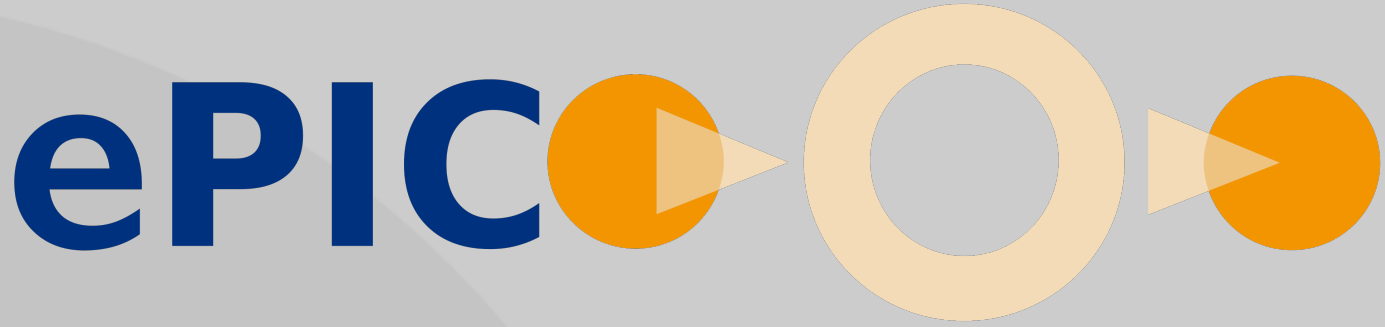
ePIC - Persistent Identifiers for eResearch

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
(GWDG)

Am Fassberg, 37077 Göttingen
ulrich.schwardmann [at] gwdg.de

21 September 2015, Paris

ePIC
Persistent Identifiers for eResearch



Persistent Identifiers for eResearch

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

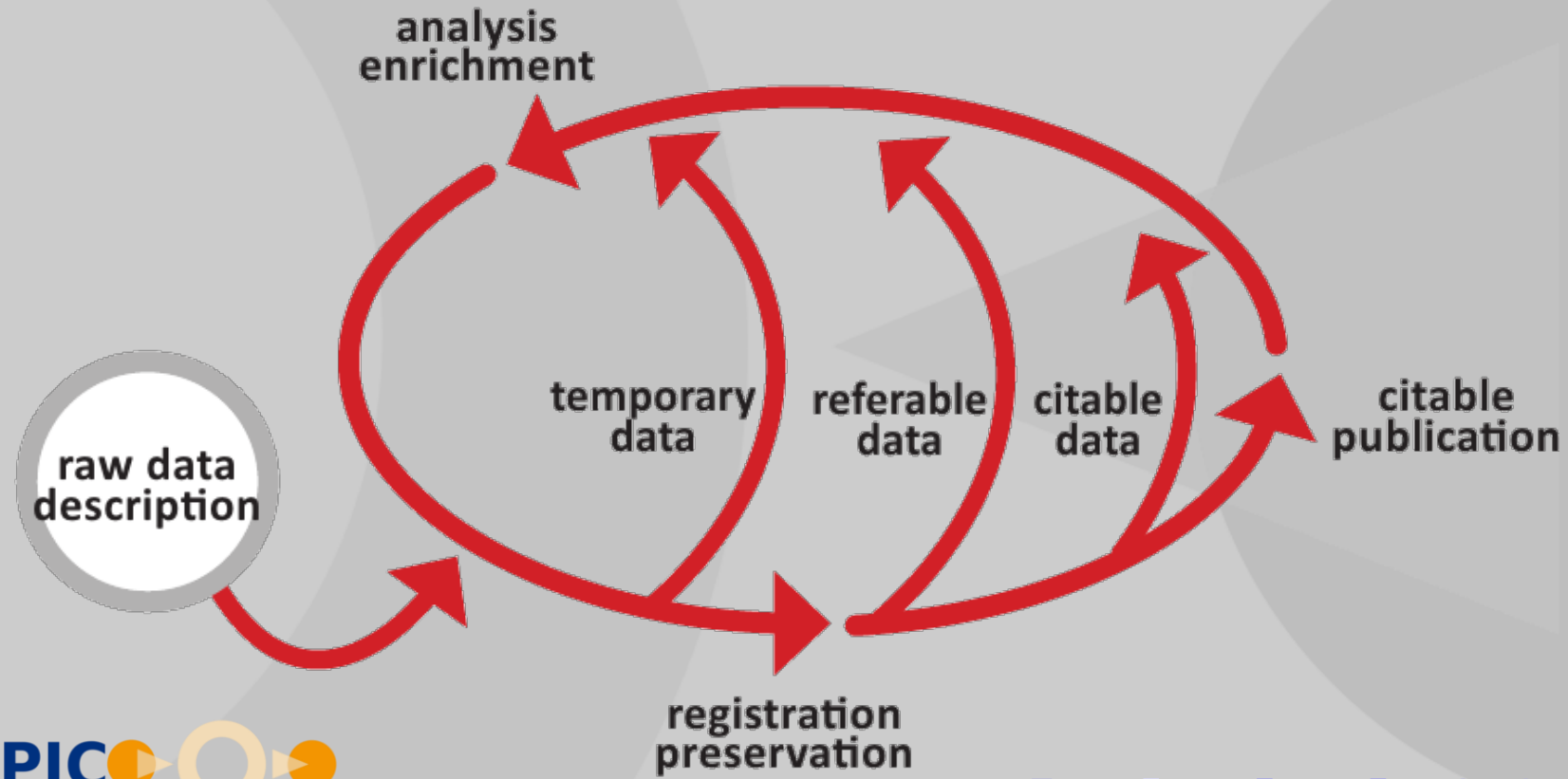
Luce (2008): *... eResearch refers to the development of, and the support for, advanced information and computational technologies to* **enhance** *all phases of research processes. ...*

ePIC provides **enhanced** Persistent Identifier
in order to **accelerate** the Research Life Cycle

The Research Data Life Cycle

data intensive research is highly collaborative

- scientists share data already in an early state
- ad hoc techniques for sharing are often prohibitive
- reliable references can accellerate the Research Life Cycle



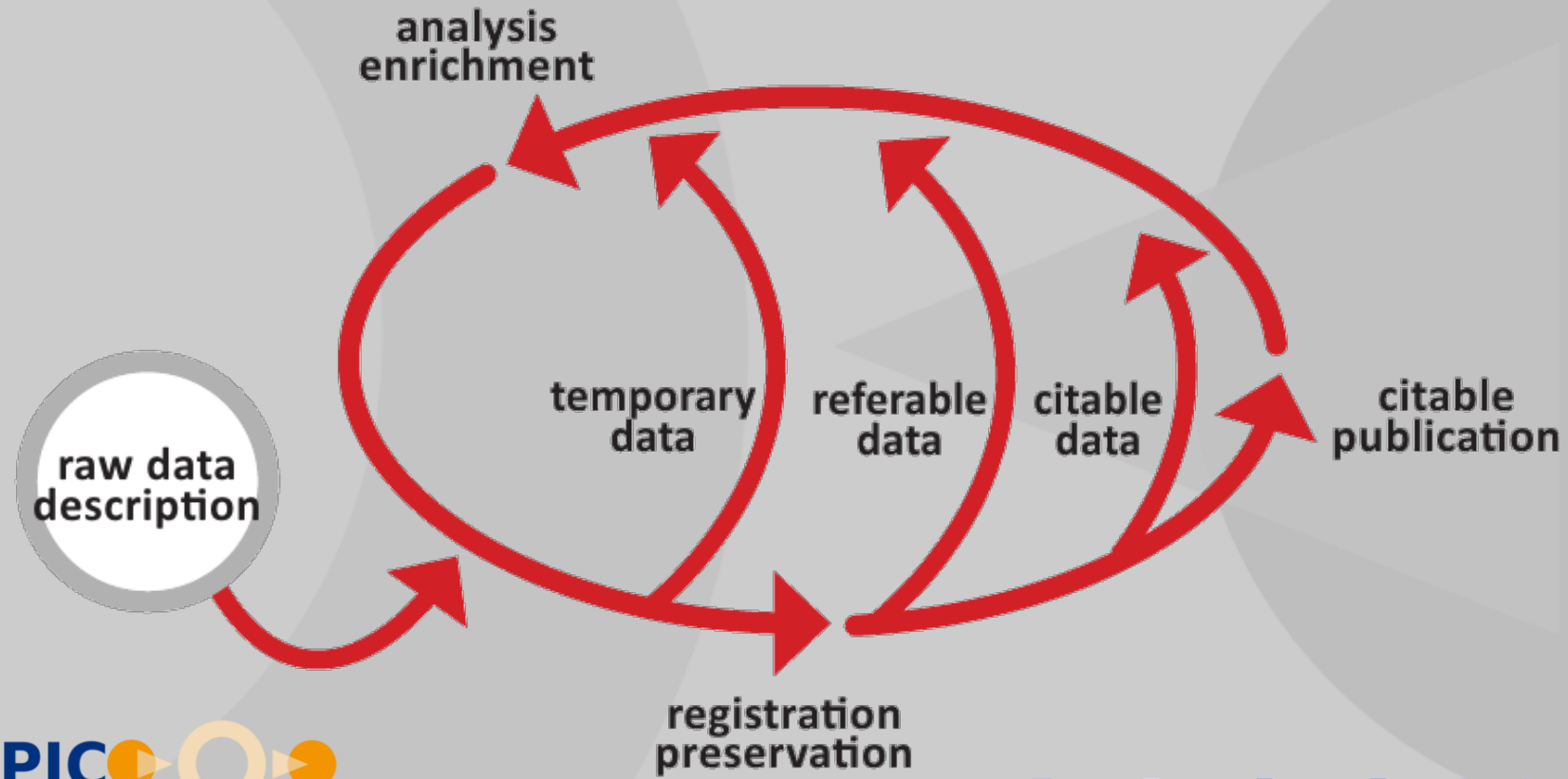
ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

ePIC
Persistent Identifiers for eResearch

The Research Data Life Cycle

data intensive research is highly collaborative

- scientists share data already in an early state
- ad hoc techniques for sharing are often prohibitive
- reliable references can accellerate the Research Life Cycle



ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

ePIC
Persistent Identifiers for eResearch

The ePIC Members

build a network of currently six strong scientific service providers signed a contract

- to ensure a reliable and persistent identifier infrastructure
- devoted to the needs of the research community at large.

Mayor focus: the referability of data

- with finer granularity and
- for sharing during the research process

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

The ePIC Members

- signed a contract to ensure
  - high availabitity for the PID services and
  - a long term perspective at organisational level
- have agreed on a Quality of Service Level,
- have implemented a redundancy scheme
- share the same service, the same API
- and the same framework of policies

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

The ePIC Members

- ePIC has minted already more then 30 million PIDs
- uses currently about 30 prefixes
- two with more than 8 million, three with more than 2 million PIDs
- ePIC takes an important part in the global Handle System
- GWDG is on behalf of ePIC a Multi Primary Adminstistrator in **DONA**
- GWDG provides on behalf of ePIC a **DONA MPA GHR** and a **Handle Proxy Server**
- ePIC is interested to run a data type registry as part of **Cordra**

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

ePIC
Persistent Identifiers for eResearch

Quality of Service

- Conditions of Operation
  - user management, privacy protection and secrecy
- incident management and monitoring
- support system with agreed responsabilties
- certification of ePIC PID services
- several policies for PID minting and update agreed
  - others are still under discussion
- quality of resolution
  - audits can be requested

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

Sharing Data in Research

- data sharing of early results requires
  - a reliable framework of trust
  - transparent and standardized policies
  - registration for referable data
    - strong coupling between data and metadata
    - stable references
  - but also a review procedure to delete data

PIDs can be the pivot to fulfil these requirements

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

Flexilibity

- ePIC PIDs can be enhanced with policies for sharing data for
  - stable references
  - transparent embargo and deletion rules
  - transparent provenance
  - direct access to data and metadata
- flexible rules for data registration granularity
- performance of resolution and minting
  - depending on needs for speed and policy

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

Types

- are additional metadata stored in the PID database
- intended to be directly accessible independent of any redirection
- typical cases are
  - checksum
  - mime type (incl. version)
  - embargo time
  - expiration date
  - add. metadata file
  - basic Dublin Core

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

Granularity

digital objects shared with other scientists for investigation have *finer granularity*

- use cases are
  - single experiments
  - simulation output and/or parameter sets
  - single files, tables, pictures, single scanned pages or video/audio sequences
  - snapshots of sensor outputs (dynamic data)
  - software and software versions
- in some cases these sets of digital objects are highly structured
  - and accessible by parameterized services
  - here also *templates* or *fragment identifiers* can be a solution

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

Templates or Fragment Identifier

rules for strings appended to the PID (see IETF RFC 6570)

- often used to address *service functions* operating on digital objects
- the template implementation in the handle system is simply a rewrite rule
- delimiter and replacement is configurable at prefix level
- *example*
  - delimiter is `@`, which is replaced by `?`
    `11858/00-ZZZZ-0000-0001-CCD1-4@aaa=bbb&ccc=ddd`
  - translates into:
    `http://wwwuser.gwdg.de/~tkalman/downloads/formtest.php?aaa=bbb&ccc=ddd`
- be **careful**: fragment identifier are much **less persistent** then the PIDs itself
- the rewrite rule can be much more complex:
  - replace semantic string elements like URLs by other strings
  - use delimiter strings instead of characters
  - ...

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

Variation in the Needs for Close Metadata

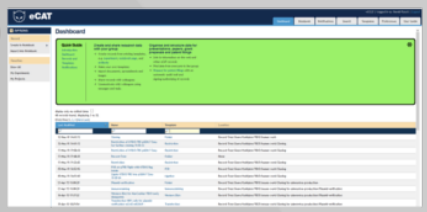
properties of digital objects and there needs for close metadata

	checksum	checksum fixed	metadata pointer	author	basic Dublin Core	mime type & version	creation date	expiration date	embargoes	versioning	templates & services	fine data granularity
edition	*		*		*					*	*	
object of edition	*			*			*		*		*	*
electronic lab notebook	*		*		*					*	*	
notebook entry		*	*	*					*			*
experiment output	*			*		*	*	*	*		*	*
image library			*		*							
image	*			*		*	*	*		*	*	*
repository entry	*			*	*	*		*		*	*	

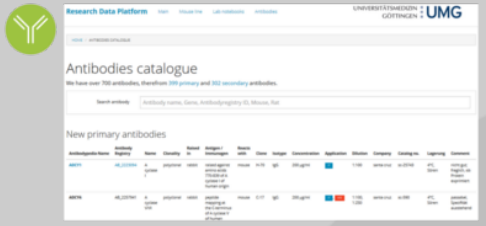
ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases

CRC1002: Modulatory Units in Heart Failure

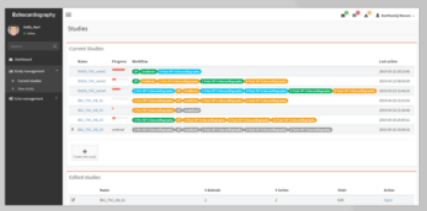
Collaborative Research Centre funded by: *Deutsche Forschungsgemeinschaft*



- template development
- based on existing standards



- Adaptation to PSI-MI/MIAPAR
- Antibody Registry/Antibodypedia



- automatic EPIC PIDs for datasets
- links to Biological databases

- PubMed standard and Indexing Number (PMID)
- PubMed Central referencing number (PMCID)
- Journal Digital Object Identifier (DOI)

- Unique name (EPIC PID)/URL

ePIC
Persistent Identifiers for eResearch

ePIC
Mission
Trust and Reliability
Research Data
PIDs for Data Intensive Research
Flexilibity
Granularity
Use Cases