Global Digital Object Cloud (DOC) - A Guiding Vision

11 September 2016

Larry Lannom, Peter Wittenburg

Background

It is now widely agreed in the Research Data Alliance (RDA) and beyond that concept of the "Digital Object (DO)" is central to proper data management, access and use. A DO has a bit sequence that can be stored in multiple repositories and is associated with a Persistent Identifier (PID) and quality metadata [1]. The persistent identifier, augmented with usable attributes about the DO, can provide identification, location, and other functions. The metadata description has many different functions amongst which is offering information allowing proper interpretation and reuse of the bit sequences. Due to the claim of persistence PID records are increasingly seen in a binding role, i.e., storing persistently all the necessary actionable references to the locations of the bit sequences, to the metadata and other useful information. Given such a domain of registered DOs it is compelling for users to just deal with PIDs and metadata as widely as possible. We call this layer of virtualization the Global Digital Object Cloud (DOC), which is based on the ideas of a Digital Object Architecture [2] and fully compliant with the FAIR principles [3].

Global Digital Object Cloud

The concept of the Global Digital Object Cloud (DOC) is illustrated in Figure 1 below. Key to this model are Digital Objects, which comprise a virtualization layer on top of various network resources and services, much as files and databases currently are virtualization layers on top of raw computer storage and sets of standard processes, but in this case extended to the network level. Each Digital Object is persistently identified, with the persistence guarantee dependent on use case, such that every object on the network can be referenced and, given adequate permissions, can be the subject of stated operations, including but not limited to raw access. Further, the objects are described and typed by metadata such that their structure can be well understood through the mechanisms of type registries. These objects are shown in the second panel from the left in Figure 1.

The ability for clients to call, access, and act upon these objects is provided by the network services shown in panel three of Figure 1. Here the set of repositories, registries, and identifier resolution services provide the structure and processing that enable the array of storage and specific data management services, shown as the rightmost and lowest level panel of Figure 1, to be consistently addressed as coherent objects. The repositories essentially serve as unifying portals into the lower level storage and heterogeneous information management technologies and

they have the responsibility of presenting the consistent structured object view to clients, regardless of the details of the underlying data stores and management systems. The registries provide discovery services for objects, by providing searchable metadata, as well as providing the information needed to access and act upon objects, through registration of types and related services that are relevant to specific types. The identifier resolution services enable the objects to be directly addressed, regardless of their current state and location, again depending on permissions as well as the capabilities of the individual objects and the repositories providing the portal service.

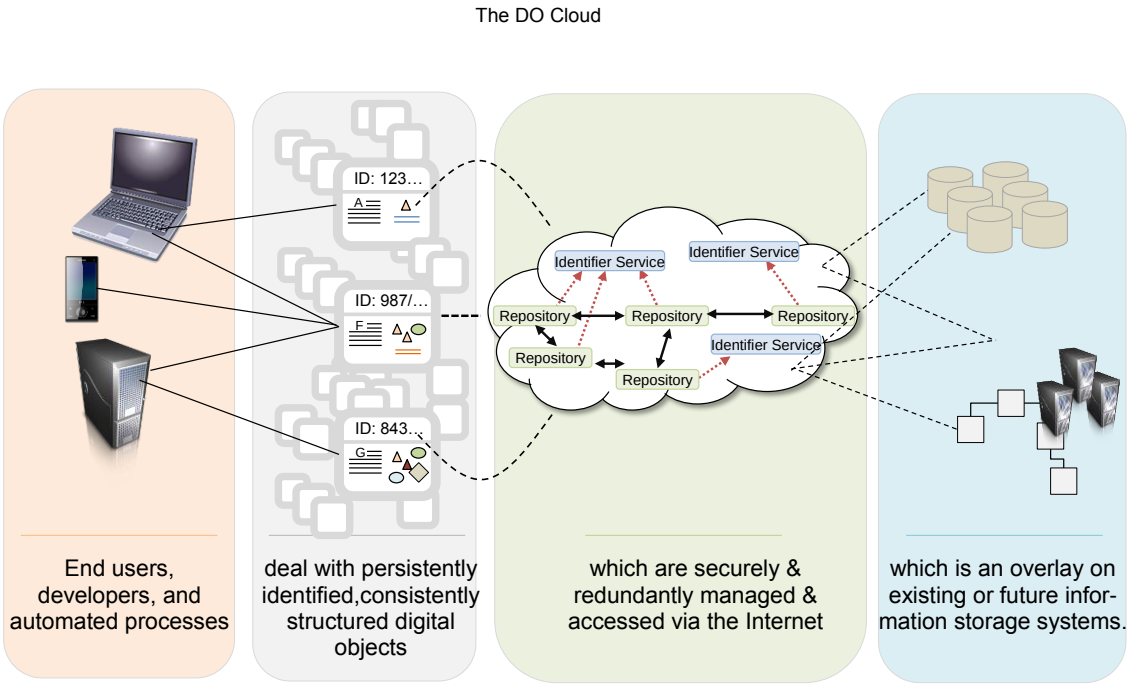

Figure 1

Note that registries, repositories, and resolution services are functional categories and not necessarily software categories. Metadata registries, for example, are best regarded as metadata object repositories and the same software-based service can provide both object discovery and access services.

Necessary Steps

While we are far away from a complete implementation of this DOC model, we find that it is necessary to define such a goal, and to keep it in mind as we work towards revolutionizing our up to now inefficient approaches. One of the Data Fabric configuration options currently being pursued, for example, is a PID Centric approach to data management [4] and it embodies several principles from the model, i.e., persistent identification, the centrality of well-managed repositories, and type registries. Some communities, such as the global climate modeling community which has been relying widely on assigning PIDs to all their Digital Objects for many years now, want to take the next step towards a DOC, i.e., they are designing

powerful operators to implement global infrastructure based on the sketched virtualization.

Summary

This DOC model does not, of course, solve any problems on its own but serves instead to guide our work in the direction of a non-proprietary and highly efficient data management and access infrastructure where globally available and stable PIDs are anchors for all activities. DOC serves as a framework into which multiple lower-level data storage and management solutions can be placed as well as accommodating a wide variety of applications. Users deal with a domain of registered Digital Objects where they primarily just deal with entities such as PIDs and metadata as long as they do not start calculations on the bit sequences themselves. But it is the task of the DOC middleware 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. DOC based on these principles opens up completely new perspectives that also have the power to involve industry in providing compliant software.

It is obvious though that in such PID centric models we are highly dependent on a functioning, highly available and powerful PID infrastructure which allows everyone to register and resolve PIDs to meaningful information about the DOs. Similarly to how we now rely on a worldwide IP-based infrastructure to connect computers, we need to be able to rely on a worldwide infrastructure for PIDs. The Handle System, now owned by the Swiss DONA Foundation and governed by an international board is with its multinational root nodes and numerous service providers, including DOI providers, is a strong candidate for such an infrastructure.

Given the existence and broad support for such an infrastructure trust and validation, for example, could be applied at the object level instead of depending on a variety of underlying heterogeneous systems. As has been shown by Crossref, for example, a multitude of valuable services can be implemented on top of the PID infrastructure.

Interest in and support for such a global Digital Object Cloud emerged during the last year in the discussions in the RDA Data Fabric Interest Group. It needs further elaboration and thus we welcome collaboration on its development by interested communities.

[1] DFT Model: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

[2] Robert, E. Kahn: The Architectural Evolution of the Internet, http://www.cnri.reston.va.us/papers/Architectural_Evolution_Internet_17Nov10.pdf

[3] FAIR Principles: https://www.force11.org/group/fairgroup/fairprinciples

[4] PID Centric Approach: https://rd-alliance.org/group/data-fabric-ig/wiki/df-configuration-pid-centric-data-management-and-access.html