



PID-Systems for Digital Objects

Ulrich Schwardmann
GWDG, Göttingen, Germany



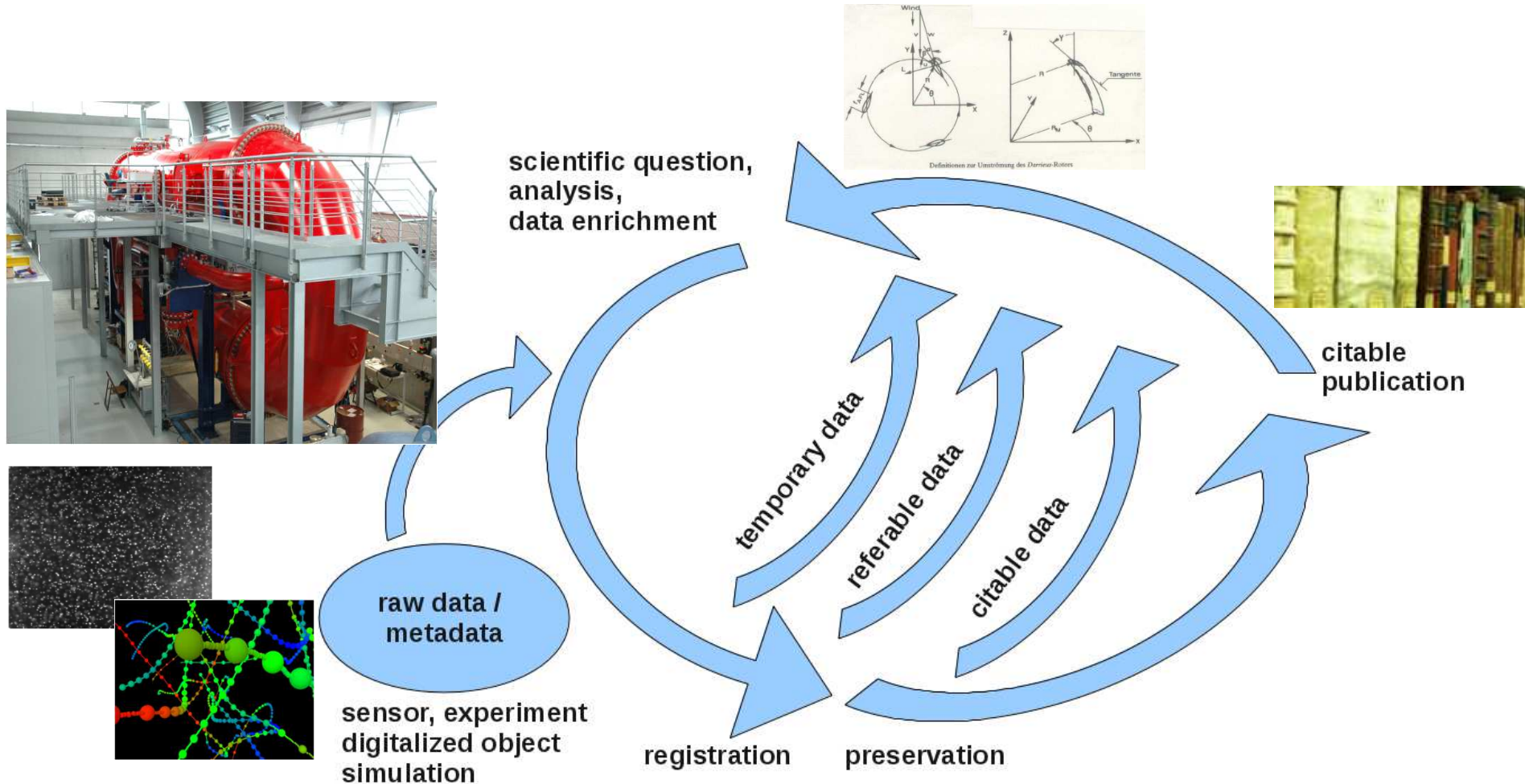


Outline of the talk

- Why PIDs?
 - Reuse in the Life Cycle of Data Objects
- How PIDs work?
 - Abstraction and Redirection
- What are the difficulties?
 - Ensure Resolution and Point to Objects
- Which system meets the criteria?
 - The handle system
- Who provides the infrastructure?
 - EPIC

Data Life Cycle in Scientific Work

Citation and Reuse





Persistent Identification for Data Sets

- Scientific data is often more than just a file
 - Series of sensor output, video, virtual machine, ...
 - Sensor output without gauge information is useless
- Metadata after registration
- Regard this collection of files as one object
- For reuse the object has to be located
 - Collaboration, data citation, LTP, ...

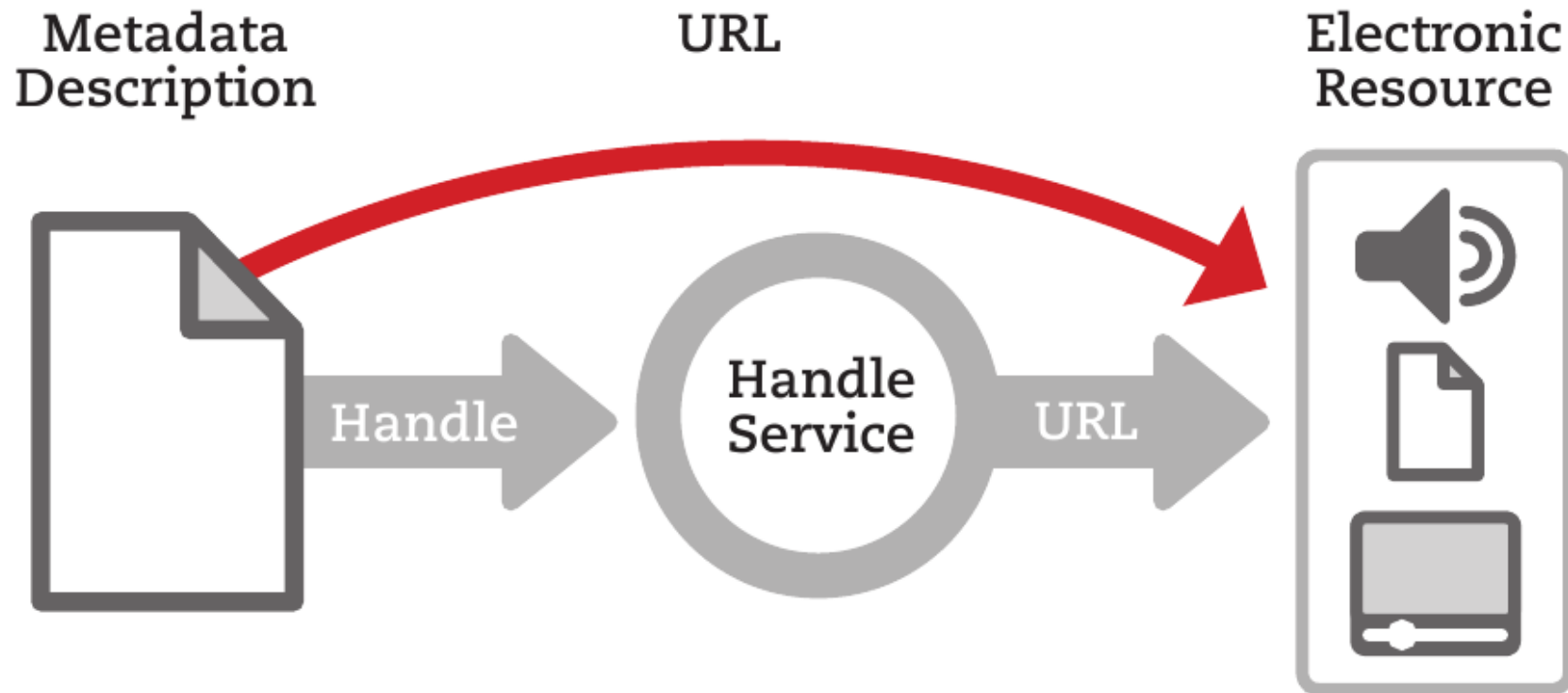


Persistent Identification, How?

- Main problem: the data location may change ...
- W3C: use URLs, but don't change the location
 - because: no central solution for the decentral WWW
 - called “cool URIs”: ten years stability?
 - !But! you should not use the exact place of a book in a shelf as the shelfmark
- PURL: persistent URLs, based on HTTP-redirection
 - no data object awareness
 - central solution or administration/ownership unsolved



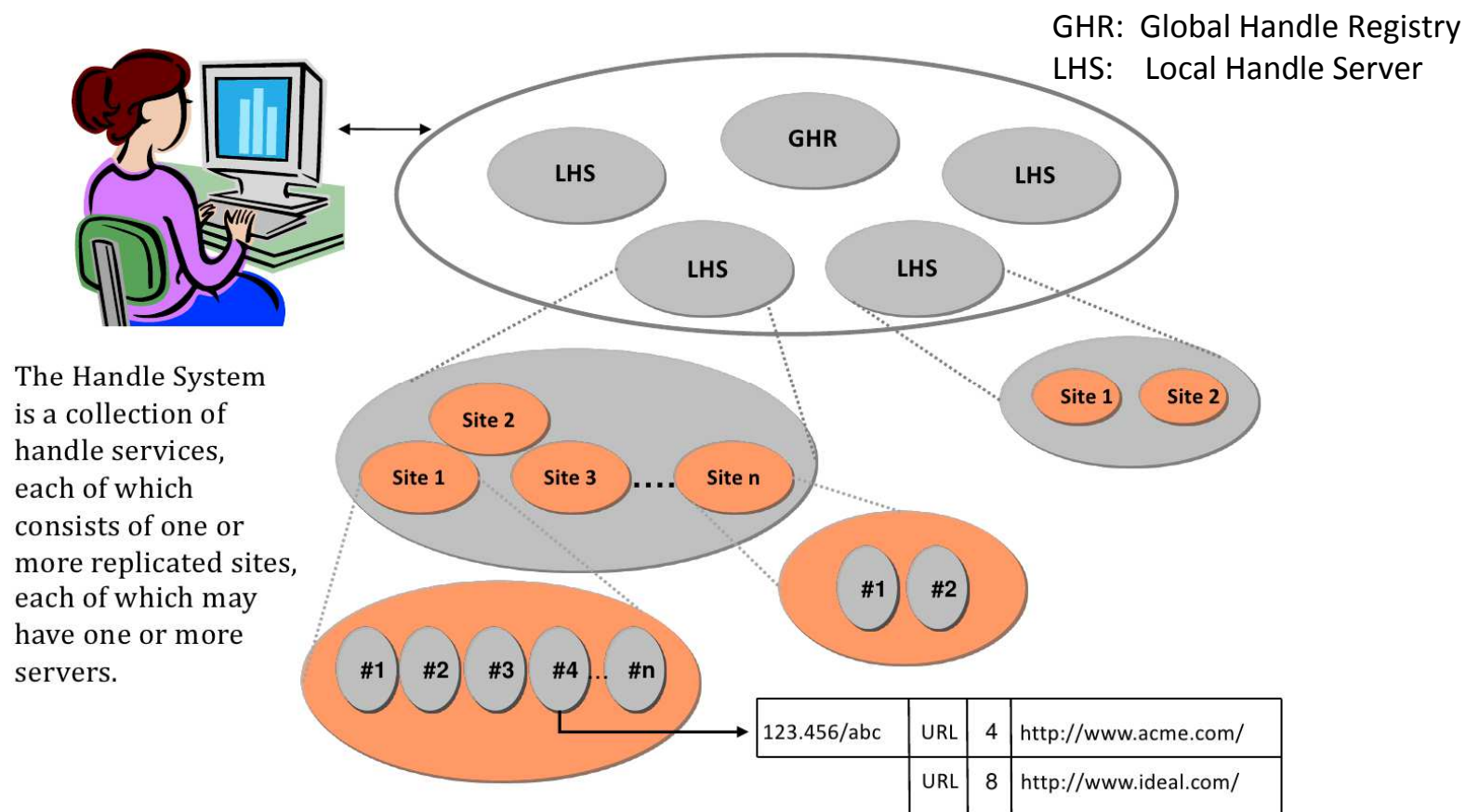
Redirection



- URN, ARK, Handle, DOI, PURL (by HTTP-redirect)
- Critical: Resolution
- Important: awareness of data object character

The Handle system

Handle Resolution





The Handle System

- Highly redundant resolution system
 - LHS resolves for one prefix
 - Secondary LHS is mirror of Primary
 - Answer given by the fastest LHS
- Single point of failure: GHR?
 - GHR is a cluster of servers
 - Secondary GHR at GWDG and in China (and Australia in near future)

Handle Persistent Identifier API

11858/00-ZZZZ-0000-0005-BD17-

Prefix / Suffix
A

Create Verbose Handle

PID request service

URL:
Filesize: (in bytes)
Checksum: (optional, see below)
Checksum is write-once, lower case hex: 'md5:32digits' or 'sha1:40digits'
Title:
Author(s):
Publication date: (see below)
Expiry date: (optional, see below)
Dates can be in format yyyy, yyyy-mm-dd or yyyy-mm-dd hh:mm:ss
Metadata URL:
Suffix: (user defined, optional)
Confirm in XML: ☒ (default: HTML)
Enter data above and

XML-Output of
Handle content

```
-<pidserviceresponse>  
  <action>view</action>  
  -<Handle>  
    <pid>11858/00-ZZZZ-0000-0005-BD17-A</pid>  
    <url>  
      http://www.gwdg.de/~parallel/parallelrechner/Hardware_Ueberblick.pdf  
    </url>  
    <size>-1</size>  
    <pubdate>2010-11-12</pubdate>  
    <title>Parallelrechner Hardware Ueberblick der GWDG</title>  
    <authors>Ulrich Schwarzmann</authors>  
    -<metadata_url>  
      http://www.gwdg.de/~parallel/parallelrechner/Hardware_Ueberblick.xml  
    </metadata_url>  
    <creator>demo2</creator>  
  </Handle>  
  <User>  
    <uid canCreate="no" canModify="no">anonymous</uid>  
  </User>  
</pidserviceresponse>
```

New faster and more generic API
is in preparation (Rel.: Q2 2012)



Handle: Features

- supports several fields
 - Can be used as pointer to data objects
 - allows pointers to replication sites
- supports pointers to sub-structures of objects called fragments:

11858/00-ZZZZ-0000-0005-BD17-A@type=pdf&page=1

resolves to

http://www.gwdg.de/Hardware_Ueberblick.pdf?type=pdf&page=1

which must be supported by the server of the resource



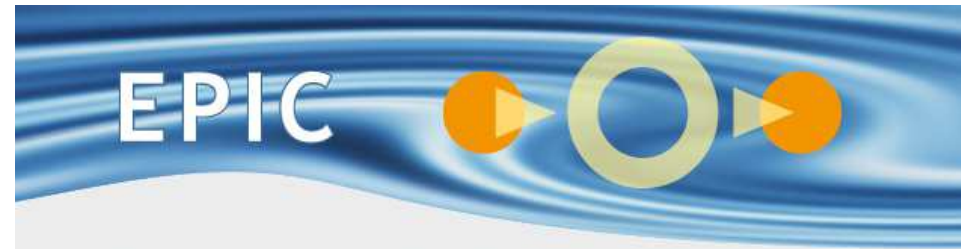
EPIC

Mission: ... setting up and maintaining a joint service with high availability for registering, storing and resolving persistent identifiers based on handles for the scientific community on a non-profit basis. (MoU)

Candidates:

DKRZ, CINECA, CINES, STFC, BSC, INGV, DESY, PSNC

Collaboration with: DataCite



EPIC Partners:



GWDG
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen



SARA
Reken- en Netwerkdiensten



CSC
IT Center for Science