Persistent Identifiers for Scientific Data at CSCS

Other Conference Item

Author(s):
Valle, Mario

Publication date:
2019-09-13

Permanent link:
https://doi.org/10.3929/ethz-b-000365505

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

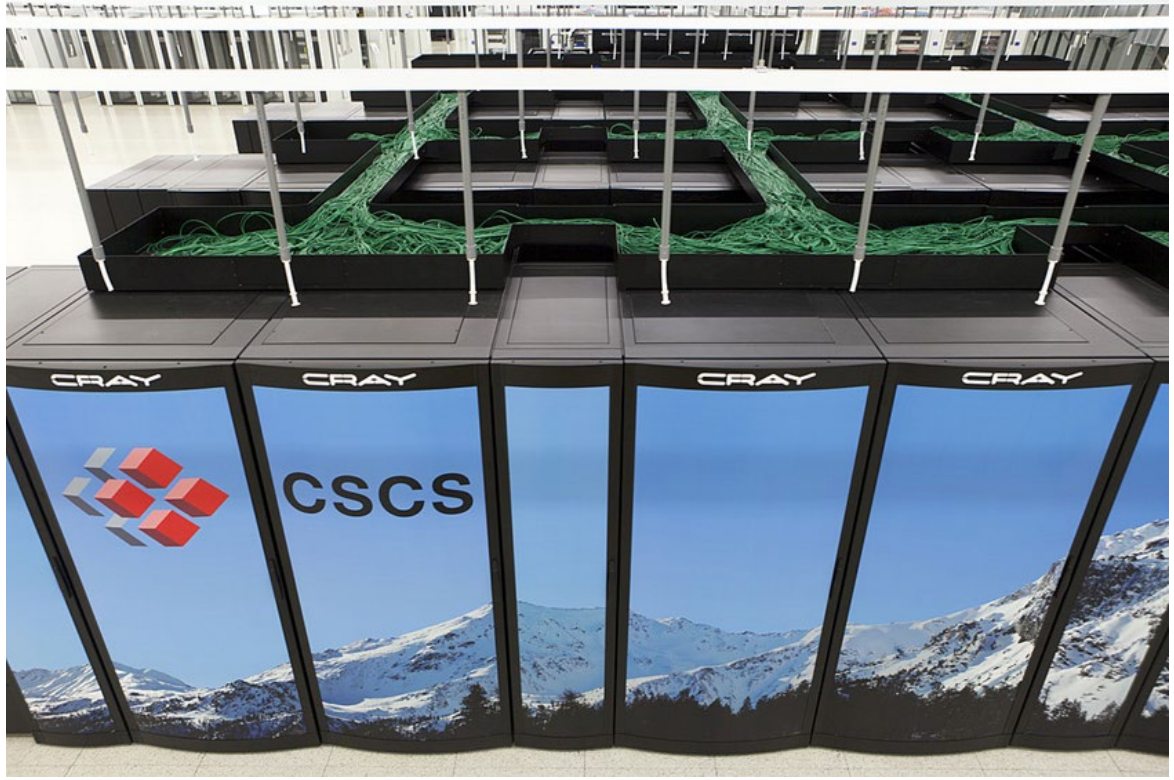
Persistent Identifiers for Scientific Data at CSCS

Persistent Identifiers in Research – ETH Zürich event

Mario Valle, CSCS

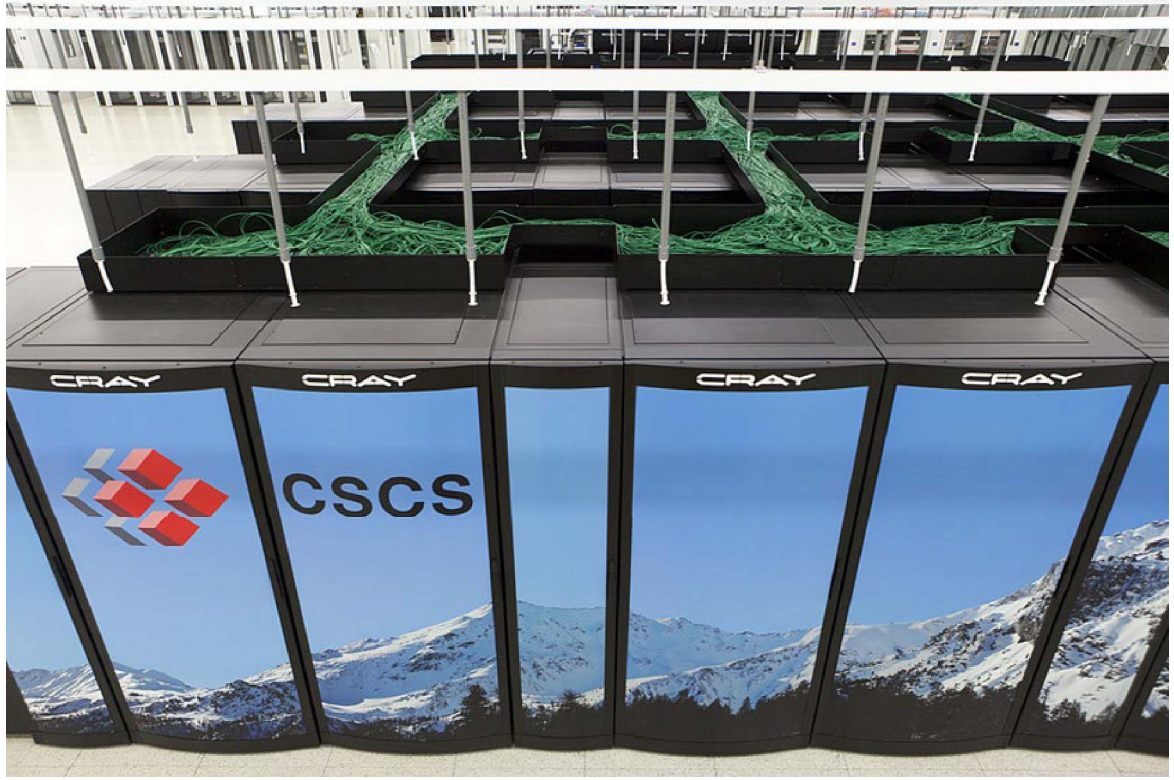
September 13, 2019

CSCS Mission



Founded in 1991, CSCS, the Swiss National Supercomputing Centre, develops and provides the **key supercomputing capabilities** required to solve important problems to science and/or society. …

CSCS Mission



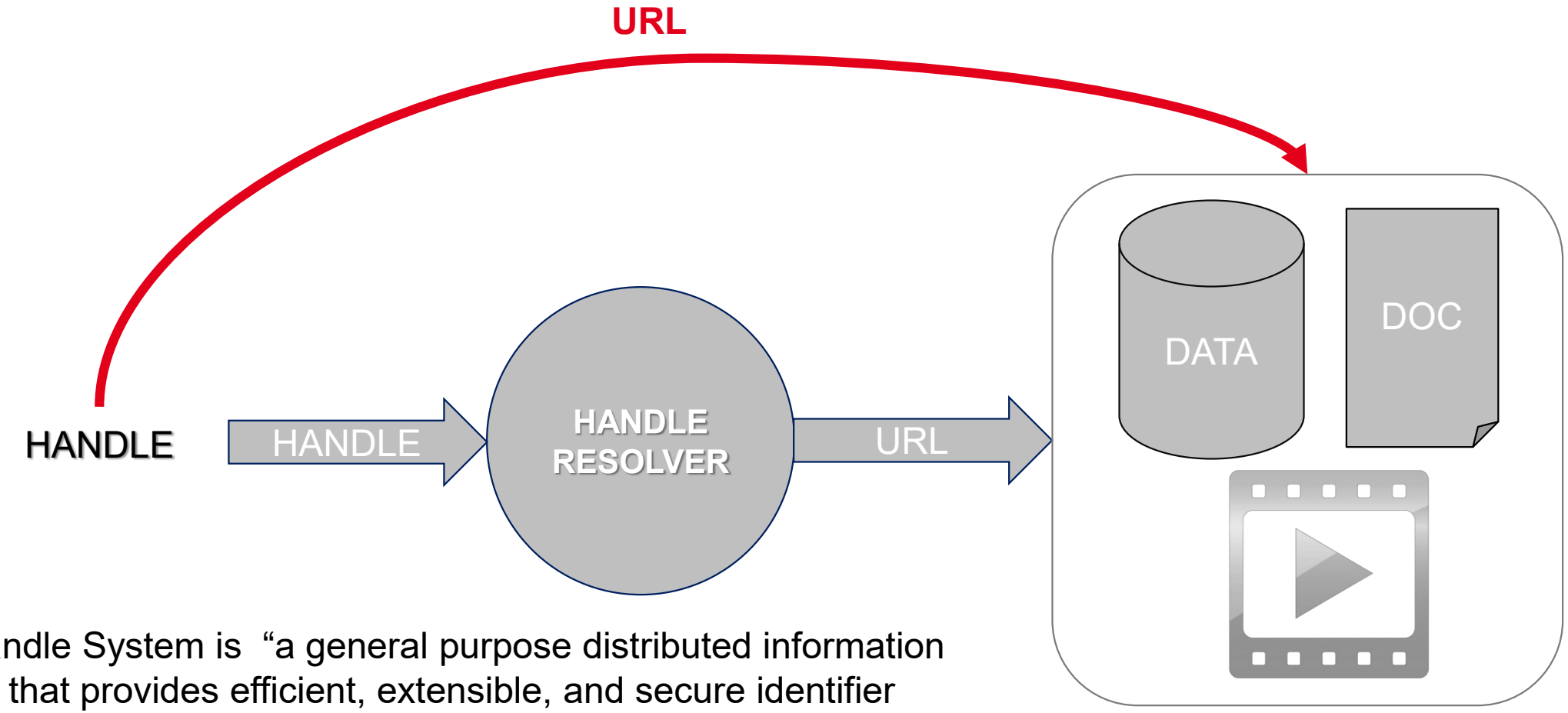
Founded in 1991, CSCS, the Swiss National Supercomputing Centre, develops and provides the **key supercomputing capabilities** required to solve important problems to science and/or society. …

Means also: Data Management, Data Analytics, FAIR support…

Prerequisite for good data science

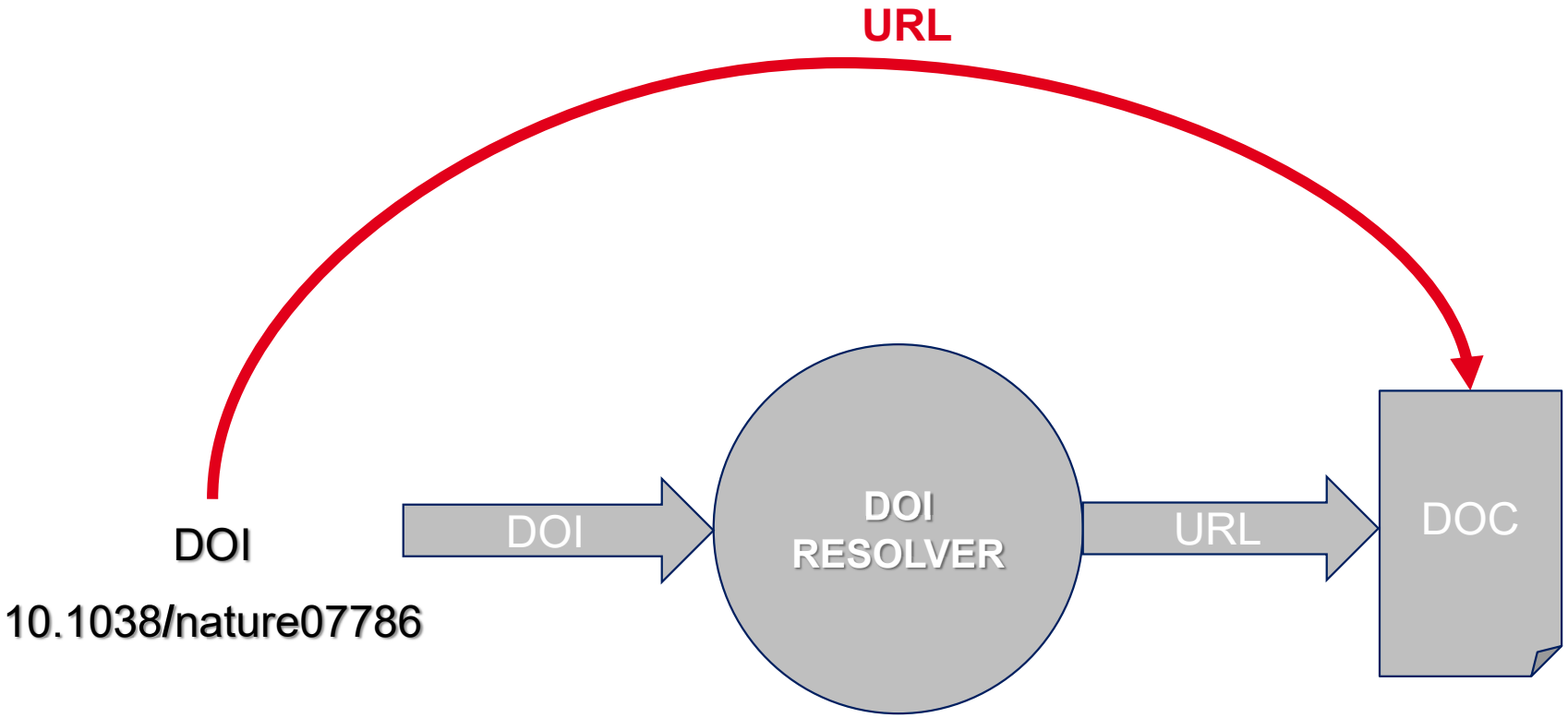
- Data should be **unambiguously** and **certainly identified** (by something that depends on data content and not location)
- A persistent identifier is a handle for any type of dataset. Identifies data objects regardless of their physical location
- A persistent identifier should be permanent and reliably associated to its dataset
- A persistent identifier carries more information than a generic UUID because it can associate metadata to the dataset

Base of every handle system (DOI, PID, URN, ARK, PURL, ISBN…)

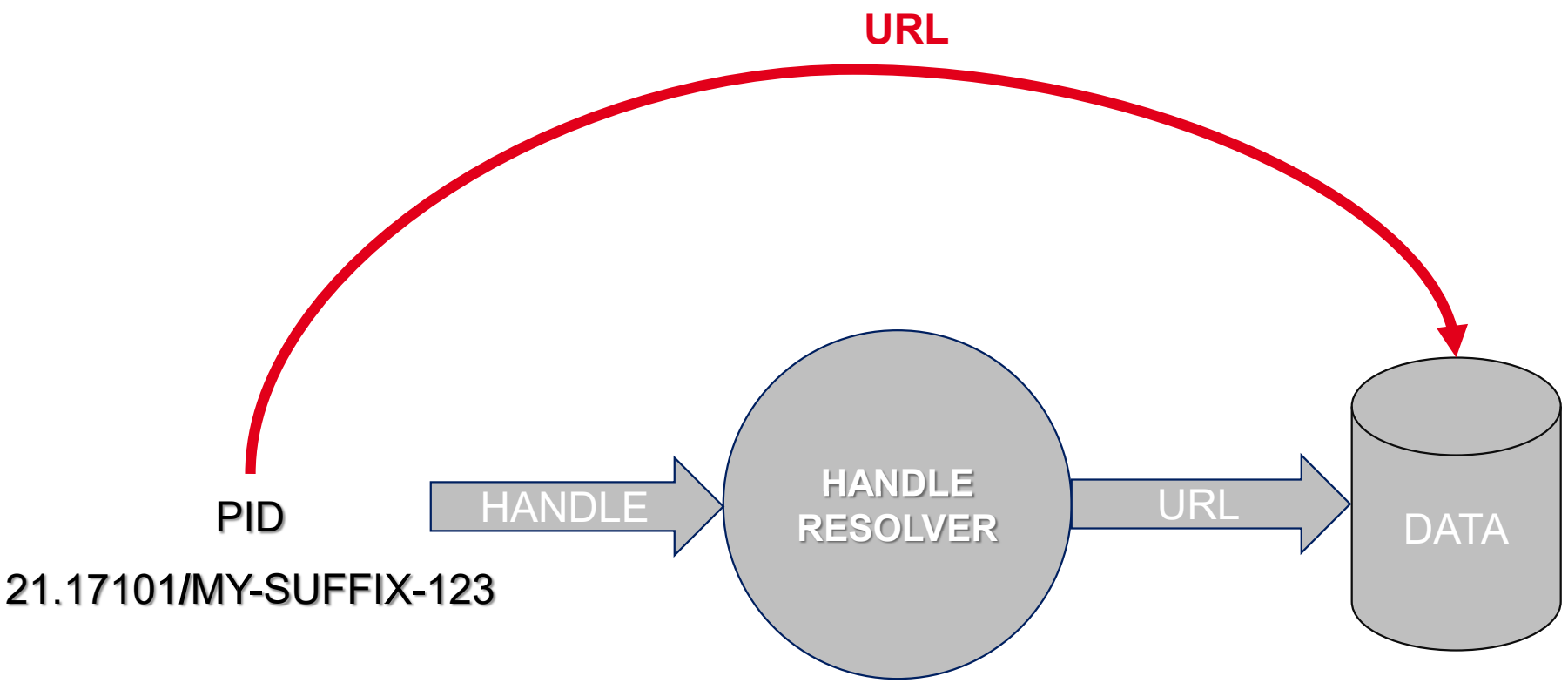


The Handle System is "a general purpose distributed information system that provides efficient, extensible, and secure identifier and resolution services for use on networks such as the Internet."

The DOI resolving process



The PID resolving process



The **21.** identifies ePIC PID; **10.** identifies DOI. Both handled by DONA foundation (https://www.dona.net/mpas)

ePIC consortium for Persistent Identifiers (PID)

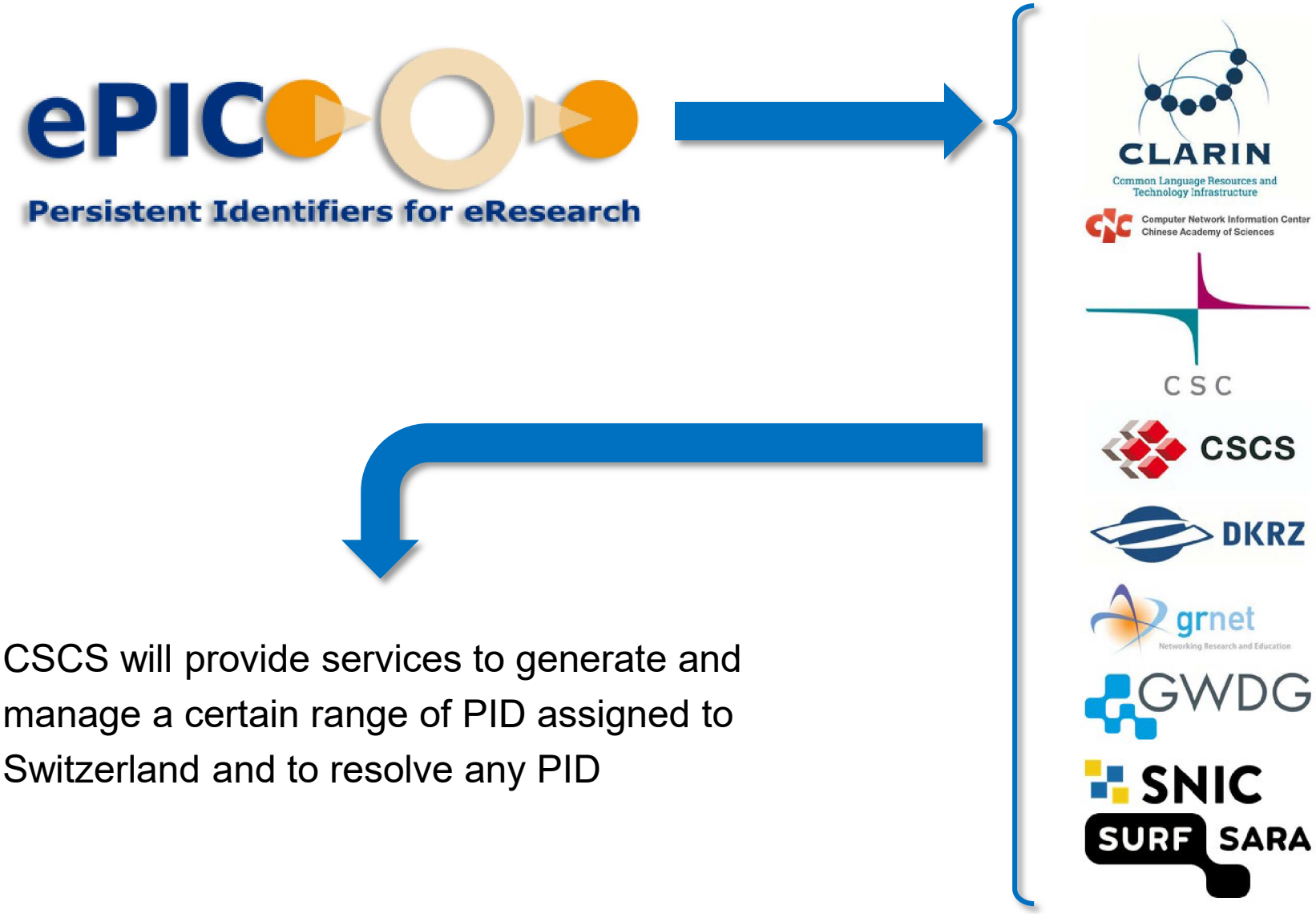
ePIC
Persistent Identifiers for eResearch

https://www.pidconsortium.eu/

"The eResearch Persistent Identifier Consortium (ePIC) offers a service to create, manage, and resolve persistent identifiers (PID). The increasing amount of research data, the variety of the usage profiles and the international exchange within different infrastructures demand to uniquely assign the data with a PID with a high degree of flexibility and robustness. ePIC offers a reliable mechanism to guarantee these features of persistent identifiers."

Excerpt from a poster at RDA 3rd Plenary Meeting

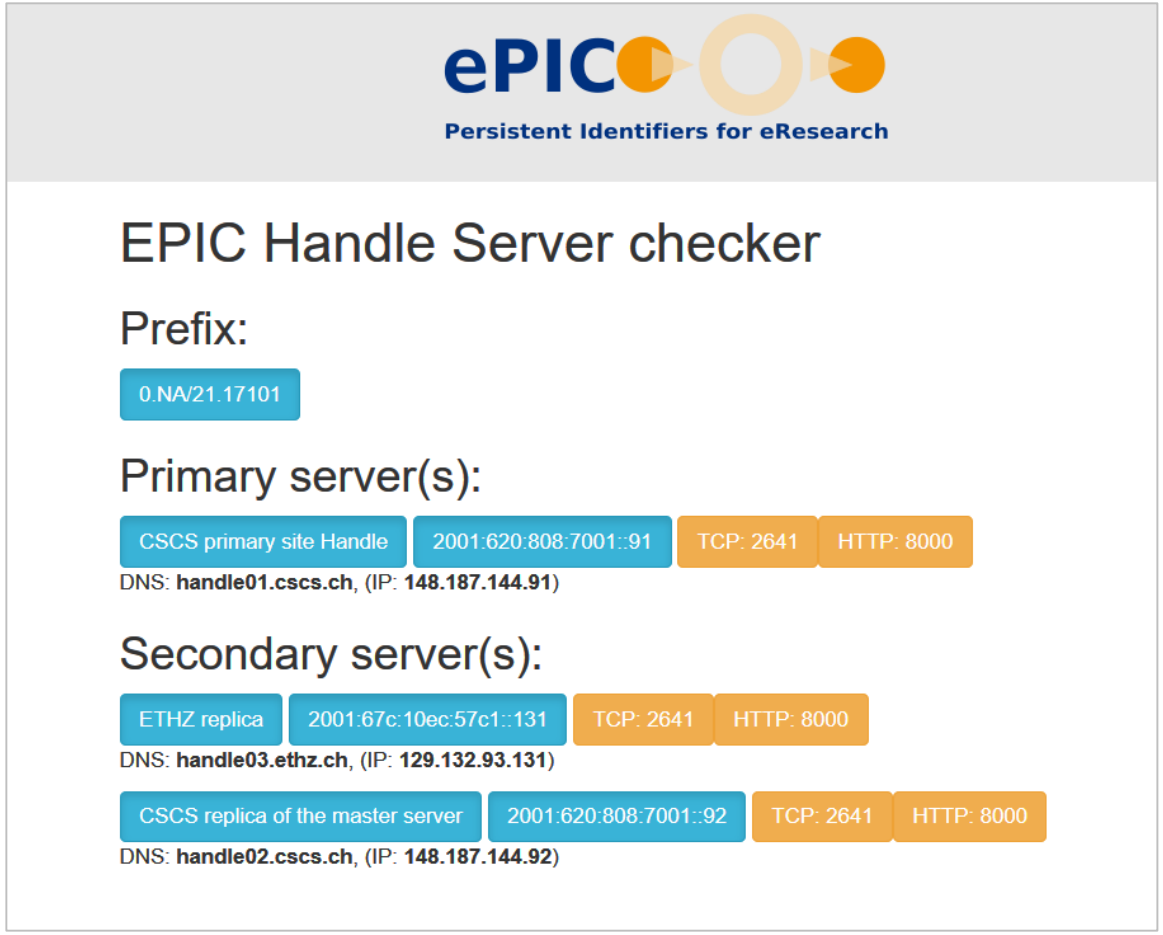
CSCS is part of the ePIC consortium (since Sept. 2018)



CSCS will provide services to generate and manage a certain range of PID assigned to Switzerland and to resolve any PID

Infrastructure at CSCS

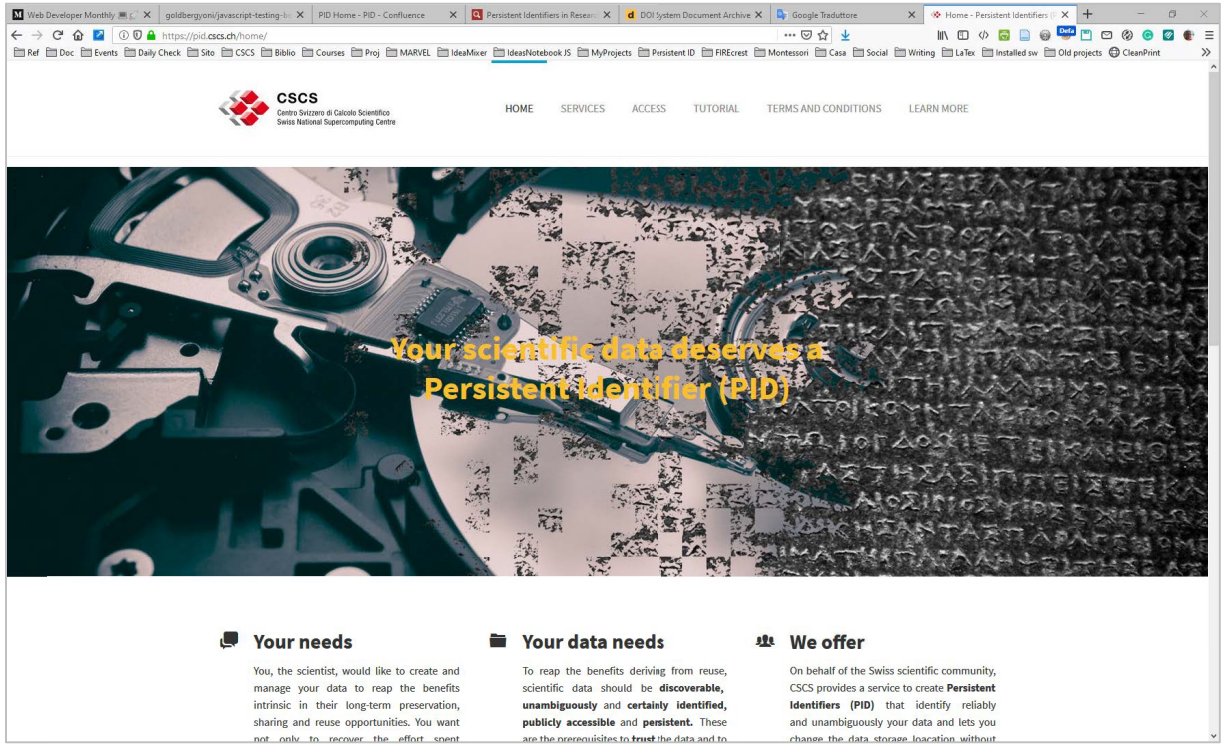
- For reliability and availability CSCS has three handle servers integrated in the ePIC infrastructure
- To resolve a PID use the master resolver: `https://hdl.handle.net/21.17101/SUFFIX`
- We manage two prefixes:
  - 21.17101 (persistent)
  - 21.T17999 (testing)
- We could provide to research institutions they own prefix:
  - From 21.17102 ↑ counting up
  - From 21.T17998 ↓ counting down



User access to CSCS PID infrastructure

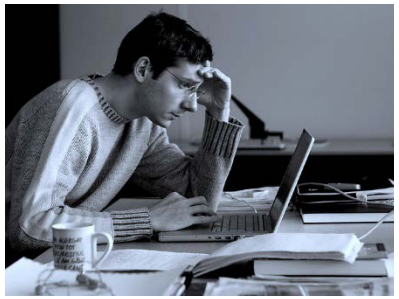
- Through the portal `pid.cscs.ch` (currently on invitation only)
- The PID system could be accessed through its API (docs on: `http://www.handle.net/`)

```
$ curl -s \
https://hdl.handle.net/api/handles/21.T17999/12345-54321?pretty=true
{
  "responseCode": 1,
  "handle": "21.T17999/12345-54321",
  "values": [
    {
      "index": 1,
      "type": "URL",
      "data": {
        "format": "string",
        "value": https://cloud.cscs.ch/owncloud/index.php/s/4xi37uW1HsK91cy"
      },
      "ttl": 86400,
      "timestamp": "2018-10-31T14:22:50Z"
    },
    ...
}
$
```



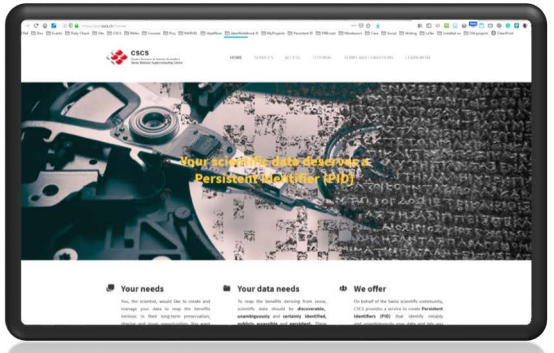
- Creation and management of PIDs through the portal and through the API

Next use case: long term storage at CSCS



Simulation results

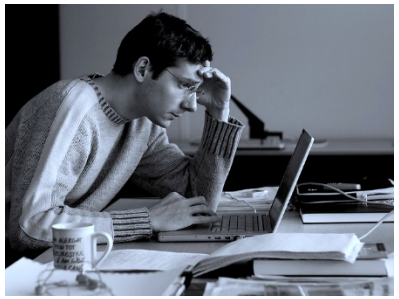
PID



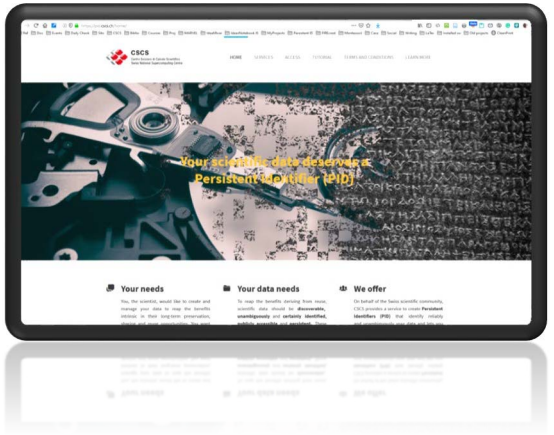
CSCS Long-term archiving user interface

Access simulation results using PID

Next use case: long term storage at CSCS



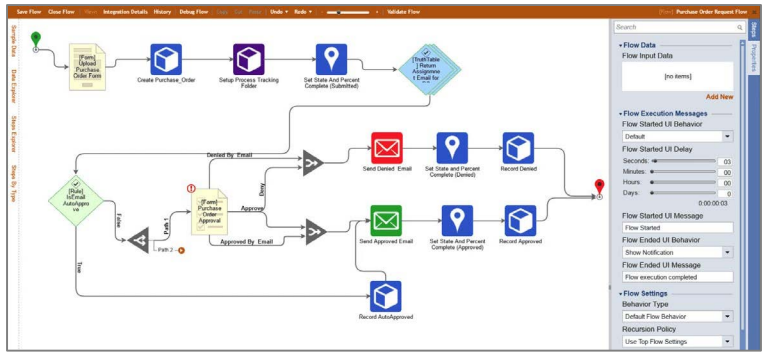
Access simulation results using PID



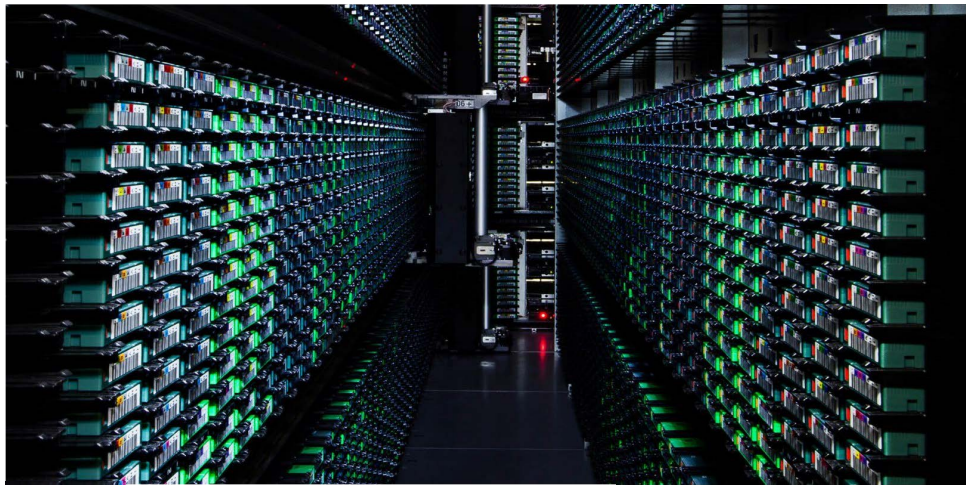
CSCS Long-term archiving user interface

Data migration

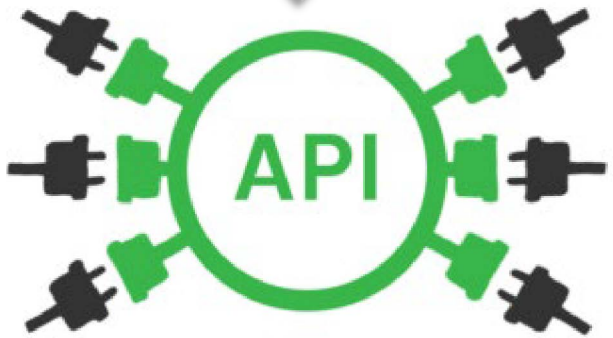
Next Use Case: Automatically associate PID to files



Result location



PID

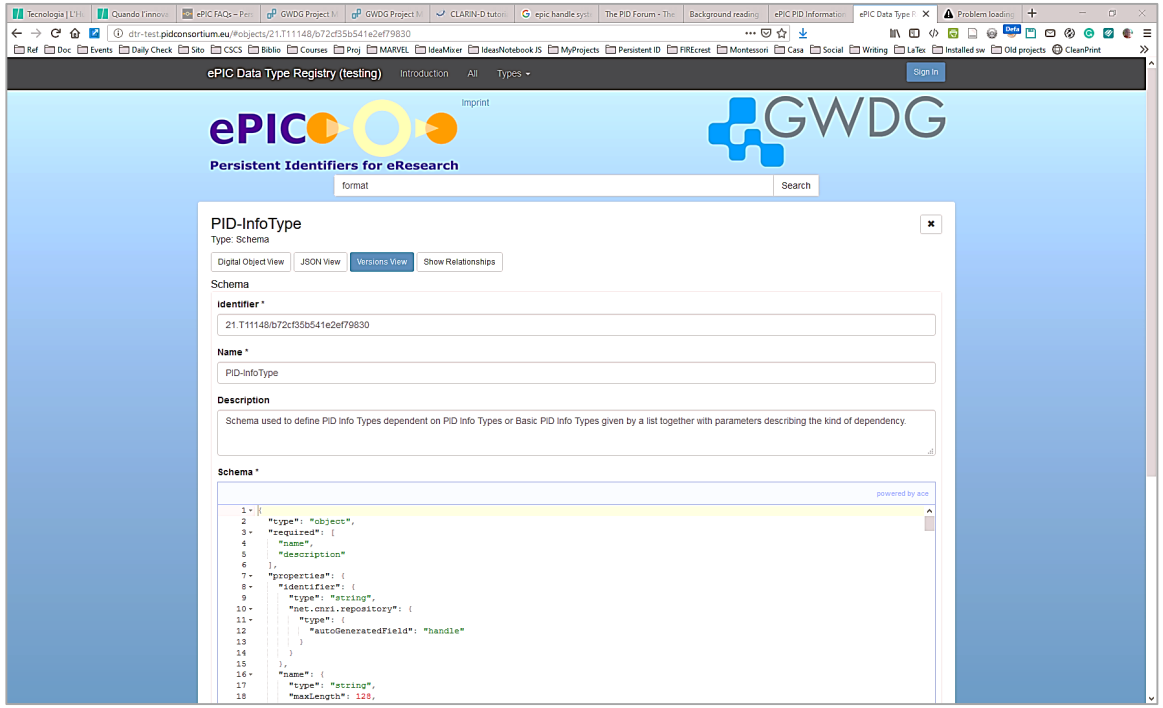


CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Last Use Case: scientific metadata

- Sufficiently flexible to cover ample set of scientific data needs
- Make use of ontologies to extend the metadata set
- Enhance the “Interoperable & Reusable” in FAIR
- Strong dependency on science domains
- The user could associate a set of metadata to a PID
- The user can run queries on metadata to obtain a list of PID



DOI comes with an established set of metadata

doi2bib — give us a DOI
and we will do our best to get you the BibTeX entry

10.1107/S0108767310026395 get BibTeX

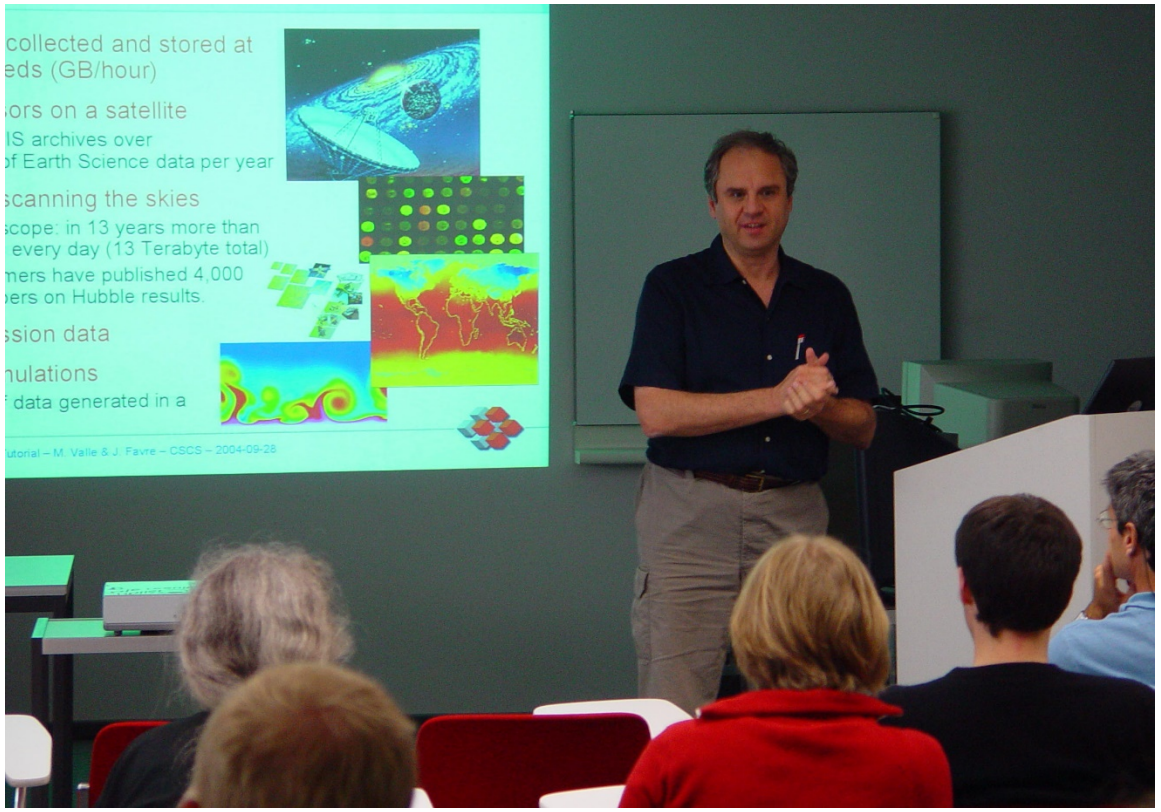
```
@article{Valle2010,
  doi = {10.1107/s0108767310026395},
  url = {https://doi.org/10.1107/s0108767310026395},
  year  = {2010},
  month = {aug},
  publisher = {International Union of Crystallography ({IUCr})},
  volume = {66},
  number = {5},
  pages = {507--517},
  author = {Mario Valle and Artem R. Oganov},
  title = {Crystal fingerprint space {\textendash} a novel paradigm for studying crystal-s
  journal = {Acta Crystallographica Section A Foundations of Crystallography}
}
```

https://doi.org/10.1107/s0108767310026395

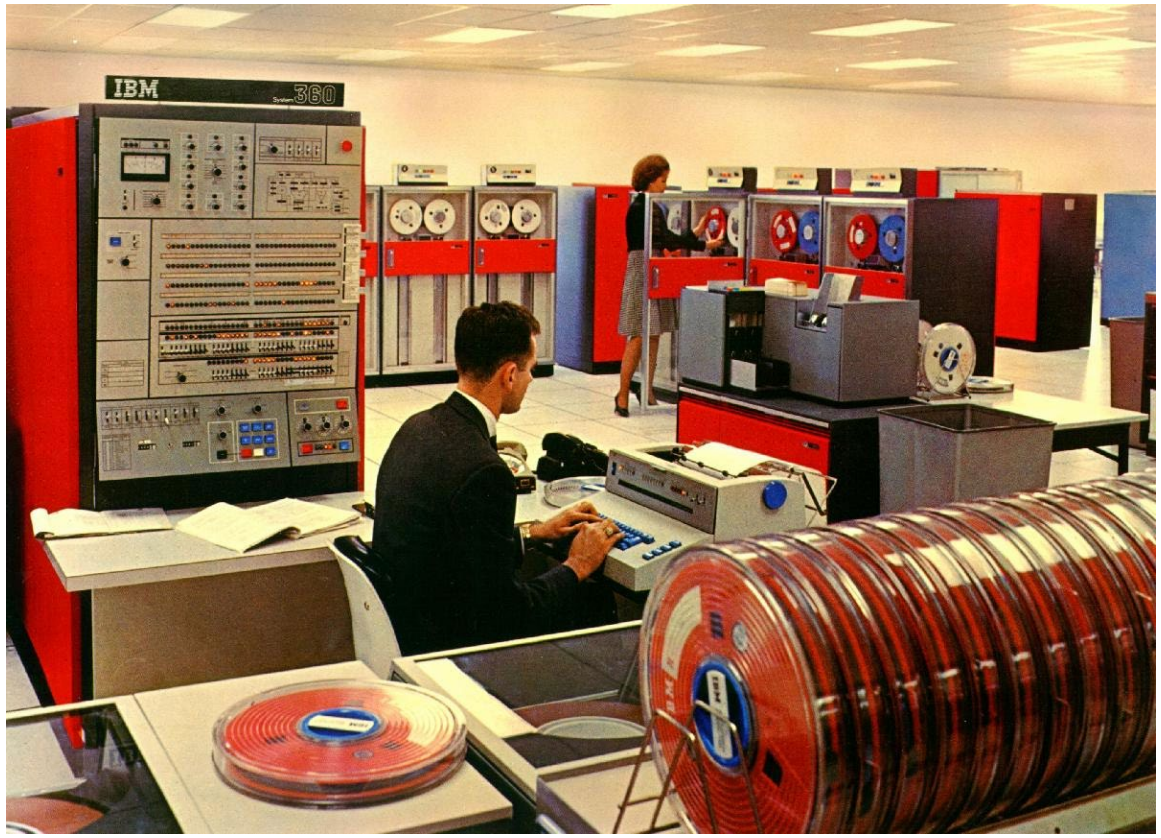
Copy Bib to Clipboard Copy URL to Clipboard

Main CSCS goal: find Use Cases, experiment with users

- I'm the point of contact for PID ideas, suggestions and project specific requests
- I'm collecting use cases to suggest how this technology could help Swiss scientist's work
- Contact me at: **`mvalle@cscs.ch`**



Where all Handle Systems stops?



- All handle systems (PID, DOI, etc.) are glorified DNS systems
  - DNS receives www.google.com and returns: `172.217.16.36`
  - DOI receives `10.1038/nature07786` and returns: https://www.nature.com/articles/nature07786
- Who enforces that returned URL is valid?
- Who enforces file content has not tampered with? Versioning?
- Who enforces file has not moved without updating its PID or DOI?
- Who protects file (or publication) from being deleted?

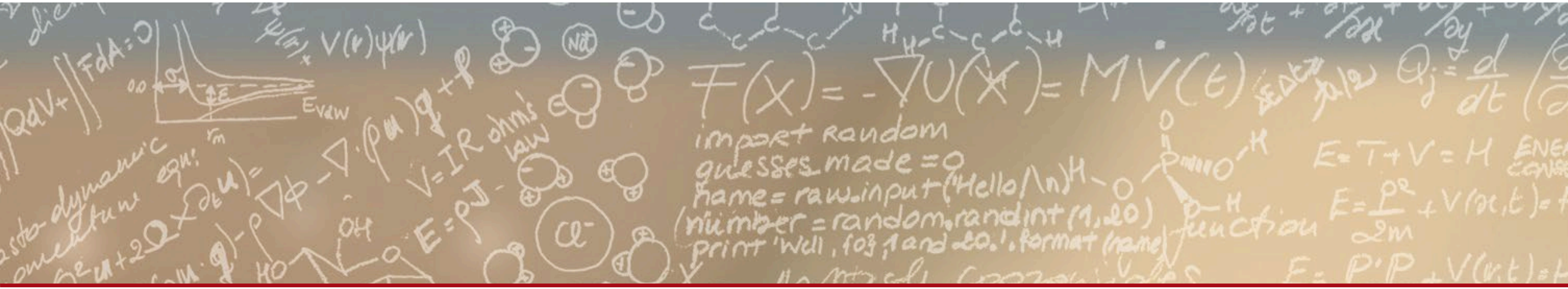
A human (cultural) problem needs a human solution

PID future at CSCS

- We want to create awareness and, hopefully, build a Swiss community interested in this aspect of data management
- Continue collecting and implementing PID Use Cases
- Continue working with other ePIC Consortium members to contribute to PID maturity and to simplify PID usage and metadata management and search
- Study ePIC collaborations:
  - With ORCID **Project RIPEN.** This project proposes the use of JSON Web tokens (JWTs) for collecting authenticated user permissions and to delegate them from one system to another.
  - With EU **Project FREYA.** The project aims to extend the infrastructure for persistent identifiers (PIDs) as a core component of open research, in the EU and globally.

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



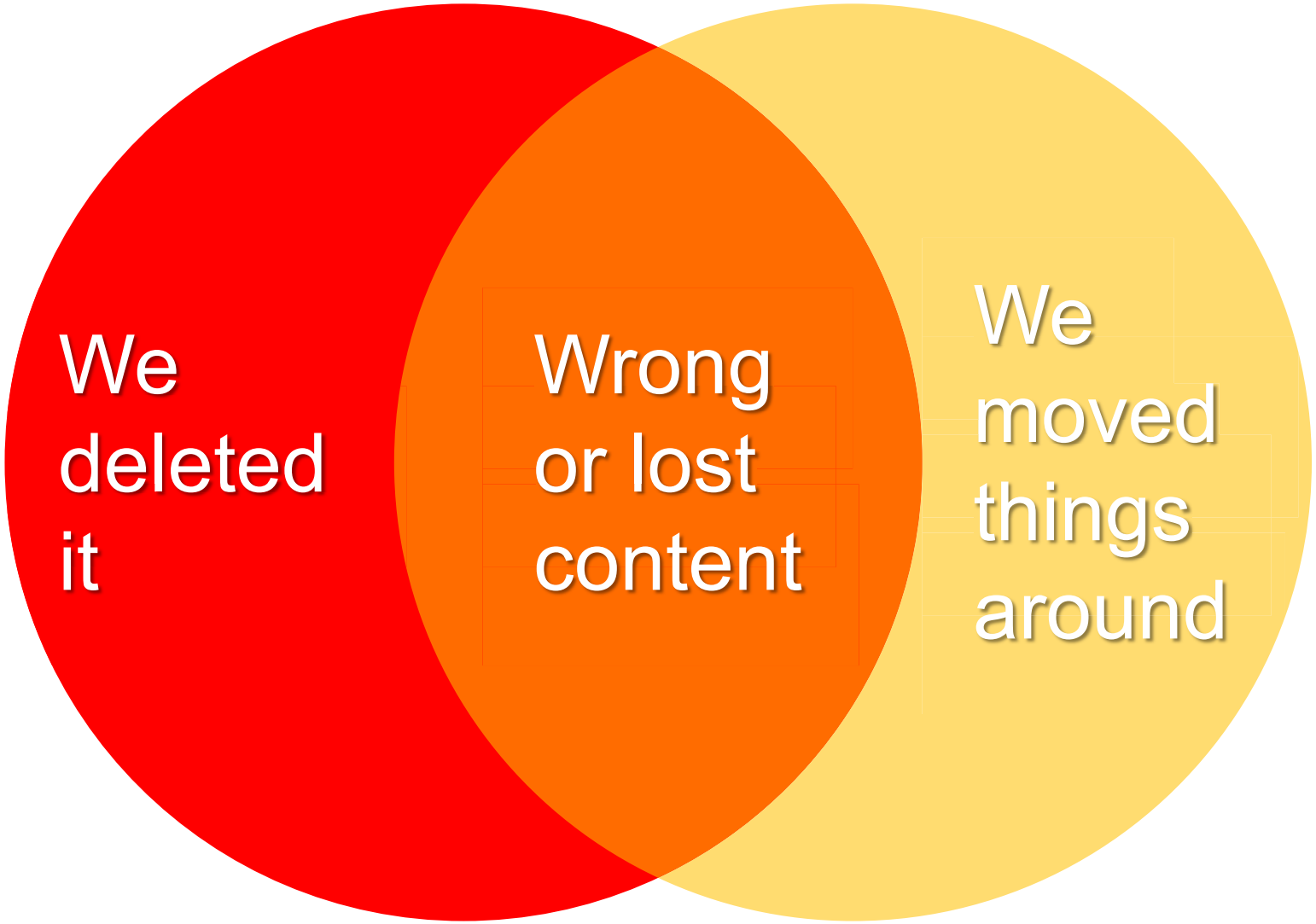
Thank you for your attention.

cscsch

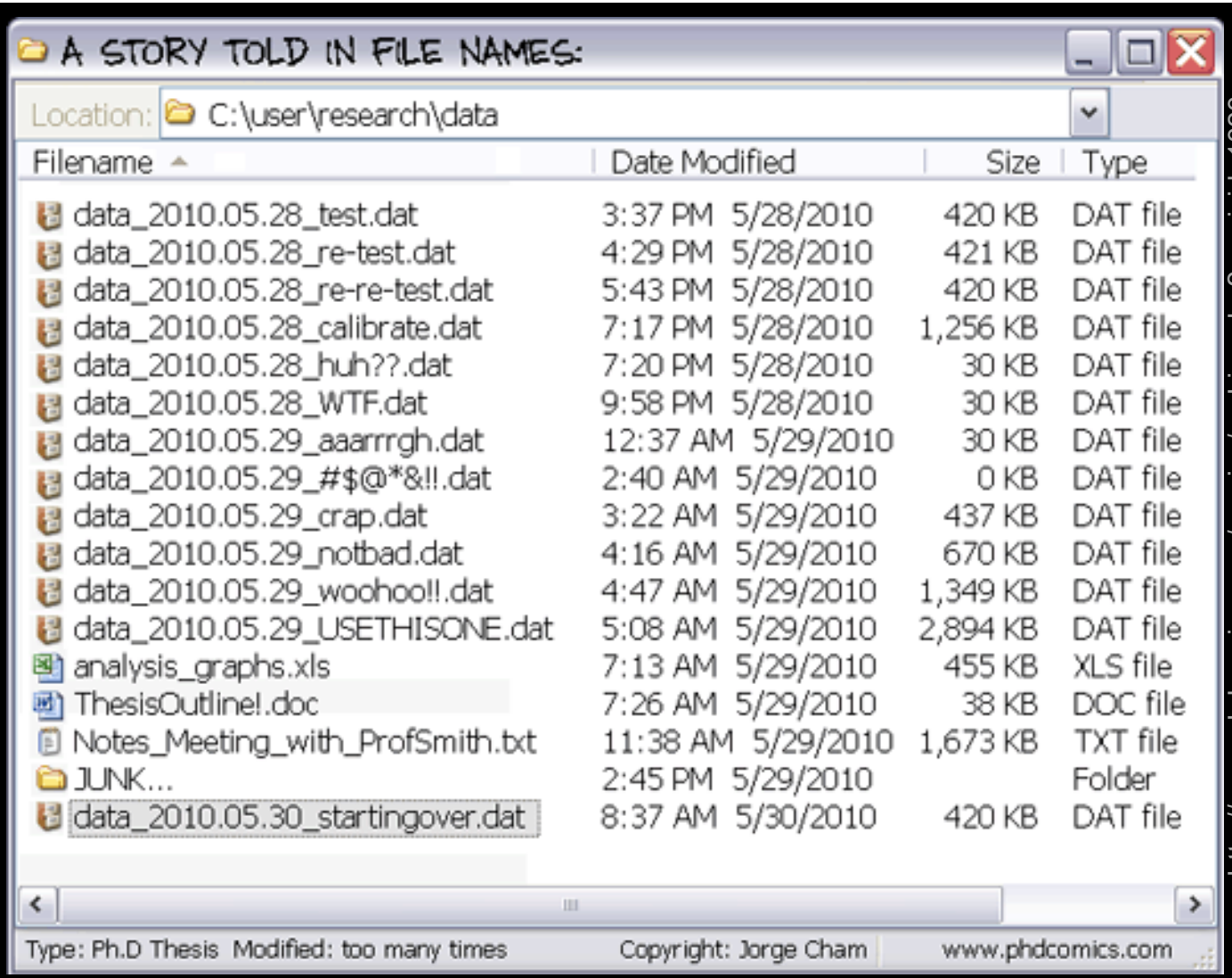
PID needs a social infrastructure

- PID Infrastructure maintained by a dedicated and reliable team
- Provided by a non-profit organization
- Governed by international boards
- Based on open standards

A human problem needs a human solution



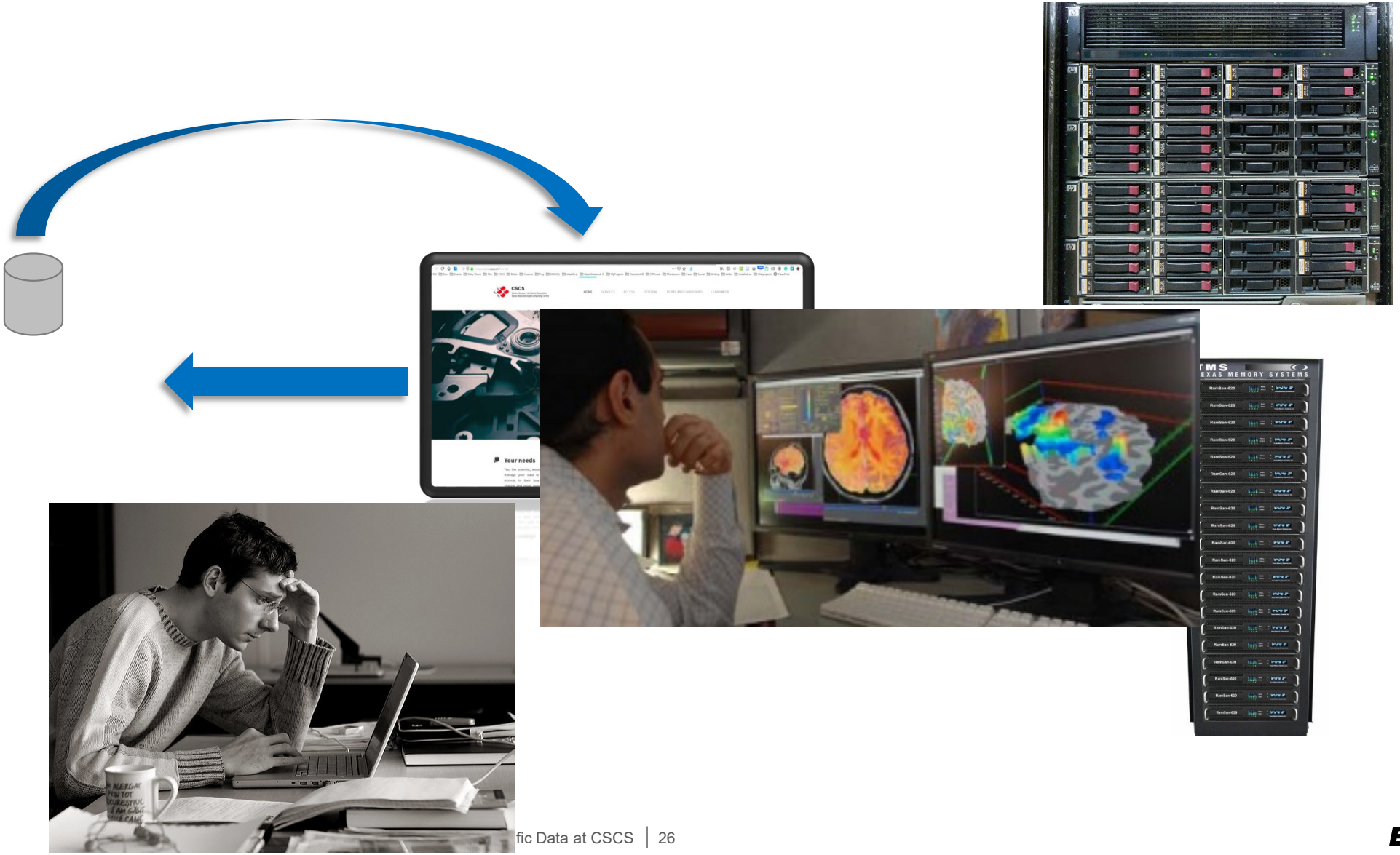
Not to say data management leaves (often) a lot to be desired…



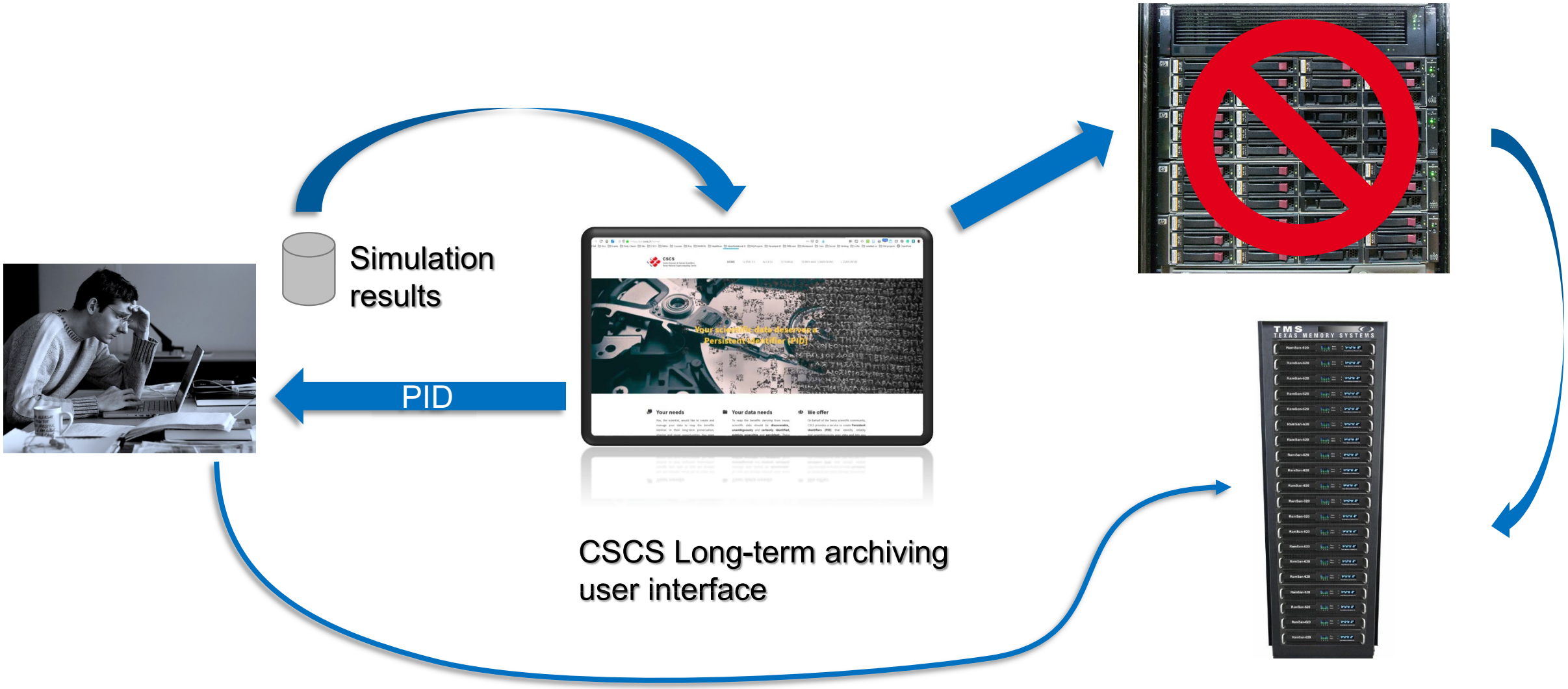
http://www.phdcomics.com/comics/archive.php?comicid=1323

Next use case: long term storage at CSCS

Next use case: long term storage at CSCS



Next use case: long term storage at CSCS



- DOI for publications, PID for data
- DOI has metadata for publications, PID more general
- (not sure of this): DOI: managed by International DOI Foundation (IDF), a not-for-profit membership organization that is the governance and management body for the federation of Registration Agencies providing Digital Object Identifier (DOI) services and registration. PID is managed by the ePIC consortium.
- Both depend from DONA for the first part of prefix (10. vs. 21.)

Handle frameworks comparison

	Standard	Robust Software	Resolution System	Resolution Type	Security Admin	Assoc Info	Cost
URL	RFC2616	no	yes (DNS)	single	no	no	no
URN:ISSN	ISO2397	no	no	?	no	no	no
URN:ISBN	ISO2108	no	no	?	no	no	no
URN:NBN	RFC3188	no	no	?	no	no	?
PURL	no	no	yes	single	no	no	no
Handle	RFC3650	yes	yes	multiple	yes	yes	little
DOI	Z39.84…	yes	yes (Handle)	multiple	yes	yes	large
ARK	no	no	(yes)	multiple	(no)	yes	?
info URI	RFC3668	no	no	?	no	no	no
XRI	yes	no	no	?	no	?	?