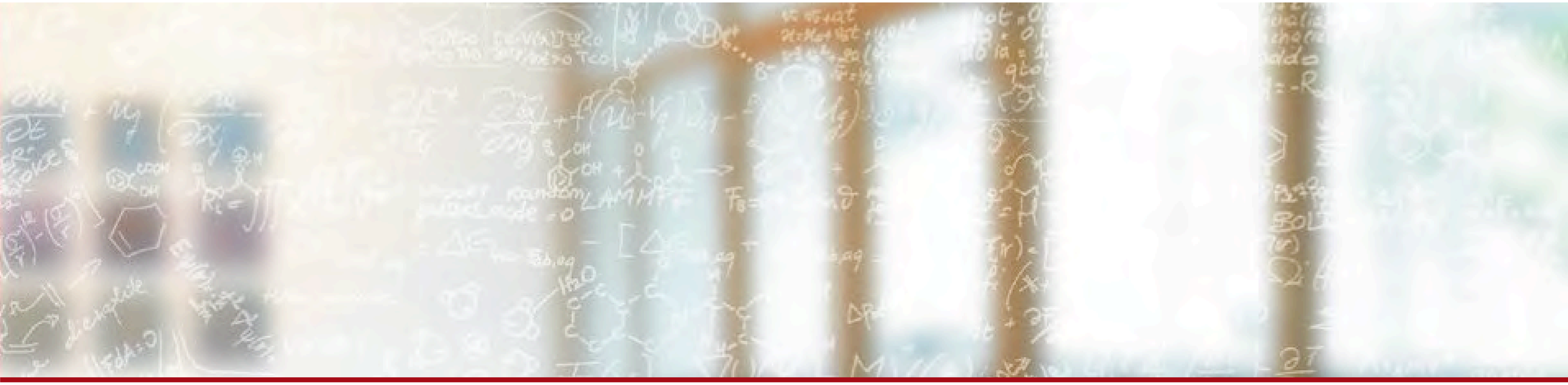
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Your Data Deserves a Persistent Identifier (PID)

Swiss Data Science Center – Lausanne
Mario Valle, CSCS
November 29, 2018

A truism: we produce a lot of data

“Our ability to capture and store data far outpaces our ability to process and exploit it. This growing challenge has produced a phenomenon we call the **data tombs**, or data stores that are effectively write-only; **data is deposited to merely rest in peace**, since in all likelihood it will never be accessed again. Data tombs also represent missed opportunities.”

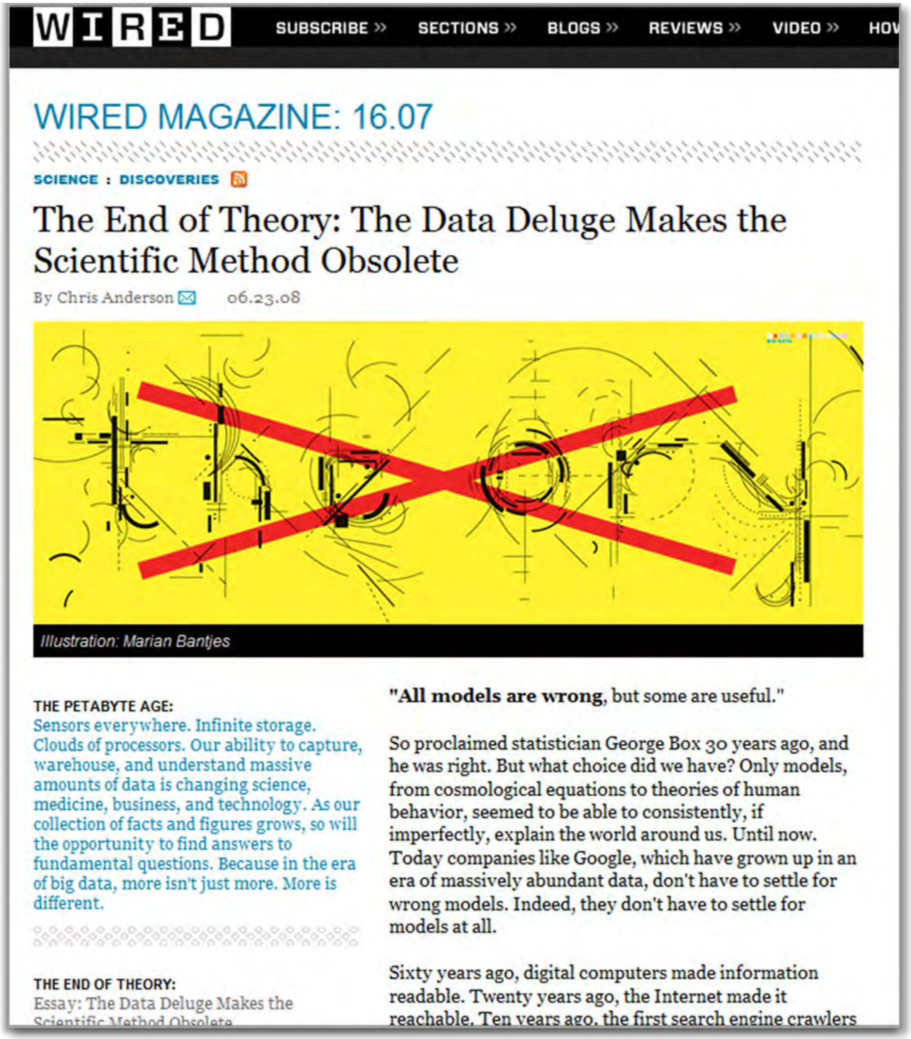
Usama Fayyad – Yahoo! Research Laboratories

We could benefit from data use, reuse and recycle

- Astronomy and Astrophysics Virtual Observatories (e.g., EURO-VO)
- Data reanalysis (common at CERN and in climate science)
- Discovery by browsing (a.k.a. Google science)
- Find correlations between data and metadata (e.g., OMEGA project for bio-imaging of virion movement in cells)
- Providing context for other data
- Stimulate new usage patterns (paradigms)

Controversial, but nonetheless a new data paradigm

- Many think that having abundant data means they do not need a scientific theory behind.
- For example: Google can translate languages without actually “knowing” them.
- Without taking to the extreme this attack to the scientific method, we are already pointing in that direction.
- Think. Every kind of Machine Learning, deep or not, substitutes the need of a model with the availability of training data.



WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different.

THE END OF THEORY:
Essay: The Data Deluge Makes the Scientific Method Obsolete

"**All models are wrong**, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers

www.wired.com/science/discoveries/magazine/16-07/pb_theory

SNF (and other) requests about data publication…

- … but these are only the tip of the iceberg.
- Scientists want to structure data use, reuse and recycle from the beginning, when data is created. They don’t want to attach the problem at the end, when the work is published.
- Also they don’t want more bureaucracy or rules to comply with, without perceived benefits for their science.

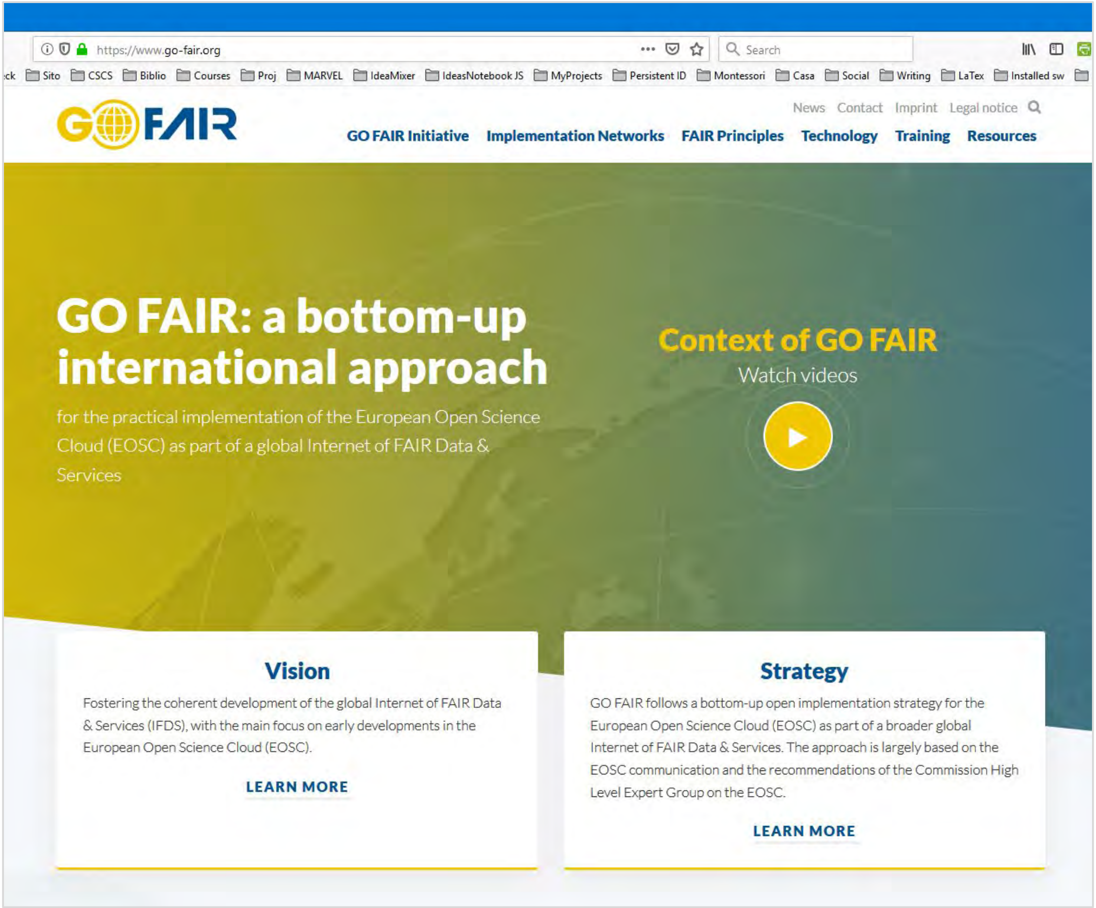
Essential prerequisites to make all this happens

- Data should be **discoverable** (by associated metadata or by public catalogs. Kudos to Google for its Dataset Search)
- Data should be **unambiguously** and **certainly identified** (by something that depends on data content and not location and is the basis of authorship assignment)
- Data should be **publicly accessible** and **persistent** (should not disappear when researcher moves to another university. Public does not means free. After discovery there may be an authorization step)
- Data should be **trusted** (i.e., it is what it claim to be, authorship is clear, metadata are verified)

In other words: data should be FAIR

FAIR data is data which meets standards of:

- **F**indability
- **A**ccessibility
- **I**nteroperability
- **R**eusability


https://www.go-fair.org
Search
ck Sito CSCS Biblio Courses Proj MARVEL IdeaMixer IdeasNotebook JS MyProjects Persistent ID Montessori Casa Social Writing LaTex Installed sw
GO FAIR
News Contact Imprint Legal notice
GO FAIR Initiative Implementation Networks FAIR Principles Technology Training Resources

GO FAIR: a bottom-up international approach

for the practical implementation of the European Open Science Cloud (EOSC) as part of a global Internet of FAIR Data & Services

Context of GO FAIR

Watch videos

Vision

Fostering the coherent development of the global Internet of FAIR Data & Services (IFDS), with the main focus on early developments in the European Open Science Cloud (EOSC).

LEARN MORE

Strategy

GO FAIR follows a bottom-up open implementation strategy for the European Open Science Cloud (EOSC) as part of a broader global Internet of FAIR Data & Services. The approach is largely based on the EOSC communication and the recommendations of the Commission High Level Expert Group on the EOSC.

LEARN MORE

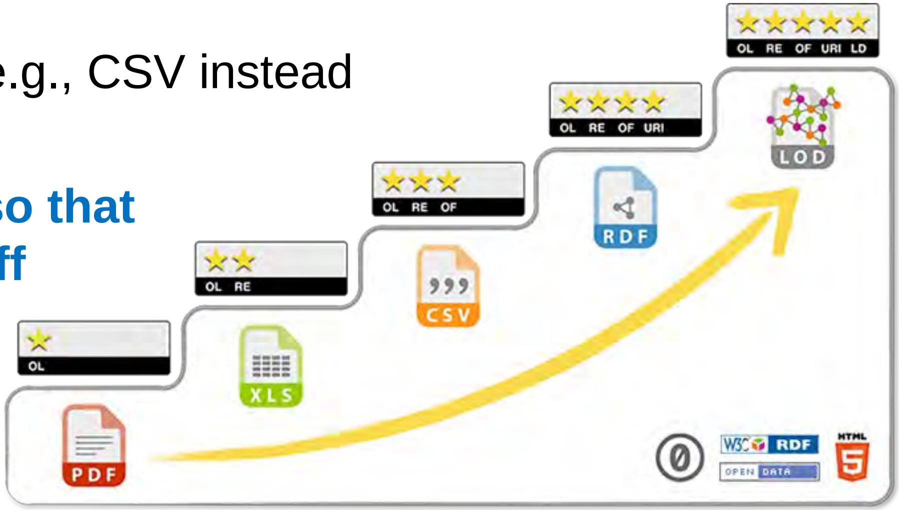
```
(https://www.nature.com/articles/sdata201618) or DOI: 10.1038/sdata.2016.18
```

Another step after FAIR is Linked Open Data

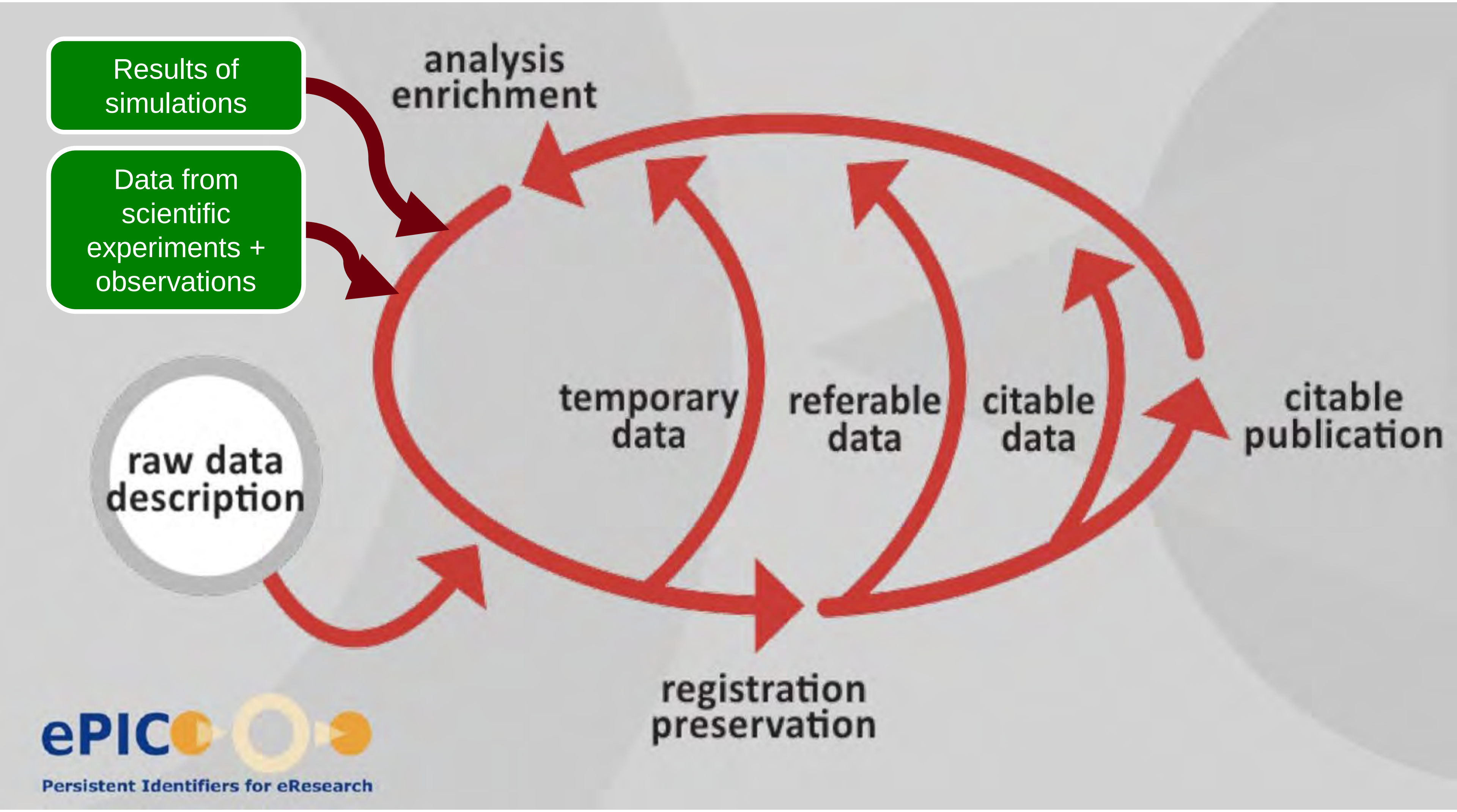
The 5-stars deployment scheme for Linked Open Data proposed by Tim Berners-Lee

`(https://5stardata.info/en/)`

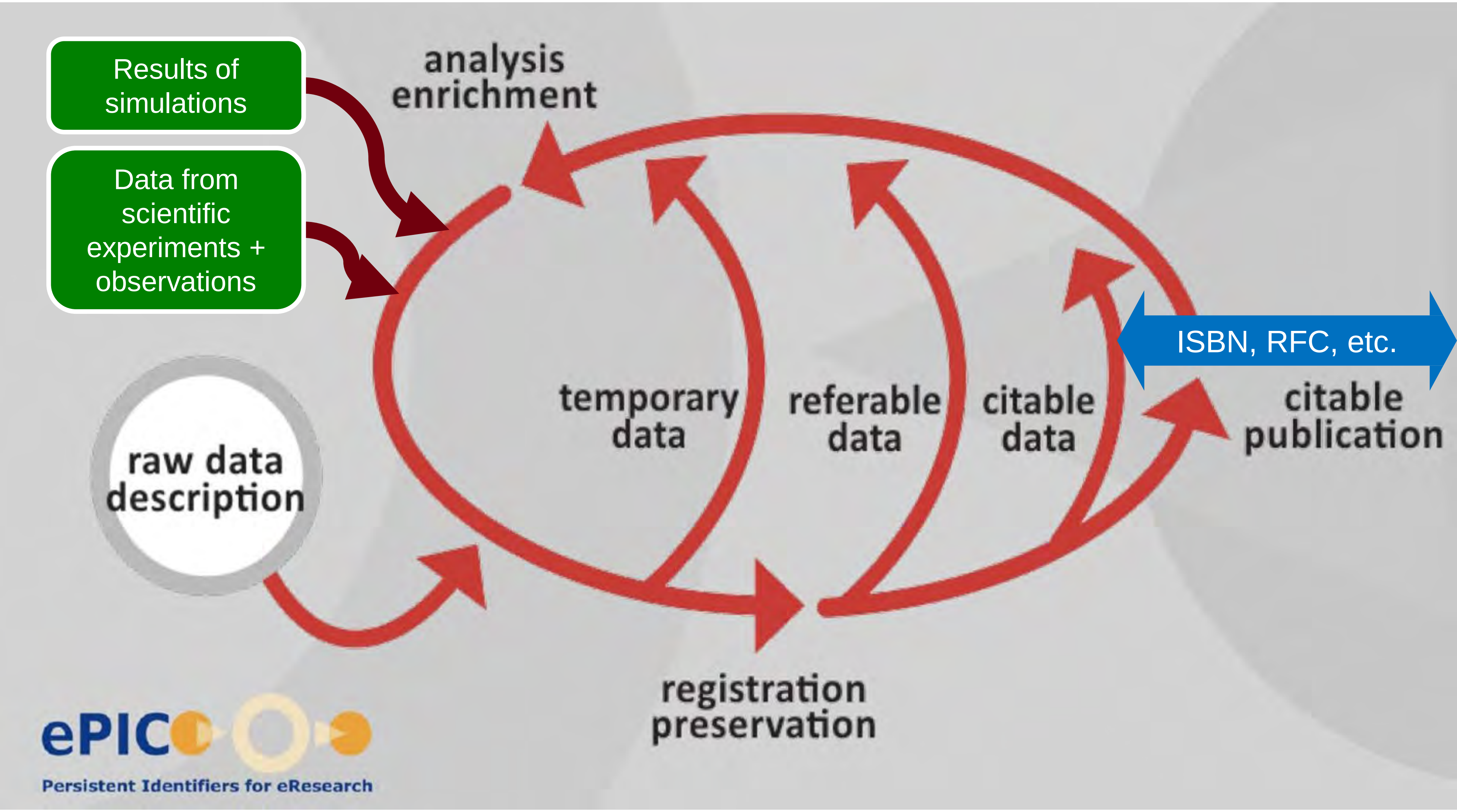
- ★ Make your stuff available on the Web (whatever format) under an open license
- ★★ Make it available as structured data (e.g., Excel instead of image scan of a table)
- ★★★ Use non-proprietary formats (e.g., CSV instead of Excel)
- ★★★★ **Use URIs to denote things, so that people can point at your stuff**
- ★★★★★ Link your data to other data to provide context



Citing Data in Science in all their instantiations



Citing Data in Science in all their instantiations

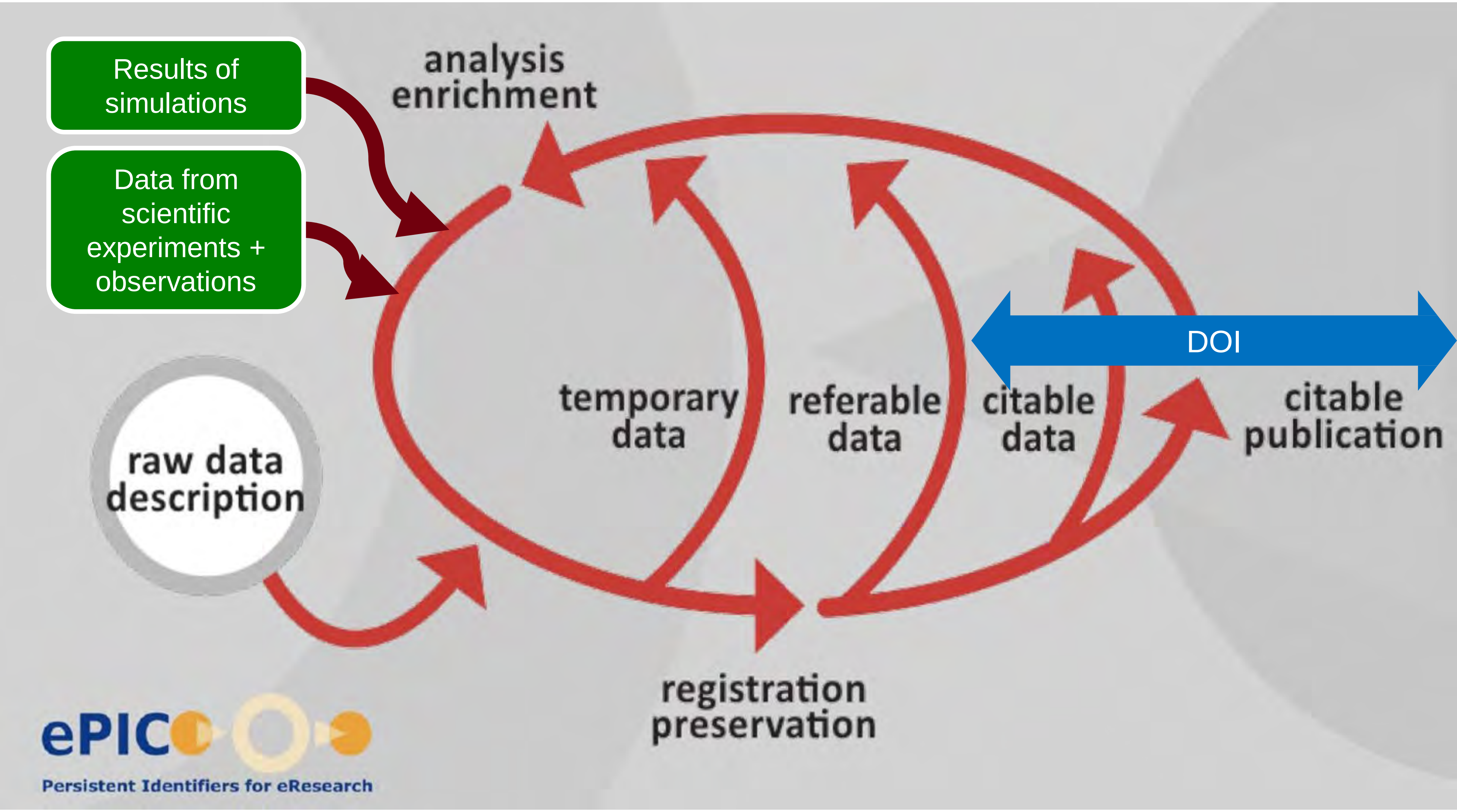


Results of simulations

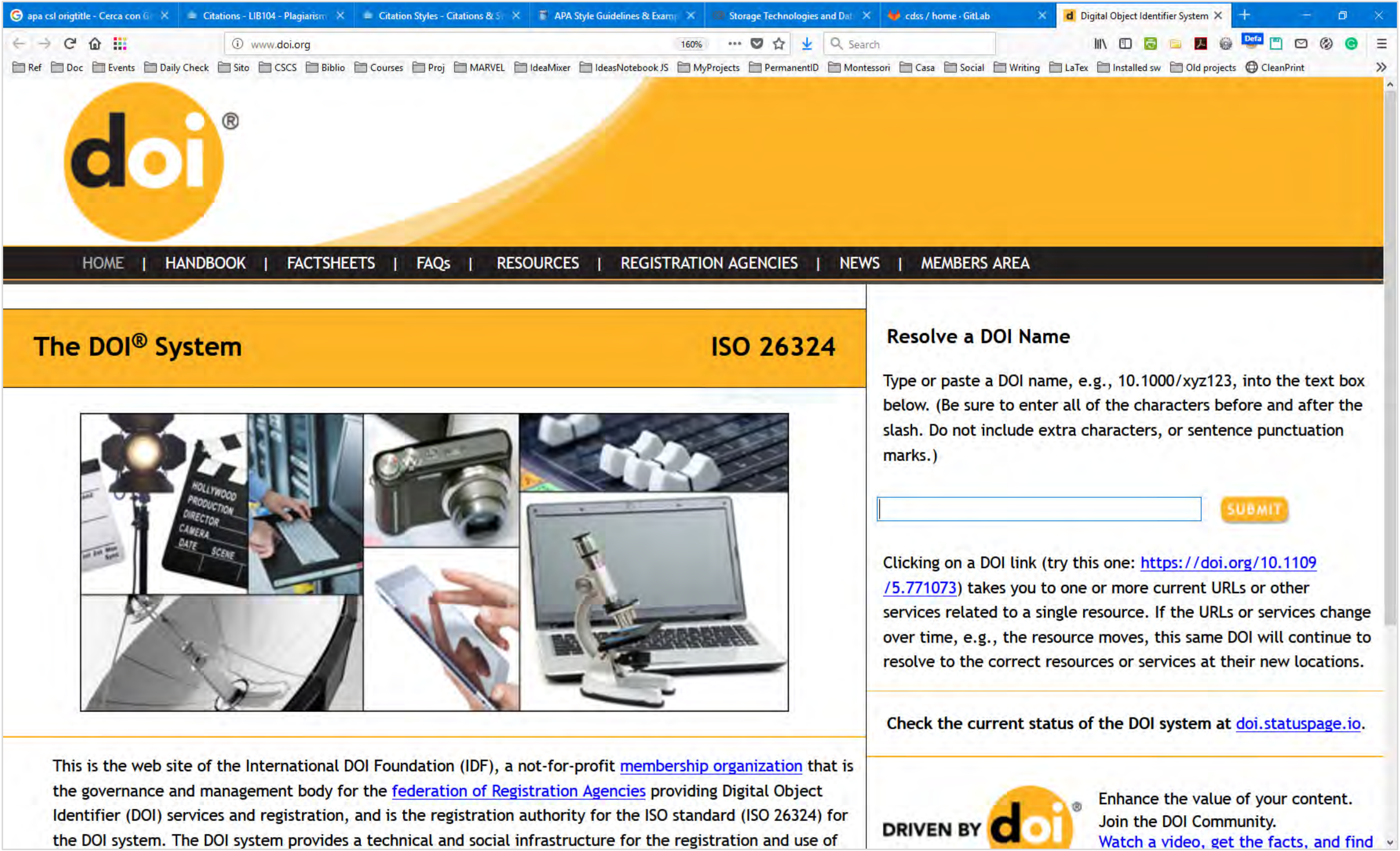
Data from scientific experiments + observations

ISBN, RFC, etc.

Citing Data in Science in all their instantiations



Publications solved these problems introducing DOI


apa csl origtitle - Cerca con G
Citations - LIB104 - Plagiarism
Citation Styles - Citations & S
APA Style Guidelines & Examp
Storage Technologies and Dat
cdss / home · GitLab
Digital Object Identifier System
www.doi.org
160%
Search
Ref
Doc
Events
Daily Check
Sito
CSCS
Biblio
Courses
Proj
MARVEL
IdeaMixer
IdeasNotebook JS
MyProjects
PermanentID
Montessori
Casa
Social
Writing
LaTex
Installed sw
Old projects
CleanPrint
doi®
HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | REGISTRATION AGENCIES | NEWS | MEMBERS AREA
HOLLYWOOD PRODUCTION DIRECTOR CAMERA DATE SCENE
SUBMIT
DRIVEN BY doi®

The DOI® System — ISO 26324

This is the web site of the 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, and is the registration authority for the ISO standard (ISO 26324) for the DOI system. The DOI system provides a technical and social infrastructure for the registration and use of

Resolve a DOI Name

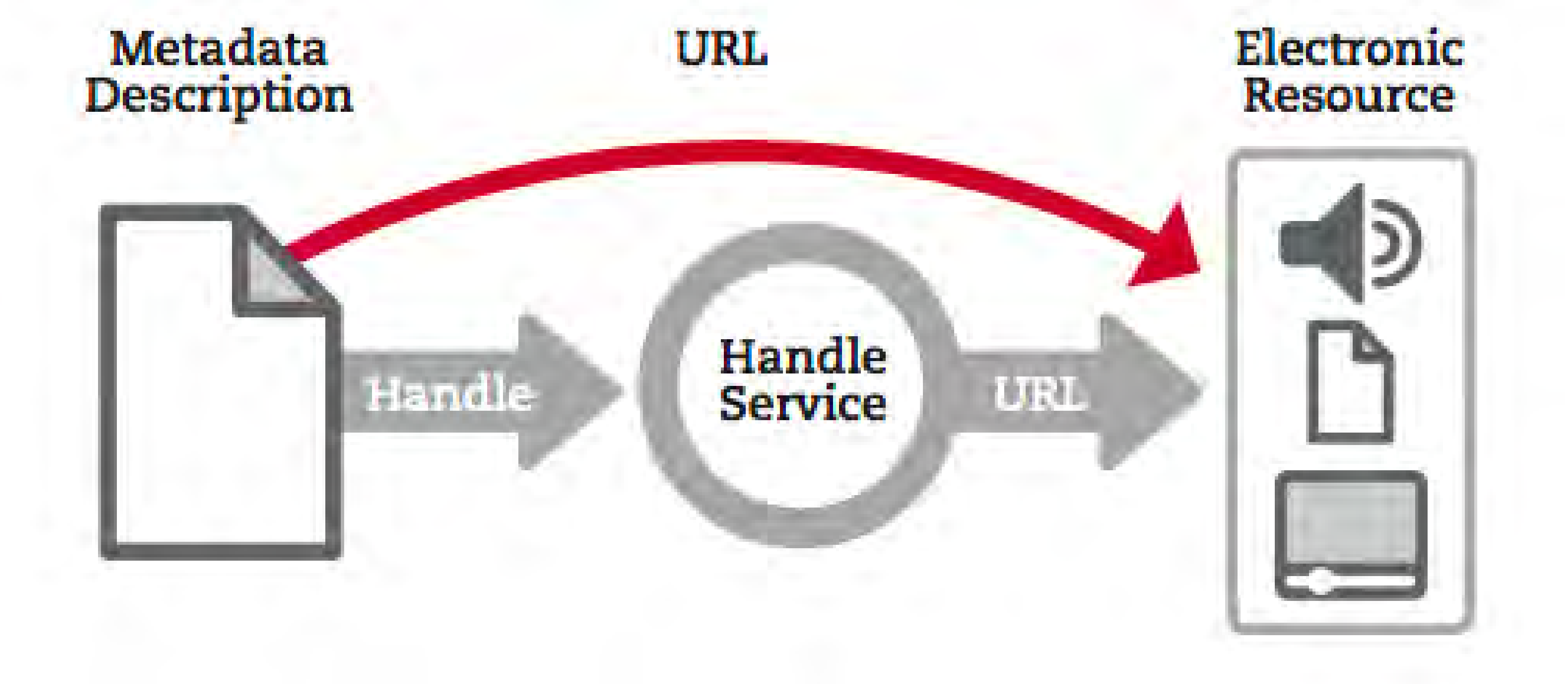
Type or paste a DOI name, e.g., 10.1000/xyz123, into the text box below. (Be sure to enter all of the characters before and after the slash. Do not include extra characters, or sentence punctuation marks.)

Clicking on a DOI link (try this one: https://doi.org/10.1109/5.771073) takes you to one or more current URLs or other services related to a single resource. If the URLs or services change over time, e.g., the resource moves, this same DOI will continue to resolve to the correct resources or services at their new locations.

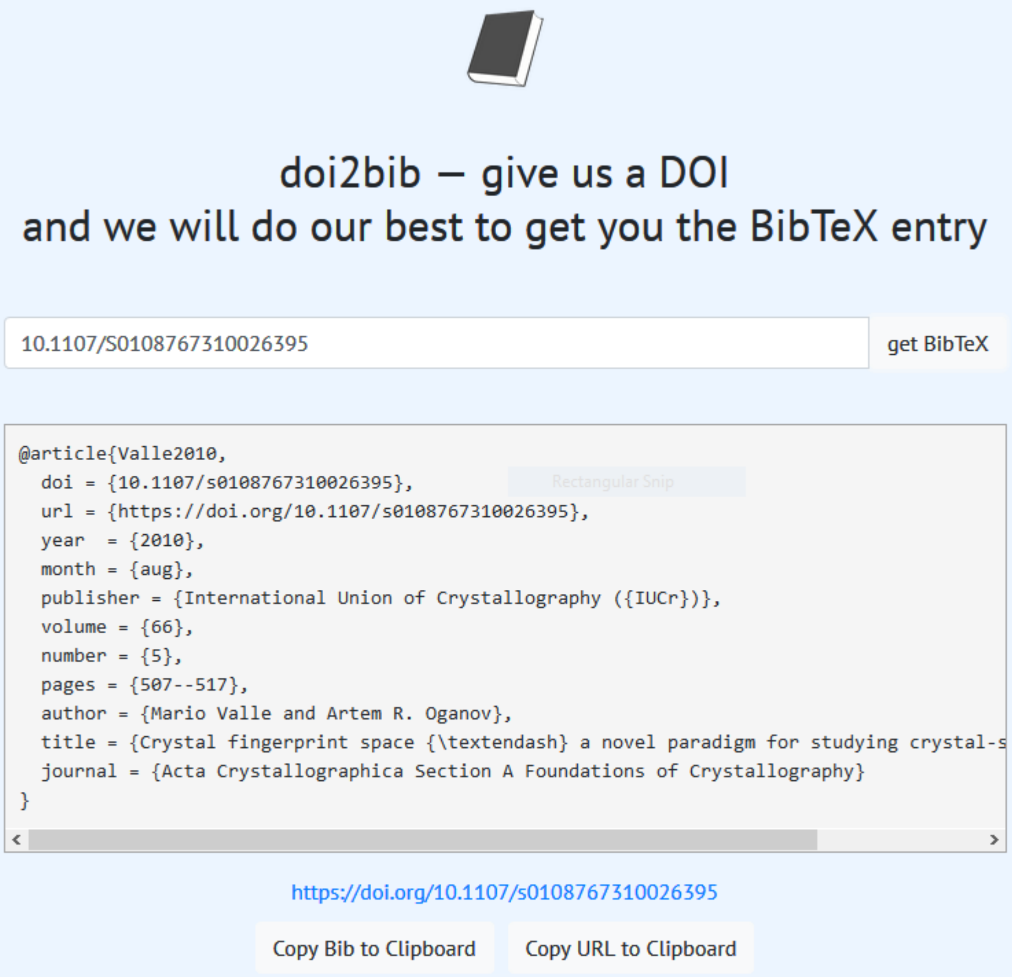
Check the current status of the DOI system at doi.statuspage.io.

Enhance the value of your content.
Join the DOI Community.
Watch a video, get the facts, and find

Base of every handle system (e.g., DOI)



DOI comes with an established set of metadata

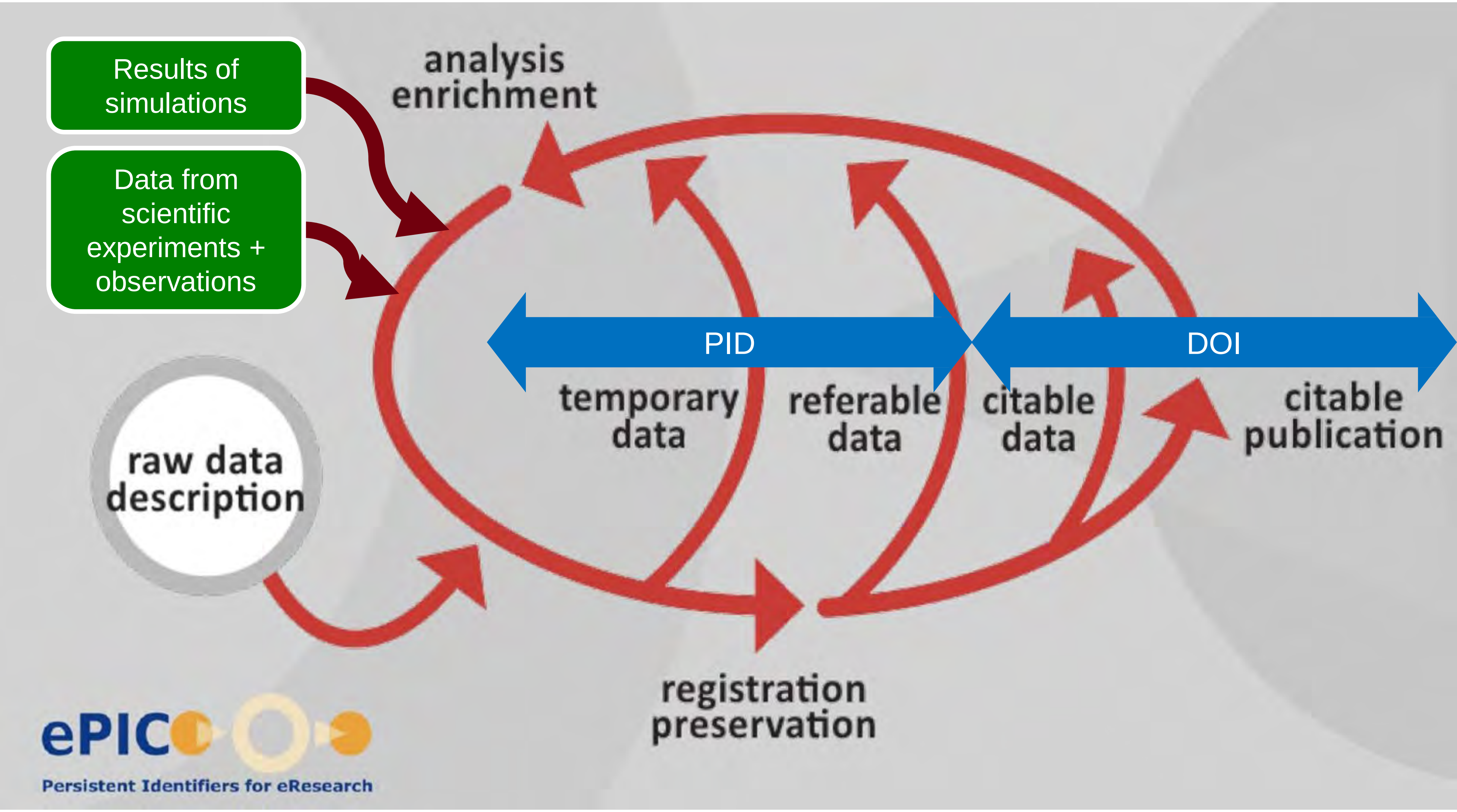


doi2bib – give us a DOI
and we will do our best to get you the BibTeX entry

```
@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

Citing the full spectrum of Scientific Data



Persistent Identifiers (PID) to cover the rest

- A Persistent Identifier (PID) identifies data objects regardless of their location, associate metadata to them and claim authorship.
- The PID infrastructure provides, at least, the following services:
  - Create PID and keep track of them.
  - Resolve a PID to the corresponding location.

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)

ePIC
Persistent Identifiers for eResearch

CLARIN
Common Language Resources and Technology Infrastructure

CNIC
Computer Network Information Center
Chinese Academy of Sciences

CSC

CSCS

DKRZ

grnet
Networking Research and Education

GWDG

SNIC

SURF SARA

CSCS will provide (March 2019) a service to generate and manage a certain range of PID assigned to Switzerland and to resolve any PID

Structure of a PID

- A PID is a string with the following structure:
  - `<PREFIX>/<SUFFIX>`
- `<PREFIX>`
  - `21.nnnnn`
  - Where “21.” identifies a PID (note that DOI starts with “10.”)
  - “nnnnn” five digits identifying the namespace (could be composed by country and institution IDs for example, but in general it is opaque)
  - CSCS manages the **21.17101** prefix
- `<SUFFIX>`
  - Can be any unique string inside the namespace. But preferred as: `PRE-0000-0000-0000-0-POST`
  - An optional PRE UTF-8 string
  - An UUID with check digit (Universally Unique Identifier. It enables distributed systems to uniquely identify information without significant central coordination).
  - An optional POST UTF-8 string

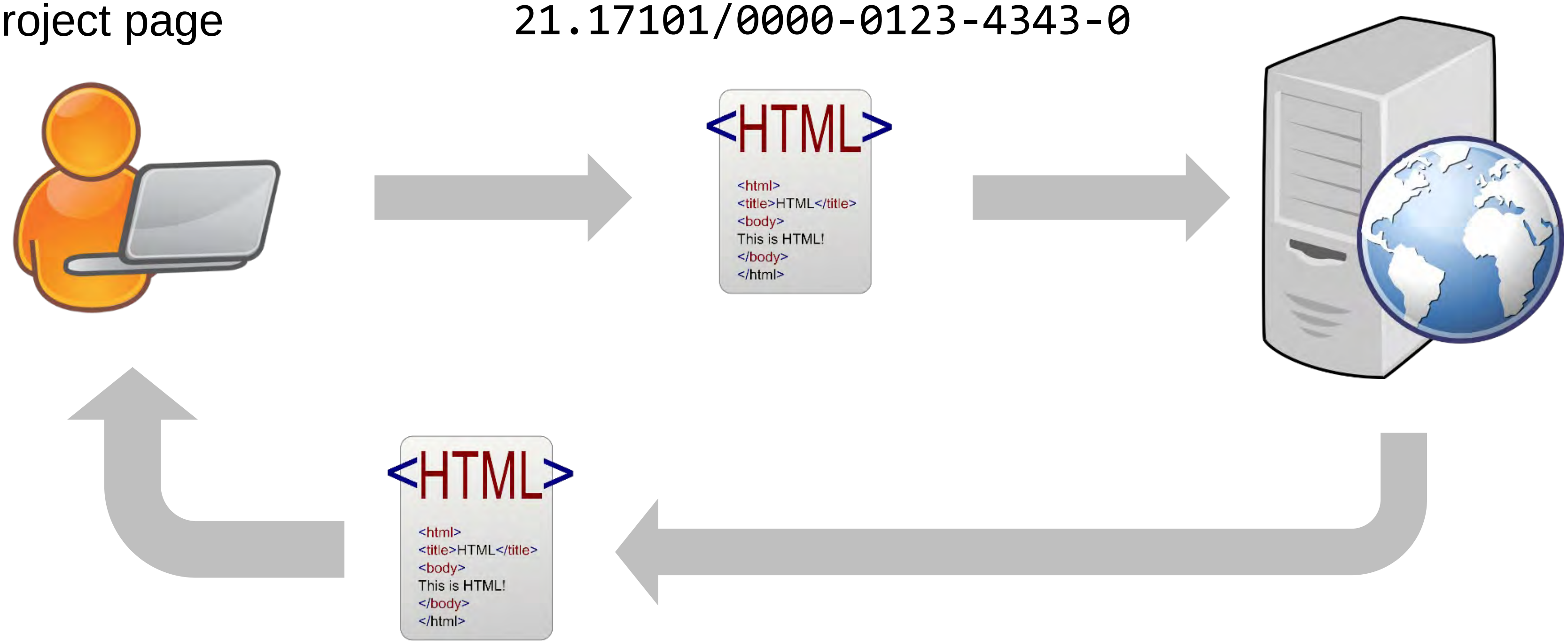
Temporary or Test PID

- We can generate and manage not only persistent PID, but also temporary (or test) PID
- DOI does not have this capability
- Only difference: the `<PREFIX>` format is `21.Tnnnnn`
- CSCS manages the **21.T17999** prefix for temporary PID
- The differences between Persistent and Temporary PID are:
  - A Persistent PID should always resolve to an URL. If the corresponding data has been removed, it should resolve to a page that states the data is missing. The PID itself could never be deleted.
  - A Temporary PID instead could be deleted anytime and normally has associated a TTL (Time To Live) value.

PID Resolution

User access some project page

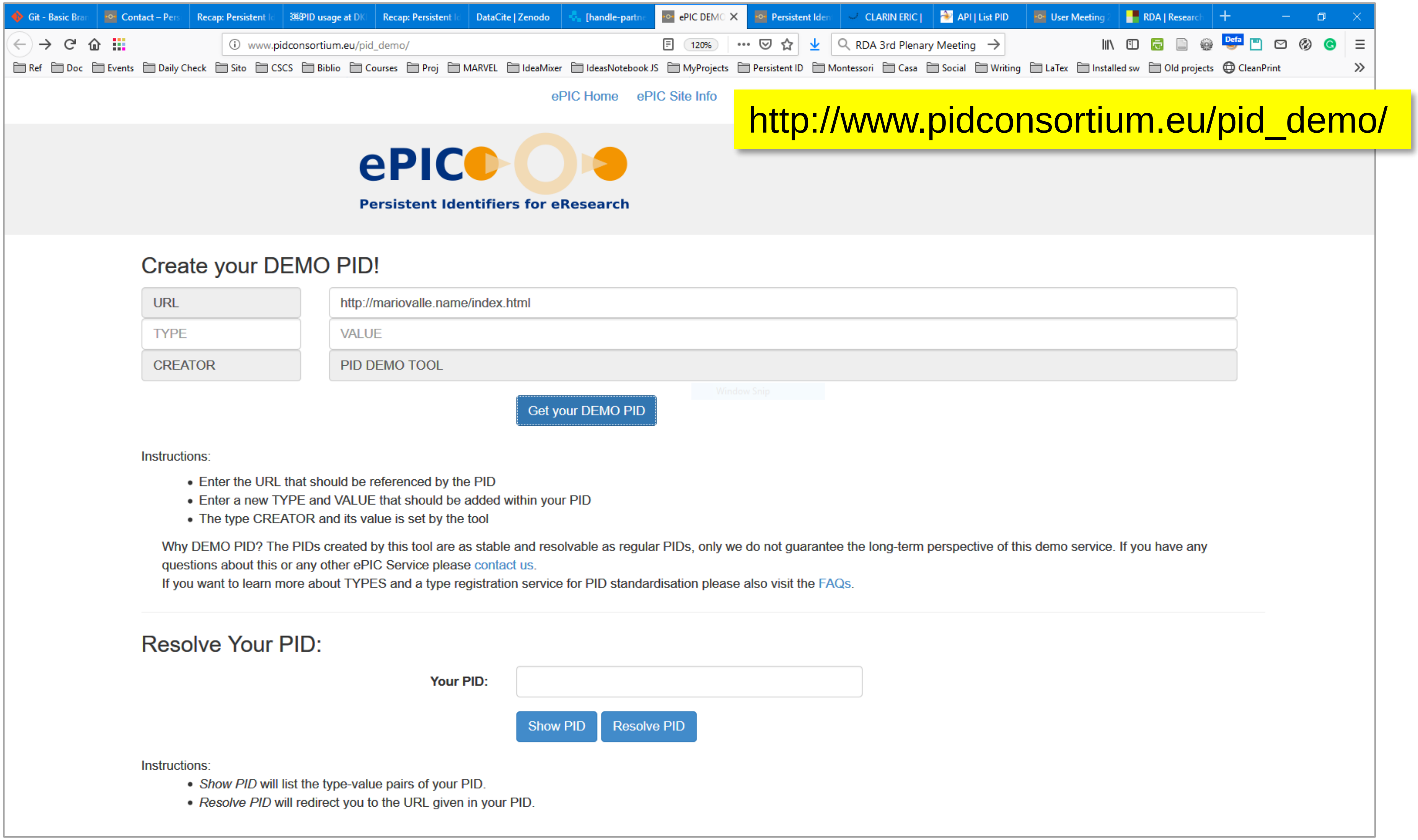
User clicks on a PID present there:
`21.17101/0000-0123-4343-0`



User download or access the data file from the page

Resolver returns and redirect the user to:
`https://cscs.ch/data/proj1/file.html`

PID Demo page



http://www.pidconsortium.eu/pid_demo/

Create your DEMO PID!

URL	http://mariovalle.name/index.html
TYPE	VALUE
CREATOR	PID DEMO TOOL

Instructions:

- Enter the URL that should be referenced by the PID
- Enter a new TYPE and VALUE that should be added within your PID
- The type CREATOR and its value is set by the tool

Why DEMO PID? The PIDs created by this tool are as stable and resolvable as regular PIDs, only we do not guarantee the long-term perspective of this demo service. If you have any questions about this or any other ePIC Service please contact us.
If you want to learn more about TYPES and a type registration service for PID standardisation please also visit the FAQs.

Resolve Your PID:

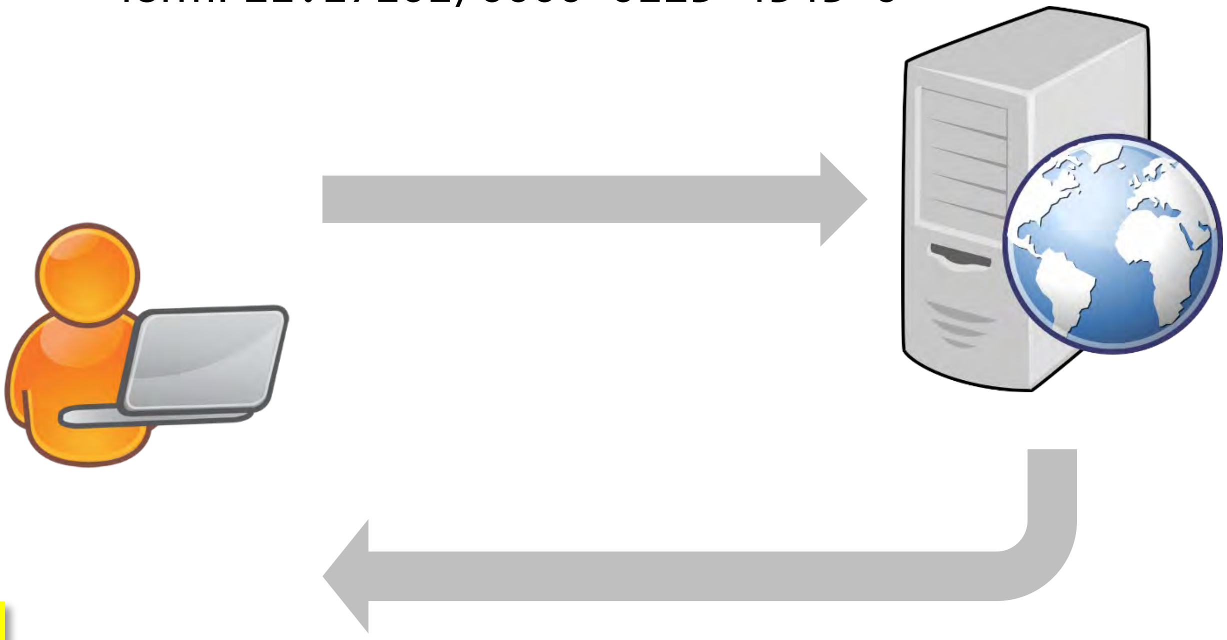
Your PID:

Instructions:

- *Show PID* will list the type-value pairs of your PID.
- *Resolve PID* will redirect you to the URL given in your PID.

PID Resolution

User enter a PID on the resolver web form: `21.17101/0000-0123-4343-0`

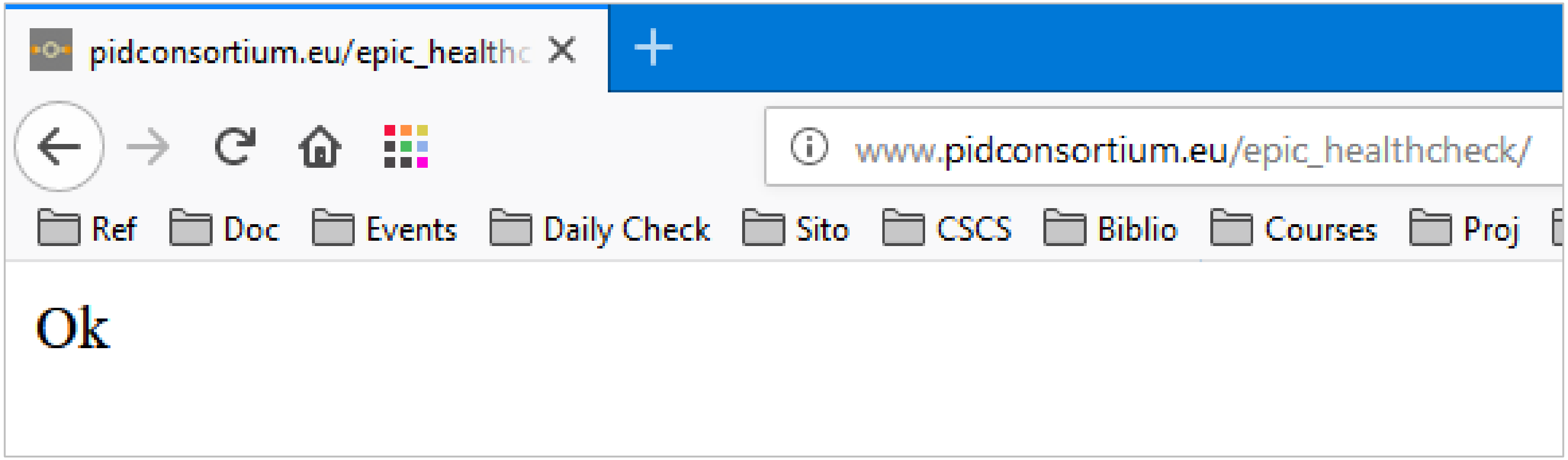


Note: These works also for non-CSCS managed prefixes.

Resolver returns: `https://cscs.ch/data/proj1/file.html`

The CSCS assigned prefix works

https://hdl.handle.net/21.17101/EPIC_HEALTHCHECK

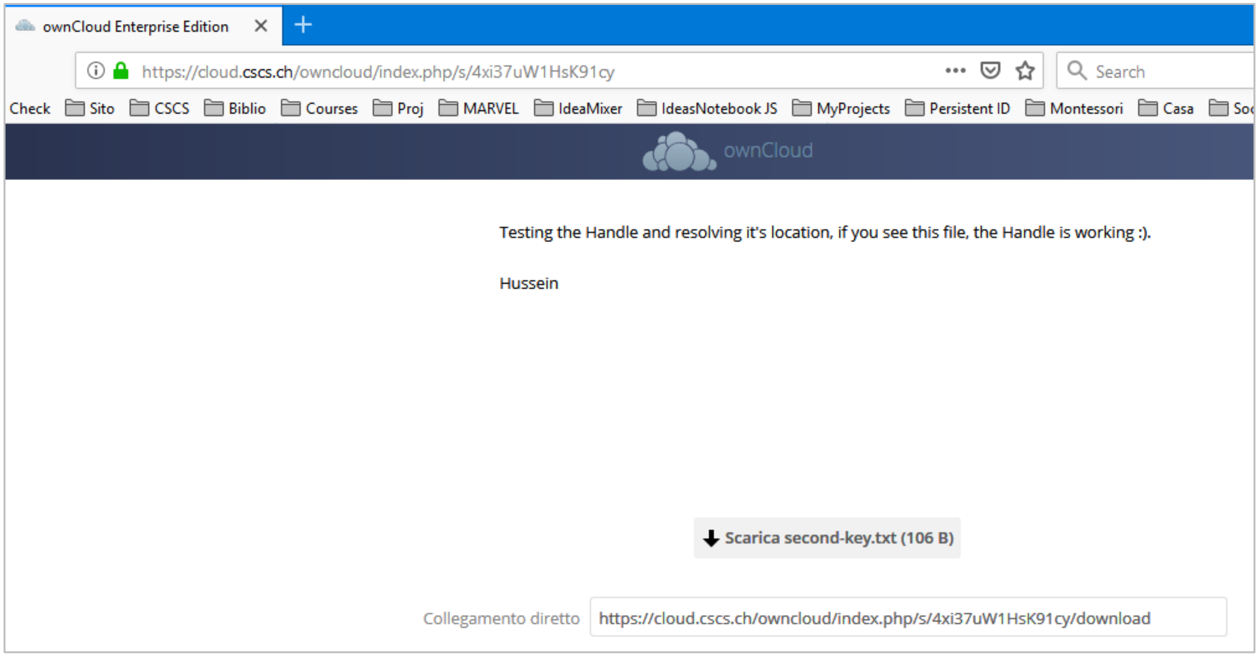


pidconsortium.eu/epic_healthc
www.pidconsortium.eu/epic_healthcheck/
Ref
Doc
Events
Daily Check
Sito
CSCS
Biblio
Courses
Proj

Ok

Also our temporary prefix works

```
https://hdl.handle.net/21.T17999/12345-54321
```


ownCloud Enterprise Edition
https://cloud.cscs.ch/owncloud/index.php/s/4xi37uW1HsK91cy
Search
Check
Sito
CSCS
Biblio
Courses
Proj
MARVEL
IdeaMixer
IdeasNotebook JS
MyProjects
Persistent ID
Montessori
Casa
Soc
ownCloud
Scarica second-key.txt (106 B)

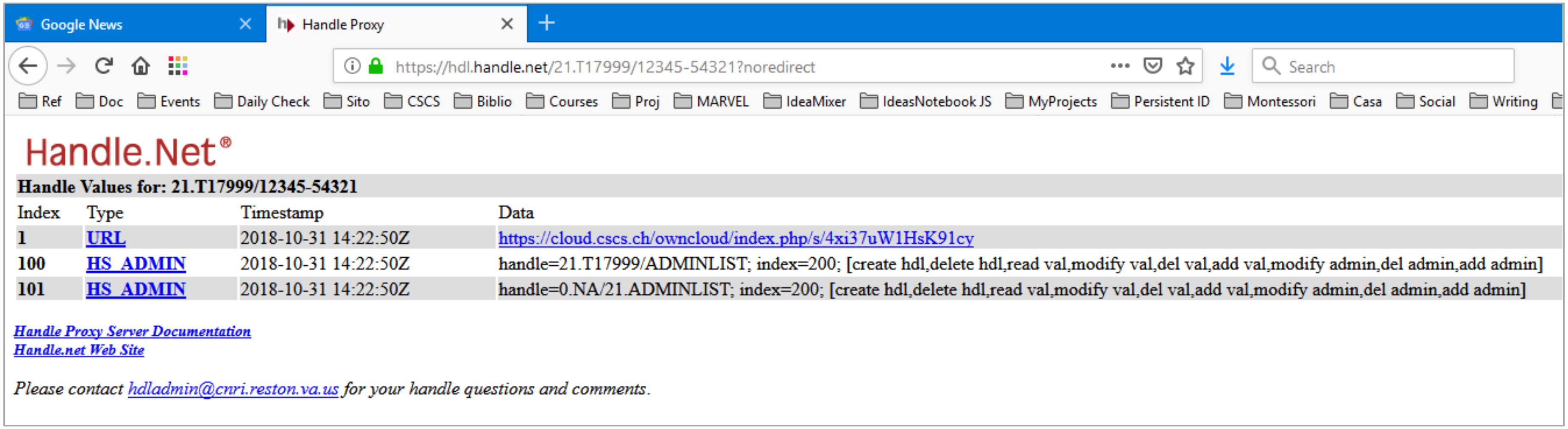
Testing the Handle and resolving it's location, if you see this file, the Handle is working :).

Hussein

Collegamento diretto https://cloud.cscs.ch/owncloud/index.php/s/4xi37uW1HsK91cy/download

What is the record behind this PID?

```
https://hdl.handle.net/21.T17999/12345-54321?noredirect
```



Handle.Net®

Handle Values for: 21.T17999/12345-54321

Index	Type	Timestamp	Data
1	URL	2018-10-31 14:22:50Z	https://cloud.cscs.ch/owncloud/index.php/s/4xi37uW1HsK91cy
100	HS_ADMIN	2018-10-31 14:22:50Z	handle=21.T17999/ADMINLIST; index=200; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]
101	HS_ADMIN	2018-10-31 14:22:50Z	handle=0.NA/21.ADMINLIST; index=200; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]

Handle Proxy Server Documentation
Handle.net Web Site

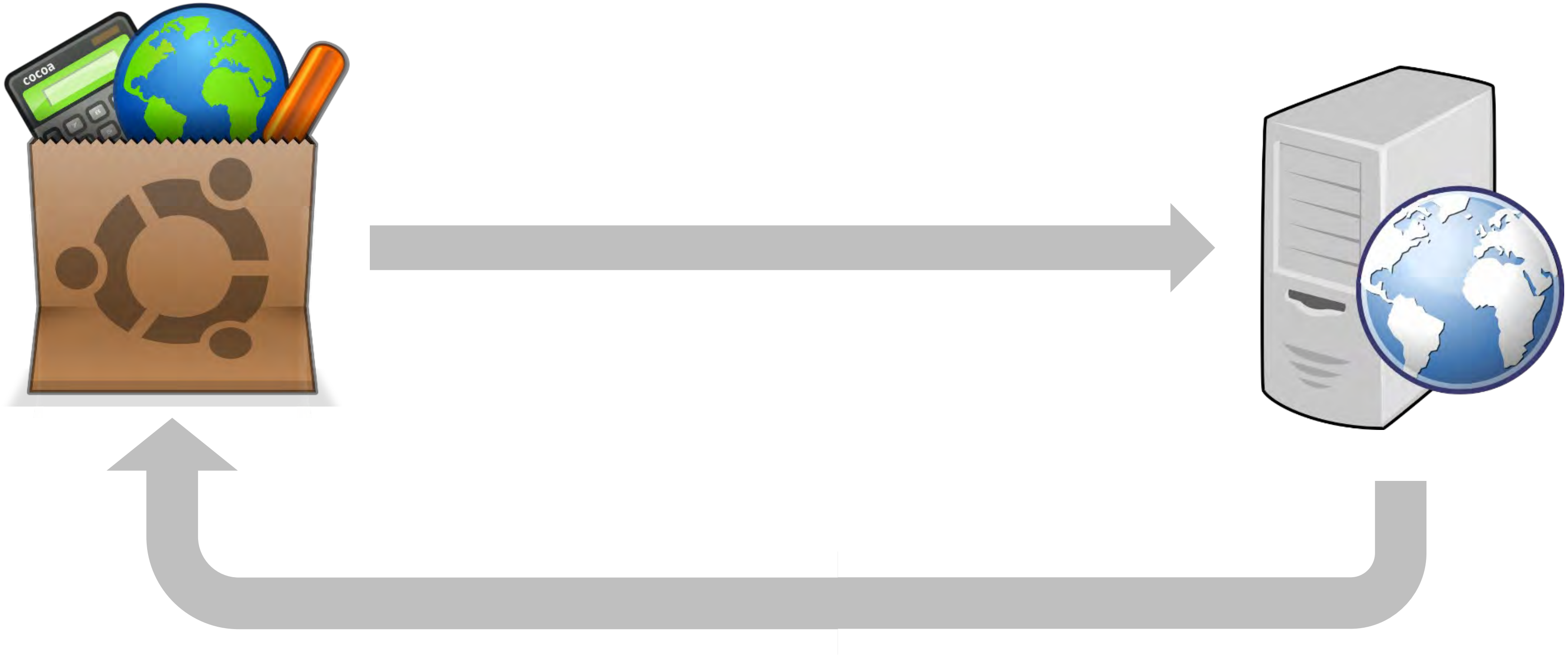
Please contact hdladmin@cnri.reston.va.us for your handle questions and comments.

Access through the Handle System API

```
$ 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"
    },
    ...
}
$
```

PID Resolution from API

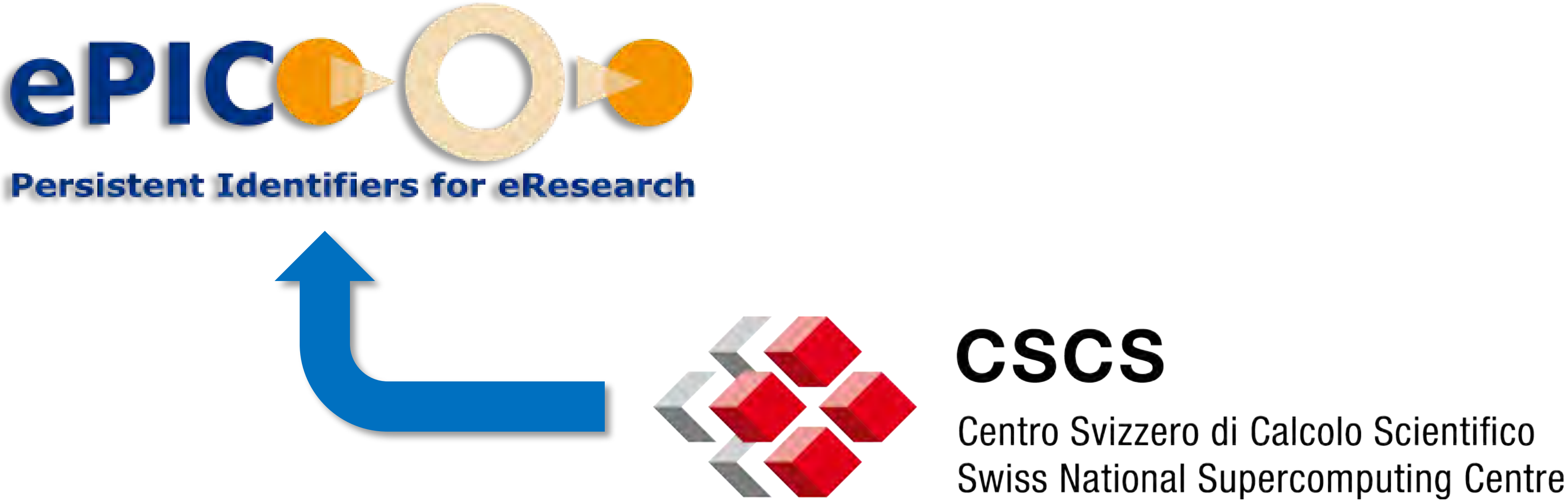
One application accesses resolver API via a GET request:
`https://hdl.handle.net/api/handles/21.T17999/12345-54321`
and ask for direct access to the data file



Application accesses the data file

Resolver API returns:
`https://cloud.cscs.ch/owncloud/...`

CSCS has a roadmap to comply with ePIC consortium requirements



ePIC

Persistent Identifiers for eResearch

CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

CSCS PID levels of service

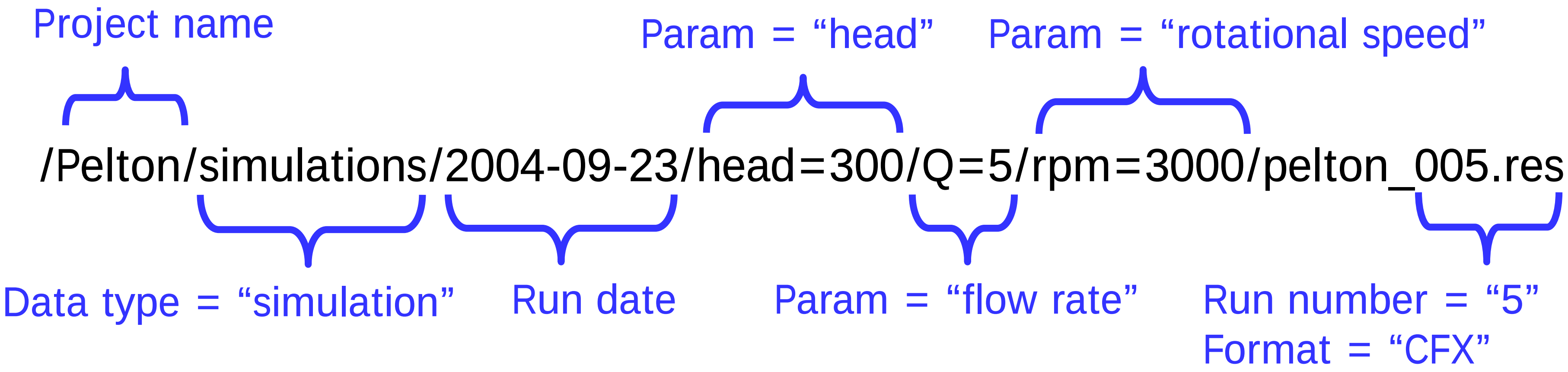
- **Level 1 – Basic PID creation/resolution**
  - End February 2019
  - PID creation initially in a CSCS namespace, plan to provide institution-specific namespaces
  - Resolution for any issued PID (not only from CSCS)
  - User editing of resolved URL and minimal metadata
  - Documentation and support
- **Level 2 – Storage at CSCS**
  - Tentatively June 2019
  - Persistent Identifiers demand Persistent Objects
  - CSCS provides a public, persistent storage space
  - Data ingested, for example, with a Dropbox-like mechanism (user deposits the file in a directory, and receives a PID for it).

CSCS PID levels of service (cont.)

- **Level 3 – Metadata search**
  - Not planned yet
  - The user could associate an ample set of metadata to a PID
  - The user can run queries on metadata to obtain a list of PID
- **Level 4 – Scientific Use Cases**
  - On going
  - Consultancy on specific Scientific Use Cases and HPC projects related to large amount of data
- **Level 5 – Future requirements**
  - On going
  - CSCS will track evolution of PID to be prepared and to implement new functionalities and services

A detour on the importance of metadata

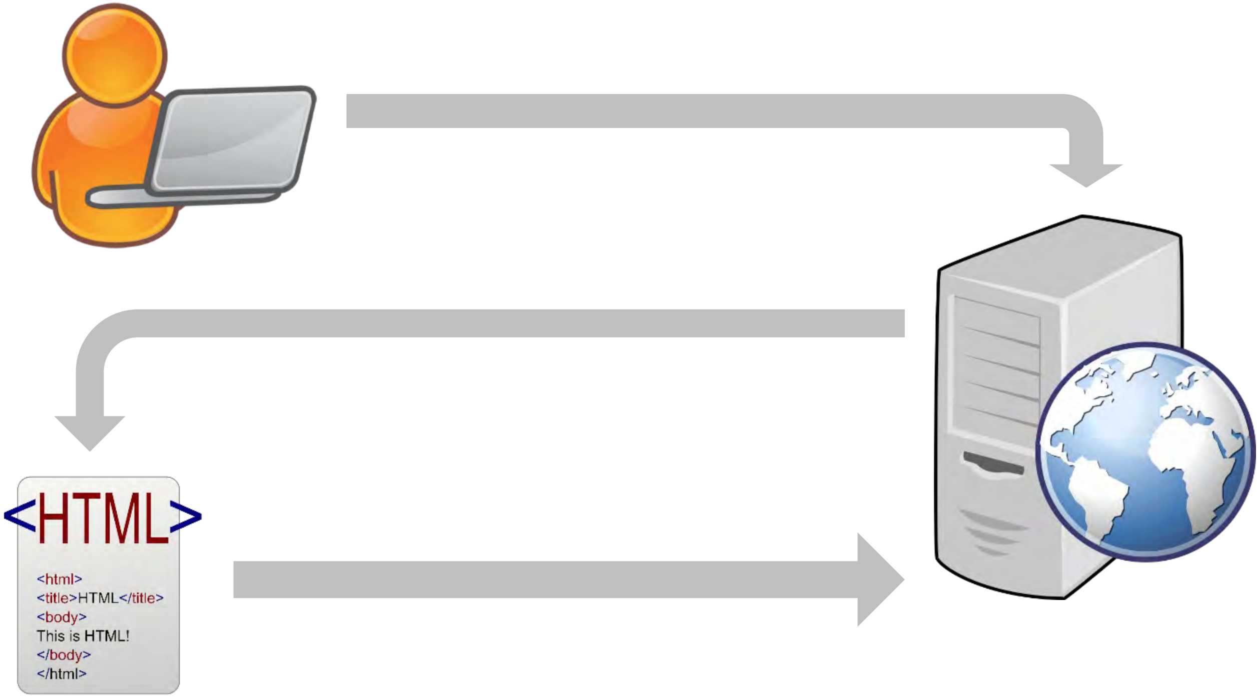
- Researchers try hard to record somewhere useful information about their data
- Metadata importance has grown so there is as much value in retrieving metadata as the object itself
- When I started at CSCS in my first project found those information recorded using very “ad-hoc” methods:



/Pelton/simulations/2004-09-23/head=300/Q=5/rpm=3000/pelton_005.res

PID Metadata Search & Resolution

User searches for PIDs on the resolver web form by entering:
`project=Climate&date=2009-09-09&var=ozone`



```
<html>
<title>HTML</title>
<body>
This is HTML!
</body>
</html>
```

The metadata catalog returns a list of PID

As before the user selects and retrieves the data file it is interested in

Few technicalities on metadata storage

- How metadata are stored could influence how they are used in applications
- SQL database (e.g., Postgress, MySQL)
  - Fixed schema
  - Tricks to store unlimited K/V pairs
    (TABLE mdataKey: key, mdataValue: value – many-to-many)
  - Query: SQL
- NoSQL database (e.g., MongoDB)
  - No schema
  - Metadata are JSON objects {pid: pid1, key1: value1, key2: value2, …}
  - Query: db.pids.find({key: value})
- Triple store (aka RDF databases e.g. Apache Jena)
  - Triples (<subj> <property> <object>) plus ontology (private or shared?)
  - Things identified by URI. **URI ⇔ https://resolver.cscs.ch/PID**
  - Query: SPARQL

The unpleasant side of PIDs

- The ePIC CSCS membership costs. Ergo, CSCS should operate this service at least recovering these costs (plus hw, personnel, machine time, etc.)
- Not yet defined what will cost and how much. But probably:
  - Creation/Resolution only: no fee
  - Then bundled inside persistent storage offering.
  - Idem for metadata management

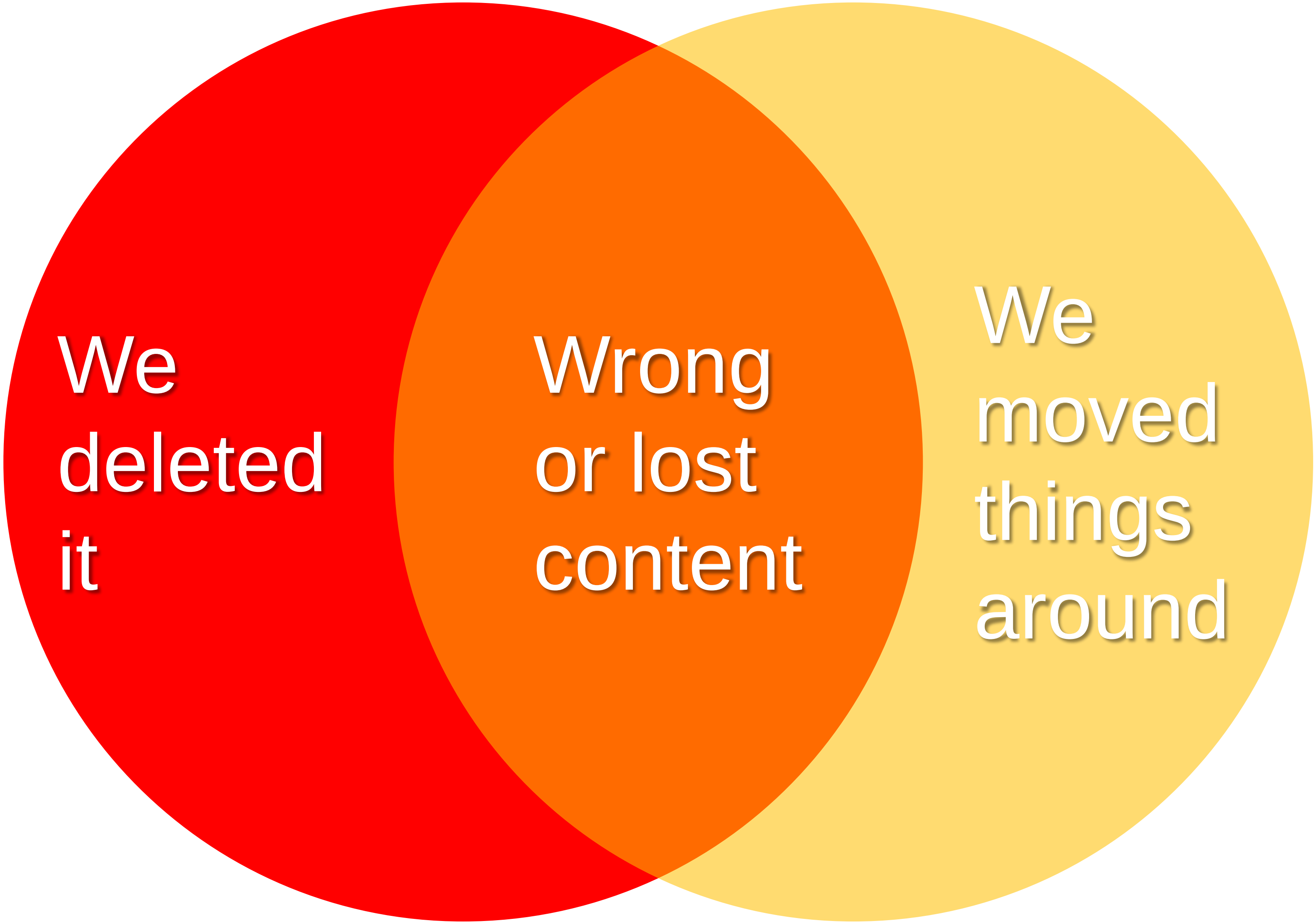
Collecting good, real life scientific use cases

- Integration with Provenance tracking
- Link component of an experiment in a Laboratory Notebook
- Integration with Workflow management
- Data publication and research validation
- Long term storage, migration from disk to tape (or openBIS → Repositories)
- Substituting custom references for data fragments (e.g., database record)
- Identifier for Docker images
- Identify standard training data for Deep Learning (e.g., the MNIST handwritten digits)

There are questions

- PID that resolve to multiple objects
- PID for resources with a registered datatype
- Should we insure the PID resolves? Should we insure the file has not changed?
- How to verify the PID has been created by who stated so?

A human problem needs a human solution

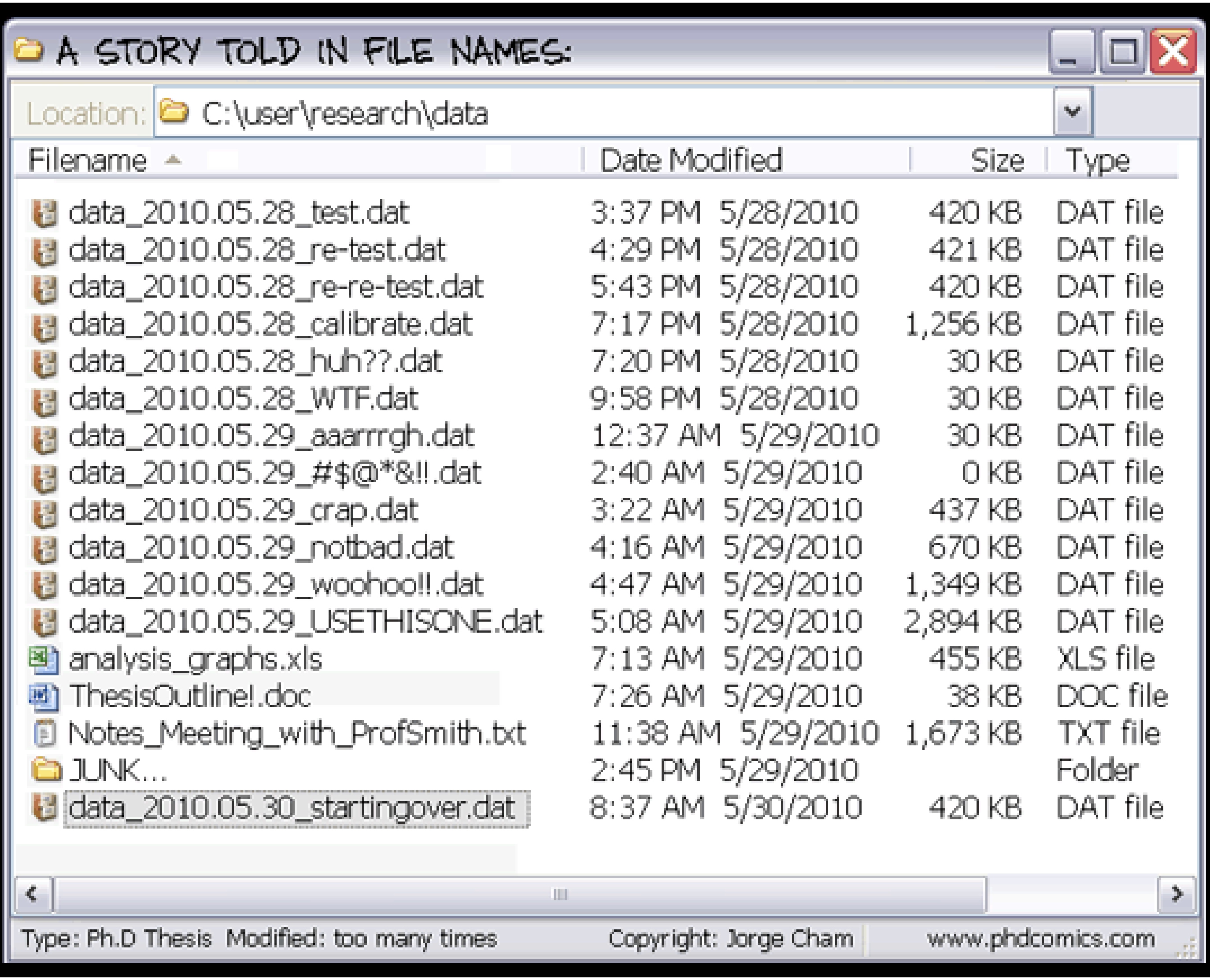


We deleted it

Wrong or lost content

We moved things around

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



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified		Size	Type
data_2010.05.28_test.dat	3:37 PM	5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM	5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM	5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM	5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM	5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM	5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM	5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM	5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM	5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM	5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM	5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM	5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM	5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM	5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM	5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM	5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM	5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

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

A human (cultural) problem needs a human solution

Data mining:

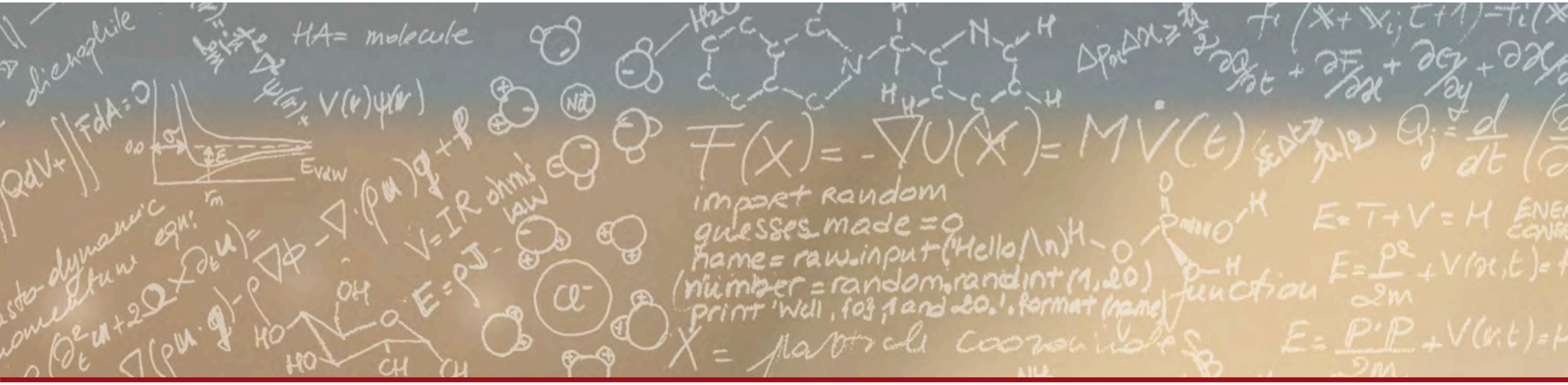
“my data is mine,
and your data is mine”

PID needs a social infrastructure

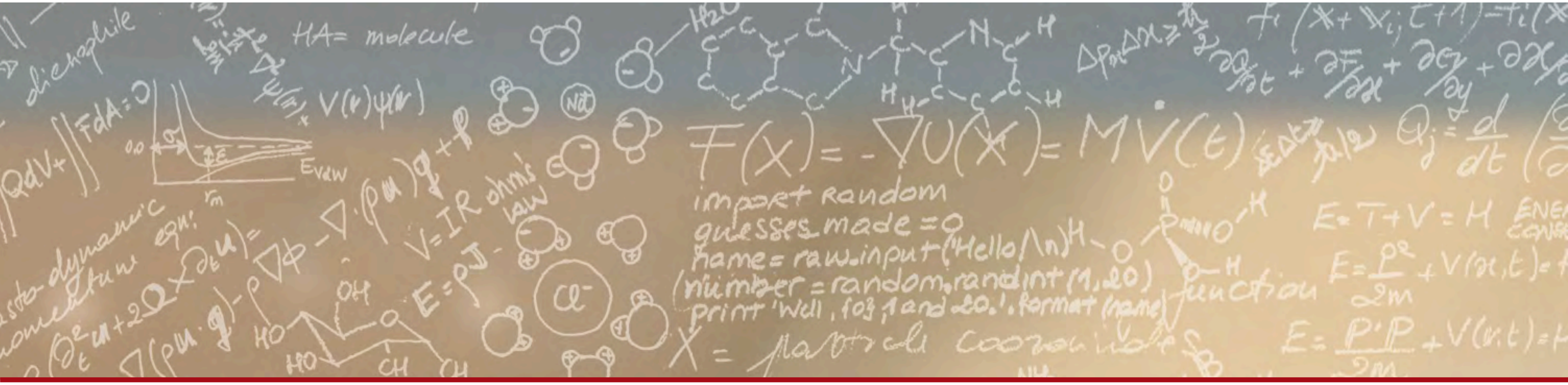
- PID Infrastructure maintained by a dedicated and reliable team
- Provided by a non-profit organization
- Governed by international boards
- Based on open standards

Creating awareness and community

- I’m the point of contact for PID ideas, suggestions and project specific requests
- I want to create awareness and hopefully create a Swiss community interested in this aspect of data management
- I’m collecting use cases to suggest how this technology could help Swiss scientist’s work


HA= molecule
V=IR ohm's law
import Random
guesses_made = 0
name = raw_input('Hello\n')
number = random.randint(1,20)
print 'Well, {0}, 1 and 20.'.format(name)
function
E=T+V=H

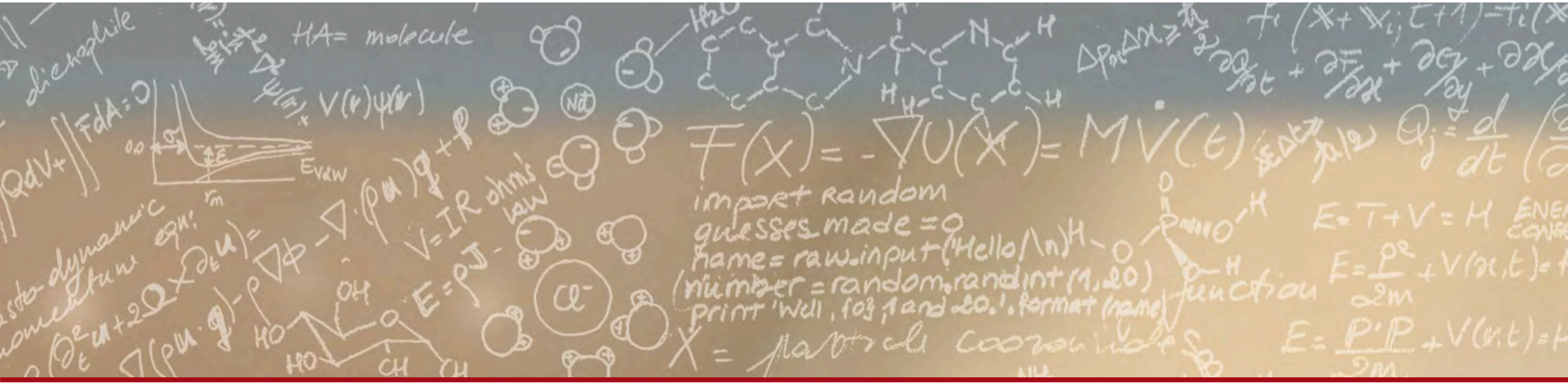
Thank you for your attention!


HA= molecule
V(r)ψ(r)
V=IR ohm's law
E=ρJ
import Random
guesses_made = 0
name= raw_input('Hello\n')
number = random.randint(1,20)
print 'Well, {0}, 1 and 20.'.format(name)
function
E= T+V = H

Now we have some time, so I am…

...awaiting your valuable contributions:

questions, curiosities, ideas,
something that resonates with your work...


HA= molecule
V=IR ohm's law
import Random
guesses_made =0
name= raw.input('Hello/\n)
number =random.randint(1,20)
print 'Well, {0}, 1 and 20.'.format(name)
function
E=T+V=H

Now we have truly finished.
Thank you for your contributions!