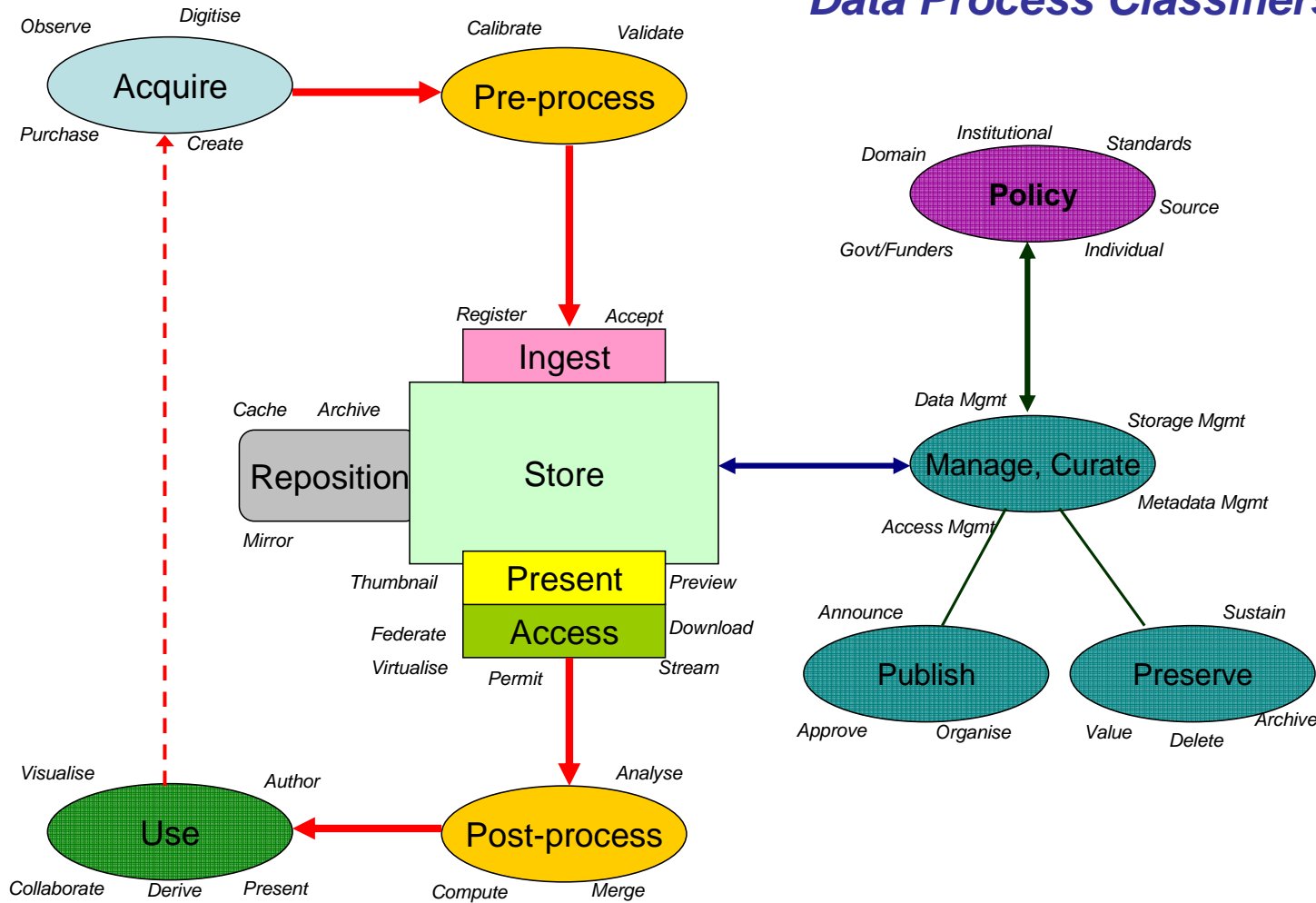


Data Process Classifiers



© 2006 ARSR_AERES Project Team

This diagram seeks to set out a structure for the *classification* of data processes, i.e. a reasonably standard set of terms that can be used for comparisons and analysis of common requirements. These in turn could lead to the identification of appropriate standards for various processes, and the development of best-practise guidelines. It could be used by end-users to map out their processes, by service providers and educators to elicit information from data-related activities, and by service providers to identify coverage of support activities. The core terms are within the shapes; the additional terms around the shapes provide examples or clarifications.

Guidelines

- The various elements classify *processes*, i.e. they are (generally) verbs that identify *capabilities*, and not facilities or infrastructure.
- The processes only describe *actions* that are undertaken, and do not identify or align with responsibilities for the provision of those processes. It is generally difficult to identify hard boundaries for the processes, and hence hard boundaries for responsibilities.
- Each process identified here is “optional”, i.e. may be absent from the process path, or the path may not complete the full cycle. Some activities may cover only the first half of the process chain, from *acquire* to *store*, others only cover the latter half from *store* to *use*. It is also quite feasible to have a process from acquisition to dissemination that does not *store* data at all.
- Arrows indicate information flows (so for example, data, policy, management, control, influence, education, etc.).
- There are many more implicit links than are shown here. For example the influence of management, particularly policy, preservation and publication requirements, is across all processes, not just the *store* element. In many activities, management and policies have strong influence through the storage point (e.g. a repository), and this influence then flows to all other process (on acquisition or dissemination), but this is not always the case. Similarly, policy influences and education will be across the data being managed and the processes themselves.
- A full representation of a project may repeat the diagram more than once, indicated by the link from *use* to *acquire*. One needs to focus on what is being ‘managed’ each time round; if it is different data, it is a different activity and needs to have its own classification done (which may have overlap, repetition, re-use and common policy frameworks – or not). An example might be a project that acquires survey data on paper then has a data-entry phase where the survey data is digitised, and the paper data is archived. These are actually two separate process flows, one on the paper data, one on the digital data, with potentially separate frameworks, policies and capability requirements, where the output of the paper-flow feeds the input of the digital data flow. This may not be as compact as a single diagram with multiple phases, but is usually more complete and simpler to interpret.
- Some broader capabilities can have hooks in more than one place. For example, managing access has three specific points identified: *policy*, which sets the framework for access permissions, *access management* which is where the policy is mapped to explicit (machine readable perhaps) rules, and *permit* as part of the *access* process where the actual control is exercised, i.e. access is allowed or prevented.
- Each process element is likely to be itself quite rich, and often can be broken into further sub-elements
- Meta-layer concepts such “repository”, “acquisition-phase”, “maintenance-phase”, “dissemination-phase”, “management drivers”, “responsibility boundary” and so on, can be overlaid if that assists the interpretation of the diagram in any particular use-case. Defining boundaries for these concepts though is again a difficulty.

Process terms

- **Acquire**: besides the self-explanatory aspects, one needs to be aware that data may be undergoing a second or further cycle of processes, e.g. from analogue (non-digital) data such as paper, tape, film, and physical objects. Such data may inherit e.g. policy frameworks from the original cycle.
- **Pre-Process**: is generally used to describe processes that “clean-up” data and ensure its validity before undergoing the expense of moving it into some kind of storage capability, and/or making it accessible.
- **Ingest**: is a formal, potentially legal, process, apart from moving data onto physical storage say. It is a process that notifies the storage capability through some registration mechanism that new data is there to be managed. It can be seen as a responsibility boundary, and so there may be an acceptance element.
- **Store**: should be seen as the simple process of storage itself (e.g. as a file, a record, within a simple or complex hierarchical infrastructure), but is the hub for management of the data and the cross-over from acquisition to dissemination. As a capability it needs to be managed, just as the data itself is managed.
- **Reposition**: describes processes that create a copy of the data for various reasons. Most data movement occurs for *Performance*, *Protection* and *Policy*, so e.g. caching for quicker access, mirroring for risk-management purposes, and moving data across a policy boundary e.g. where a physical facility is not accessible due to policy, or where data has to be sanitised (e.g. anonymised, de-identified) to be made accessible.
- **Present**: refers to processes that create ‘presentation’ versions of the data. These might be preview copies like thumbnails or extracts from larger files. It is a process that creates a specific derivative product, not a new product.
- **Access**: is defined as the process where data leaves a storage capability and is delivered to another process. This could include interfaces such as web, ftp, sql, rtsp, etc. and is a crucial policy-enforcement point for permitting or denying access. It also encompasses processes that support *virtualised* access to data, i.e. where the requirements of underlying infrastructure such as hardware and software are hidden from view (a web interface to a database, a POSIX interface to a specific HSM installation), and access to *federated* data, i.e. where (some/all of the) data is stored in multiple locations. The latter links with the repositioning process, usually through some replica registry.
- **Post-Process**: is subtly different from *Use*. It covers processes that e.g. combine or modify data to generate outputs that are used by end-users. These could include processes for data fusion, analysis, modeling, etc.

- **Use:** is the ultimate target of the data cycle, and covers processes that an end-user undertakes with the final data. These can include areas such as visualisation, collaboration around data, authoring papers for publication and/or presentation and so on. The outputs of this process element may themselves become inputs for another process cycle, so e.g. publications derived from data are themselves stored in a repository, ideally with references to the source data that led to the publication.
- **Manage/Curate:** covers a broad range of management processes. These can be broken down into several sub-headings. Storage management deals with the underlying infrastructure (hardware, software, processes). Data management deals with issues surrounding the actual data (e.g. files, formats, etc.). Metadata management deals with curation aspects (metadata editing, annotation, etc.) as well as standards, schemas and metadata capture/storage. Process management deals with issues surrounding e.g. the ownership and maintenance of supported processes. Access management addresses the process of turning policy into something that can be implemented and enforced.
- **Publish:** is used in a formal, potentially legal, fashion to describe the ‘publication’ of data (not of papers derived from the data). It is the set of processes that make data formally accessible to end-users. These usually include additional management elements, to organise data into a publishable structure, a formal approval process, and announcement of accessibility. These could include technical aspects such as URNs/URIs and discovery tools.
- **Preserve:** covers processes that ensure the longer-term accessibility of the data. These include the technical aspects such as archival standards for data, metadata, associated software and algorithms, and storage, and the ability to sustain the various processes and infrastructure. They also include processes that allow a community to define “longer-term”, implying community valuation of data, and potentially the deletion of some data.
- **Policy:** is what ultimately controls *all* of the process elements. Policy influences the management of and access to data, processes and physical infrastructure, and policy sources can be extremely diverse, as indicated. Ultimate access to any piece of data may touch multiple policies. It should be noted that policies themselves need to be informed by the various processes, to avoid becoming either meaningless or a complete block to end-users or data providers. New use-cases arise all the time, and may identify policy shortcomings. New technologies may mean a policy becomes unenforceable, or has to be enforced in a different fashion. So there needs to be a process for communication in both directions.