Identifier Infrastructure Usage for Global Climate Reporting

IoT Week 2017, Geneva

Tobias Weigel
Deutsches Klimarechenzentrum (DKRZ)
World Data Center for Climate (WDCC)

Scientific driver: Global climate modelling

https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6

Scientific driver: Global climate modelling

- Operational phase ca. 2017-2021+

- Community-driven, aligned with IPCC AR6

- Global data volume in order of 100-250 PB
  - full replication impossible!

Eyring, Bony, Meehl, Senior, Stevens, Stouffer, Taylor: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geosci. Model Dev., 9, 1937-1958, 2016. doi:10.5194/gmd-9-1937-2016

The climate data life-cycle
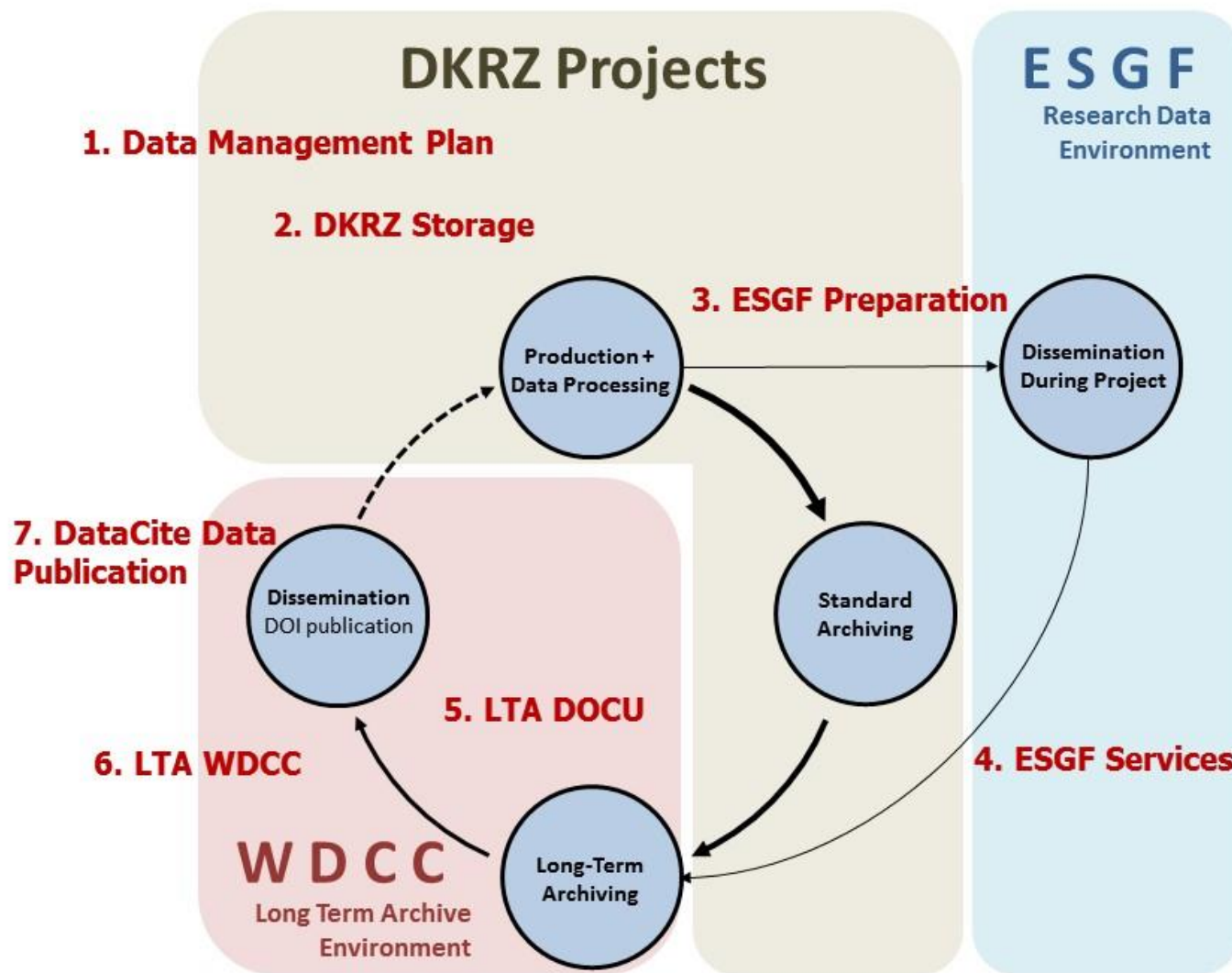
M. Lautenschlager

The Earth System Grid Federation

http://esgf.llnl.gov
http://esgf-data.dkrz.de
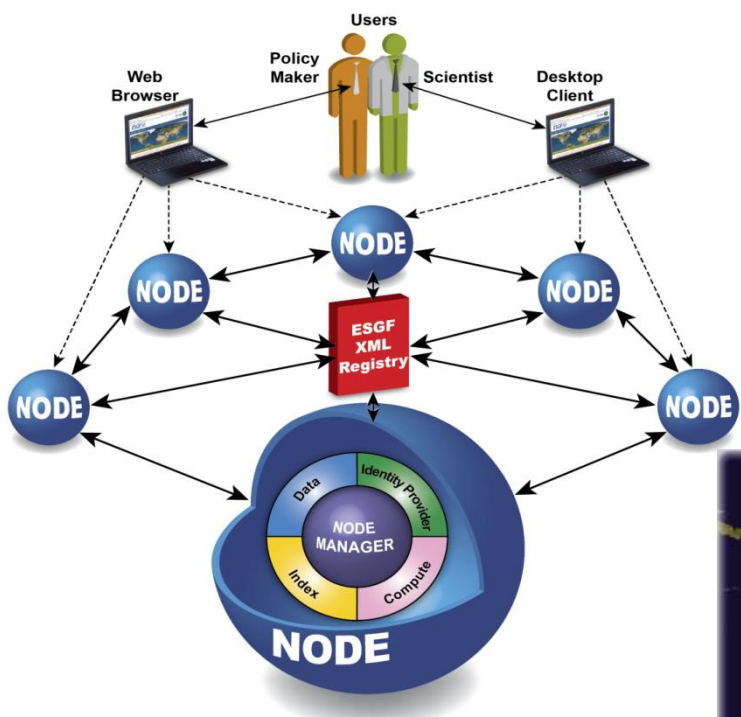
D. Williams (LLNL); U.S. DOE 2017. 6th Annual Earth System Grid Federation Face-to-Face Conference Report. DOE/SC-0188. U.S. Department of Energy Office of Science

DKRZ technical infrastructure and ESGF
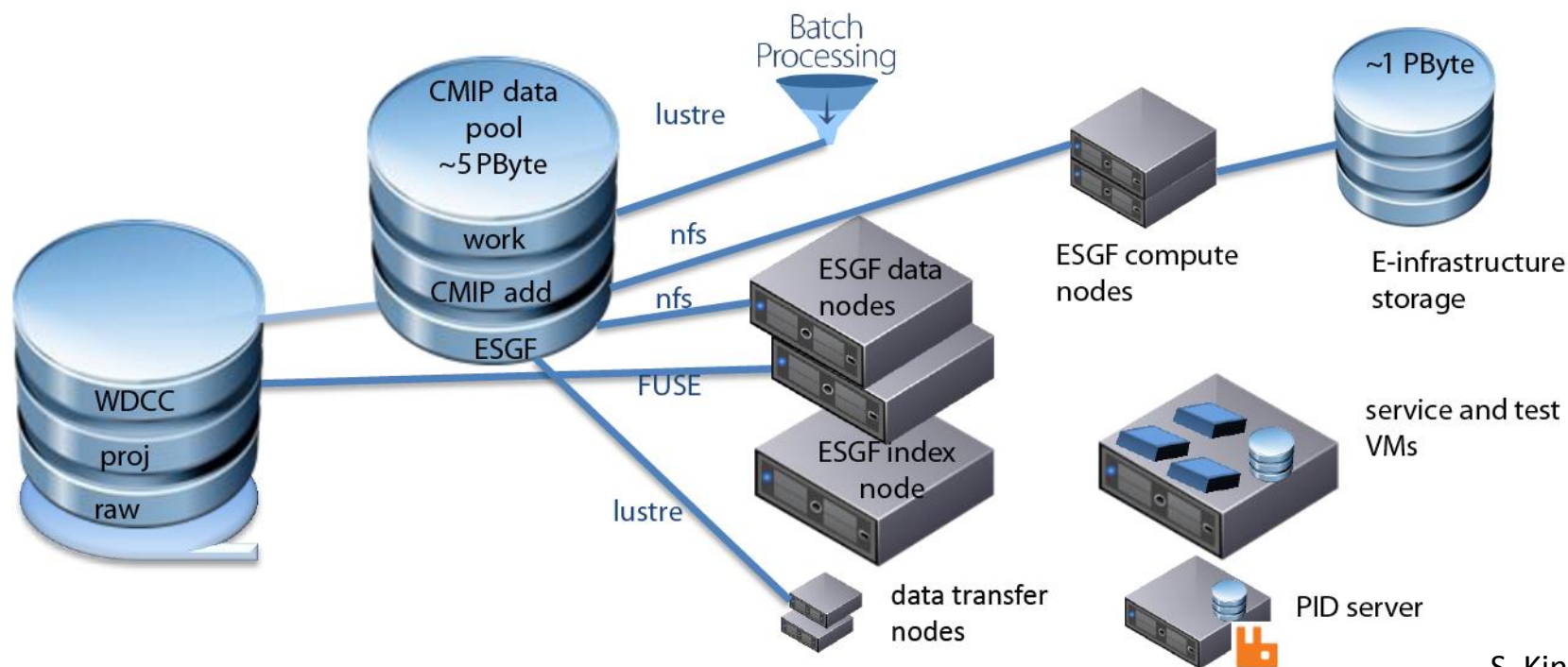
S. Kindermann

Making it scalable requires additional effort

The PID registration and metadata update tasks are pushed to a message queueing system facilitating high availability and scalability...

...and then processed asynchronously.
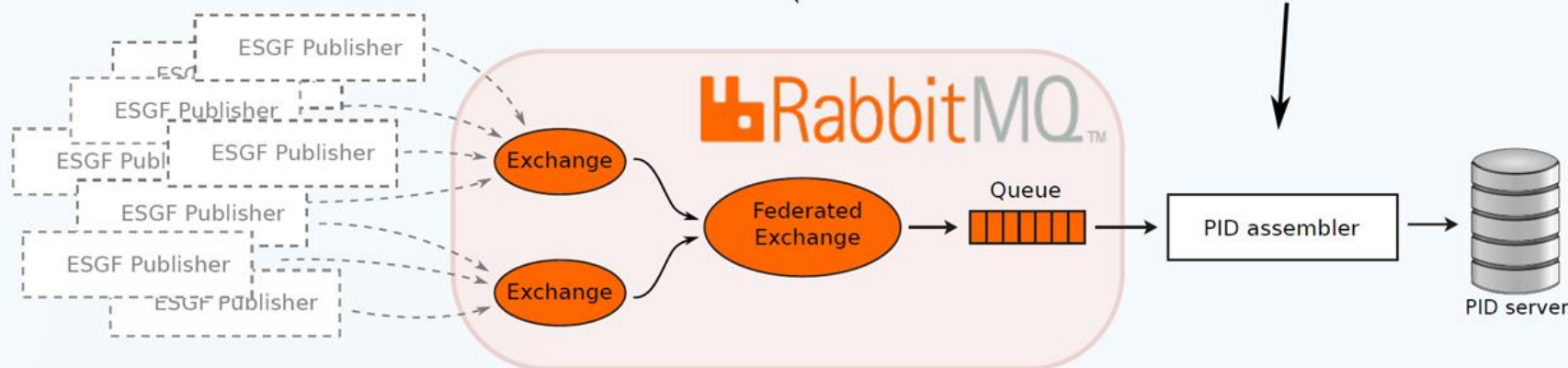
Buurman, Weigel, Juckes, Lautenschlager, Kindermann:
Persistent Identifiers for CMIP6 in the Earth System Grid Federation,
EGU 2016

Properties stored in Handle records for ESGF

Files	Datasets
URL	URL
aggregation_level	aggregation_level
url_replica	replaced_by
tracking_ID	replaces
checksum	errata_IDs
is_part_of	has_parts
DRS_ID	DRS_ID
file_size	
file_name	

Kernel Information Profiles

Why do we care? What is the long-term strategy?
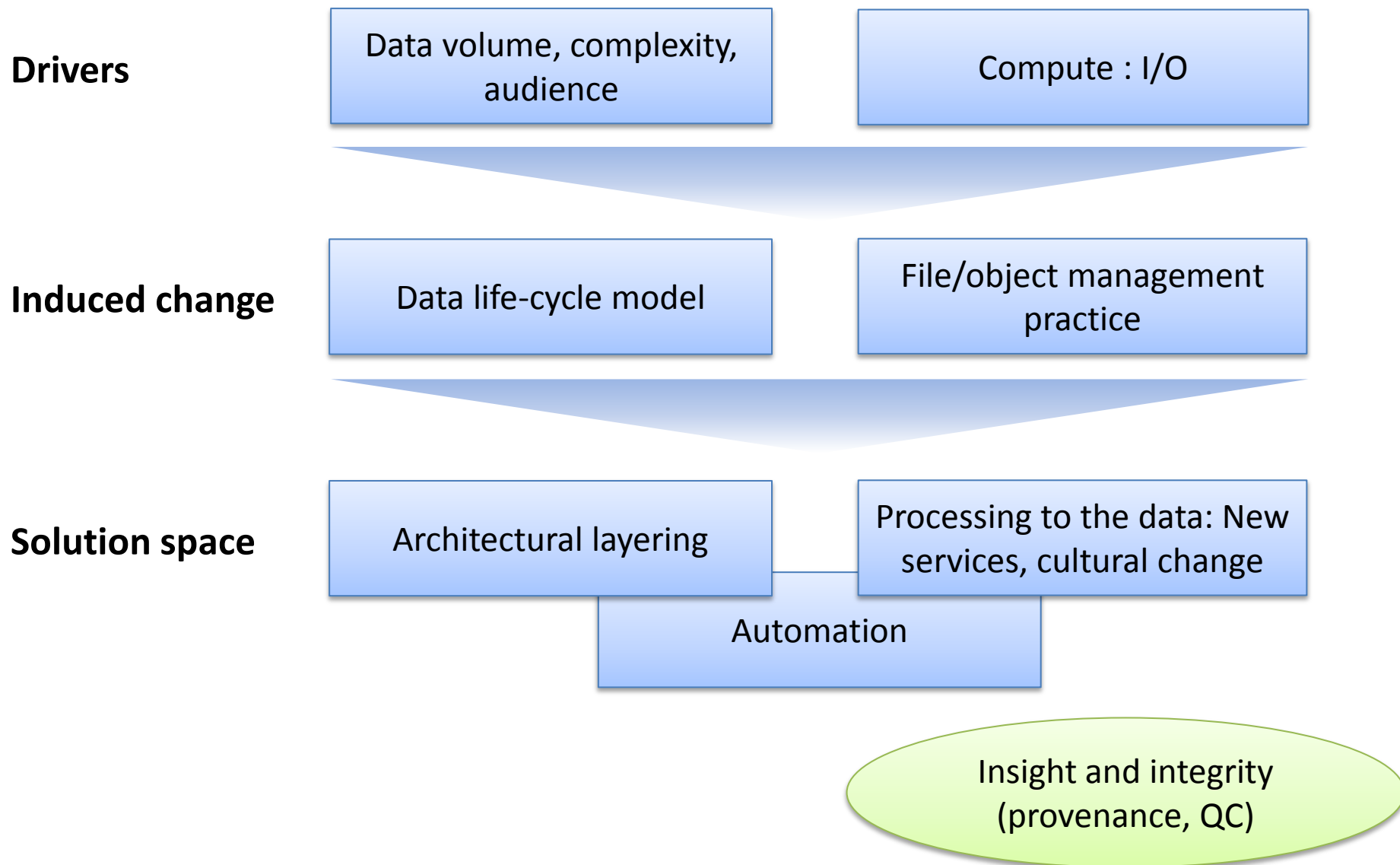
Drivers

| Data volume, complexity, audience | Compute : I/O |

Induced change

| Data life-cycle model | File/object management practice |

Solution space

| Architectural layering | Processing to the data: New services, cultural change |

Automation

Insight and integrity (provenance, QC)

The users' reality...

Type-Triggered Automated Processing (T-TAP)

netCDF-Files

Collection

<Metadata>
(xml)

?
(third-party input)

script

Processing service
(WPS)

Agent

output

multiple types, e.g. netcdf, xml, linked data, text reports, PROV record

✓ well-defined ways to publish it (automatically)
✓ possible repacking into a new collection

described in DTR
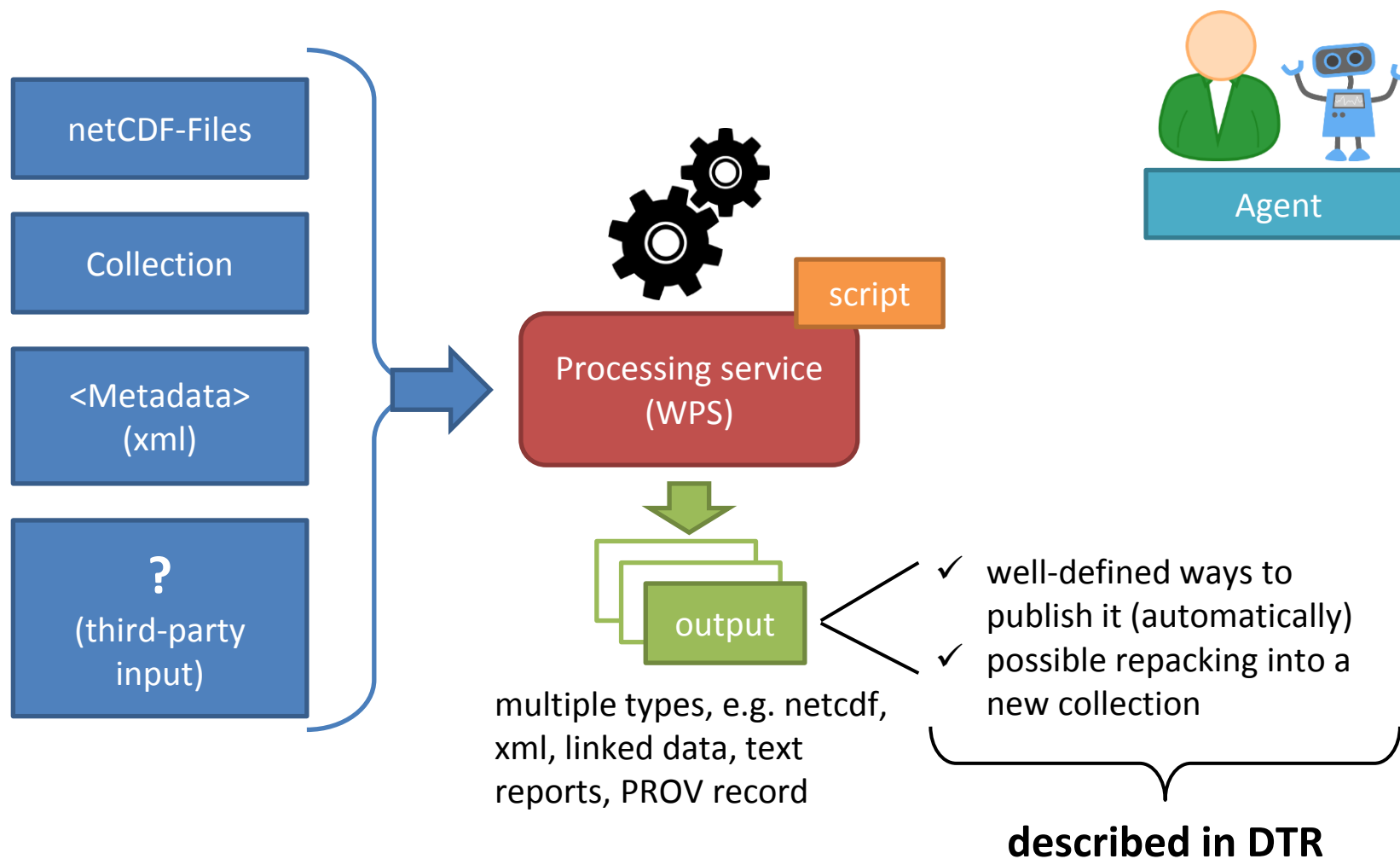
Data processing perspectives

- Climate data analytics service (for EOSC)
  - Cluster-based, 2 pilot implementations, 2018+
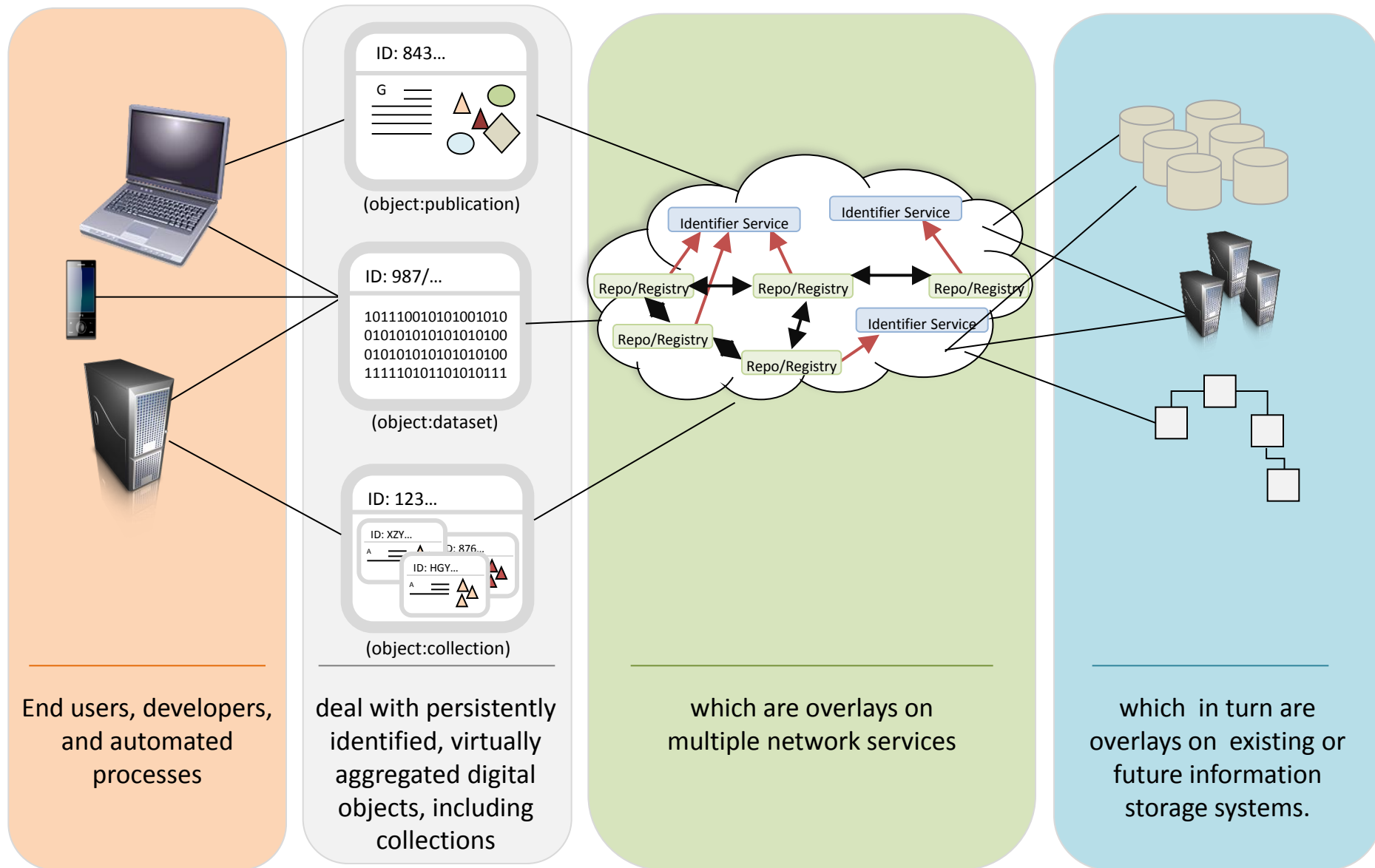

- Copernicus Climate Change Service (C3S)
  - coordinated by ECMWF, operational 2018+
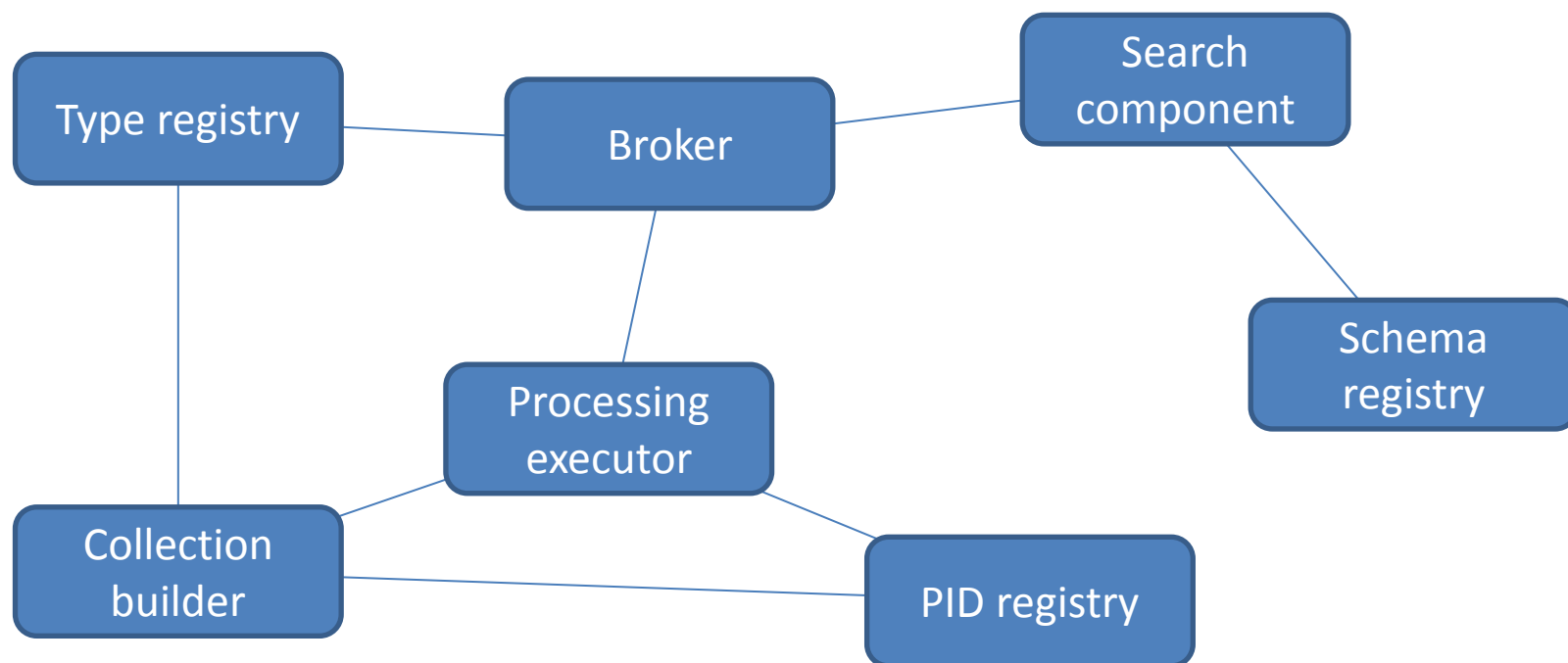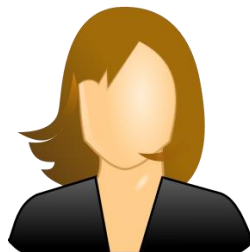- WPS-based service ecosystem with multiple deployments

Global Digital Object Cloud (GDOC)

ID: 843…
(object:publication)

ID: 987/…
101110010101001010
010101010101010100
010101010101010100
111110101101010111
(object:dataset)

ID: 123…
ID: XZY…
ID: 876…
ID: HGY…
(object:collection)

Identifier Service

Identifier Service

Identifier Service

Repo/Registry

Repo/Registry

Repo/Registry

Repo/Registry

Repo/Registry

End users, developers, and automated processes

deal with persistently identified, virtually aggregated digital objects, including collections

which are overlays on multiple network services

which in turn are overlays on existing or future information storage systems.

L. Lannom / DFIG

GDOC and reusable data service components

Thank you for your attention.

weigel@dkrz.de