About Digital Repositories

**draft version**

Peter, ?

There are different interpretations of the term "repositories" dependent on the context in which experts are working. This note is meant to present some views about repositories, derive some requirements extending the Core Trust Seal towards FAIRness of data and the Digital Object model, and present a definition.

# Views

A few views about repositories will be discussed first.

## Cultural/Societal View

There is the overarching question what the essential pillars will be to manage our "cultural/societal" memory over decades in similar ways as the libraries and archives did and still do it for physical objects. The domain of digital objects has completely different rules and the media for storing (storage media) and for doing curation of (curating tools) digital objects are subjects of fast changes due to rapid aging and innovation. Nevertheless, we have the big task to preserve our cultural heritage of which scientific objects are an essential part for the future. It is widely agreed that trustworthy digital repositories take over a key role in this respect and that trustworthiness needs to be documented by widely accepted certification processes as they currently emerge such as the CoreTrustSeal.

## Functional View

The tasks of repositories can be classified in three overlapping areas coupled with specific skills: data scientists, data manager/stewards and computer science/IT. In current practice, one will not find a clear mapping of functions to educational backgrounds, since some data scientists have become excellent managers or software developer etc.

|  |  |  |  |
| --- | --- | --- | --- |
| Task/Function | data scientists | data manager | it specialist |
| data analytics | X |  |  |
| data semantics | X |  |  |
| semantic mapping/transformation | X |  |  |
| metadata semantics | X | X |  |
| data access | X | X |  |
| user interfaces | X | X |  |
| user interaction/support | X | X |  |
| legal/ethical/licensing issues | X | X |  |
| organisational embedding | X | X |  |
| data formats/structures |  | X | X |
| metadata formats/structures |  | X | X |
| metadata exporting |  | X | X |
| format transformations |  | X | X |
| preservation |  | X | X |
| software designs |  | X | X |
| standards/protocols |  | X | X |
| storage system |  |  | X |
| server system |  |  | X |
| network system |  |  | X |
| software development |  |  | X |
| data protection |  |  | X |
| fast indexing |  |  | X |

## Phasing View

Many of these tasks have to be carried out in different phases. For data/metadata semantics we can for example indicate a number of phases such as designing a semantic domain to be covered including the definition of concepts, creating a formal framework for the semantics, creating instances according to the framework, maintaining the definitions and the framework, curating the instances, transforming instances, deleting instances etc.

In all these phases some of the above mentioned skills are required.

## Digital Object View

The notion of Digital Object is increasingly often accepted and a repository itself can be seen as a DO, i.e. it has a (complex) collection of data and other content, it has a PID and it is being described by various types of metadata. RDA DFT has defined a digital collection in a recursive way allowing creating complex collections and the holding of a repository can be seen as representing such a complex collection. This way of looking at repositories will facilitate automation.

## Organisational View

We can only describe a few typical cases[[1]](#footnote-1), in reality there are many types of mixed forms. We only refer here to those forms that deal with scientific data and not all those forms that deal with metadata of different types knowing that there is overlap. There is still much dynamics with respect to organisational forms and one can expect that there will be changes in future in the sense of distinguishing roles more clearly enabling economy of scale factors.

### Project Related Repositories

These repositories have been designed and setup in very close collaboration between scientists, managers/curators and IT experts to meet the requirements of a project that later often takes responsibility for a broader community. These repositories often take care of most of the above mentioned functions. Partly these repositories emerge as close collaborations between a few departments where the partners take over different roles, but to the users they appear as a single unit. Typically these repositories have people with different expertise (scientists, archivists/curators, IT experts) and it is generally known that these experts can closely interact due to overlapping knowledge and therefor a deep understanding. There are a number of positive characteristics of such repositories such as closeness to the researchers and therefore excellent services also with respect to special software addressing direct scientific wishes. The disadvantage may be that economy of scale factors are difficult to realise.

### Domain Related Repositories

These repositories take over responsibility for a broader scientific domain and are usually larger units also covering various tasks as mentioned above. Therefore, they also have experts with different skills in the team, but also do outsourcing of some of the functions. They still have a close relationship with the scientific users although they need to cover a broader list of wishes. In general they can realise economy of scale effects better than the project related repositories.

### Organisation Repositories

Organisation repositories are being setup by scientific organisations or governments to address the needs to their researchers and to take care that they have excellent services to foster the scientific work. There are different models of how organisations have set up such repositories. Here we only can sketch a few typical models as examples.

* The Max Planck Society established a Computing & Data Facility to help about 85 institutes working in different scientific domains in their data management and processing needs. In addition to offering storage and computing facilities, it is requested to closely collaborate with different scientists that work on demanding data-intensive science projects. The facility mostly takes care of computer science and IT aspects which in several cases includes methodology support such as in machine learning.
* The national computing and data centre in Finland (CSC) has the task to offer storage and compute capacity to Finnish researchers. But in addition it is partner in large scientific infrastructure projects and engages scientists from the corresponding domains with the intention to have a close relationship with the scientists.
* In the Netherlands there are a few institutions that share the job. Surf-SARA is meant to serve data and compute services. The eScience Centre carries out joint projects in data-intensive science in close collaboration between its computer science experts and the involved researchers. An institution such as DANS is meant to take care of data archiving, metadata aspects, certification aspects and some curation tasks. In addition, it has an extended role in social sciences and the humanities.
* In Australia one has basically two institutions that share the tasks at national level. An infrastructural centre offering storage and compute facilities and doing data intensive projects together with scientists. The other centre is focussing on data valuation, metadata harvesting and metadata enhancement/curation amongst others. Recently trends could be seen to merge these two services.

### Generic Repositories

Generic repositories are those that offer their services basically to all scientific disciplines based on a business model. These can be services such as ZENODO or B2SHARE that allow people to drop their documents and mainly small data objects (long tail of data). But one can also mention commercial offers such as Dryad and FigShare which are small companies that emerged from science, or cloud offers from the big companies such as Amazon, Google, Microsoft etc. Often, these offers are restricted to those tasks that can be carried out by managers and IT experts, i.e. in general there is no discipline knowledge available.

### Libraries/Archives/Museums

These traditional institutions get often the task to move their holding into the digital domain and extend their scope to digital data. Due to limited budgets they often cannot extend their staff with all kinds of different expertise and thus often look for collaborations with other institutions, organisations or companies for software and storage solutions. Of course they have deep discipline knowledge.

### Summary

Summarising we can state that there are three areas of knowledge (data science, data management, computer science/IT) that are needed to set up a fully functioning repository. In some way all three areas need to be covered whatever organisational form finally is being chosen. It can be assumed that the landscape of repositories will change over time. Economy of scale factors and ensuring organisational competitiveness will remain very important drivers.

# 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 taken into account.

## Core Trust Seal Rules

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 compliant or FAIR supporting requirements for repositories, i.e. trustworthy and FAIR 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.

As can be seen there is some overlap between FAIR principles and CTS, but FAIR principles go much more about the details of the data/metadata. While FAIR principles for example speak about "sufficiently rich metadata, CTS speaks about "appropriate metadata" to support understanding and use of data. These requirements are vaguely formulated and equivalent. We can observe a clear trend towards convergence.

An additional requirement emerges from some RDA groups active in the Data Fabric discussions (DFT, PIT, Kernel Information, DTR, MD, Dynamic Citation) that show a way to implement the FAIR principles and to fulfil some of the CTS requirements.

* Repositories need to ensure that they organise their stored entities in form of Digital Objects which includes a binding between its elements (PID, types of metadata, bit sequence).
* Repositories need to support a wide range of different metadata types[[2]](#footnote-2) (descriptive, system, scientific, provenance, access permissions, transactions, etc.)
* Repositories need to use a PID system that guarantees global and persistent resolution of PIDs that are amended by typed attributes that contain for example the binding parameters and other state metadata for example enabling authenticity and integrity checks.
* PID attribute types need to be defined and registered in open Data Type Registries.
* Repositories need to check the stability of references in case of dynamic data.

All these criteria emerging from FAIR and RDA work would require the repository to take care of the organisation of data, of the adherence to standards, of the explicitness of the used structures and concepts, etc. which would require an extension of the task list. Some repositories already include for example software libraries with routines to check format compliance during ingest. To come to a stable and accessible data landscape we assume that there will be a development of certification standards such as CTS so that they will include the corresponding checks.

1. The categories are similar to the ones found in CTS as examples. [↑](#footnote-ref-1)
2. The metadata types are not yet clearly specified. [↑](#footnote-ref-2)