



A Knowledge Base for Persistent Identifiers

Wim Hugo
DANS



Themis Zamani
GRNET



Co-funded by
the European Union



Project Teams and Contributors

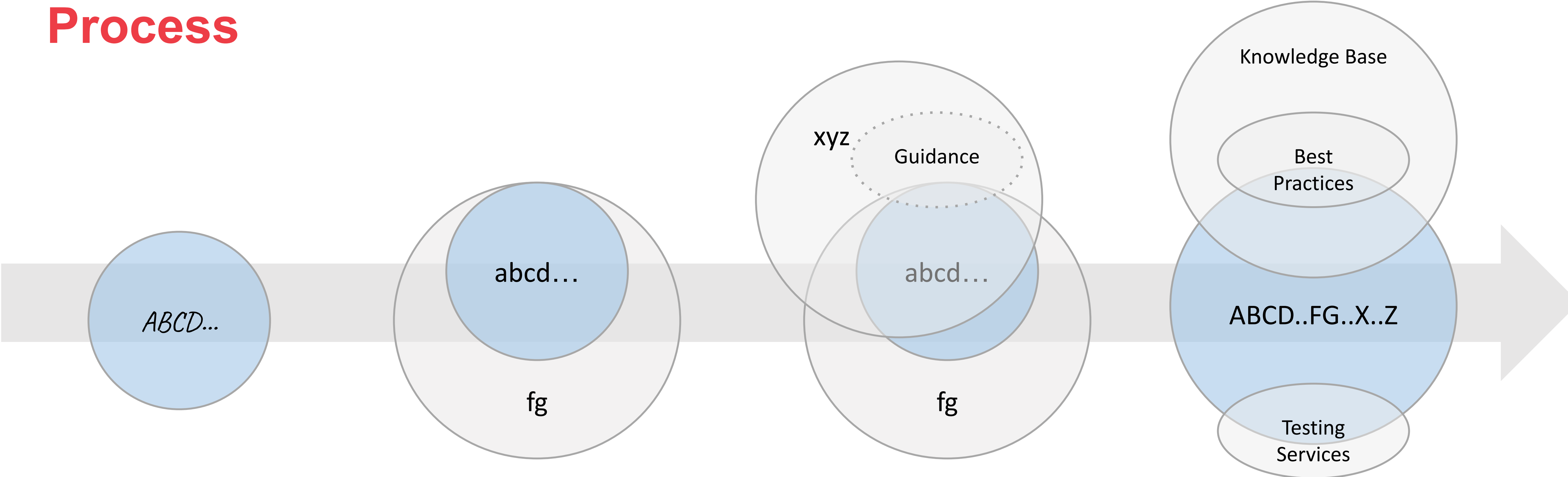
Herbert vd Sompel, Wilko Steinhoff, Fjodor van Rijsselberg, René van Horik, Nina Heinmann, Eoghan O'Neill, Laurens Tobias (DANS), Natascha van Lieshout, Giacomo Cannizzaro (SURF), Matt Buys (DataCite), Chris Ariyo, Harri Hirvonsalo, Joonas Kaseniemi (CSC), Juha Hakkala (Finnish National Library), Kostas Kaggelidis, Chryssa Thermolia, Fotios Basios (GRNET), Tilman Happek (GWDG), Paolo Manghi (OpenAIRE)

Collaboration between two EU-funded projects



Context and Process

Process



European Commission,
Directorate-General for
Research and Innovation,
Hellström, M., Heughebaert,
A., Kotarski, R., et al., **A
Persistent Identifier (PID)
policy for the European Open
Science Cloud (EOSC)**,
Publications Office, 2020,
<https://data.europa.eu/doi/10.2777/926037>

Hugo, W., Steinhoff, W., Turner,
D., Buys, M., & Zamani, T.
(2023). **D2.1 Compliance
Assessment Specification**.
Zenodo.
<https://doi.org/10.5281/zenodo.10067253>



Work in Progress: FAIR-IMPACT
WP3

Community Expectations
Use Cases
Workflows
PID Policies
Best Practices



Work in Progress: FAIR-IMPACT
WP3

**PID Knowledge Base
integrated with CAT**



Example

C1: Minimum Operations Are Available

				Provider	SHOULD
#	Principle or Objective	Suggested Criterion	Description	Metric	Benchmark
C1	Preferred Unambiguous Interoperability	Minimum Operations Are Available	Service providers SHOULD provide a common Application Programming Interface to interact with PIDs, supporting a minimum set of operations (create, resolve and modify PID and PID Kernel Information)	$\sum T1, n$	$=0 \rightarrow 0$ $>1 \rightarrow 1$

#	Test	Description	Type	Method	Guidance
T1,1	CREATE	Create a PID and provide kernel information: API exists and evidence (URL) is available	Binary	Yes = 1 No = 0	G1
T1,2	UPDATE	Update kernel information for existing PID: API exists and evidence (URL) is available	Binary	Yes = 1 No = 0	G1
T1,3	Resolution Service	Resolution API (URL) or URI Pattern exists, evidence is provided	Binary	Yes = 1 No = 0	G1
G1	One may extend the tests to recognise typical and popular standards for API implementation, such as REST, SmartAPI, and the like.				

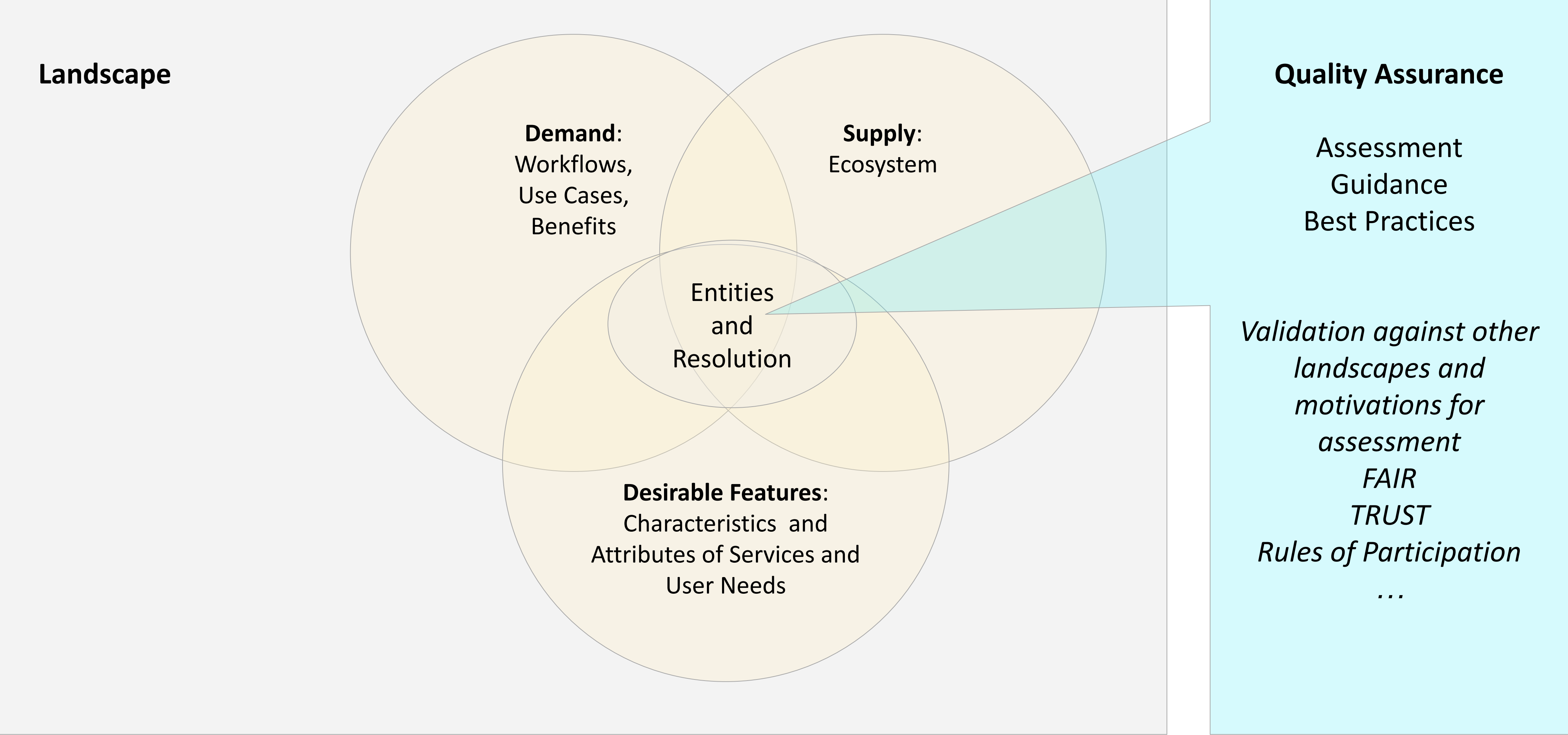
Resolvable

Mutable

PiD Knowledge Base

Necessary to Encode and Characterise the
Supply, Demand, Matching, and Quality
Assurance for the PiD Landscape and its
Stakeholders

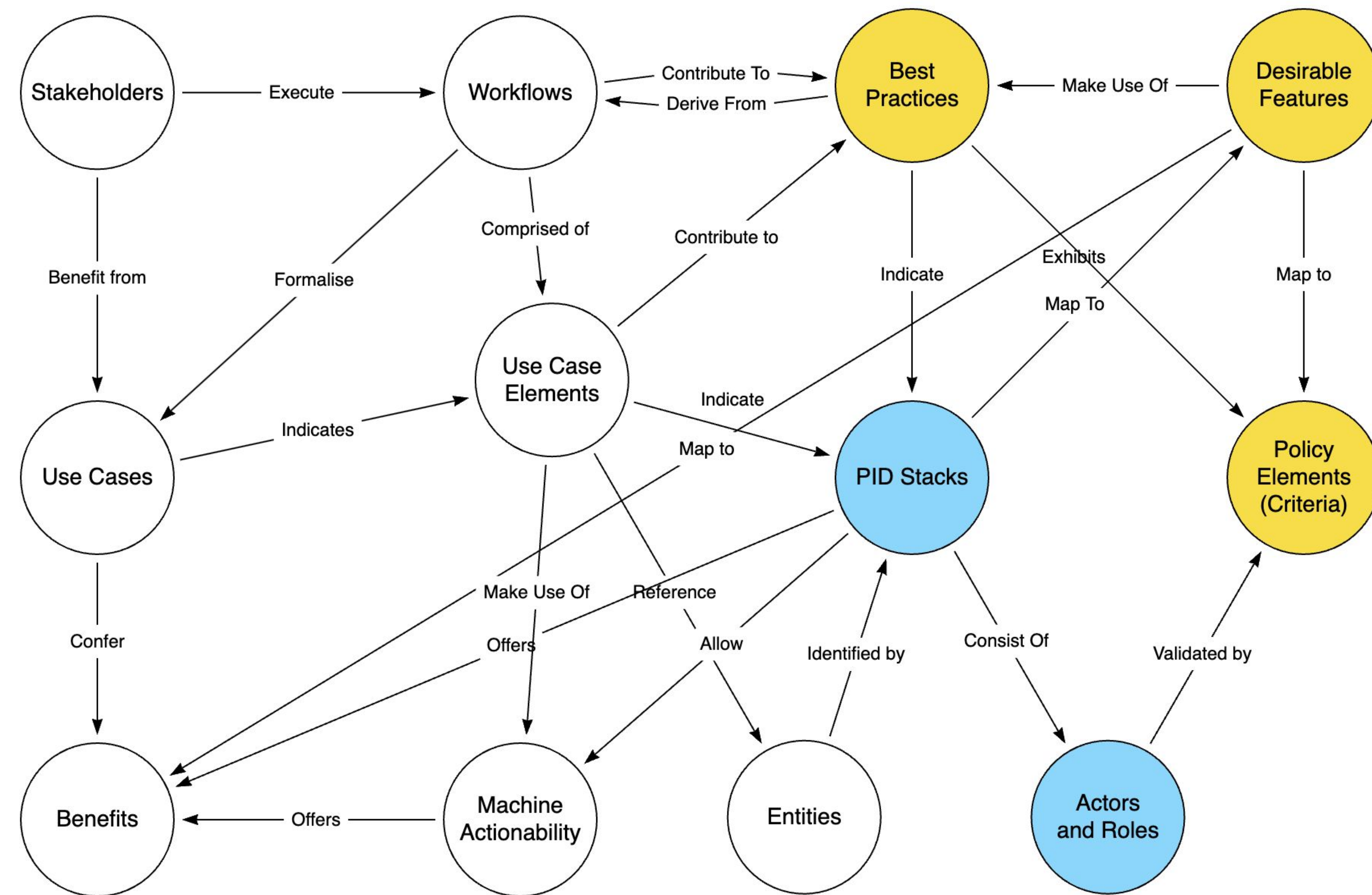
A Portfolio of Interrelated Information Models



Landscape

Stakeholders and their use of PIDs are described in an information model for the PID Landscape. The model defines relationships between workflows and use cases, best practices and desirable features, policies, and the services used to obtain benefits.

These end users interact with 'PID Stacks' (next slide) and sometimes rely on content negotiation, inflection, Signposting and/ or multiple resolution (MR) capabilities of those stacks for automation and machine actionability.

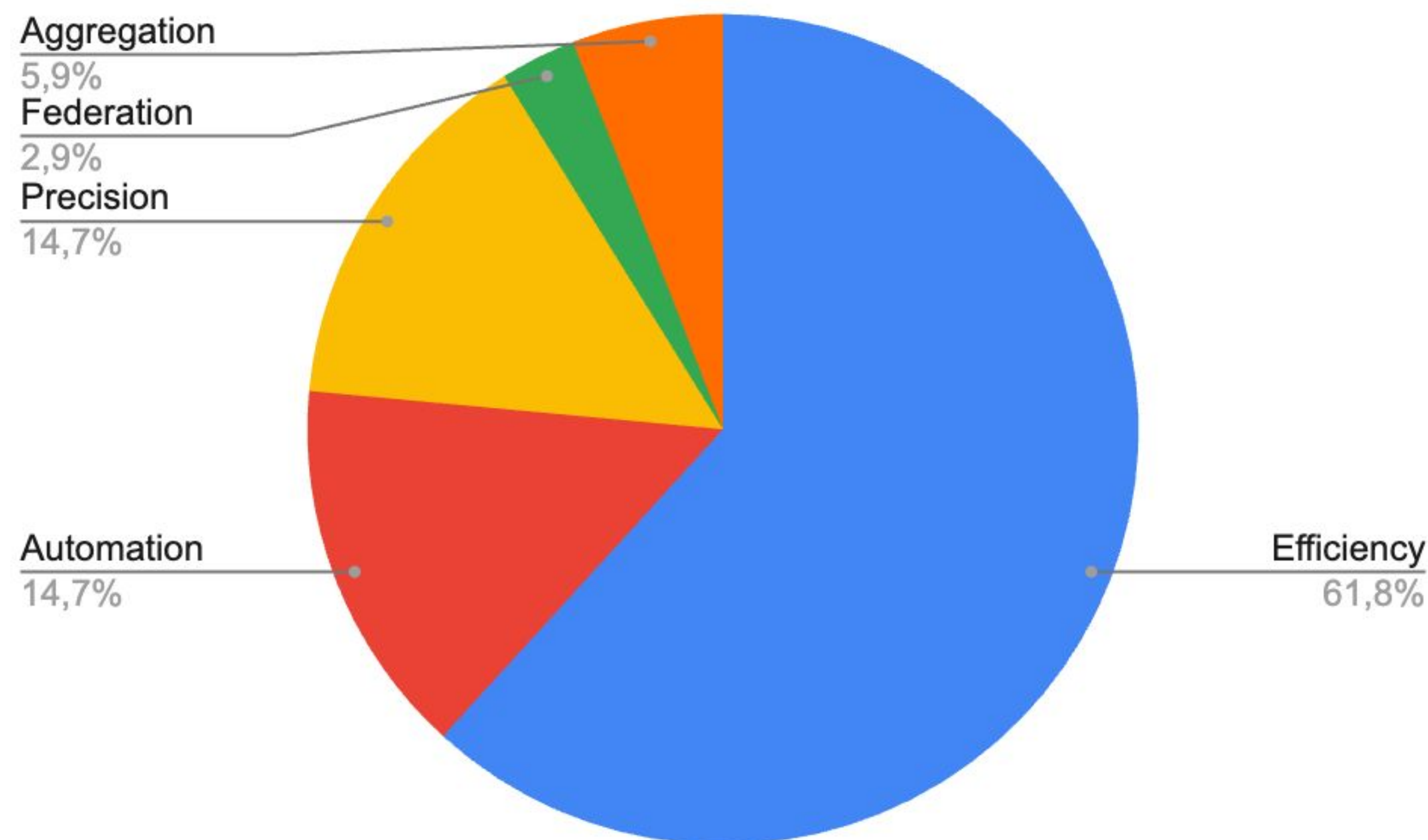


Use Cases and Workflows: Expected Benefits

The main reason for conceiving resolvable digital identifiers that are persistent is an attempt to alleviate the effects of changes in the location or availability of web resources.

This is a pervasive problem even in seemingly well-managed scholarly environments, and has been well documented and to some extent analysed.

In the research environment, the direct consequences of so-called 'link rot' include lack of accurate citation, poor reproducibility, additional work for all involved in the research cycle, and a gradual loss of research outputs that are made available as digital asset. Persistent identifiers alleviate these issues, and also bring additional benefits that are not linked to the primary aim of improving the availability of web resources.



Typical Benefits of PID Use

Generic Use Cases and Use Case Elements

Examples

“Researchers and institutions pass PIDs for previous grants, outputs, organisations, people, and projects to grant application systems.”

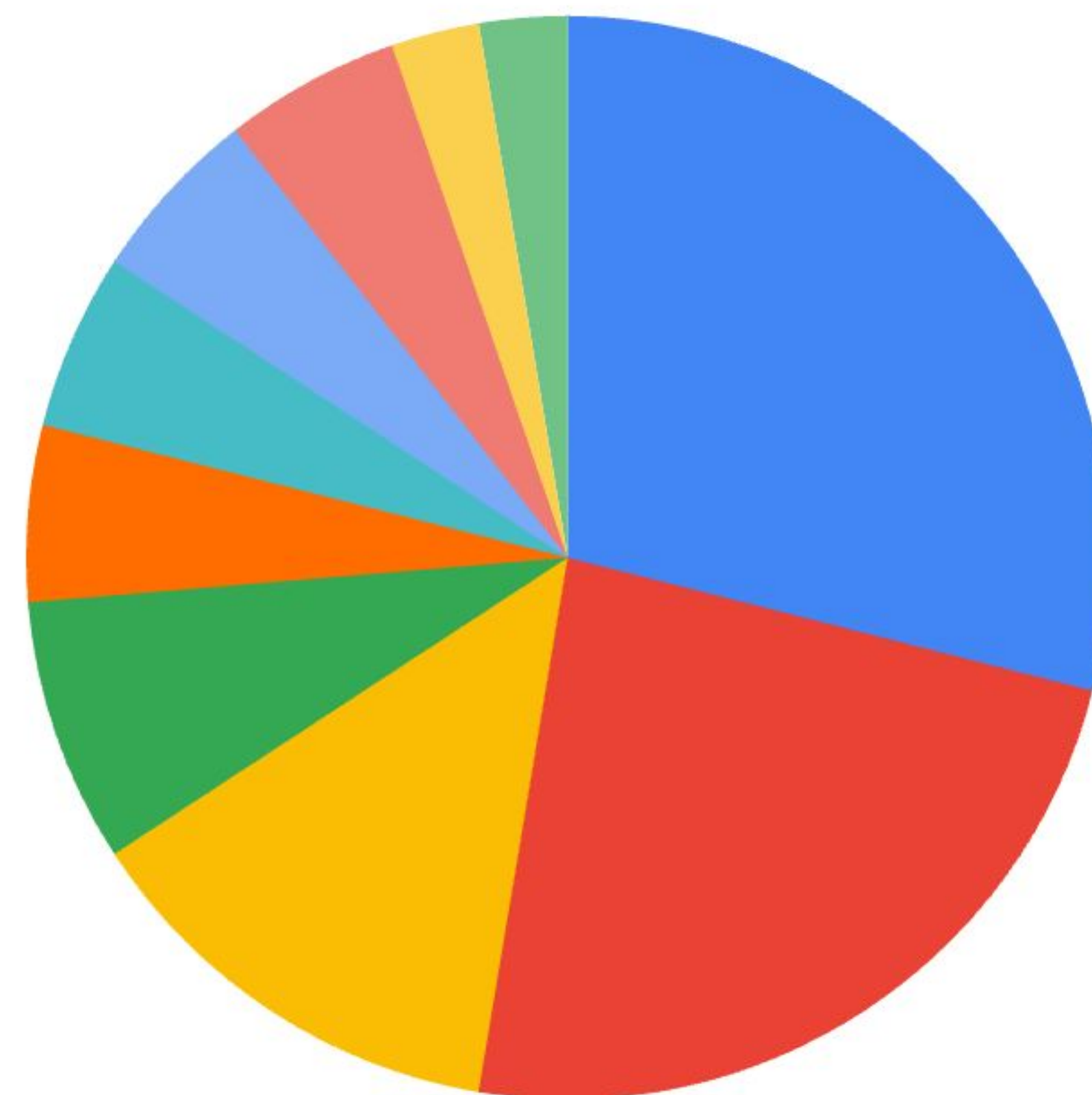
“Researchers share their ORCID iD when submitting new outputs and connect to ROR IDs for institutional affiliations and grant DOIs for funding.”

“Registration of research output is necessary to report to funders [...] for monitoring and evaluation of research (e.g. according to SEP or BKO protocols).”

"They can also much more easily see whether researchers have met mandated obligations for open access publishing and open data sharing."

Top 10 use case elements: literature review

- Identification of research outputs
- Identification of organisations
- Identification of researchers
- Identification of datasets
- Attribution of affiliation
- Attribution of works
- Identification of annotations
- Identification of specimens
- Deduplication of researchers
- Identification of authors



163 publications/ websites
39 use cases identified to date
27 common elements identified
3 detailed use cases (FAIR-IMPACT) to be processed

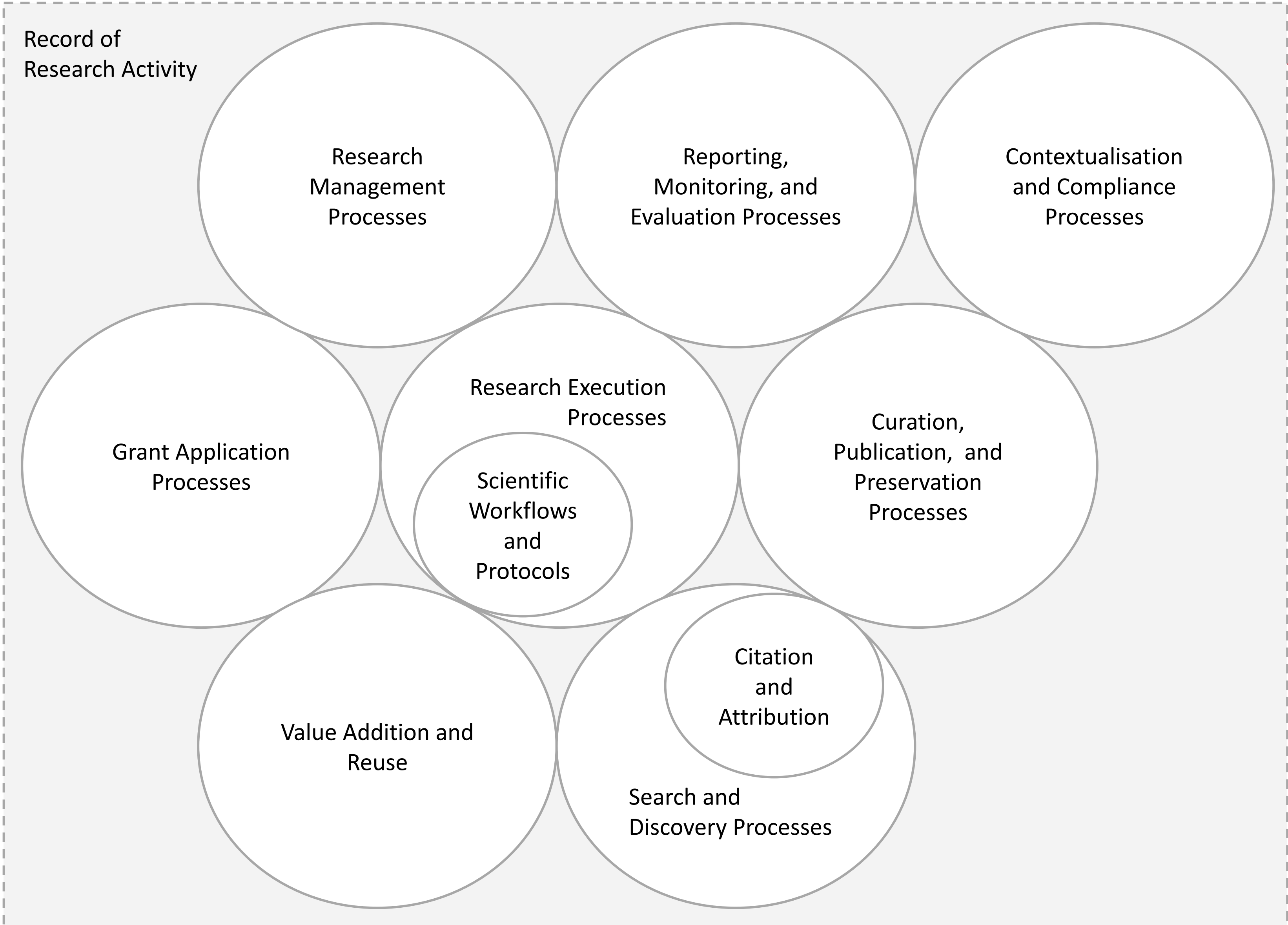
Research Life Cycle Activities: Generic Workflows

Derived from literature

Matching use cases and use case elements to workflow classifications

A workflow can include multiple use cases

*Informal observation:
Some workflows and use cases are less often documented and represented in literature on PID use (contextualisation, compliance, value addition, ...)*



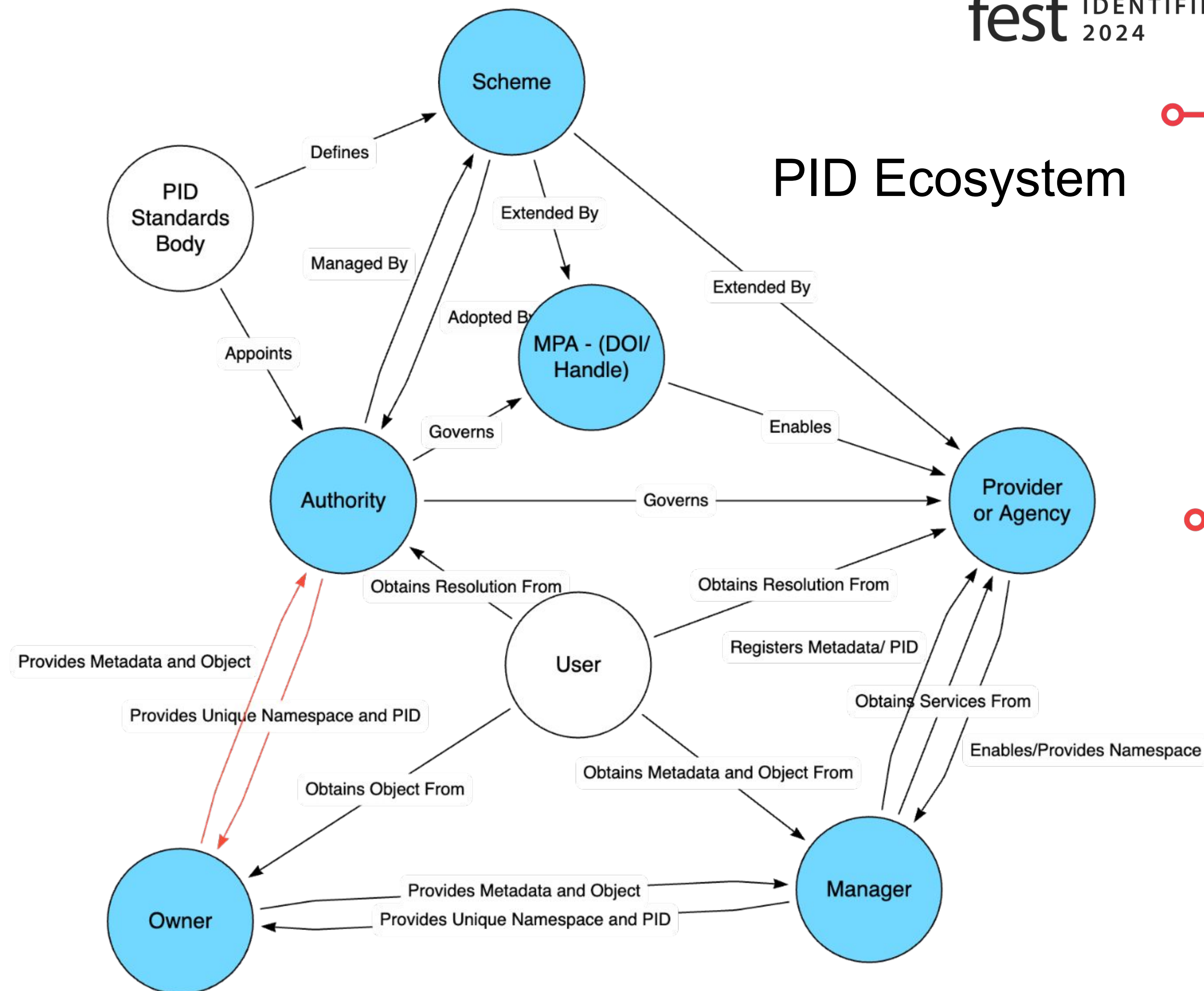
Ecosystem and PID Stacks

The EOSC PID Policy defines a number of Actors (Roles).

These were refined, extended, and generalised via validation against a large number of PID Services.

PID Stacks are composites of Actors that are needed to deliver a Service to Managers, Owners and Users - and stacks often re-use instances of actors at higher levels in the stack.

Note that an **Owner**, in some cases, can obtain PIDs *what appears to be* directly from an Authority (e.g. ARK, SWHID, PURL, ...) - but in reality, the organisation or service assumes more than one role.



PID Stacks

PID Stacks are best described by looking at some examples. The table on the right shows a number of PID Stacks built from the Handle System. It includes popular stacks such as the **DataCite** and **CrossRef** DOIs, as well as **ePIC**.

There are many other stacks in the ecosystem.

Scheme	Authority	MPA	Provider (Registration Agencies)
Handle System	DONA Foundation	Corporation for National Research Initiatives (CNRI)	Not investigated
Handle System	DONA Foundation	Coalition for Handle Services – China	Not investigated
Handle System	DONA Foundation	GDWG/ ePIC	See A.4.6
Handle System	DONA Foundation	CTIC	Not investigated
Handle System	DONA Foundation	MISADI	Not investigated
Handle System	DONA Foundation	Smart Africa Alliance	Not investigated
Handle System	DONA Foundation	Tunisian Internet Alliance	Not investigated
Handle System	DONA Foundation	RosTelecom	Not investigated
Handle System	DONA Foundation	International DOI Foundation	Airiti
Handle System	DONA Foundation	International DOI Foundation	BSI Identify
Handle System	DONA Foundation	International DOI Foundation	Chinese DOI
Handle System	DONA Foundation	International DOI Foundation	CNKI
Handle System	DONA Foundation	International DOI Foundation	CrossRef
Handle System	DONA Foundation	International DOI Foundation	DataCite
Handle System	DONA Foundation	International DOI Foundation	EIDR
Handle System	DONA Foundation	International DOI Foundation	HAND
Handle System	DONA Foundation	International DOI Foundation	JaLC
Handle System	DONA Foundation	International DOI Foundation	KISTI
Handle System	DONA Foundation	International DOI Foundation	mEDRA
Handle System	DONA Foundation	International DOI Foundation	EU-OP

The entities referenced by persistent identifiers can be classified in terms of whether they are objects or an abstract concept agreed upon by humans. Objects can be physical objects with a digital representation or purely digital.. Within this broad classification, we have attempted to standardise the entities encountered in the literature review in a consistent manner.

There are two additional dimensions to the classification that are not currently reflected in the diagram on the next page:

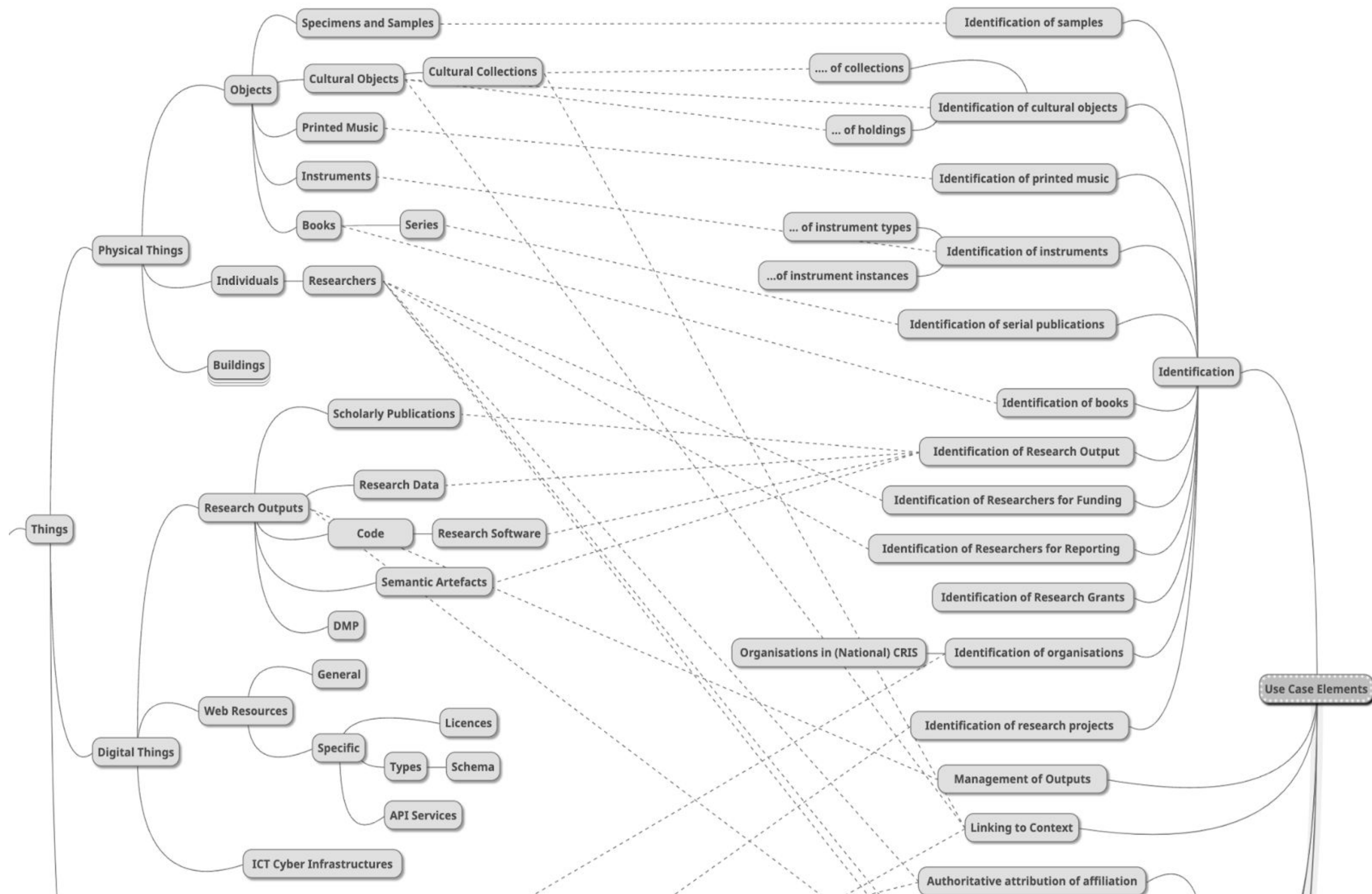
1. There will be specialisation in many cases to accommodate ***domain or industry sector*** differences (for example, the identifiers and associated metadata for instruments in two scientific domains may not be the same).
2. In some cases, a distinction needs to be made between a ***type and an instance of that type***. Scientific instruments again provide a good example - there is an instrument type (make and model) that must be unambiguously identifiable, but there are potentially many instances of the instrument (as reflected by distinct serial numbers). The instances have different or additional characteristics - for example each of them can have one or more calibrations that are unique to an instance, and change over time - leading to ***versions*** of the instance.

Despite this, the classification will be useful as a mechanism for grouping and classifying PID Stacks and could be considered as an additional facet of service of PID Stack classification in the [Compliance Assessment Toolkit](#) being developed by FAIRCORE4EOSC.

Entities

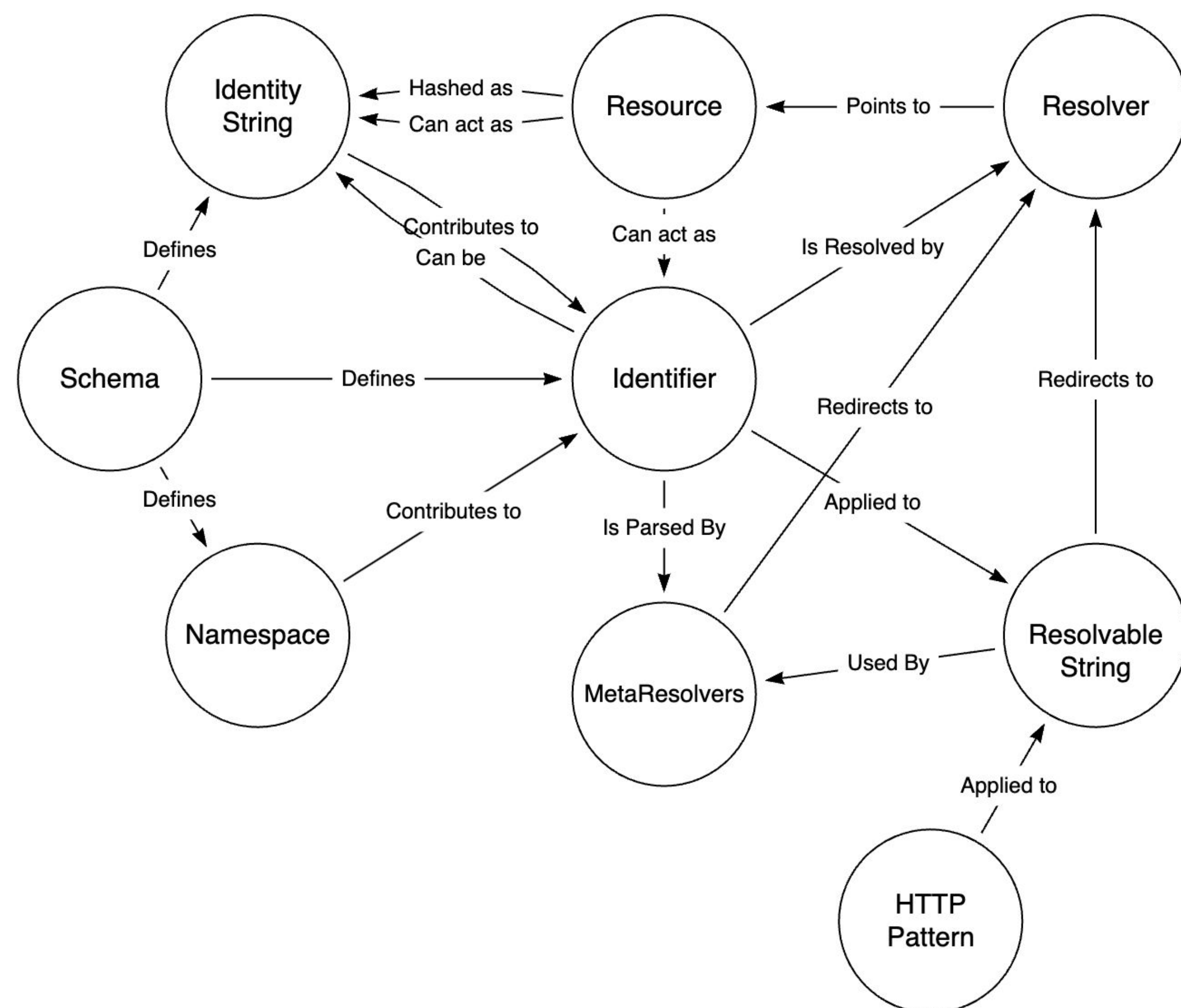
In the process of merging this with the [FABIO Ontology](#) for bibliographic objects, which covers a comprehensive list of the entities in scope for the research library field. Also included - [COAR](#), [BIBFRAME](#), and [DCMI](#).

The Entities relate to two other vocabularies: **Use Cases** (shown) and **PID Stacks**



Entity classification, with mappings to use case elements (excerpt). [Full version](#).

Landscape

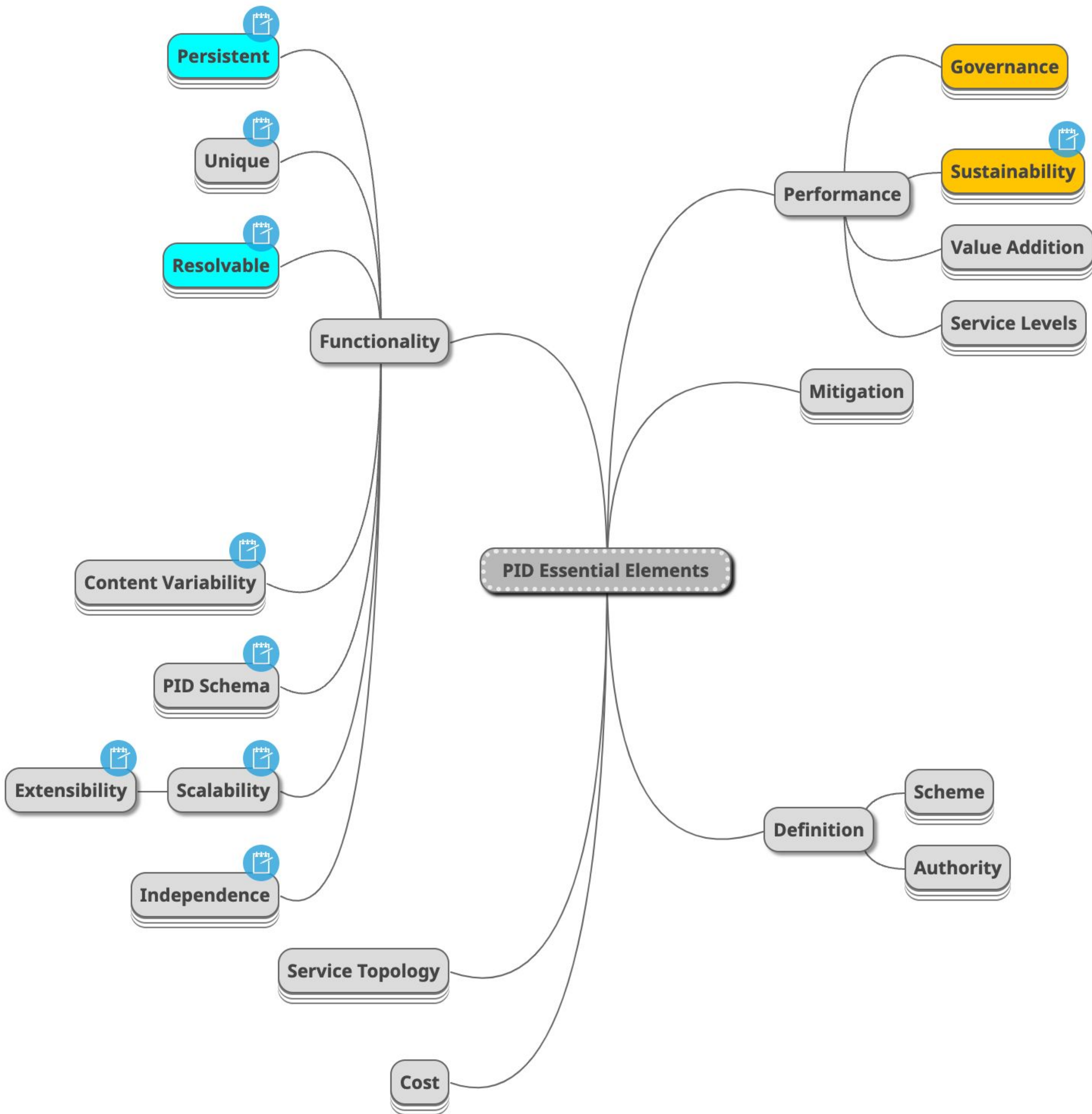


Identifiers can be composed in a number of ways, and depending on the schema in use, namespaces and uniqueness are guaranteed at various levels of granularity - global to local.

identifiers can then be resolved via several mechanisms, ranging from single point centralised resolution to federated or cascading models.

The same resolution target can be referenced by more than one PID, raising interesting questions. There is also an important use case not imagined originally: finding all PIDs

Elements, Characteristics and Attributes of PID Stacks

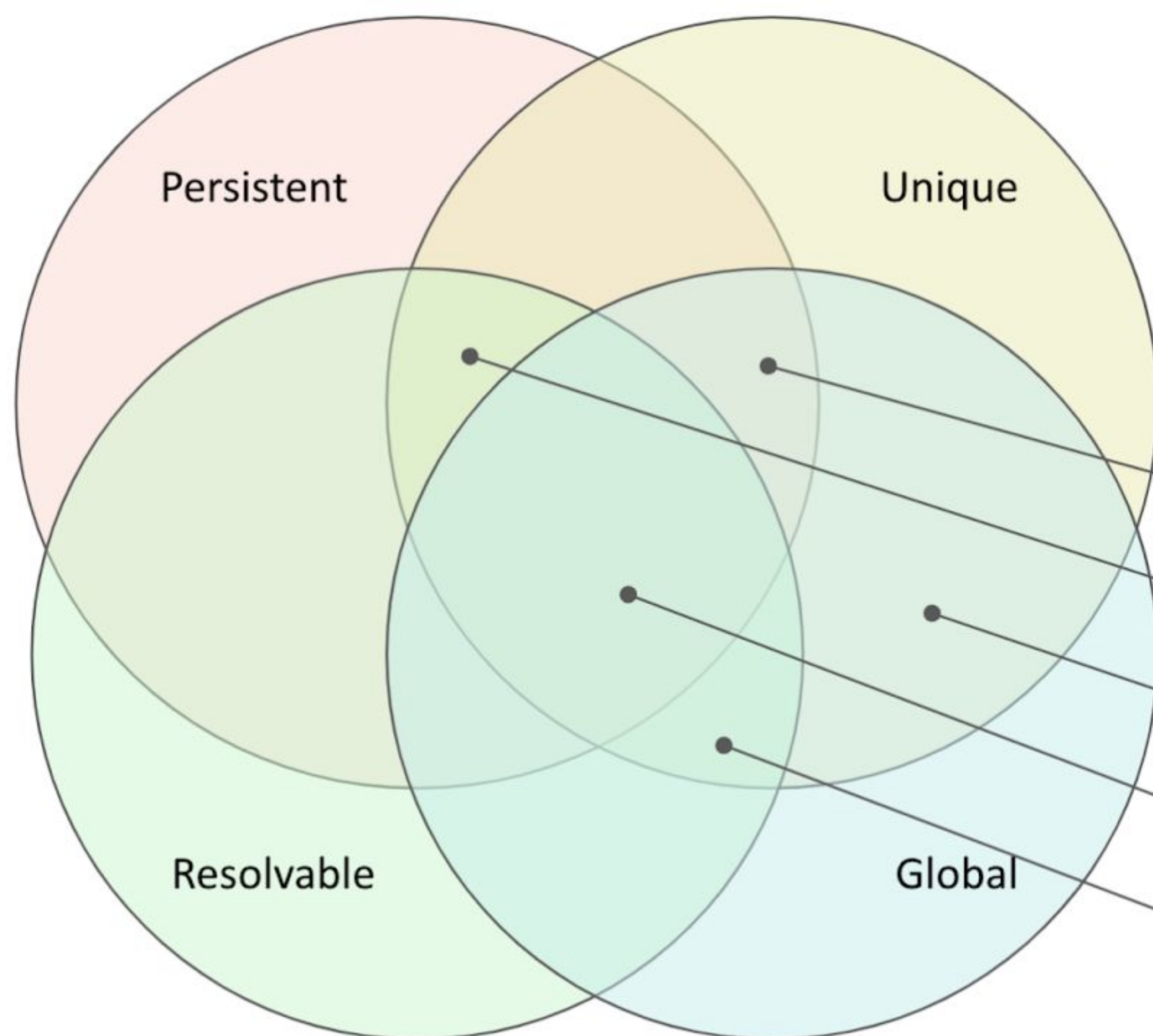


This inventory and classification is in development by the FAIR-IMPACT project. It consolidates the expectations of the community, assertions and features advertised by PID Stacks, and the content of the EOSC PID Policy into a multi-level hierarchy.

[Live version](#)

Note: work in progress!

More About GUPRI



GUPRI: Globally unique, persistent, resolvable identifier
In practice:

- Globally unique: some namespace within DNS
- Persistent: requires effort and management
- Resolvable: HTTP
- Identifier: unambiguous within a namespace - implies an authority of some kind

URN:DEV, ...

URN:NBN, ...

LSID, PURL, ...

ORCID, ARK, DOI, IGSN, ...

Web URLs: DBPedia, GitHub, ...

Monitoring Persistence and Resolvability Over Time

- Complex and Ambitious Goal
- Three interrelated tests:
 - **Persistence: half-life or median time**
 - Not simple to measure accurately
 - **Resolvability: percentage resolution** across PID Stack Elements
 - Mode of access determines HTTP Status
 - **Content Drift:**
 - Exact equivalence (hash/ checksum)
 - but the same version
 - but not the same version
 - Semantically similar (benchmark needed)

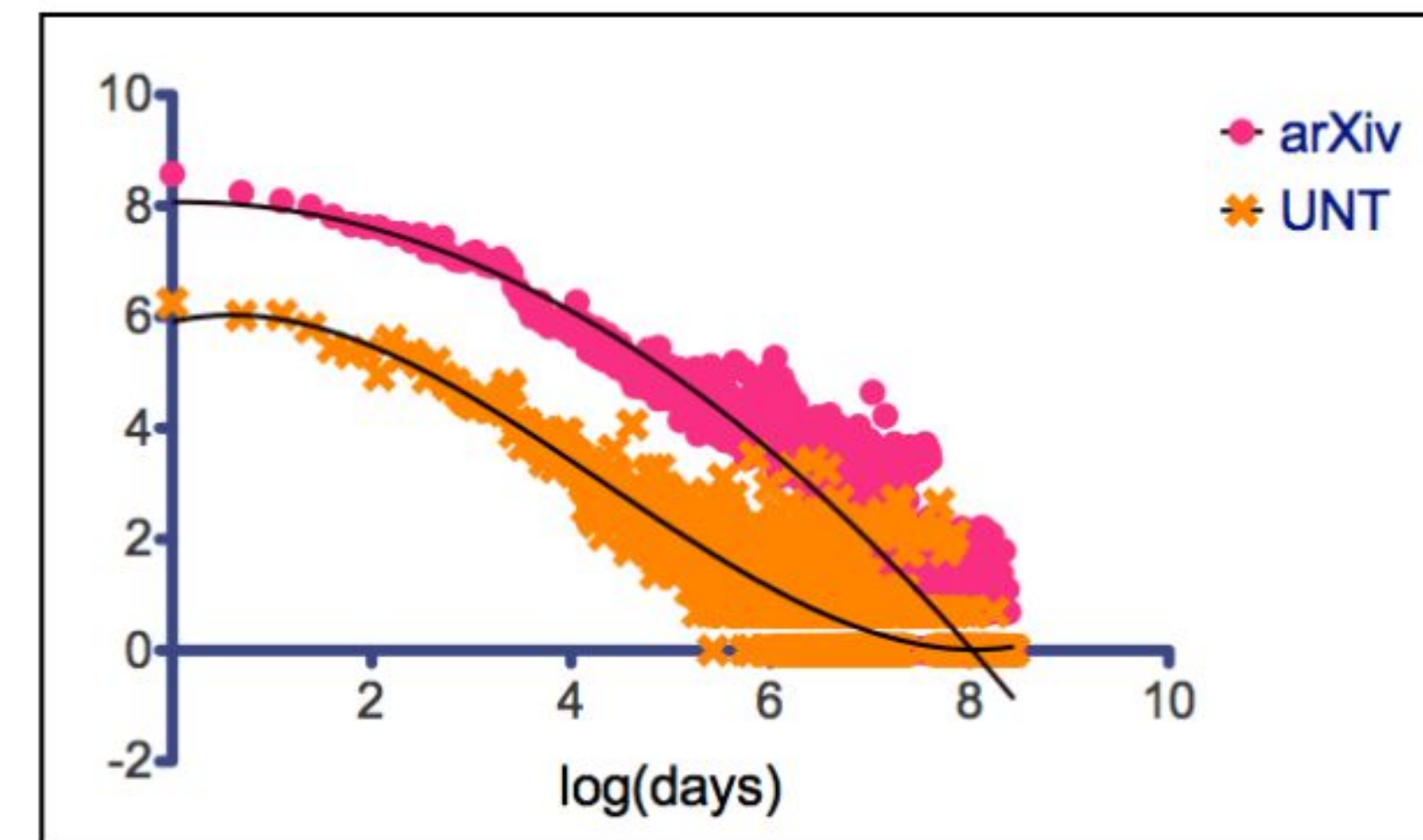


Figure 5. log() of publication/archive date difference vs log() of count of such URLs

Sanderson, R., Phillips, M., van de Sompel, H. Analyzing the Persistence of Referenced Web Resources with Memento, Open Repositories 2011 Conference, 2011

<https://arxiv.org/abs/1105.3459v1>

<https://doi.org/10.48550/arXiv.1105.3459>

Eve, M. (2024) 'Digital Scholarly Journals Are Poorly Preserved: A Study of 7 Million Articles', Journal of Librarianship and Scholarly Communication. 12(1) <https://doi.org/10.31274/jlsc.16288>

Sources of Best Practices

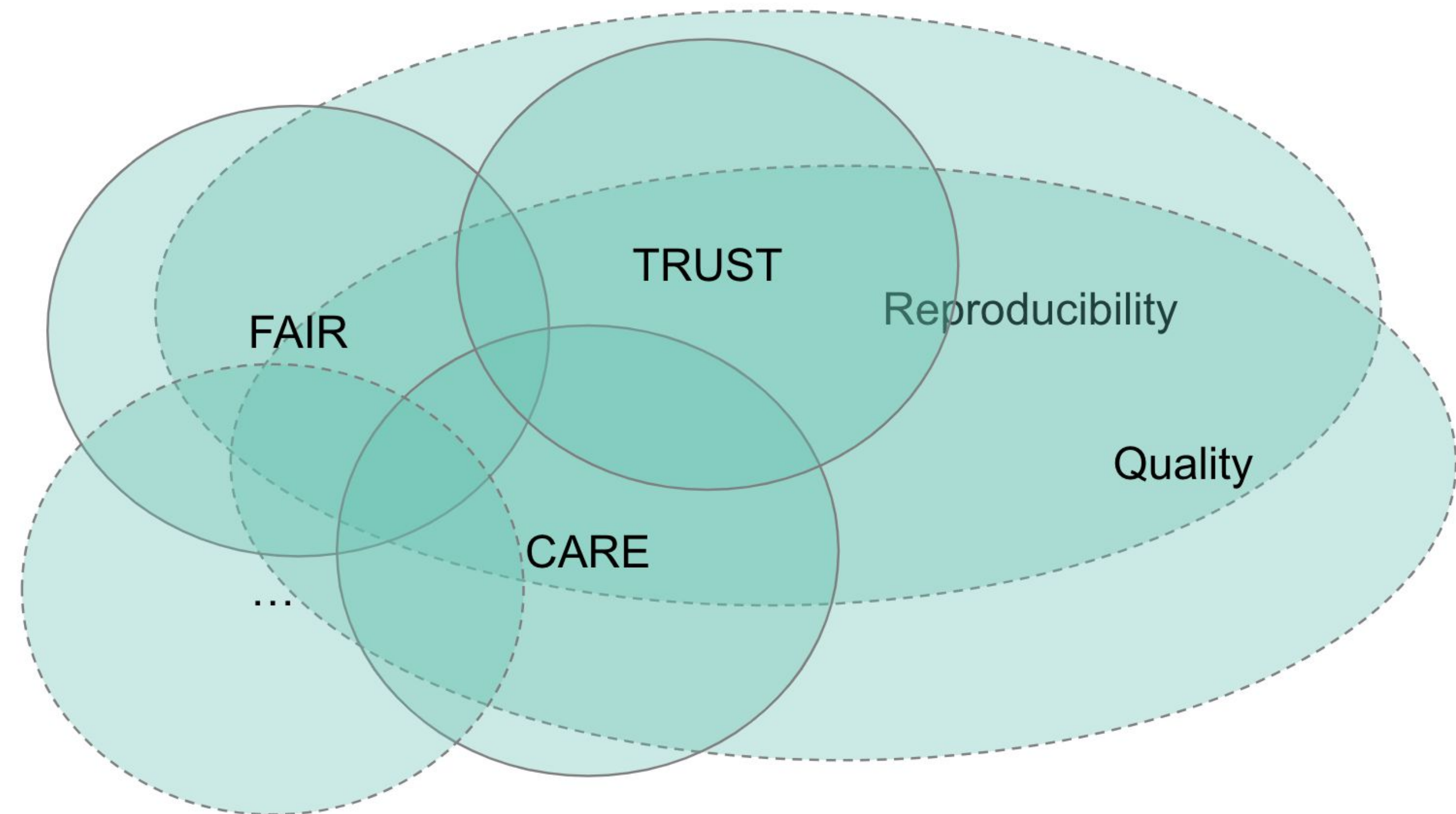
- EOSC PID Policy
- FAIR-IMPACT
 - Review of Data and PID Policies (National, Institutional)
 - RDA Outputs and recommendations
 - Review of PID Stack documentation (recommended practices from providers, ...)
 - Published use of PIDs in workflows and specific use cases



Assessment and Compliance

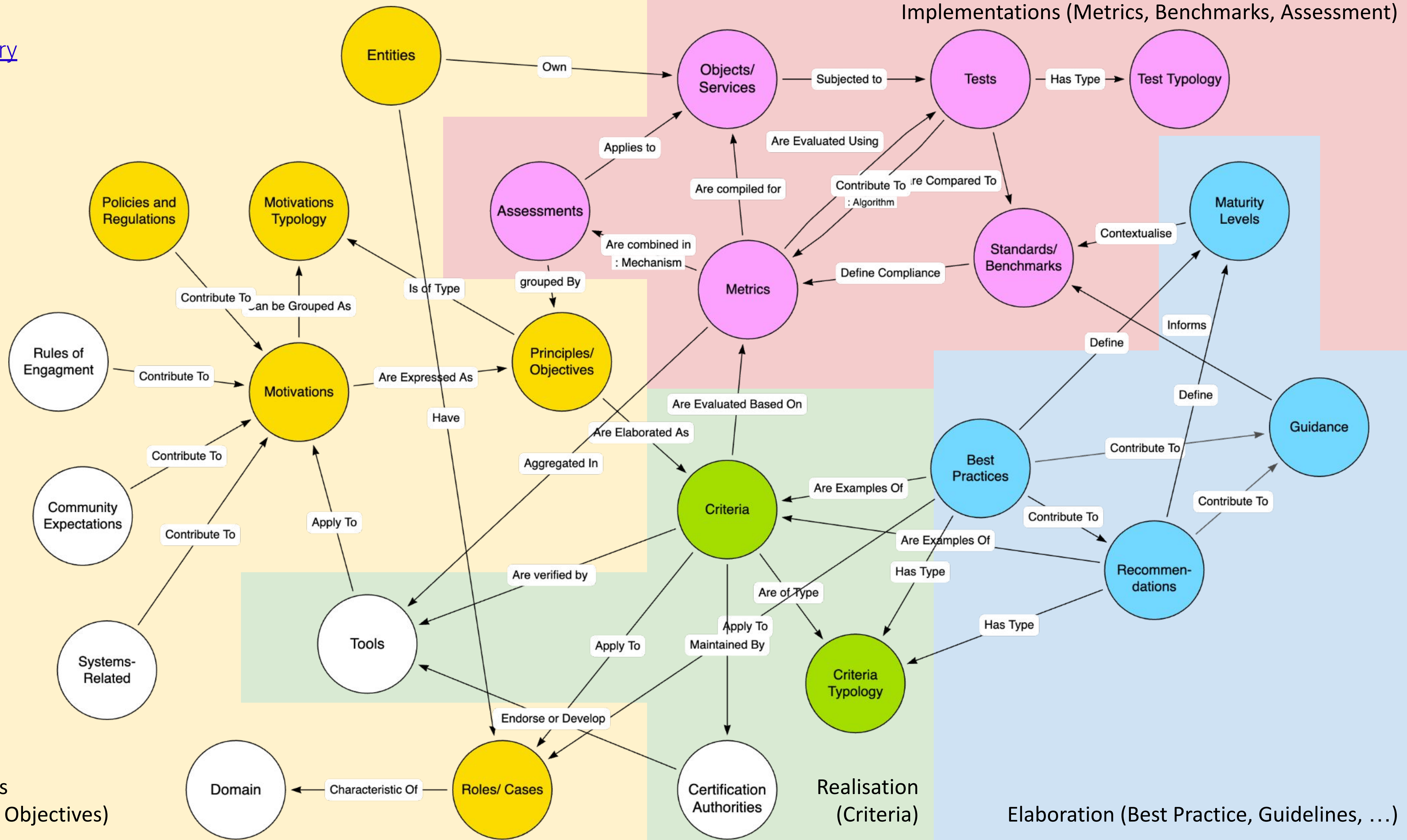
Assessment and compliance against some set of agreed criteria are elements of quality assurance - there are several additional mechanisms for QA that we do not cover in detail here.

For Assessment of PID Service quality, we have started with the EOSC PID Policy, but we have also asked whether there are any additional desirable features expected by the community that could form part of an extended PID Quality Assessment.



FAIRCORE4EOSC has developed and continues to refine a conceptual model for compliance assessment, and has evaluated a variety of case studies to date to validate the model.

Full Story





CAT
EOSC Compliance
Assessment Toolkit

eOSC | FAIRCORE4EOSC

Enabling a FAIR EOSC ecosystem

PiD
fest
CONFERENCE
ON PERSISTENT
IDENTIFIERS
2024

EOSC PID Policy compliance for all actors in the ecosystem

Compliance Assessment Toolkit

Themis Zamani (GRNET)



CAT
EOSC Compliance
Assessment Toolkit

An easy to use Toolkit to measure compliance (ex. with PID Policy - 1st version) based on the results of Conceptual Model (definitions, relationships)

PiD
fest
CONFERENCE
ON PERSISTENT
IDENTIFIERS
2024

design

based on the landscape of compliance assessment and recent development of a compliance ecosystem for TRUST and FAIR



develop

based on a set of definitions for the concepts involved and the relationships between them

The Value of EOSC PID Policy Compliance

For Authorities and Providers: It is important to back up public claims of EOSC Policy Compliance and of performance/ features with publicly available information

All assessments are initially self-assessments - envisaged that we assist all interested PID Stacks with this. *The EOSC PID Policy envisages a compliance authority.*

For Managers: Optional disclosure of policy compliance for purposes of trust and community acceptance

All assessments are initially self-assessments and ***there is no obligation to publish*** these assessments for the foreseeable future - publish if it is useful.

For Owners and Users: Matching PID services with use cases and expected benefits

Find PID services and assess their characteristics, features, and how it matches use cases - whether they are policy compliant or not.

Assessments

 **FAIRCORE4EOSC**
Core Components Supporting a FAIR EOSC

Compliance Assessment Toolkit

Login

PROFILE

VALIDATIONS

ASSESSMENTS

 assessments

+ Create New

≡ View Your Assessments

Read about different actors in the ecosystem before starting.



PID Authority

⋮

View public assessments





PID Service Provider

⋮

View public assessments





PID Manager

⋮

View public assessments



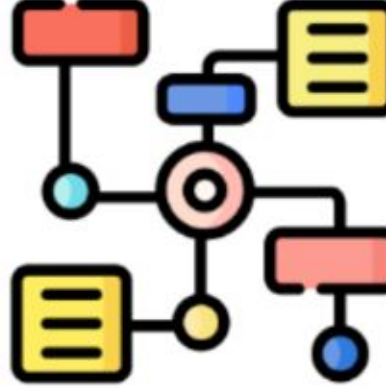


PID Owner

⋮

View public assessments





PID Scheme

⋮

View public assessments



Contact and Service

Contact

Terms and Conditions

Data Licenses

Privacy

About CAT

Our Team

GitHub

Disclaimer

Follow Us

 Newsletter

 Youtube

 Twitter

Partners

 Data Archiving and Networked Services





 Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

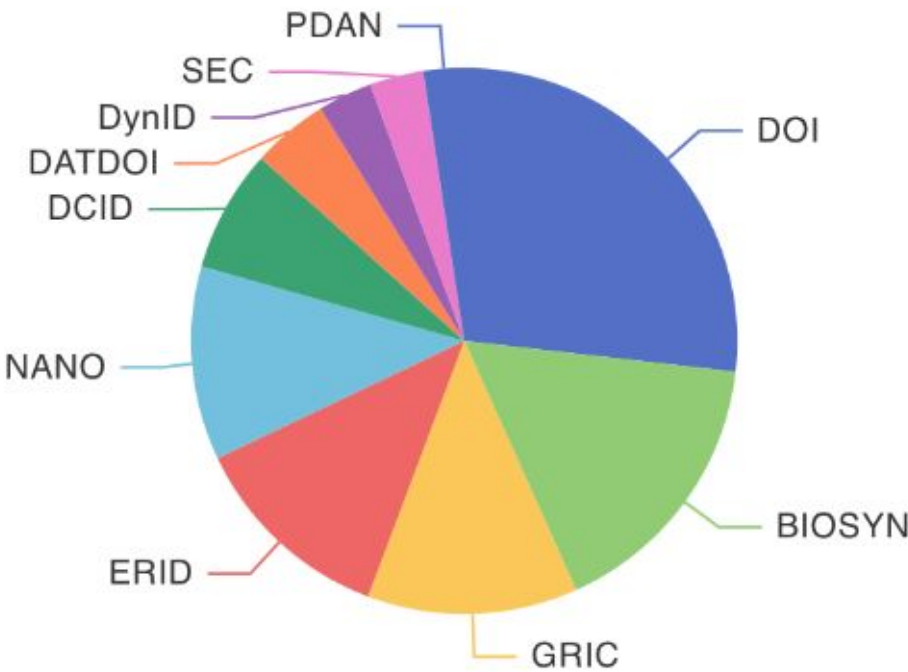
Version: **0.2.0** Commit: [6f916a23594e492804f3ca70ca74483a4589ea2b](#) 📅: 2023-11-16T12:16:15Z

PID Selection

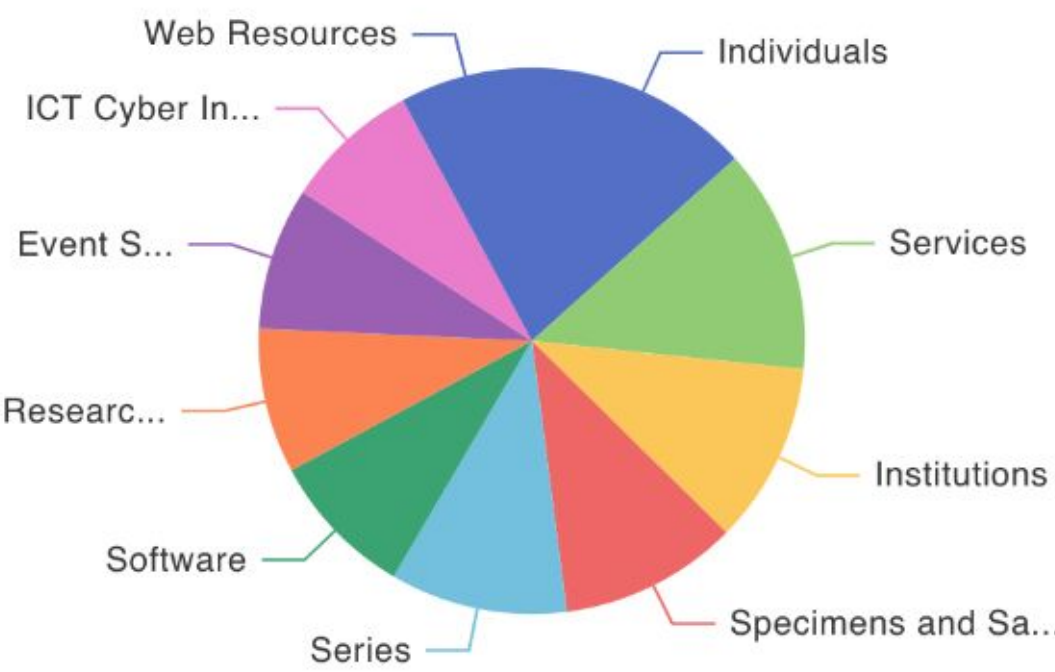
Scalability

100,000 - 1,000,000	153070
More than 100,000,000	65877
10,000 - 100,000	62658
1,000 - 10,000	61528
1,000,000 - 10,000,000	38607
10,000,000 - 100,000,000	18133
Less than 1000	9514

PID stack



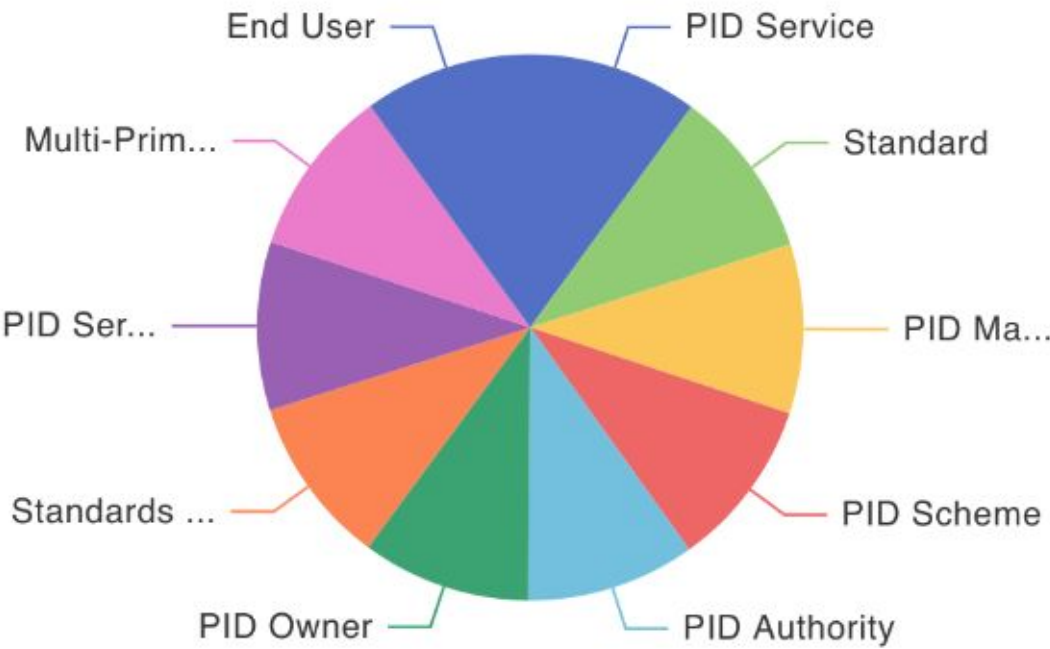
Entity



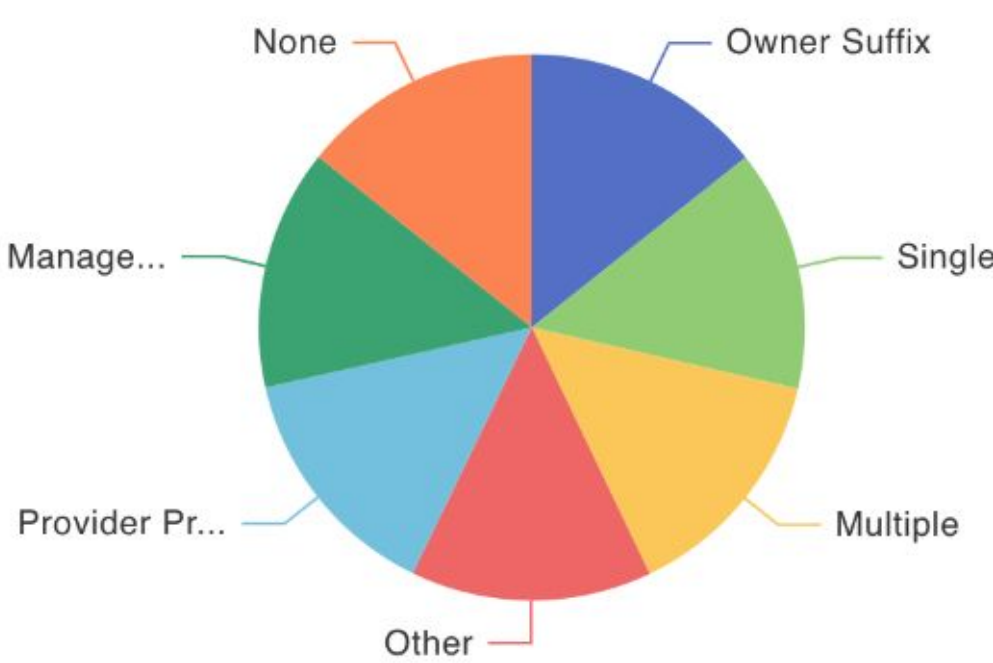
Readability

Both	136512
Human	136473
Machine	136402

Role



Namespace



Create a New Assessment



Assessment : Assess and document the current state of compliance of your subject based on rules extracted from the defined policy (e.g the PID Policy)

- each actor different assessment
- multiple
- Follows the assessment methodology

 **eosc**

FAIRCORE4EOSC

Core Components Supporting a FAIR EOSC

Compliance Assessment Toolkit

edabef...@einfra.grnet.gr

PROFILE

VALIDATIONS

ASSESSMENTS

create assessment

Step 1. Actor

Step 2. Submission

Step 3. Assessment

As a first step you need to select one of your validated actor roles in the following organisations:

☒ PID Manager at Data Archiving and Networked Services

← Prev

Next →

Create

Close

Contact and Service

Contact

Terms and Conditions

Data Licenses

Privacy

About CAT

Our Team

GitHub

Disclaimer

Follow Us

Newsletter

Youtube

Twitter

Partners

Data Archiving and Networked Services

 **DANS**

 **grnet**

 **DataCite**

 **GWDG**

Version: **0.2.0** Commit: 6f916a23594e492804f3ca70ca74483a4589ea2b

 2023-11-16T12:16:15Z

Object or Service Being Assessed



Subject : The subject that the User wants to assess (service, infrastructure , url etc)

 multiple

eoscFAIRCORE4EOSC
Core Components Supporting a FAIR EOSC
Compliance Assessment Toolkit

edabef...@infra.grnet.gr

PROFILEVALIDATIONSASSESSMENTS

create assessment

Step 1. ActorStep 2. SubmissionStep 3. Assessment

General Info

Submitter

Subject of assessment (Object, Entity or Service)

Select an existing subject from previous assessments or define a new one

Subject (*)wdfadfa-DANS EASY Repository-Repository ServiceClear Selection

Subject ID (*)wdfadfa

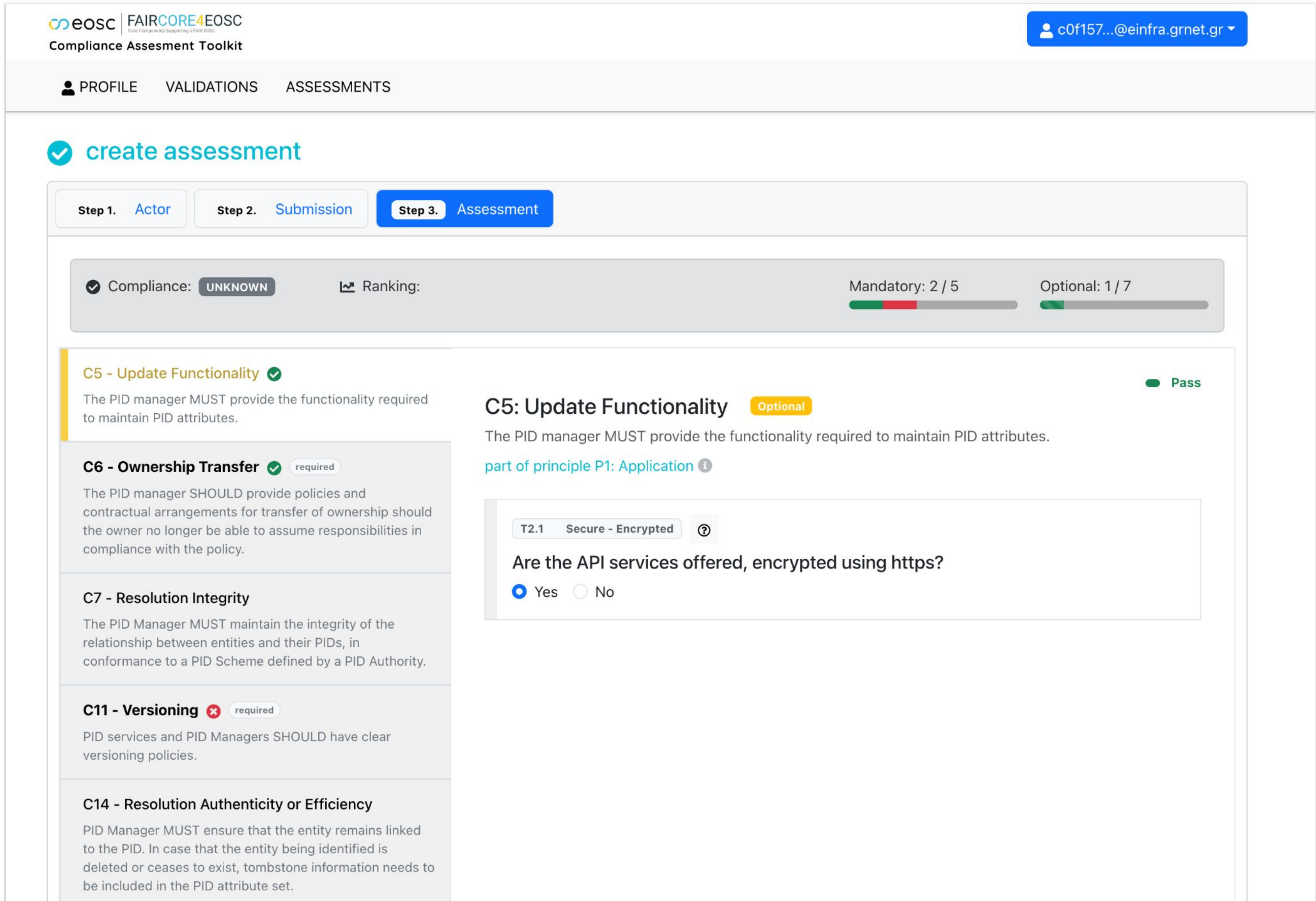
Subject Name (*)DANS EASY Repository

Subject Type (*)Repository Service

Rights, Licencing or Re-use

PrevNextCreateClose

Principle	P1: Application
Criterion	C5 - Update Functionality
Metric	<pre>"type": "number", "benchmark": { "equal_greater_than": 1 },</pre>
Test	T2.1 Secure - Encrypted Are the API services offered, encrypted using https?



Mandatory/ Optional Criteria

Mandatory: 0 / 13

Optional: 0 / 13

Compliance: UNKNOWN

Ranking:

Criterion

Optional

Required

Metric

>1

>=2

Test

TEST

Mandatory: 11 / 13

Optional: 12 / 13

Compliance: PASS

Ranking: 0

eosc | FAIRCORE4EOSC

Compliance Assessment Toolkit

c0f157...@einfra.grnet.gr

PROFILE

VALIDATIONS

ASSESSMENTS

create assessment

Step 1. Actor

Step 2. Submission

Step 3. Assessment

Compliance: UNKNOWN

Ranking:

Mandatory: 2 / 5

Optional: 1 / 7

C5 - Update Functionality

The PID manager MUST provide the functionality required to maintain PID attributes.

C5: Update Functionality

The PID manager MUST provide the functionality required to maintain PID attributes.

C6 - Ownership Transfer

The PID manager SHOULD provide policies and contractual arrangements for transfer of ownership should the owner no longer be able to assume responsibilities

C11 - Versioning

PID services and PID Managers SHOULD have clear versioning policies.

assessments

Import

Filter

Create New

Name	Type	Compliance	Ranking	Published	Subject Type	Subject Name	Organisation	Created On	Action
DANS EASY	eosc pid policy	UNKNOWN	7	Private	Repository Service	DANS EASY Repository	Data Archiving and Networked Services	2023-11-20	
DANS EASY	eosc pid policy	UNKNOWN	UNKNOWN	Private	Repository Service	DANS EASY Repository	Data Archiving and Networked Services	2023-11-20	

<<

<

1 of 1

>

>>

Back

Guidance

eosc

FAIRCORE4EOSC

Core Components Supporting a FAIR EOSC

Compliance Assessment Toolkit

c0f157...@einfra.grnet.gr

PROFILE

VALIDATIONS

ASSESSMENTS

✓

create assessment

Step 1. Actor

Step 2. Submission

Step 3. Assessment

✓

Compliance: UNKNOWN

Ranking:

Mandatory: 2 / 5

Optional: 1 / 7

C5 - Update Functionality ✓

The PID manager MUST provide the functionality required to maintain PID attributes.

C6 - Ownership Transfer ✓ required

The PID manager SHOULD provide policies and contractual arrangements for transfer of ownership should the owner no longer be able to assume responsibilities in compliance with the policy.

C7 - Resolution Integrity

The PID Manager MUST maintain the integrity of the relationship between entities and their PIDs, in conformance to a PID Scheme defined by a PID Authority.

C11 - Versioning ✗ required

PID services and PID Managers SHOULD have clear versioning policies.

C14 - Resolution Authenticity or Efficiency

PID Manager MUST ensure that the entity remains linked to the PID. In case that the entity being identified is deleted or ceases to exist, tombstone information needs to be included in the PID attribute set.

C5: Update Functionality Optional ⓘ

The PID manager MUST provide the functionality

part of principle P1: Application ⓘ

Are the API services offered, encrypted using https?

Yes

No

PID application depends on unambiguous ownership, proper maintenance, and unambiguous identification of the entity being referenced.

Guidance G4

Inventory of public evidence of processes and operations. Subjective evaluation of the completeness of the inventory compared to the infrastructures stated products and services.

New Features / Ideas - Next Steps

- Give the ability to users to create their own Policy
 - Either copy an existing one
 - Create a new one based on a library of principles, criteria, metrics, tests
- More Guidance
 - General Guidance
 - Help topics menu
- Use of a Vocabulary Service
- Support a Providers Registry
- Monitor the PID ecosystem

Visit and test the service



<https://cat.argo.grnet.gr>

Thanks! Questions?

Correspondence:
wim.hugo@dans.knaw.nl