

e-Research Context

R1. CITATION OF DATASET

Physicist uses an online collaboration dataset to create a new dataset and a paper. The datasets are not necessarily persistent or publication-quality.

- Author accesses collaboration dataset through identifier. (Note: this may be the Grid and not a Handle)
- Author generates a publishable dataset out of the collab dataset.
- Publishable dataset deposited in open access repository with global persistent ID.
- Author writes paper, which links to publishable dataset.
- Author publishes paper, online or paper.
- Six months later, reader clicks on hyperlink to publication dataset.
- Dataset gateway shows up on browser.
- Ten years later, reader clicks on hyperlink to publication dataset.
- Archival stub shows up on browser.

R2. IDENTIFY INDIVIDUAL DATA—COUNTING

A physicist wants an identifier minted for every electron detected during an experiment (= a timespan during which equipment is left in steady state recording data). This amounts to a new identifier every microsecond. This stretches the capacity of a global identifier system.

- Researcher has a counter in experimental local system that increments whenever a new electron is located.
- Counter values act as index to electron within local system.
- Counter is (subsequently) mapped to a global identifier, for each such counter value.
- Local services interacting with the local identifier can likewise be mapped to global services interacting with global identifiers; e.g. if electron 7777 is mapped to global ID 1.2.3/635, `VELOCITY(7777)` may be mapped to <http://www.example.com/velocimeter?id=1.2.3/635> (with the service resolving the global to the local identifier). Alternatively the local service invocation may be bundled with the global identifier (<http://www.handle.net/hdl/1.2.3/635?tellvelocity>)

Note

It is not clear this use scenario is realistic in e-Research; physicists prefer a global identifier just for the dataset, and any access to components of the dataset are restricted to local access. (This is the *functional granularity* problem for identifiers.) The use case may be more applicable in e-Learning/simulations, e.g. tracking events in a virtual learning environment.

The use scenario plays out differently depending on the time constraints of turnaround. One can turn on a counter for stars as well as electrons; but the speed requirements for the latter are more restrictive, and affect any workflow implementing them.

R3. IDENTIFY INDIVIDUAL DATA—INDEXED POINTS

A physicist wants identifiers to be registered for data which are not identified discretely, but through indices along a continuum: e.g. not “electron #7777”, but “the electron at time 4.64, x=73.6, y=62.1”. (Alternative: an instant needs to be selected on a video recording for annotation.)

- Researcher analyses data and identifies points along a continuum within the data domain as things to be identified.
- Access to the referent on the local system is through a service parameterised on location and data domain. The data domain is a digital object (which may be generated by simulation or measurement), but within that object there are no discrete digital objects to refer to—the domain is a continuum.
- Nonetheless, a global identifier can be associated with this local parameterised locator: global ID 1.2.3/635 resolves to <http://www.example.com/graphreading?experiment=6162&t=4.64&x=73.6&y=62.1>. (The experiment parameter is the data domain, which is a digital object; the other parameters pick a referent out of a continuum.)

Note

While identifiers abhor a continuum, this use scenario is tractable: the choice of thing to identifier is the requester’s, and it is their responsibility to make the provided resolution work. Using location as a discriminator attribute for identification is exactly what IDs resolving to locators (URLs) do in cyber- rather than physical space, and takes advantage of physics: discrete objects occupy a unique locality at any time.

We can do two things with datasets and identifiers: have either IDs for just the data sets, and a consistent indexing scheme operating on the dataset; or else identify the individual data points of interest as well as the datasets. In theory, there is no limit to the number of identifiers that

can be created. (Web-Actionable URIs are constrained only by particular browser technologies—see <http://www.boutell.com/newfaq/misc/urllength.html> ; but even the smallest limit imposed Internet Explorer, 200 characters, allows $36^{200} = 10^{312}$ alphanumeric strings. In practice though, the more identifiers, the more contaminated the search space is: exposing a billion identifiers' metadata for full-text search is not a good idea because of the probability of collisions, as well as any performance constraints.

R4. IDENTIFY DATA RANGES IN CONTINUUM

Nick Thieberger at Melbourne Uni is writing a paper on Efate, language of Vanuatu. His citations of Efate text are hyperlinked to time indexed spans of an video recording of the speaker. Accessing the video in turn gives the user access to time-indexed annotations. In both cases, the annotations refer to a time span, not a single point. The time span can be located through a locator parameterised on digital object (the video), starting time, and ending time. But that locator defines its own (virtual) digital object, and can be identified by a global identifier.

- A video (or other continuum of data) goes online with an identifier. In e-research terms, it constitutes a dataset.
- Researcher identifies a continuous range of the dataset (a time span) as a referent of interest: it is the time span of a speaker uttering a specific sentence.
- The range has a locator constructed for it, given the dataset identifier and endpoints; e.g. for dataset 1.2.3/456, <http://www.example.com/timeslice?clip=1.2.3%2F456&t1=1.02&t2=3.72> .
- The range can have metadata, and for that matter data, associated with it: annotation, transcription, gloss, Labanotation.
- The video span needs to have a persistent locator, as it will be referred to in a publication.
- The anchoring of data to the video span should not be contingent on the particular service doing the time-indexing. Anchoring metadata is most reliably done through an identifier for a virtual digital object, instead of a service request through a specific service host—unless the service has a guaranteed persistence comparable to that for the identifier.
- The span is assigned a persistent identifier.

Note

Pretty much the same use scenario applies if you're doing classroom research, and annotating video with observations

This scenario is similar to disaggregation, though it involves a continuum rather than discrete objects. What is constructed is a virtual data object: a transformation mediated through a service. In e-research terms: this is the virtual data grid.

A physicist picking 3 random discrete points out of an experimental continuum is doing a similar kind of disaggregation. So is an astronomer indexing an observation session, and narrowing the data down to a particular window through time indices and polar coordinates.

Having an identifier rather than a locator service for a span allows the span to be treated as a discrete digital object, which makes attaching metadata to it much easier.

A service is needed for the data transformation that generates the virtual data object. The persistence of an identifier to the virtual data object depends on the persistence of the transformation service, and this depends on enforcing the service provider contract. This problem is outside the control of the identifier manager, as is the maintenance of things identified in general.

R5. DOCUMENT RELATIONS—SCIENCE

The relation between digital objects in the e-Research domain is more elaborate than in e-Learning. Studies use datasets; and beyond being derivative works of other studies, studies can also be responses/follow-ons of other studies, and cite other studies.

- A physicist writes a paper, with identifier 1.2.3/55. This paper uses publicly accessible dataset 1.2.3/636. It is a follow-up to someone else's paper 1.4.2/62, which was based on dataset 1.4.2/12. It cites several more papers, including 1.4.2/62.
- The citation tree for the new paper can be discovered through a citation tracking service operating on the identifiers. But a citation tracking service does not isolate the causal relations between documents: A is a draft of B, a repurposing of B ("based on"), a customisation of B, a response to B. Nor does it tie documents to datasets.
- A relationship service is therefore set up to provide such information. The service allows discovery of the causal chain of documents, so it will publish the fact that 1.2.3/55 responds to 1.4.2/62, and is based on data in 1.2.3/636. (The links between original and response are bidirectional on presentation.)

Note

In e-learning, relations between learning objects are restricted to copy, derivation, redescription; aggregation, disaggregation (Scott Wilson, "tins

of beans” presentation for CETIS). The physics repertoire is drawn from CSMDM (CCLRC Scientific Meta Data Model):

<http://epubs.cclrc.ac.uk/work-details?w=30324> , Section 9.1.1.3. The relations (with suggested FRBR equivalents) are as follows:

- “Use by” relates dataset to paper. As CSMDM points out, datasets can include surveys in the social sciences. FRBR relation: Complement or Supplement. (Complement means one cannot make sense without the other, and is an insert not an add-on; the FRBR examples are musical: librettos and operas, or concertos and cadenzas.)
- “Derived from” applied to datasets is the reverse of “Use by”. Applied from paper to paper, it is like Repurpose/Customise in e-Learning. FRBR: Adaptation.
- “Prior Study” is a Presuppose relation of paper to paper; it is similar to Prerequisite in e-Learning, and is thematically privileged instance of “Cites”. FRBR: Successor (or if you’re being candid, Imitation).
- “Follow on Study” is a Response relation of paper to paper. FRBR: Successor.
- “Parent Study” is not defined by CSMDM.
- There is provision in CSMDM for a “Method” field specifying the link between objects further (presumably how it was derived, though that is not explained).

CSMDM was driven by a JISC requirement for institutional accountability and bibliometrics, rather than bibliographical requirements and discovery like FRBR’s relation model. However bibliometrics are increasingly important, and very useful for discovery anyway.

R6. DATASET RELATIONS

Australia publishes a dataset for electron crystallography. Three years later, Argentina publishes its own dataset which it claims supersedes the Australian data, because the Australians didn’t prevent radiation damage of the proteins in the samples. The claimed relation between the datasets is exposed through a service operating on dataset identifiers. Argentina admits that transmembrane proteins were not severely affected, so studies concentrating on those proteins and using the Australian data may still be valid. The applicability of the claim relation also needs to be exposed.

- A Dataset has a persistent identifier.
- Datasets are discoverable through global identifiers.
- There may be a Supersedes relation between an old and a new dataset if they have the same purpose and domain. This is a version relation, though the authority of the new data may be distinct from the old.

- The claim of Supersedes, or any other relation between datasets, is itself open to query, and originates with a particular authority.
- The claim of Supersedes may be specific to only certain domains; some experiments may still be able to use it without problem. So the versioning relation needs to have metadata not only about the authority making the metadata claim, but also about the applicability of the claim.
- Data objects can be based on other data objects; e.g. aggregation, or calibration established in a previous study.

Note

See discussion under UPDATED DRAFT—KILL ORIGINAL.

R7. DATA HOLDINGS

The CSMDM (CCLRC Scientific Meta Data Model) defines Data Holdings as the collection of data objects supporting a study: "There is one Data Holding (DH) related to each Investigation in the Study. The Data Holding contains the Data Collections & Atomic Data Object (DC&ADO's)." Each Data Collection consists of discreet Atomic Data Objects. Since the data holding can be heterogeneous and is assembled on the spot for different studies, data holdings are an aggregation of objects.

- For a study of electron crystallography, I make reference to two locally managed datasets, one dataset managed by a partner institution, and a couple of atomic data objects purpose-gathered for the study. The datasets are collections of atomic data objects themselves.
- Each atomic data object has a global identifier.
- Data Collections have their own identifiers, and a membership service which reports whether a given object is part of the collection.
- Because data collections and atomic objects are digital objects with identifiers, one-off aggregations of such objects can be created readily through those identifiers.
- Each paper has a collection of data (data holdings) it is based on. This is an aggregation of digital objects, which may be assigned its own identifier.
- Once the data holdings aggregation is defined, it can have attributes assigned it as metadata (through its identifier), possibly with attribute inheritance: "raw/intermediate/final", and search engine ranking.

R8. IDENTIFIER FOR STORAGE RESOURCE BROKERED OBJECTS

Storage Resource Brokers (SRBs) are used in e-Research to provide a virtual file interface, including replicas (appropriate copies), metadata, distributed storage, object-specific authorization, and a separation of logical from physical files. The expected model for access to files in SRBs is via either a data repository or the SRB virtual file system (requiring an SRB client and protocol): the discovery request, browse or search, comes in to the broker, which discovers the resource.

SRB files have each their own GUID (global unique identifier), which are global and survive file location changes, but are repository-specific—they change when the object is uploaded to a new repository. Access to a specific Item requires an SRB URI (http://www.sdsc.edu/srb/index.php/SRB_URI), which contains domains and passwords, encoded either in the URI or as a login profile. Access to a logical name for a resource (Manifestation) still involves a locator—the specific SRB service query on that name. Moreover, logical names can be just as meaningful and situation-specific as filenames in general, leading to known problems.

With a globally actionable identifier, the GUID and the SRB service call invoking it can be encapsulated in the one package. Logical names mapping to files with GUIDs can be mapped to a global, arbitrary identifier. Metadata can be associated with the object outside the SRB system, in addition to the metadata already captured within the SRB. The identifier can also allow direct access to the SRB, potentially using the GUID, rather than mediating it through a repository. The object thereby becomes interoperable with the non-SRB world.

- A digital object, and associated replicas and metadata, are created and uploaded into an SRB.
- An identifier is assigned to resolve to the SRB URI of a particular digital object. This identifier can now be used outside the SRB.
- The digital object maintainer moves the item to a different local repository, possibly in a different SRB federation. The object gets a new GUID by definition. The identifier changes resolution accordingly. The identifier remains persistent, while the GUID does not.
- The object maintainer arranges for the object to be housed in two different SRB federations. The instances can be linked through a service either treating them as resolutions of the same identifier, or as two identifiers connected through as service as manifestations of the same work.
- An identifier is assigned to resolve to the logical name of a digital object.

Note

GUIDs are already persistent and provide appropriate copy deliver, so the main reason to assign them a discrete persistent identifier is access to a global namespace and interoperability outside SRB services.

- Being outside the SRB system, the identifier can interoperate with other, global metadata services, such as a global versioning service.
 - As an example: structural (= resource) metadata about the object (e.g. MIME type) can be discovered from the repository housing the object, or publicly documented object naming conventions, even if that metadata is not available from SRB.
- The identifier can be arbitrary, whether or not the logical name is meaningful. This lets it survive semantic shift.
- An identifier is assigned to an SRB metadata query. This creates a virtual collection of whatever SRB objects happen to match at any given time.
- The query identifier is persistent, even though its content model may shift in time.
- The query identifier is persistent, even though the query protocol or metadata model may shift in time.

R9. DATASET LIFECYCLE MANAGEMENT

APSR survey indicates that not all data sets are guaranteed long-term preservation, because of the appreciable size of resources involved. Datasets are contractually mandated by the funding authority to be made available, but only for a limited time. Once that period expires, a value decision needs to be made on what data is preserved. The decision may not take into account all possible stakeholders (particularly long-tail).

The determination of future use must be robust enough to cope with: changes in location of dataset (including both changes of repository location and move to different repository); changes in location of user; off-line access to the dataset (e.g. use of downloaded copy past dataset expiration, or transfer of downloaded copy to another party). When the decision comes to delete the dataset, a reasonable effort must be made to notify potential stakeholders. An identifier-base scheme only partly addresses the problems

- A researcher accesses a dataset. As part of their initial access, they are asked to subscribe to notifications of changes in the lifecycle of the dataset. If they do not, they risk losing access to the data long-term.
- The subscription takes the form of a mapping between a persistent identifier of the dataset, and a reasonably persistent locator of the user.

- A new stage in the lifecycle of the dataset is initiated. This includes deletion, access restriction, and service enhancement. It would not normally include relocation of the data or the identifier, as a persistent identifier should cope with such changes.
- The subscribers to the dataset retain their association with the dataset, because of the persistent identifier; so they can still be discovered.
- The subscribers are alerted as to the new stage coming up through their persistent locators.



R10. RESEARCH WORKFLOW MANAGEMENT

In e-Research (from a RAMS perspective), research activity takes place within a particular workflow, which can be managed to some extent. A workflow can be a sequence of steps, bound together by having the one goal: an experiment or study. The study in turn is situated in a research environment: one or more hosting institutions, one or more researchers, potentially a virtual organization, one or more research grants or programmes. Each of these entities can be considered virtual objects, and assigned identifiers keyed to metadata. The entities can be placed in hierarchical relations, with inheritance of metadata attributes.

As a result, research workflows can be managed through identifiers. A study, tagged with an identifier, can be identified with a research programme, and the study can be tracked through its research enterprise lifecycle: the deliverables at each stage (grant plan, grant submission, drafts, data collection incl. questionnaires, datasets, outcome dissemination) are associated with the study explicitly, allowing discovery and audit trails. Auditing can include RQF assessment as well as peer review, supervisor review, and institutional review (e.g. ethics). The submission of project outputs to institutional repositories can be substantially automated. The metadata gathered in research workflow management can be exposed to the public as enhanced metadata for the project outputs. Workflows can be flexibly configured and run online, given the metadata associated with the identifiers e.g. staged collaborative analysis and discussion).

- A research enterprise sequence is initiated.
- The sequence is assigned an identifier and exposed by the RAMS server.
- Metadata about the sequence is gathered and recorded; this includes the association of the study with other virtual entities, such as programmes, institutions, researcher profiles, and grants.
- Access profiles are formulated on the basis of the available metadata and entity relations.
- All artifacts generated within the sequence are keyed to the sequence identifier.

- The artifacts involved in the sequence are deposited in a repository.
- An audit of the sequence is possible, involving retrieving from the repository all artifacts associated with the sequence identifier.

© Copyright 2007	Legal 	Privacy		Powered by 
------------------	--	---------	--	---

The PILIN project is funded by the Australian Commonwealth Department of Education, Science and Training, ([DEST](#)) under the Systemic Infrastructure Initiative ([SII](#)) as part of the Commonwealth Government's Backing Australia's Ability – An Innovation Action Plan for the Future ([BAA](#)) under the ARROW Project.