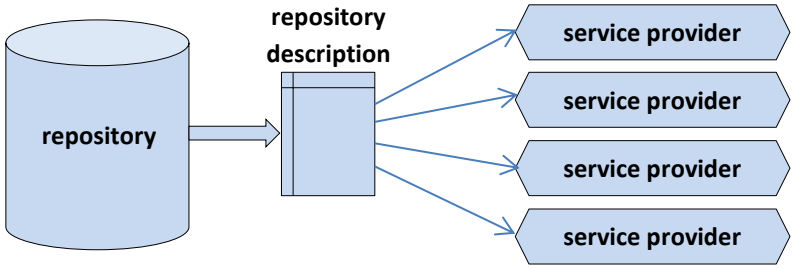
5. Federating Repositories

Repositories will increasingly often be members of trust federations of different sort for a variety of purposes some of which are mentioned in the following list:

- Repositories may want to exchange their digital objects to have several copies and thus to increase the chance of data survival.
- Repositories may want to exchange data to improve access efficiency and for load balancing reasons.
- Repositories may want to exchange the metadata of the hosted digital objects to improve visibility and findability.
- Repositories want to exchange their characteristics and services to inform users about their missions, properties and processes.
- Repositories may want to exchange their access rights records and license rules to facilitate access in distributed scenarios.
- Repositories already make use of distributed authentication mechanisms.

For all these types of federations different agreements and security measures are required. For the exchange of metadata records it was common practice to adhere to the OAI-PMH protocol [25] to allow any service provider to harvest the records and build search portals. For a service such as re3data [26], which is very successful, repository descriptions have to be submitted. re3data 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*." For federations that exchange data for redundancy purposes, which can include sensitive data, such as in EUDAT's B2-services even more detailed and precise information is required.

The current situation is not efficient, since for every service provider different, but partly overlapping sets of attributes are required in idiosyncratic formats and with OAI-PMH we simply have one exchange protocol that is dedicated to typical Dublin-Core type of metadata. We lack generic mechanisms which would increase efficiency for repositories. With ResourceSync [27] an attempt has been made to define a more generic way of exchanging different types of information, yet it is not obvious whether the community will accept this suggestion. With respect to informing about the characteristics and processes of repositories each federation uses different ways which



makes it an urgent task for RDA, for example, to specify a schema which includes all relevant information allowing different service providers to select those attributes which they need for their specific services. The diagram shows this efficient scenario where a repository exports its characteristics and processes enabling different service providers to extract exactly the information that is needed to run their services. This description could include information about the mission of the repository (for re3data type of services), ports for certain services (for metadata harvesting), formal specifications of processes (for automatic CTS and FAIRmetrics checks), PKI information (for secure data exchange), human readable information about hotlines (for emergency cases) and many more.

6. Conclusions

Trustworthy repositories as care takers of our cultural, societal and scientific digital memory are key pillars of the emerging eco-system of research/data infrastructures and they need to guide data processing across many steps of their life-cycle. This important role has been recognised in most of the distributed ESFRI and other large research infrastructure initiatives. As a consequence several working and interest groups in RDA, driven by data practitioners from these research infrastructures, are focusing on different aspects of repositories. One of the results of these efforts in RDA is the wide agreement to focus about data organisation aspects and to define the term FAIR Digital Object to overcome inefficiencies. Any large initiative such as EOSC at European level or the corresponding national initiatives now evolving need to build on the knowledge about repositories. However, these repositories need to adapt their processes due to new requirements.

Repositories need to address a variety of functions to fulfil their crucial role. Dependent on available skills and tasks different forms of mappings on organisational structures will be used. Some combine different skills to address a wide variety of functions within one team and institution, others share the work load between different institutions employing experts with different skills. A split that can often be found is the one between care takers of bit sequence management and those that have knowledge about content and thus do management/stewardship of metadata aspects. Yet, we cannot speak about well-established structures and aspects such as skills, but also economy-of-scale factors are critical, since maintaining trustworthy repositories requires considerable funding support.

To improve quality and trustworthiness the repositories need to fulfil requirements. They are mainly defined by the CoreTrustSeal rule set, which was developed in RDA and can now be seen as a world-wide standard, and the evolving FAIR Maturity Indicators, which are now being discussed amongst different stakeholders in an RDA group []. But also the need to join a variety of trust federations for different purposes ranging from metadata exchange up to the exchange of sensitive data poses new requirements for repositories. Therefore, it will be of great importance to come to efficient solutions:

- For improving the data organisation and implementing FAIR the concept of FAIR Digital Objects is crucial.
- For the assessment of quality and trustworthiness it will be important to establish widely agreed FAIR Maturity Indicators, to look for ways to automatize the auditing and to reduce the overlap between thee different rule sets.
- For joining federations it will be essential to come to unified protocols such as the DO Interface Protocol which will offer a unified exchange of Digital Objects independent of the way data is organised in the different repositories. But more effort is required to define additional standards.

Funders such as research organisations and governments need to see the crucial role of repositories when designing the research infrastructure eco-system and need to understand that sustainable funding support will be necessary to maintain our cultural/societal/scientific digital memory.

References

[1] https://www.esfri.eu/esfri-events/esfri-ris-eosc-liaison-workshop
[2] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf
[3] http://doi.org/10.23728/b2share.166a074bff614a31b05e9df5bfd9809d
[4] http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
[5] https://www.rd-alliance.org/trust-principles-trustworthy-data-repositories-%E2%80%93-update
[6] http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0
[7] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[8] https://www.esfri.eu/
[9] https://www.bbc.com/news/science-environment-31450389
[10] https://www.dataone.org/
[11] https://www.coretrustseal.org/
[12] http://dobes.mpi.nl/
[13] http://www.nomad-repository.eu/
[14] https://www.rcsb.org/
[15] https://zenodo.org/
[16] https://b2share.eudat.eu/
[17] https://datadryad.org/stash
[18] https://figshare.com/
[19] https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries
[20] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
[21] https://www.coretrustseal.org/about/history/data-seal-of-approval/
[22] https://www.icsu-wds.org/
[23] https://www.rd-alliance.org/wg-fair-data-maturity-model-rda-13th-plenary-meeting
[24] https://www.nature.com/articles/sdata2018118
[25] https://www.openarchives.org/pmh/
[26] https://www.re3data.org/
[27] http://www.openarchives.org/rs/toc

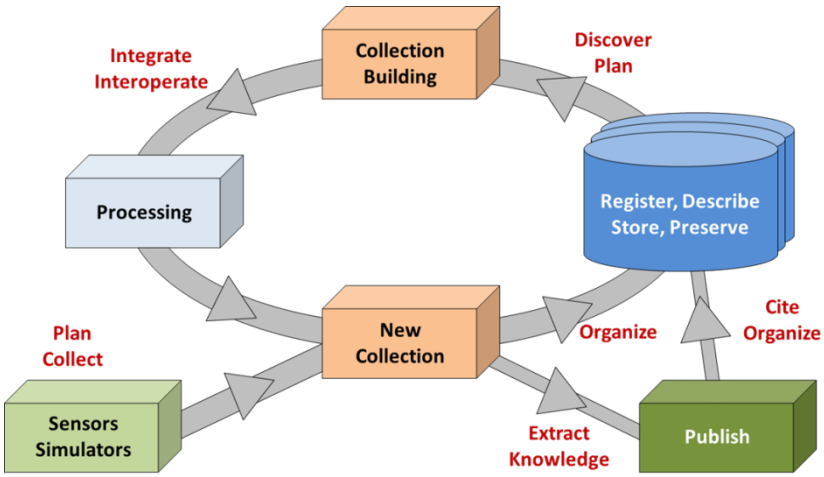
Appendix A

Definitions of Repositories in 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.

Appendix B

RDA Groups that deal with repositories in some way		
I may have missed some and also my comments may be not fully correct. Please check yourself.		
RDA Group	Chairs	Comments
Research Data Collections WG	Bridget Almas, Frederik Baumgardt, Tobias Weigel, Thomas Zastrow	Developed an API for Research Collections. In an abstract sense a repository is a collection, i.e. the API can inspire discussions on APIs for repositories.
Research Data Repository Interoperability WG	Thomas Jejkal, David Wilcox	Looking at requirements for repository interoperability, i.e. for ways to easily exchange data for example.
Archives and Records Professionals for Research Data IG	Rebecca Grant, Laura Molloy, Sarah Ramdeen	Come from the world of archives and bring in their experience. Archives are repositories - traditionally with physical collections
Data Fabric IG	Jianhui Li, Tobias Weigel	Discussed a lot about the roles of repositories in the fabrics and their requirements. A spin-off group is implementing an API for Digital Objects the content of which is stored in repositories, etc.
Data Foundations and Terminology IG	Gary Berg-Cross, Raphael Ritz	describing the DO Core Model which is now called FAIR Digital Objects
Domain Repositories Interest Group	Kerstin Lehnert, Peter Doorn	Developed a core model for Digital Objects where repositories are a key element. Also suggested some terminology.
Libraries for Research Data IG	Birgit Schmidt, Andi Ogier, Marta Teperek, Juliane Schneider	Libraries traditionally have much experience in storing results of science. They are players in data as well, mostly referring to long-tail data.
National Data Services IG	Adrian Burton, Mark Leggott, Christine Kirkpatrick	National Services often have repository functions as key element.
Physical Samples and Collections in the Research Data Ecosystem IG	Kerstin Lehnert, Lesley Wyborn, Simon Cox, Jens Klump	Discuss how to integrate physical entities stored in museums and archives into the digital domain.
Preservation e-Infrastructure IG	David Giaretta, Jamie Shiers, Ruth Duerr	were working on the preservation aspects, something some repositories have to take care of. not sure whether the group is active.
Preservation Tools, Techniques, and Policies	Ruth Duerr, Michael Hildreth, Peter Cornwell	Discuss methods for preservation which is of interest for repositoris
RDA/WDS Certification of Digital Repositories IG	Michael Diepenbroek, Ingrid Dillo, Mustapha Mokrane	Discuss methods for the certification of Repositories, result is the CoreTrustSeal
Repository Platforms for Research Data IG	Ralph Müller-Pfefferkorn, Robert Downs, João Rocha da Silva, Mike Jones	Discuss software platforms that could be used by repositories such as D-SPACE, FEDORA, etc.