WP3 Registry of tools & services
WP5 Proofs of concept
dArceo and PLATON archival services
EUDAT services

Maciej Brzeźniak, Norbert Meyer, PSNC

Task 3.3
registry of services and tools

Maciej Brzeźniak, Norbert Meyer, PSNC

D3.3 Registry of services and tools (1)

- **Task 3.3. Registry of services and tools: definintion in DoW:**
  - This task will **define the structure of a registry** of services and tools used by the stakeholder community for digital preservation for DCH.
  - It will put in place a pilot registry… described in D3.3.
  - Deliverable D3.3 at M12 (10PMs): Registry of services:
    - … presents a registry of the services available to support preservation activities, with particular regard to the services that can better fit the requirements of the DCH sector.
- **D3.1's „Mind the gap" section states:**
  - There are a **few hundred software tools** on offer to support automation of preservation tasks…
  - **… yet their support status,… interoperability…, level of documentation, … quality and reliability are poorly documented.**
  - … There continues to be **inadequate support for decision making**, selecting, testing and benchmarking tools for preservation.
  - **There is a need for a registry of preservation services with clearly applied metrics which makes tools easy to compare.**

D3.3 Registry of services and tools (2)

DCH RP Digital Cultural Heritage Roadmap for Preservation

❑ Task discussion (1) – questions:

- **Registry:**
  - **Purpose / target audience?**
    - Help stakeholders to understand SoA and make decisions
    - Help ourselves ;] to identify gaps vs the roadmap
  - **Scope**:
    - 100s of tools?
    - Is it doable? Does it make sense?
  - **Inputs/Contributors**:
    - WP3 partners vs all project partners?
  - **Lifecycle**:
    - are we going to update the registry – how/when?
  - **Practicalities:**
    - Collaboration tools to be used: wiki is a 1st guess?

D3.3 Registry of services and tools (3)

DCH RP — Digital Cultural Heritage Roadmap for Preservation

- **Task discussion (2) – some proposals:**

- **Purpose:**
  - **Help stakeholders** to: understand the SoA, in order to:
    - Make educated decisions in complex & changing environment…
    - …based e.g. on indications what to use for particular purposes
  - **Help ourselves ;]** to: understand the SoA, in order to:
    - Identify gaps: roadmap vs existing tools/services
    - Investigate if/how cloud/grid can be used for DP of CH
  - **Registry vs roadmap:**
    - Roadmap says where we are going
    - Registry lists the current solutions and benchmarks them?

- **Scope - we should:**
  - **Focus a bit =>** on **best tools?** => selection criteria?
    - open source vs commercial
    - ‚cloud/grid-friendly' vs ‚silo-ed'
    - novel vs mature/well-established
    - compliance to standards? (what standards?)
  - **Consider also quality** not only features **=> *metrics?***
    - *See above*

D3.3 Registry of services and tools (4)

- **Task discussion (3) – some proposals:**

- **Sources of input:**
  - **Initial inputs:**
    - **DC-NET's report on SoA:**
      - » Digital Preservation Services State of the Art Analysis
    - **DCH-RP D3.1 1st version:**
      - » Study on a Roadmap for Preservation
    - *What else?*
  - **Contributors/Further inputs:**
    - All partners asked to add tools/services names/descriptions
    - …using a collaboration platform -> wiki?
    - *Will it work? (10PMs assigned in total)*

- **Lifecycle:**
  - **Made once for every version of the roadmap**
  - Registry should be done with the roadmap in mind…
  - Constant, on-going updates – realistic? *(I would say no)*

WP5
proofs of concept

Maciej Brzeźniak, Norbert Meyer, PSNC

WP5's proof of concept overall plan

- **Proof of concept organisation in Poland:**
  - **1st phase: go with the existing tools:**
    - CMS + DP + CD + LTS ☺
      - dArceo/dLibra/dLab – for content mgmt, delivery and preservation
      - PLATON Popular Archival Services – for long-term storage
    - Examine them vs roadmap – see where we are
    - Feedback to be sent to:
      - WP3's roadmap
      - Tools/services development plans/roadmaps
      - Infrastructure owners/policy makers
  - **2nd phase: see how projected changes impact us:**
    - Evaluate other tools or combination of local/general or established vs novel tools?
      - *e.g. eCSG with PLATON back-end*
    - Evaluate different models for offering existing tools/services : comparison study, user trial
      - *e.g. can we put dArceo/dLibra/dLab into cloud/grid?*

WP5's proof of concept DP tools & services in Poland (1)

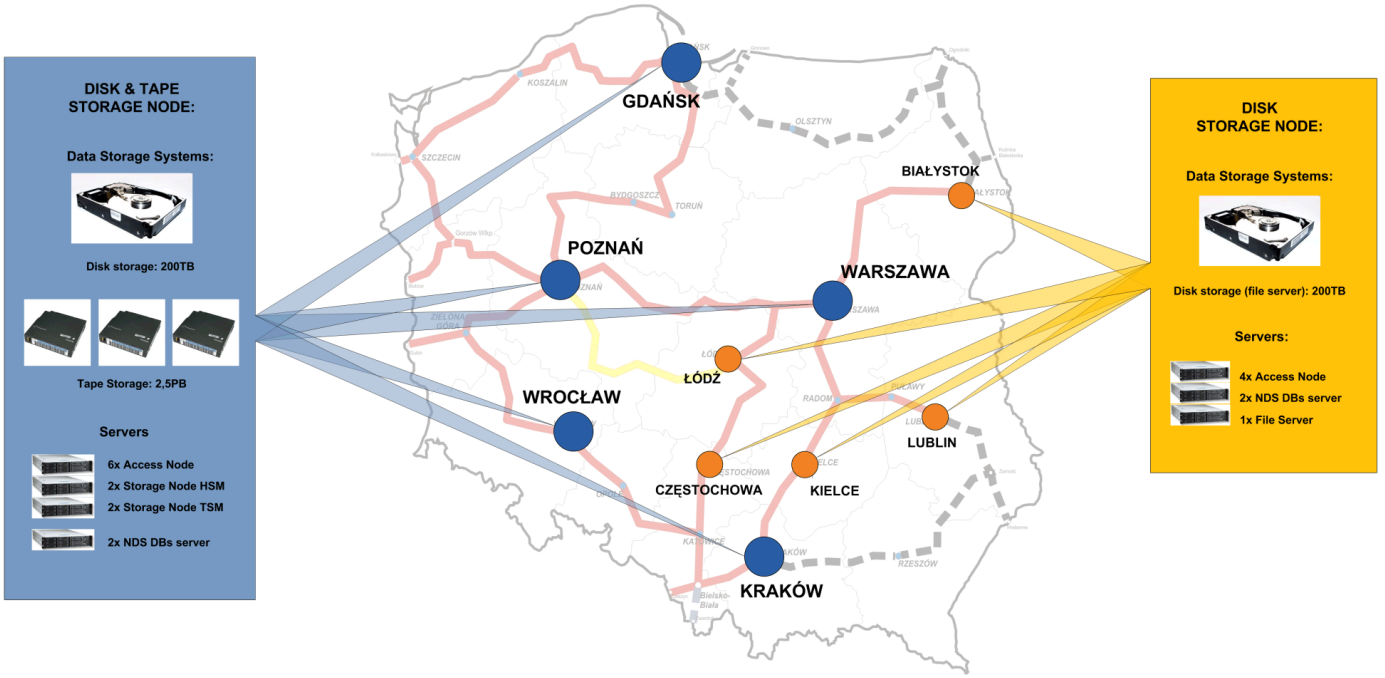
- **dArceo/dLibra/dLab – a complete toolkit to manage OAIS compliant workflow:**
  - dArceo: managing archival versions of data
    - specificaly devoted to long-term preservation, including: format migration, integrity checks etc.
    - De-facto standard in the country - based on long-running collaboration with 100+ DLs and CH institutions in Poland
    - Follows OAIS model, standard-based:
      - METS, PREMIS, TextMD, DocumentMD, VMD, MIX, AES57;
      - Source data in formats: PDF/A, TIFF, JPEG200, MPEG-4
    - By-design:
      - Modular
      - cloud- and grid-friendly

WP5's proof of concept DP tools & services in Poland (2)

- **PLATON Archiving Services – production service for long-term data storage**
  - Result of National Data Storage project (2yr R&D)
  - Deployed in distributed storage infrastructure of PIONIER network (Polish NREN):
    - 12+PBs tapes, 2+PBs disks
    - 5 HSMs with filesystem frontend
    - 5 file servers
    - Dozens of servers, 100s VMs

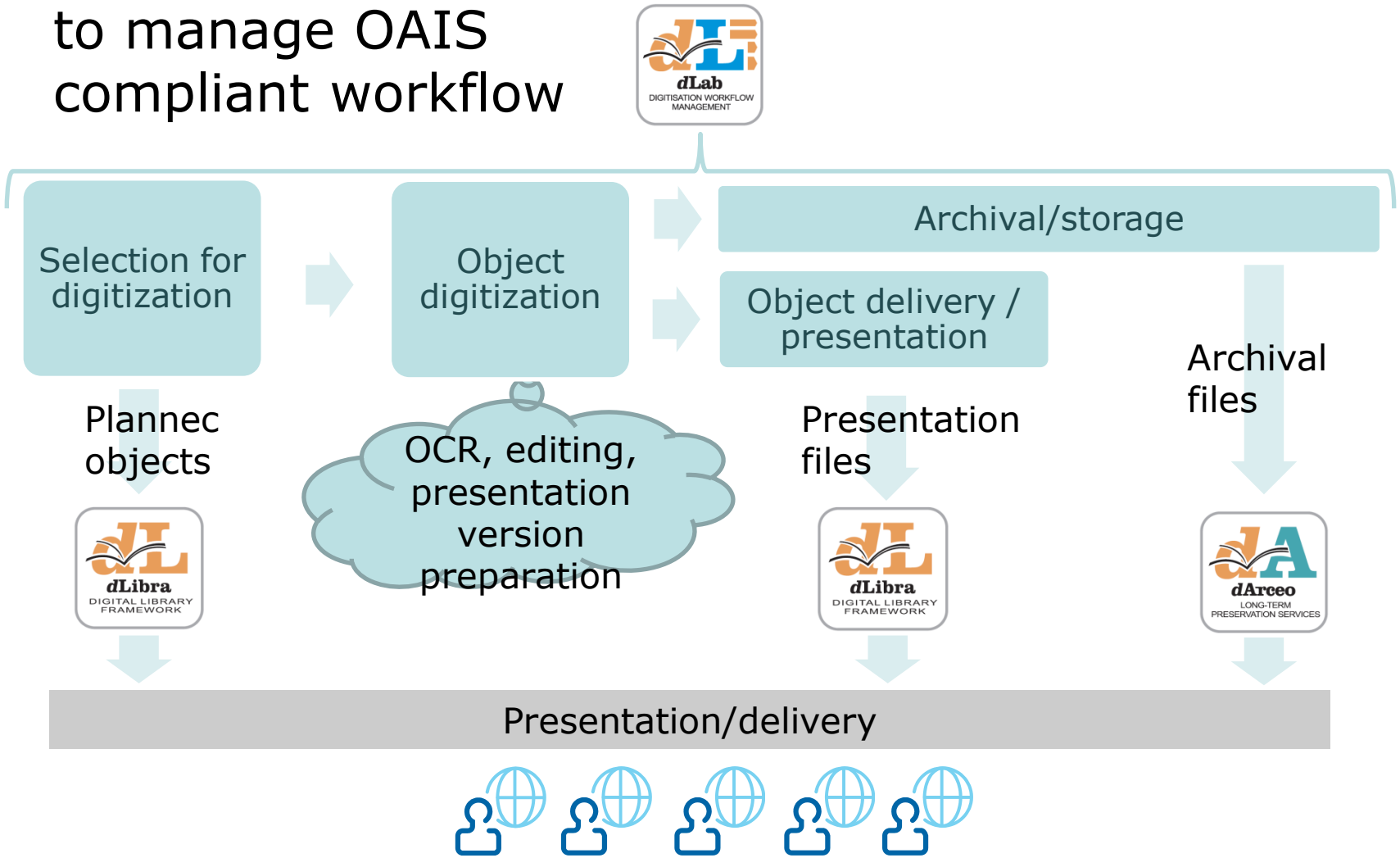
WP5's proof of concept DP tools & services in Poland (2)

- **PLATON Archiving Services – production service for long-term data storage**
  - Result of National Data Storage project (2yr R&D)
  - Deployed in distributed storage infrastructure of PIONIER network (Polish NREN):



WP5's proof of concept dArceo/dLibra/dLab

- a complete toolkit to manage OAIS compliant workflow



WP5's proof of concept dArceo elaborated (1)

- **dArceo – devoted to long-term preservation**
  - Services:
    - Source Data Manager (SDM)
    - Source Data Monitor (SDMo)
    - OAI-PMH Repository
    - Data Manipulation Services (DMS)
    - Data Migration and Conversion Manager (DMCM)
    - Rights Manager (RM)
    - Services Register (SR) / Notification Manager (NM)

- **Modular architecture:**
  - Can scale beyond single system
  - Particular services can be run in the cloud / grid

WP5's proof of concept dArceo elaborated (2)

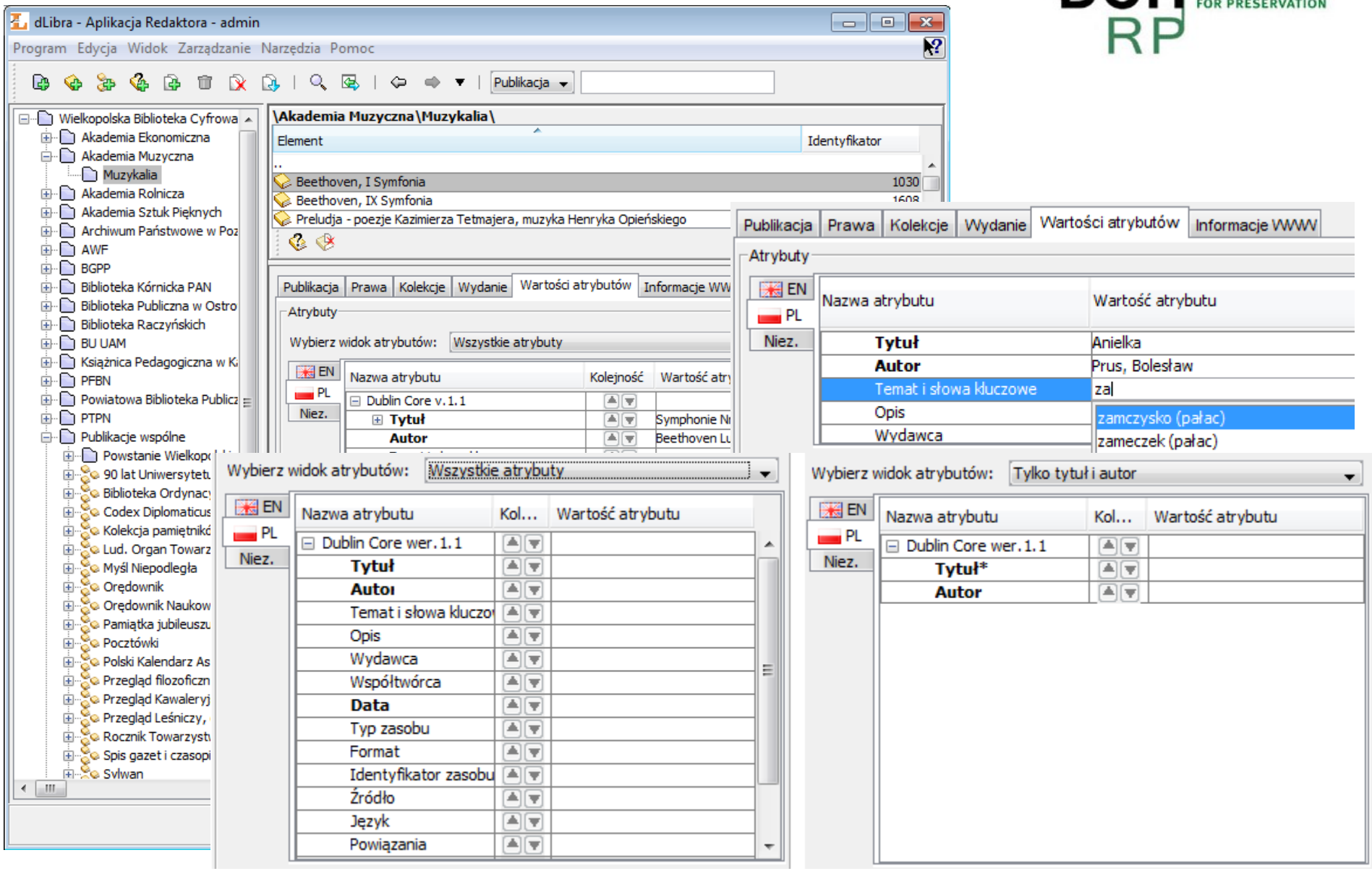
- **dArceo: Features interesting for PoC: interoperability**
  - Supported back-ends:
    - Server's local filesystem
    - sFTP,
    - PLATON Archival Service
  - Integration options:
    - PLATON Archival Services (current approach)
    - Other storage back-ends: Grids and Clouds
    - External authN/authZ – e.g. Shibboleth-enabled
    - Could be delivered as private/public IaaS/SaaS cloud

WP5's proof of concept dLibra elaborated (1)

- **dLibra – deals with content / collections delivery, data&meta-data lookup/ searching content access, presentation versions management, editing and reading**
  - ± 100 digital libraries,
  - 100s memory institutions
  - 1,1M+ of digital objects
  - 98% content delivered via services based on dLibra

  - **Some novel features :**
  - Access performance optimisation
    - on the fly content conversion of content
    - Mobile devices optimisation
    - Streaming and progressive access

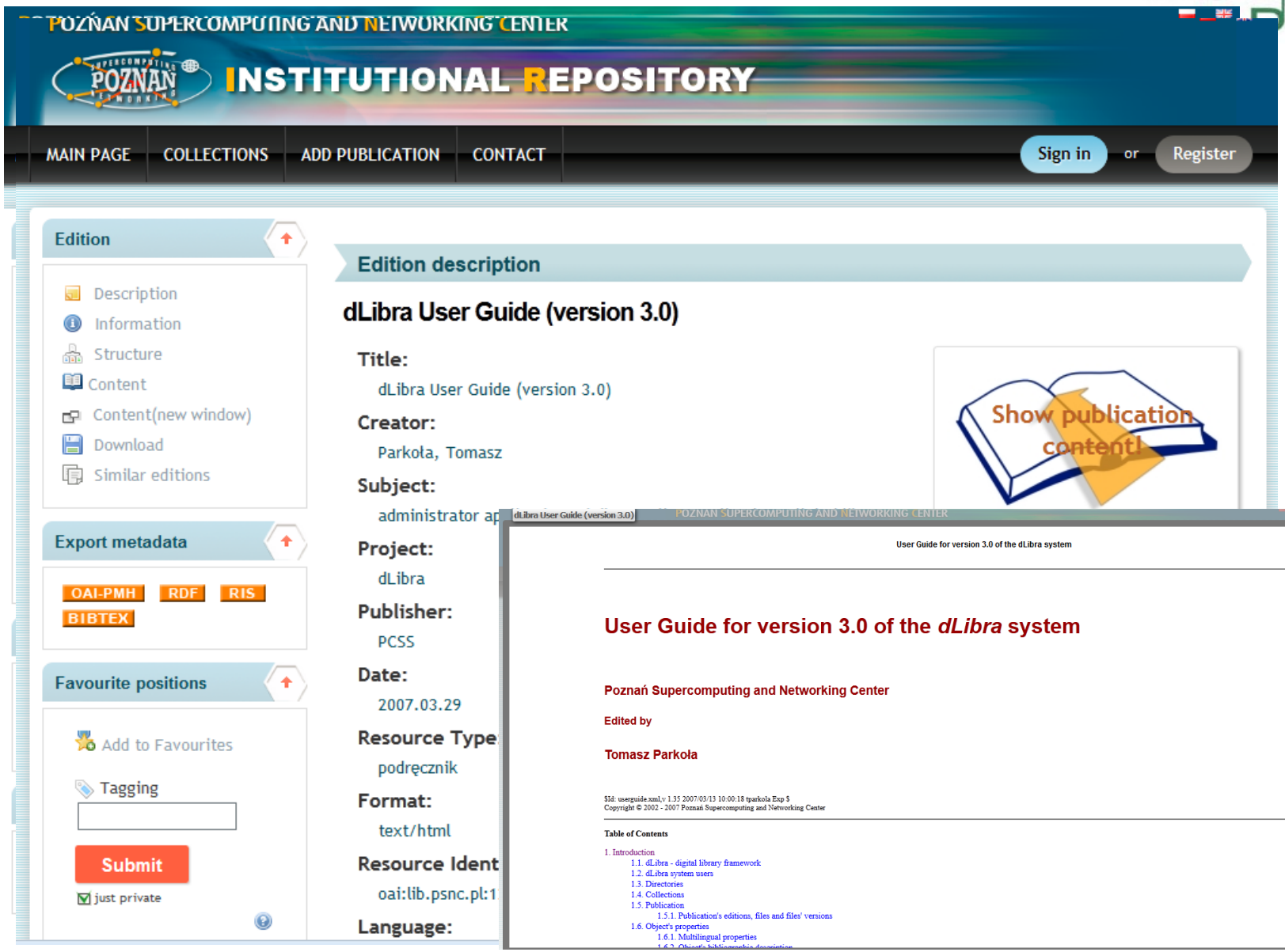
dLibra look & feel (1) editor's view



dLibra look & feel (2) reader's view

dLibra
DIGITAL LIBRARY FRAMEWORK

DCH
DIGITAL CULTURAL HERITAGE ROADMAP FOR PRESERVATION



WP5's proof of concept dLab (1)

- **dLab – overall**
  - automates workflow
  - Puts together dLibra and dArceo
- **Controls**
  - Ingesting Master and presentation-optimised versions of data to the systems (dArceo, dLibra)
  - Automated conversion to presentation version (e.g. ->PDF)
  - Automated archival of ingested data
  - Monitoring the progress of the tasks, reporting, restarting etc.

WP5's proof of concept PLATON Archive Services elaborated (1)

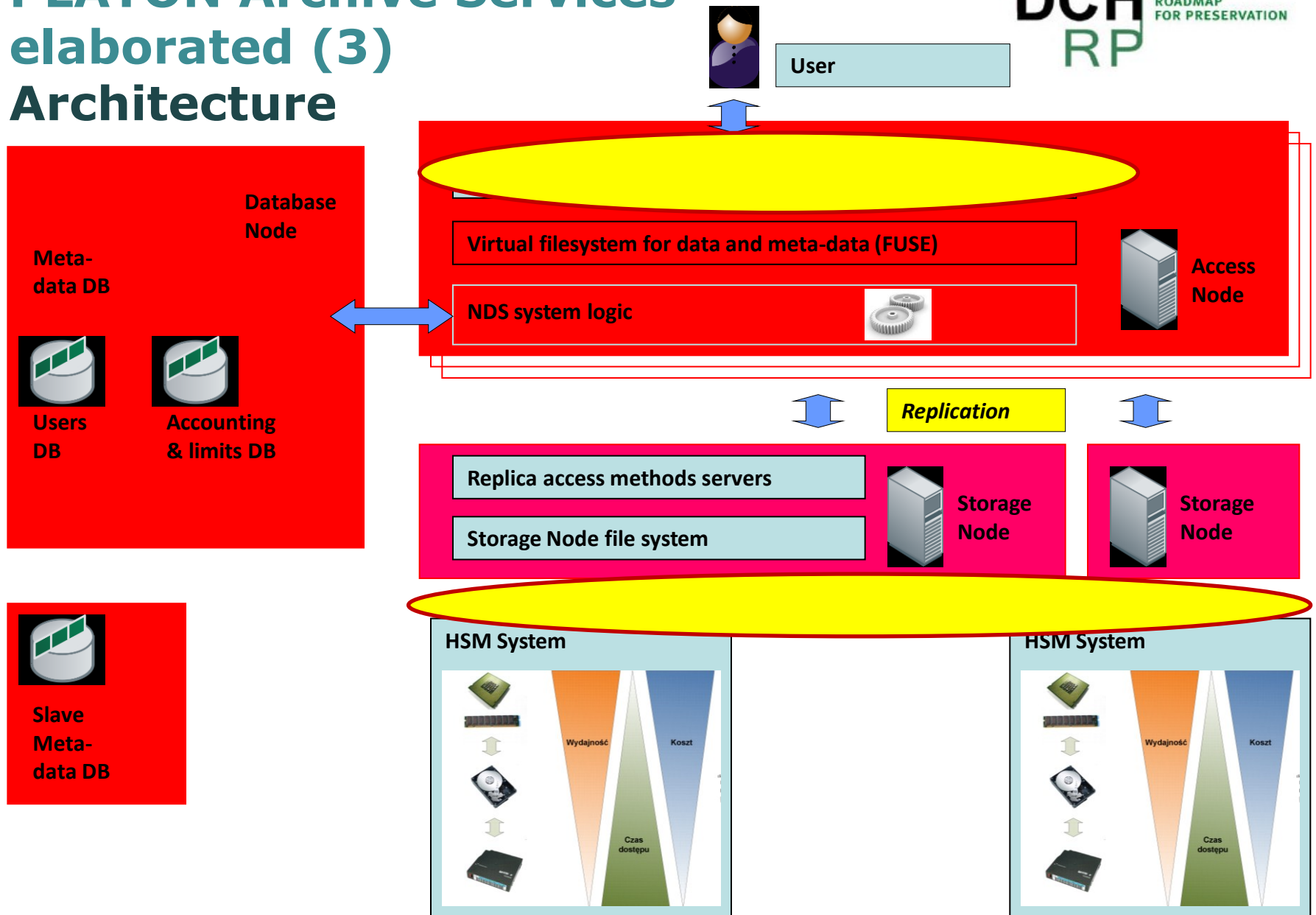
- **Features/architecture:**
  - Reliable data storage with efficient access
  - Abstract logical view for users
  - Automated on-the fly, transparent and reliable data replication
    - At least 2 geographically distant replicas
    - User has the logical view (he may also see details)
  - Long-term storage features:
    - Background integrity checks,
    - Replica fault detection, automated replica reconstruction
    - Storage technology migration handled by service provider – again transparent to users

WP5's proof of concept PLATON Archive Services elaborated (2)

- **Front-end interfaces:**
  - sFTP, WebDaV, GridFTP
- **Back-end storage:**
  - So-called Storage Nodes are POSIX filesystem storage elementy from the point of view of Popular Archival Services software

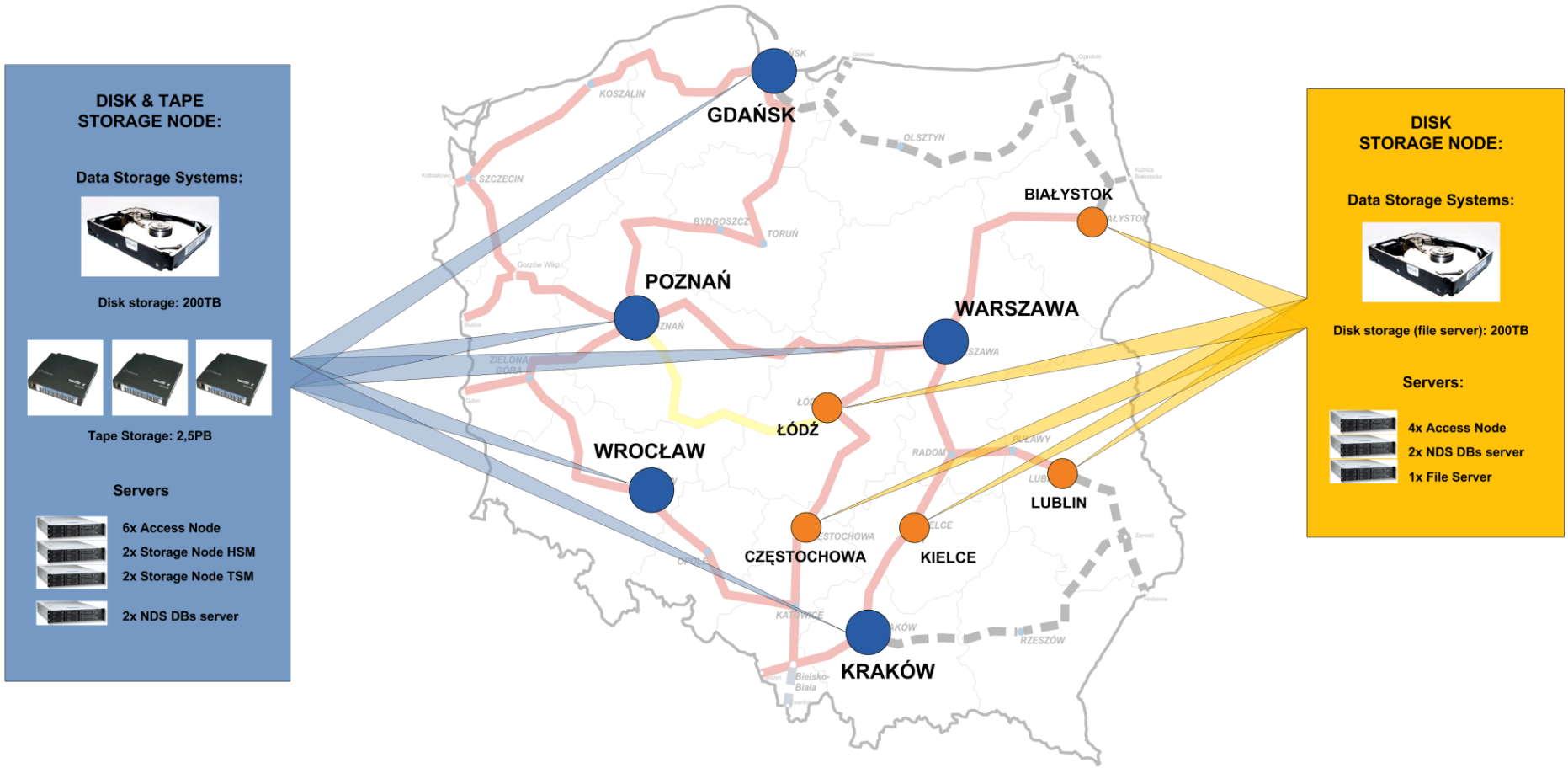
WP5's proof of concept PLATON Archive Services elaborated (3)

Architecture



WP5's proof of concept PLATON Archive Services elaborated (4)

Service deployment and infrastructure

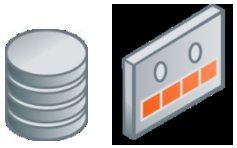
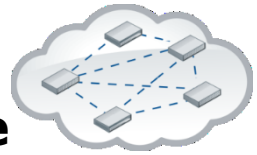


WP5's proof of concept PLATON Archive Services elaborated (4)

Service deployment and infrastructure

WP5's proof of concept
Possible combinations

- **High-level services:**
  - dArceo/ dLibra/ dLab 
  - eCSG 
  - Can we put these into cloud? 

- **Low-level storage services:**
  - **LTS:** reliable, long-term storage 
  - **Cloud storage** 
    - S3-compliant,
    - CDMI-compliant
  - **Grid storage** 
    - GridFTP
    - FTS
    - DPM

EUDAT <-> DCH-RP

Maciej Brzeźniak, Norbert Meyer, PSNC

EUDAT – overall information

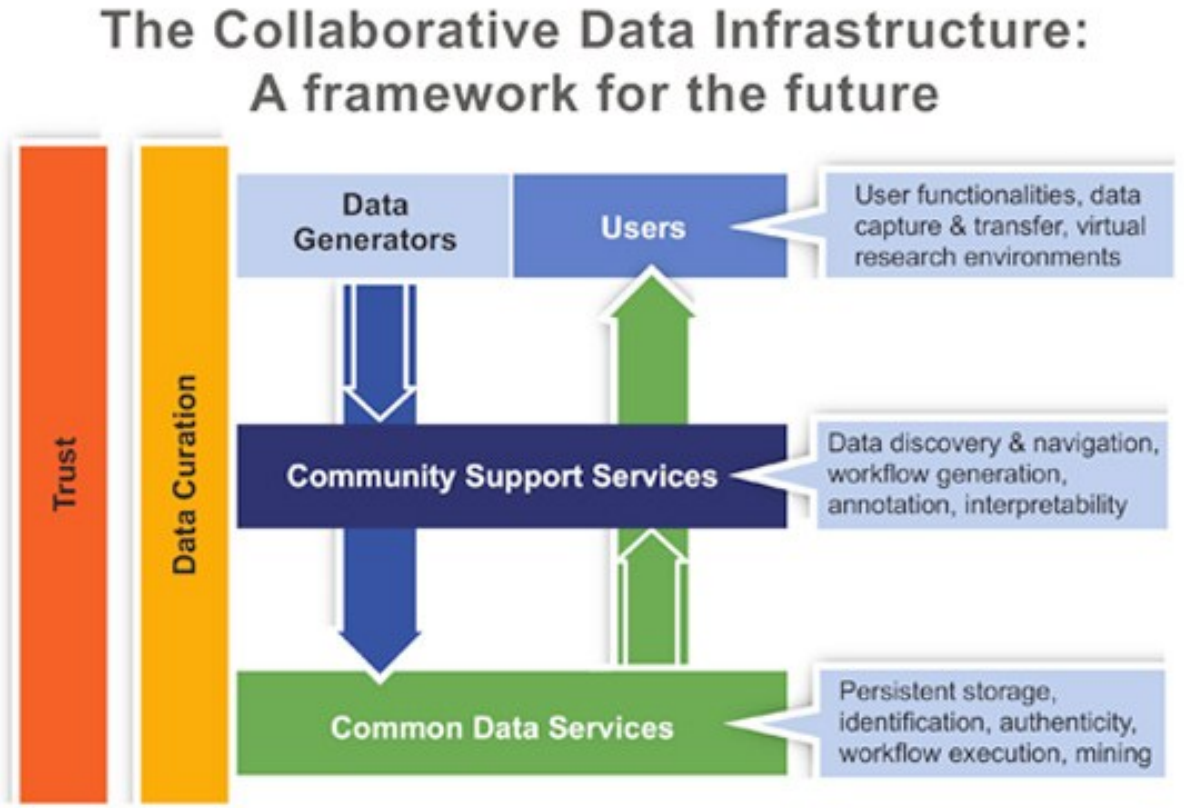
❑ EUDAT – European Data Infrastructure

- **Vision**: to support a Collaborative Data Infrastructure
- **Aims**:
  - Provide a **sustainable platform** of **technologies**, **tools** and **services** driven by user needs
  - Engage users in defining/shaping a platform for shared services
  - Support **data-intensive, multi-disciplinary research**:
    - Humanities and Social Science: CLARIN
    - But also: earth science (ENES, Earth system modelling; EPOS: European Plane Observing System), ecology (LifeWatch), Virtual Physiological Human (VPH)
  - Deliver **common low-level services** that are required to provide the level of interoperation and trust of data
  - Ensure that the data **infrastructure is robust/scalable** (able to address ‚data tsumami')
  - **Build community/domain-specific services** on top of the common services with participation of users

EUDAT – overall information

❑ EUDAT – European Data Infrastructure

- Common vs community-/domain-specific services



Source: High Level Expert Group on Scientific Data, Riding the wave, 2010.

EUDAT – overall information

EUDAT's services:

- ***Note:*** *EUDAT works in* ***iterations*** *addressing* ***urgent needs first***

- **Services ready for production**
  - **Data Replication:**
    - automated replication (iRODS)
    - PID registration (EPIC)
  - **Data Staging:**
    - Staging data from user community premises/systems (iRODS)
    - to computing systems, e.g. PRACE's HPC centres (GridFTP, FTS)

- **Services under development:**
  - **Meta-Data Service:**
    - Harvesting (OAI-PMH), searching; semantic mapping etc.
    - D-NET, CKAN, LUCENE Solr, others
  - **AAI services:**
    - iRODS-Shibboleth
    - Contrail, Moonshot
  - **Simple Store service:**
    - Invenio…

EUDAT – overall information

EUDAT
European Data Infrastructure

DCH RP
DIGITAL CULTURAL HERITAGE ROADMAP FOR PRESERVATION

❑ EUDAT partners and numbers:

AUSTRIA
umweltbundesamt

CZECH REPUBLIC

FRANCE
CINES, CERFACS, maatG

ITALY
CINECA, INGV

SPAIN
Barcelona Supercomputing Center, Red IRIS

NETHERLANDS
SURF SARA

POLAND
POZNAN

SWEDEN
SNIC

NORWAY
UNINETT Sigma

FINLAND
CSC

UNITED KINGDOM
Science & Technology Facilities Council, epcc, UCL

SWITZERLAND
CERN

GERMANY
DKRZ, JÜLICH, KIT, Eberhard Karls Universität Tübingen, Max-Planck-Institut für Meteorologie, rzg