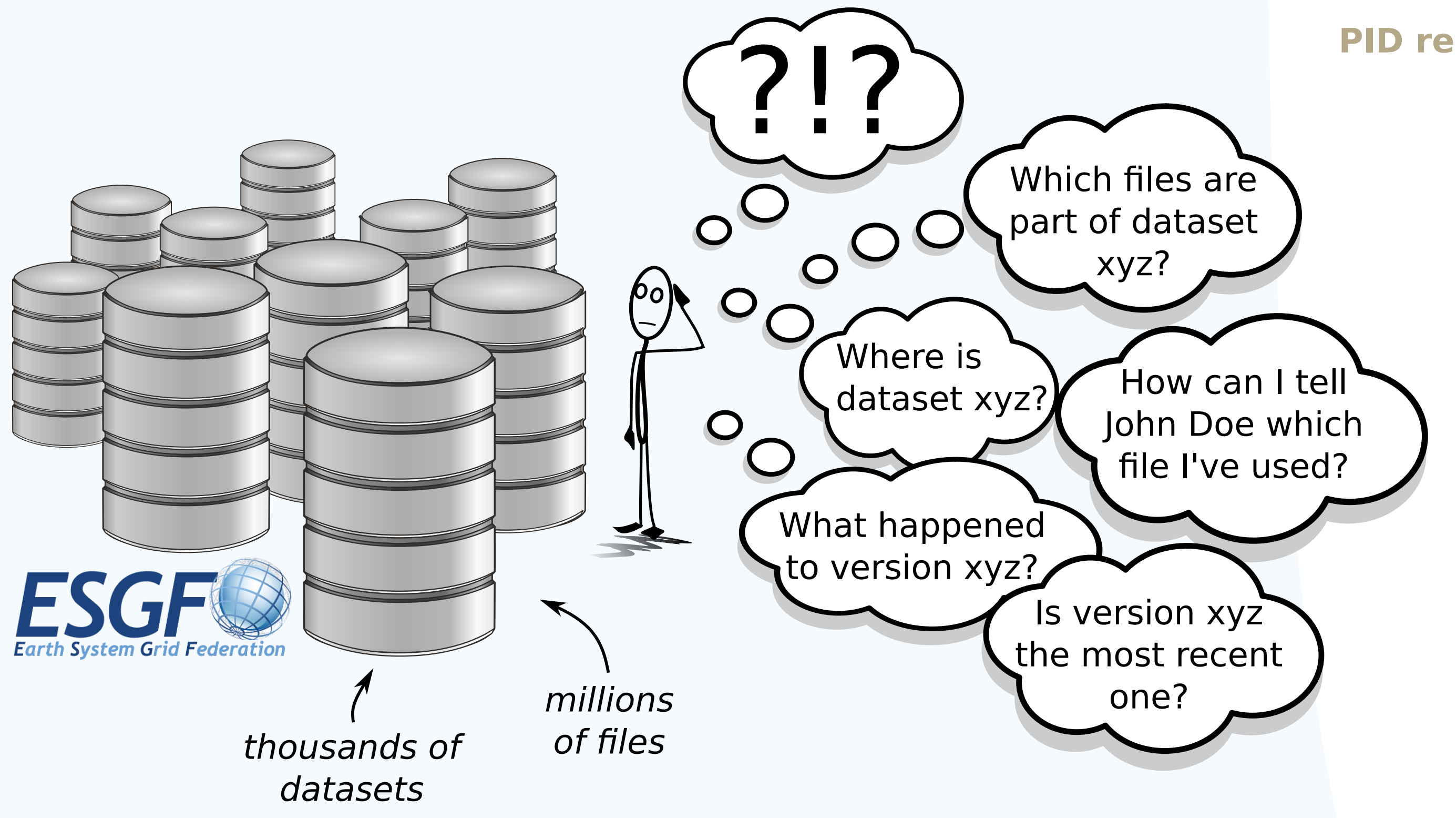
EGU2016-13297
ESSI 2.6

Persistent Identifiers for CMIP6 data in the Earth System Grid Federation

M. Buurman[1], T. Weigel[1], M. Juckes[2], M. Lautenschlager[1], S. Kindermann[1]
[1]Deutsches Klimarechenzentrum,[2]STFC/British Atmospheric Data Centre

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

Problem:
Much data.
Very much data.


?!?
Which files are part of dataset xyz?
Where is dataset xyz?
How can I tell John Doe which file I've used?
What happened to version xyz?
Is version xyz the most recent one?
ESGF
Earth System Grid Federation
thousands of datasets
millions of files

The Earth System Grid Federation (ESGF) is a distributed data infrastructure that will provide access to the CMIP6 experiment data.

Coupled Model Intercomparison Project

How to keep track of it?

Each dataset is hosted at a single data centre, but can have one or several backups (replicas) at other data centres.

Over the course of the CMIP6 operational phase, datasets may be retracted and replaced by newer versions that consist of completely or partly new files.

Handle.Net®

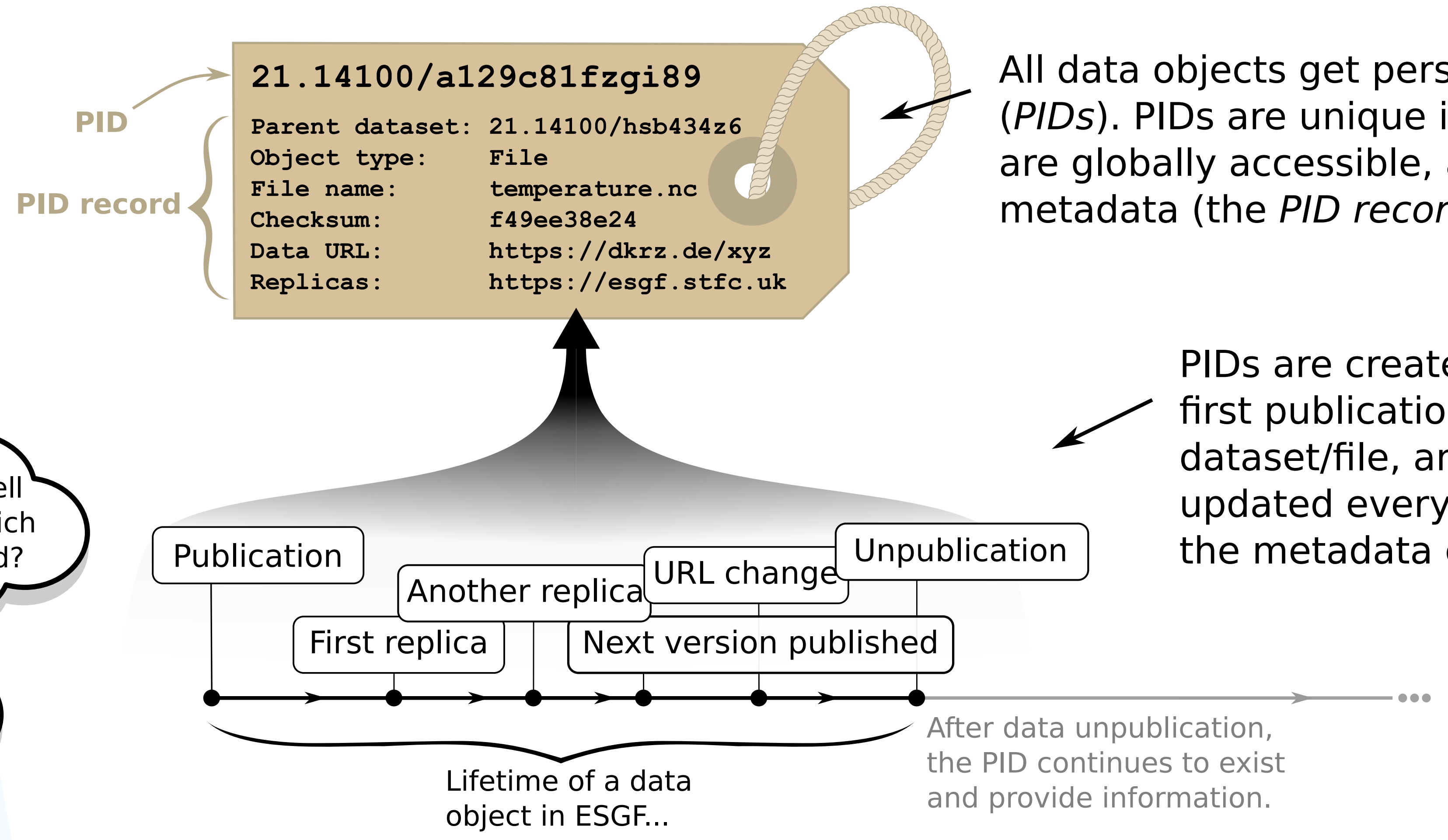
The Handle System

The persistent identifiers used in ESGF/CMIP6 are called *Handles*. The Handle System is an implementation of the PID concept, which is also underlying the Digital Object Identifiers (DOIs).

While technically the same, handles can be employed more flexibly than DOIs.

CMIP6 Example Handle: *21.14100/efeca96c-3220-3c2b-bef7-3f6cad083e05_EGU*

Solution: Persistent Identifiers...



```
21.14100/a129c81fzgi89
Parent dataset: 21.14100/hsb434z6
Object type:    File
File name:      temperature.nc
Checksum:       f49ee38e24
Data URL:       https://dkrz.de/xyz
Replicas:       https://esgf.stfc.uk
```

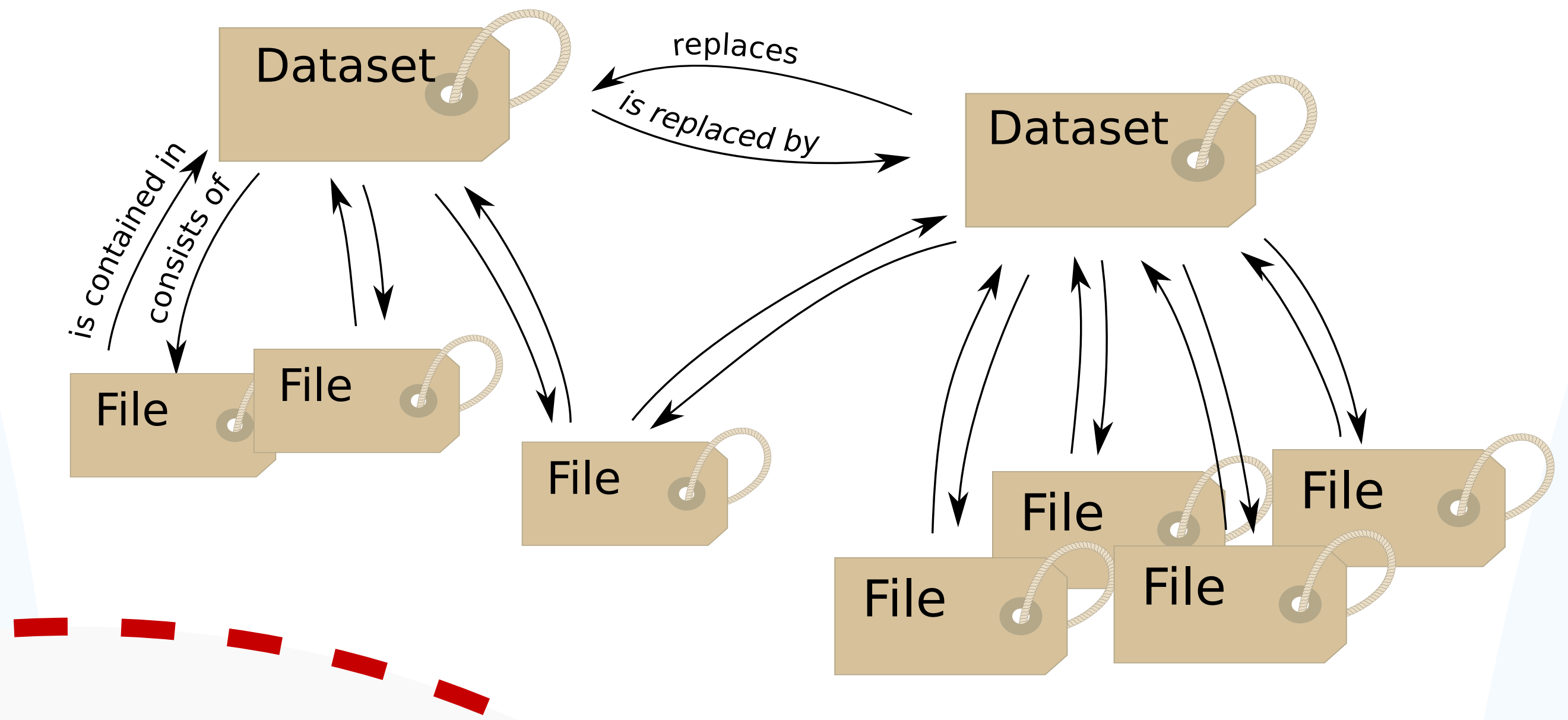
Lifetime of a data object in ESGF...

After data unpublication, the PID continues to exist and provide information.

All data objects get persistent identifiers (*PIDs*). PIDs are unique identifiers that are globally accessible, along with some metadata (the *PID record*).

PIDs are created during first publication of a dataset/file, and updated every time the metadata changes.

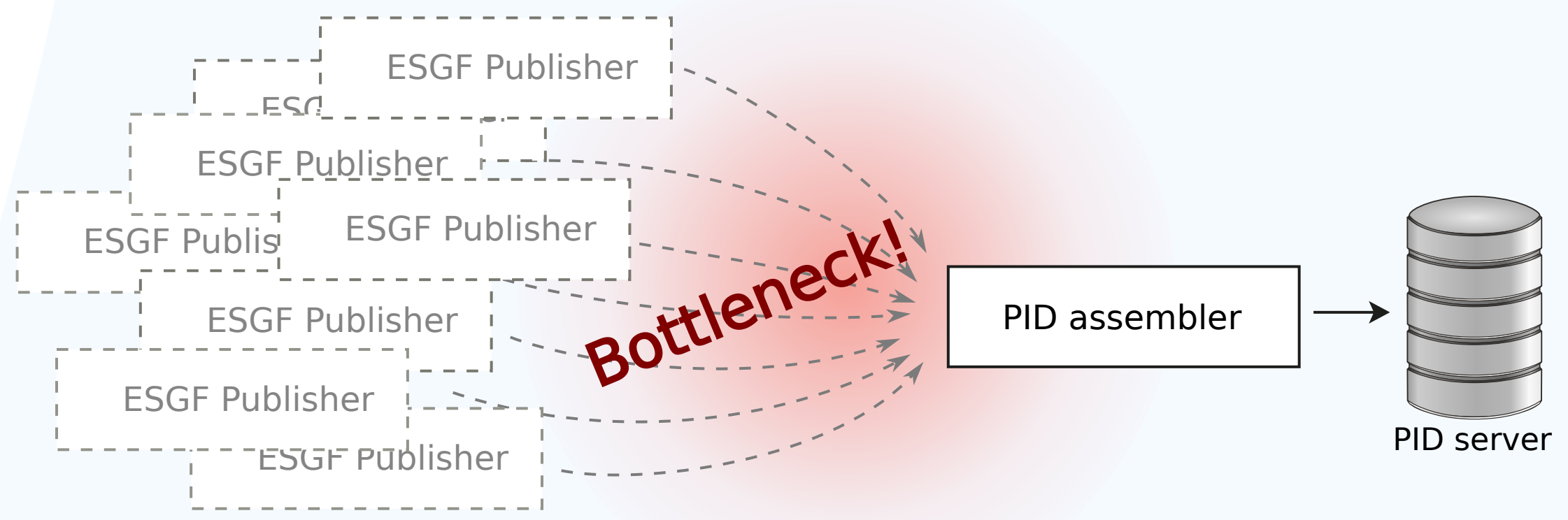
The PIDs are interlinked to keep track of the relationships between data objects:



Implementation

The PID creation is embedded in the ESGF data publication process.

Assembling the metadata records and registering the PIDs on a central server is a potential performance bottleneck:

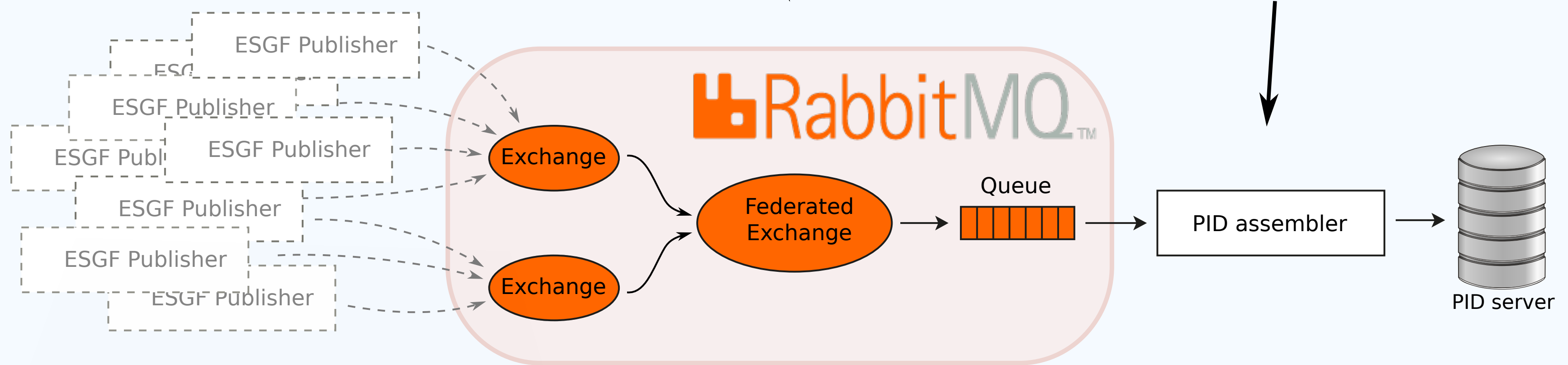


Millions of data objects may be published in a short time when the CMIP6 experiment phase begins!

Instead:

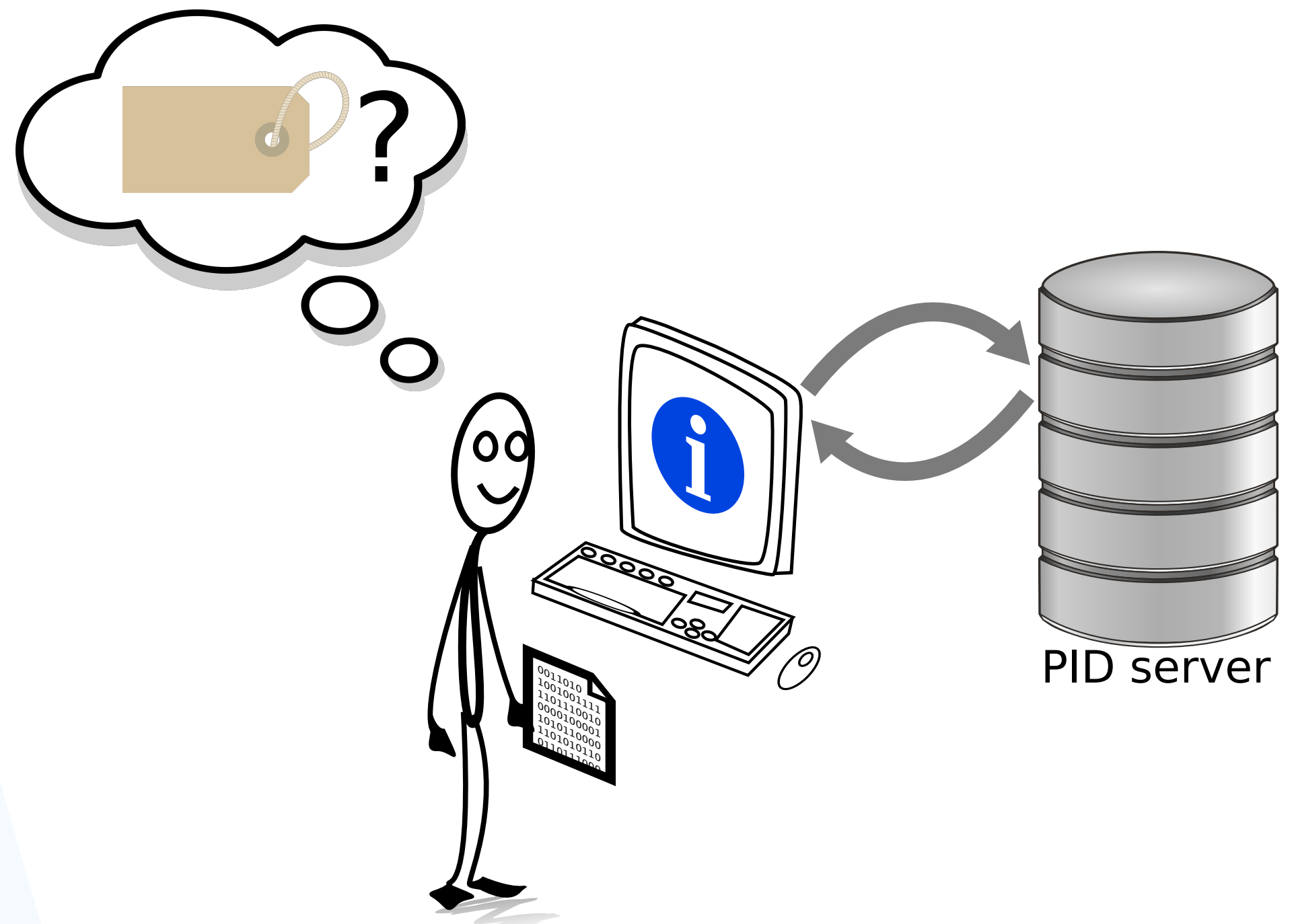
The PID registration and metadata update tasks are pushed to a message queueing system facilitating high availability and scalability...

...and then processed asynchronously.

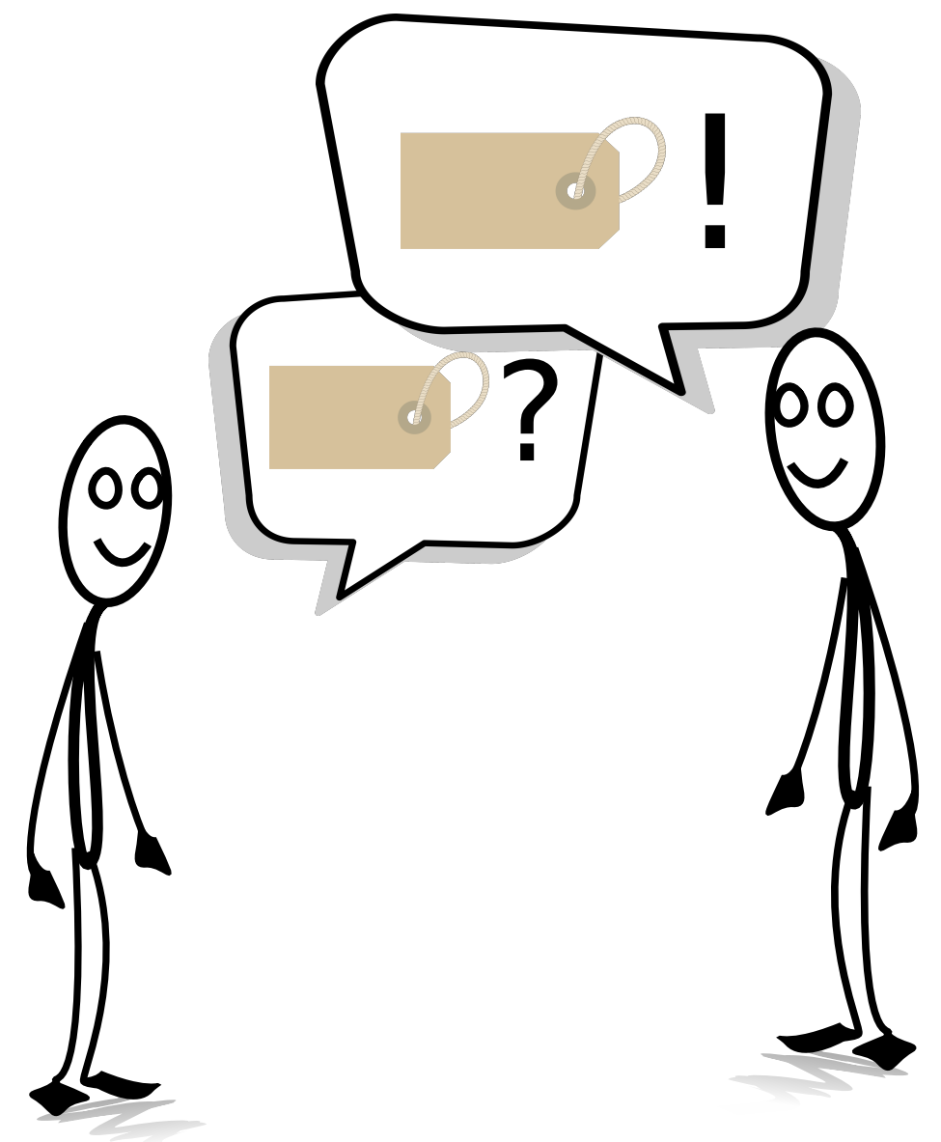


This leads to a slight delay in PID registration but avoids blocking resources at the data centres and slowing down data publication.

Benefits



Scientists can retrieve additional information about the data they are working with (e.g. more recent versions)...



Scientists are able to communicate precisely and on a very fine granularity about data...

Contact: buurman@dkrz.de
http://www.dkrz.de