PID System for eResearch

EPIC – the European Persistant Identifier Consortium

Ulrich Schwardmann

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Am Fassberg, 37077 Göttingen
ulrich.schwardmann@gwdg.de

IZA/Gesis/RatSWD-WS
Persistent Identifiers for the Social Sciences
Bonn, 2. Februar

PID System for eResearch

Content

1. Consortium for a PId System for eResearch
2. PIds 4 eResearch
3. Users and Usage
4. Conclusion and Outlook

EPIC

European Persistant Identifier Consortium

- is dedicated to providing a persistant identifier (PId) service
- main scope is European scientific and cultural heritage communities
- is a consortium of three mayor European scientific computing centers
  - with solid backing of national funding authorities
  - and long experience in providing reliable, safe and secure services and technical sustainability
  - all partners have a structure similar to a company
  - can provide SLAs
  - are involved in several big eScience projects
  - have signed a MoU to provide a PId system for the scientific community

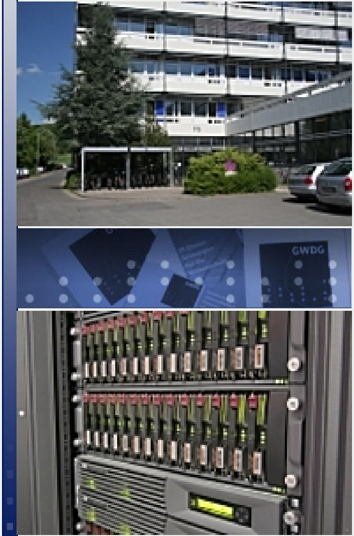
GWDG

Partners of EPIC

- GWDG is a corporate facility of the Max-Planck-Gesellschaft and the Georg-August University of Göttingen.
- for both it operates as a computer center, for the MPG it is furthermore IT competence center.
- GWDG was founded in 1970 as company.
- is located in Göttingen
- It operates on a non-profit principle
- 25,000 users
- 1000 scientific HPC users
- Staff: about 80 employees

GWDG

Partners of EPIC



- main topics
  - high performance computing
  - high performance networking
  - infrastructure services
  - IT consulting
- partner in several escience & grid projects
  - Dariah-DE
  - Clarin
  - D-Grid DGSI
- leading role in:
  - instant-grid
  - optinum-grid
  - goegrid
  - kopal

SARA

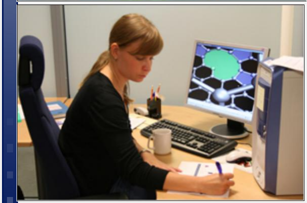
Partners of EPIC

sara
Computing & Networking Services

- SARA Computing and Networking Services is an advanced ICT service center
- that supplies – since more than 30 years – a complete package of
  - high performance computing and
  - visualization
  - high performance networking and
  - infrastructure services.
- is located in Amsterdam
- Among SARA's customers are the business community and scientific, educational, and government institutions.

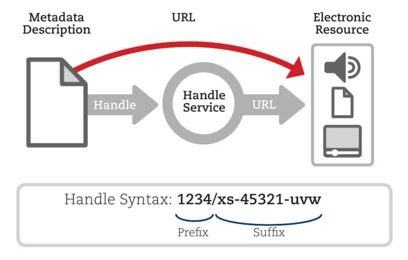
CSC

Partners of EPIC



- CSC, as part of the Finnish national research structure, develops and offers high-quality information technology services
- CSC founded in 1970, reorganized as a company in 1993
- Operates on a non-profit principle
- Facilities in Espoo, close to Otaniemi campus of Helsinki University
- Staff 180
- 3000 researchers use CSC's computing capacity

What kind of PIds provides EPIC?



- technology basis is the handle system
- the syntax therefore contains a prefix and a suffix
- a field in the suffix relates to a organisational unit
- no meaningful strings are involved
- the PId can be resolved:
  - by user transparent HTTP redirection to associated URL
  - by dedicated software embedded into client applications
- EPIC does not provide a repository for data and metadata

EPIC API for the creation of PIds

- realized as web page (https://handle.gwdg.de/pidservice/) and webservice (REST)
    - a user administration: realized as web page and interface to the backend data base
    - creation, modification and search of PIds
    - all requests as HTTP and XML response
- the EPIC PId contains additional auxiliary information mandatory
    - URL
    - author, title, creator
    - publication and expiration date
- not mandatory
    - meta data URL
    - checksum (MD5,SHA-1), file size
    - easy to implement: pointers to first, next, last version

How reliable is the EPIC service?

- basis is the handle system already used by many organisations
- the handle system exists since almost twenty years
- it is highly scalable and safe by the use of multiple local and global server
- a global handle server for Europe is established for Europe at GWDG
- the stability and funding of the partner organisations stands for a long term reliability

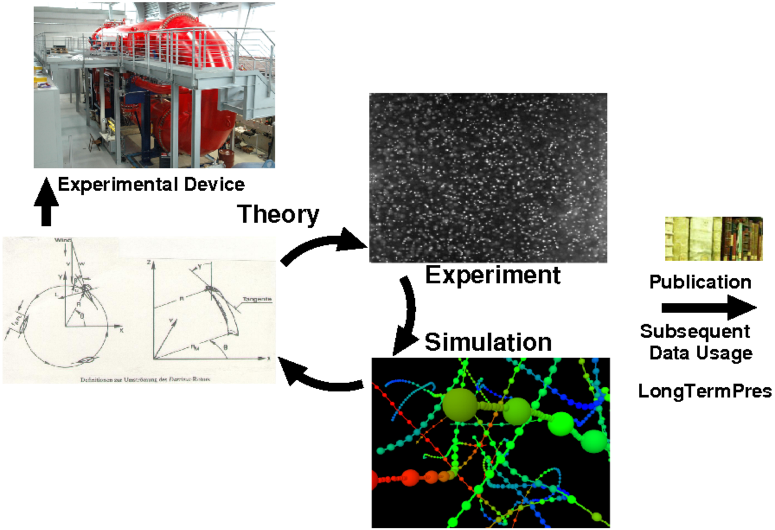
EPIC – what does it cost?

- the infrastructure and the cost should be completely under control of the scientific community
- at the moment there are no costs for the basic service
- the business model is based on COFUR:
  Cost Of Fulfilled User Request
- it is expected, that the service and infrastructure cost are neglectible (creation, resolution)
- software development for extension of the PId service API will be funded by projects or on the need of big institutions

User Communities of EPIC

- MPG, Max Planck Society
- CLARIN, Common Language Resources and Technology Infrastructure
- Dariah-DE, Digital Research Infrastructure for the Arts and Humanities
- SUB, Niedersächsische Staats- and Universitätsbibliothek Göttingen
- CATCH, Continuous Access To Cultural Heritage (no decision yet)
- ...

the scientific workflow of a wind channel



Experimental Device

Theory

Experiment

Simulation

Publication

Subsequent Data Usage

LongTermPres

the scientific workflow of archeological explorations

- archeological explorations are destructive
- each step has to be documented (protocols, recordings, scans, photographs)
- additionally there is increasingly more sensoric (seismic etc.) data
- these documents are more and more stored as digital data
- all these documents have to be identified uniquely
- again the choice and granularity of the objects identified by PIDs should be a scientific decision
- at one exploration site this could mean hundreds of PIDs per day.

persistance of data vs. identifier

- there is a growing amount of data in science
- scientists do not know a priori
  which data is worth to be kept
- a posteriori a persistent identifier for
  referenced data is certainly needed
- but before in their working groups they need to
  - uniquely identify the data
  - move the data to other places and responsibilities
- a priori the metadata generation can be automatized
  a posteriori this is much harder
- the PId can be a link between and reference for both
- PIds itsself are persistent, but they can be invalidated
  - if their data is never referenced by any published entity
  - this can be proven automatically in a digital world
  - this decision is part of the scientific workflow

benefits of PId for the scientific workflow

- the references can survive the whole scientific life cycle
- automatic processes can link data and metadata
- easy references for collaborative work
- easy references for archiving
- automatic processes can aid the decision about which data can be thrown away

prerequesites of PId for the scientific workflow

PId are and have to be part of the scientific process

- choice and granularity of PId is a scientific question
- this decision is only possible with very cheap PId
- because lots of them are created and most potentially wasted
- the costs has to be completely under scientific control
- reliability and security is a crucial matter

Future of PId

a personal opinion

probably there will be several PId systems and several ID schemes for different purposes and communities

- but all will share common principles:
  - redirection for location independence
  - heterogineity of access to (meta-)data
  - reliable institutional backing
  - open source software basis
  - hierarchical but decentralized resolution
- they will differ in
  - their requirements for persistency of the underlying data
  - their identifier syntax
  - their cost and business model
- possible(??): a common standardized resolution process and API

Outlook for EPIC

what has to be done additionally in the future:

- unify PID service API of different existing prefixes
- more detailed API specification
- verify URLs in PIDs (checksum and crawler)
- fragment/parameter support (comes with handle v7.0)
- versions support
- multiple URLs per PID (easier with handle v7.0)
  - identify same content with multiple resolutions
- batch operations
- support integration and migration of existing collections

Thanks for your attention

http://pidconsortium.eu

EPIC User Forum
Amsterdam, Middle of April

Questions ??