National Initiatives for Open Science in Europe

Technical aspects of ORDM: PIDs, tools and best practices

Themis Zamani, GRNET

Facts = Users + Data

Good day,

I am interested **in using PID** to the repository of data that we are handling.

What do you have to do to sign me up?

Best regards

Raw data

Please provide me with an account I can use to **obtain PIDs for an application** we're creating to store and publish our University's working **papers**.

Kind regards,

Papers

Hello support !

I'm writing to you to get some technical support for a project. The task at hand is to issue data products persistent, unique identifiers, using your service. There are some technical issues around doing this which I would like to request some support for.

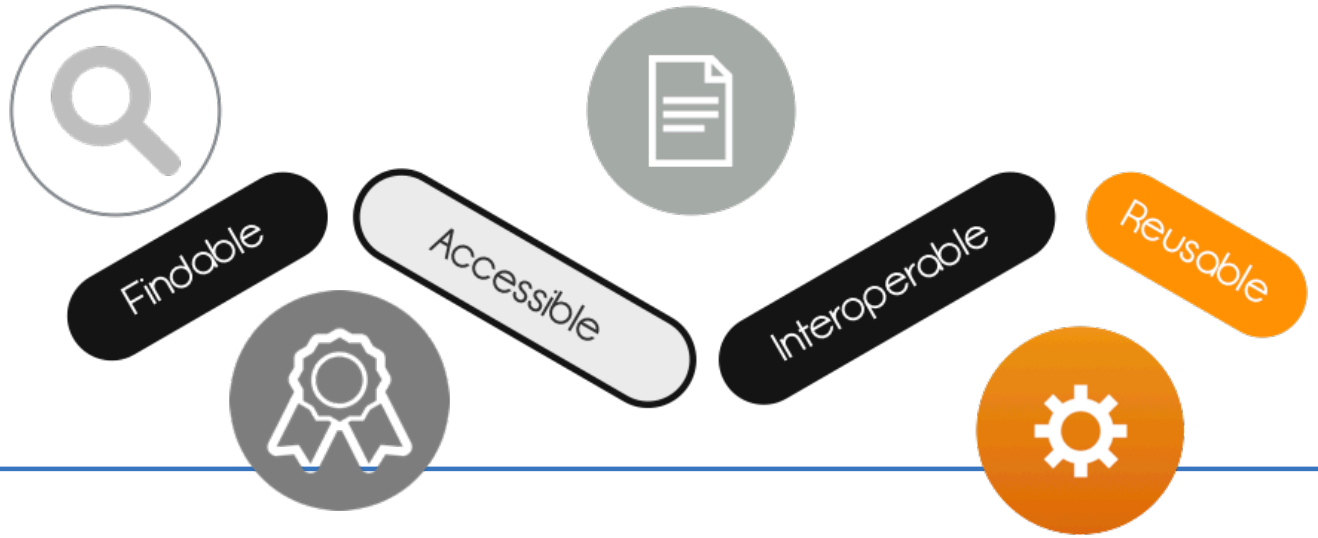
metadata

Different Users - with different types of data

Raw data **Papers** **metadata**

with a common goal

optimise the utility of the data



Content

- What are persistent identifiers?
- Why use persistent identifiers?
- Different persistent identifier systems
- The Handle System
- Initial Persistent Identifier (PID) policy for EOSC
- PID for your data
- Tools
- Policies
- Use cases

Persistent identifiers

For all, for specific domains?

Science & Data – Data & Science

Data generation is getting easier/cheaper

Complexity-shift from data generation to data processing & analysis

The amount of **data output is increasing**, quality is getting better

How to stimulate **reuse** and enable **reproducibility**?

Make the data world a better place

Data Principles

Data needs to be

- Findable — Easy to find
- Accessible — Stored for long term
- Interoperable — Ready to be combined with other datasets
- Reusable — Ready to be used and to be processed

by humans as well as computer systems;

with well-defined license and access

The solution : Persistent Identifiers

What is it?

An identifier is a unique identification code that is applied to “something”, so that the “something” can be unambiguously referenced.

A Persistent Identifier is an identifier that is **effectively permanently** **assigned to an object**.

Science & Data – Data & Science

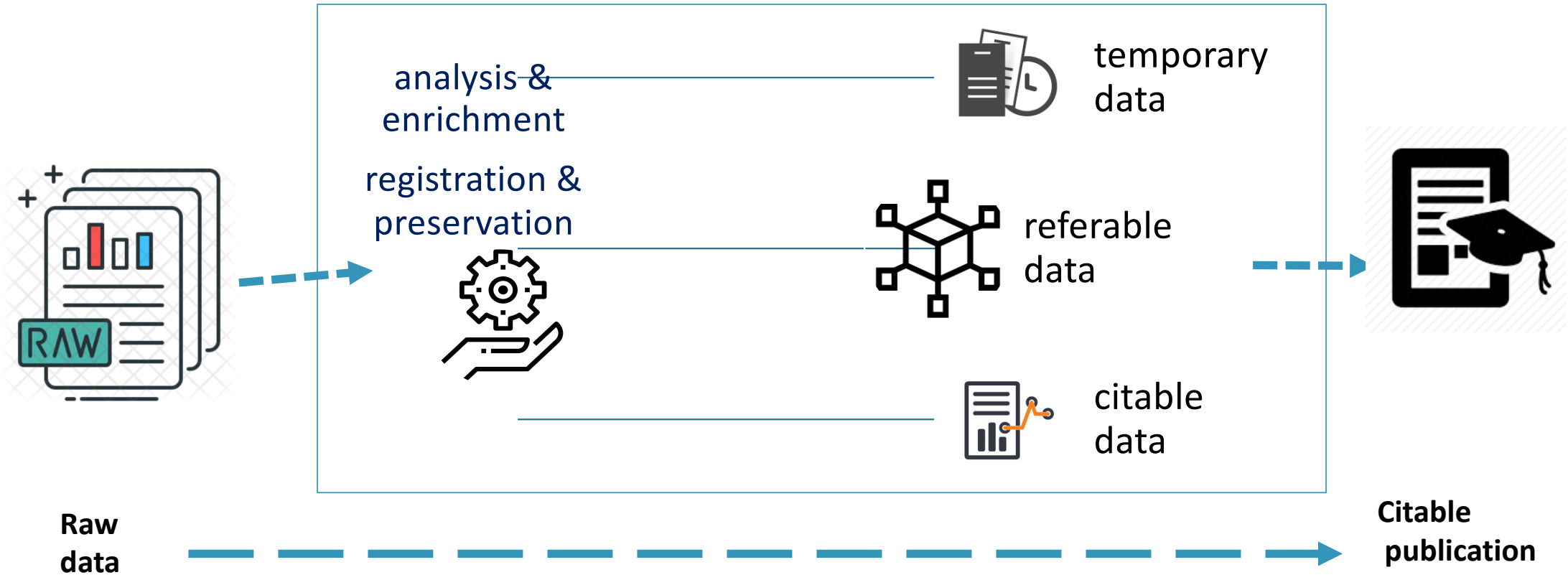
- PIDs are increasingly important and are being applied almost everywhere across sectors and disciplines, and for all types of digital objects.
- "sectors" covers science, industry, governments, health care, etc.
- Data management experts are becoming increasingly dependent on the availability of functioning persistent identifiers

FAIR DATA - FAIR guiding principles

- Persistent identifiers are well established in the research lifecycle
- Are part of The FAIR guiding principles:
  - ***a unique and persistent identifier should be assigned such that the data can be unequivocally referenced and cited in research communications.***
- Central to the realisation of FAIR are FAIR Digital Objects, which may represent data, software or other research resources. These digital objects must be accompanied by persistent identifiers, metadata and contextual documentation to enable discovery, citation and reuse.
- FAIR:
  - To be Findable: F1. (meta)data are assigned a globally unique and persistent identifier
  - To be Accessible: A1. (meta)data are retrievable by their identifier using a standardized communications protocol *(stored, accessed and/or downloaded with well-defined license and access)*

FAIR data are data which meet principles

Data Creation Cycle



Why not use URLs?

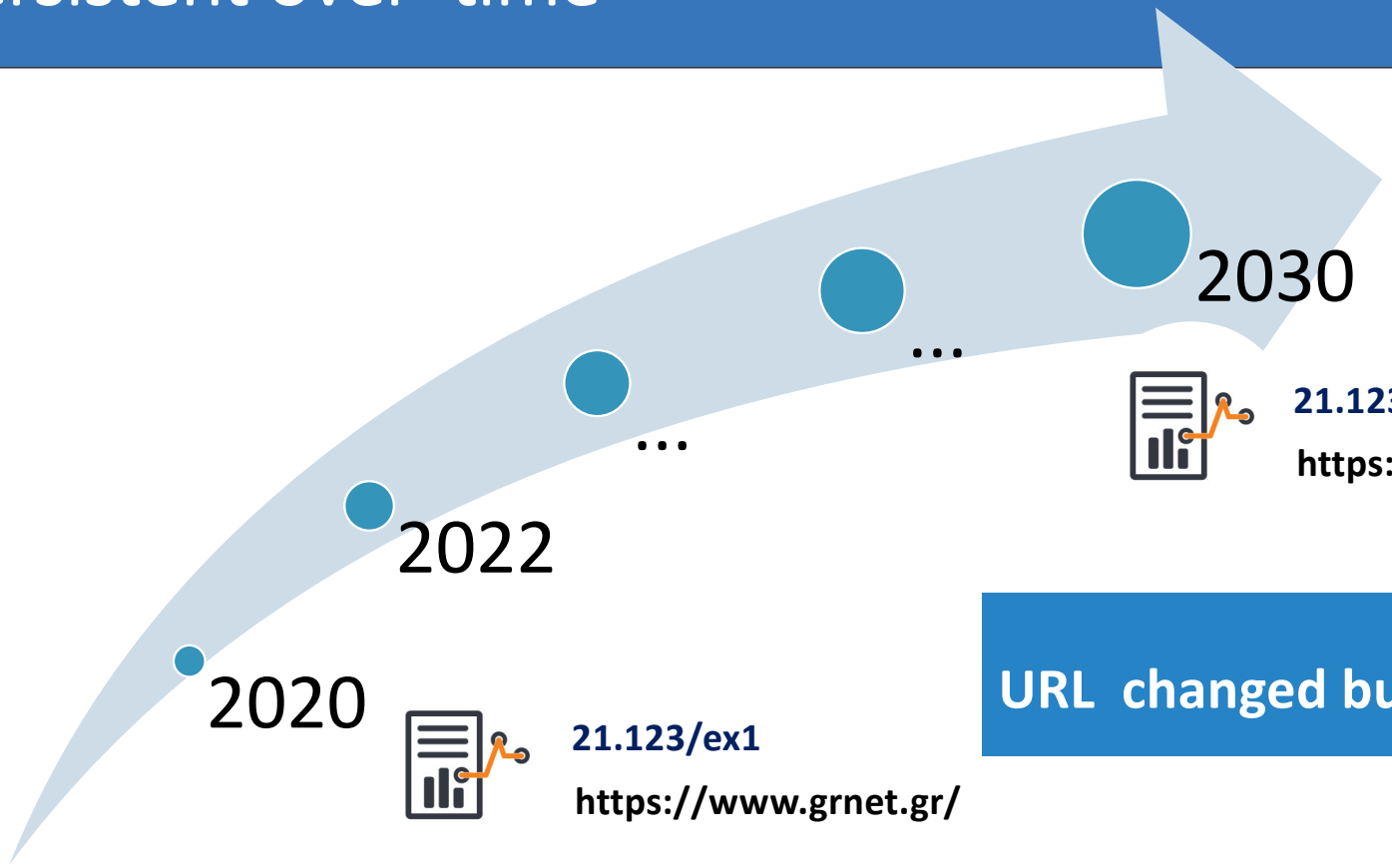
The URL specifies the **location**, on a **particular server**, from which the **resource** could be **retrieved**. **Strictly network locations** for digital resources.

BUT

- domain may change
- resource may be relocated
- link may change

In the long term **URLs a year later, often no longer work**

Persistent over time



21.123/ex1

https://www.grnet.gr/

21.123/ex1

https://www.grnet.gr/moved/

URL changed but PID is the same

Structure of a Persistent Identifier

points to a resource

21.123/ex1

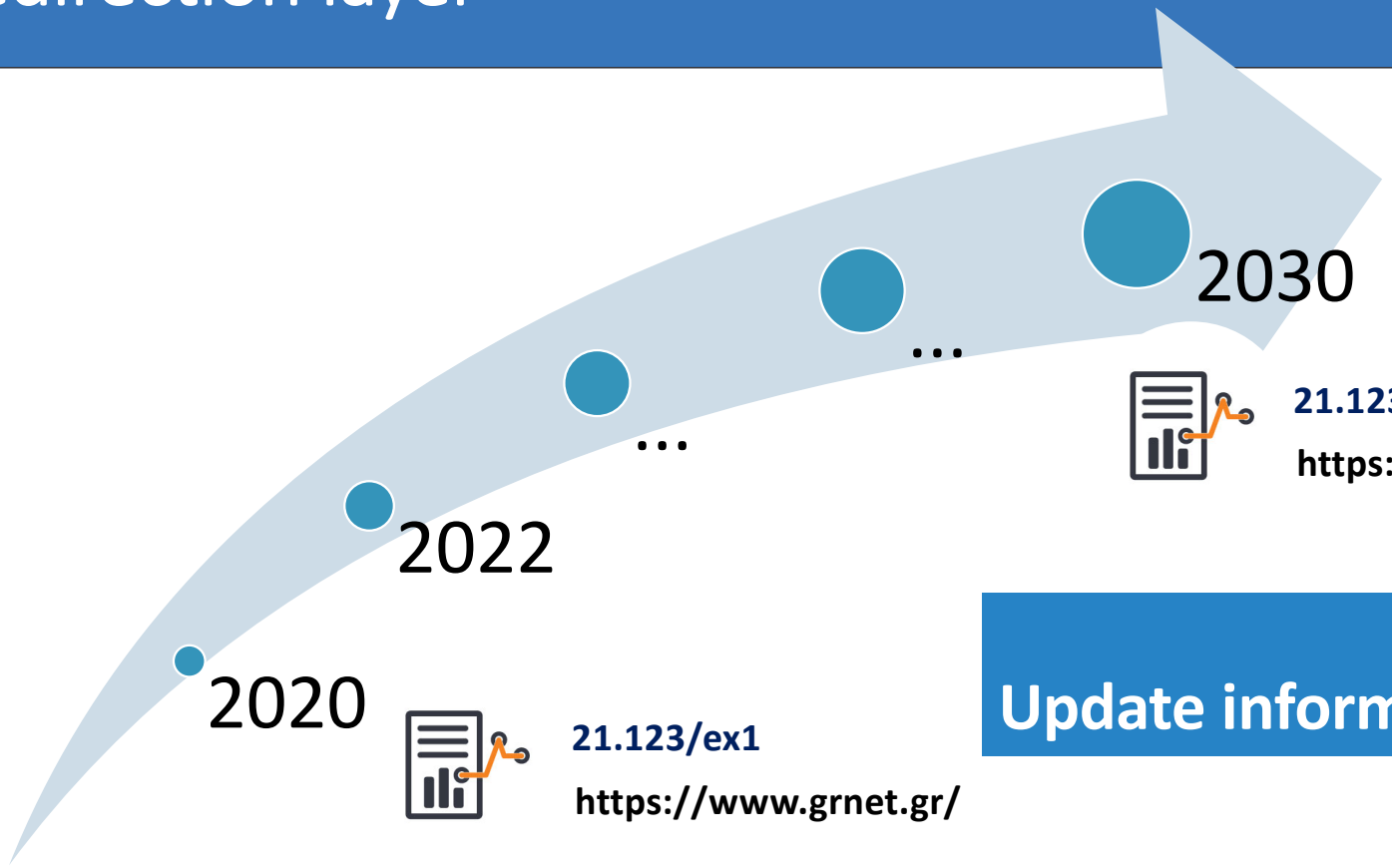
Is globally unique

21.123/ex1

Prefix: designates administrative domain, comes from an issuing instance
Suffix: unique in the realm of the prefix

Once the PID is created, the resource is globally addressable.

Redirection layer



Update information Redirection

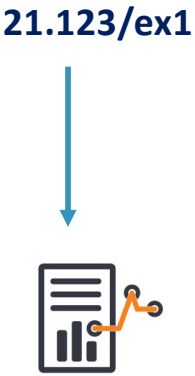
21.123/ex1	https://hdl.handle.net/21.123/ex1	https://www.grnet.gr
String	URL	Digital Object

Why can Persistent Identifiers help?

"A persistent identifier (PID) is a long-lasting reference to a digital object—a single file or set of files."

Not a URL

- Identifier **points to a resource** with no actual knowledge of the resource
- **Responsibility** of the PID owner to keep it up-to-date when the resource changes



PID Costs

- **New Effort is introduced**
  - Do you know what you want to do ?
  - Do you know what you want to reference?
  - What is the granularity your are going to use?
    - Files?
    - Raw data?
  - Coordination across organisations

- **New effort to support persistency**
  - Enforce discipline among your organization
  - Technology necessary but not sufficient

- **Ok based on that**
  - Does it worth it?
  - Your data need PID?

PID systems

Persistent Identifier structure

- Every persistent identifier consists of two parts: its prefix and a unique local name under the prefix known as its suffix
  - Prefix - designates administrative domain, is generated by an issuer, which makes sure that all prefixes are unique
  - Suffix - local name must be unique under its prefix.

- The uniqueness of a prefix and the local name under that prefix ensure that any identifier is globally unique within the context of the System.

< PREFIX > / < SUFFIX >
(e.g. 21.15111/123456745)

Type of Systems

- There are different PID types for different kinds of resources.
  - for people
    - (researchers, authors, contributors, such as ORCIDs, ISNIs)..
  - for objects
    - (publications, data, software, such as URNs, DOIs, ARKs, Handle)

ORCID: Distinguish yourself

- Name Ambiguity Is a Problem
  - Shared names
  - Different versions (full name vs. initials)
  - Transliteration
- ORCID provides a persistent digital identifier (an ORCID iD) for a researcher.
- **ORCID** Through integration in key research workflows such as manuscript and grant submission, ORCID supports automated linkages between the researcher and his professional activities, ensuring that the work is appropriately attributed and discoverable
- **Researcher** can connect his iD with the professional information — affiliations, grants, publications, peer review, and more. He can use the iD to share the information with other systems.

PID Systems

Based on: Handle System

a Persistent URLs (PURLs)

purl: GPO/gpo46189

Cost: no

Metadata: No additional metadata

b EPIC

*hdl:*21.15102/*123*

Cost: annual fee per prefix

Metadata: Associate any metadata

c Archival Resource Key (ARK)

ark: /12025/654xz321

Cost: no

Metadata: ERC (Electronic Resource Citation) metadata

https://n2t.net/ark:/12345/x98765

d Digital Object Identifier (DOI)

DOI: 10.1000/182

Cost: fee per DOI + annual fee

Metadata: The INDECS schema, stored in separate database

The Handle System

The Handle System

- The **Handle System** is a technology specification for **assigning, managing, and resolving** persistent identifiers for digital objects and other resources.
- The protocols specified enable a distributed computer system to **store identifiers** (names, known as Handles) of digital resources and **resolve those Handles** to the information necessary to **locate, access**, and otherwise make use of the **resources**.
- That information **can be changed** as needed to reflect the **current state or location of the identified resource** without changing the Handle.

The Handle System

- The main goal of the Handle system is to contribute to persistence.
- The Handle system is:
  - reliable
  - scalable
  - flexible
  - trusted
  - built on open architecture
  - transparent

A Handle Record

PID – handle: 10232/1234

Actionable PID (URL/resolving): http://hdl.handle.net/10232/1234

Handle	Data Type/KEY	Index	Handle data	Timestamp
10232/1234	URL	1	https://www.eudat.eu/ex1	2014-04-09 12:46:53Z
	DOMAIN	2	EUDAT	2014-04-09 12:46:53Z
	HS_ADMIN	100	eudat/user1	2014-04-09 12:46:53Z

Handle System®

Handle: 11239/GRNET-PIDS

Handle Values for: 11239/GRNET-PIDS

Index	Type	Timestamp	Data
1	URL	Wed Apr 09 201 15:46:53 EES	http://epic.grnet.gr/
2	INST	Thu Mar 20 2014 12:09:42 EET	CLARIN-EL
100	HS_ADMIN	Thu Mar 20 2014 12:09:42 EET	07F30000000A302E4E412F31313233390000012C

Handle System Web Site

ePIC and PIDs

GRNET is a member of ePIC consortium

ePIC

- Support Persistent Identifiers for eResearch
- ePIC is a consortium of European partners that provides an identifier system for the research community.
- ePIC is setup as a highly reliable, persistent and high performance service through a network of strong data centers.

Its focus is the registration of data with persistent identifiers

- in an early state of the scientific process
- where lots of data is generated and has to become referable
- where at least some metadata should be persistently accessible

PID System

but how does the PID system work ?

PID Service

register and manage PIDs for digital objects

API

PID Resolution Service

service to guarantee reliable resolution of the PIDs. Forwarding the user to the resource. Handle to achieve a redundant and load-balanced setup between the data centres.

PID Replication

replicate the database of Handles to guarantee an robust and high-availability PID resolution function

Global Handle Mirror

A mirror of the Global Handle in Europe, so as to assure the resolution of prefixes in Europe.

Ticketing | Support | Monitoring

1st Policy

Initial Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC)

EOSCFAIR Executive Board Working Group

EOSCArchitecture Executive Board Working Group

Initial Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC)

DOI for version 1: https://doi.org/10.5281/zenodo.3574203

This policy was authored by representatives of the EOSC FAIR Working Group and EOSC Architecture Working group. See Appendix 3 for details. This initial policy was released in December 2019 for community feedback and comment.

We welcome responses to and comments on this first version. Please share them with the wider community on https://pidforum.org or email them to us pid-policy@eoscsecretariat.eu. We understand that some areas may require discussion and encourage you to have that discussion on https://pidforum.org. That is where we will also provide details of opportunities for face-to-face feedback and discussion.

We will develop a second version for March 2020 and a final policy will be delivered to the EOSC Governance Board in October 2020.

1. Rationale

1.1. This Persistent Identifier (PID) policy is written for senior decision makers within potential EOSC service and infrastructure providers, and will be of interest to all EOSC stakeholders. It defines a set of expectations about what persistent identifiers will be used in support of a functioning environment of FAIR research. Requirements of providers and the basic services they offer are also outlined. The policy will be approved by EOSC governance, who will also oversee its implementation. The implementation will be guided through recommendations on the PID Technical Architecture which the EOSC Architecture Working Group will provide.

Generic PID Definitions

- Globally Unique
  - a PID name should comply to a syntax that is controlled to avoid clashes, for instance by having namespaces that are governed by a single authority
- Persistent
  - The PID should be managed and governed in such a way that it can be trusted by the community to remain unique and resolvable for the long term
  - The syntax of the PID should also consider persistence, and it is recommended to not include semantics in the identifier string where semantics may change over time
  - The referent should also be stable, whether it is a FAIR Digital Object or digital representation of a physical entity
- Resolvable
  - A PID is resolvable when it allows both human and machine users
  - globally resolvable, the PID needs to be part of a namespace defined by a syntax that is controlled by an Authority

Roles and Responsibilities

- PID Authority (Role). A controller who is responsible for maintaining the rules for defining the integrity of PIDs within a PID Scheme.
- PID Scheme (Component). A set of rules and standards defining the nature of a class of PIDs. This would include a set of lexical formatting rules for PIDs within a namespace.
- PID Service Provider (Role). An organisation which provides PID services in conformance to a PID Scheme, subject to its PID Authority.
- PID Service (Component). A service that creates and processes PIDs and their associated metadata which conforms to a PID Scheme. Service types can include: PID Issuing, PID Resolution, PID Search, PID Metadata, PID Linking, PID Graph, Citation services.

Roles and Responsibilities

- PID Manager (Role). PID Managers have responsibilities to maintain the integrity of the relationship between entities and their PIDs, in conformance to a PID Scheme defined by a PID Authority.

- PID Owner (Role). An actor (an organisation or individual) who has the authority to create a PID, assign PID to an Entity, provide and maintain accurate Kernel Information for the PID

- End User (Role). The end user of PID services.

PID applications

- Applications require secure mechanisms built in PID Infrastructures and some applications require encryption of PIDs to protect activities.
- The owner is responsible for proper management of PIDs and to keep the attribute set up-to-date
- Granularity of PIDs is very much dependent on the communities and it will change over time. Multiple levels of granularity should be supported by the PID ecosystem and linking between levels of granularity should be fostered.
- PIDs should support versioning.
- PIDs should not be re-assigned or deleted.

PID types

- PIDs can identify many different research entities. These can be born digital (e.g. documents, data, software, services - otherwise known as digital objects - and collections made of them), physical (e.g. people, instruments, artefacts, samples), or conceptual (e.g. organisations, projects, vocabularies).
- Physical and conceptual research entities must be represented via a digital representation (e.g. Landing Page, metadata, attribute set, database index) to have a presence in the digital landscape. All digital representations should be FAIR Digital Objects.
- Classes of research entities may need different attribute sets a PID is resolved to. It is the responsibility of a community of practice to define and document these attribute sets (profiles)

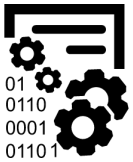
PID for your data

Use of Persistent Identifier

Use persistent identifier for your data

- Managing data online, includes managing the persistent identifier (PID) for the data.
- Synchronize PID, Data during creation, maintenance, update and deletion of your data!
- PID should always
  - be updated to point to the new location (URL).
  - continue to provide the latest information about the resource.

Data Life Cycle



has a life cycle, which involves it going online, accessed by users

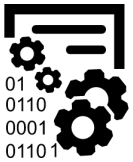
Publish online

Move to another location

used for research

- Published online: http://www.test.com/test.html
- Other users may cite, access, re-use this url
- Relocate the resource at http://www.example.com/
- Other users are not informed -> 404

Data Life Cycle + PID life



has a life cycle, which involves it going online, accessed by users

Publish online

Move to another location

used for research

Pid service	Register PID	Update PID	Get PID
Handle System	Resolve PID	Resolve PID	Resolve PID

Tools you can use

B2handle , Handle api

What is B2HANDLE?

- B2HANDLE is EUDAT's PID service based on
  - Handle as technology
  - EPIC as federation

- B2HANDLE offers:
  - Assignment of prefix via one of the EUDAT partners
  - Hosting of PIDs, i.e. operation and maintenance of Handle servers and technical services
  - Replication and safe-keeping of PIDs via the EPIC federation
  - Resolution mechanism based on Handle
  - Easy maintenance and programmatic resolving of PIDs by the B2HANDLE Python library for general interaction with Handle servers

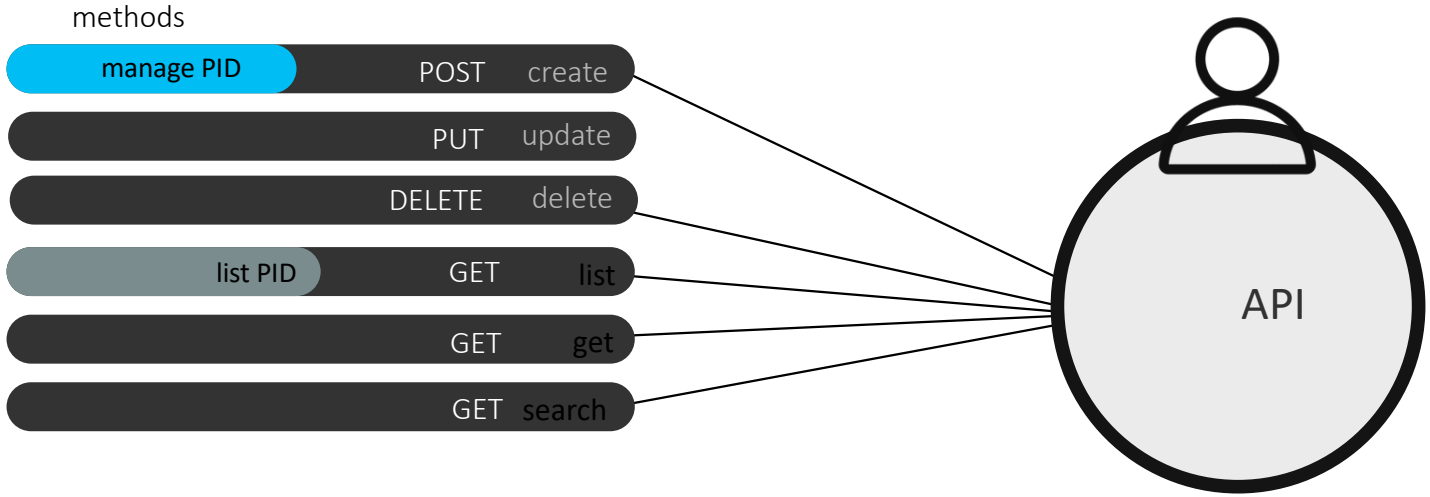
EUDAT B2HANDLE

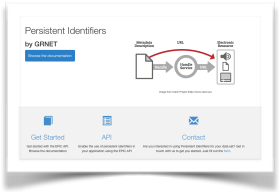
- Features
  - Unified technical interface for access, registration and modification of PIDs and PID records.
  - Operational management of PID namespaces (prefixes) transparent to user communities, including replication.
  - Local reverse-lookups to retrieve Handles based on target location or other fields.
  - Client-side application support through a Python library.
  - Compatible with common standards and practices (e.g. RDA).

Documentation	https://eudat-b2safe.github.io/B2HANDLE/
Github	https://github.com/EUDAT-B2SAFE/B2HANDLE

Handle API

The main interface to register & manage PIDs.

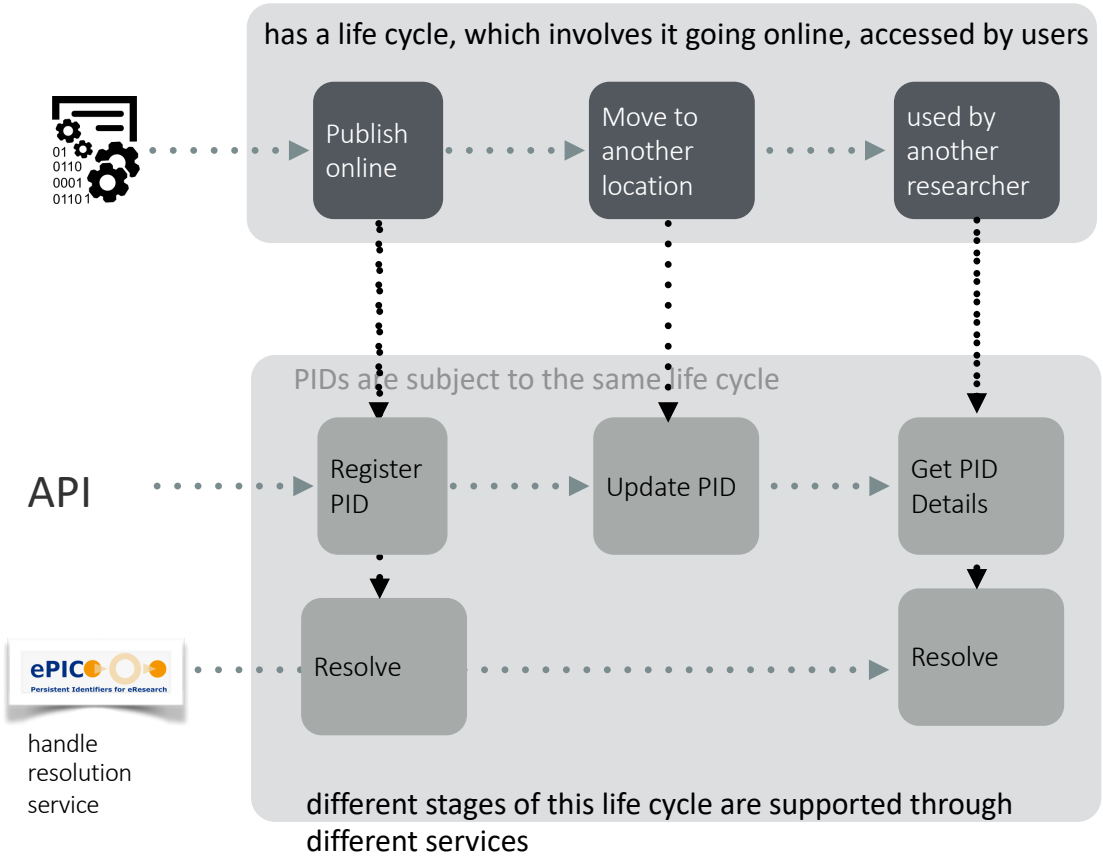




doc.pidconsortium.eu
epic.grnet.gr

A RESTful web service, using the HTTP application protocol.

HANDLE API



resource relocation

Urls - broken links

PID

- points to a resource
- globally unique
- persistent over time

Policy Document

When should I use PID’s

How may I use a PID

When you have a PID use it:

- To cite the data behind the PID:
  - In publications
  - On web-pages
- Include actionable PIDs in linked data
- Retrieve the data:
  - By using the corresponding resolver
  - Via the actionable PID

Citing data sets Ideas

- A user downloads a data set, do a new analysis, publishes the result
  - Mint a PID
  - The PID should resolve to a landing page with information about the data
  - The PID should resolve to the same data.

- A scientist uses and wants to cite two years' worth of data from a 10-year time series.
  - A PID for the whole data set series
  - You may split to chunks based on time. Then the scientist could refer to the time and area so that others can retrieve the same subset.

Accessing cited data sets

- A researcher reads an article, sees a data citation, and wants to use the same data
  - Mint a PID for the data
  - The PID should resolve to a landing page that provides
  - : The PID should resolve to the exact data

- The same researcher wants to use the data from another article. The data set is not available any more.
  - The PID should redirect to a page explaining why the data are not there anymore.

Deleting data sets

- A data set series is deleted
  - The PID should redirect to a page explaining why the data are not there anymore.
- Moving data to another server
  - The PID remains the same. IT should be server agnostic,

Policy Document: When to use PID's

while testing and implementing

There is no one-size fits all strategy for implementing PIDs

The problems may be the same **BUT** each system. community addresses PIDs in its own way (administrative, technical)

Create a Policy Document of What & When

Analyse the use of PIDs, create a policy for the management

What

to register

When

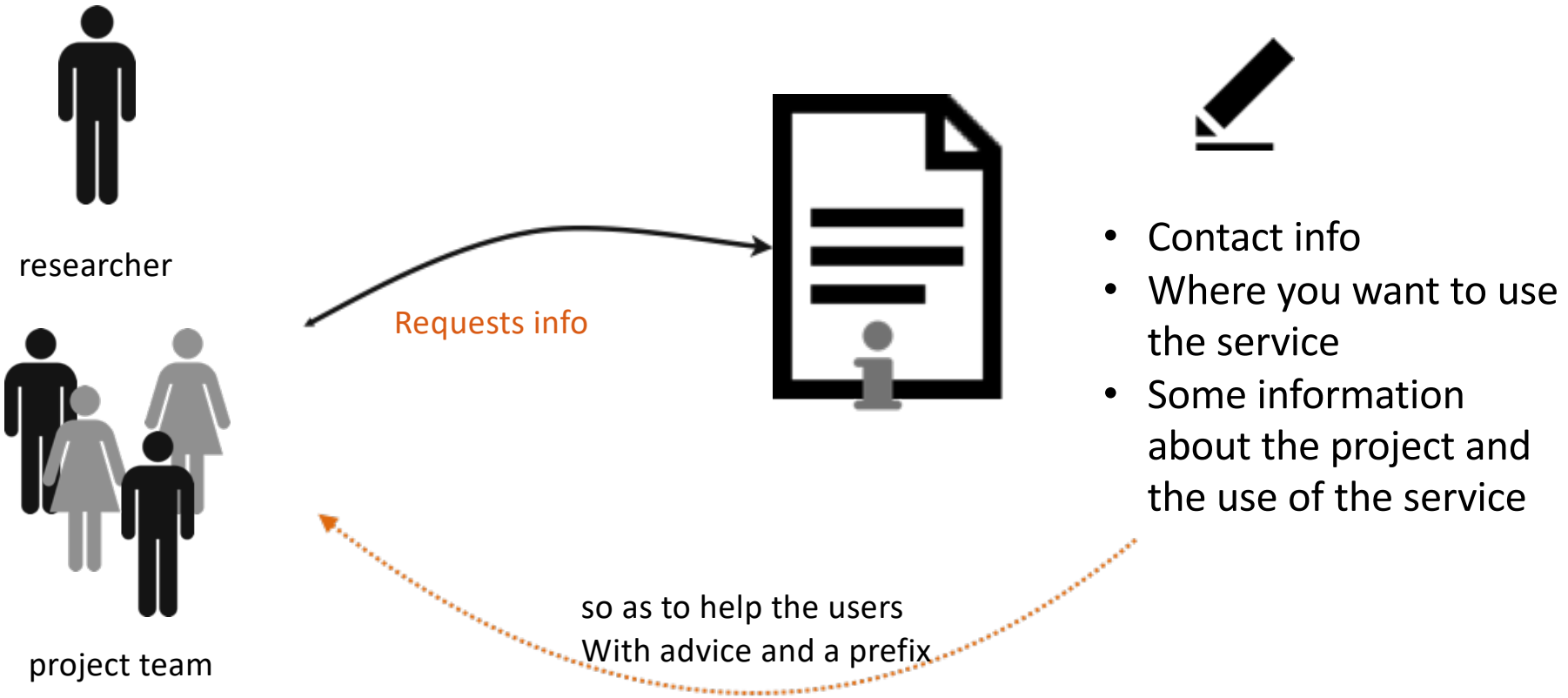
data management life cycle

- Which data objects need a PID (collections, files., metadata records)?
- What kinds of data are likely to stay online long enough?
- What kinds of data are likely to be linked to ?
- What kinds of data are likely to be analysed/processed with tools?
- What will happen after data goes off-line?

A prefix

Is it easy to acquire a prefix?

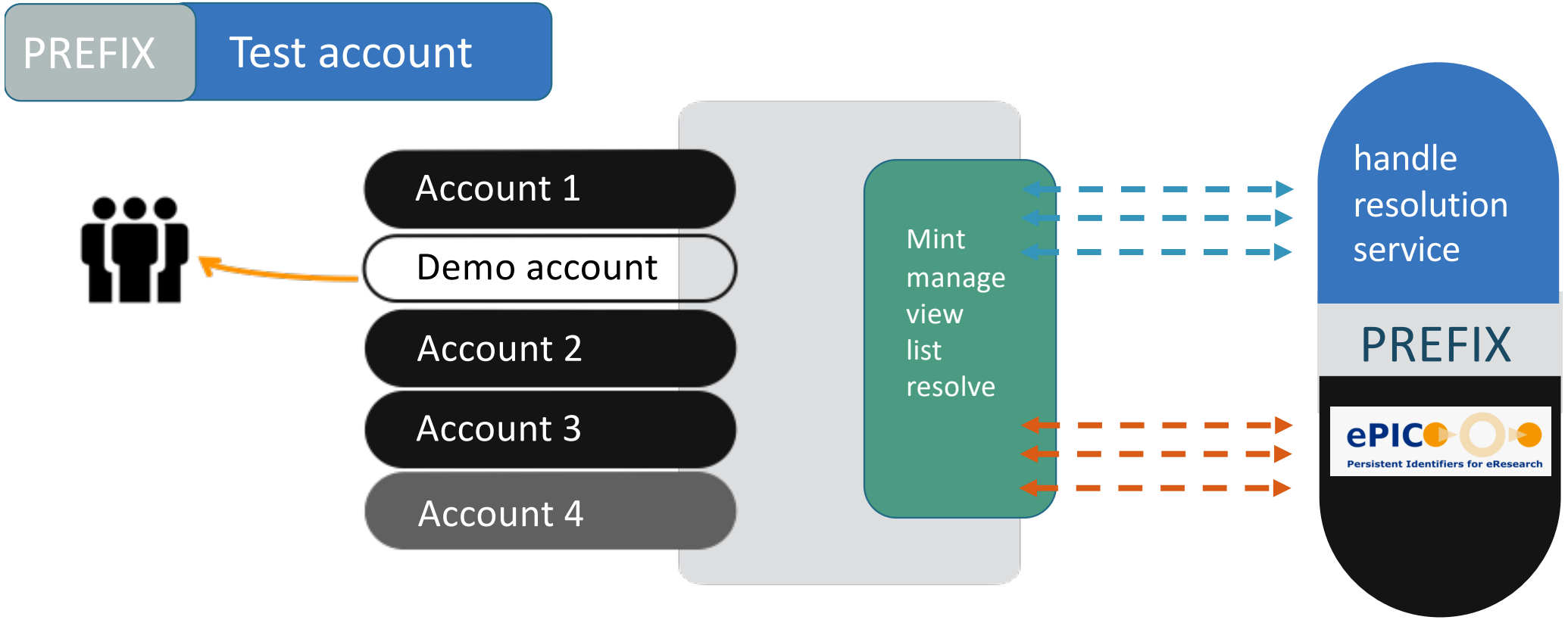
Request a prefix



- Contact info
- Where you want to use the service
- Some information about the project and the use of the service

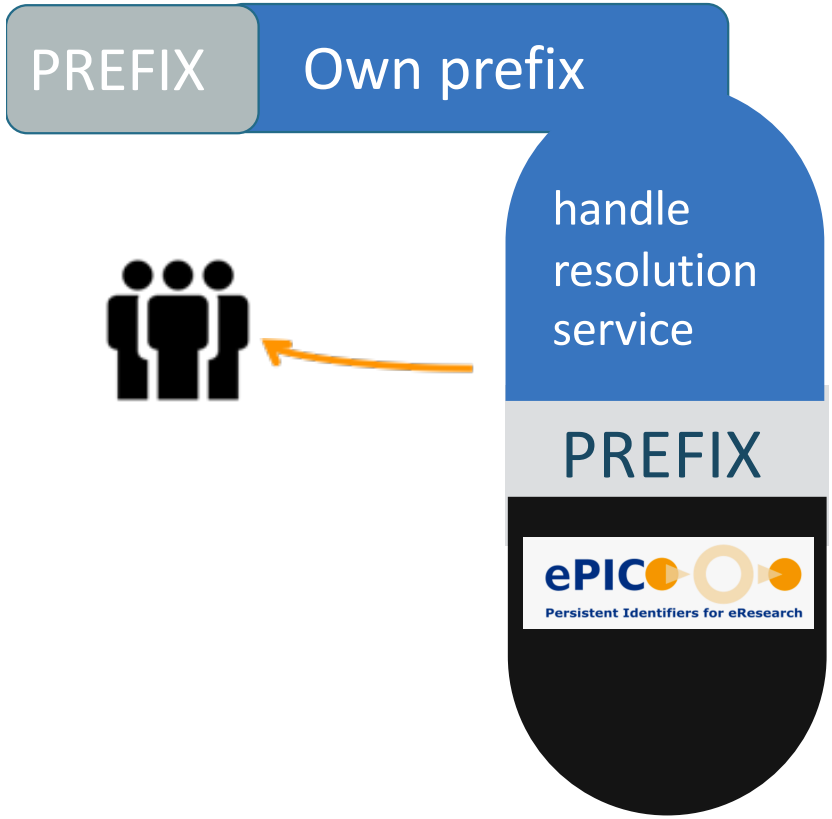
Demo Account

New Demo Account Creation



When you finish

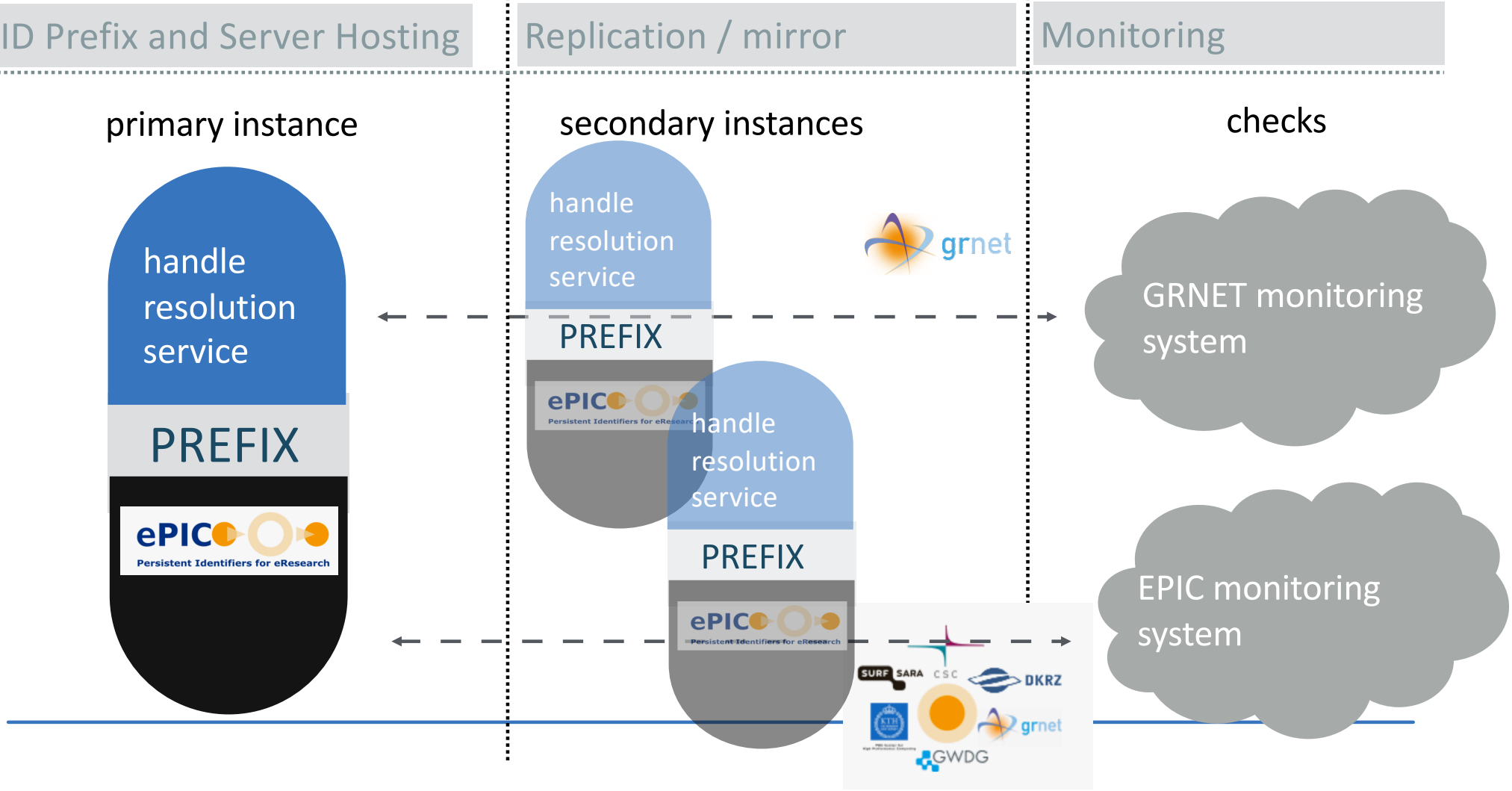
PID Prefix and Server Hosting



GRNET setups a PID Prefix and Server on behalf of a scientific institution or community.

service instance	policies
with its own service instance	with its own policies: • naming • case sensitive (yes,, no) • delete policy • part identifiers support

GRNET PID Service



Uses Cases

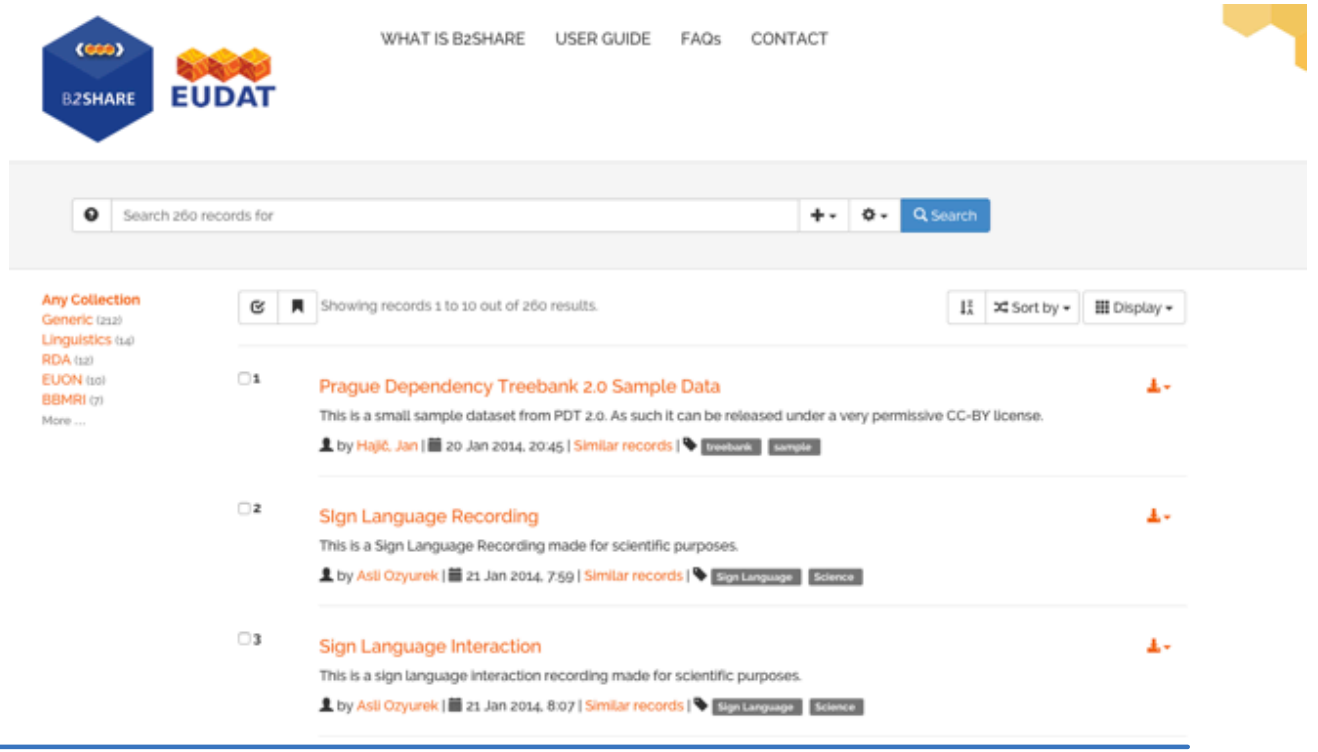
Examples with PIDs

Use Case 1

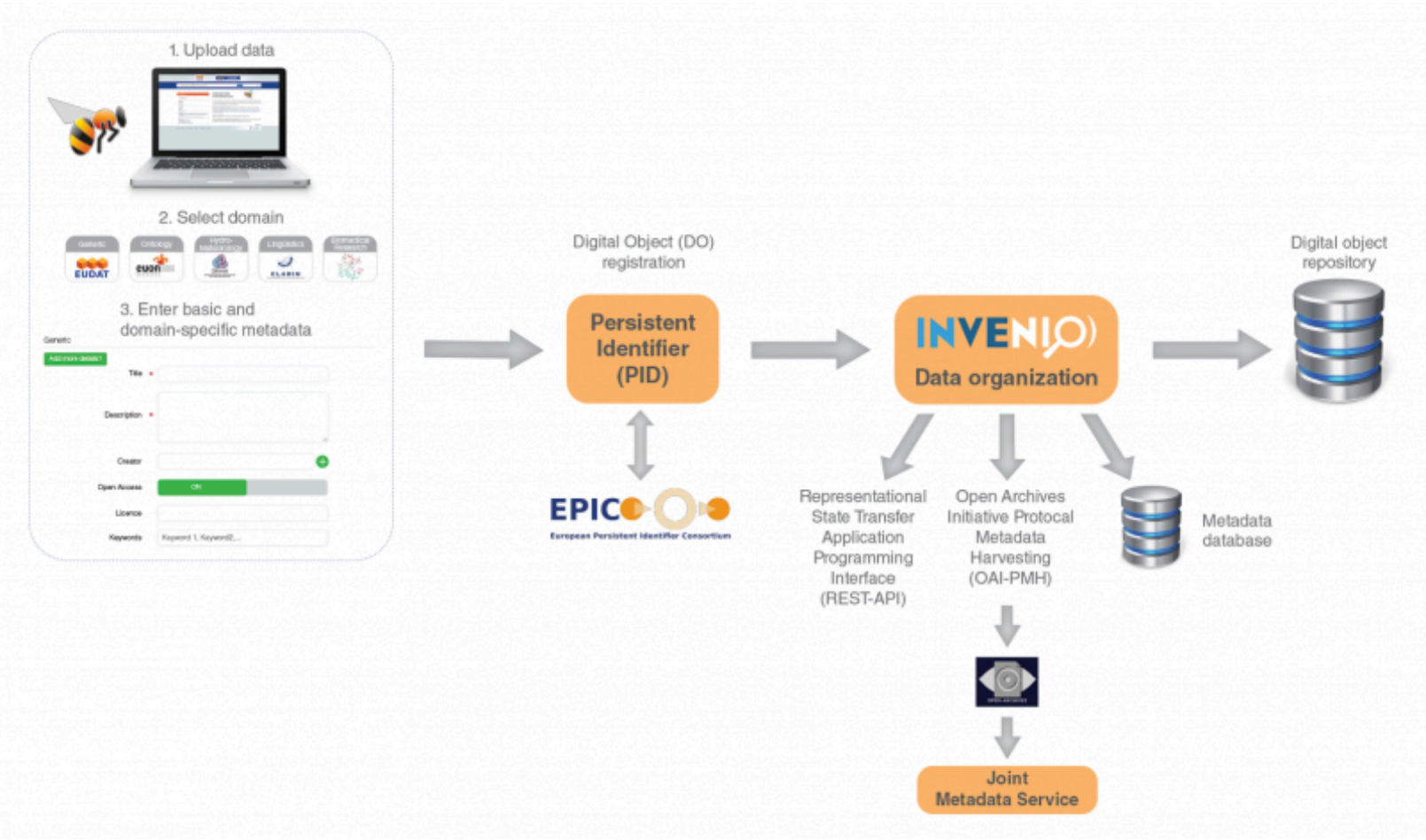
EUDAT B2SHARE

EUDAT B2SHARE

B2SHARE is a user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen scientists to store and share small-scale research data from diverse contexts.



EUDAT B2SHARE



- **Store**: facilitates research data storage
- **Preserve**: guarantees long-term persistence of data
- **Share**: allows data, results or ideas to be shared worldwide

EUDAT B2SHARE

Keywords: parkinson Dopaminergic;

DOI: 10.23728/b2share.ad7fb8cda95f401e80d6b5140bdbb093
PID: 11304/9a5fd2ef-ca7b-43d2-ab18-c7450d0c70d3

Files

Name	Size
11-parkinson eng.docx	24.16KB

Basic metadata

Open Access	True
License	Creative Commons Attribution (CC-BY)
URL	http://creativecommons.org/licenses/by/4.0/
Contact Email	alkoclar@gmail.com
Publication Date	2020-01-04

The persistent identifier for the resource

Handle.Net®

Handle Values for: 11304/9a5fd2ef-ca7b-43d2-ab18-c7450d0c70d3

Index	Type	Timestamp	Data
1	URL	2020-05-01 09:33:44Z	https://b2share.eudat.eu/records/ad7fb8cda95f401e80d6b5140bdbl
2	EUDAT/PROFILE_VERSION	2020-05-01 09:33:44Z	1
3	EUDAT/FIXED_CONTENT	2020-05-01 09:33:44Z	False
100	HS_ADMIN	2020-05-01 09:33:44Z	handle=0.NA/11304; index=200; [create hdl,delete hdl,read val,mo

Handle Proxy Server Documentation
Handle.net Web Site

Please contact hdladmin@cnri.reston.va.us for your handle questions and comments.

Use Case 2

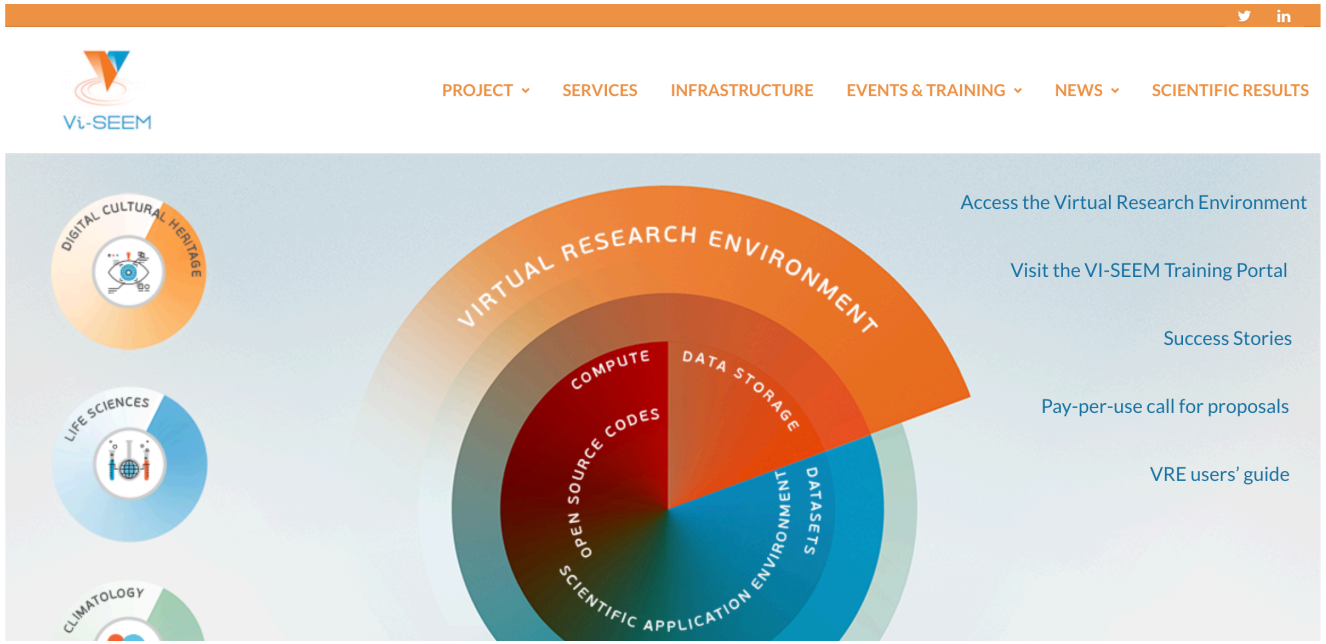
Vi-SEEM

Vi-SEEM

NI4OS Europe

Vi-SEEM Project

Vi-SEEM aims at creating a unique Virtual Research Environment (VRE) in Southeast Europe and the Eastern Mediterranean (SEEM), in order to facilitate regional interdisciplinary collaboration, with special focus on the scientific communities of Life Sciences, Climatology and Digital Cultural Heritage.



VI-SEEM Repository

- A repository to host data Communities
  - VI-SEEM Climate Sciences
  - VI-SEEM Digital Cultural Heritage
  - VI-SEEM Life Sciences
  - VI-SEEM Project Community

4
Communities

255
Submissions

Vi-SEEM: Persistent Identifiers

When should I use a persistent identifier?

Data

Life Cycle

- Create
- Update
- Publish
- Analysed with tools
- Archived
- Delete

Vi-SEEM

Publication status

- Internal
- Ingested
- Published

Vi-SEEM

Recommendations

- Own a prefix
- Associate PID with metadata record
- Use PIDs for files:
  - accessible via internet
  - stable
  - worth to be accessed directly

Vi-SEEM: Persistent Identifiers Solution

Own a prefix

21.15102

Technical Solution

Created our own solution to support Dspace.

Pattern for PID

prefix / VISEEM - UNIQUE ID

21.15102/VISEEM-182

ACIQLife Dataset for city of Sofia

No Thumbnail

Annually averaged output fields of the Air Quality Indices for the territory of city of Sofia

URI
http://hdl.handle.net/21.15102/VISEEM-182

Collections
ACIQLife [2]

View/Open
annual-Sf.nc (711.8Kb)

Date
2016-05-19

Author
Gadzhev, Georgi
Georgieva, Ivelina
Ganev, Kostadin
Syrakov, Dimiter

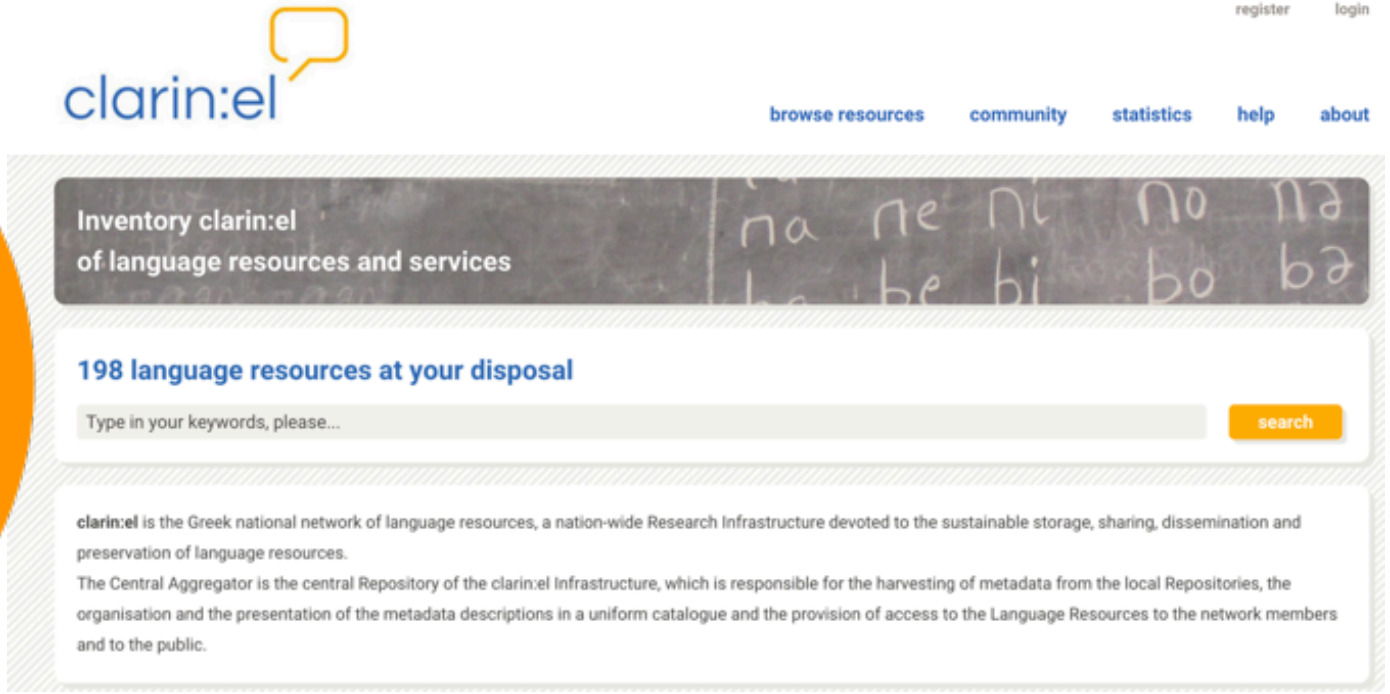
Use Case 3

Clarin:el

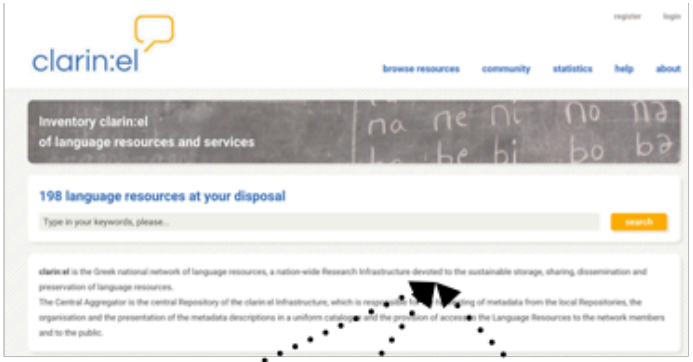
Clarin:el Project

Common Language Resources and Technology Infrastructure

The greek consortium of CLARIN which aims to provide easy and sustainable access to **digital language data** (in written, spoken, video or multimodal form), and advanced tools to discover, explore, exploit, annotate, analyse or combine them, wherever they are located.

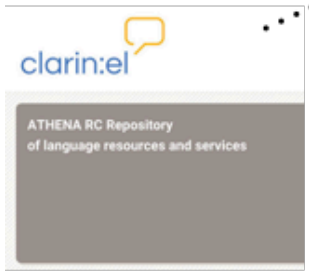


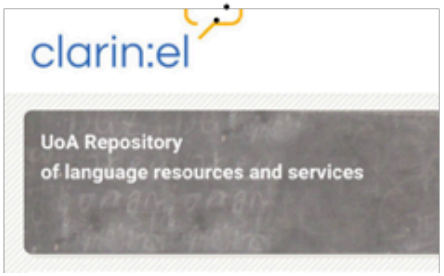
Clarin:el Project

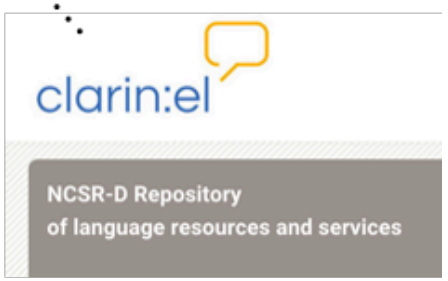


Aggregator

responsible for the harvesting of metadata from the local Repositories







local repositories

Persistent Identifiers

When should I use a persistent identifier?

Data

Life Cycle

- Create
- Update
- Publish
- Analysed with tools
- Archived
- Delete

Clarin:el

Publication status

- Internal
- Group Internal
- Ingested
- Published

Clarin:el

Recommendations

- Own a prefix
- Associate PID with metadata record
- Use PID or part identifiers for non - metadata files
- Use PIDs for files:
  - accessible via internet
  - stable
  - worth to be accessed directly

Clarin:el : Persistent Identifiers Solution

Own a prefix

11500

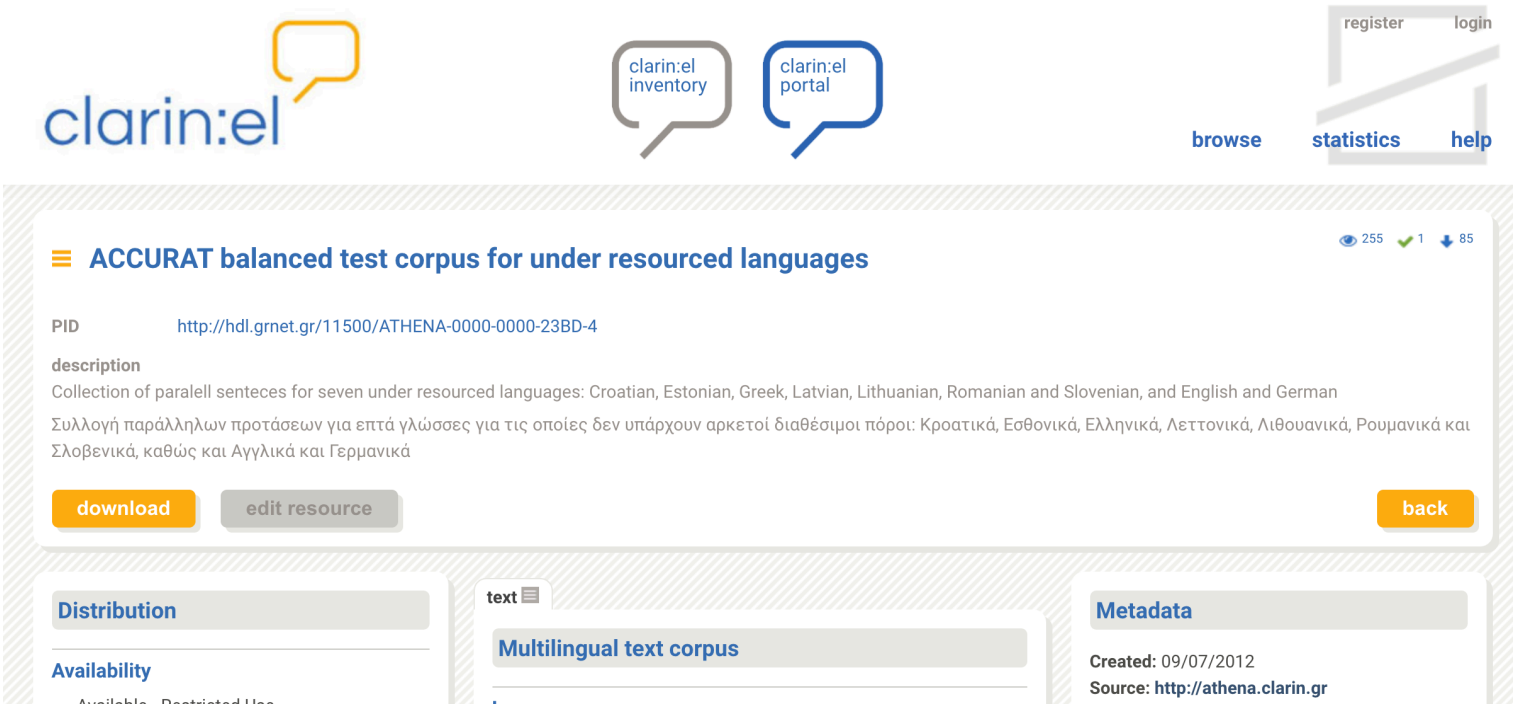
Technical Solution

Created our own python solution by using the epic api.

Pattern for PID

prefix / [REPO NAME] - UNIQUE ID

11500/ATHENA-0000-0000-23BD-4



Thanks!

https://ni4os.eu/

https://twitter.com/NI4OS_eu

https://www.facebook.com/NI4OS/