EUDAT

A European Collaborative Data Infrastructure

Daan Broeder
The Language Archive – MPI for Psycholinguistics
CLARIN, DASISH, EUDAT

SEVENTH FRAMEWORK PROGRAMME

EUDAT

OpenAire Interoperability Workshop
Braga, Feb. 8, 2013

EUDAT Key facts

Project Name	EUDAT – European Data
Start date	1st October 2011
Duration	36 months
Budget	16,3 M€ (including 9,3 M€ from the EC)
EC call	Call 9 (INFRA-2011-1.2.2): Data infrastructure for e-Science (11.2010)
Participants	25 partners from 13 countries (national data centers, technology providers, research communities, and funding agencies)
Objectives	"To deliver cost-efficient and high quality Collaborative Data Infrastructure (CDI) with the capacity and capability for meeting researchers' needs in a flexible and sustainable way, across geographical and disciplinary boundaries."

EUDAT

Consortium


CINES


epcc


UNINETT
sigma


Science & Technology
Facilities Council


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación


sara
Computing & Networking Services


SNIC


DKRZ


CERFACS


JÜLICH
FORSCHUNGSZENTRUM


CINECA


rzg
IPP
RECHEN-
ZENTRUM
GARCHING


umweltbundesamt
PERSPEKTIVEN FÜR UMWELT & GESELLSCHAFT


Max-Planck-Institut
für Meteorologie


POZNAN


CERN


CSC


Max Planck Institute for Psycholinguistics


KIT
Karlsruhe Institute of Technology


maatG


INGV

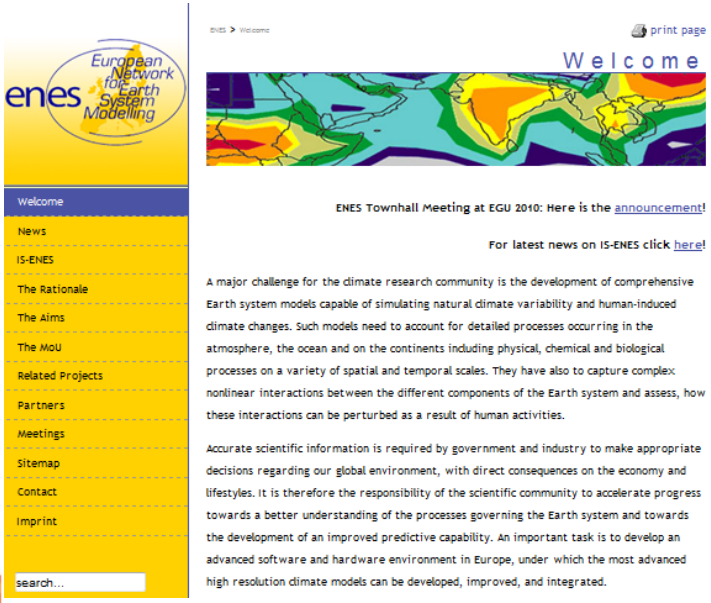

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN


UCL


EUDAT

First EUDAT Communities

ESFRI



EUDAT

EUDAT Core Service Areas

Community-oriented services

- Simple Data Acces and upload
- Long term preservation
- Shared workspaces
- Execution and workflow (data mining, etc.)
- Joint metadata and data visibility

Enabling services (making use of existing services where possible

- Persistent identifier service (EPIC, DataCite)
- Federated AAI service
- Network Services
- Monitoring and accounting

Core services are building blocks of EUDAT's Common Data Infrastructure
mainly included on bottom layer of data services

EUDAT

First Service Cases

- List of 6 service/use cases identified

  - Safe replication: Allow communities to safely replicate data to selected data centers for storage and do this in a robust, reliable and highly available way.

  - Dynamic replication: Perform (HPC) computations on the replicated data. Move (part of) the safely replicated data to a workspace close to powerful machines and move the results back into the archives.

  - Metadata: A joint metadata domain for all data that is stored by EUDAT data centers by harvesting metadata records for all data objects from the communities. Allow to have a catalogue to demonstrate what EUDAT stores, and to have a registry which can be used for automatic operations such as data mining.

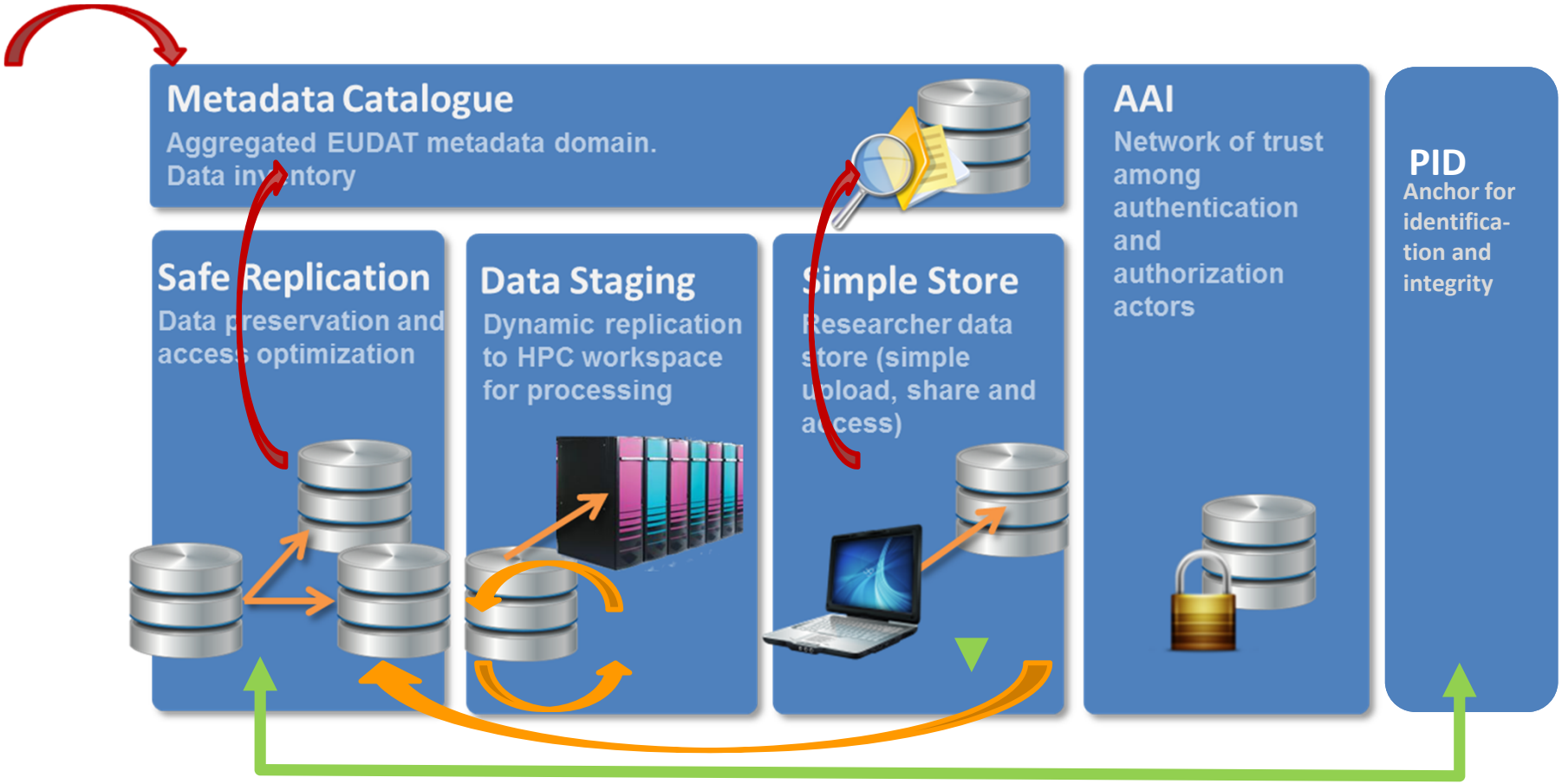
  - Research data store: A function that will help researchers mediated by the participating communities to upload and store data which is not part of the officially handled data sets of the community.

  - AAI: A solution for a working AAI system in a federation scenario.

  - PID: a robust, highly available and effective PID system that can be used within the communities and by EUDAT.

EUDAT

EUDAT – Architecture



Metadata

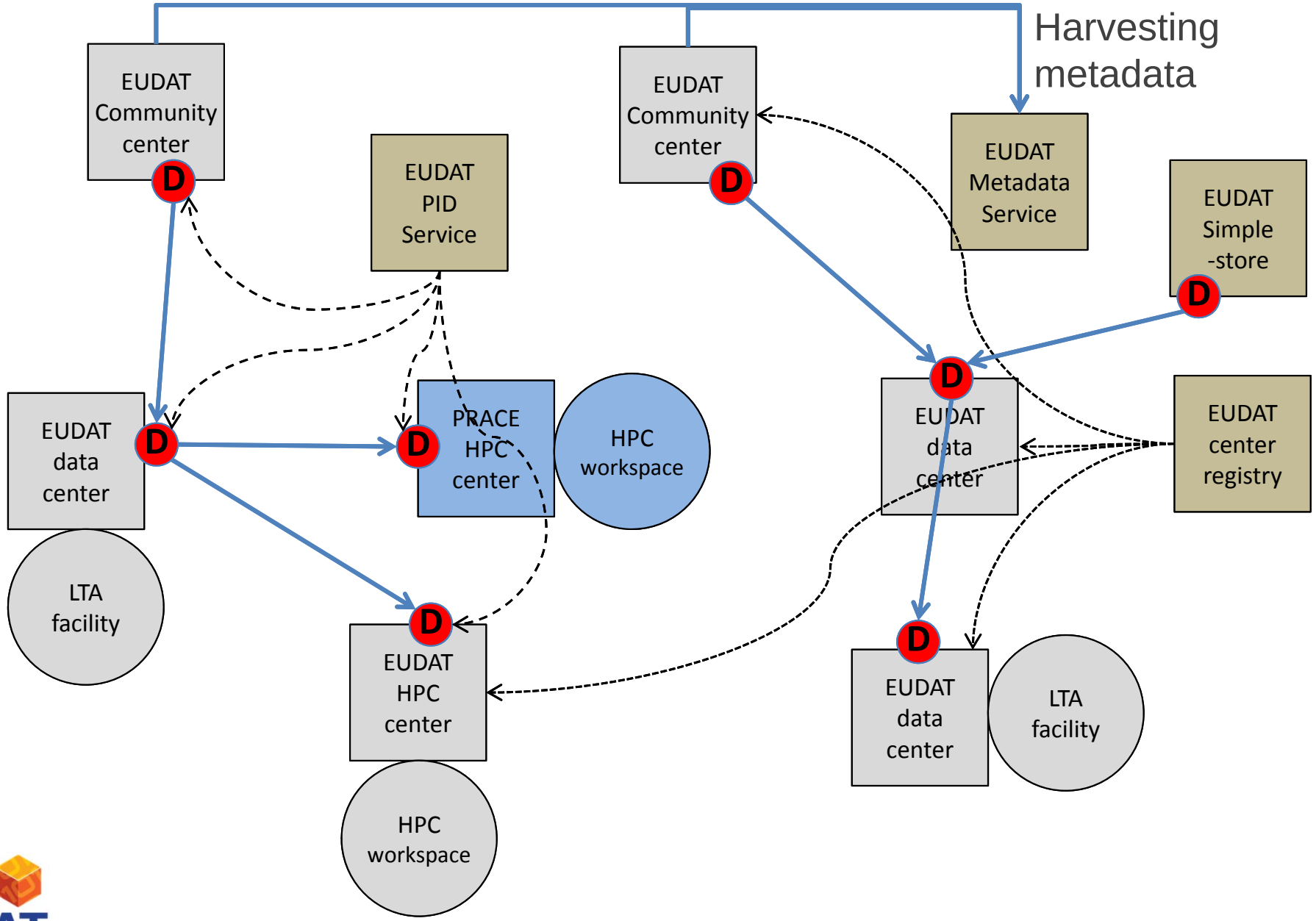
Data

PID

EUDAT

SEVENTH FRAMEWORK PROGRAMME

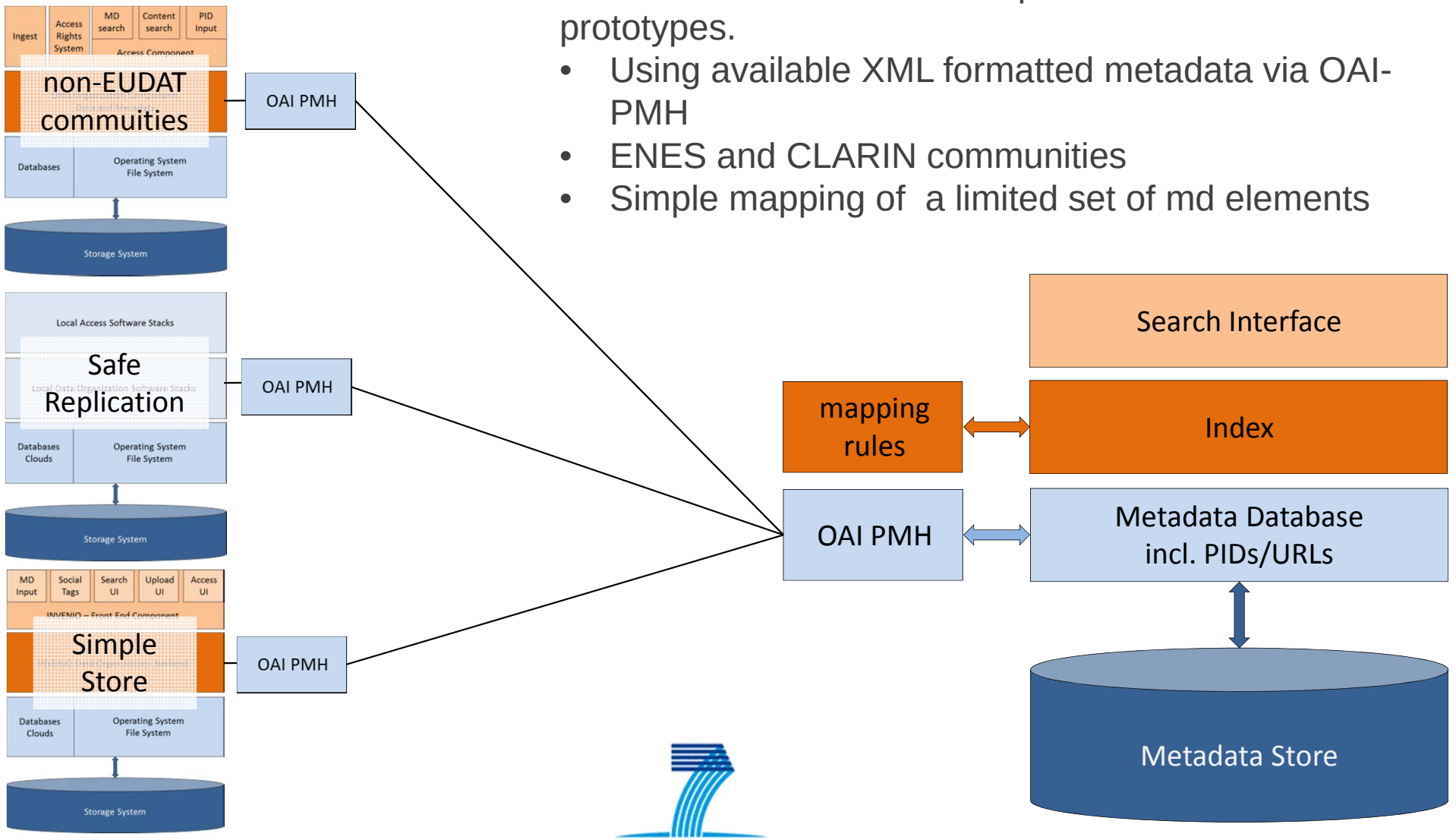
EUDAT Data & Community Centers



EUDAT – Architecture Joint Metadata Catalogue

Joint metadata will be developed as a series of prototypes.

- Using available XML formatted metadata via OAI-PMH
- ENES and CLARIN communities
- Simple mapping of a limited set of md elements



SEVENTH FRAMEWORK PROGRAMME

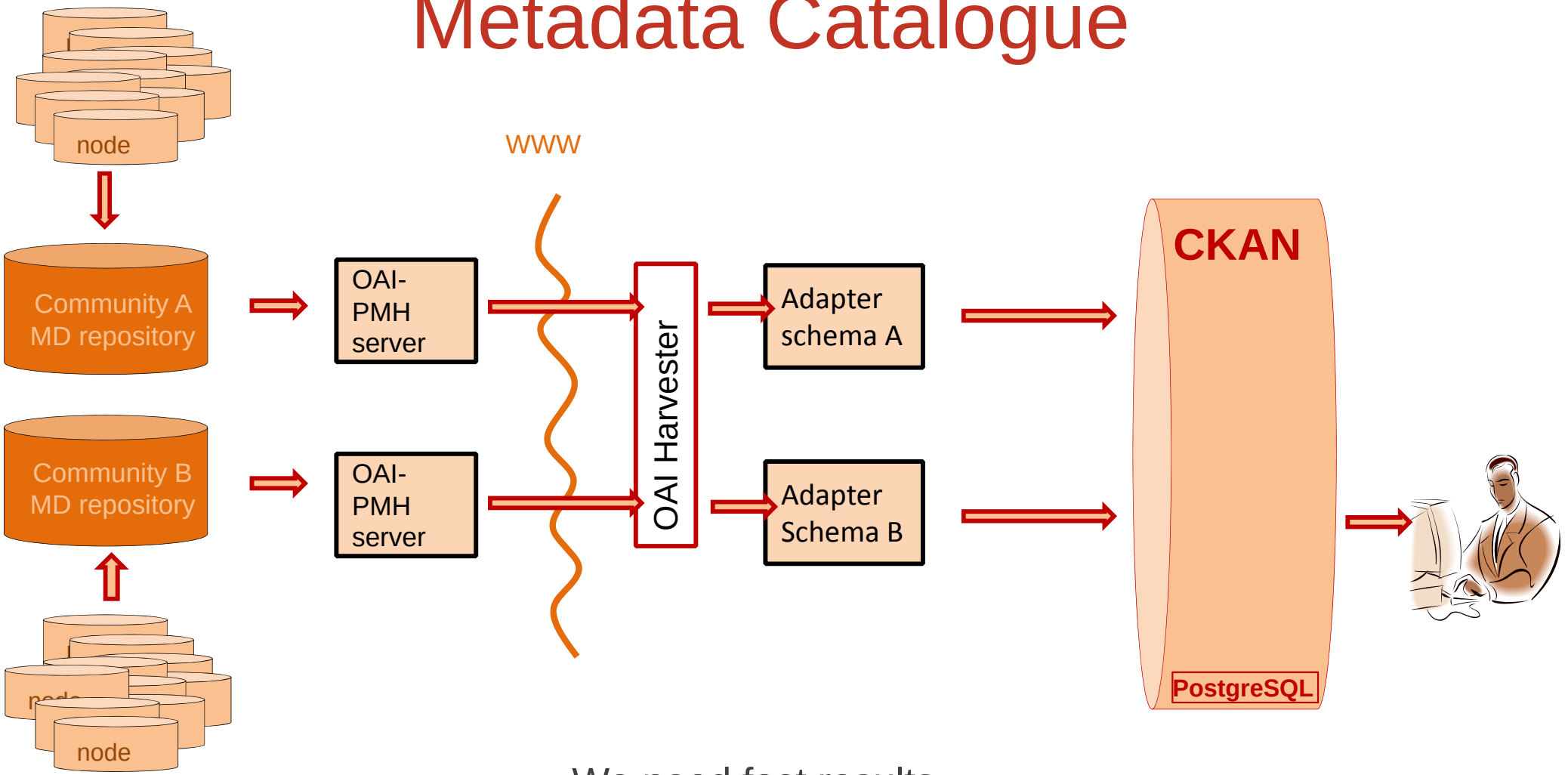
EUDAT

User perspective joint metadata catalogue

- From the user's perspective the following points are considered important:
  - What metadata terminology can be used in the catalogue interface
  - What are the browsable/searchable dimensions
  - How is the data presented e.g. what visualization options are available
  - What added value can there be?
    - Cross community data
    - Commenting facility (also for metadata quality improvement)
    - Direct visibility of data uploaded in the EasyStore
    - Good visualization of different community data types
    - Linking to original metadata records and resource
    - Harmonized data citation facility
    - Export as LoD
- *Consider also its use for the community own purposes*
  - *Starting communities, that do not yet have a proper catalogue*

EUDAT

Metadata Catalogue

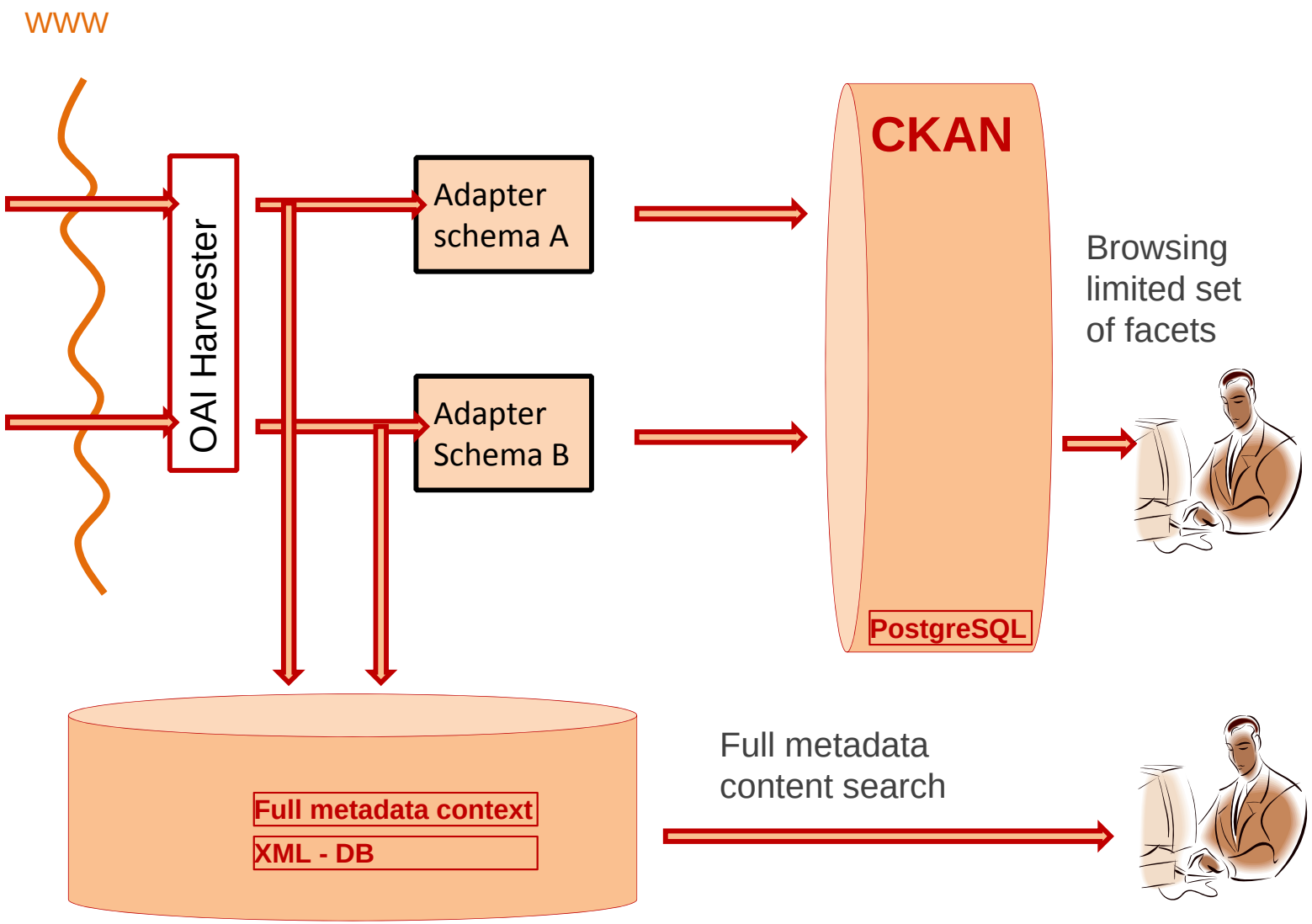


We need fast results

- Based on community practices
- No new schema developments
- Using metadata that is already out there
- Using existing and proven technology if possible

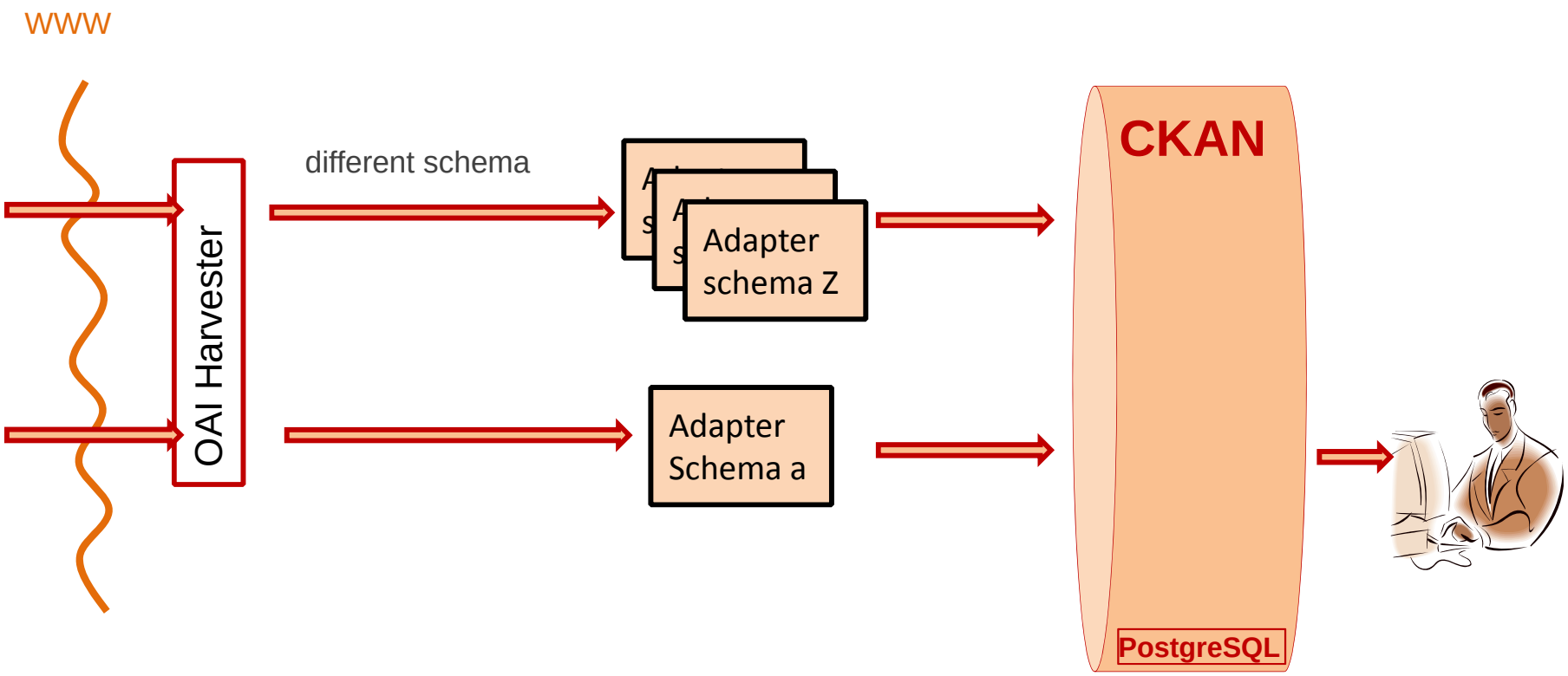
EUDAT

More Functionality



EUDAT

CLARIN particulars



CLARIN uses CMDI as a metadata infrastructure
Users can specify metadata schema as they see fit to describe a particular resource type.
This results in many (>>100) metadata schemas

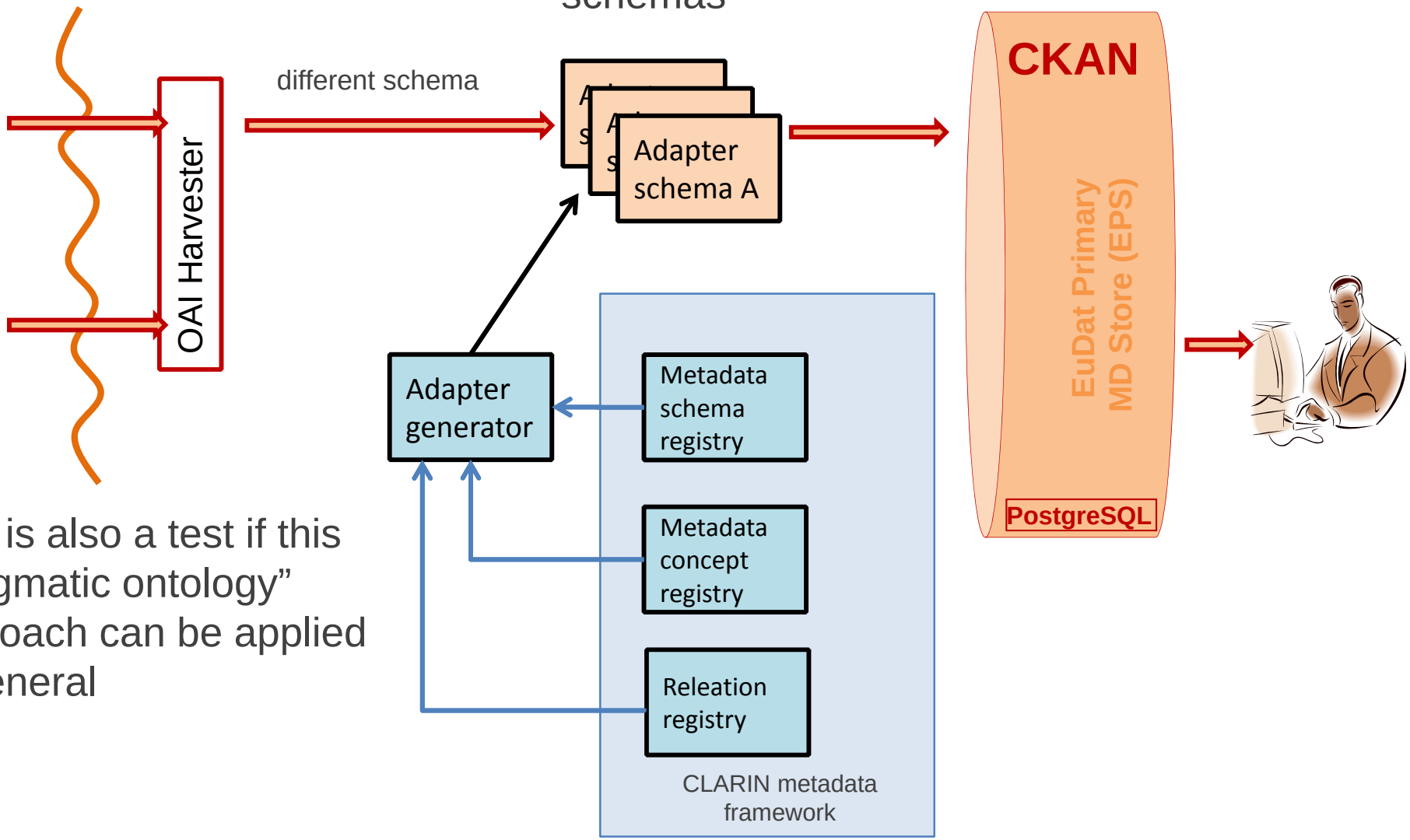
EUDAT

CLARIN particulars

CLARIN uses CMDI as a metadata infrastructure

Users can specify metadata schema as they see fit to describe a particular resource type.

This results in many (>>100) metadata schemas



This is also a test if this “pragmatic ontology” approach can be applied in general

EUDAT & Interoperability

In this context we use a ‘narrow’ definition from the IEEE Glossary that defines interoperability as:

> *the ability of two or more systems or components to exchange information and to use the information that has been exchanged*

Interoperability consequences within EUDAT between data and community centers with maybe some consequences for the community practices

- Use of iRods as a data replication software component.
  - Used within that a set of policies (general) governing the exchange of data between the centers.
- Use of FIM and/or X509 certificates to access EUDAT services
- Use of a specific PID system; with some added information items
  - EPIC, DataCite both based on Handle System technology
- Use of OAI-PMH harvesting of metadata:
  - possibly developing an EUDAT standardized use of a metadata concept and relation registry
- *Other technology choices for data-infra components as CKAN as a metadata catalogue en INVENIO as CMS for the SimpleStore are not so intruisive since they are central facilities that can be replaced*

EUDAT

Interoperability of ideas

- There are many communities active with data management
- Developing ideas, tools & infrastructure or both
- To make ideas come together or at least get informed about differences. We need to talk!

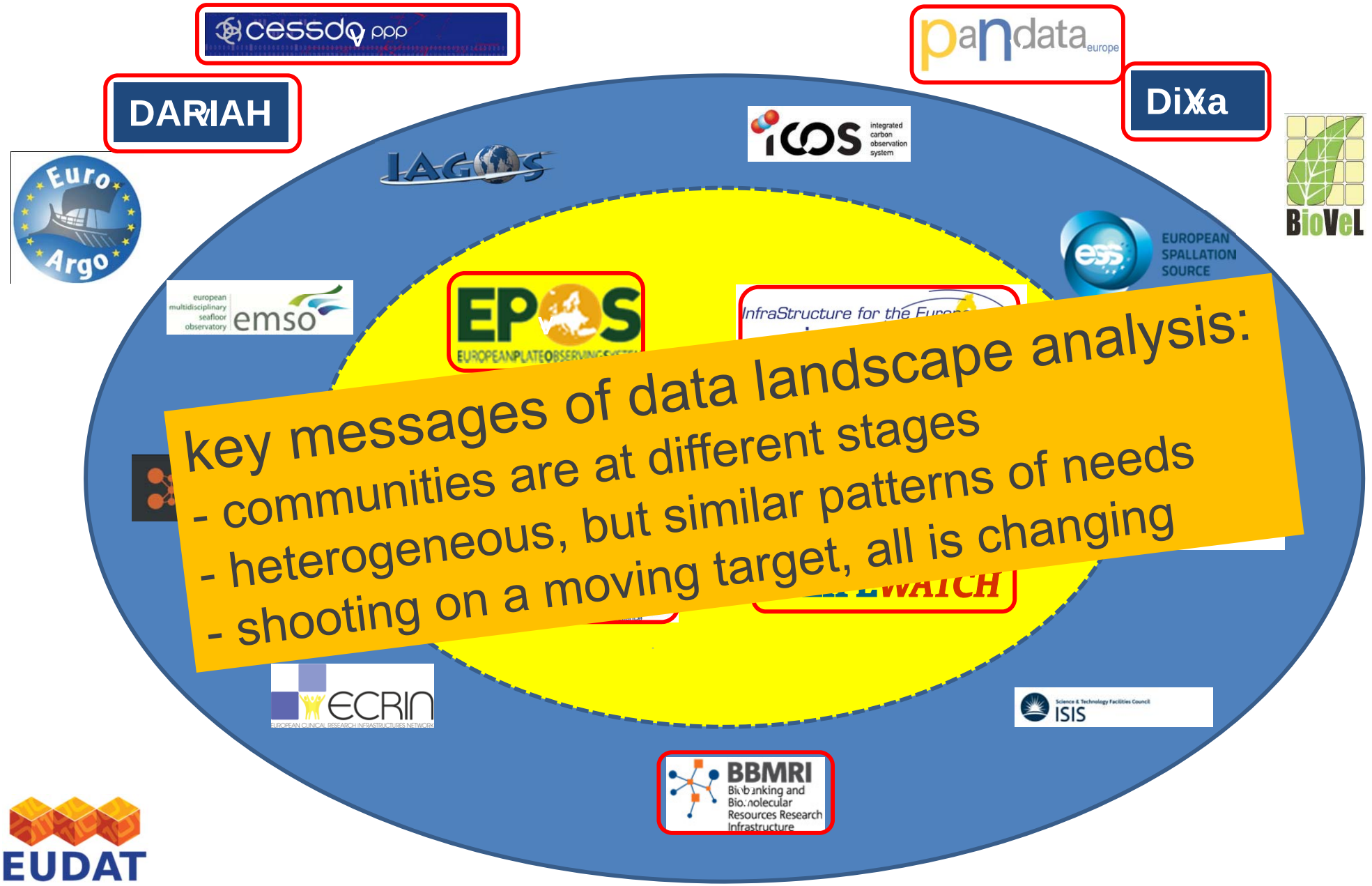
- RDA – Research Data Aliance
  - Supported by EC, US, Aus
  - EUDAT, OpenAire via the iCordi project
  - Bottom up, community driven, (including data-scientists)

EUDAT

Thank you for your attention

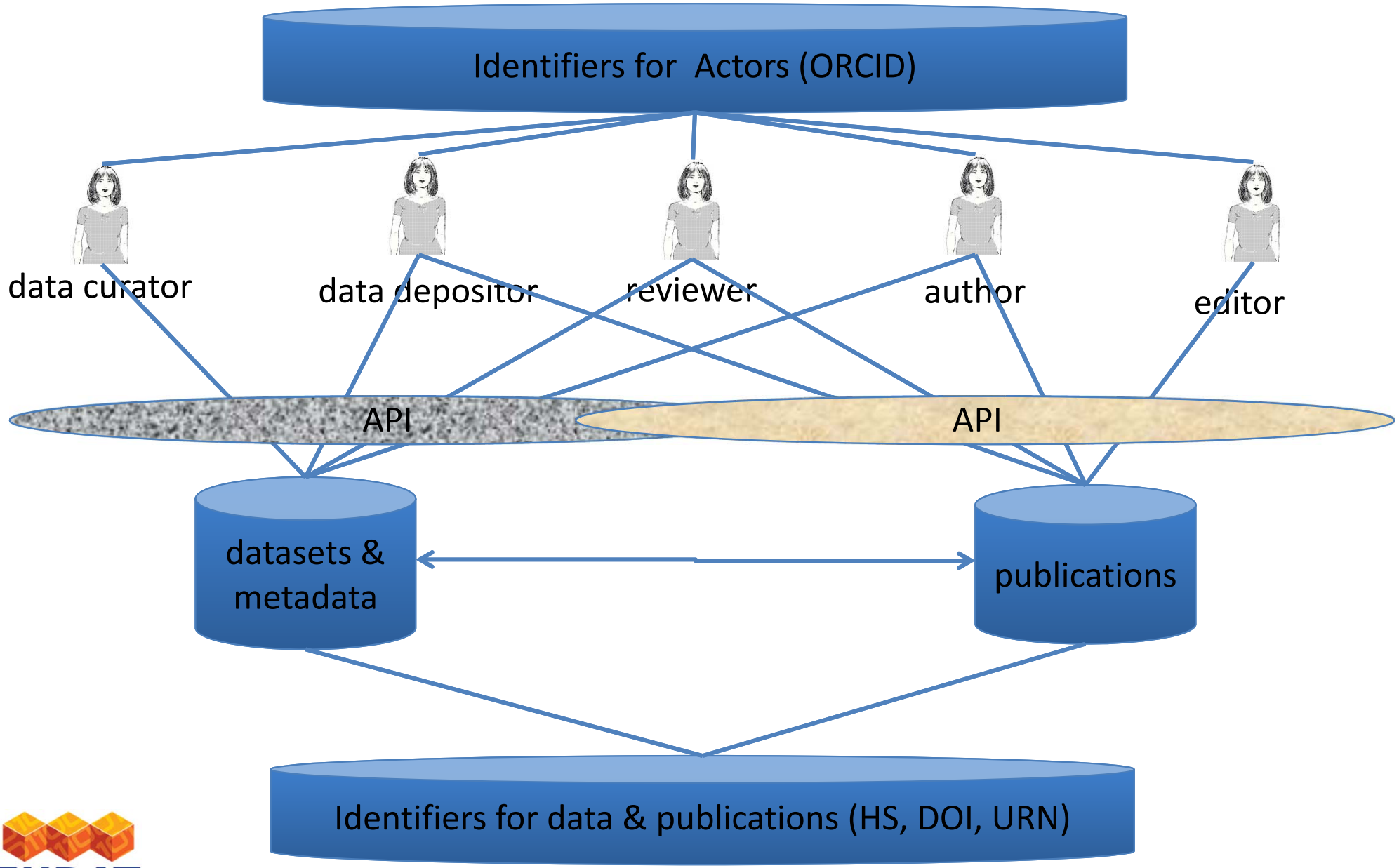
EUDAT

EUDAT Services for Communities!?



EUDAT

Interlinking data and publications



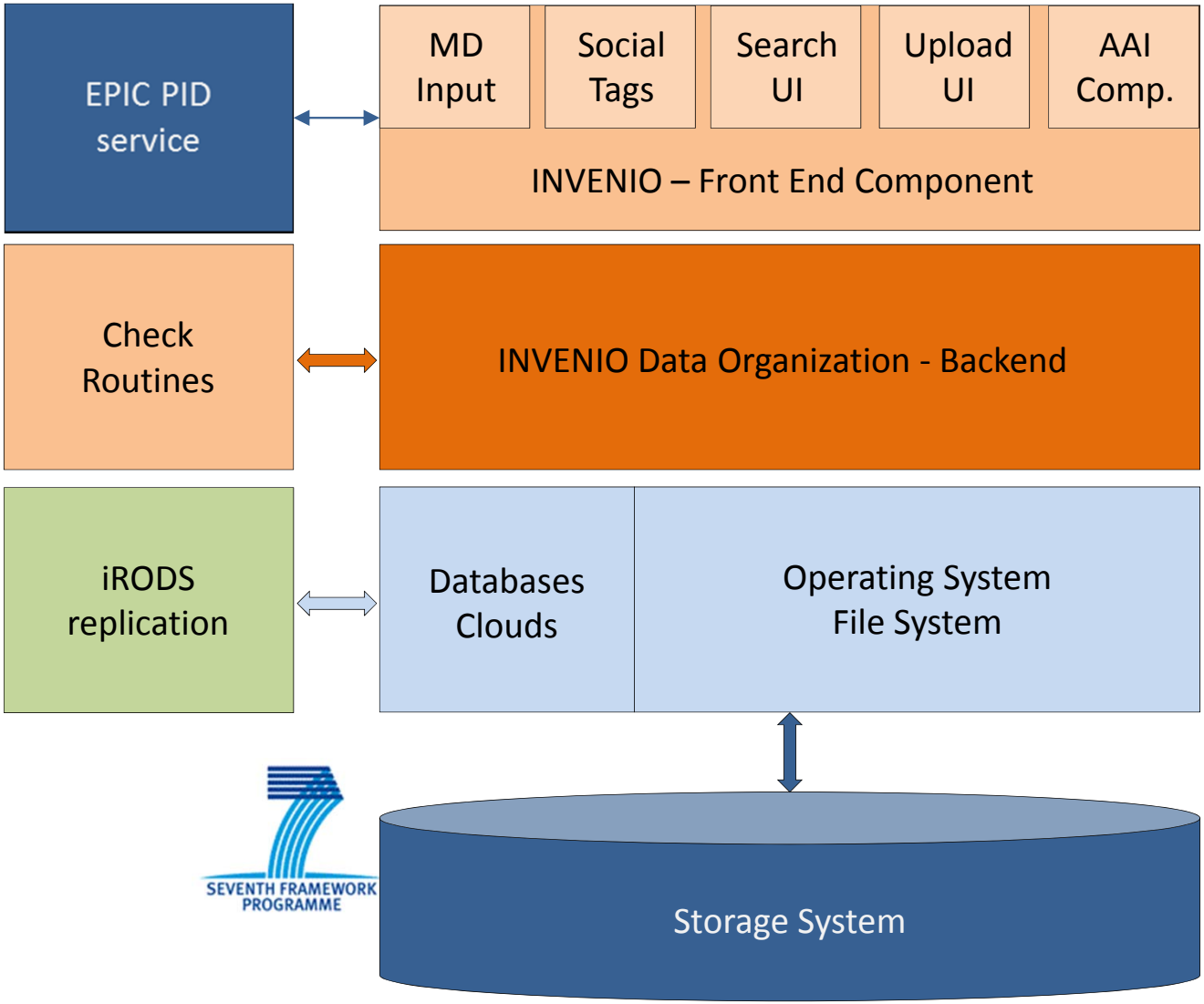
EUDAT

EUDAT – Architecture Simple Store (SS)

The architecture of the Simple Store service is driven by the Invenio backend machinery and an adapted and optimized front end that has all functionality as required by EUDAT. The backend system makes use of underlying standard components, i.e. all metadata, social tag data etc. is stored in a SQL database with references to files stored in a file system or cloud dependent on the kind of component the data center is using.

Important for EUDAT is that we

- will not interfere with the setup at the hosting data center
- will not be dependent on Invenio but be able to export all metadata and data to new technology if necessary
- will associate PIDs with all objects once uploaded
- will carry out checks on the uploaded objects to facilitate later curation and access (check routines can come from libraries such as JHOVE or are self-defined)
- will be able to make use of the SR service to replicate all stored data (microservices to support this need to be developed)
- AAI is supporting many users from communities registered with IdPs and beyond.



EUDAT

Interoperability

From Wikipedia, the free encyclopedia

Interoperability is the ability of diverse systems and organizations to work together (inter-operate). The term is often used in a technical systems engineering sense, or alternatively in a broad sense, taking into account social, political, and organizational factors that impact system to system performance.

While **interoperability** was initially defined for IT systems or services and only allows for information to be exchanged (see definition below), a more generic definition could be this one:

> *Interoperability is a property of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, without any restricted access or implementation.*

This generalized definition can then be used on any system, not only information technology system. It defines several criteria that can be used to discriminate between systems that are "really" inter-operable and systems that are sold as such but are not because they don't respect one of the aforementioned criteria, namely:

- non-disclosure of one or several interfaces
- implementation or access restriction built in the product/system/service

The IEEE Glossary defines interoperability as:

> *the ability of two or more systems or components to exchange information and to use the information that has been exchanged.*[1]

James A. O'Brien and George M. Marakas define interoperability as:[2]

> *Being able to accomplish end-user applications using different types of computer systems, operating systems, and application software, interconnected by different types of local and wide area networks.*

EUDAT