Application Examples for Handle System Usage

Frank Toussaint, Tobias Weigel, Hannes Thiemann, Heinke Höck, Martina Stockhause, and Michael Lautenschlager

ID: 1477066
IN23C – 1523
AGU FM 2012

Persistent Identifiers for scholarly data

Why Persistent Identifiers (PIDs) ?

- PIDs to identify & access digital entities independently of the storage location
- PIDs to join data and different levels (& types) of metadata (MD): e.g., discovery MD and use MD
- PIDs to join data sets and data subsets of different granularity
- PIDs to trace data provenance: links to raw data, to a specific version of data producing software (models, error handling, automatic calibration,...), to processing descriptions
- PIDs to organise data handling, e.g., data replication or long term archival

What are a Handle's characteristics ?

- Unique, non-ambiguously discoverable/resolvable, updateable and updated resource pointer, persistent independently of the resource
- Open system (like Domain Name System), easy access

Further general information on PID: see Poster IN23C – 1524 (Weigel et al.)

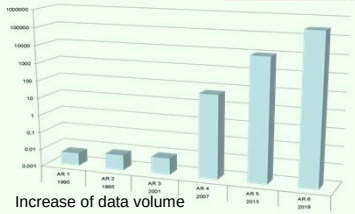
The World Data Centre for Climate

...is hosted by Deutsches Klimarechenzentrum GmbH (DKRZ) in Hamburg, Germany.

WDCC offers climate model data for web download to the world wide scientific community.

DKRZ's mission is to provide computing and storage capacities for the German Earth System Science Community.

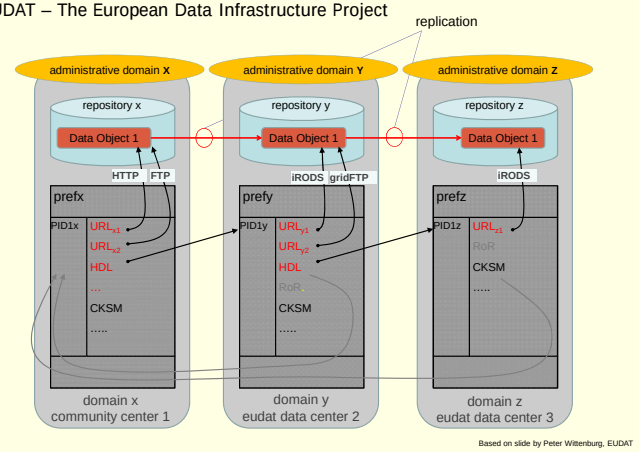
The DKRZ and the High Performance Storage System (HPSS)



Increase of data volume
for the IPCC Assessment Reports (partly estimated)

EUDAT – An example for Handle usage

EUDAT – The European Data Infrastructure Project



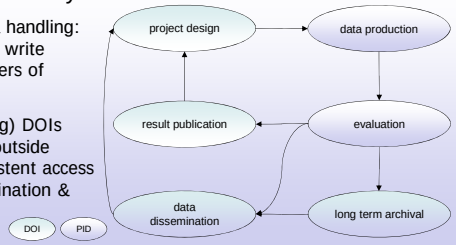
Data replication in EUDAT

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

Handles in the data's life cycle

- PID for efficient data handling: performant read and write access to high numbers of data identifiers
- DataCite (datacite.org) DOIs as identifiers in the outside world: reliable, persistent access e.g., for data dissemination & citations

General PID and DOI in the Data Life Cycle



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 Identifiers

- Identifies publications and makes them citable
  (from the International DOI Foundation)

ORCID – The Open Researcher & Contributor ID

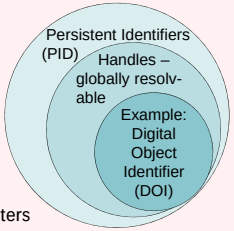
- Identifies persons in R&D (from Orcid Inc.)

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)



PIDs in Earth System Science Projects: see Poster IN23C-1525 (Stockhause et al.)

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.

toussaint@dkrz.de, World Data Centre for Climate (wdc-climate.de)
at German Climate Computing Centre (DKRZ, dkrz.de)