INTEROPERABILITY, PIDs and EPOS

Keith G. Jeffery

WP07 Leader

Interoperability is the ability of heterogeneous computer systems, or components of them, to work together. Simple examples are: (1) a dataset created by system A is readable and ideally processable (after any appropriate conversion) by system B; (2) System A utilises software components copied from System B (an example of re-use); (3) a user usually utilising a workflow or process in system A can utilise that same workflow or process in system B. More complex examples are: (4) System A during execution accesses and executes software within System B; (5) a particular instance of a system can be redeployed dynamically across more than one heterogeneous computing platform such as CLOUDs or GRIDs.

The overall aim of interoperability is to make heterogeneous (and usually distributed and distant) systems appear to the end-user as if they are his/her "home system". This is one aspect of virtualisation in which physical assets are described by metadata as digital objects to be managed digitally. One approach to interoperability (the so-called broker approach) is to write software to convert between every pair of DDSS object types. It is obvious that – given n DDSS object types, this requires n*(n-1) – almost n**2 – brokers. In EPOS we take the approach of conversion to a single superset canonical format which means that we require only n convertors.

In EPOS the ICS Catalog holds DDSS information provided by the TCS in CERIF (Common European Research Information Format: an EU recommendation to Member States). The TCS provide the metadata they have for their own purposes and agree with the ICS team the mapping from their formats to CERIF. This is not a trivial task but once done provides the basis for the convertors and thus for interoperation. Commonly this mapping exposes that additional metadata elements are required from the TCS since the commonly-used metadata formats do not have all the required information necessary for interoperability as described above. This is because generally these formats were designed for data processing within a closed domain and/or in one organisation whereas EPOS-IP is multi-domain and multi-organisation. CERIF provides the required information for interoperability at all the levels mentioned above and also provides for interoperability between user and user, and users and managers. Thus an end-user can locate one or more datasets, can analyse and visualise it/them (by downloading to a local computing resource or – by agreement – at the source computing resources). A user can locate appropriate software and
execute it at their "home computing facility" or - by agreement - at the location of the original software. A user can create a workflow involving data, software components and computing resources (or even including activating instrumentation) all of which is executed at several appropriate computing resources providing the end-user with monitoring information on the progress of the execution until the result are delivered. In short, interoperation is what makes EPOS a reality.

**PIDs** are permanent identifiers. Usually PID includes the concept of UID (Unique identifier). One problem is that there are many schemes for the allocation and management of UIDs/PIDs. Not only do many agencies allocate their own PID/UIDs (think of social security identifier, driving licence identifier, employee identifier... for a person) but these IDs are allocated for specific roles. Fortunately, CERIF - used for the EPOS ICS catalog - has the concept of federated IDs so can provide logical relationships between different IDs allocated by different schemes and agencies which purport to identify the same digital object (such as a dataset, software component, person (as a user), organisation, computing resource, scholarly publication etc.). PID/UIDs are very useful for quick identification of a digital object within a computer system; however, from an end-user point of view the PID/ID should be invisible since the end-user will utilise assets (datasets, software...) selected by a combination of attributes (such as geographical area, temporal duration, keywords...). In fact, within EPOS-IP we assign an internal EPOS PID/UID so that we can assure referential and functional integrity of the data and thus ensure interoperation.

The figure illustrates the EPOS-IP ICT architecture: it is clear immediately that the metadata catalog "drives" the other components by providing the information required to interoperate among the TCS, ICS-D, CES and external e-Infrastructures.