Data Fabric Interest Group

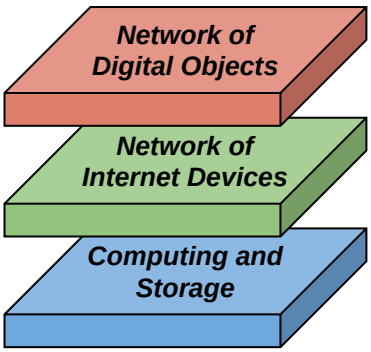
Summary of Virtual Layer Recommendations

Data volumes and complexity of scientific data are increasing rapidly, and cross-border (disciplines, countries) re-usage of such data is becoming more and more common in most research fields. Open data as an agreed default model, backed by FAIR principles to ensure data is 'Findable, Accessible, Interoperable, and Re-usable', together with Data Management Plans to encourage stronger curation and stewardship have all emerged as community policies underpinning research data sharing without barriers. At the core of operationalizing this ambition is recognition and consensus that:

- All Digital Objects (DO) must be registered in trustworthy repositories to make them findable, accessible and persistent; and,
- All such registered DOs must have assigned Persistent Identifiers (PID) and Metadata (MD) to improve interoperability and re-use.

Presently, the number of research and data infrastructures worldwide is growing, with many similar components re-invented in different variations. Due to this fragmentation, the costs of building and maintaining infrastructure are higher than they should be; as are costs for data re-use. This trend must be counteracted by identifying common components and means of interoperability as a new conceptual framework that creates a new global momentum on data infrastructure interoperability, akin to the creation of the Internet itself. This is the aim of the **"Global Digital Object Network"**.

The Network of Digital Objects as a new virtual layer



Computing and storage components have existed from the beginning of modern computing. They have undergone technological and architectural changes, but the basic concepts remain stable. In the late 20th century, a first virtualised network of devices was introduced by stating that the computers we use need not be local but can be accessed over wide areas via network protocols. The Internet Protocol Addressing system identifies all devices in the network so they can exchange messages using protocols and registries. Together these two layers provide us today with almost unlimited cloud computing capacity with near universal access to dynamically configurable, shared compute and storage capabilities. Yet, to support data-driven science, this is no longer sufficient. We must address data organisation, typing and re-use facilitation to be able to fully realise the value of data, network, compute and storage. We are now in the phase where we must provide a new layer of network wide virtualisation that interlinks data and other artefacts with the help of PIDs, new protocols and new registries.

A digital object (DO) is represented by a bitstream, is referenced and identified by a persistent identifier and has properties that are described by metadata. Digital objects can be data, collections of data, metadata, software, configurations, provenance, etc. A DO can also be a surrogate for a physical object.

How do we realize the integrated virtual layer?

The following components will be required to implement the FAIR principles and realize a fully integrated virtual digital object layer:

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 DOs 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 and collectively populate the digital object network layer.

Further background information to be found in the discussion document:
https://www.rd-alliance.org/group/data-fabric-ig/wiki/recommendations-implementing-virtual-layer-management-complete-life-cycle