# Summary

This report presents an analysis of the first data fabric recommendation [2]. The first section describes provides a summary of the Global Digital Object Cloud proposal. Then the second section presents a structural analysis of the elements of the GDOC Virtual Layer.

# Global Digital Object Cloud proposal of the RDA DF IG

The Global Digital Object Cloud proposal[2] consists of a common conceptual framework and a description of a set of tools and components aimed at revolutionising the management of scientific data.

The conceptual framework is designed to support a set of data management principles, and the implementation of the Global Digital Object Cloud (GDOC) model. The set of tools and components are described as the components of the virtual layer that provides access to the objects of the GDOC.

## GDOC principles

The set of GDOC principles are designed to align with the FAIR data principles[5], extending the requirements of data management to the objects that enable it.

The GDOC model principles are:

* Creation of non-proprietary and highly efficient data management and access infrastructure
* Globally available and stable PIDs are anchors for all activities.
* Allow multiple lower-level data storage and management solutions to plug into the framework along with useful data processing applications.
* The primary layer exposed to users is based on registered Digital Objects accessed through stable PIDs and metadata infrastructure, except when users want to start calculations on the bit sequences themselves.
* Set up a middleware layer to transparently access and operate on the bit sequences and for example sort out which copy should best be taken or whether the data should go to the algorithms or vice versa.

## GDOC Virtual Layer

The set of tools and components described consists of 10 registries and a repository type which constitute the virtual layer of the GDOC. The descriptions are not explicit about the characteristics of each component. In fact, it does not even provide a definition for each of the components included. However, most of the definitions can be found on the TeDT [3], or derived from context. However, a clear definition of each should be provided, because some terms are closely related and could be confused. Additionally, root concepts, for instance repository and registry, should also be clearly defined to facilitate the definition of components which are instances of them.

The components of the virtual layer are derived from a set of 10 requirements:

1. a network of **trustworthy repositories** (T-REP) that are available for every researcher to register, store and manage data, that have a clear interface to access DOs and that are certified to guarantee a certain quality of service,
2. a **trustworthy registry of such T-REPs** that is human and machine readable to enable efficient and goal-driven access,
3. a **system to register and resolve PIDs** available for every researcher that we can rely on and that offers adequate security mechanisms,
4. a **system to register types of DO**s allowing machines to relate actions with types as the basis of automating data processing,
5. a **definition of a set of core types** that are being used to describe the state of DOs enabling machine action,
6. a **system to register metadata schemas and metadata descriptions** to enable re-use and machine processing,
7. a **system to register concepts and concept vocabularies** to enable re-use and machine processing,
8. a **system of authorisation record registries** to enable efficient access control in large federations of repositories,
9. **a system of license registries** to efficiently deal with licenses and their acceptance,
10. an **ecosystem of tools and operating procedures** that enable data service providers to efficiently manage digital objects as part of their daily business and collectively populate the digital object network layer and to put data into context (software, publications, etc.).

# Analysis of the GDOC Virtual Layer

This section presents three analysis steps which attempt to (1) distinguishing the entities in the GDOC Virtual Layer Model, (2) elicit the relationships between those entities, and (3) suggest possible refinements.

## The Elements of the GDOC VL

The advance in the creation of each of these components is defined in Table 1. The DF IG recommendation lacks precise definitions. For this review, definitions were supplemented with those found in the RDA TeDT [3], or the ENVRI RM, when not otherwise defined (see Glossary).

Table 1 Status of the implementation of GDOC Virtual Layer

| **VL Concept** | **Status** |
| --- | --- |
| Trustworthy Repositories (1) | it is known how to set up DFT certified repositories supporting a high level of FAIR-ness, it is a matter of consequent realisation and funding support to make the network available for everyone |
| Assessment of Repositories (1) | DSA/WDS is in use and available, but probably will evolve over time and applied by all repositories which want to appear as serious repositories |
| Repository Registry System (2) | with re3data a registry is available for human processing, a system needs to be added for machine processing and an RDA WG has been started |
| Repository API (1,2) | a new RDA group and others have defined an API or are in process of defining one based on earlier experiences |
| PID registry and resolving system (3) | the Handle System12 is tested and available, service providers need better funding support |
| Type registry system (4) | the specifications have been worked out by an RDA WG, but still much work will be necessary to make it professionally usable |
| set of core types (5) | this is currently being investigated based on earlier work within an RDA WG |
| MD schema registry system (6) | there are many registries - one has been setup by an RDA WG, we need to turn the many flavours to a usable and maintained system |
| MD registry system (6) | Same |
| Concept Registry (7) | same |
| Vocabulary Registry (7) | same |
| Authorisation Record Registry (8) | this needs to be developed urgently |
| License Registry (9) | this needs to be developed urgently, there is a first test solution in Finland |
| Ecosystem of Tools (10) | this does not yet exist, but is urgently needed, it will include a wide range of services from simple operations such as "replication" to "complex brokering" |

The clear definition of terms allowed the identification of three types of elements in the GDOC model: components, collections and tasks. From the term definitions, 11 of the terms can be classified as components, 2 are collections and one is a process/task.

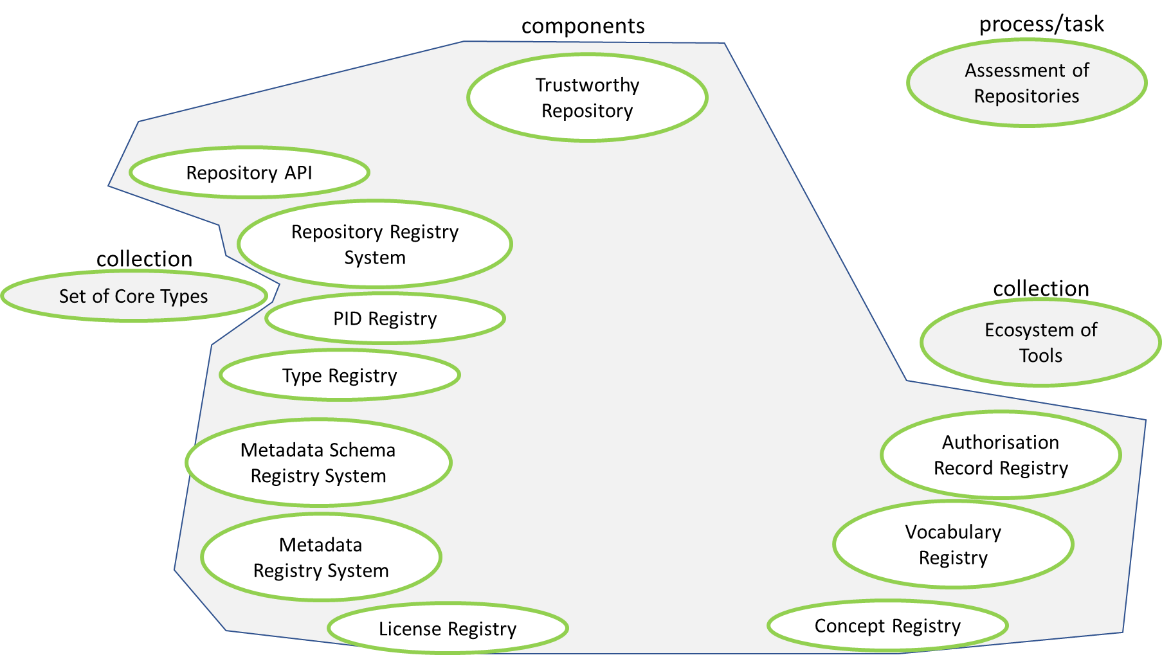


Figure 1 Step 1 Analysis of definitions and classification of GDOC VL elements.

## The relationships between the elements of the GDOC VL

The second step aimed to establish concrete relationships between the elements of the GDOC VL. In a first attempt the relationships between the elements were drawn as indicated in Figure 2. The definition of relationships is mainly structural, and it is imprecise, because the proposal lacks a clear specification of meaningful functional relations between the VL elements.

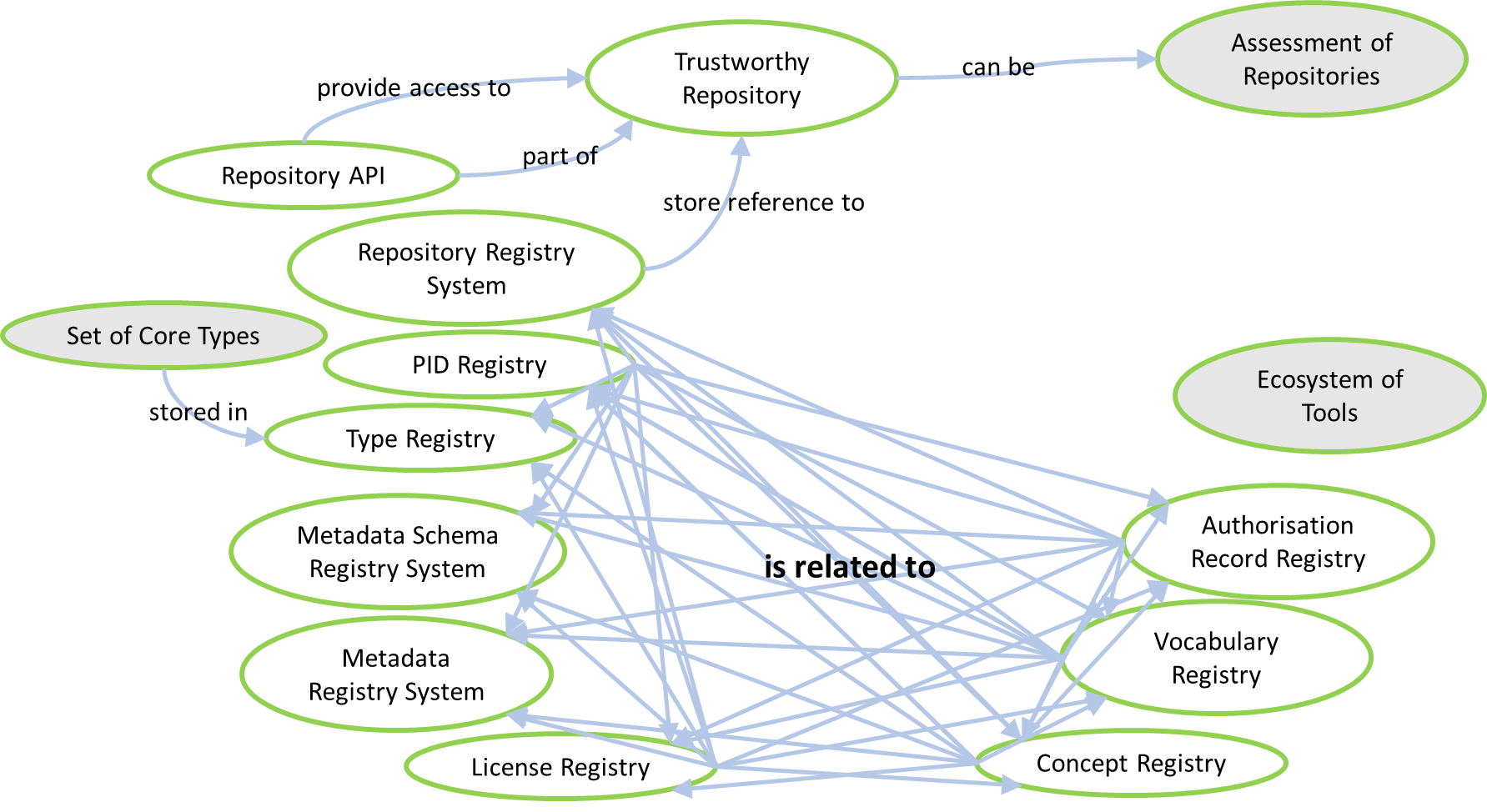


Figure 2 Step 2: Initial definition of relationships between GDOC VL elements

## Refinements and improvements

The complex relationships between the different kinds of registries, and the improved connection between the Assessment of Repositories task requires the addition of two bridge terms: Assessed and Registry. Assessed is defined as a property of trustworthy repository, while Registry is defined as a parent class to all registry objects[[1]](#footnote-1). For completeness, two more elements are proposed: Repository and Tool Registry. The Repository class was added as the parent class of Trustworthy Repository. Tool Registry was added as a registry to store references to the tools in the Ecosystem of Tools.

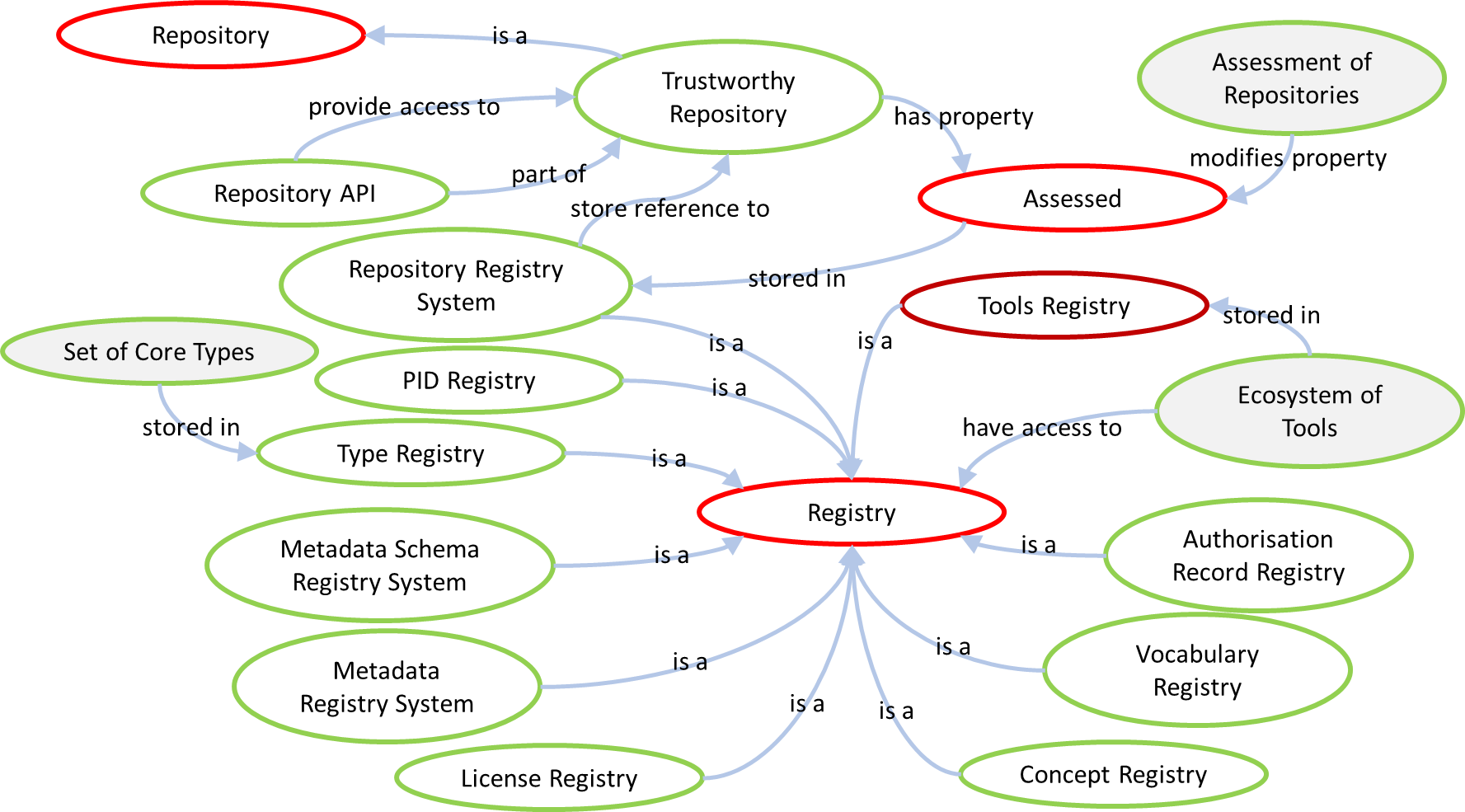


Figure 3 Step 3: Refinement of relationships

# Suggested improvements

The analysis of the GDOC Model, also suggests some improvements to the DF IG recommendation which are: a set of referenced definitions to be added to the proposal, a set of additional terms that can enhance the model, a set of structural relationships, and adding functional relationships between terms.

# Acronyms

**DF IG:** Data Fabric Interest Group of the RDA

**DFT IG:** Data Foundation and Terminology Interest Group (DFT IG) of the RDA

**DSA/WDS:** Data Seal of Approval/World Data System

**ENVRI RM:** Environmental Research Infrastructure Reference Model

**GDOC:** Global Digital Object Cloud

**RDA:** Research Data Alliance

**TeD-T:** term definition tool of the RDA DFT IG

# Glossary

**Assessment of Repositories:** a set of harmonized Common Procedures for certification of repositories at the core level, drawing from the procedures already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (ICSU­WDS)[[2]](#footnote-2).

**Authentication and** **Authorisation Record Registry****:** a system of authentication and authorisation records designed to enable efficient access control in large federations of repositories. The Authentication and Authorisation Record Registry provide secure and standardized mechanisms (SAML) to interact with an authentication system verifying and validating a user's identity[[3]](#footnote-3).

**Catalogue System:** a special type of storage system designed to support building logical structures for classifying data and metadata[[4]](#footnote-4).

**Concept Registry:** A type of special type of registry to store and manage data concepts which are elementary descriptors in a linguistic structure or an annotation scheme[[5]](#footnote-5).

**Core Types:** core types are used to describe the state of DOs enabling machine action[[6]](#footnote-6).

**Data Type Registry:** A type of registry for data types supporting their standardization, uniqueness and discoverability[[7]](#footnote-7).

**Digital Object:** (1) A digital object is composed of structured sequence of bits/bytes. As an object it is named. The bit sequence (see definition) realizing the object can be identified & accessed by a unique and persistent identifier or by use of referencing attributes describing its properties**.** (2) Digital Object is also called a Digital Entity defined as “machine-independent data structure consisting of one or more elements in digital form that can be parsed by different information systems; the structure helps to enable interoperability among diverse information systems in the Internet.”[[8]](#footnote-8)

**Digital Repository:** is an infrastructure component that is able to store, manage and curate Digital Objects and return their bitstreams when a request is being issued[[9]](#footnote-9).

**Ecosystem of Tools:** a set of registries of tools and operating procedures designed to enable data service providers to efficiently manage digital objects as part of their daily business and collectively populate the digital object network layer and to put data into context (software, publications, etc.)[[10]](#footnote-10).

**License Registry:** a system of linking digital objects to licensing agreements designed to efficiently manage licenses and their acceptance when accessing distributed digital objects[[11]](#footnote-11).

**Metadata Registry:** a system to enable management and publishing of metadata schema declarations, application profile declarations, and value space declarations. Because a given metadata schema or application profile may evolve, metadata registries are used to maintain relationships among a schema's various versions to promote semantic and machine interoperability over time[[12]](#footnote-12).

**Metadata Schema:** definition of a set of metadata elements and the rules governing the use of metadata elements to describe a resource[[13]](#footnote-13).

**Metadata State:** an object property that determines the set of all sequences of actions (or traces) in which the metadata object can participate, at a given instant in time (as defined in ODP, ISO/IEC 10746-2)[[14]](#footnote-14).

**Persistent Data State:** an object property that determines the set of all sequences of actions (or traces) in which the object can participate, at a given instant in time (as defined in ODP, ISO/IEC 10746-2)[[15]](#footnote-15).

**PID:** a Persistent Identifier (PID) is a long-lasting reference to a document, file, web page, or other object[[16]](#footnote-16).

**PID Generator:** a system which generates persistent global unique identifiers for a digital object[[17]](#footnote-17). The PID Generator can be implemented as part of a PID Service or a PID Manager

**PID Manager:** a system or service that assigns persistent global unique identifiers to data and metadata products. The Manager invokes an external entity, the PID Service, to register the PIDs. The manager maintains a local catalogue of PIDs that are being used to reference data and metadata. If the data or metadata in the RI change location or are removed, the PID manager updates this information locally and informs the PID Service[[18]](#footnote-18).

**PID Registry:** an information system for registering PIDs[[19]](#footnote-19).

**PID Resolving System:** The combination/integration of PID Services/Registries and PID Managers that are configured to facilitate resolving PIDs[[20]](#footnote-20)

**PID Service:** a public system or service which can generate and assign persistent global unique identifiers (PIDs). The PID Service also maintains a public registry of PIDs for digital objects[[21]](#footnote-21).

**Repository API:** Application programming interface for a Digital Repository, allowing the creation of applications that access the features and data of the registry[[22]](#footnote-22).

**Repository Registry:** is a type of registry that collects useful information about repositories for human consumption in order that depositors and users can easily find where to go to for their data needs[[23]](#footnote-23).

**Trustworthy Repository:** a digital repository that undertakes (undergoes) regularly quality assessments successfully such as Data Seal of Approval/World Data Systems *(these two initiatives harmonized their requirements)* [[24]](#footnote-24) . Slight variation on the definition of Trusted Repository from [3]

**Vocabulary Registry:** Vocabulary is a body or listing of a group of "terms" used by a community for particular domain purposes. Vocabularies usually are annotated by explanatory definitions as to meaning and purpose and links to related terms. Vocabularies are usually arranged alphabetically within a lexicon or glossary where definitions are provided although as part of communication definitions may not have been explicitly provided[[25]](#footnote-25).

# References

1. ENVRI (2017). ENVRI Reference Model V2.2, Consulted on: 20/12/2017, Available at: https://envri.eu/rm
2. RDA Data Fabric Interest Group (2017). Recommendations for Implementing a Virtual Layer for Management of the Complete Life Cycle of Scientific Data, V10. Consulted on: 19/12/2017, Available at: https://www.rd-alliance.org/system/files/DataFabric\_SupportingOutput\_recommendation-aug-2017-v10.pdf
3. RDA Data Foundation and Terminology Interest Group (2016). Term Definition Tool, consulted on 20/12/2017, available at: https://smw-rda.esc.rzg.mpg.de/index.php/Main\_Page
4. RDA Repository Audit and Certification DSA–WDS Partnership WG (2017). Repository Audit and Certification Catalogues, DSA–WDS Partnership Working Group Catalogue of Common Procedures.
5. Wilkinson, M. D. et. al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, Article number: 160018. doi:10.1038/sdata.2016.18.

1. According to Keith Jeffery the registries defined in the GDOC Model are the result of an effort to bring together and acknowledge the outputs and contributions of different RDA working groups [↑](#footnote-ref-1)
2. [4] p. 1 [↑](#footnote-ref-2)
3. Undefined, derived from [2] p.5 and p. 16 [↑](#footnote-ref-3)
4. [1] at: https://confluence.egi.eu/display/EC/SV+Community+Roles#SVCommunityRoles-roles\_curRolesintheDataCurationCommunity. [↑](#footnote-ref-4)
5. Undefined, derived from the definition of Category Registry, [3] at: https://smw-rda.esc.rzg.mpg.de/index.php?title=Category\_Registry [↑](#footnote-ref-5)
6. Undefined, derived from [2] p. 5. [↑](#footnote-ref-6)
7. [3] at: https://smw-rda.esc.rzg.mpg.de/index.php?title=Data\_type\_registry [↑](#footnote-ref-7)
8. [3] at: https://smw-rda.esc.rzg.mpg.de/index.php?title=Digital\_Object [↑](#footnote-ref-8)
9. [2] p. 13, [3] at: https://smw-rda.esc.rzg.mpg.de/index.php?title=Digital\_repository [↑](#footnote-ref-9)
10. Undefined, derived from [2] p. 5 [↑](#footnote-ref-10)
11. Undefined, derived from [2] p. 5 [↑](#footnote-ref-11)
12. [3] at: https://smw-rda.esc.rzg.mpg.de/index.php?title=Metadata\_Registry [↑](#footnote-ref-12)
13. [1] at https://confluence.egi.eu/display/EC/IV+Information+Objects#IVInformationObjects-infobj\_metadatametadata [↑](#footnote-ref-13)
14. [1] at https://confluence.egi.eu/display/EC/IV+Information+Objects#IVInformationObjects-metadatastate [↑](#footnote-ref-14)
15. [1] at https://confluence.egi.eu/display/EC/IV+Information+Objects#IVInformationObjects-persistentdatastate [↑](#footnote-ref-15)
16. Undefined, taken from: https://en.wikipedia.org/wiki/Persistent\_identifier [↑](#footnote-ref-16)
17. [1] at https://confluence.egi.eu/display/EC/Appendix+B+Terminology+and+Glossary [↑](#footnote-ref-17)
18. [1] at https://confluence.egi.eu/display/EC/SV+Community+Roles#SVCommunityRoles-roles\_curRolesintheDataCurationCommunity [↑](#footnote-ref-18)
19. [1] at https://confluence.egi.eu/display/EC/Appendix+B+Terminology+and+Glossary [↑](#footnote-ref-19)
20. Undefined, derived from PID Service, PID Manager, and PID Registry definitions [↑](#footnote-ref-20)
21. [1] at https://confluence.egi.eu/display/EC/SV+Community+Roles#SVCommunityRoles-roles\_curRolesintheDataCurationCommunity [↑](#footnote-ref-21)
22. Undefined, derived from: http://www.dictionary.com/browse/api?s=t [↑](#footnote-ref-22)
23. [3] at: https://smw-rda.esc.rzg.mpg.de/index.php?title=Repository\_Registry [↑](#footnote-ref-23)
24. [2] p. 13 [↑](#footnote-ref-24)
25. Undefined, derived from the definition of Category Registry, [3] at: https://smw-rda.esc.rzg.mpg.de/index.php?title=Vocabulary [↑](#footnote-ref-25)