Separation of Concerns: PID Information Types and Domain Metadata

Tobias Weigel
DKRZ /
Universität Hamburg
Germany
weigel@dkrz.de

Timothy DiLauro
Data Conservancy /
Johns Hopkins University
USA
timmo@jhu.edu

Abstract

This brief article aims to define a pragmatic separation of concerns between metadata activities and the typed information associated with Persistent Identifiers. The illustrative model used is that of a black box or envelope metaphor. This distinction is important for ongoing debates within respective communities as well as in the working groups of the Research Data Alliance.

Keywords: Metadata; persistent Identifiers

A separation of concerns

From a data archive's viewpoint, a useful metaphor is that of the "black box" or "envelope": Data management is increasingly done by machinery rather than human users. So the machinery must know what to do with the boxes that come in through various channels, but it cannot open them for various reasons (for example due to performance and scalability requirements). We propose that metadata is a concern that is – from this particular view of automated data management – located inside the black box. A metadata description may actually be a black box object that must be managed just like all the others. Still, some information must be written on the outside of the box to be interpreted by the machinery (see Figure 1). This information, closely associated with a persistent Identifier (PID) and thus also called a PID record, may be a subset of metadata, but it may also contain additional information not interesting as domain metadata.

These metaphors also work well with a technological layer stack view that is also envisioned in the original Handle System design (Kahn and Wilensky, 2006) similarly to the IP stack (Clark, 1988). Conceptual architectures such as the OSI model use distinct layers of abstraction. Understanding neighboring layers is not required; layers are defined through clear interfaces at the boundaries and can be changed independently. Consequently, PID information forms a lower

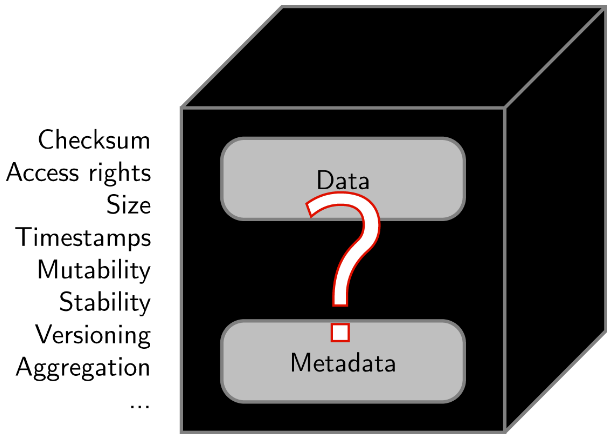

FIG. 1. The content of a black box is of less concern to machine agents managing it, who will instead look at properties available outside of the box.

layer, while metadata and all services working with them form a higher layer. Both layers are used by different actors, and there may be some transition of information between them. Services working independently in the framework of an individual layer will benefit from such an architectural division.

To understand the particular requirements and also the scoping boundaries for using PID information and more fully-fledged metadata, we need to give some practical examples. The EUDAT project[1] is using PID records to bind distributed replicas of data objects together. The iRODS[2]-based infrastructure services, which are machine agents, generate and act according to the PID record information without having to look at metadata objects. As indicated earlier, another important scenario under discussion regards the coordinated management of data and metadata objects where each of them receives its own PID. These PIDs might be linked together through the PID records, enabling efficient service operations such as replication and resource lookup. The varying levels of granularity commonly found in data and metadata hierarchies could also be expressed more formally by linking PIDs from different hierarchy levels together, e.g. linking single metadata records to construct aggregations of multiple data objects. Moreover, more conceptual PIDs might be created that then identify these aggregated data and metadata conglomerates.

More advanced scenarios include versioning and provenance tracing. In versioning scenarios, an infrastructure service would be contacted by a user (another machine agent, but potentially also a human user) with a request to replace an obsolete object with a more recent version of the same object. The infrastructure will have to trust the requesting agent on the question of "sameness" of objects, since it would not necessarily be able to verify this claim on itself. Accordingly, it will replace the older object in the data infrastructure, and, given such policies are in effect, assign a new PID to it and link this PID with the identifier of the obsolete object. By doing so, essential parts of the versioning history of an object are preserved, even though the obsolete object may be discarded. This can also be classified with distinct primary and secondary levels of preservation (Weigel et al., 2013). A provenance tracing scenario would look similar to this, though there may be differences regarding the cardinality of relations and the policies of keeping objects.

In particular these latter two scenarios impose questions of which pieces of information are located outside or inside the box and where potential overlaps exist. For now, the envelope metaphor appears to be useful to guide practical developments and support detail decisions.

Acknowledgements go to the participants of the CAMP-4-DATA workshop whose feedback helped to clarify the examples and point out further paths of action.

References

Clark, David (1988): The design philosophy of the DARPA internet protocols. SIGCOMM Comput. Commun. Rev., vol. 18, no. 4 (pp. 106-114). doi:10.1145/52325.52336

Kahn, Robert. Robert Wilensky (2006): A framework for distributed digital object services. International Journal on Digital Libraries, vol. 6, no. 2 (pp. 115-123). doi:10.1007/s00799-005-0128-x

Weigel, Tobias. Michael Lautenschlager, Frank Toussaint, Stephan Kindermann (2013): A framework for extended persistent identification of scientific assets. Data Science Journal, vol. 12, pp. 10-12, 2013. doi:10.2481/dsj.12-036

[1] http://www.eudat.eu

[2] integrated Rule-Oriented Data management System, http://www.irods.org