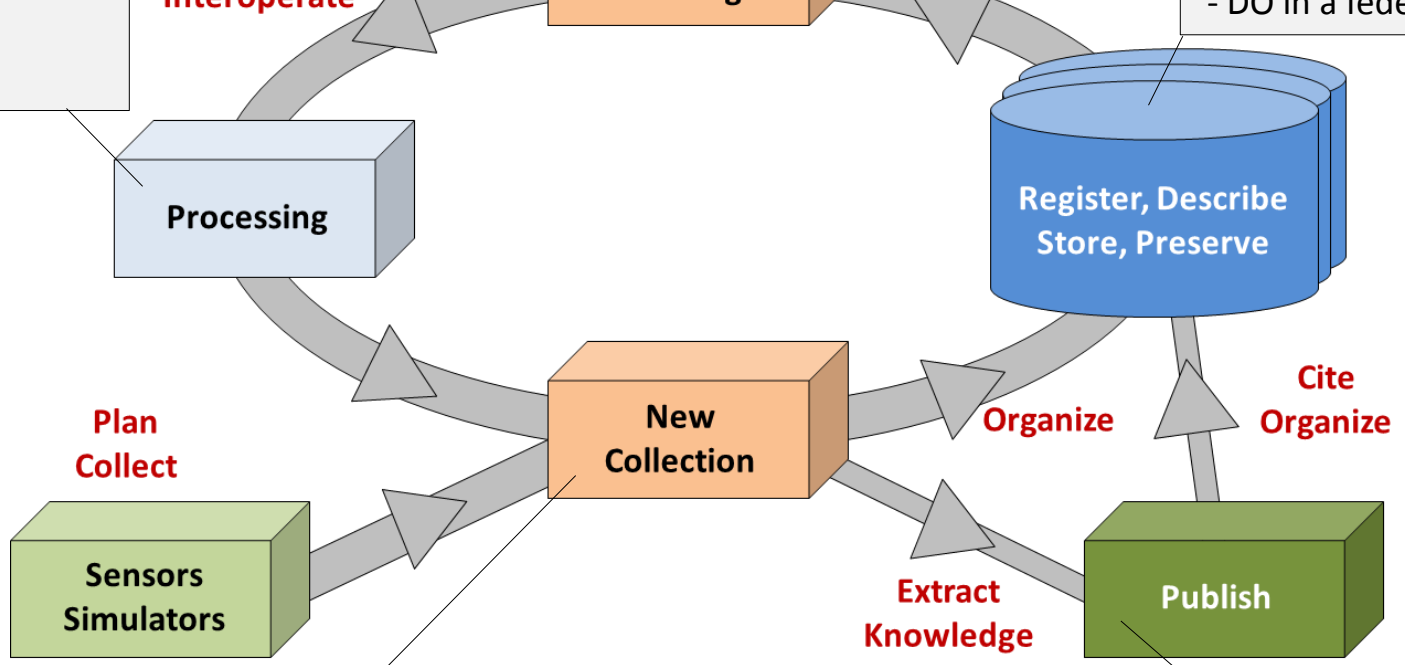
MAX PLANCK COMPUTING & DATA FACILITY

Usage of PIDs from a Lab Perspective

Peter Wittenburg (MPCDF)
former MPI/DOBES Archive, CLARIN, EUDAT

The Lab Situation: “endless” cycle



- combine data to get new insights
- often based on ad-hoc decisions
- CB should be a virtual operation

processing can mean
- management,
- curation
- transformaton
- selection
- analytics
- etc.

- DO stored in some rep in some organisation
- DO have PID and MD
- DO in a federation

- new data is organised in some form (files, db, cloud objects, etc.)
- to be ingested into a structured rep

- in case of evidences people think of publishing a paper
- then context needs to be known, i.e. publishing data

diagram from RDA Data Fabric IG

Lab-Situation: practices

- ~ 120 interviews in different departments

  RDA Europe: Data Practices Analysis
  http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f
- Biologist: 75 % of my time I am a data manager
- Michael Brodie (MIT): 80% of DS time is data management
- practices are not efficient and too costly
  - ad-hoc scripts and manual operations dominate
  - documentation is not sufficient
- broad agreement about need of changes
  - lack of reproducibility is an issue for DS
  - but how – don‘t make it a burocratic act – need tool support
    - automation of workflows not easy
      - too many exceptions, too unpredictive
      - lack the experts to build parametrised workflows
    - manual PID registration hardly feasible
    - PID integration into ad-hoc scripts not common

Old ways don’t work any longer

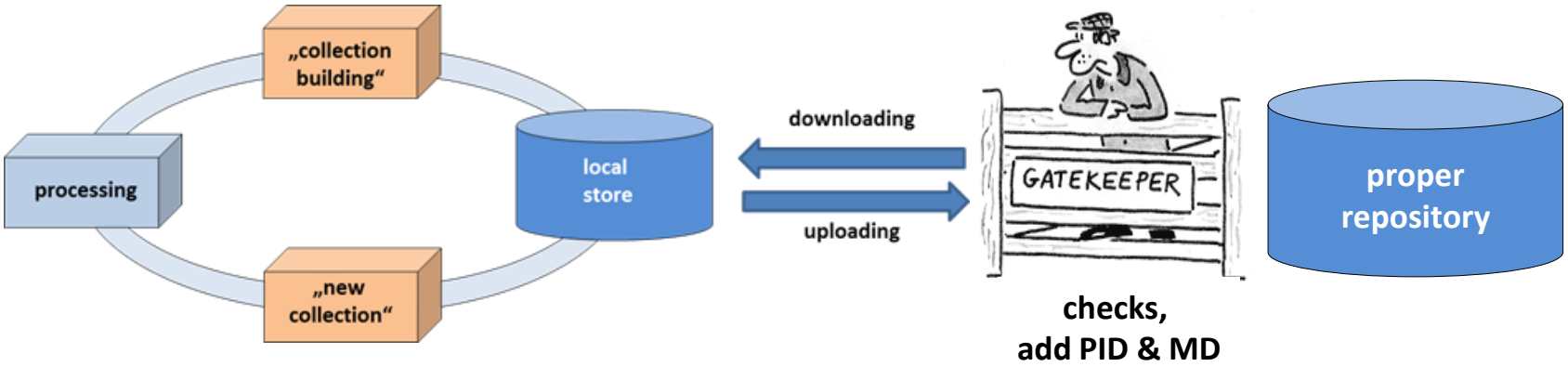
- as long as experts work as individuals or in small groups file systems, local databases, etc. may still work
  - in case of data moves everyone can be informed
- volume and in particular complexity change the game even in groups
  - after few weeks people don‘t know anymore which versions were combined for a specific purpose, etc.
- of course sharing with larger groups changes the game as well
  - impossibility to follow DM actions

- what is a feasible solution?
  - have an easy-to-use structured repository to upload and register your data and thus create referencable collections
  - registration implies fixing, PID creation, minimal MD creation
  - registration of workflow generated data is simple
  - registration of ad-hoc generated data can be simple

  - is that already data publishing? It is an explicit step and costs effort!
  - who runs the repository? It should fulfill some criteria!

Broken Cycle Solutions

- scientists rely on “safe and trustworthy” repositories
- but their processing is separated from the repositories



- advantages:

  lots of operations in flexible style, continuation of what people have worked out over many years, no extra burdon by PIDs, MD etc.
- disadvantages:

  no proper tracking (provenance), extra step with error potential, late registration, etc.

Example 1: MPI/DOBES Repository

- about 80 TB from 200 TB
- 4 external dynamic copies
- all accessible and re-usable
- using a Handle System
- flexible MD schema
- simple web-interface
- some support
- experimentalists not yet interested

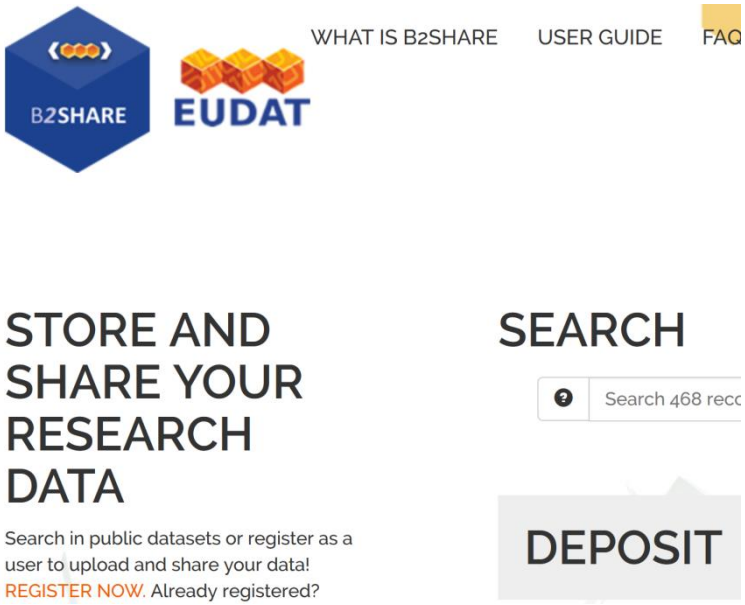
- about 200 MPI and DOBES researchers create data about languages of the world
- what is the gain
  - managed reference collection with stable references and lt preservation
  - easy combination to new collections to find patterns in languages
  - media fragment referencing in papers
  - visibility
  - point of attraction for many additional researchers (economy of scale)

Example 2: NOMAD Repository

- huge amount of simulation results
- no one has an overview
- virtual and physical aggregation of result data sets
- all receive DOI and get some MD
- MD generation not trivial

- many material scientists world-wide create simulation data
- almost endless compound space
- what is the gain?
  - overview about what has been computed already worldwide
  - searching for specific patterns by combining results is possible
  - categorisation of materials by doing analysis
  - management, curation and long term preservation

Example 3: EUDAT B2Share

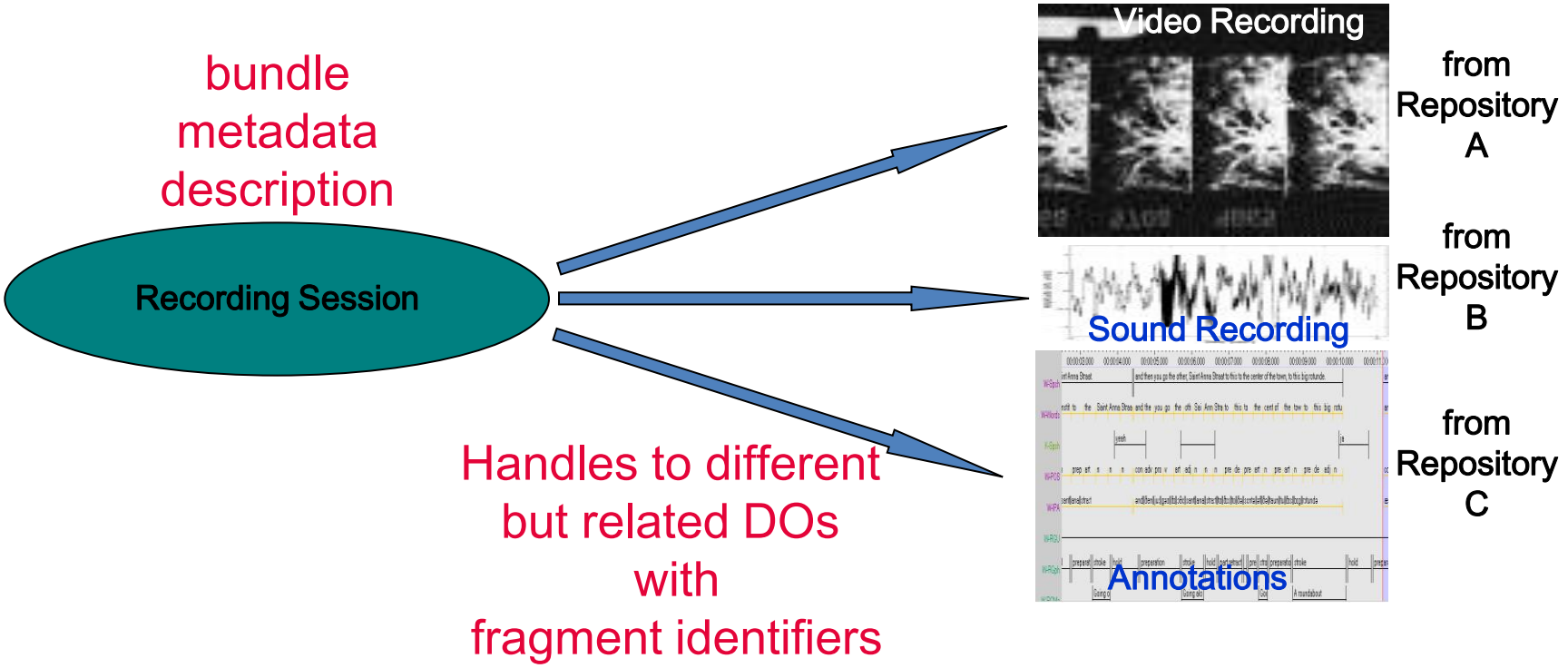


- all kinds of uploaded data from various communities
- all receive an EPIC PID and get some MD
- thus all stored objects are citable
- usage is up to users by download
- OpenAIRE has similar offer (ZENODO)
- and several others

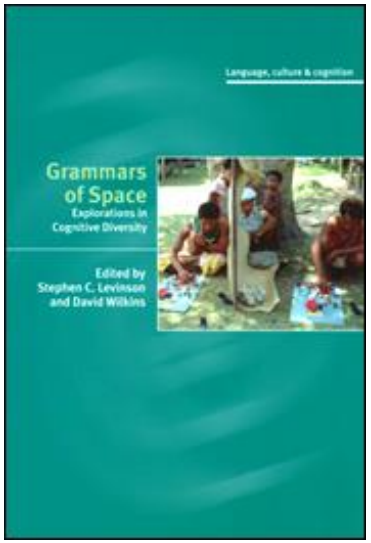
- there is so much data out there from people not directly associated with a project or so that need to be stored
- big question: why should anyone trust some folks they do not know
- what is the gain:
  - well-managed repository, searchability and accessibility
  - big capacity store
  - referencibility

DOBES/MPI Usage 1

- assume that we have a recording of an extinct language and some annotations that tell us what someone said about medicine etc
- researchers create relations that need to be preserved

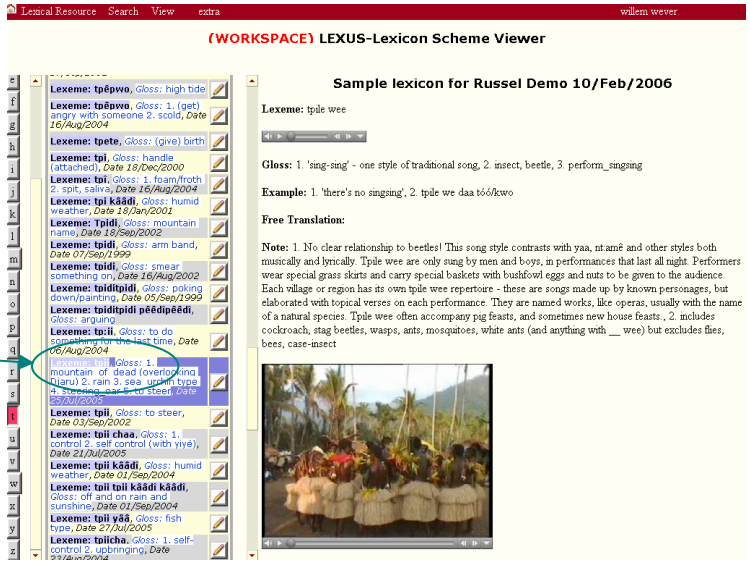


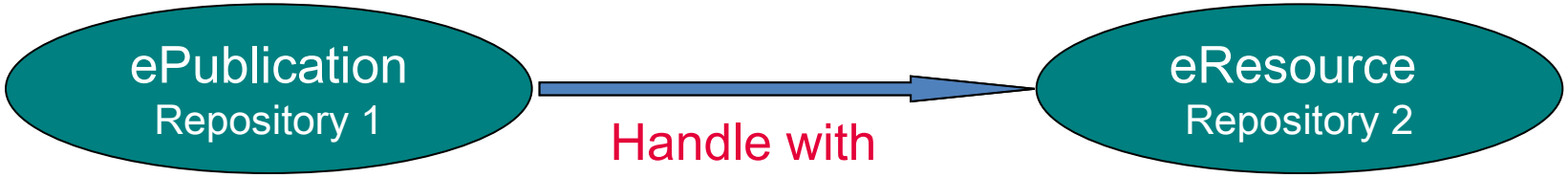
DOBES/MPI Usage 2



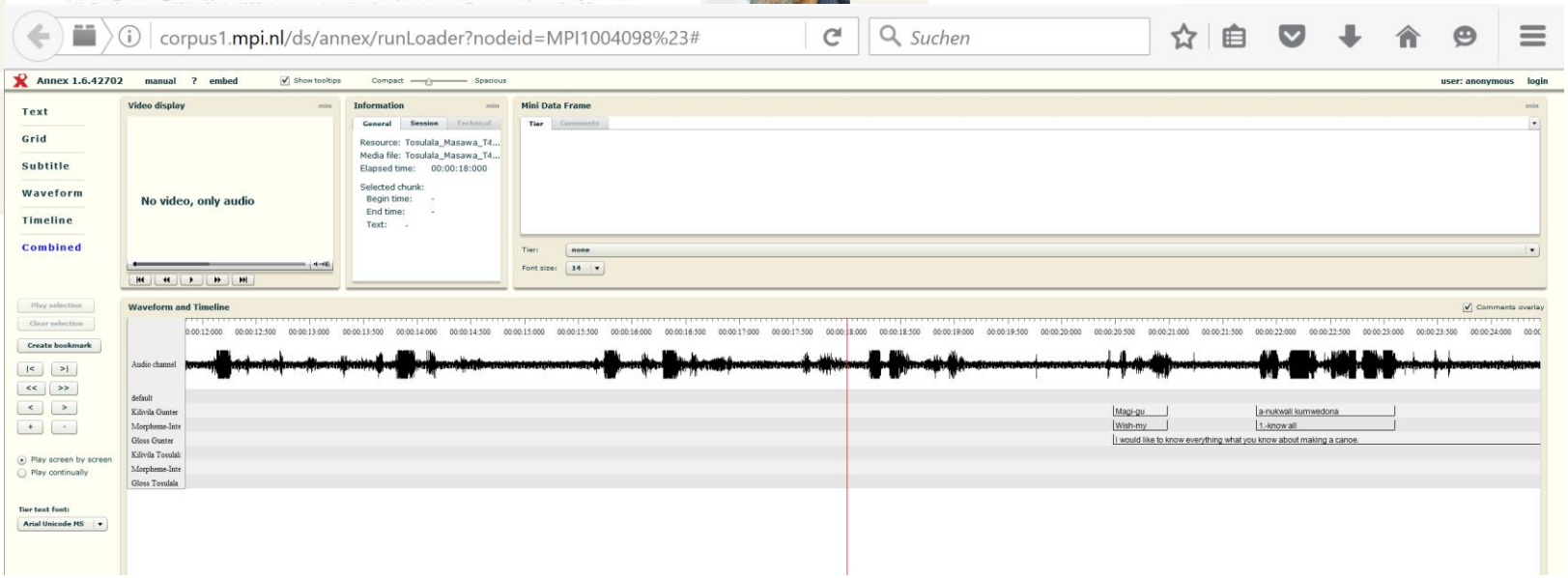
Biological and cultural processes have evolved together, in a symbiotic spiral; they are now indissolubly linked, with human survival unlikely without such culturally produced aids as clothing, cooked food, and tools. The twelve original essays collected in this volume take an evolutionary perspective on human culture, examining the emergence of culture in evolution and the underlying role of brain and cognition. The essay authors, all internationally prominent researchers in their fields, draw on the cognitive sciences -- including linguistics, developmental psychology, and cognition -- to develop conceptual and methodological tools for understanding the interaction of culture and genome. They go beyond the "how" -- the questions of behavioral mechanisms -- to address the "why" -- the evolutionary origin of our psychological functioning. What was the "X-factor," the magic ingredient of culture -- the element that took humans out of the general run of mammals and other highly social organisms?

Several essays identify specific behavioral and functional factors that could account for human culture, including the capacity for "mind reading" that underlies social and cultural learning and the nature of morality and inhibitions, while others emphasize multiple partially independent factors -- planning, technology, learning, and language. The X-factor, these essays suggest, is a set of cognitive adaptations for culture.





DOBES/MPI Usage 3

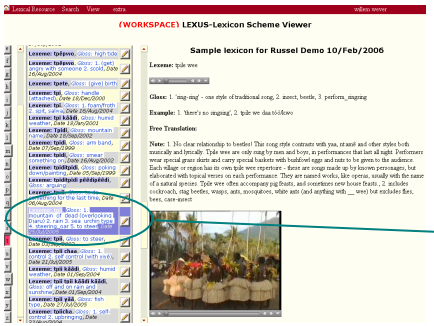


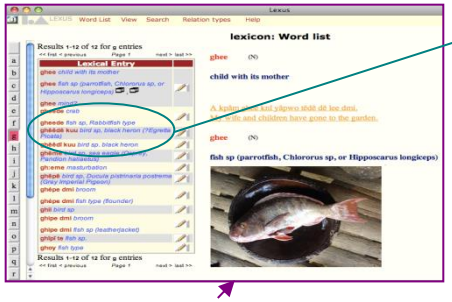
ePublication
Repository 1

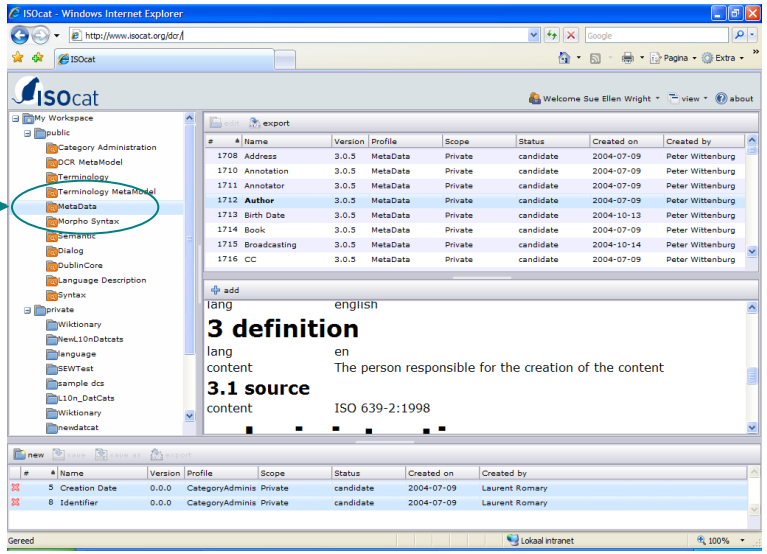
→

eResource
Repository 2

DOBES/MPI Usage 4







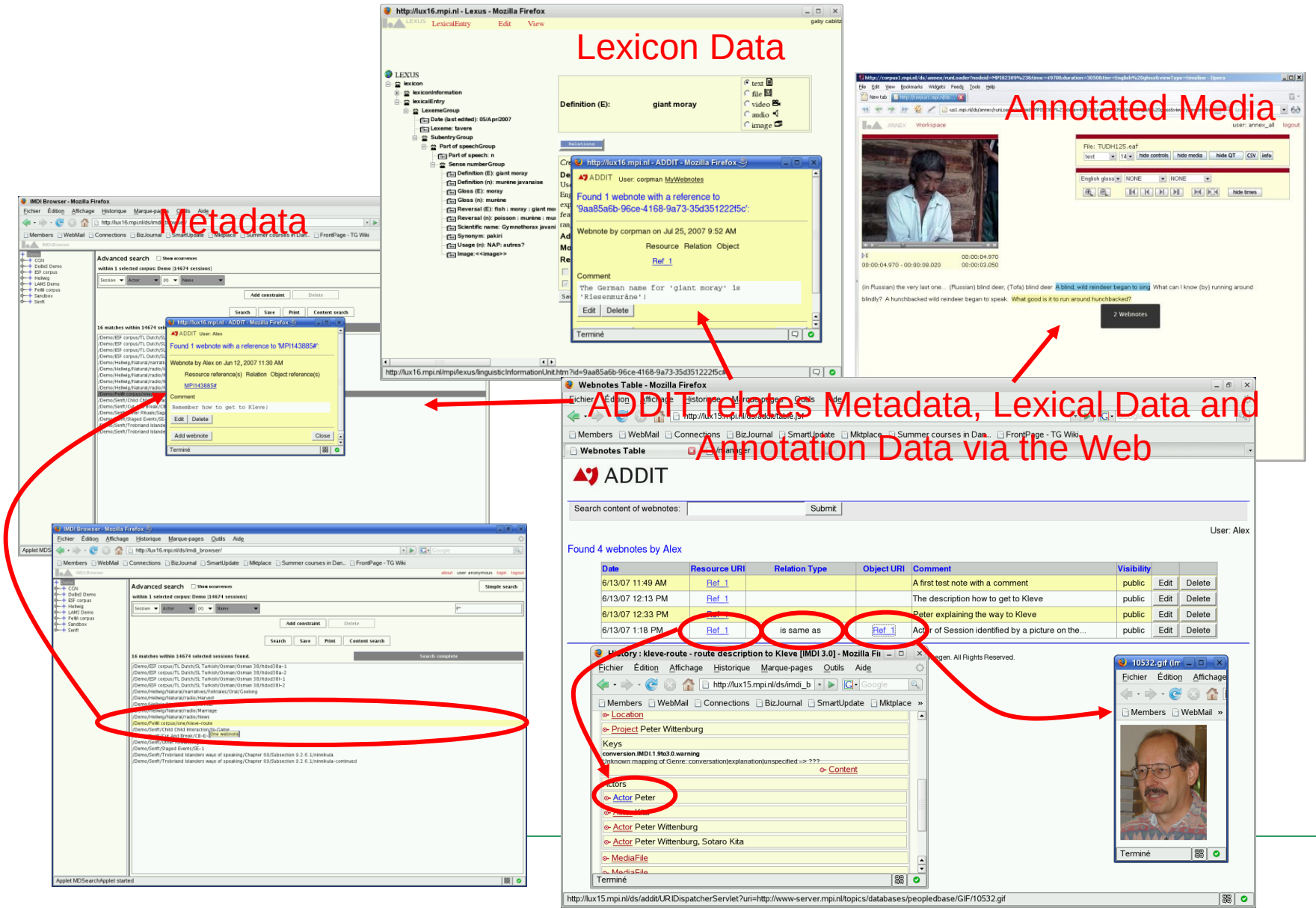
eResource1
Repository 1

eResource2
Repository 2

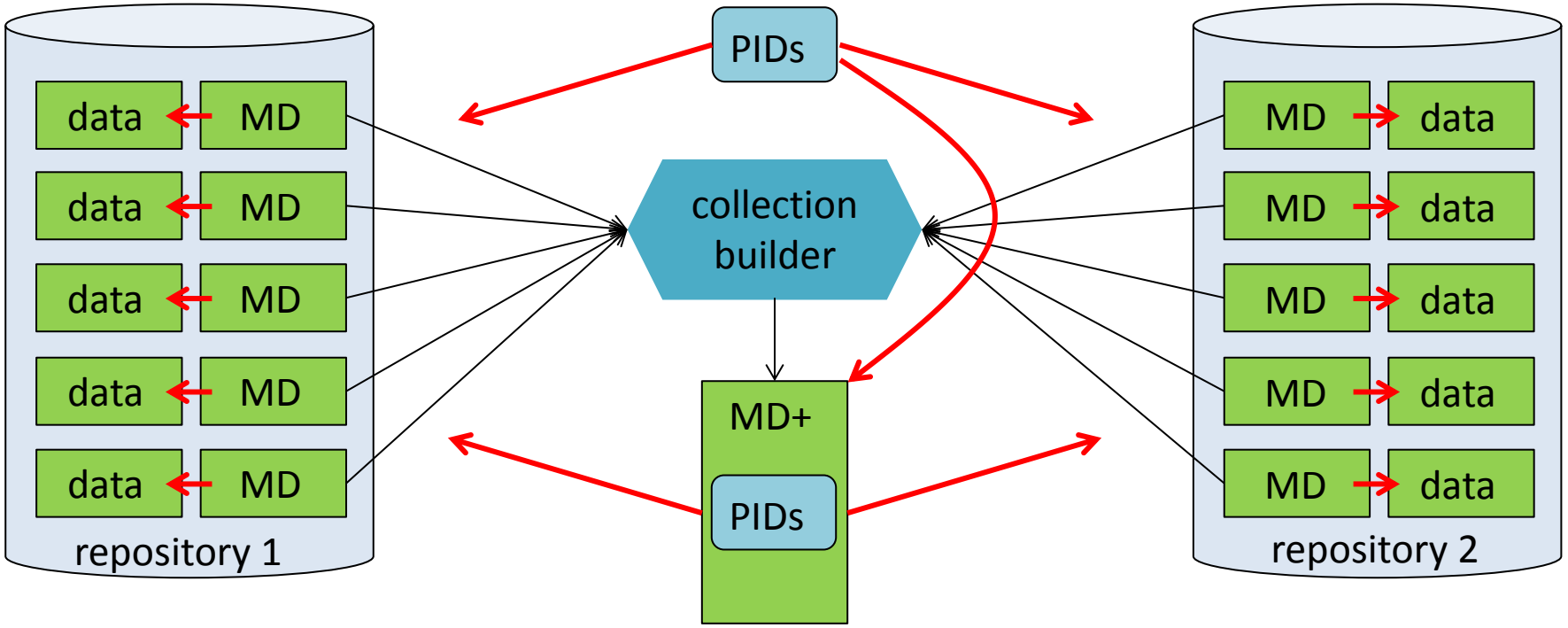
Ontology
open registry

here we used URLs
is that ok?

DOBES/MPI Usage 5

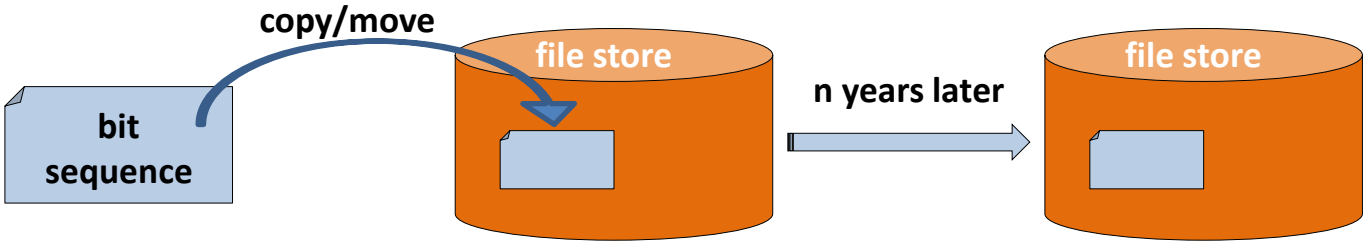


CLARIN Usage 6



Collection building across repositories is essential for running analysis. PhD students f.e. create their collection which exists of a MD object having a PID and storing many PIDs to refer to its components. Such a collection is an aggregation, but has an identity, can be cited, etc.

Just referencing?



- still the same?
- management error?
- hardware error?
- cheating?

- „bit sequence“ could
  - just have an identifier (number/string/etc.)
  - have an identifier to be resolved into location (URL)
  - have an identifier with other relevant information (location, type, checksum, MD, rights, etc.)

an identity card has

- a number
- some fingerprint info to check
- some type information
- a time period specification

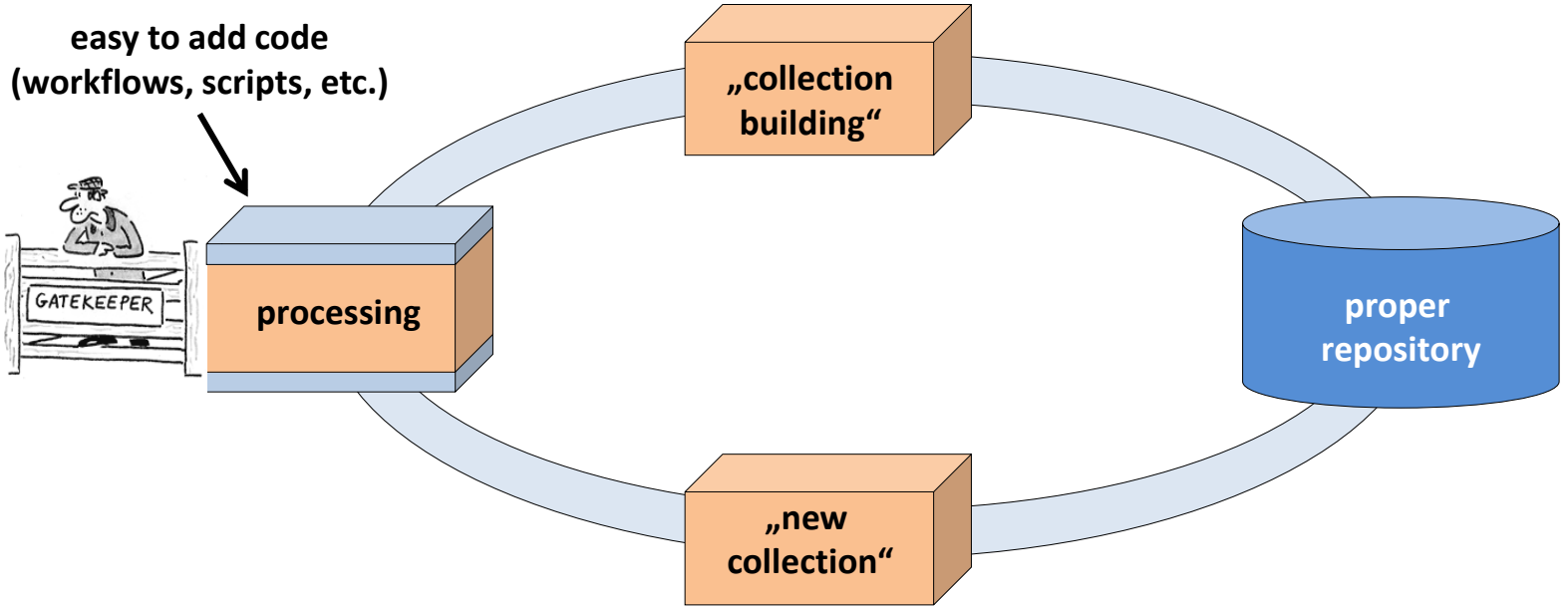
allows checking correctness

DOI vs. Handles

- why did we not use DOIs?
  - in 2003/4 we would have to pay 30 k€ per year just for our archive
  - Handles: <Prefix>"/"<HandleLocalName>
  - DOI: “10/"<DOILocalName>
    - difference is not technical – you need to establish trust
  - we asked MPG to set up a service for all 80 institutes
    - GWDG offers a service now beyond MPG
    - you need an organisation you trust
  - we asked CNRI to make Handle System independent
    - CNRI initiated the DONA foundation in Geneva
  - no principle difference as long as you have a trustworthy organisation behind technique (are IDF and MPG ok?)

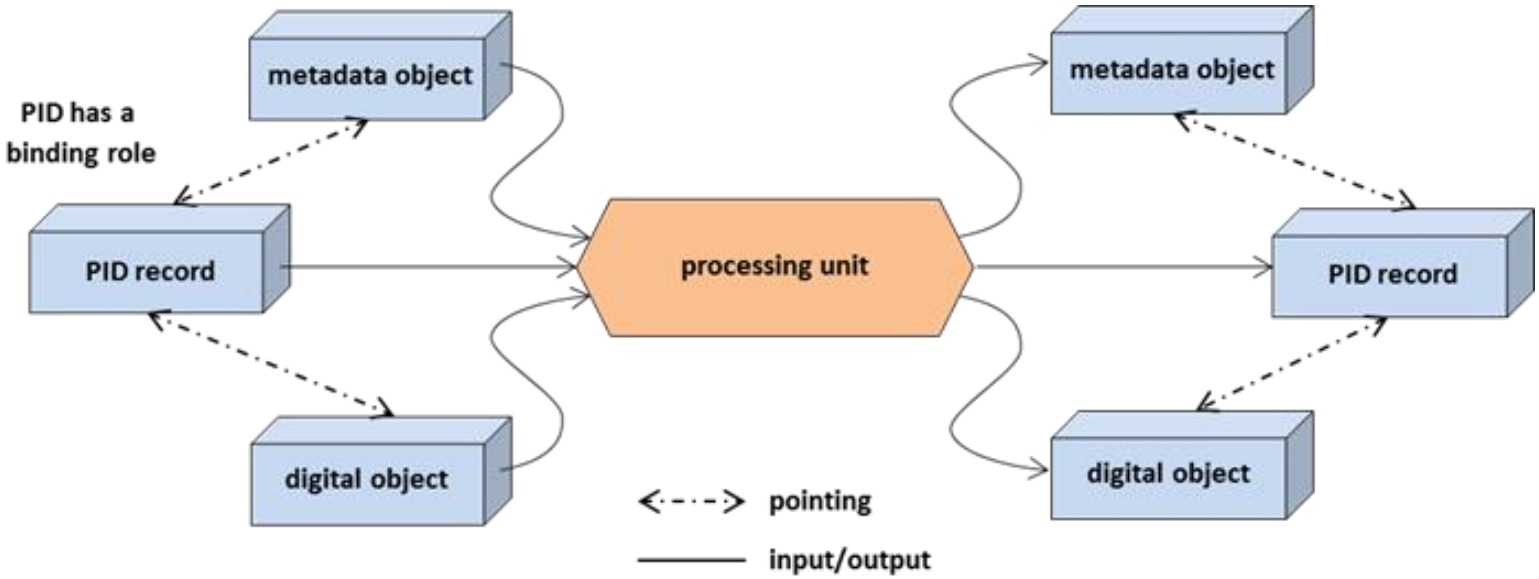
Closed Cycle Solutions

- not many implementations with „integrated gate keeper“
- often in simulations with data generated by „own“ software



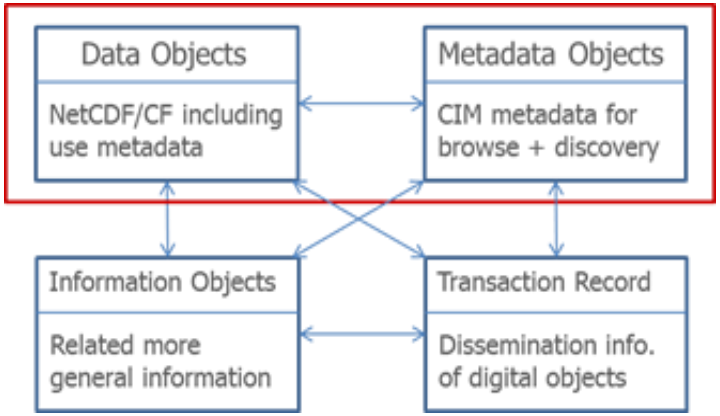
- one would not leave the domain of registered DOs

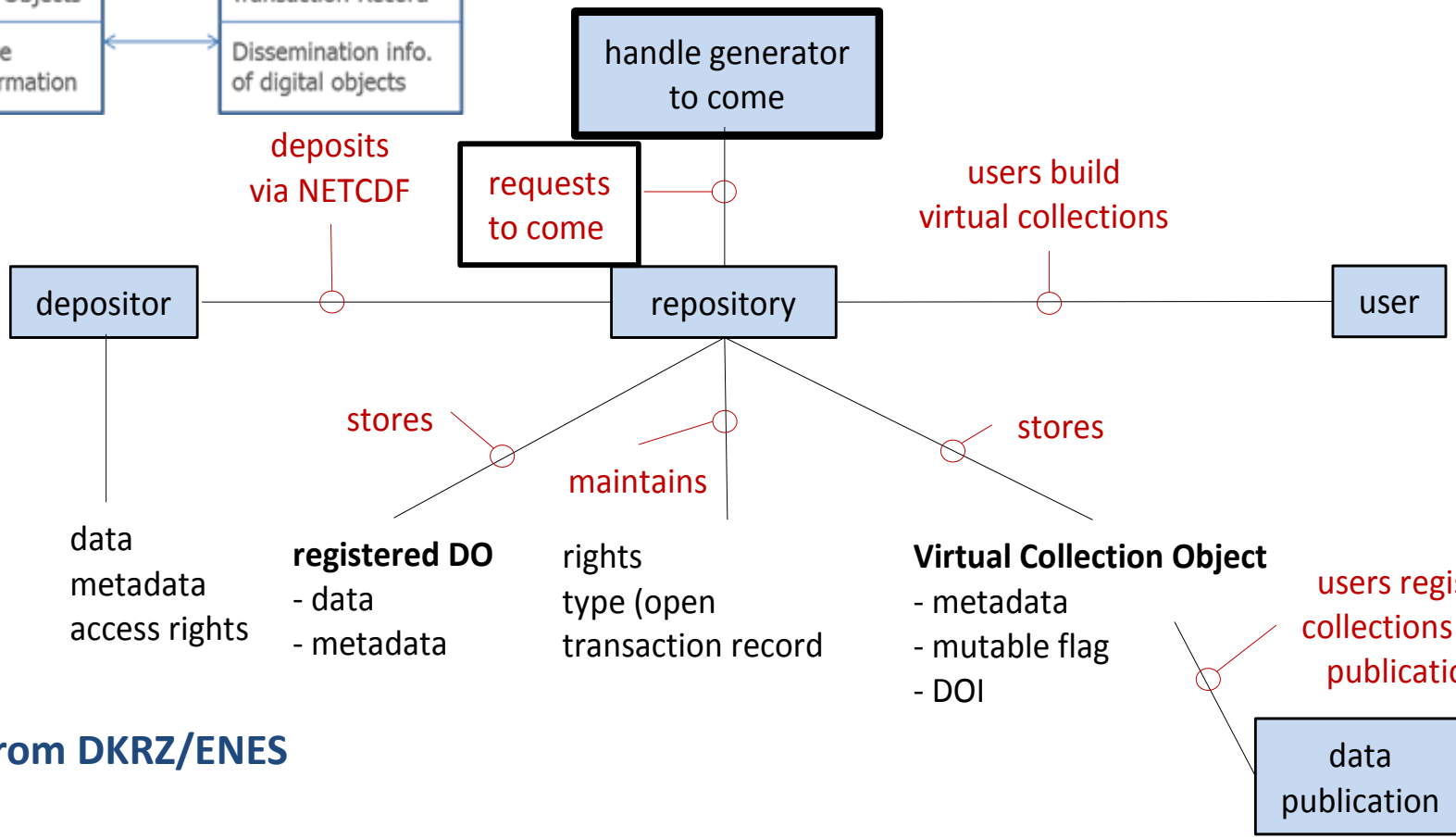
Missing Components



- shown is a suggestion for automatic processing
  - take old PID record, open MD and bit sequence
  - do some processing
  - at the end give new bit sequence an updated MD record and a new PID
  - PID can be registered automatically via API
- do this from the very beginning to get reproducibility

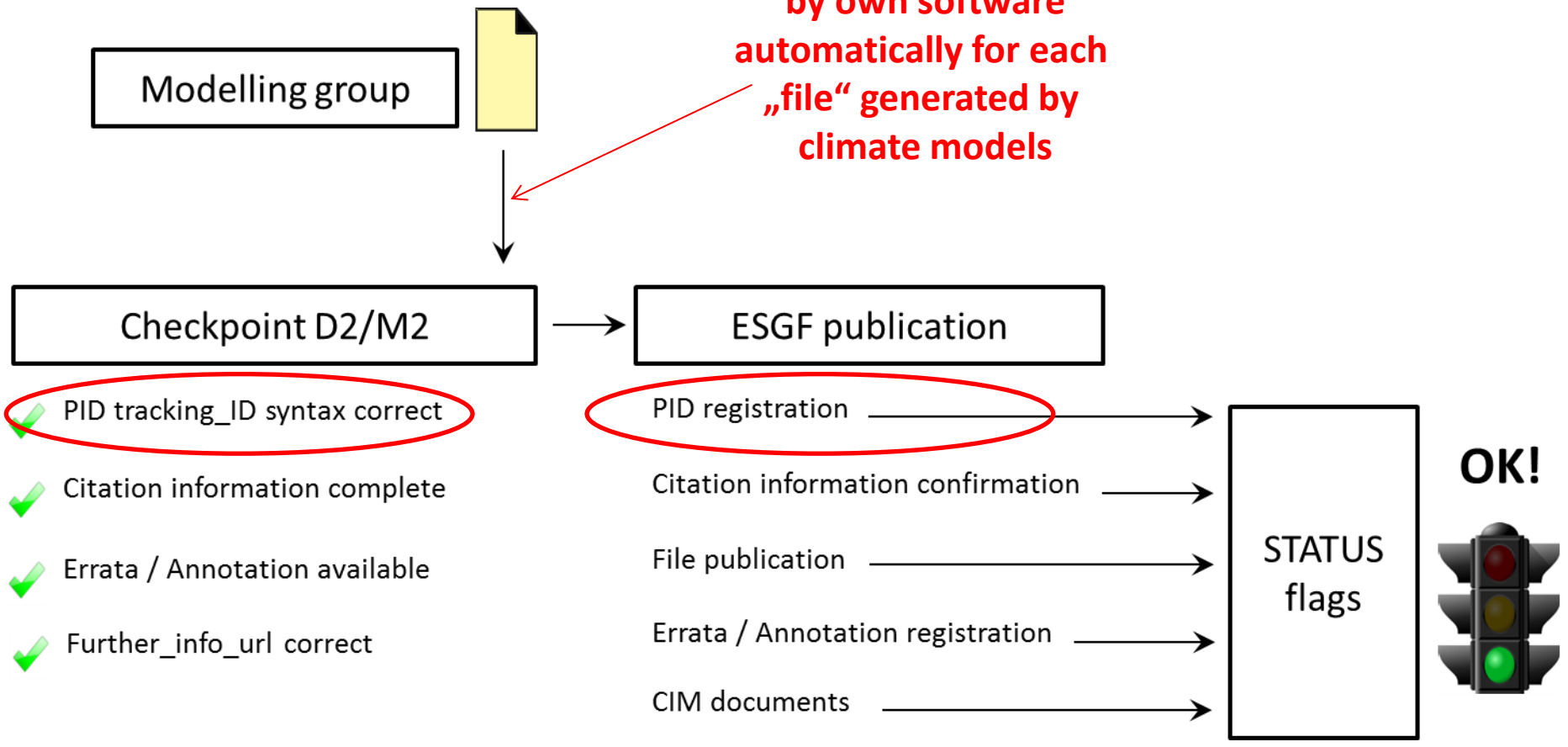
Climate Modelling Workflows





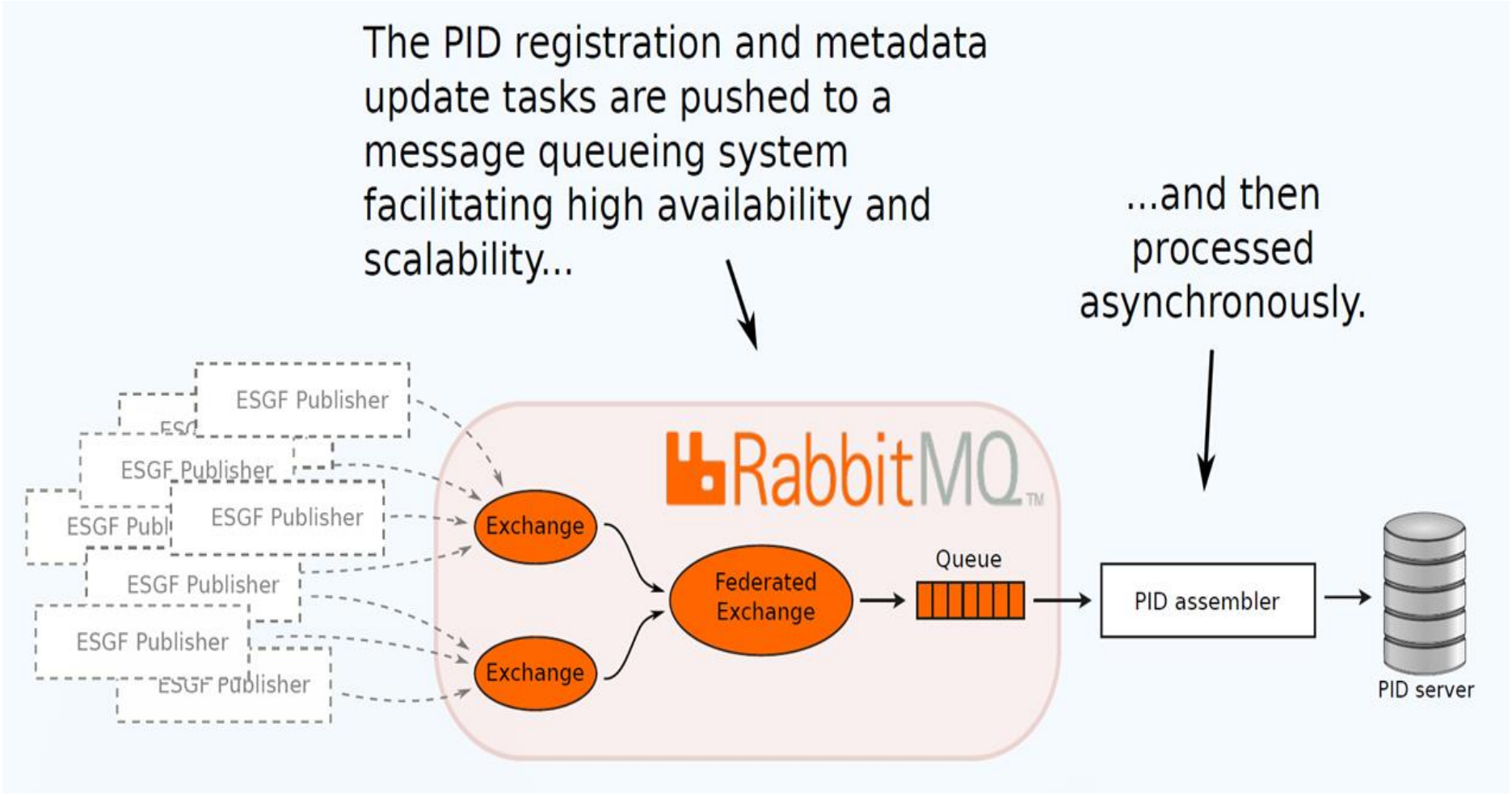
taken from DKRZ/ENES

CMIP6 PID Generation and Registration



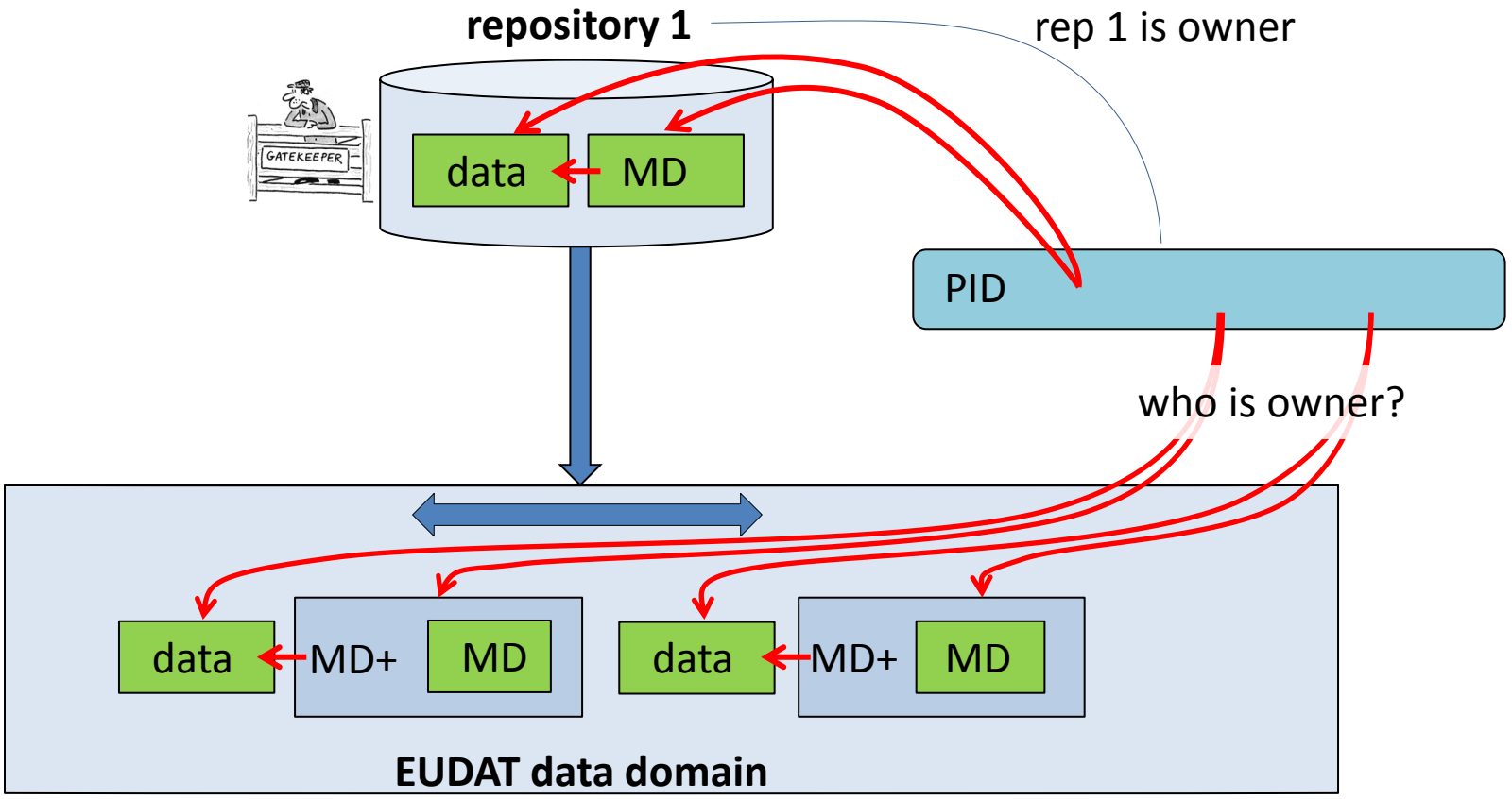
taken from DKRZ/ENES

Massive PID Registration



taken from DKRZ/ENES

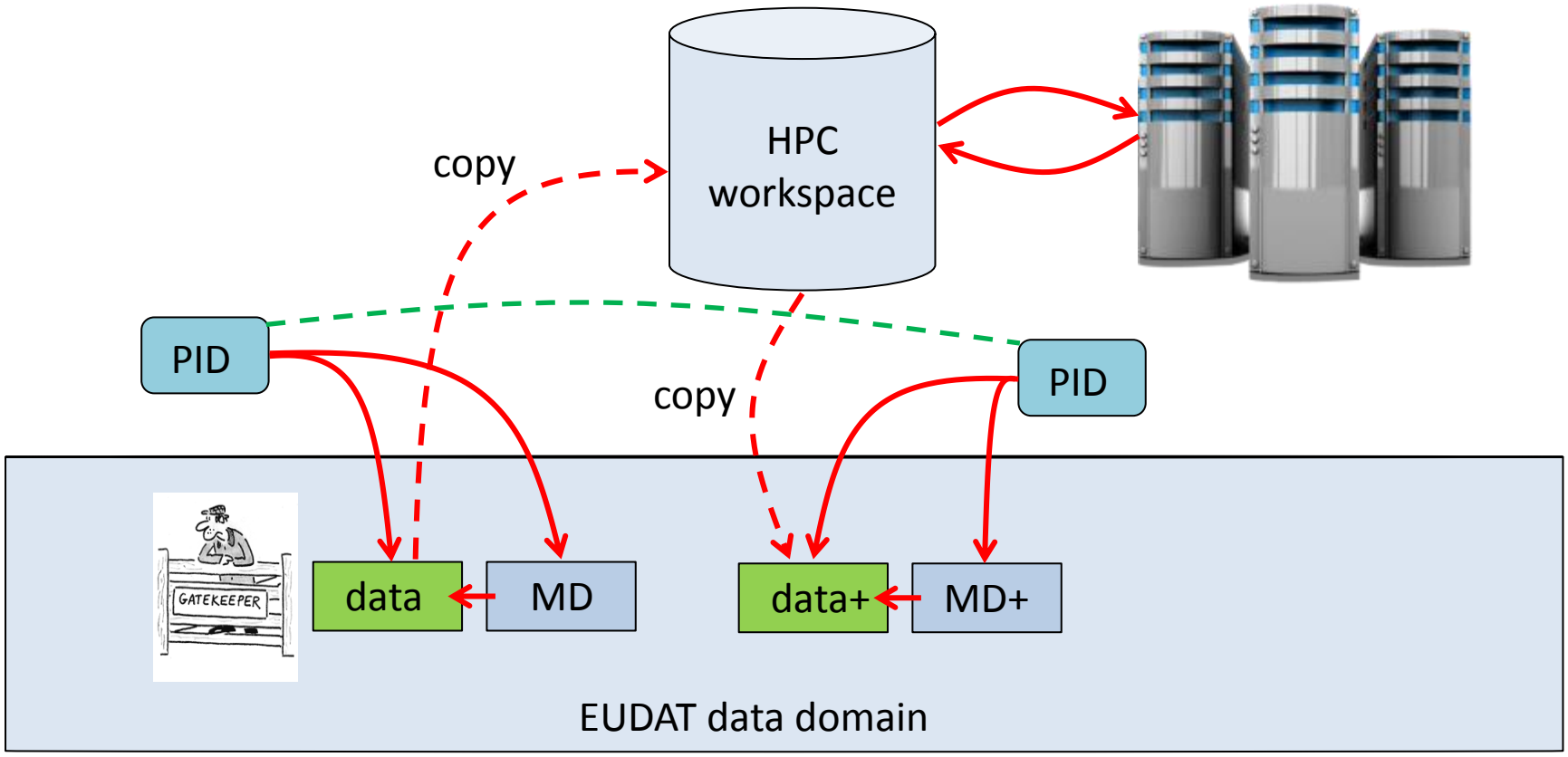
Federations are tricky (EUDAT)



replicating data from different types of data organizations from different communities = creating a coherent domain of data is not trivial

who is owner of record and who is allowed to write?

Federations are tricky (EUDAT)



also here we have a broken cycle – how to solve that?
how to relate these two DOs?

Summary

- we are creating millions of DOs in the labs
  - their relevance is not obvious at the beginning
  - but we need to reference them for various purposes
    although no quality check, no publishing act, etc.
  - APIs to automate registration are needed
  - need service providers we can trust
  - should we turn „some“ Handles into DOIs? – why and how?
- Handles/DOIs can help in achieving reproducibility
  - some started creating workflows incl. Handles/DOI
  - but common practices far from being ideal
    - it‘s the lack of experts and the many exceptions
    - will it change?
- some scenarios are difficult
  - for example in federations – ownership vs. access flexibility

Thanks for your attention.