Persistent Identifiers in Earth science data management environments

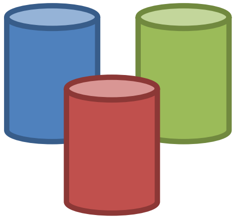
PIDs for ESGF

EGU 2014 GI-0
02 May 2014

Tobias Weigel, Martina Stockhause, Michael Lautenschlager
Deutsches Klimarechenzentrum (DKRZ)

PID usage is driven by two needs.

1. Users want to precisely reference data
2. Management of different versions and replicas by node managers



User needs in ESGF

- Refer to a specific subset of data
  - slices across one or several simulations
- There is typically no single hierarchy.

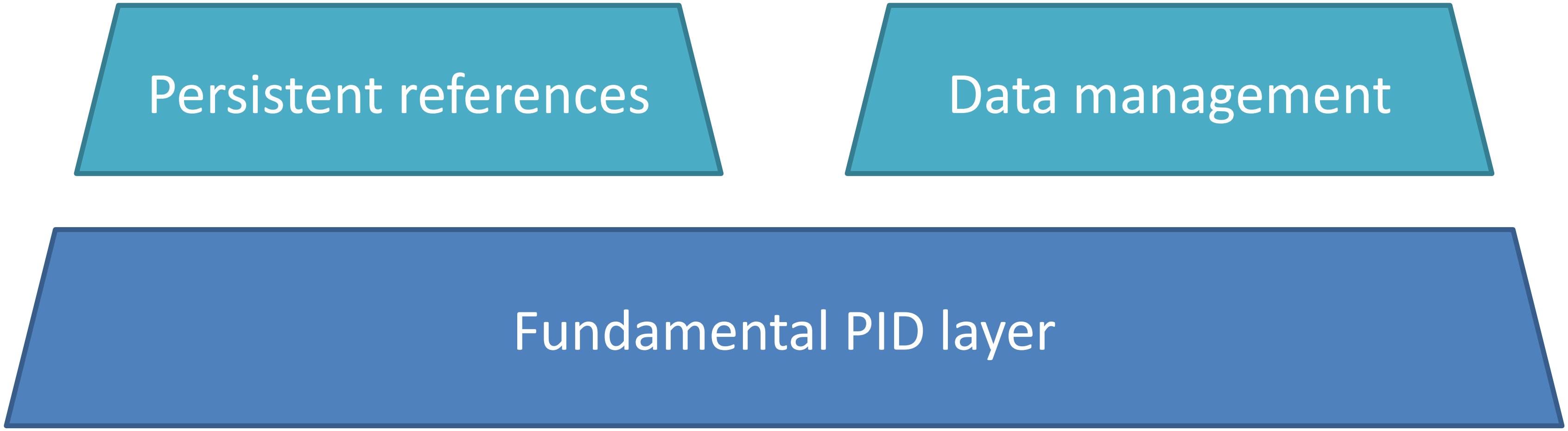
- Not to be confused with citation via a DOI.
  - prior to late QA stages and formal publication

Needs of node maintainers

- Competing and incoherent identification mechanisms in use

- Improved communication
- Improved version control
- Support in case of replication failures
- ...

Motivations differ, yet there is a common layer.



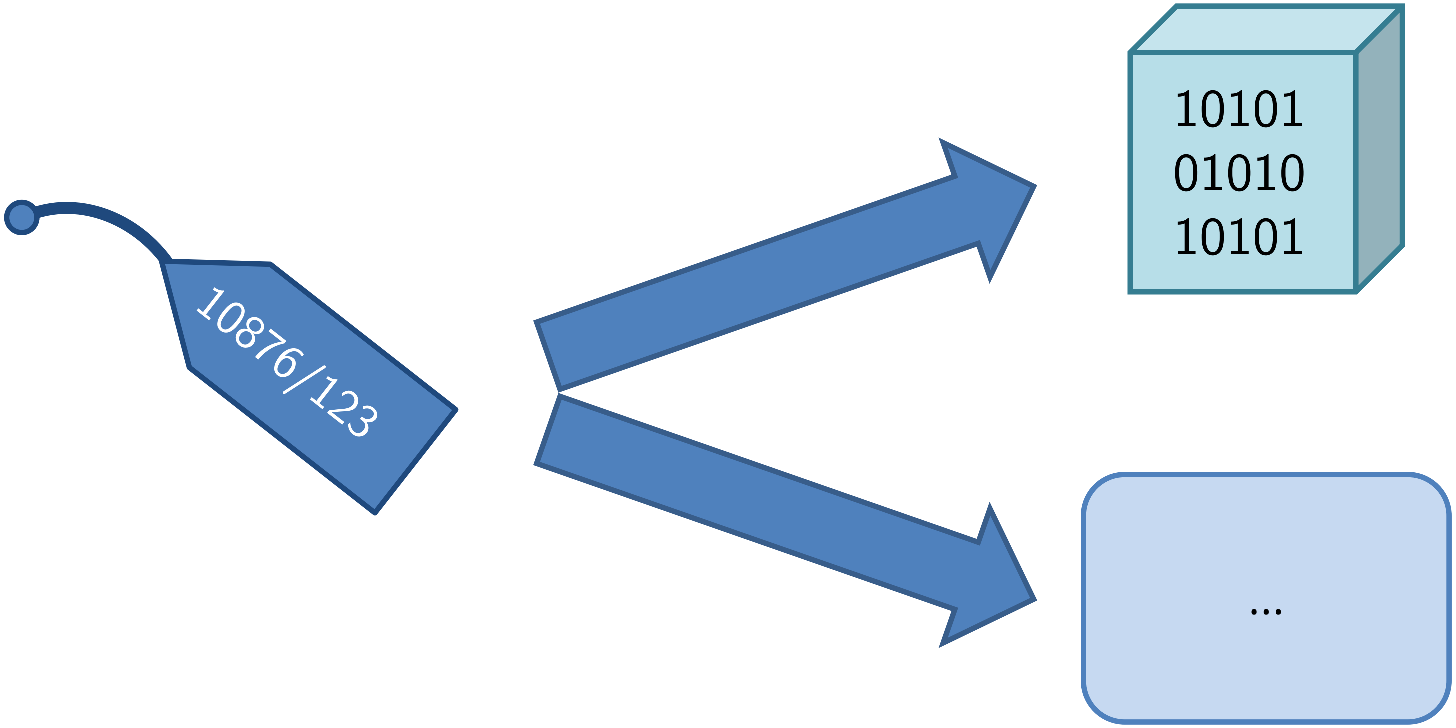
Persistent references

Data management

Fundamental PID layer

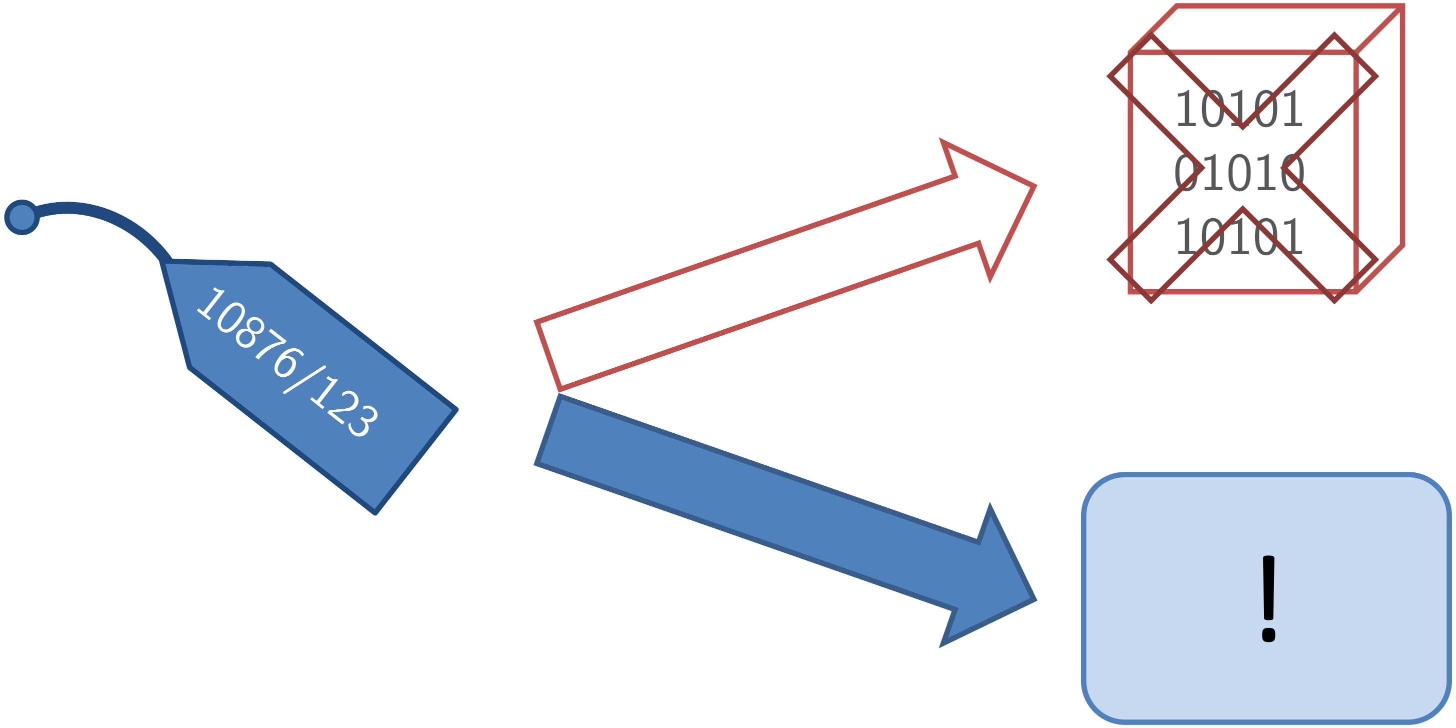
Persistency of identification

- A persistent identifier can be resolved to meaningful state information for at least as long as the resource exists.



Persistency of identification

- A persistent identifier can be resolved to meaningful state information for at least as long as the resource exists.



PID Information Types

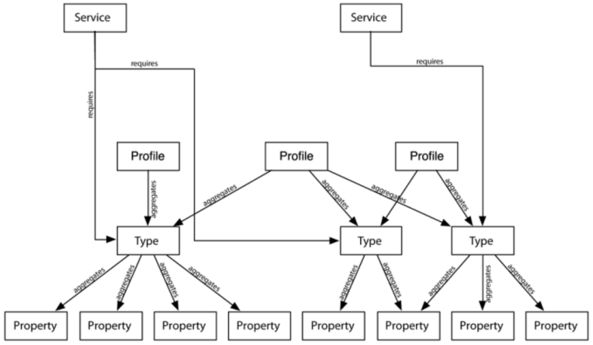
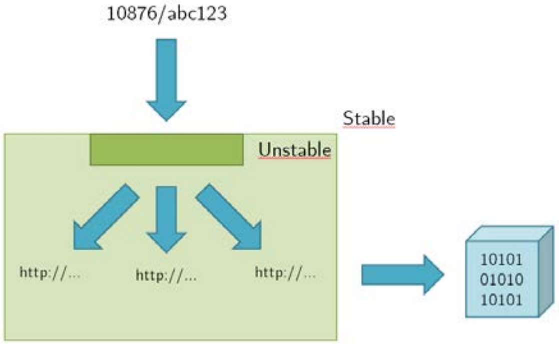
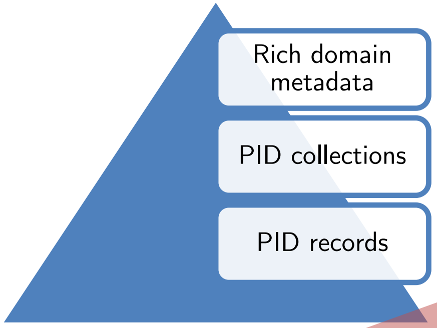


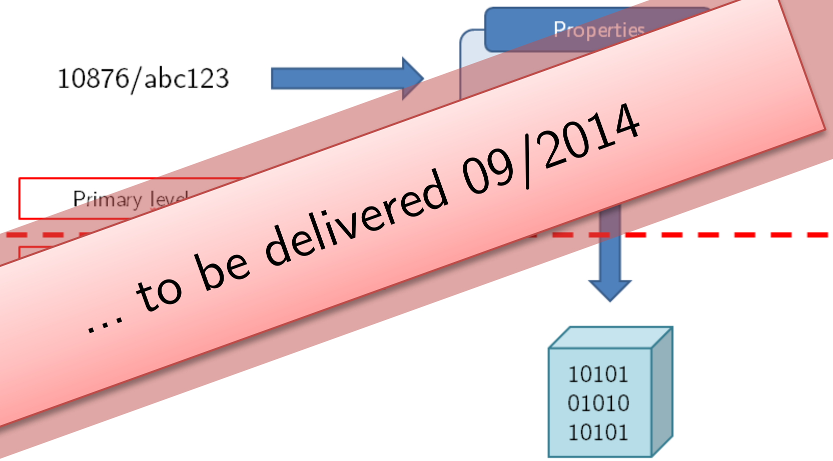
diagram courtesy of Timothy DiLauro (JHU)

http://bit.ly/1fSL78t

```
getProperties()
getAllProperties(PID)
getPropertiesOfType(PID, typeID)
getPropertyValue(PID, propertyName)
describeType(typeID)
doesPIDconformToType?(PID, typeID)
writeFullPIDrecord(PID, dict)
registerType(properties, ...)
createPIDaccordingToType(typeID,
PID, ...)
...
```







... to be delivered 09/2014

What does this mean to ESGF?

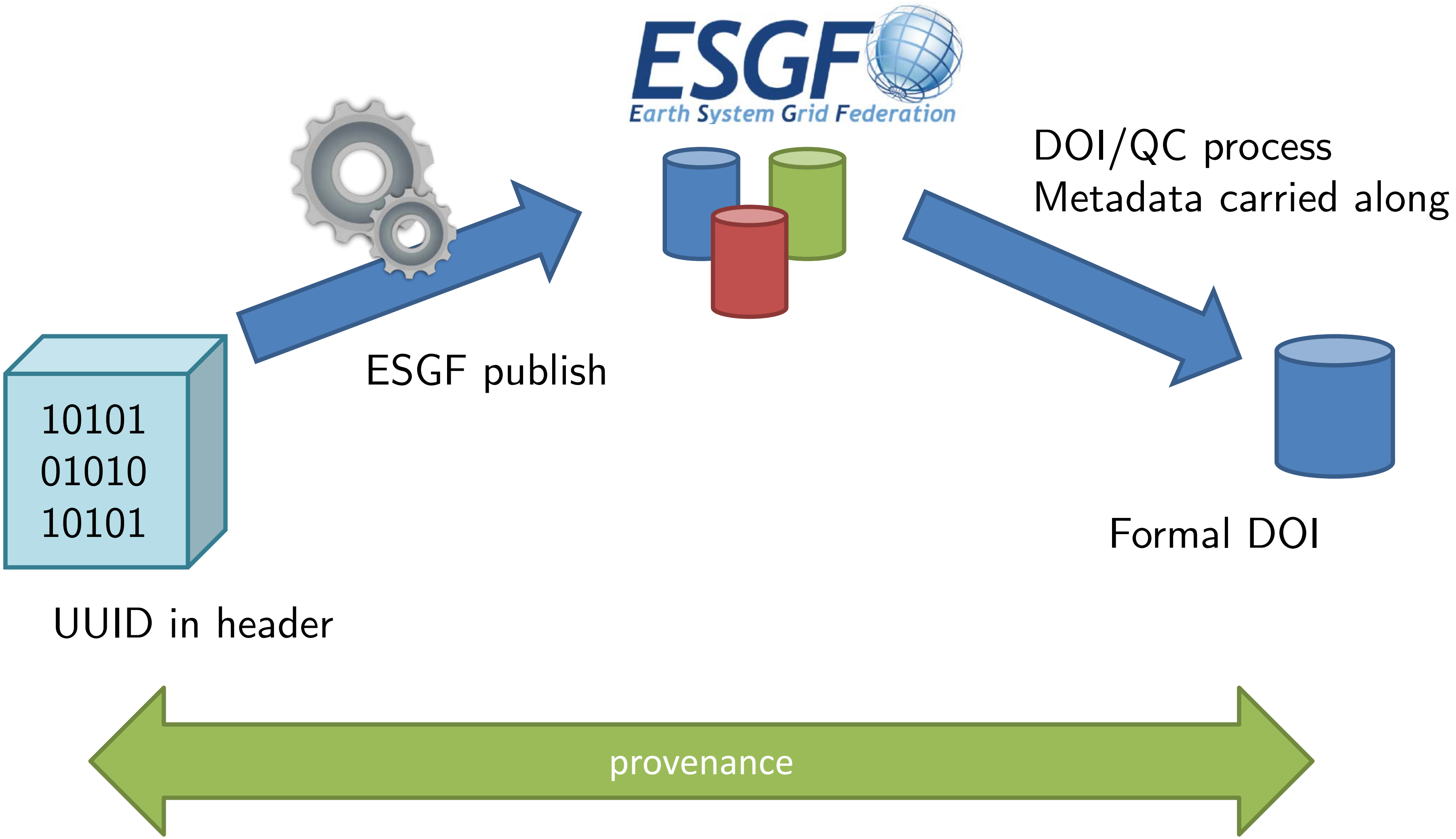
We have some prior experience.

- Existing experience from EUDAT services
- PID federation lessons learned from running distributed Handle Server nodes (EPIC)
- Some first experiments with PIDs and collections for CORDEX

How can we assign PIDs at an early stage?

- Nodes are not allowed to modify data.
- Write UUID in netcdf header during CMOR process
  - establish structure to minimize UUID collisions
- On ESGF publish: mass-register PIDs with name based on UUID

Possible PID assignment process in ESGF



UUID in header

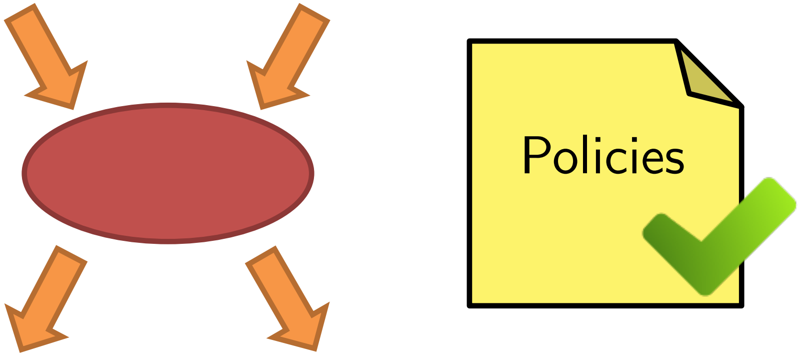
ESGF publish

DOI/QC process
Metadata carried along

Formal DOI

Next steps

- Continuing implementation and prototyping
  - particularly for CORDEX
- Agree on solid mechanisms to ensure proper identifier usage



- Detailed concepts open for discussion at next GO-ESSP meeting

Conclusions

- PIDs can address identification issues within ESGF
- There are many potential downstream use cases
- Range of previous work from concepts to practical experience
- Some costs involved in terms of QA
- Detailed concepts to be developed closely with ESGF developer community

- Thank you for your attention.

DKRZ DEUTSCHES KLIMARECHENZENTRUM

ESGF Earth System Grid Federation

RDA RESEARCH DATA ALLIANCE

UHH Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Weigel, Lautenschlager, Toussaint, Kindermann (2013): A Framework for Extended Persistent Identification of Scientific Assets. Data Science Journal, Vol. 12, pp 10-22. http://dx.doi.org/10.2481/dsj.12-036
- Weigel, Kindermann, Lautenschlager (2014): Actionable Persistent Identifier Collections. Data Science Journal, Vol. 12, pp. 191-206. http://dx.doi.org/10.2481/dsj.12-058
- Toussaint, Stockhause, Weigel, Höck, Lautenschlager (2013): Application of Handles in the European Data Project EUDAT. EGU General Assembly, EGU 2013-5475