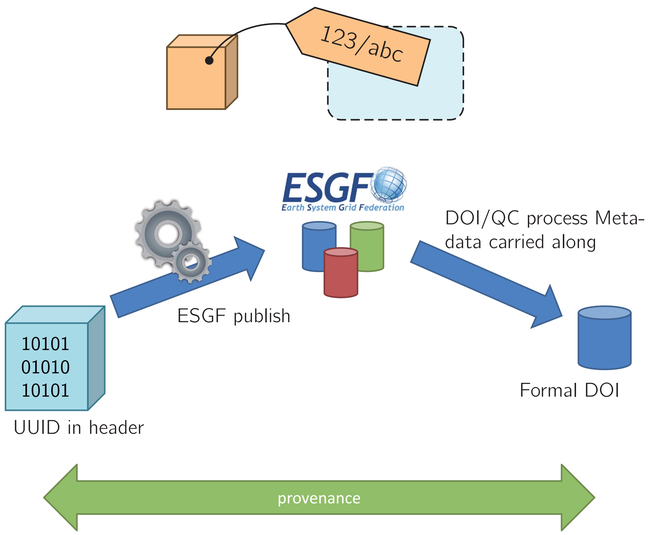
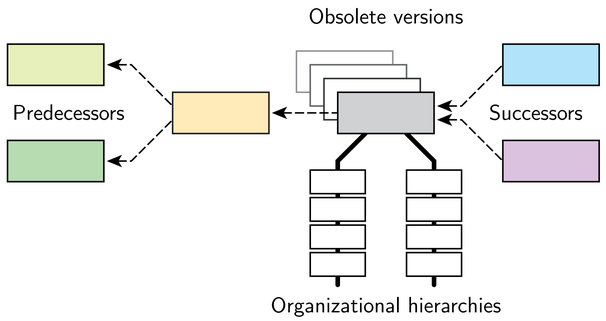
Collections and user tools for utilization of persistent identifiers in cyberinfrastructures

American Geophysical Union Fall Meeting 2014
Poster IN21B-3709 / 21819
Leveraging Enabling Technologies and Architectures to Enable Data Intensive Science I

Tobias Weigel[1,2]
[1]Deutsches Klimarechenzentrum, [2]Universität Hamburg
weigel@dkrz.de



Persistent Identifiers (PIDs) are assigned to objects with unclear preservation status and kept beyond object lifetime. Maintaining a PID also covers essential properties associated with it and relations to other identified objects. Objects occurr in massive numbers and their use is a priori unknown.

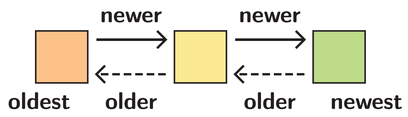


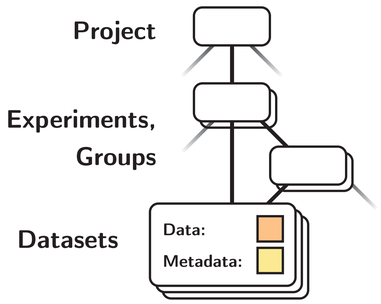
Persistent state information covers dimensions of organizational hierarchies for data, previous and later versions and successors or predecessors in other infrastructures.

Information is structures through typed links or actionable collections, which are PID-based instances of common abstract data types such as lists and sets. In general, state information should be static to minimize maintenance costs.

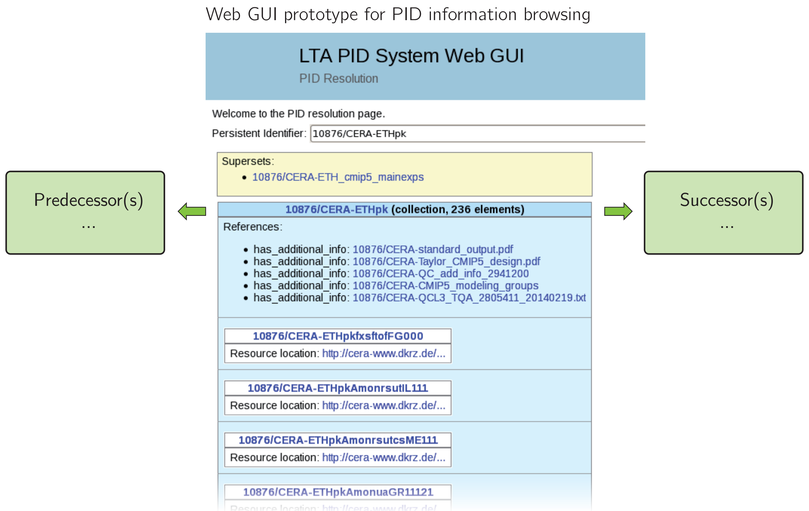
Major PID Information Types API methods

```
GET /pid/{identifier}?filter_by_type=...
  &filter_by_property=...

GET /property/{identifier}
GET /type/{identifier}

GET /peek/{identifier}
```





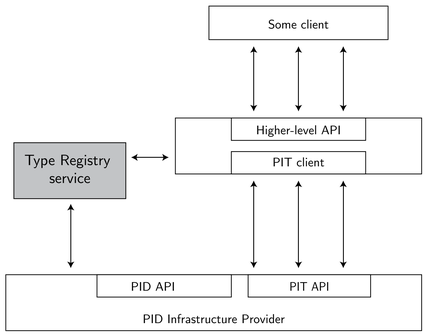
Web GUI prototype for PID information browsing



Access to associated information can be provided possibly independent from particular PID infrastructures through the RDA PID Information Types API. The essential properties providing linkage and other information should be stored in a Type Registry, for which there is a working prototype developed by CNRI and available as an RDA outcome.

The long-term adoption strategy for PID information types relies on PID infrastructure providers offering the unified API and a type registration process. Type registration can be done by individual communities, for which ESGF makes an exemplary case.

PIDs should be assigned to end-products of complex workflows and on ingest to other systems. A provenance exploration tool can visualize the connections between executed workflows, distribution infrastructure and subsequent re-use workflows, providing cross-system provenance traces.



RDA Europe report
on the PID Information Types WG

It is not required to harmonize individual workflow metadata schemas; users should be redirected to existing workflow provenance interfaces as appropriate.

Careful planning must be conducted with policies that ensure that the provenance traces do not break. PIDs to deleted objects must be kept and marked as tombstones. PID assignment does not come at zero costs, yet the maintenance effort must be minimized.

Other possible tools that exploit PID-based information in the context of future ESGF development include:

Version finder service: provides basic information on tombstones and redirects users to latest versions

Collection builder: extends the notion of a download shopping cart by automatic PID collection creation as a surrogate for custom data hierarchy slices

References

T. Weigel, S. Kindermann, M. Lautenschlager (2013): Actionable Persistent Identifier Collections. Data Science Journal, Vol. 12, pp. 191-206. doi:10.2481/dsj.12-058

T. Weigel, M. Lautenschlager, F. Toussaint, S. Kindermann (2013): A Framework for Extended Persistent Identification of Scientific Assets. Data Science Journal, Vol. 12, pp. 10-22. doi:10.2481/dsj.12-036

T. Weigel, T. DiLauro (2013): Separation of Concerns: PID Information Types and Domain Metadata. CAMP-4-DATA workshop, iPRES/DC-2013, Lisbon.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DKRZ
DEUTSCHES
KLIMARECHENZENTRUM