Application of Handles in the European Data Project EUDAT

Frank Toussaint[1], Martina Stockhause[1,2], Tobias Weigel[1,3], Heinke Höck[1], and Michael Lautenschlager[1]

EGU 2013-5475
ESSI 2.4

Persistent Identifiers (PID) as Basis of the EUDAT Infrastructure

Increasing quantities of data lead more and more to automation of data handling. This needs to be closely linked to automated data and metadata handling, as well.
In EUDAT (European Data), a European project for interdisciplinary, collaborative data infrastructures, Handles will serve as persistent identifiers to keep track of data and metadata.

In the safe replication services, processes of creation, movement, and deletion of data objects and their replicas will be tracked and guided by use of PIDs. It was decided that all data objects handled need a PID. Wherever a PID is accompanied by storage information, this will be updated automatically via, e.g., a web service when the data is moved or replicated.

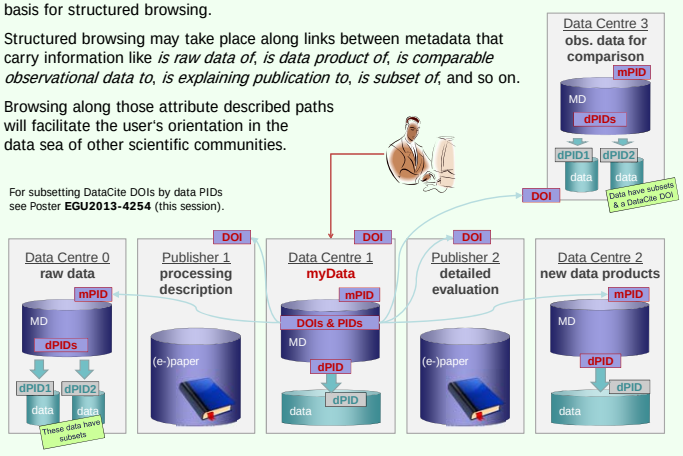
In addition, a normalisation of metadata (MD) structures and semantics can enable the user to get from one to another MD object easily by PID pointers to predecessors, successors, etc.

Surfing metadata via Persistent Identifiers

For cross community data usage mutual search functionality is needed. However, to know what to search, the first step is browsing metadata, to get an idea what data are availabe. As browsing in an unordered list is inefficient, data links by PIDs can be basis for structured browsing.

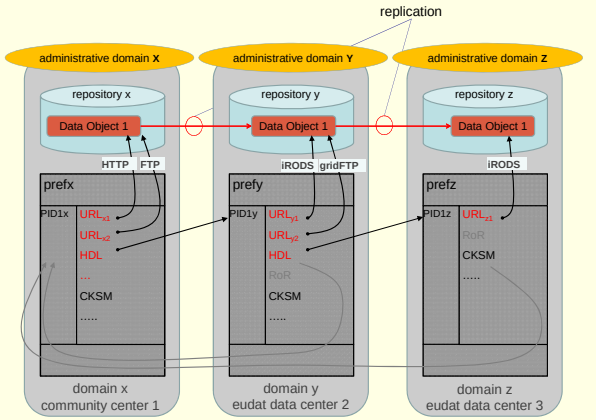
Structured browsing may take place along links between metadata that carry information like *is raw data of*, *is data product of*, *is comparable observational data to*, *is explaining publication to*, *is subset of*, and so on.

Browsing along those attribute described paths will facilitate the user's orientation in the data sea of other scientific communities.



For subsetting DataCite DOIs by data PIDs see Poster **EGU2013-4254** (this session).

PID service as a basic element of the safe replication

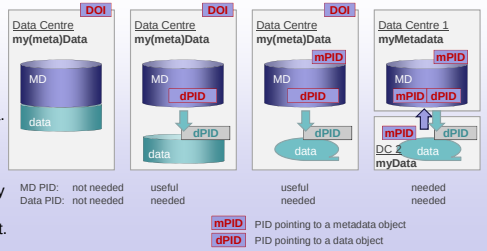


Data replication example in EUDAT

- EUDAT will take iRODS as basis for services like save data replication.
- The internal access to data will be achieved by several different protocols like, e.g., http, iRods, ftp, or gridftp.
- Example replication:
  A detailed replication mechanism relying on PID has been proposed in the EUDAT Project (www.eudat.eu).

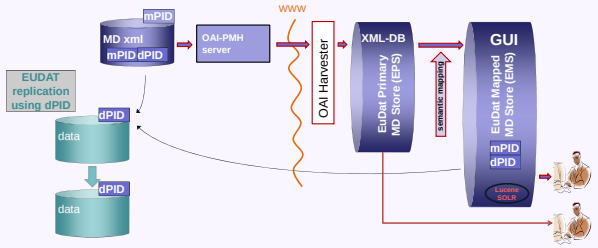
Why Persistent Identifiers for metadata?

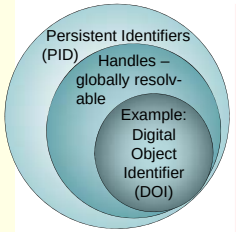
Many of a data object's metadata are kept close to the data. They mostly belong to the so called use metadata like, e.g., descriptions of formats and coordinate systems. However, often there are separate data objects, containing more general metadata. These discovery metadata may comprise general information on content or data producer, references to papers or software, etc. To keep the connection between data and metadata object, both can be linked by mutual pointers held in PID. This requires, that PIDs are assigned to both, data object and metadata object.



Persistent Identifiers in the EUDAT joint metadata domain

In EUDAT metadata from different sources of different communities will be harvested and made available for search. In the EUDAT Primary Store they are kept without specific order with a free text search installed. After a semantic mapping they will be searchable in another graphical user interface (GUI).
During the harvesting process, PID can be fixed to the data. This later can support browsing through the MD along links like predecessor, successor, and other relations that might be coded in the PID data.





More examples for Identifiers

- **URI** – Uniform Resource Identifier – not necessarily globally resolvable
  identifies: anything, consists of printable ASCII
  structure: <scheme>://<authority>/<path>?<query>#<fragment>
- **URN** – Uniform Resource Name – a URI in a defined name space
  identifies: anything, not directly resolvable, example: urn:isbn:3827370191
- **URL** – Uniform Resource Locator – fragile, example: ftp://foo.org/ab.c
  identifies: the (present) location of anything
- **IRI** – Internationalized Resource Identifier – like URI but includes Unicode
- **Purl** – persistent URL of OCLC (Online Computer Library Center)
  identifies: internet resources
- **UUID** – Universally Unique Identifier of OSF (Open Software Foundation)
  identifies a resource, but are not sufficient to locate it
  different versions exist, based on hex codes or readable names

More examples for Handles relevant for publications in Earth System Research

DOI – The Digital Object Identifier (doi.org, for Data Publications: DataCite.org)
- Identifies publications & makes them citable (from Int'l DOI Foundation)

ORCID – The Open Researcher & Contributor ID
- Identifies persons in R&D (from Orcid Inc.)

ISNI (ISO 27729)
- Identifies: persons, legal entities, fictional characters (from ISO, see isni.org)

IGSN - International Geo Sample Number
- Identifies samples of the natural environment (from IGSN e.V., igsn.org)

The remaining question: How to keep the meta data up to date???
- Archives' commitment to updating (like today in case of DOIs) – at least on location changes and deletions ➔ the data object needs to know its own PID!
- Any standardisation and centralisation makes automation easier and facilitates data curation.

[1]World Data Centre for Climate (wdc-climate.de) at German Climate Computing Centre (DKRZ), toussaint@dkrz.de
[2]Max Planck Institute of Meteorology, Hamburg, Germany, [3]University of Hamburg, Hamburg, Germany