DKRZ
DEUTSCHES
KLIMARECHENZENTRUM

Web Services and Handle Infrastructure

WDCC's[1] Contributions to International Projects

Gregory Foell, Tobias Weigel, Frank Toussaint, Stephan Kindermann, Michael Lautenschlager

ExArch[2]: Climate analytics on distributed exascale data archives

Motivation

Climate science demands on data management are growing rapidly as climate models grow in the precision with which they depict spatial structures and in the completeness with which they describe a vast range of physical processes.

This project explores the challenges of developing a software management infrastructure which will scale to the multi-exabyte archives of climate data which are likely to be crucial to major policy decisions in by the end of the decade. Support for automated processing of the archived data and metadata will be essential.

Climate Data Operators (CDO)

CDO[3] is a collection of command line Operators to manipulate and analyse Climate and NWP model Data.

The ExArch approach

The ExArch Web Processing Service brings CDO to an exascale archive.

Features & Benefits

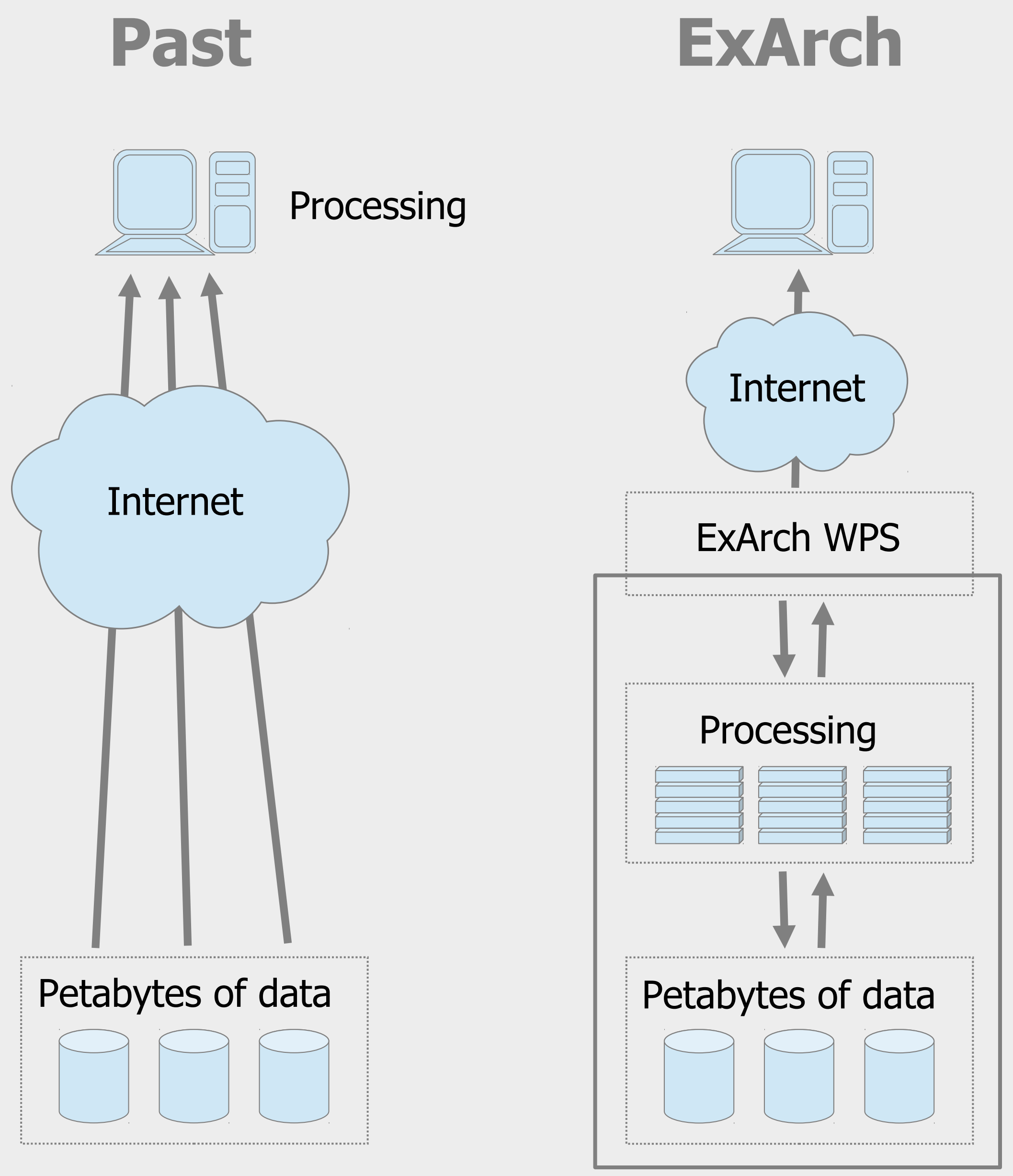
- Flexible transparent and uniform task management due to standardised Web Processing Services.
- Performance enhancement in consequence of distributed computing using dedicated hardware and fast local data access.

Partners

Centro Euro-Mediterraneo per i Cambiamenti Climatici
University of Toronto
British Atmospheric Data Centre
UCLA
Institut Pierre Simon Laplace
DKRZ Deutsches Klimarechenzentrum
Princeton University

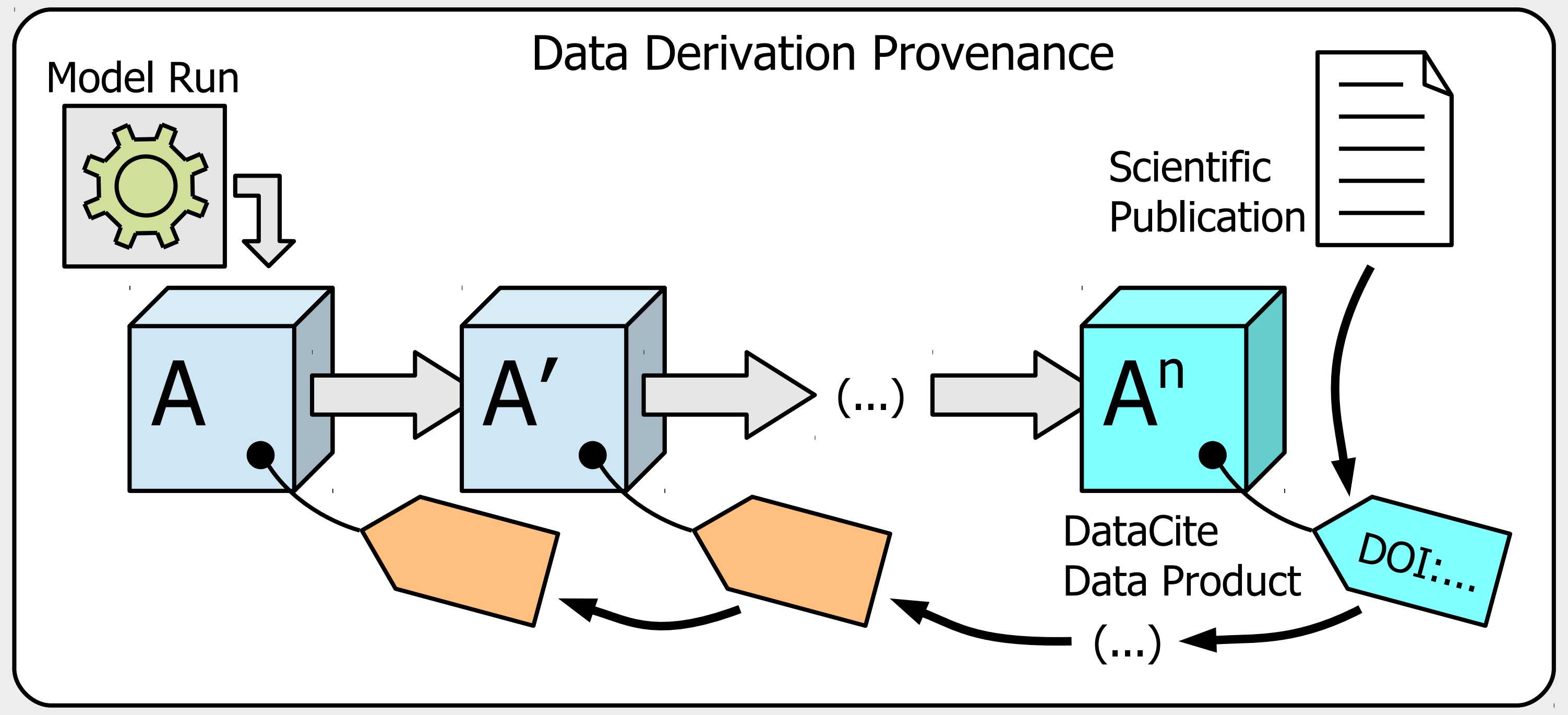
Quality control

Quality control will become increasingly important in an exascale computing context. Researchers will be dealing with millions of data files from multiple sources and will need to know whether the files satisfy a range of basic quality criteria. Hence ExArch will provide a flexible and extensible quality control system.



[2] http://proj.badc.rl.ac.uk/exarch
[3] https://code.zmaw.de/projects/cdo

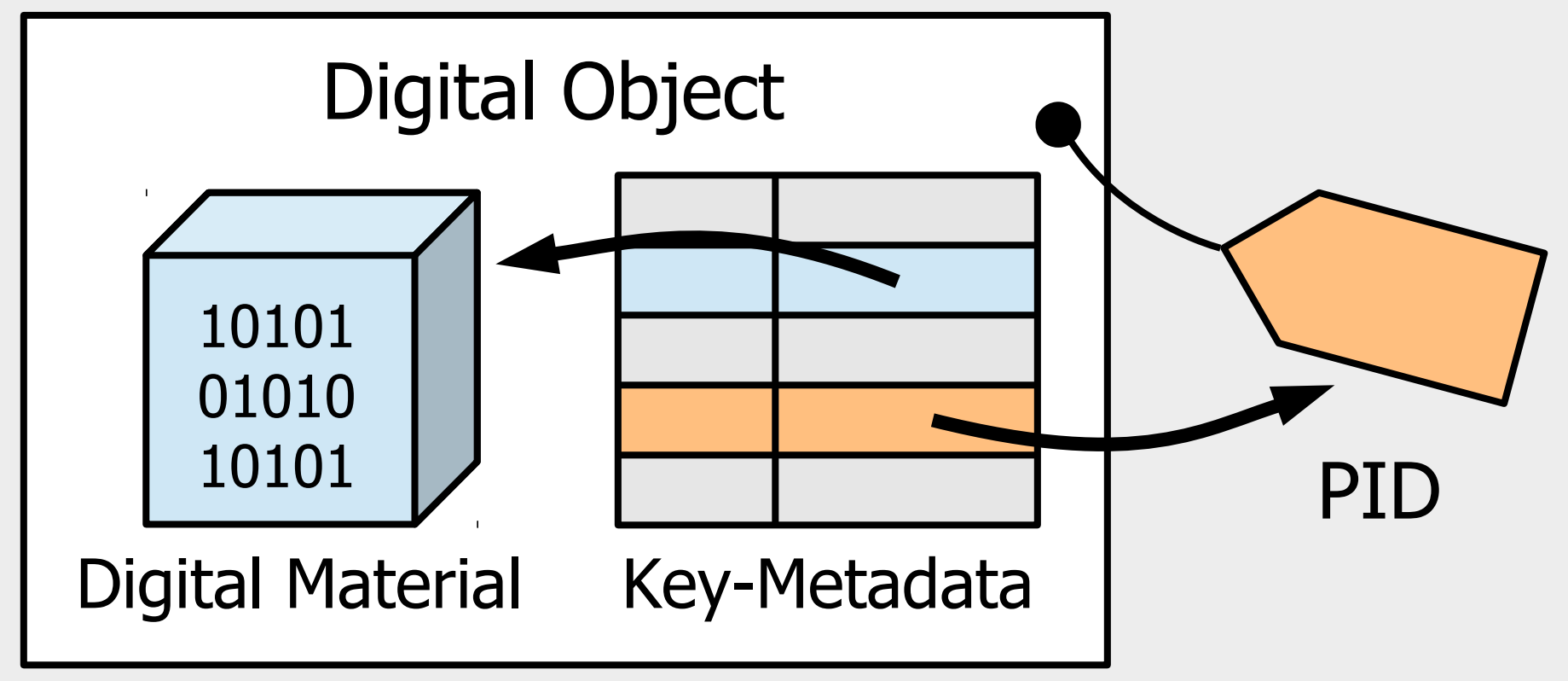
Cross-project persistent identifiers for provenance tracing



Motivation: Capturing data derivation provenance

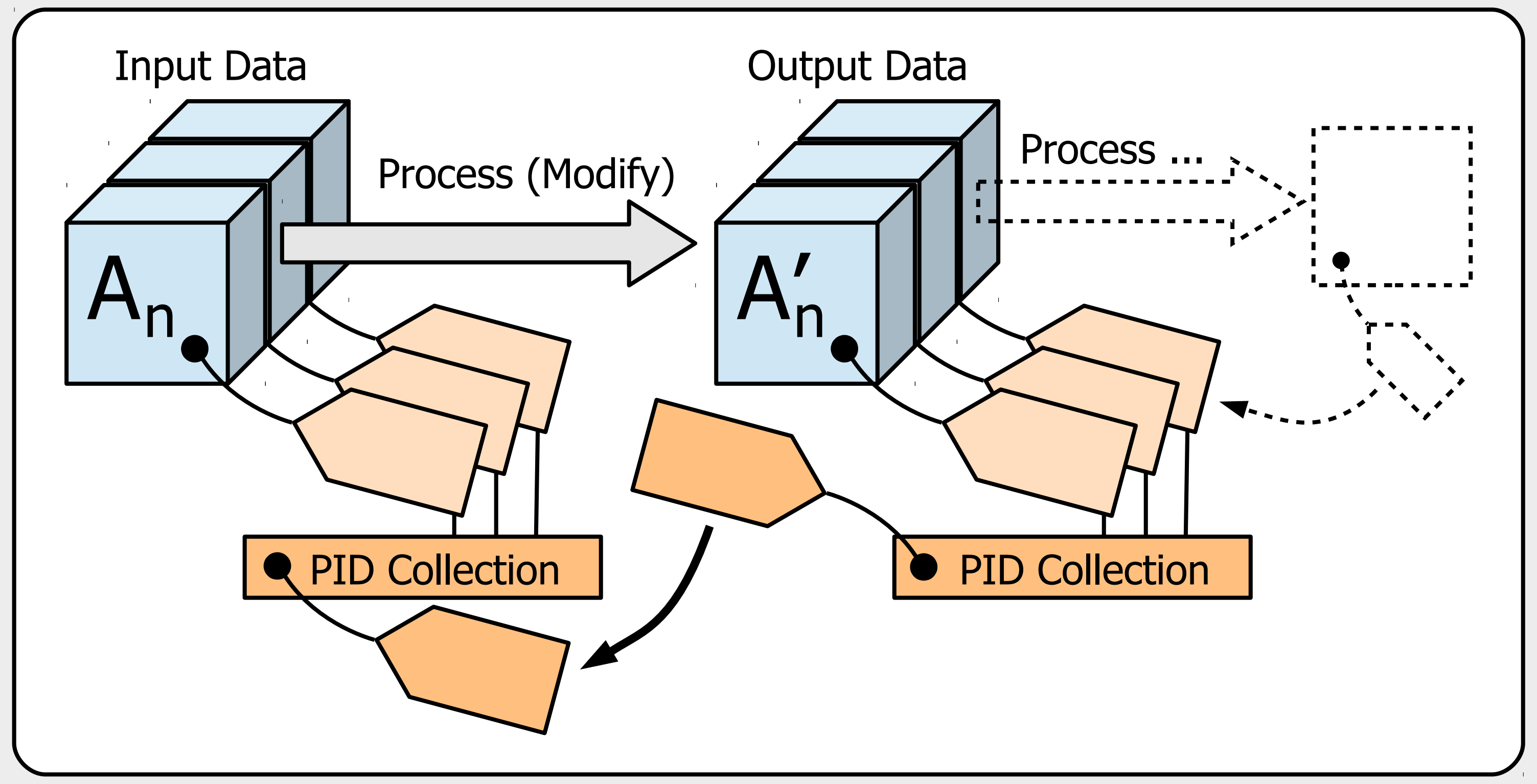
The motivation is to record data derivation provenance for purposes of scientific verifiability and correct data usage by downstream data users, preferably in a way that does not require an end-user scientist's interaction.

Digital Objects (Kahn & Wilensky 2006[4]) are logical constructs consisting of digital material and key-metadata, which includes a globally unique *handle* or persistent identifier (PID). The Handle System® is a distributed database for storing key-metadata and resolving handles.



Method

The upcoming persistent identifier (PID) infrastructure of EPIC can be used to record steps of data derivation provenance as part of PID annotations. All information required to describe a data product's provenance must be available from the PID infrastructure and independent of external systems. It must therefore be recorded within the bounds of PID key-metadata. In contrast to the original Digital Object theory, key-metadata must be maintained even if data products are deleted.



Linking processed data through PID collections

ExArch processing typically works on time series which consist of a large number of individual files to produce a number of output files. All of these receive their own PIDs, but are also aggregated in two PID collections, which are then linked together to formalize the provenance. PID collections are realized completely within the PID infrastructure through key-metadata fields. The necessary PID operations can be performed by either an extended CDO or the WPS.

If further processing takes place, the PID chain is extended accordingly, limited to subsets where applicable. Since the PID infrastructure is independent of project-specific infrastructures, provenance may even be recorded if the processing takes place outside ExArch. PID creation and maintenance is an infrastructural service that does not require end-user interaction.

Outlook

This poster presents ongoing work which is not only relevant to ExArch, but also to EPIC and other projects. Manifold applications for PIDs exist across projects and even domains, yet the underlying infrastructures are only just being set up now. Enhancing widely used tools such as CDO provides possibilities for provenance tracking outside ExArch.

[4] R. Kahn, R. Wilensky: A framework for distributed digital object services. Int. J. Digit. Lib., Vol. 6, No 2. (2006), pp. 115-123, doi:10.1007/s00799-005-0128-x
[5] European Persistent Identifier Consortium, http://www.pid-consortium.eu

[1] World Data Centre for Climate at Deutsches Klimarechenzentrum,
Bundesstraße 45a, 20146 Hamburg

http://www.dkrz.de
foell@dkrz.de weigel@dkrz.de

XL145
Abstract identification number: EGU2012-9124