Project no. 269977

APARSEN
Alliance for Permanent Access to the Records of Science Network

Instrument: Network of Excellence

Thematic Priority: ICT 6-4.1 – Digital Libraries and Digital Preservation

D22.1 PERSISTENT IDENTIFIERS INTEROPERABILITY FRAMEWORK

Document identifier:	**APARSEN-REP-D22_1-01-1_9**
Due Date:	2011-12-31
Submission Date:	2012-04-30
Work package:	WP22
Partners:	FRD,UNITN,UESSEX,CERN,STM,DANS,DNB,BL,GLOBIT,FORTH,BL,MRI,APA
WP Lead Partner:	Fondazione Rinascimento Digitale (FRD)
Document status	FINAL
URN	urn:nbn:de:101-20140516102

Abstract:

This deliverable provides a PI Interoperability Framework Reference Model that addresses functions, roles and responsibilities needed to make heterogeneous PI systems interoperable.

Delivery Type	REPORT
Author(s)	Emanuele Bellini (FRD), Chiara Cirinnà (FRD), Maurizi Lunghi (FRD), Barbara Bazzanella (UNITN), Paolo Bouquet (UNITN), Salvatore Mele (CERN), Hervé L'Hours (UESSEX), Yannis Tzitzikas (FORTH), Marjan Grootveld (DANS), Rene van Horik (DANS), Gerald Jaeschke (GLOBIT), Eefke Smit (STM), Sabine Schrimpf (DNB), Kirnn Kaur (BL), Heikki Helin (CSC), David Giaretta (APA)
Approval	David Giaretta/Simon Lambert
Summary	This deliverable provides a PI Interoperability Framework Reference Model that addresses functions, roles and responsibilities needed to make heterogeneous PI systems interoperable
Keyword List	Persistent identifiers
Availability	☒ PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0_5	2012-01-11	Initial version	Emanuele Bellini
1_0	2012-02-12	Small updates for consistency and name change	David Giaretta
1_1	2012-01-14	changed images of ontology relations, considered the scalability issue, added the executive summary	Emanuele Bellini
1_2	2012-01-15	minor changes to the executive summary	Emanuele Bellini
1_3	2012-01-16	Revised Exec Summary - to be delivered	Emanuele Bellini
1_4	2012-03-26	Deliverable refined	Emanuele Bellini
1_7	2012-04-26	Deliverable with table of related WP and follow up paragraph	Emanuele Bellini
1_8	2012-04-28	Minor format changes	David Giaretta
1_9	2012-08-15	Further minor format changes after acceptance by EC	Simon Lambert

Project information

Project acronym:	**APARSEN**
Project full title:	**Alliance for Permanent Access to the Records of Science Network**
Proposal/Contract no.:	**269977**

Project Co-ordinator: Simon Lambert/David Giaretta

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

CONTENT

1 INTRODUCTION 9

1.1 PERSISTENT IDENTIFERS AND THE INTEROPERABILITY CHALLENGE 9
1.1.1 Persistence and Interoperability 10
1.2 WORKPACKAGE 22, IDENTIFIERS AND CITABILITY: OBJECTIVES AND TASK RELATIONS 10
1.3 SCOPE OF THE DELIVERABLE 11

2 METHODOLOGY 12

2.1 STEP1: INITIAL INPUTS 12
2.2 STEP 2: SCENARIOS AND USE CASES 12
2.3 STEP 3: INTEROPERABILITY FRAMEWORK REFRENCE MODEL 13
2.4 STEP4: PI SYSTEMS BENCHMARK 13

3 PIS INTEROPERABILITY FRAMEWORK DESIGN 14

3.1 INTRODUCTION 14
3.2 PI SYSTEMS USER REQUIREMENTS 15
3.2.1 Current use of PIs 15
3.3 TRUSTED PIS DEFINITION 18
3.4 DEFINING PIS INTEROPERABILITY 20
3.5 IF SCENARIO DEFINITION 22
3.5.1 Scenarios on Citability and Metrics services 22
3.5.2 Scenarios on Global Resolution Services (GRS) 25
3.5.3 Scenarios on Digital Object Certification 26
3.6 USE CASES FOR INTEROPERABILITY OF PIS 27

4 PI INTEROPERABILITY FRAMEWORK REFERENCE MODEL 32

4.1 WHAT IS A REFERENCE MODEL AND WHY DO WE NEED IT? 32
4.2 THE REFERENCE MODEL 32
4.3 THE MAIN CONCEPTS 35
4.3.1 Resource 35
4.3.2 Digital Object 36
4.3.3 Author 36
4.3.4 Institution 37
4.3.5 PI (PI) 37
4.3.6 PI Domain 38
4.3.7 Policy 39
4.3.8 Resolver 40
4.3.9 User/Actor 40
4.3.10 Conceptual Map 40

5 PIS INTEROPERABILITY: RELATED PROJECTS 42

6 PI SYSTEMS BENCHMARKING 45

6.1 BENCHMARK CRITERIA ASSESSMENT 45
6.1.1 PI systems for Digital Objects benchmark 48
6.1.2 PI systems for Authors benchmarking 54

7 ARTICULATION WITH THE REST APARSEN WPS AND TASKS 58

8 CONCLUSION AND NEXT STEPS 60

8.1 INTEROPERABILITY FRAMEWORK FOLLOW UP 60

9 REFERENCES 61

SURVEY RESULTS AND ANALYSIS 62

PI SYSTEMS FEATURES ANALYSIS 96

GLOSSARY 106

EXECUTIVE SUMMARY

Scope	This work aims to investigate the interoperability issues between PIs and proposes a general Interoperability Framework (IF) as a starting point to design new solutions to support interoperability.
Context	The rapid increase of digital assets in recent years, especially in the context of e-science, has made clear that digital identifiers are crucial for preserving, managing, accessing and re-using huge amounts of data over time. The functionality to unambiguously locate and access digital resources, associate them with the related authors and other relevant entities (e.g. institutions, research groups, projects) is becoming essential to allow the citation, retrieval and preservation of cultural and intellectual resources.. Some notable solutions for identifying digital resources have been proposed in different domains like Libraries, Publishers, Science, and several standards are currently at a mature stage of development (e.g. DOI, Handle, NBN, ARK, Scopus Id, ResearcherId, VIAF, etc.) , but significant weak points still remain making persistent identification a complex problem which involves a large number of stakeholders who sometimes have opposing views on many of the issues that need to be addressed.
Interoperability challenge	Since the PI field is a fragmentary landscape and a unique global identification solution is far from being adopted, the challenge is to establish an IF among the current PI solutions to enable the persistent access, reuse and exchange of information through the use of existing identifiers and associated resources across different systems, locations and services. To tackle this challenge this work provides a Reference Model to support PI Domains (PID) in providing their PI-resource associations with a shared semantic model, enabling new services to discover new relationships and make inferences on digital resources.
Preparatory steps	As a first step, the work started surveying current uses and approaches by different user communities of identification systems and analyzing the main features of these systems in order to explore the main issues related to PI interoperability. The survey and benchmarking analysis have been translated into concrete scenarios and use cases to serve as requirements for IF designing the model.
Scope	This work aims to investigate the interoperability issues between PIs and proposes a general Interoperability Framework (IF) as a starting point to design new solutions to support interoperability.
Framework definition	The IF is a conceptual framework that defines the PIDs as the combination of technology, policies and decisions implemented by a user community interested in the preservation and use of digital objects and related PI systems. Since there are several ongoing PI initiatives, it has been necessary to define a list of criteria for determining the level of Trust required by a PID to be eligible for the IF. Thus, the IF will not argue about DP policies or the trustworthiness of the relations but assumes correct any statement coming from a Trusted PID. According to this approach, the IF assigns also to the PID the responsibility of guaranteeing suitable policies like, for example, the content selection/granularity criteria, scalability of the system, the Trusted Digital Repositories policies and certification, the trustworthiness of the PI management, and so on. Each PID is free to choose the best solution and we trust them for the correctness that we assume.
Benchmark analysis	Since the IF works only with Trusted PIDs, it was necessary to perform a benchmark analysis, based on the criteria indentified for determining the trustworthiness of the PID and select accordingly which PIDs are eligible for the IF. The results of the benchmark assessment have highlighted the existence of some critical issues that in some cases (e.g. DOI, Handle), can be easily solved with appropriate policies, in other cases (e.g. PURL, Cool URI,), are difficult to manage because of the scope of the PI system that does not fit

	with the IF Trustworthy requirements.
Interoperability Knowledge Base	The IF is described by a Reference Model that formalizes the main concepts of the IF identifying entities and relations in a so-called Interoperability Knowledge Base (IKB). In this way, the IKB defines the fundamental relations between the entities in play in the PID, creating a layer of accessible knowledge on which interoperability services can be built thanks of the explicit representation of these relations.
Trusted PIDs populate the IKB	All the information that is necessary to populate the IKB must be provided by the Trusted PIDs when they bring an entity into our scenario. The Trusted PIDs will populate the IKB with their entities presenting these contents though an API according to the types of information requested by the framework: for example for any digital object the PID, in addition to some description metadata, can declare existing multiple PI for the same object (e.g., DOI, NBN), any relation with other objects and any PI for persons or institutions known by the PID.
Building new services	Once all the contents and their relations from different Trusted PIDs (currently not interoperable and absolutely isolated) are visible and accessible through a common interface provided by the IF, final users can exploit them to create any type of interoperability service by accessing all the contents and their relevant relations which are exposed in a unique semantic format from different PIDs. In this way, we envision that it will be possible to overcome the limits, which impede the current PI systems to interoperate in a effective way.

1 INTRODUCTION

In recent years, the rising growth of scientific and non-scientific digital data is resulting in an increasing number of digital objects and resources that has to be managed, creating a new set of opportunities and challenges in the realm of science and culture in general. The possibility of accessing a massive amount of scientific and cultural data in digital format, the increasing linkage across authors and their publications, the development of new and much more powerful metrics for assessing impact of scientific production are only some of the opportunities that can be created in this data-intensive environment. However, this scenario has led to the emergence of new challenges, like for example digital preservation, data integration, quality assessment and provenance. These challenges become magnified in global contexts where resources are distributed across systems and standards, and the movement of data across disciplines and organizations is very intensive.

This imposes the need for implementing solutions, which allow identifying digital resources in a global and interoperable way across these boundaries, making different systems able to communicate and operate together in an efficient way. One area of interoperability that has been scarcely investigated is between identifiers and particularly PIs (PI).

Since different kinds of identifiers are in use across different stakeholder communities and systems, and multiple identifiers can be available and used within the same system, a reasonable solution is to guarantee interoperability across different identifier systems as well as develop services common to more than one system.

This report aims to investigate the interoperability issues between PIs and proposes a general Interoperability Framework (IF) as a starting point to design solutions to support interoperability

1.1 PERSISTENT IDENTIFERS AND THE INTEROPERABILITY CHALLENGE

The persistent identification of digital objects (e.g. articles, datasets, images, stream of data) and non-digital objects (namely real-world entities, like authors, institutions but also teams, geographic locations and so on) is becoming a crucial issue for the whole information society. The functionality to unambiguously locate and access digital resources, associate them with the related authors and other relevant entities (e.g. institutions, research groups, projects) is becoming essential to allow the citation, retrieval and preservation of cultural and intellectual resources. The rapid increase of digital assets in recent years, especially in the context of e-science, has made this dependency even stronger, making clear that digital identifiers are crucial for preserving, managing, accessing and re-using huge amounts of data over time.

This is especially true if we consider that today valuable scientific and cultural resources increasingly reside on network-based systems like the Web, encouraging the development of new effective solutions to allow the use of these resources into the future and across many different boundaries (i.e. geographical, organizational, cultural, disciplinary). The implementation of a system for persistent identification of digital and non-digital objects is the first fundamental step to this purpose, becoming a crucial prerequisite for sustained and reliable resource discovery, citation and re-use.

On the one hand, it is well-known that the use of URLs (which have been adopted from the birth of the Web to identify and reference network resources) can not be considered per se a reliable approach to address the long term identification and access of digital resources due to the fact that URLs serve the combined purpose of identifying a resource and describing its location. If the resource is moved to another location, the previous URL is no longer useful to access the resource. For this reason, the use of PIs has become the most popular solution to preserve access to a digital resource regardless of its location, by associating the PI with the correct current location, when the resource is moved.

Some notable solutions for identifying digital objects have been proposed in different domains and several standards are currently at a mature stage of development, like the Uniform Resource Name (URN), the digital object identifier (DOI), the persistent URL (PURL), the Archival Resource Key (ARK). Unique identifiers for authors are still not commonly used but some author identifier systems and initiatives have started to emerge in the last years, such as AuthorClaim, Scopus Author ID, Researcher ID, arXiv Author ID, ORCID. Recent efforts are also focusing on the development of a

standard for uniquely identifying institutions, as addressed by the NISO Institutional Identifier Working Group in the context of information supply chain, but it is worth to notice that PI systems for organizations are at a very immature stage.

On the other hand, despite the increasing awareness and interest for PIs significant weak points still remain making persistent identification a complex problem which involves a large number of stakeholders who sometimes have opposing views on many of the issues that need to be addressed. In particular assurance about the persistence of any of the identifier systems, specifically their resolvers, is lacking.

For instance, user communities such as librarians, archivists, researches, publishers, funding agencies have different visions and approaches to PI concepts, different legal and business models, different requirements and policies. The effect of this differentiation is that some identifier systems turned out to address better the needs of certain communities (and consequently are widely adopted by these communities) but many local solutions are still largely in use to address specific requirements. This makes clear that a discussion on PIs cannot only focus on the technical aspects of assigning PIs to digital resources, but needs to consider the complexity of the entire spectrum of responsibilities and requirements which underlie the development and maintenance of an identifier system. Each of these requirements involves the commitment of many stakeholders to maintain an appropriate infrastructure, to agree on policies, responsibilities, rights and restrictions. Long term funding commitments are, in general, impossible to obtain from funders. This may explain the fragmentation of the current landscape of PI systems and the difficulty of making these identifier systems interoperable.

Since a unique global identification solution is far from being adopted, the challenge is to establish an IF among the current PI solutions to enable the persistent access, reuse and exchange of information through the use of existing identifiers and associated resources across different systems, locations and services.

Through the IF, the identifiers assigned in one context can be encountered, and re-used, in another context, system or time and to access services outside the direct control of the PI assigner.

1.1.1 Persistence and Interoperability

Since a resource can be identified by (i.e. registered in) different identifier systems, the IF can be used to better guarantee the access to the resource (and related metadata), by using different PIs and through multiple ways of access. In other words one avoids putting reliance on any single system.

In addition, the IF should provide a general architecture on which services tailored on specific community requirements can be implemented facing the organizational and social complexity that characterizes the current PIs landscape.

These premises suggest that the identifier interoperability becomes another fundamental step to ensure reliable and long-term access and re-use of digital resources, enabling their persistence and facilitating their preservation.

1.2 WORKPACKAGE 22, IDENTIFIERS AND CITABILITY: OBJECTIVES AND TASK RELATIONS

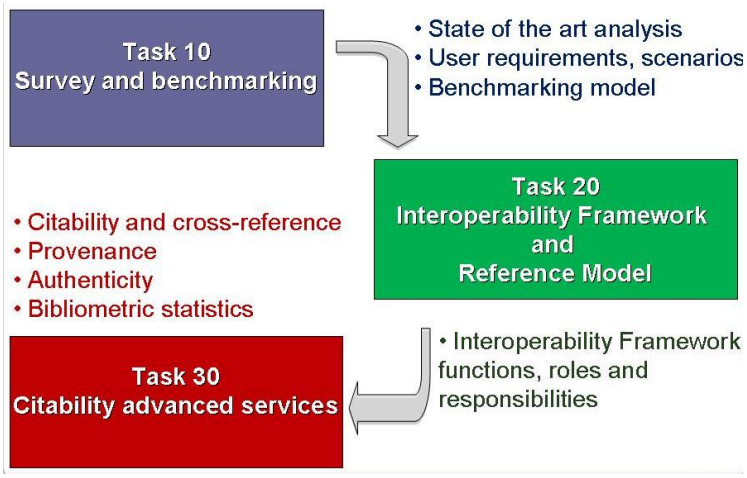
To start facing the issues underlying the interoperability challenge described above, the Work package 22 has three main objectives:

a) to provide an overview of the current PI systems and criteria for evaluation;

b) to design a reference model to describe an interoperability framework;

c) to define community-driven added value services.

The WP22 consists of three tasks. The relations among these tasks are described by the following diagram.



The **Task 10** has 3 main objectives: 1) to define a complete state of art of the available PI systems; 2) to provide an analysis of user requirements, use cases and scenarios, to outline a set of criteria in terms of organisational framework, object management, infrastructure and security; 3) to provide a benchmarking model, with criteria for evaluation, to support user communities and institutions in the identification of the appropriate PI system.

The **Task 20** is focused on modelling an IF for PI systems which addresses functions, roles and responsibilities to allow interoperability among these systems.

The **Task 30** aims at designing some advanced services for resources identified by different PI systems, such as services for citability, cross-referencing, quality assessment, citation metrics and evaluating the user satisfaction about these services.

1.3 SCOPE OF THE DELIVERABLE

Essentially, the present deliverable covers the first two tasks:

a) **Defining the user requirements**: the aim is to provide a current overview of the state of the art of the current PI systems through 1) a benchmark of the main PI systems according to a set of features suggested in previous studies, projects and the experiences in the field of the institutions involved in this work 2) a survey submitted to different communities. This work is the result of the Task 10.

b) **Designing a PIs IF**: the objective is to address all the identified requirements by providing a PIs IF to drive the implementation of added-value services which fulfil the user requirements. This work is the result of the Task 20.

Since the aim of first task was to collect preliminary information about the use of identifier systems and related practices, needs and requirements to orient the design of the IF, in this report we first present the IF, followed by the description of the all steps that created the ground for the design of the model.

2 METHODOLOGY

The workflow followed in the present deliverable consists of 4 steps: 1) Initial inputs including a) Survey on PI systems b) desk research and feature analysis c) APARSEN community and relevant stakeholders inputs; 2) Scenarios and use cases definition 3) Interoperability Framework design and 4) Benchmark analysis. Each step includes a number of outcomes that should be completed before moving toward the next step. The steps details are explained in the following paragraphs.

2.1 STEP1: INITIAL INPUTS

PI systems survey (see ANNEX I)

In order to gain a clearer understanding of the current state of the use of PI systems by different user communities, we developed a questionnaire that has been disseminated to the partners belonging to the APARSEN network of excellence[1]. Our intent was to explore existing **practices, requirements and resources** for persistent identification as well as to identify real **needs, gaps and challenges** related to the use of PIs systems. The complete results of the questionnaire are reported in Annex I. The analysis of the results provided a valuable starting point to define some basic requirements for future identifier solutions and the development of the IF, which is part of the WP22 expected-outcome.

Desk research (see ANNEX II)

Desk research was conducted to identify relevant features, which characterize the main current PI systems and may have an impact on interoperability. This analysis was also useful to understand weaknesses and strengths of each PI system in relation to the user expectations about digital preservation. For instance, some systems provide functionalities to the users that potentially prevent the use of such PI systems as a reliable tool for guaranteeing long-term access to the resources. This aspect is crucial to consider the eligibility of these systems for an interoperability common framework.

PI systems to define a small set of criteria to assess the eligibility of these systems to the IF. This assessment was the goal of the benchmarking analysis.

The results of the desk research activity and the correspondent feature analysis are reporter in the Annex II.

APARSEN partners, Key players and stakeholders inputs

Several APARSEN partners that are involved directly in PIs projects or services such as STM (DOI), CERN (ORCID), DNB (NBN:DE), DANS(NBN:NL), FRD (NBN:IT) and so forth or are users of these services, since they manage institutional repositories (usually universities and research institutions) or scientific datasets, have contributed to the writing of this document. Thus, this work is the results of a good synthesis of different points of view by addressing different requirements in a unique PI IF Reference Model.

Other key players such as DataCite, SURF Foundation, National Library of Sweden, National Library of Australia, National Library of Finland, CrossRef, IETF NBN Working Group and so on, have been interviewed during workshops and meetings such as that organized by Knowledge Exchange on "Exploring interoperability of Persistent Object Identifier systems" which produced an important contribution to the identifier interoperability issue through the so called Den Hague Manifesto[2]. The point of view and the suggestions of these stakeholders has been taken into account during the entire work.

2.2 STEP 2: SCENARIOS AND USE CASES

[1] http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/

[2] http://www.knowledge-exchange.info/Default.aspx?ID=62&M=News&NewsID=124

First, scenarios have been defined to introduce and concretize the interoperability concepts and requirements, by providing a number of envisioning uses for the IF following the Scenario Based design technique. These scenarios have been then translated into more simple use cases, a schematic definition useful for identifying entities, their relations, functionalities and so forth. The results of this phase have been used as input for the modelling phase.

2.3 STEP 3: INTEROPERABILITY FRAMEWORK REFRENCE MODEL

The aim was to propose a conceptual framework addressing the identifier interoperability issues, which have been identified in the survey and benchmarking phases and have been translated into concrete scenarios and use cases to serve as requirements for designing the model in the second phase of the present work. The framework describes the entities of our domain, their relations and dependencies, the main functionalities and a minimal sets of concepts in order to enable the development of specific implementations (i.e. interoperability services).

2.4 STEP4: PI SYSTEMS BENCHMARK

The current PI systems for digital objects and authors have been assessed based on a number of criteria that specify some crucial aspects for interoperability. These criteria have been adopted to identify the trusted PI systems. The objective was to provide a clear overview of the current identification solutions and practices, highlighting the main differences among the PI systems in order to identify which systems may be considered eligible for the IF. This activity was conducted as an expert review by the WP22 partners.

3 PIs INTEROPERABILITY FRAMEWORK DESIGN

3.1 INTRODUCTION

Persistent and unambiguous identification of digital resources is a fundamental issue for enabling their granted accessibility and re-usability over time.

On the one hand, it is well-known that Internet resources have a short average life and their identification and persistent location pose complex challenges affecting both technological and organizational aspects. On the other hand, the common approach of using URLs to identify and reference digital resources can not be considered a reliable approach because URLs are not durable identifiers, being location-based identifiers, and they lack trustworthiness on the generation process of the URL associated to that resource. Therefore the URL-based identification can lead to the loss of access to information or under-use of available data [5].

Many initiatives and technological solutions - such as Uniform Resource Names (URN), Persistent URL (PURL), Archival Resource Key (ARK), Handle System with its Digital Object Identifier (DOI) implementation, the Library of Congress Control Number (LCCN) and so on - have been proposed as possible solutions. However, the survey conducted in the context of the APARSEN project (see ANNEX I of the present document for detailed results) shows that a restricted number of approaches such as DOI (33%), Handle system (29%), URN-based identifiers (25%) seem to prevail at European level.

The survey results show also that content management is a fragmentary landscape. While the most common and cross-domain approach seems to be the Open Archive, on the other hand the institutions that do not adopt it, use a number of different solutions. In particular, this means that there are a number of different metadata schema that describe the resource in different ways. The problem of mapping a metadata schema to another is well known and affects the interoperability possibility. Moreover, the differentiation among access systems and repository systems suggests that the IF can not be focused on specific solutions or approaches and has to remain at a high level of description.

This result confirms the change of focus from technological solutions (for instance the simple URL redirect performed by URL rewrite engines) toward upper level solutions, where policies, business, social and political issues are taken into account.

This scenario suggests also that it is not viable to impose a unique PI technology for every community by promoting a top-down standardization. In fact, in the current practice, each system progressively satisfies a specific community (or few communities) with specific identification requirements. Moreover, it is commonly accepted that during the resource life-cycle, a digital object can receive more then one PI (e.g a DOI is assigned to the copy published by a publisher and an Handle is assigned to the self-archived post-print when is submitted to an Institutional Repository) according to the identification requirements of the institution that manages a copy of the object. Therefore, these PI services are becoming progressively complementary, even if some overlaps still remain.

Therefore, the current challenge is to setup an IF among these systems in order to allow the implementation of services required by specific communities.

The results of the APARSEN survey indicate that citability (76,53%), global resolution service (61%) and PI resolution service to the resource and digital object certification (both 55% ca.) are considered the most interesting services among the proposed options. These options were presented in order to understand the user expectations towards a class of underlying services. For instance, the citability option includes services directly related to the citation capability, but also metrics services (selected by a 30% of users). Taking into account the APARSEN survey results, we have decided to cluster the new proposed services into 3 macro categories: 1) **Citability and Metrics** 2) **Global Resolution services** and 3) **Digital object certification**.

In synthesis, the IF that we propose aims to define a reference model to describe actors, relations and other entities involved in the PI management in order to create the conditions for the implementation of the services defined in the scenarios presented in the section 3.

3.2 PI SYSTEMS USER REQUIREMENTS

In this section, an extract of the results of the questionnaire on PI systems for digital objects, authors and organizations is reported with a focus on the results which turned out to be more relevant for the design of the IF. The complete and detailed analysis of the results is provided in the Annex I. However, according to the results about the provenance of the survey participants presented in the ANNEX I, it is worth to notice that these considerations are applicable only at European level.

3.2.1 Current use of PIs

A first analysis was conducted to investigate the current use of PI systems for digital objects, authors and institutions among different stakeholder communities. The results show that the DOI, Handle System and URN based identifiers are the most popular PI systems for digital objects. Other systems, like PURL and ARK, are used by a minority of participants. From this result we gain a first indication on which systems have to be considered to be included into the IF.

The survey results show also that PI systems for identifying authors are scarcely adopted. The major initiatives are underused or unknown. In any case, the IF has to assume the existence of Author ID systems, but avoiding a focus on specific implementations.

A very similar result to the previous one has been found for organizations. However, the answers of the participants indicate that there are no specific PI initiatives for organizations. In fact, the majority of the respondents reported that no system is adopted to identify their organizations. Globally, a fragmentary picture emerges where PI systems adopted for digital objects are shyly adopted for institutions. Following the same approach holds for author PI systems, the IF assumes the existence of Organization PI systems avoiding a focus on specific implementations.

Limits for adoption of PI systems

One of the objectives of the survey was to investigate the limits experienced in using PI systems for digital objects. In the Annex I are reported some expected results such as "Locally defined" and "no standard associated" referred to internal identifiers solutions. On the other hand, the answer "low adoption" referred to DOI and URN by a number of responders is quite surprising, but it can be explained if it is referred to a specific user community. A certain PI system can be scarcely adopted within a given community and consequently it can be perceived as poorly adopted by the members of this community. For instance, since the analysis by stakeholder groups shows that DOI is the most common PI system used by universities, research organizations, archives and publishers and URN is the most widely used PI system by libraries (see the ANNEX I), it is possible that a library choosing DOI might have perceived as low adopted that solution within its reference community.

This scenario shows that PI systems are becoming increasingly oriented towards a specific community, indicating that an IF that allows a cross-community and cross-system communication is clearly needed.

Regarding the PI for Authors, the results show that many organizations do not use PI systems for authors and creators mainly because they do not consider it as a key issue or because of lack of awareness. In general, looking at the explanations reported by those who selected the "Other" option, we can say that users perceive a certain level of immaturity for systems for author identification which concerns services, trust and authority.

If we compare the obstacles that the respondents reported about the use of PI systems for authors with those about the use of PI systems for organizations, we can notice that the two most frequently selected obstacles are the same: the lack of awareness and the fact that the use of PI systems is not considered a key issue for the organization. This result confirms that one of the main actions of intervention to promote agreement across the different stakeholder communities about the adoption of PI systems should start from increasing the level of awareness about the available systems and their potential positive effects. However, the users of PI systems for organizations seem to be slightly more aware of the potential beneficial impact of using PIs for information discovery compared to users of PIs systems for authors.

Granularity and versioning

The recent growth of the amount and complexity of data is an important aspect that has to be considered dealing with the scalability, granularity and versioning issues. This has been confirmed by the results of the questionnaire, showing that a finer capability of a PI system to identify and recover part of digital objects is required.

Moreover, the survey results indicate also that the most common approach for content versioning is linking a new version to the original version through metadata, followed by the practise of considering the new version as an autonomous object. The use of naming rules is less common among the participants.

According to these results, the IF should include those PI systems that support the scalability, granularity and versioning issues working mainly at metadata level.

Technology, organization of the service, scope and naming rules

We investigated the user requirements in 4 domains: technology, organization of the service, scope and naming rules. In terms of technology, our results indicate that users prefer to adopt a system that represents a standard de facto, widely adopted and based on an open source infrastructure. This was an interesting input in defining the criteria to evaluate as eligible for the IF the PI system (Trusted PI).

About the organization of the service, distributed naming authority and supported by an institution with a mandate were the preferred options. In terms of scope, the respondents reported to prefer systems open to any digital objects and cross-community. Finally, about naming rules opaque identifiers that support deep granularity are preferred than semantic identifiers supporting low-level granularity. No relevant differences were found between the stakeholder groups in the requirements for adopting a PI system for digital objects.

Services

Table 2 shows that citability is the most important service associated to the use of PIs, followed by services which support resolution (i.e. global resolution services, resolution to the resource or to metadata). More than half of participants reported services for digital object certification among the required services. According to the stakeholders analysis (see ANNEX I) seems that if citabilty is a desired service for all the stakeholder groups, aspects related to the resolution mechanisms are more relevant for libraries, archives and publishers, while aspects related to certification (and metrics) are more important for universities and research organizations.

Moreover, differently from the expectations, the PI basic services are those most required. The so-called “advanced services” that were considered most important for the IF instead, have received less votes. According to this result, the framework design has to take into account also the objective to empower the basic PI services in addition to set up the conditions for developing new advanced services.

	Value	Freq	Percent %
Mostly PI Basic services	**Citability**	**76**	**74**
	Global resolution service	**62**	**60**
	PI resolution service to the resource	57	55
	Digital Object certification	55	53
	PI resolution service to metadata	50	48
Mostly PI future advanced services	Association of PI to multiple location (URLs)	41	40
	Metrics	31	30
	Multiple association name	27	26
	Link digital object to dynamic dataset	19	19
	Others	3	3

Table 2: PI associated services

The PIs Analysis produced a very clear pattern of results, which can be summarized as follows:

- There are some weaknesses of some PI systems about ensuring persistence, which concern the possibility to retrieve and access a resource in a reliable way along the time.
- Some PI systems require a fee, instead others are totally free of charge. This implies that there are different modalities of access to the PI services and different level of service provided. For instance the DOI manages the ONIX metadata set that supports the DRM of the content. Instead other PI systems (usually those that are free) provide a basic resolution service. In order to make these systems interoperable it is necessary to define a common level of service among PI systems.
- Textual documents and images are the most common digital objects managed by the respondents, which are reported to require a high level of granularity to identify them.
- There is a different level of maturity between the more advanced systems for digital objects and the gradually emerging solutions for authors and organizations.
- For digital objects we are assisting to a convergence toward few major systems: DOI and Handle, used mainly by universities, research organizations and publishers and URN, largely adopted by libraries. However, internal solutions are largely diffuse, especially by libraries and archives.
- About authors and organizations, there is still a lack of use of PI systems or, at best, a wide use of internal solutions, mainly due to a lack of awareness or added-value services tailored on the specific needs of the stakeholders.
- High level of heterogeneity has been reported in the adopted metadata schema, repository systems and access systems, showing the complexity of the required interoperability levels.
- There is a substantial agreement among the stakeholders about the requirements for a PI system for digital objects. It should be: a de facto standard, widely adopted, based on an open source infrastructure, characterized by a distributed naming authority, supported by an institution with a mandate, open to any digital objects, cross-community, adopting opaque identifiers and support deep granularity.
- Stakeholders agree also on the main requirements for a PI system for authors: it should be cross-discipline, managed by public/government institutions and transnational.
- Preservation practices are quite diffuse between the stakeholder communities. The main common practises are transformation, multiple copies and offsite storage.
- There is a disparity between the reported need of preserving digital resources permanently and the scarce commitment in terms of funding and sustainability. When funds are available to this purpose, they come mainly from the organization budget.

3.3 TRUSTED PIS DEFINITION

In order to design a reliable IF among PI systems, we have to define the criteria that a PI system should have to be eligible to the framework. A PI IF has to be reliable to enable the development of advanced services. Thus, only those PI domains that match a definition of trust will be take into account as a component of the framework.

In order to define the trusted PIDs, we introduced a small set of criteria distinguishing between mandatory (M) and optional (O) criteria. The following criteria are adopted to decide if a PI domain is trusted and eligible for the IF.

The definition of these criteria has been suggested by several studies such as, PIs for Cultural Heritage DPE briefing paper [15], NESTOR reports on trustworthiness of PI systems [14], A Policy Checklist for Enabling Persistence of Identifiers [10], the results of the ERPANET [3] and DCC [4]workshops.

1. **Having at least one Registration Agency (RA).**
   Within a PI domain it is necessary that a RA is established to assign and maintain the association PI- digital resource. This criterion is considered mandatory in our trust assessment (M).

2. **Having one Resolver accessible on the Internet.**
   To meet this criterion a resolver able to resolve a PI has to be accessible on the web. This criterion includes also the capability of a PI to be resolved to a single object such as webpage or file, or to both object and metadata or to multiple objects, such as different formats of the same objects, or different content types, through the same PI. We consider this criterion mandatory (M).

3. **Uniqueness of the assigned PIs within the PI domain**.
   The RA has to guarantee that a PI is univocally assigned to a digital resource within the PI domain. In fact, since a PI is essentially a string, the uniqueness can be assured only within a domain of reference served by a defined RA. This criterion is considered mandatory (M).

4. **Guaranteeing the persistence of the assigned PIs**.
   Each RA has to guarantee the persistence of the generated PI in terms of preventing the following possible actions:

   a) *String modification*: indicates the PI string update. This kind of updating procedure is not allowed according to our definition of a trusted system.

   b) *Deletion*: indicates the possibility of deleting a PI once it has been created and assigned. This is another process that must be avoided to guarantee trust.

   c) Lack of *sustainability*: indicates that a RA is not able to guarantee its commitment to maintain a PI as far as the identified resource exists. Managing identifiers in a sustainable way is another requisite for a trusted PI domain.

   The point a) and b) can be addressed at a functional level of the PI service but they depend on the PI domain policies; the point c) is related to the sustainability of the PI service and the PI domain business model. This criterion is considered mandatory (M).

5. **User communities, which implement the PID should implement policies for digital preservation (e.g. trusted digital repositories).**

[3] ERPANET workshop Persistent Identifiers Thursday 17th - Friday 18th June 2004-University College Cork, Cork, Ireland www.erpanet.org/events/2004/cork/index.php

[4] DCC Workshop on Persistent Identifiers 30 June – 1 July 2005 Wolfson Medical Building, University of Glasgow http://www.dcc.ac.uk/events/pi-2005/

It is well known that the main objective of a PI is to provide a reliable access to digital resources in the long term. Thus, if on the one side the RA has to guarantee the persistence of the PIs and their association with the identified digital resources, on the other side, PIs should be used to identify stable and preserved digital resources. The content providers should manage their contents with repositories compliant with standards and common criteria of trustworthiness[5] and implement digital preservation strategies for the resources identified by a PI. This criterion is considered optional (O), since content providers manage resources with different life cycles and they can also adopt different commitment to preserve their contents in respect to other institutions. For instance, as shown in the survey results (Annex I), there are several cases where preservation strategies are not still in place despite the PIs are widely used.

6. Reliable resolution.

One of the crucial functionalities of a PI system is ensuring that the resolution results of a PI are always the same across time. The definition of the meaning of *the same* statement is critical, since different domains may manage digital resources at a different level of granularity and require that a PI is generated and assigned to different levels of abstraction of a digital resource.
For instance, the PDF version of an article and the HTML version of the same article can be considered an "equivalent manifestations" of the same object within the DOI domain, while they would receive two different identifiers in the NBN domain. Again, if a digital resource is subjected to digital preservation strategies, such as transformation, the results can be considered equivalent manifestations in a domain but not in another.
In fact, in the CrossRef DOI service there is only a guideline, namely "Assign new CrossRef DOIs to content in a way that will ensure that a reader following the citation will see something as close to what the original author cited as is possible."[6]
According to this, the reliability of resolution is referred to guarantee, provided by a PI domain, that the resolution of a PI points to *the same* resource along the time, according to the similarity definition adopted by a PI community. This criterion is considered mandatory (M).

7. Uncoupling the PIs from the resolver.

This criterion is crucial and it is referred to the PI generation rule defined by a PI system. To be eligible for the IF a PI system has to be based on identifiers whose syntax does not include the URL of the resolver or the content provider in the string. For instance, the NBN syntax definition does not include the URL of the associated NBN resolver. This feature is necessary because the URL of the resolver itself can change. Thus, if a part of the PI string specifies the URL of the resolver domain, all the PIs which contain the original URL will become invalid, in case the resolution service is moved to another domain. This criterion is considered mandatory (M) in the proposed IF.

8. Managing the relations between the PIs within the domain.

This criterion identifies the possibility to specify the linkage between resources within the PI domain through explicit relations between their identifiers. For example, a PI domain can make

[5] Examples of Trusted digital repository criteria are: *Date Seal of Approval*: http://www.datasealofapproval.org/, *Nestor Catalogue of Criteria for Trusted Digital Repositories*: http://files.d-nb.de/nestor/materialien/nestor_mat_08-eng.pdf, *Trusted Digital Repositories: Attributes and Responsibilities*, http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf - *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC): http://wiki.digitalrepositoryauditandcertification.org/pub/Main/ReferenceInputDocuments/trac.pdf- ISO/DIS 16363: http://public.ccsds.org/publications/archive/652x0m1.pdf, ISO/DIS 16919 http://wiki.digitalrepositoryauditandcertification.org/pub/Main/WebHome/RequirementsForBodiesProvidingAuditAndCertification-SecRev1.doc

[6] http://www.crossref.org/CrossTech/2010/02/does_a_crossref_doi_identify_a.html

explicit the part-of relation between resources embedding this linkage within the PI string, or using metadata. An example of this kind of relation is that exists between a resource and the collection of which it is part. This criterion is considered optional (O) in our framework, but it represents an added value that can speed up the implementation of interoperability services.

We are aware that there are other features and criteria which can be considered in a Trusted PI definition. A critical example is scalability. A PI system that aims to identify an increasing number of objects on Internet (i.e. a global distributed system) must also handle the scalability to be considered Trusted. In fact, scalability is one of the basic requirements for the long term sustainability of every PI service. However, evaluating the scalability can be difficult for many PI systems.

As an example, a centralized service can be based on a scalable grid infrastructure. Therefore, if from an external point of view based on the centralization of the system we could say that the system is not scalable, but this evaluation could be wrong on the basis of the specific technological infrastructure of the system. PURL system is an example of this difficulty. The service is based on DNS that it is obviously scalable but it is difficult to obtain enough information about how the "redirect" service is actually implemented. Thus, the main reason why we have not included the scalability as a criterion is due to the variability of the possible technical implementations of a system, and the difficulties in obtaining sufficient information about the technical implementation for making an accurate assessment. The difficulties of obtaining definitive results on such a criterion represent an ongoing concern that has been taken into account in the present work.

3.4 DEFINING PIs INTEROPERABILITY

Interoperability is an essential feature for federated information architectures which operate in heterogeneous settings and over time. However, the use of the concept is very heterogeneous: interoperability is conceived in an object-related or in a functional perspective, from a user's or an institutional perspective, in terms of multilingualism or of technical means and protocols. Moreover, interoperability is conceived at different levels of abstraction: from the bitstream level up to the semantic interoperability level.

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

The ISO/IEC 2382 Information Technology Vocabulary defines interoperability as “the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units.”

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

“the ability of two or more systems or components to exchange information and to use the information that has been exchanged” (IEEE glossary[7]).

Presently, it is commonly accepted that during the resource lifecycle, a digital object can receive more than one PI according to the identification requirements of the institution that manages a copy of the object. For instance, an Handle is assigned to a resource stored in an Institutional Repository (IR) registered to the Handle System registry; if a copy of this resource is also deposited in the National Library, the resource will receive an NBN, and so forth.

Similarly, a researcher receives a different ID every time he moves from an institution (e.g.

[7] http://www.ieee.org/education_careers/education/standards/standards_glossary.html

university) to another. Given this panorama, one of the main challenges for persistent identification is to setup an IF among different systems in order to allow the implementation of those services which are required by specific communities.

Identifier interoperability is necessary for different purposes such as:

1. Metadata interoperability (since metadata often define a relationship which somebody claims to exist between two referents);
2. The creation of standards to define the relationships which occur between the referents of different identifiers;
3. The creation of services between different systems, e.g., discovery of "related content" items; compiling multimedia objects, etc.

Interoperability can be considered on different abstraction levels, within a continuum ranging from a very concrete to a very abstract perspective. For instance, Stefan Gradman, in [3], identifies four layers for digital library interoperability. These layers are: Technical (common tools, interfaces and infrastructures), Syntactic (allowing the interchange of metadata and protocols elements), Functional (based on common set of service) and Semantic (allowing to access similar classes of objects and services across multiple domains). Similar to this approach, Norman Paskin, in [4], identify three levels for PIs interoperability:

1. **Syntactic interoperabilit**y. The ability of systems to process a syntactic string and recognise it (and initiate actions) as an identifier even if different syntax schemes are used in different systems.

2. **Semantic interoperability**. The ability of systems to determine if two identifiers denote precisely the same referent; and if not, how the two referents are related.

3. **Community interoperability**. The ability of systems to collaborate and communicate using identifiers whilst respecting any rights and restrictions on usage of data associated with those identifiers in the systems.

These aspects form dependent layers: community interoperability is only possible if semantic interoperability is ensured; semantic interoperability is only possible if syntactic interoperability is enabled.

Another definition of interoperability in the context of identifier management infrastructures can be derived from the PLIN project[8] "A component is interoperable if an action can operate on the component from outside the curation boundary of the identifier management system. The action must follow a well-defined interface, which is known outside the curation boundary. If a component is not interoperable, then only the own identifier management system infrastructure can be used to operate on it. If the action uses a publicly documented interface through an open protocol such as Web **services**, it is interoperable."

Both the definitions capture an essential aspect of interoperability, highlighting that interoperability is not only a technical issue, but it deals with boundaries of responsibilities which regulate the curation of a resource and its identifier/s during their entire lifecycle. This includes, for example, responsibilities for the preservation, the description and the identification of the resource, and responsibilities for the resolution of the identifier, which refers to the resource itself. Since different content providers and identifiers domains may have very different curation boundaries and responsibilities, an since resources can be moved across these boundaries, an IF should be as flexible to provide continued interoperability even when the holders or the nature of these responsibilities may change.

[8] http://www.pilin.net.au/Project_Documents/Glossary.htm

3.5 IF SCENARIO DEFINITION

The pervasiveness of the designed services is in fact one of the key factors for obtain an extensive consensus building and for the long term sustainability of the service itself. To this end, a number of scenarios coming from different communities are provided.

This document gathers a number of envisioning scenarios written assuming the IF as a black box. Each scenario thought a storytelling technique, describes a possible interaction between a user and the IF in a particular context. The context is a description of the issues to be solved and the interaction is a description of how the system is able to support the user in solving that issues. This information has driven the IF design.

Methodology: Scenario Based Design

In order to face the open issues described above, we aim to design an IF following the Scenario Based Design (SBD) principle [1]. Scenarios are a vocabulary for coordinating the central tasks of system development, understanding people's needs, envisioning new activities and technologies, designing effective systems and software, and drawing general lessons from systems as they are developed and used.

The basic idea behind scenario-based methods is that descriptions of people using technology are essential in discussing and analyzing how the technology is (or could be) used to reshape their activities. A secondary advantage is that scenario descriptions can be created before a system is built and its impacts felt. According to [1], there are three different use modalities of the scenario approach:

1) Analysing the activity [2] for structuring data harvested from observation of user tasks.
2) Prototyping for envisioning the future task and stimulate the design process.
3) Evaluation for testing existing solutions.

The approach works at Prototyping (2) level assuming the existence of an IF as a black box. The methodology consists in considering the enabling technologies as given and, through the storytelling techniques, describing future services in action, the possible interaction with the users, etc.

3.5.1 Scenarios on Citability and Metrics services

[Citability] Scenario I: Data citation and access

Tom is a researcher in Physics at the University of Florence and he is involved in an experimental research for monitoring the air pollution trend in Florence. Through sensors installed in different points of the city, air samples are acquired during pre-defined slots of time. After a defined control time, the samples acquired are organized in a dataset and submitted to an open access repository for environmental data managed by National Research Council. Each dataset submitted receives a PI that allows its citability and IPR protection.

The repository submission process requires the PI for the laboratory, which has created the collection and the PI for each related author declared. Once a dataset has been submitted, it becomes immediately visible and accessible to the scientific community under a IPR definition.

After six months, Tom submits a comparative study about air pollution to a conference. Other laboratories have already published papers in the past based on other datasets and Tom wants to show the differences between the results of the present study with those of previous studies. In particular, he aims to identify the distribution parameters and fit the collected data to a statistical distribution. To this end, he cites different PIs of old datasets that are stored in a different repository.

In order to verify whether the proposed model is effectively the best model to describe the data, the peer reviewers decide to resolve the PIs and analyze the old datasets.

Following the PIs, the reviewers can obtain an aggregated representation of different digital objects related to the analysis, such as the datasets, each related publications that are stored in different locations and published by different publishers, and in some cases a simulation of the data acquisition

[6]. Since some datasets are very old and they are no longer available at the original location, the PI resolution system could suggest the reviewer to follow an alternative PI to retrieve a copy of the same dataset available at another location. In this way all the information is available for the review analysis.

[Citability] Scenario II - Access to the appropriate copy

According to the Berlin declaration on Open Access, the University of Florence has set up an institutional repository where each researcher can self-archive a free copy of a digital document, in order to maximize its accessibility, usage and citation impact. Researchers usually archive peer reviewed research journal and conference articles or theses.

Mark is a researcher at Distributed Systems and Internet Technology Laboratory (DISIT) of University of Florence and he aims to submit an article to a journal. The publisher is currently registered at the online publisher SHERPA RoMEO website and Mark can easily verify that he is allowed to self-archive the preprint version of his paper.

Mark knows that the publication of a manuscript in a peer-reviewed journal usually takes a lot of months from the time of the initial submission, but he needs to quickly circulate the results within his scientific community in order to receive early feedback from their peers which may be helpful in revising and preparing further experiments. To this end, he decides to self-archive the preprint article in the Open Access institutional repositories of his University. Once submitted, the article receives an Handle from the repository and an NBN when is forwarded to the National Library for the long term preservation deposit.

After some feedback coming from the scientific community, Marks finalizes his article and submits a new version to the journal.

In the meanwhile Lucas, a researcher on the same field of Mark, is looking for new approaches for digital libraries applications. He searches on different portals and catalogues and finally, he finds the bibliographic information of Mark's open access article and decides to read it. Resolving the Handle identifier included in the bibliographic metadata, a common splash page provides the metadata information with the link to the copy of the resources stored into the IR and the NBN identifiers that allows the access to the copy stored to the National Library. The splash page informs the user that both copies are effectively available.

Lucas finds the article very interesting but he notices that it is a preprint and some conclusions seem to him not well defined. Lucas wants to read the final version of the paper (published) in order to verify if some lacks and issues are fixed in the new version, thanks the peer-review process. Lucas saves the PI into a feeds reader that will inform the researcher if related resources have been published.

After some moths, the feed reader shows to Lucas a link to the published copy waited by Mark and decides to buy the article.

[Citability] Scenario III - Access to fragments and linked resources

The Dutch Veteran Institute has interviewed Dutch veteran military staff who have carried out missions abroad in the last decades. These audiovisual interviews are online accessible for research purposes such as studying the use of metaphors in the way the interviewees talk about violence, or their mental images of friend and foe. Rachel, a linguist who enthusiastically uses this material, decides to publish her findings with references to the actual interviews. Because a typical interview lasts two hours, she applies an online "segmentation tool" to select the relevant segments. In fact, Rachel makes a so-called "Enriched Publication", which in her case consists of both text and audiovisual material, and of course the links between them. The examples she can provide this way literally speak to one's imagination. Before publishing, Rachel has cleared with the Veteran Institute that these fragments can be presented as Open Access.

The complete interviews are stored in a Trusted Digital Repository (TDR), provided with a PI. The segmentation tool can retrieve the PI and add a substring to it that pertains to the segment. Thanks to

clear policies which different parties (e.g. the TDR, the segmentation tool, the video streaming server) adhere to, Rachel can be sure that the fragment will be accessible in the long run. Moreover, the Veteran Institute as owner of the interviews, can be assured that accessing an Open Access fragment doesn't imply accessing a Restricted Access complete video.

Rachel receives positive comments on her Enriched Publication. She already knows a spectacular Youtube video that she would like to include in her next project, but first she needs to find out how to deal with this legally and how to guarantee permanence of the resource.

[Metrics] Scenario IV: Author career tracking and evaluation (Number of papers published by the Author x)

a) The University of Florence makes a public selection, trough exams and evaluations of publications. The candidates have to provide the list of their publications. In order to avoid potential frauds, the commission requires a proof of the existence of those publications. Many candidates have been working since 20 years and the first articles published at the beginning of their careers are no longer available at the original location. Thanks to the IF, the commission is able to check the presence and the authenticity of the publications using the research ID assigned by the institution where the candidate works or generated by an international initiative such as ORCID.

b) An industry is looking for an expert on nanotechnologies to improve its industrial process and rise up its market competitiveness. The human resource manager requires for each candidate the CV with the list of the publications exclusively related to the topic. Mark is a good candidate for the job position because he is a brilliant researcher and has submitted a number of papers during his research career to different journals and conferences. He has worked also in different European universities and his papers are stored in different institutional repositories. Up to now, he has not a complete list of his publications available and the common indexes such as DBLP Computer Science Bibliography cannot provide a complete overview of the situation because some publication submitted to Open Access journal are not indexed. Mark decides to query the System inserting his Author ID number. The system provides to Mark the complete list of his publications without duplications.

[Metrics] Scenario V - Researcher profiling with a collaborative index (number of papers produced in cooperation with other authors)

The department of Information Engineering of the University of Florence publishes a call for a post-doc position. Four candidates apply for the position. The dept. director needs a person with leadership abilities, high technical profile and collaborative skills. The candidate will be involved in a technical project with several partners and he/she will be responsible for some development tasks. From the interviewer point of view, the candidate profiles seem to be very similar. To solve this situation the dept. director decides to seek more information about the collaborative skills of the candidates by query the IF. The system requires an identification token (can be any PI system) of the candidates. The system returns a collaborative index calculated on the basis of the number of papers written by the candidate in collaboration with other people during his research activities. The system can give more detailed information providing an International Collaboration index based on the number of collaborations with international laboratories.

[Citability] Scenario VI Association of multiple author PIs with a single person

The premise here is that authors change organisation affiliation over time or may have multiple organisations (funding body, university, project) at one time and may therefore be assigned several disconnected author PIs.

a) An Archive is ingesting a data collection from a given author and adds that author to their local database with the persistent author ID provided. The Author ID is used to identify other related publications from the author for presentation to end-users. The Archive enters the Author ID into the Interoperability System and a number of possible matches are returned which might represent the same author. The Archive is able to contact the author directly during ingest negotiations and confirms

which of the other Author ID's represent the same person. In this way the Archive can 'link' the IDs and the Archive's presentation of related publications is greatly enriched.

b) Ideally a mechanism exists for the Archive to submit the linked identifiers back to the interoperability system and the system can respond to future author queries with increased confidence. The Organisation ID of the Archive could be presented alongside the associated author IDs on subsequent searches to clarify the provenance of the correlation between the IDs.

3.5.2 Scenarios on Global Resolution Services (GRS)

[GRS] Scenario I: Unique resolution interface: finding digital objects or author information using metadata

John wants to find various documents about a theory produced by Mike Mills. The problem encountered by John is that, whenever he tries to use Google to find documents about Mike Mills, he found mostly documents about Mike Mills, the film producer.

John remembers to have heard about a new interoperability service engine, where each entity is uniquely identified and linked to scientific documents mentioning the entity. John accesses to this service and types keywords such as the name of the entity. The query is sent to the IF. The system knows that only the name of the entity produces ambiguous results. The system returns a list of entities matching the given name, but with additional information for each of the entity, such as the domain of expertise or related entities. In the list, John can found Mike Mills, the film producer, but also Mike Mills, the scientist. John selects the right entity, and the search engine returns a ranked list of documents mentioning the entity. John discovers new documents such as the personal home page of Mike Mills, and finds that he has published various papers about his theory. From this result screen John can navigate to the website of various publishers and retrieve the documents he was looking for. In this way, John could discover new sources of information that, with a normal search engine, he would have never found (or maybe after a day of searching and browsing).

In short: the user enters a keyword query to look for an object or a person. The system presents a list of possibly matching of objects/persons with short descriptions, ranked by relevance. Short description is based on the metadata about the resource. Each results contains links to the repositories where the digital object or more information about a person can be retrieved.

[GRS] Scenario II: Unique resolution interface: finding aggregated information on a digital objects or author using an identifier

John wants to find a paper wrote by Mike Mills of which he knows the identifier used in one repository. He would like to find the paper itself, but also know if new versions of the same paper are available or if the author has developed his previous work further.

The problem is that, whenever John tries to use Google or the publisher website he can only find information about that specific paper and he can not understand if there are other versions of the same paper or if there are new developments of the theory presented in the paper he knows.

John remembers to have heard about a new interoperability service engine, where each entity is uniquely identified and linked to scientific documents mentioning the entity. John goes to this service and types in the identifier about the paper he knows. The query is sent to IF. The system returns aggregated information about the paper from different sources, the link to the resolver of the identifier he knows, but also in addition different links to related material all described by metadata which lead John to find the information he was looking for.

In short: given a PI a user wants to retrieve information about the object or author identified by the submitted identifier. The identifier has been issued by one of the systems registered to the IF. The system shows an aggregated view of the registered metadata available for that object or author and links to related material and to resolver sites. In addition, the system also provides a view of the aggregated information available about this entity and offers different resolvers by linking to the publisher websites.

[GRS] Scenario III: Finding information about a resource's authenticity and availability

Paul Weller is a researcher at the University of Amsterdam. His current research interest is the behaviour of politicians before elections and after they are elected. To investigate this behaviour he uses as input political programs (what did they promise before the elections) and minutes from the parliament (what did they actually do after being elected). Both resources are accessible on the Internet. Paul wants to use these resources. For Paul it is important that the resources are authentic and that they will be available not only today, but also in 10 years time.

Based on these two resources he publishes his results.

Paul wants to refer in his publication to these resources for reasons of verification and validation for his peer reviewers.

For Paul it is important to know:

- Are the resources authentic?
- Are the resources permanent accessible?
- Where can a resource be retrieved?
- How to refer to the resources?

The trusted PI infrastructure can provide Paul with answers to these questions. The PID is the key to unlock the answers for Paul's basic questions.

Paul can go to a global resolution service and retrieve information about the organization that authorized the PID. When the authorizing organization is part of the trusted PI infrastructure then Paul knows it is an organization he can trust. Paul can also retrieve the information of the organization regarding their policy on permanent access. The trusted PI infrastructure supplies Paul with good practices on how to reference the resources.

[GRS] Scenario IV: A machine using a Global Resolution Service to retrieve metadata and information about a resource's authenticity and availability

The Council of European Social Science Data Archives (CESSDA) is building up an European Research Infrastructure to give researchers better access to social science data. One of the building blocks of this infrastructure is the Data Documentation Initiative (DDI3.0) metadata standard. This standard covers the whole lifecycle of scientific data: Discovery & Planning, Initial Data Collection, Final Data Preparation & Analyses, Publication & Sharing, Long-Term Management.

On this DDI3.0 fundament there will be additional services like a Question Data Bank and Harmonization Tools. Most of these services will extract information from DDI3.0 resources in an automated manner. PIDs form an important part of the internal functioning of this infrastructure.

For a client application (a software tool) it is important to know:

- What kind of information can be expected?
- Are the resources authentic?
- Are the resources permanent accessible?
- Where can a resource be retrieved?

The trusted PI infrastructure can provide the CESSDA infrastructure with answers to these questions. The PID is the key to unlock the above questions.

For tools as well as for people it is important to have a single or unique location where all PIDs can be resolved. When resources in the trusted PI infrastructure are retrieved the tools can rely on the fact that these resources are permanent accessible and unchanged.

3.5.3 Scenarios on Digital Object Certification

[DOC] Scenario I - Authenticity

Emma wants to find documentation about a topic. She wants to create a collection of articles in order to read about the topic. Emma is interested in selecting only material coming from major journals because she trust this type of source better then others.

Emma remembers to have heard about a new interoperability service engine, where each digital

object is enriched with metadata and also information about the source. Emma accesses this service by typing in the search field keywords such as the name of the digital object she is interested into. The query is sent to IF. The system returns a list of entities matching the given name, but also additional information for each of the results, such as the domain of expertise and provenance information or related items.

Provenance information could be added to the publisher in order to certificate the material, either with the organization identifier but also with other techniques such as digital signature.

[DOC] Scenario II – Dissertation Authenticity Check

Laura needs to submit her dissertation ten days before her final exam. She had forwarded a digital copy to the University of Paris archive as administrative evidence. The University archive maintains the original copy for a long-term period. The final exam goes well and Laura decides to self archive her thesis into the University digital library for a wide dissemination. After a month, the University submits a copy to the National Library.

After several months, Laura decides to participate in a call for a position as associate professor. During a first step of the selection process, the committee need to formally check the titles declared by each candidate in their CVs. Using Laura identifier, the committee is able to retrieve a copy of her thesis from the Paris University archive. The system displays different locations where the thesis can be downloaded with additional information about content and source. They choose to consider only the copy present in the archive and not in other repositories because the one stored in the archive is considered the legal copy and for this reason provides enough evidence to claim the authenticity of the resource.

[DOC] Scenario III – Provenance

a) At the beginning of her career Kathy Mills participated to the curation of a data set, later preserved, of climate observation. Her name appears somewhere in the file. Twenty years later, this data set becomes a key element for tracking some crucial evolution in a variable long forgotten and crucial in the modelling of a very important climatic effect. All authors of the paper have retired, and Kathy name was only in the deposited data set and questions for her arise… where to find her? She's called now Catherine Smith.

b) Two large digital libraries and two institutional repositories have several, overlapping, records of a given person. If a central ORCID-like register of person/DOI exists, any partner can query the central register with their own DOI, get unambiguous person ID, go back to all other partners which will then use that ORCID ID, matched to their internal ID, to return all information which they have, so that the original four partners can dynamically federate their information, while remaining the ultimate custodians of the parts under their own responsibility.

3.6 USE CASES FOR INTEROPERABILITY OF PIS

This section aims at identifying the functional requirements of an IF for digital identifiers through a set of use cases.

The intention for the analysis of these use cases was to capture **who** (actor) does **what** (interaction) with the IF, for what purpose (**why**), **where, when and how** she/he performs the interaction and which are the main **issues** involved in the process. The set of the use cases should specify different ways to interact with the IF, and therefore define the main functionalities required by it, bounding its scope. This represents a fundamental step in designing the potential abstract architecture of the system and proposing added-value services to build on top of it. We notice that many of the use cases have been derived from the scenarios but the correspondence is not one-to-one.

Use case 1: identifier resolution

Who	A researcher in molecular biology

What	Wishes to access a paper identified by a given ID
Why	To study related work in the same research area she is interested in
Where and When	Online or library database Any time
How	Requires a resolution mechanism which ensures that the identifier can be resolved to access the corresponding resource
Issues	Implies a mechanism to find an appropriate resolver for the given ID

Use case 2: identifier resolution, broken link resolution service

Who	A physics researcher
What	Wishes to access a dataset identified by a given ID, which ceased to be resolved to a given location.
Why	To compare his results with those found in a previous study on a different dataset.
Where and When	Online Any time
How	Requires a "broken ID resolution service" which redirects the user to a matching resource (e.g. a copy of the dataset in a different location).
Issues	Implies a redirection system to find an appropriate matching resource given the ID of the original resource.

Use case 3: finding a digital object through metadata

Who	A psychologist
What	Wants to access to a digital object of which she/he does not have an ID
Why	to compare the results of his research on those founded in a previous study
Where and When	Online Any time
How	Through a keyword query
Issues	Requires a method to query by metadata ID registration agencies to get one or more IDs referring to the same entity The same service might provide the ID for the object, which can be used, for example, for citation purposes. Since multiple results can be obtained, the system should present the results with authoritative metadata to allow the user to select the right result.

Use case 4: finding related content

Who	A cognitive scientist
What	Wishes to discover and explore scientific publications or other information related to a paper, which published the results of a new theory about the role of emotions in decision-making. For instance she want to find: • Other publications on the same theory • Other publications of the same author

	• Other authors focused on the same research topic The "starting point" paper has its own digital ID.
Why	To explore related content on a topic of interest
Where and When	Online Any time
How	Requires a discovery mechanism to locate resources related to one, which the user had accessed.
Issues	Implies mechanisms for metadata linking though ID.

Use case 5: discovery of versions of the "same" digital object

Who	A social scientist
What	Wants to discriminate between different versions of the same work
Why	To access the final version of the work stored in the institutional repository
Where and When	Online Any time
How	Requires a discovery mechanism to locate all versions of a scientific work through a mechanism of linking object identifiers assigned to the different versions of the "same" object.
Issues	Implies mechanisms for metadata linking though ID.

Use case 6: discovery of fragments of the same digital object

Who	A linguist working on audiovisual content
What	Wants to discriminate between different fragments of the same digital object (e.g. audiovisual interview)
Why	To enrich her publications by citing and linking to fragments of an audiovisual work
Where and When	Online Any time
How	Requires a segmentation mechanism to segment a digital object into fragments and connect each fragment to the original object.
Issues	Implies a method to generate "sub"-identifiers for segments directly referring to the ID for the complete object. It may allow to assign accessing restrictions to specific fragments or to the object.

Use case 7: career assessment

Who	A university evaluation committee
What	Wants to check the publications of a candidate
Why	To extract a complete overview of the publications of a given candidate and assemble the candidate portfolio for evaluation purposes, even if the

	candidate in question has worked for several different institutions.
Where and When	Online Any time
How	Requires a service that links an author ID to the candidate's output based on the submitted ID but also on other IDs assigned to the same author
Issues	Implies a method to connect the author ID with other IDs from different schemas Requires a method that allows to check for authenticity of the results (provenance and authoritative links)

Use case 8: deduplication

Who	A repository
What	Wishes to ingest a digital object checking for duplications
Why	To avoid to ingest an exact duplicate of an object already ingested in the same repository
Where and When	Repository/repository federation Any time
How	Through a mechanism which uses semantic discriminative attributes
Issues	Implies a method to connect the author ID with other IDs from different schemas Requires a method that allows to check for authenticity of the results (provenance and authoritative links)

Use case 9: metadata management

Who	A digital repository
What	Metadata management
Why	• Assigning metadata and identifying them through an ID • Changing metadata (deleting the old version) Versioning metadata (preserving lineage)
Where and When	Repository/repository federation Any time
How	Through a mechanism which uses the object ID
Issues	Object ID – metadata ID pair

Use case 10: extracting relations between digital objects, authors and institutions

Who	A scientist
What	Wishes to extract all the significant relations of a given digital object with other digital objects and related entities (e.g. people, institutions)
Why	To contextualize a given work

Where and When	online Any time
How	Through a relationship service
Issues	Starting from the object ID, the service track the links with all the associated entities and uses the IDs of these entities to enrich the relationship tree

4 PI INTEROPERABILITY FRAMEWORK REFERENCE MODEL

The aim of this section is to propose a conceptual framework, i.e. a reference model, addressing the PI interoperability issues, which have been identified in the survey and benchmarking analysis and have been translated into concrete scenarios and use cases to serve as requirements for designing the model. Actually, many of the interoperability needs identified through the scenarios and use cases can be concretely addressed by the services built on top of the IF. However, the focus of the deliverable is not on designing these services, but on the definition of a common framework enabling the implementation of them. A comprehensive definition and design of the interoperability services is the goal of the task 30 of the WP 22.

4.1 WHAT IS A REFERENCE MODEL AND WHY DO WE NEED IT?

As stated in The DELOS Digital Library Reference Model [16], a reference model is "a conceptual framework that aims at capturing significant entities and their relationships in a certain universe with the goal of developing more concrete models of it".

It is worth to note that a reference model should define a common semantics, which is not tied to any specific technological implementations, standards or systems.

The model should include:

1. Concepts (entities) and their properties
2. Relationships between concepts

Combining concepts and relationships between them, we obtain propositions that are statements about the domain. The complete set of these statements represents the knowledge on the domain. This knowledge can be represented graphically through a concept map, which is a graphical tool that represents concepts inscribed in boxes and propositions as directed arrows connecting concepts.

In the context of PI interoperability, the reference model provides a common high-level framework between different PI systems to manage, represent and expose (digital) resources.

The aim is to model the relationships between the relevant entities, which are involved in these processes in such a way that this form of information can be used to support the development of interoperability services and supporting certain processes within these services. The basic idea is that a common conceptual representation is the main prerequisite to design added-value interoperability services, which can exploit the value of a scheme of representation shared and agreed across trusted systems. The framework answers the general question "How to make PI systems interoperable in order to facilitate the exchange, re-use and integration of the resources identified in these systems by different PIs"?

The model proposes a high-level solution for representing digital resources and facilitates access and re-use of these representations beyond the borders of hosting systems, enabling a new generation of cross-systems interoperability services. To this purpose, the model standardizes the relationships between the identified entities (e.g. digital objects, authors, institutions) and their PIs, creating a common layer where meaningful information from independent systems can be exchanged.

The main motivation to define a PI IF reference model is to set the foundations and identify the basic concepts within the universe of PIs systems (and their interoperability), creating the ground for developing appropriate interoperability solutions and interactions with them.

Since it is abstract in nature, the reference model can be used by system designers as a template for designing different technical interoperability solutions and services based on it. These services, we remark, are considered external to the framework in our model.

4.2 THE REFERENCE MODEL

The IF has been designed following 4 main assumptions:

1. In the IF we consider only entities identified by at least one PI.
2. Only PI Domains (PIDs) that meet some criteria **Trusted** for the IF.
3. We delegate the responsibility to define relations among resources to Trusted PIDs.
4. We don't address digital preservation issues but it is demanded to the PID management

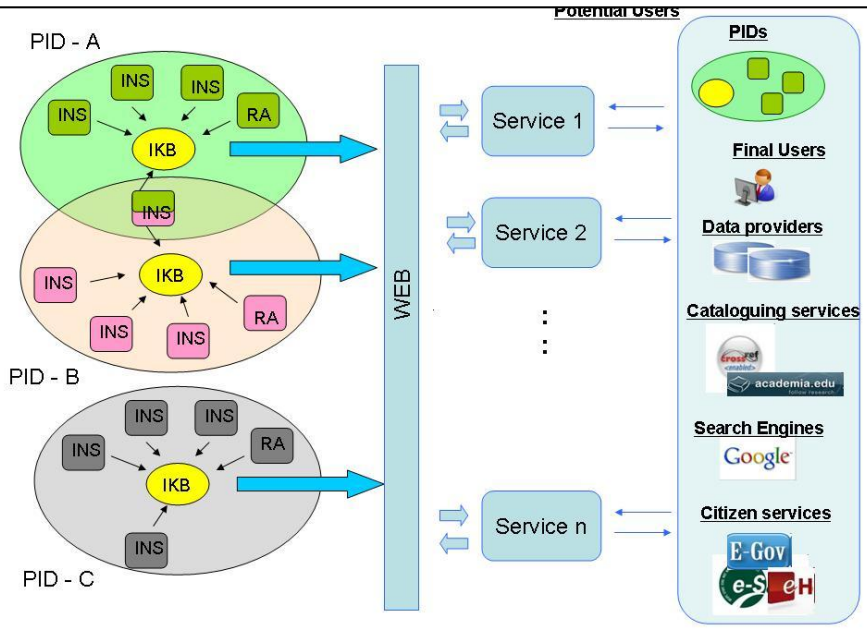


Figure 1: A graphical representation of the IF

The key actors in the IF are the so-called PI Domains (PIDs) that include in our definition:

1) The systems (namely Registration Agencies or RAs), which manage the allocation and registration of PIs and provide the necessary infrastructure to allow the registrants to declare and maintain their metadata. Since we consider only trusted PIDs, the services provided by these systems are expected to encompasses quality assurance measures to guarantee fundamental requirements such as, for example, the uniqueness of PIs within the domain or the accurate updating of the related data.
   We limit to only 3 types of PIDs based on the three different types of identified entities:
   a. PID for digital objects
   b. PID for authors
   c. PID for institutions
2) The content providers (INS in Figure **1**) that are the institutions responsible for storing, managing and preserving the access to digital contents through the use of PIs.
3) The resolver is the system that creates the link between a PI and information about the object and its current location.

According one of the main assumptions stated above, only trusted PIDs can join the framework and populate the scenario with their entities. In our model, for a PID to be trusted it must respect the benchmark criteria described in the next section. Consequently, any object, person or institution without a PI assigned by a trusted PID cannot enter in our scenario as 'entity' but, in case, only as descriptive information with no property or relation to other entities.

It is important to notice, for the purposes of the present work, that the user community board managing the PID is responsible for guaranteeing suitable policies for any aspect of the Digital Preservation Plan underpinning that system, like for example, the content selection/granularity criteria (included the FRBR 3 levels), the Trusted Digital Repositories policies and certification, the trustworthiness of the PI management, and so on.

Moreover, within each PID there can be different approaches and architectures to share roles and responsibilities among different components of the system, like the Registration Authority (RA), the Certification Authority (CA), the domain resolver, the digital repository curator and content holders, the Digital Preservation (DP) manager, and so on. The user community is free to choose the best solution and we trust them for the correctness that we assume.

The IF does not argue about DP policies and it assumes correct any statement coming from a trusted PID. The framework provides a shared conceptual infrastructure to represent the identified entities and their relations within what we call an Interoperability Knowledge Base (IKB), assuming this declared information as guaranteed by trusted PIDs.

The interoperability scenario, as we said, deals with at least 3 types of PID with specific properties for any entity. Then, we forecast some possible relations between two or more objects, or between objects and persons or institutions and their PIs. These relations must be provided by the PIDs when they bring an entity into the interoperability knowledge base. In particular, some trusted PIDs will populate the IKB with their entities presenting these contents following an API so providing specific info requested by the IF. For any digital object the PID, in addition to some descriptive metadata, should declare existing PI (e.g., DOI, NBN,…), any relation with other objects within the domain and any PI for persons or institutions known by the PID.

In this way, the IKB defines the fundamental relations between the entities in play in the domain (e.g. between objects and PIs), creating a layer of accessible knowledge on which interoperability services can be built thanks to the explicit representation of these relations (see Figure 2). Indeed, the knowledge generated independently by the trusted PIDs using the framework, will be exposed on the Web with a common semantics and format, becoming available to third parties in order to implement interoperability services by exploiting this knowledge in many different ways. In this way, the contents are visible through a common interface provided by the IF and populated by the trusted PIDs and users can create services accessing all the domains and using all the contents even if they are from different PIDs (see the use cases described before in this document) overcoming in effects a relevant limit in the current situation. The **Figure** 2 shows also that institutions that adopt more then one PI systems for their resources, for instance DOI and NBN, contribute to the IKB of the DOI PID and NBN PID with the same relation statements. Thus, IKBs present some overlapping (in Figure 2 is represented by overlapping area between PID- A and PID-B) that can be exploited as a bridge to walk across PIDs and enabling new services to discover new relationships and make inferences on digital resources.

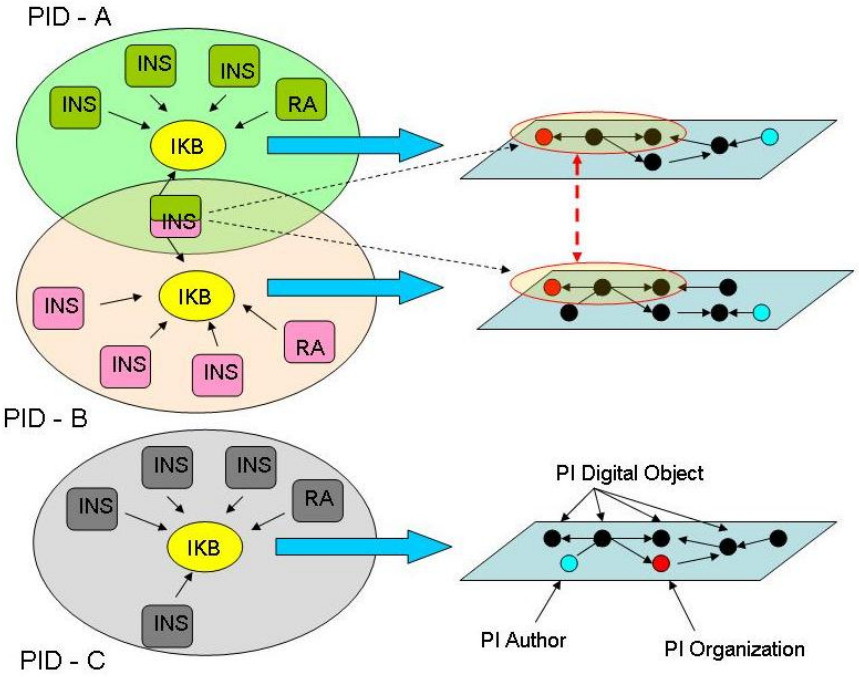


Figure 2: Significant relations established through the IF across the PID boundaries.

For example, a service for providing the complete list of publications of a given author, can exploit the network of relations between the PI of the author and the PIs of his publications, aggregating and matching the knowledge from different PIDs.

Moreover, we envision that other services could also be specifically designed to extract other kinds of relations between entities and PIs, (i.e. relations not yet explicit within a PID) by using different techniques (e.g. reasoning on metadata) and provide this information (for example in the form of probabilistic relations) to the PIDs which could use it to update their explicit relations in the framework.

The definition and the design of some of these services, which can be built on top of the IF, will be the focus of the second part of the wp22 (Task 30).

4.3 THE MAIN CONCEPTS

In this section we provide a description of the main concepts and relationships included in the framework, according to the assumptions described above.

4.3.1 Resource

Definition: A Resource is one of the most primitive concepts in the IF reference model and covers any entity that can be identified by at least one PI. Entities, which are not assigned to a PI, are not eligible for the IF. A resource is a representation of a physical or an abstract entity.

Since the concept of resource can be very different in different PID, we propose on purpose a very general definition, which encompasses the diverse range of digital resources, including such resources as objects, annotations, and metadata.

We consider three main kinds of resources in the framework:

1. Digital Objects
2. Authors
3. Institutions

Other kinds of resources can be included in future with the development of PI systems dealing other types of entities, such as events, locations and so on.

4.3.2 Digital Object

Definition: A Digital Objet is any kind of digital resource, which is identified by at least one PI assigned by a trusted PID. We don't provide a more specific definition because we rely on the definition provided by the trusted PID which has assigned the PI to the resource. Any digital object with no PI is not relevant in the IF.

All the different types of digital objects share some fundamental properties, which link digital objects to other resources or concepts. Each digital object may have the following properties:

Relationship:

1. Is identified by at least one digital object PI (<hasPI_do)
2. Is created by an Author (<created_by>);
3. Is related to other digital objects (<related_to>)
4. Is associated to Policies (<associated_to>);
5. Can be described by metadata (<has_metadata[9]>);

The first property specifies the inescapable requirement for a digital object to be included in the IF that is the fact to be identified by a PI assigned by a trusted PID.

Different PIDs can have a different conceptualization of what a digital object is and, consequently, different assignment criteria for PIs. For example, a PI can be assigned at the conceptual level of an object, at the manifestation level or at the bitstream of it. It is important to notice that these differences can have a relevant impact on the development of interoperability services, but we decided to not include this differentiation within the framework for the purpose of simplicity.

The fact that a digital object can be identified by multiple PIs is one of the most important pillars for the IF since the association of a digital object to PIs from different PIDs allows to make explicit that these PIs refer all to the same entity.

The second property connects a digital object with its creator, which must be in turn identified by at least one PI.

A digital object can be also related to other digital objects in different ways, depending on the representation scheme used by the PID. For example, a digital object can be part of a complex object (e.g. multimedia object, collection) or can be related with its previous versions.

Finally a digital object is regulated by policies, such as policies for access, citability or re-use. Making explicit the policies which regulate the entire lifecycle of a digital object is essential in order to implement some services, like, for instance, knowledge discovery services which depends on the accessibility of the resource regulated by access policies.

4.3.3 Author

Definition: An Author is a physical entity, which is the creator of a digital object and is identified by at least one PI assigned by a trusted PID. Differently from digital objects which are digital in nature, authors are physical entities which are represented through descriptions (i.e. profiles) in the digital world. Therefore, while a PI for a digital object can point directly to the object, a PI for an author does not point to the author but always to a description of him. Moreover the resource, which describes an author, is expected to change as the referent inherently changes across time. Therefore, "the sameness" property of a PI for an author means referring to the "same physical entity" (i.e. the same author and not the same unchanged digital resource), while that of a PI for a digital object means referring to the "same digital entity" (i.e. the same unchanged digital resource).

Relationships:

1. Is identified by at least one author PI (<hasPI_author>)

[9] Metadata is the generic term OAIS has a finer granularity including Representation Information, PDI, Packaging Information etc

2. Is the creator of a digital object (<creator_of>)
3. Is affiliated to an organization (<affiliated_to>)
4. Is the contributor of other digital objects (<contributor_of>)
5. Can be described via an author profile (<has_author_profile>)
6. Can be described by metadata (<has_metadata>)
7. Is regulated by policies (<regulated_by>)

Through the described properties an author is connected with digital objects and organizations. This starts to define a network of relationships between the entities of the framework, which can be exploited to implement the interoperability services.

4.3.4 Institution

Definition: An Institution is a physical entity, which affiliates authors and other human agents and is identified by at least one PI assigned by a trusted PID.

Relationships:

1. Is identified by at least one institution PI (<hasPI_inst)
2. Is the affiliation of an author (<is_affiliation_of>)
3. Can be described by metadata (<has_metadata>)
4. Is regulated by policies (<regulated_by>)

4.3.5 PI (PI)

Definition: a PI is a character string used to uniquely identify a resource within a PID regardless of where the resource is located. In the framework we distinguish between 3 kinds of PI, as shown in

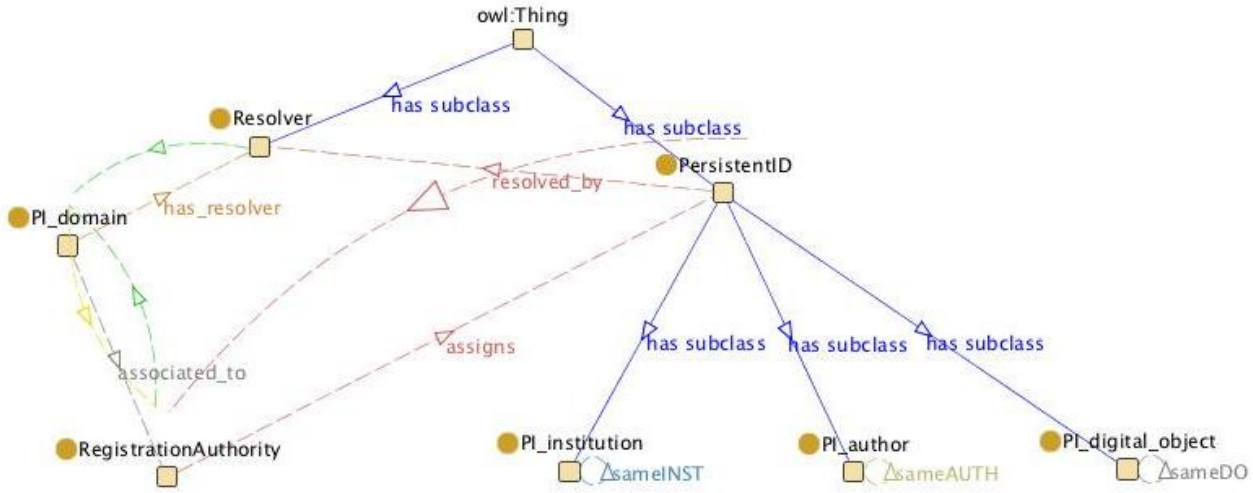


Figure 3:

1. PI for digital object
2. PI for author
3. PI for institution

Digital Object PI relationships:

1. Is the PI of a unique digital object (<is_PI_do_of>)
2. Can refer to the same digital object of another PI (<same_do>)
3. Is assigned to a digital object by a trusted Registration Authority belonging to a trusted PID (<assigned_by>)
4. Is resolved by a resolver (<resolved_by>)

Author PI relationships:

1. Is the PI of a unique author (<is_PI_author_of>)
2. Can refer to the same author of another PI (<same_author>)
3. Is assigned to an author by a trusted Registration Authority belonging to a trusted PID (<assigned_by>)
4. Is resolved by a resolver (<resolved_by>)

Institution PI relationships:

1. Is the PI of a unique institution (<is_PI_inst_of>)
2. Can refer to the same institution of another PI (<same_inst>)
3. Is assigned to an institution by a trusted Registration Authority belonging to a trusted a PID (<assigned_by>)
4. Is resolved by a resolver (<resolved_by>)

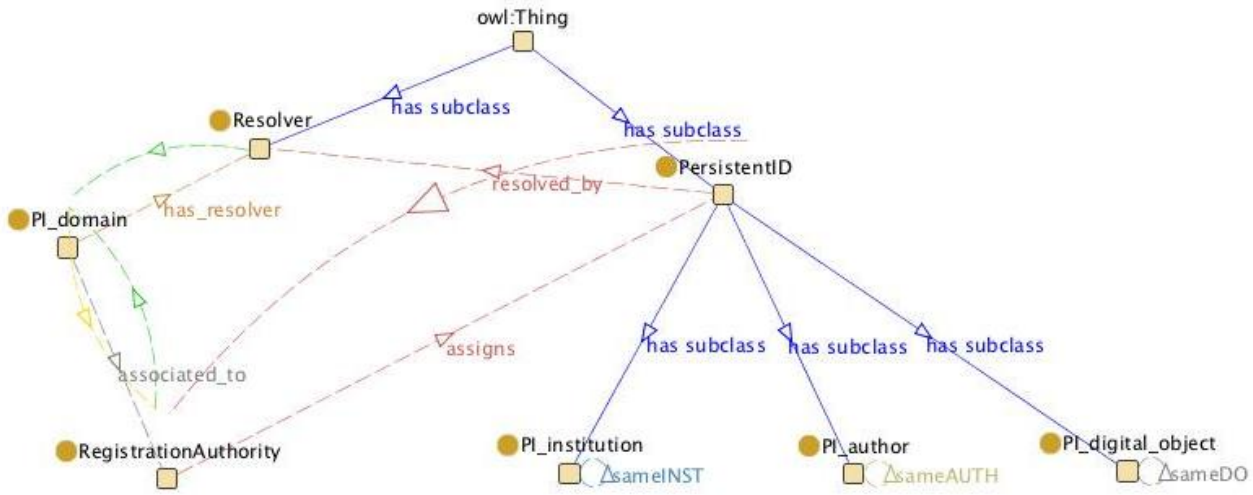


Figure 3: PIs and their fundamental relations

4.3.6 PI Domain

Definition: a PI Domain, or PID is a system of users and service providers, which manages the assignment of PIs for any type of relevant entities (e.g. digital objects, authors, institutions). Typically, these types of systems are different for different communities and specific for types of objects. PIDs must be trustable in a very long term vision. We trust PI domains implementing rules and strategy needed for a correct digital preservation.

Relationships:

1. Has a resolver (<has_resolver>)

2. Has at least one Registration Agency (<has_RA>)
3. Defines policies (<defines_policies>)

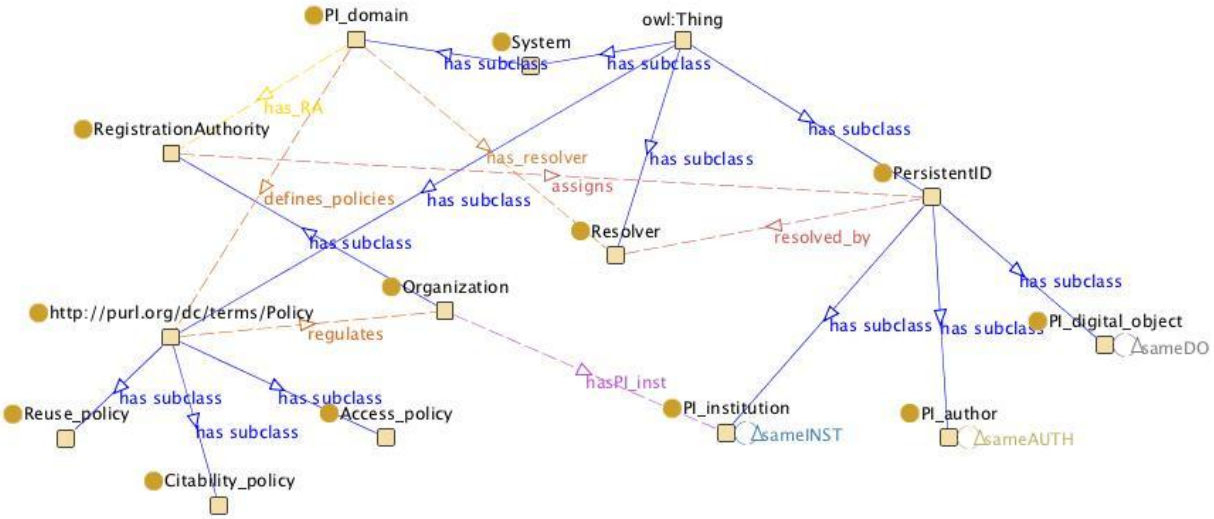


Figure 4: PID and its fundamental relations

4.3.7 Policy

Definition: the concept represents the set of conditions, rules, restrictions, terms and regulations governing the entire life cycle of a digital resource and its management within a trusted system. This domain is very broad and dynamic by nature. The representation provided by this framework does not pretend to be exhaustive especially with respect to the myriad of specific rules each PID would like to model and apply. The concept of policy captures the minimal relationships connecting it to the other relevant entities in the framework. The model is extensible and other subclasses of policies could easily be added in future.

Policies can be very different in different PIDs. Examples of policies are:

1. Policies for access
2. Policies for citability
3. Policies for re-use

Relationships:

1. Regulate a digital object (<regulate_do>)
2. Regulate an author (<regulate_author>)
3. Regulate an institution (<regulate_institution>)
4. Are defined by a PI domain (<defined_by>)

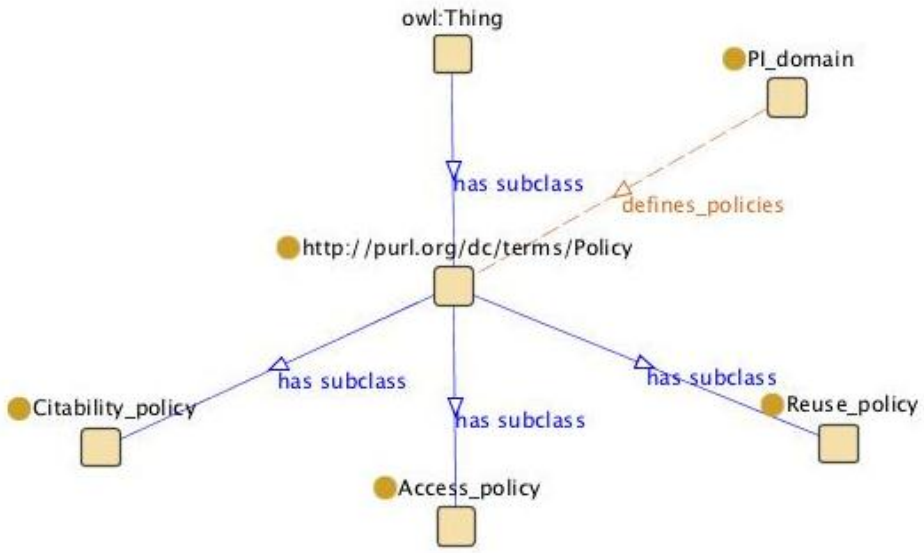


Figure 5: Policy and its fundamental relations

4.3.8 Resolver

Definition: A Resolver is a system that creates the link between a PI and information about the object and its current location of the associated object.

Relationships:

1. Is identified by a unique URL (<identified_by_URL>)
2. Belongs to a PID (<belong_PID>)

4.3.9 User/Actor

Definition: An Actor is an entity that is external to the interoperability system and interacts with it and uses the related services. Both humans and inanimate entities such as software programs can be users. A user can have the following properties:

Relationships:

1. Is identified by a user identifier (<hasUserID>)
2. Is characterized by a user profile (<hasUserProfile>)
3. Is characterized by user policies (<hasUserPolicies>)
4. Interacts with the interoperability system and services (<perform>)

4.3.10 Conceptual Map

A graphical representation of the concepts and relations of the IF is presented in Figure **5**

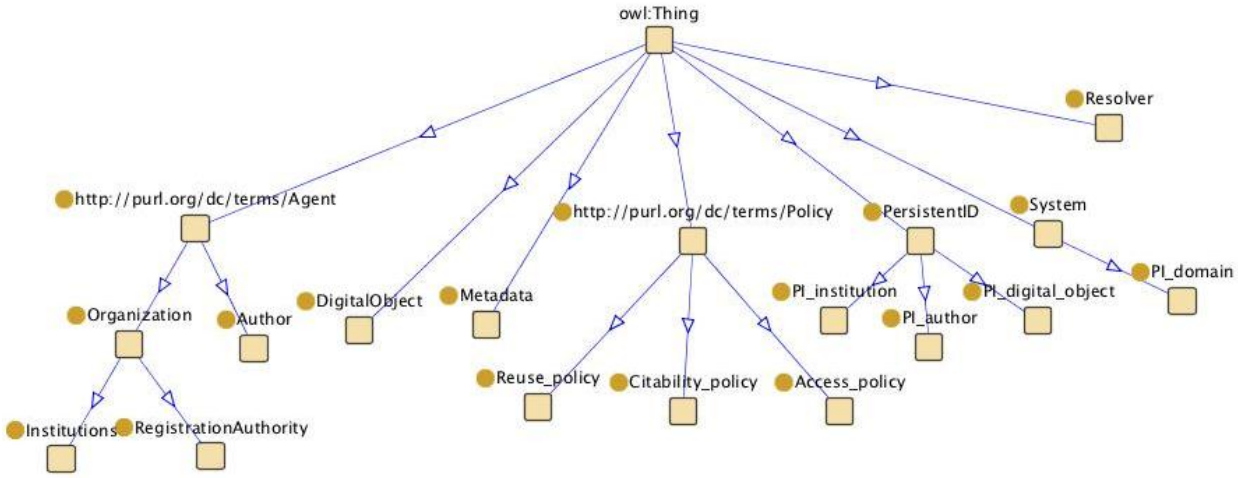


Figure 5: Conceptual map of the IF.

The Figure 6 shows which are the relations among Digital Objects, Author, and Organization entities through the PIs for Digital object, for authors and institutions.

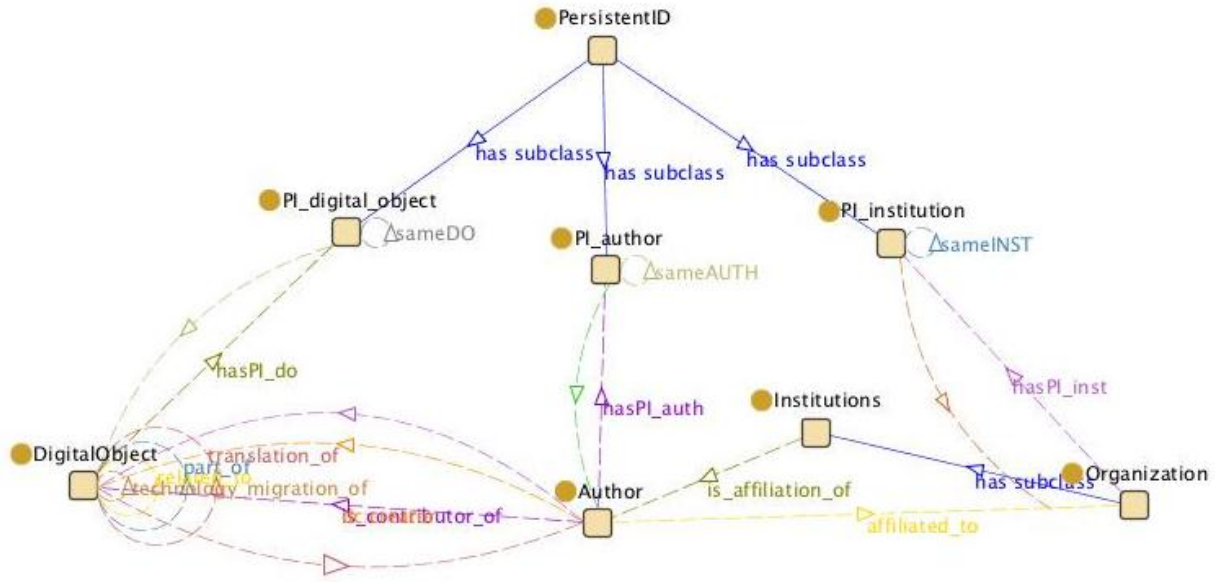


Figure 6 Relations among Digital Object, Author and Organization entities.

5 PIs INTEROPERABILITY: RELATED PROJECTS

Recently, several initiatives and projects have started to address the problem of PI interoperability and solutions have been proposed in different contexts facing some issues at identifier or metadata levels. In this section, we provide a brief description of these projects and initiatives, exploring some commonalities and differences. A description and a benchmark analysis of the main PI systems for digital objects and authors will be presented in Section 5.

The panorama that we are going to describe is quite diversified, but some distinctions can be introduced to clarify the status quo.

A first distinction can be made between national and international initiatives. Some initiatives have been emerged within a national context (e.g. LATTES in Brazil, PILIN in Australia) and some of these started as funded project on a broader geographical level (e.g. PersID). Other initiatives show their presence at international level (such as ORCID) and aim at introducing global standards for identification, creating a consortium of participating organizations. We can also distinguish initiatives limited to a specific discipline (e.g. for linguistic resources) or more generic (e.g. OKKAM).

Some projects focus exclusively on the problem of PI interoperability for digital objects (e.g. PILIN), while other initiatives address the interoperability issue for author identifiers (e.g. ORCID).

The diffusion of a given initiative can also be determined by the way in which the identifiers are assigned by the underlying ID management systems. Some governmental initiatives limit the assignment to people, that embark on an academic career (e.g. LATTES), while other systems allow the registration of any kind of entity (e.g. OKKAM).

In the following table we have classified the initiatives based on the following set of features:

- **Type**: denotes the type of initiative, in the selected initiatives it can be a consortium, it can be promoted by an institution or it may be the result of a project;
- **Identifier scope**: tells if the initiative is related to digital object identifiers (D), author identifiers (A), both (A+D) or metadata (M).
- **Funding:** indicates that the initiative is sustained by government funding.
- **Participation**: refers to the geographical extension of the initiative.

Name	Type	Identifier Scope	Funding	Participation
ORCID	Consortium	A	Private	European
PERSID	Project	A+D	Surf Fundation	European
OKKAM	Project	A+D	European Commission	European
PILIN	Project	D	University	Australia
CORES	Project	M	European Commission	European
RIDIR	Project	D	JISC	UK/European
LATTES	Institution	A	Gov	National Council Brazil

Table 1: Initiatives on identifiers interoperability

ORCID (Open Researcher and Contributor ID) www.orcid.org

ORCID stands for Open Researcher and Contributor ID. It is an initiative of key stakeholders in scholarly communication: universities, funding agencies, publishers, research institutes. It aims to build a registry of ALL active researchers: a permanent, clear and unambiguous record of scholarly communication by enabling reliable attribution of authors and contributors. ORCID will start from deposition of information from the researchers and contributor institutions, or other third parties. These profiles will contain bio/bibliographic information. Researchers will be able to create, edit, and maintain an ORCID ID and profile free of charge and will control the defined privacy settings of their own ORCID profile data. The infrastructure is designed to be interoperable with, and leverage and complement the power of, all related PI projects. ORCID is backed by over 250 organisation worldwide, who see value in their operations by the existence of such an enabling infrastructure: universities will be able to easily collate their scientific production as a whole, funding agencies will be able to trace the results of their investment, publishers will be able to enrich their submission workflows. ORCID is not limited to authors: eventually the emerging data infrastructure will also benefit from the initiative and enable unique association of persons, roles, and datasets.

PersID

The PersID project [12] aimed to support permanent access to scholarly and cultural information. PersID embraced the URN:NBN scheme for their identifiers with the objective of creating a global resolver infrastructure among NBN national domains, such as . It builds upon proven technologies and standards already in wide use, the IETF RFC3188.The initiative consists of diverse organizations that have a long-term responsibility for large amounts of publications, cultural materials and research data. There are ten project partners, mostly national libraries. PersID policy demands that the partners will ensure long term preservation of the referent of the identifiers. PersID was a project funded by SURF Foundation

OKKAM

The OKKAM project created an infrastructure called ENS (Entity Name System) for the systematic re-use of global identifiers for entities. The ENS allows to assign a global identifier to a given entity and to link this entity with other alternative identifiers. The underlying idea is that the same entity can be identified with different identifiers in different contexts. OKKAM creates its own identifiers as global identifiers to which there is a profile attached for allowing disambiguation. Part of the profile is a list of identifiers created with other systems. In this way, given an OKKAM ID or any other id for a referent it is possible to gather an entry point to the referent itself; in fact the resolver capability is left to the system which maintains the referent itself, while OKKAM provides a resolver for the global OKKAM identifiers which return the referent profile.

PILIN project [8]

The PILIN (PIs and Linking Infrastructure) project was an important national initiative led by ARROW and the University of Southern Queensland; it ended in 2007. The aim of the PILIN Project was to build a sustainable, shared identifier management infrastructure based on CNRI Handle technology to underpin sustainable global identifier infrastructure enabling persistence of identifiers and associated services. The main focus of the project was to meet the specific need, common to Australian e-learning, e-research, and e-science communities, of a sustainable identifier infrastructure to deal with the vast amount of digital assets being produced and stored in the different e-science environments.

CORES project [7]

The CORES project is a European project on metadata interoperability. The central objective of the CORES project is to encourage the sharing of metadata semantics. By creating consensus on a data model for declaring semantics of metadata terms, CORES aims at enabling existing standards to work together in an integrated, machine-understandable Semantic Web environment. To achieve this result the project started a Standards Interoperability Forum bringing together key figures from major

standardisation activities to discuss the practicalities of cross-standard interoperability. One of the main outcomes of the project is a registry environment for declaring and sharing metadata schemas based on a common model.

RIDIR (Resourcing IDentifier Interoperability for Repositories) project [9]

RIDIR is a project funded under the auspices of the Joint Information Systems Committee 'Repositories and Preservation' Programme. It is investigating the requirements for, and benefits of, the clear use of PI in order to facilitate interoperability between digital repositories of different types. The RIDIR project's main objectives were to engage with the identifier and repositories communities to understand their requirements and to build a fully working demonstrator, and to raise awareness of PI interoperability issues. The project was not about shared PI services themselves as such services would be available, but was focused on the use of PI. As a demonstration, they have created a Lost Resource Finder service, which allows a user to be redirected to the new, correct location for a resource when a PI is broken. The new location can either be specified by a repository manager, or it can be based on crowd intelligence – based on other users of the service searching for and discovering the new location of the resource within the service.

LATTES

LATTES is a governmental initiative, promoted by National Council for Scientific and Technological Development (CNPq) in Brazil. The initiative promotes the creation of a curricular information system (named LATTES CV System) that comes from a series of agreements between CNPq and institutions that have databases and Web sites for publishing information in science and technology. The aim of the system is to collect information about all institutional actors involved in scientific and technological development. The system is used for 1) the evaluation of competences of candidates in relation to scholarships and/or research support; 2) the selection of consultants, members of committees and advising groups; 3) the subsidy of the evaluation of the Brazilian post-graduation and research.

6 PI SYSTEMS BENCHMARKING

There are several ongoing PI initiatives that provide an online identification service for digital resources with the "persistence" statement in the service description. At a first glance, these PI services seem very similar, with a high level of overlapping in functionalities and type of provided service, but each of them has its own peculiar features and addresses requirements of different communities. Thus, in order to determine which of them can be consider eligible for the IF, it is necessary to analyse their features - which have been investigated during the feature analysis (see Annex II) - against the criteria which we considered basic prerequisites to set up an IF among these systems. In fact, it is necessary to include only the PI systems that support the features expected by the community addressed by the present work.

In this section the results of the benchmarking assessment are reported, distinguishing between PI systems for digital objects and PI systems for authors/creators. Since PI systems for organizations are at a very immature stage of development and are still scarcely adopted, we limited the analysis to PI systems for digital objects and authors.

6.1 BENCHMARK CRITERIA ASSESSMENT

In order to apply the criteria identified in Section 3.3, we first introduce the assessment method adopted to identify which are the PI systems that can be judged trusted and thus eligible for the IF.

1.Having at least one Registration Agency (RA)

This criterion can be assessed verifying the existence of a RA in the PID

$$RA=\begin{cases}\text{Yes}:\text{if the RA exists}\\ \text{No}:\text{otherwise}\end{cases}$$

2.Having one Resolver accessible on the Internet

This criterion can be broken down into the following sub-criteria:

2.a The existence of an accessible URL of the PI resolver:

$$\text{R-URL}=\begin{cases}\text{Yes}+\text{URL}:\text{if the resolver is availabe on the web}\\ \text{No}:\text{otherwise}\end{cases}$$

2.b Identifying the types of Resolution returned by the resolution process:

$$RO=\begin{cases}\text{O}:\text{Only Resource}\\ \text{M}:\text{Only Metadata}\\ \text{OM}:\text{Resource}+\text{metadata}\\ \text{MO}:\text{Multipleobjects}\end{cases}$$

For the Author PI system

$$RO=\begin{cases}\text{R}:\text{Resource (author information)}\\ \text{RO}:\text{Resource}+\text{Objects associated (articles, etc)}\end{cases}$$

3. Uniqueness of the assigned PIs within the PI domain

This criterion can be referred to the capability of a PI system to avoid duplications in name generation:

In particular the uniqueness can be assured if the PI string includes the authority namespace like: NBN:IT:UNIBO-1234. Instead if a PI is generated as a simple number like: 12345 the uniqueness can be easily broken. Moreover, if the resolver URL is not explicated or the descriptive label associated to the numeric string is lost, it is not possible to understand the domain of reference of the ID.

$$\text{UNI}=\begin{cases}\text{Yes}: \text{if the uniqueness is guaranteed}\\ \text{No}: \text{otherwise}\end{cases}$$

4. Guaranteeing the persistence of the assigned PIs

This criterion can be broken down into the following sub-criteria:

4.a The PI system has a functionality that allows the identifier string modification:

$$\text{MOD}=\begin{cases}\text{Yes}: \text{if the PI string modification is allowed}\\ \text{No}: \text{otherwise}\end{cases}$$

4.b PI system has a functionality that allows PI deletion. In this criterion we do not take into account if the delete function is allowed only under certain conditions. Thus we include these cases into positive cases as defined here below:

$$\text{DEL}=\begin{cases}\text{Yes}: \text{if the PI deletion is allowed}\\ \text{No}: \text{otherwise}\end{cases}$$

4.c The persistence of the PI is strictly related also to the sustainability of the PI service. Indeed it is difficult to evaluate objectively the credibility of a business model or cost model implemented by a PI system but a possible indicator useful to highlight a potential sustainability issue can be given by the following matrix:

Type of Initiative (INI)	Payment expectations
a) PI **Public** (P) initiative *requires* a **political** commitment	a fee might be not required (No)
b) PI **Not for profit** (NP) initiative *requires* at least a **cost model**	a fee should be required (Yes fee)
c) PI **Business** (B) initiative *requires* a **business model**	a fee must be required (Yes fee)

$$\text{INI}=\begin{cases}\text{P}: \text{Public}\\ \text{NP}: \text{Not for Profit}\\ \text{B}: \text{Businness}\end{cases}$$

$$\text{FEE}=\begin{cases}\text{Yes}: \text{if a fee is required}\\ \text{No}: \text{otherwise}\end{cases}$$

5. User communities, which implement the PID, should implement policies for digital preservation

This criterion can be broken down into the following sub-criteria:

5.a The PI system requires the adoption of Trusted Digital Repository (TDR) assessment criteria to a content provider before of providing access to PI service for its resources.

$$\text{TDR}=\begin{cases}\text{Yes}: \text{if the adoption of TDR criteria doption is required}\\ \text{No}: \text{otherwise}\end{cases}$$

5.b The PI system requires the adoption of digital preservation criteria to a content provider before of providing access to PI service for its resources.

$$\text{DP}=\begin{cases}\text{Yes}: \text{if digital preservation strategies is required}\\ \text{No}: \text{otherwise}\end{cases}$$

NB: These criteria (5.a and 5.b) are not applicable to the PI systems for the authors

6. Reliable resolution

This criterion can be assessed analysing the PI service policy. Thus, if a policy is stated in the PID to guarantee a reliable resolution to *the same* objects along the time, the criterion is satisfied.

The concept of *the same* relation is left to the user community managing the PID.

$$\text{RR}=\begin{cases}\text{Yes}: \text{if the PI system defines a proper policy}\\ \text{No}: \text{otherwise}\end{cases}$$

7. Uncoupling the PIs from the resolver

This criterion can be assessed analysing the syntax of the PI and its technical specification. Thus if PI string definition includes the URL of the resolver or content provider or something else that cannot directly referred to the identifier itself means that the uncoupling is not supported.

$$\text{UNCOUP}=\begin{cases}\text{Yes}: \text{if the PI is uncopuled from the resolver}\\ \text{No}: \text{otherwise}\end{cases}$$

8. Managing the relations between the PIs within the domain

This criterion can be assessed analysing the level of service provided by the PI system. As already stated above, this criterion is optional since the presence of the logical relations between digital resources or a resource and its parts is an added value for the IF but it is not considered a mandatory requirement in the present assessment. These relations can be defined with metadata or included in the PI syntax, but the specific implementation is not important for the assessment. This criterion can be broken down into the following sub-criteria:

8.a The PI system manage the relations between digital resources

$$\text{RE-TO-RE}=\begin{cases}\text{Yes} + \text{method}: \text{if the PI system mmanage the resorice to resource relation}\\ \text{No}: \text{otherwise}\end{cases}$$

8.b The PI system manage relations between a digital resource and its parts

$$\text{PART-OF}=\begin{cases}\text{Yes}+\text{method}:\text{if the PI systemnmanage the part of relation}\\ \text{No}:\text{otherwise}\end{cases}$$

This criterion is considered not applicable for the PI systems for Author

6.1.1 PI systems for Digital Objects benchmark

The feature analysis results (see Annex II) highlight some differences among the systems that determinate their level of service and trustworthiness. These systems address different requirements according to their designated community,[10] but in some cases, these communities have requirements which appear too different respect to those considered by the present analysis. In particular, some PI initiatives or systems that aim to provide a PI service does not fit with the Trusted PIs definition adopted in this work because of the level of service required by the served community is different. These systems are out of the scope of the following analysis

Starting form the results obtained from the feature analysis in the Annex II, we have selected mostly the same PI systems and we have assessed these systems against the criteria defined above. This assessment allowed us to identify which are the systems that can contribute to the IF in a trustable way.

Digital Object Identifier

- **RA**: There are several RAs in the PID such as MEDRA, CrossRef, ISTIC, and so forth[11].
- **R-URL**: http://dx.doi.org/
- **RO**: DOI can resolve to metadata and digital objects or to multiple objects [12].
- **UNI**: The uniqueness of the DOI is assured.
- **MOD**: The DOI string modification is not allowed.
- **DEL**: The DOI deletion is allowed [13].
- **INI**: The DOI is a no profit foundation but the RAs are business-oriented institutions (for instance **MEDRA** is an enterprise).
- **FEE**: a fee is required.
- **TDR**: It is not required any particular requirements to the content provider. The object identified can be stored everywhere[14].
- **DP**: It is not required any particular digital preservation strategies to the content provider for its digital resources.
- **RR**: A policy is defined to maintain the proper level of similarity every time an object identified by a DOI is updated.
- **UNCOUP**: the DOI string is not built including the URL of the resolver. For instance in the URL http://dx.doi.org/10.123/456 is the DOI is represented by the string:10.123/456 456 and is uncoupled by the URL of the resolver.
- **RE-TO-RE:** The DOI system allows the relation among digital objects through ONIX metadata schema (for instance 82 MMP.51Relation code[15])

[10] OAIS Designated community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.

[11] http://www.doi.org/registration_agencies.html

[12] http://www.medra.org/en/MR.htm

[13] http://admin.doi.org/DOI/admin_help.html

[14] http://www.medra.org/en/terms.htm

- **PART-OF:** The DOI system allows to make explicit of the dependency between the object and its sub-parts through ONIX metadata schema (for instance 81 MMP.51Relation code[16])

Handle

- **RA**: The Global Handle Registry is managed by the Corporation for National Research Initiatives (CNRI).
- **R-URL**: http://hdl.handle.net/
- **RO**: Handle can resolve to metadata and digital objects or to multiple objects [17].
- **UNI**: the uniqueness of the Handle is assured.
- **MOD**: The Handle system allows the update of the PI.
- **DEL**: The Handle system allows the deletion [18]of the PI.
- **INI**: The CNRI is a not-for-profit organization.
- **FEE**: a fee is required[19].
- **TDR**: It is not required any particular requirements to the content provider. The object identified can be stored everywhere[20].
- **DP**: It is not required any particular digital preservation strategies to the content provider for its digital resources.
- **RR**: Not enough information available to assess this criterion
- **UNCOUP**: the Handle name is not built with the URL of the resolver. For instance in the URL http://dx.doi.org/10.123/456 is the DOI is represented by the string:10.123/456
- **RE-TO-RE:** The Handle system does not allow the relation among digital objects through metadata **PART-OF:** The Handle system does not support the explication of the dependency between the object and its sub-parts

ARK

- **RA**: Each Institution has the responsibility of its ARK identifiers. A central registry with the NMAH registration is maintained by the CDLIB.
- **R-URL**: In the ARK specification, the resolution service is represented to the NMAH. In the URL http://rutgers.edu/ark:/12025/654xz321 the NMAH is: http://rutgers.edu/
- **RO**: ARK supports the multiple resolutions.
- **UNI**: The uniqueness of the ARK is assured.
- **MOD**: There is no explicit information about that. Therefore we assume that an ark cannot be modified.
- **DEL**: The is no explicit information about that. Therefore we assume that an ark cannot be deleted.
- **INI**: It is a public initiative.
- **FEE**: The system requires a cost of implementation and maintenance sustained by the content providers itself.

[15] http://www.medra.org/stdoc/en/041110_monographic_doi_metadata_p.pdf
[16] http://www.medra.org/stdoc/en/041110_monographic_doi_metadata_p.pdf
[17] http://www.handle.net/faq.html
[18] http://www.handle.net/hs_manual/server_manual_5.html
[19] http://www.handle.net/registration_agreement.html
[20] http://www.handle.net/hs_manual/server_manual_5.html

- **TDR** and DP: It is not required a particular policy to the content providers to assign an ark to its resources but the system encourages the adoption of policies and guidelines to ensure the stability of the objects identified[21]
- **RR**: an update of the object is considered a new object and receives a new ARK identifier. The **versioning** can be expressed in the identifier itself.
- **UNCOUP**: In the following: http://rutgers.edu/ark:/12025/654xz321 the ark identifier is: ark:/12025/654xz321.
- **RE-TO-RE**: ark:/12025/654.20v.78g.f55; ark:/12025/654.321xz; ark:/12025/654.44 are variants of each other.
- **PART- OF**: the ARK system manages the "part of" identification as follows: the ARK of the form ark:/12025/654/xz/321 is equivalent to publishing three ARKs: a)ark:/12025/654/xz/321, b)ark:/12025/654/xz, c)ark:/12025/654 together with a declaration that the first object is contained in the second object, and that the second object is contained in the third[22].

DataCite

- **RA**: DataCite registration agency http://.datacite.org
- **R-URL**: See DOI definition
- **RO**: See DOI definition.
- **UNI**: See DOI definition.
- **MOD**: See DOI definition.
- **DEL**: The DOI generated by DataCite cannot be deleted, but it is possible to deactivate the dataset associated to a certain DOI[23].
- **INI**: It is a not for profit organization.
- **FEE**: A fee is required.
- **TDR+DP**: To obtain a DataCite account a content provided has to guarantee that the datasets are stored and managed "such that persistent access to the data can be provided"[24].
- **RR**: The DataCite identifies stable dataset deposited on the data archives.
- **UNCOUP, RE-TO-RE and PART-OF:** the identifiers generated by DataCite are DOI, therefore, the consideration done for DOI in general are here valid for DataCite as well.

NBN/DNB

- **RA**: The RA is the German National Library.
- **R-URL**: http://nbn-resolving.de/
- **RO**: The NBN points to the digital object directly.
- **UNI**: the uniqueness of the NBN is assured.
- **MOD**: The NBN modification is not allowed.
- **DEL**: The NBN deletion is not allowed.
- **INI**: This is a public/government initiative.
- **FEE**: The service is free of charge for the content providers.

[21] https://confluence.ucop.edu/display/Curation/CDL+Policies
[22] http://tools.ietf.org/html/draft-kunze-ark-15
[23] https://mds.datacite.org/static/apidoc
[24] http://cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/datacite-canada/info-prospective-clients.html#q12

- **TDR**: The URN object are items archived by German National Library and "objects which will be administrated on certified document servers (i.e. endeavours by DINI) […]"[25].
- **DP**: "[…] with a perspective to making them permanently available via long term archiving"[26].
- **RR**: Every change applied to a resource generates a new resource that receives a new NBN.
- **UNCOUP**: The NBN syntax does not include the URL of the resolves. Thus a valid s NBN syntax is: URN:NBN:DE:GBV:089-3321752945.
- **RE-TO-RE**: For instance the versioning is not managed.
- **PART-OF**: The NBN:DE service uses the fragment in the syntax to identify the part of a resource.

NBN/IT

- **RA**: The RA is the Biblioteca Nazionale Centrale di Firenze (BNCF).
- **R-URL**: http://nbn.depositolegale.it
- **RO**: The NBN resolves to an intermediate page with metadata and digital objects.
- **MOD**: The NBN string modification is not allowed.
- **DEL**: The NBN deletion is not allowed.
- **INI**: This is a public/government initiative.
- **FEE**: The service is free of charge for the content providers.
- **TDR**: The NBN is assigned to those re sources of content providers who signed an agreement for legal deposit of their content to the National Library.
- **DP**: The adoption of a digital preservation strategy is not required to the content providers but a **NBN** is assigned only to the resources legally deposited at National Library and thus under a digital preservation strategy.
- **RR**: Every change applied to a resource generates a new resource that receives a new NBN.
- **UNCOUP**: The NBN syntax does not include the URL of the resolves. Thus a valid NBN syntax is: NBN:IT:UNIBO-12345.
- **RE-TO-RE**: The versioning relation is not managed.
- **PART-OF**: The part of relation is not managed.

PURL

- **RA**: The Registration authority is OCLC.
- **R-URL**: http://purl.oclc.org
- **RO**: the PURL resolves to the object only.
- **UNI**: the uniqueness of the PURL is assured.
- **MOD**: The PURL string modification is allowed[27].
- **DEL**: The purl deletion is allowed. In particular, deleted PURLs continue to exist on a PURL server but are modified to return http 410 status code.[28]
- **INI**: It is a no profit initiatives.
- **FEE**: The service is free of charge.
- **TDR**: Specific policies are not required to content providers.

[25] http://www.persistent-identifier.de/english/3352-EPICUR_policy.php
[26] http://www.persistent-identifier.de/english/3352-EPICUR_policy.php
[27] http://purl.oclc.org/docs/help.html#purlmodify
[28] http://purl.oclc.org/docs/help.html#purldelete

- **DP**: It is not required a particular digital preservation strategies to the content providers.
- **RR**: There is no guarantee that the object identified by the PUR will be the same because of this control is delegated to the final users.
- **UNCOUP**: The PURL is an URL itself, therefore it is not possible to uncouple the resolver from the identifier.
- **RE-TO-RE**: Relations between the identified resources are not managed.
- **PARTOF**: Relations between a resource and its parts are not managed.

Cool URI

- **RA**: A Registration authority is not defined since the Cool URI concept assumes that each institution guarantees the persistence of their URLs.
- **R-URL**: There is not a resolver, the Cool URI is URL and use the DNS infrastructure.
- **RO**: the Cool URI "identifies" an object published on the Web without a shared and controlled community guidelines, modalities, etc.
- **UNI**: the uniqueness of the Cool URI is not assured.
- **MOD**: The Cool URI string modification is allowed because they are simply URL thus any change of server domain, or folder or file format, affects the Cool URI string.
- **DEL**: The Cool URI deletion is allowed for the same motivation expressed in the MOD criterion. In fact a file can be no longer available on the institution server since it has been deleted (or moved) by the system administrator because of a system upgrade, etc.
- **INI**: It is not based on a specific initiative.
- **FEE**: The service is free of charge.
- **TDR**: It is not required a particular policy to the content providers. The unique recommendation is related to the URI design where the use of the date to build an URI is strongly suggested[29]. The result is an URL like this: http://www.w3.org/1998/12/01/chairs.
- **DP**: It is not required a particular digital preservation strategies to the content providers.
- **RR**: There is no guarantee that the object identified by the URI will be the same because of this control is delegated only to the Cool URI owners, in fact the trust is not mediated by a third-party agency.
- **UNCOUP**: The Cool URI is an URL itself, therefore it is not possible to uncouple the resolver from the identifier.
- **RE-TO-RE**: Relations between the identified resources are not managed.
- **PART-OF**: Relations between a resource and its parts are not managed.

In Table 2 the benchmark assessment results are reported. The table allows to determine if a PI system is compliant with the Trusted PI definition and thus eligible to be a part of the IF.

PI System	RA	R-URL	RO	UNI	MOD	DEL	INI	FEE	TDR	DP	RR	UNCOUP	RETORE	PARTOF
DOI	Y	Y	MO	Y	N	Y	B	Y	N	N	Y	Y	Y	Y
Handle	Y	Y	MO	Y	N	Y	NP	Y	N	N	Y	Y	N	N
ARK	Y	Y	MO	Y	N	N	P	N	Y	Y	Y	Y	Y	Y
DataCite	Y	Y	MO	Y	N	N	NP	Y	Y	Y	Y	Y	N	N

[29] T.Berned Lee- Cool URI don't change http://www.w3.org/Provider/Style/URI

NBN/DNB	Y	Y	OM	Y	N	N	P	N	Y	Y	Y	Y	N	Y
NBN/IT	Y	Y	OM	Y	N	N	P	N	Y	Y	Y	Y	N	N
PURL	Y	Y	O	Y	Y	Y	NP	N	N	N	N	N	N	N
Cool URI	N	N	O	Y	Y	Y	-	N	N	N	N	N	N	N

Table 1: Benchmark for Digital Object Identifier Systems

Similarly to the results highlighted by the Feature matrix for Digital Object Identifier Systems in Annex II, the table puts in evidence which are the requirements of a PI system to be eligible for the IF. The first two criteria are addressed by all the PI systems with the only exception of the Cool URI system. In fact, the existence of a Registration Authority of reference is a basic requirement to set up a reliable persistent identification service. A RA represents the third-party authority necessary to set up and maintain a coherent and homogeneous level of service and policies in a PID.

Moreover, an identification service is complete if allows users to continuously verify the trustworthiness of the PI-resource association. This can be done only via an online accessible resolution service. The table shows that all the considered PI systems have a resolver accessible online. For the Cool URI system, the resolver is represented by the DNS table. Thus, according to the resolver definition provided in the Reference Model, we cannot consider this as a proper resolver.

Another of the requirements to be eligible for the IF is the capability of a PI system of managing not only the access to the identified resource, but also its descriptive metadata. In this case, the PURL and Cool URI provide this simple access to the object. The other system provide also access to metadata and object (OM) or multiple objects (MO).

The set of persistence criteria (MOD, DEL, INI and FEE) have to be treated carefully. In fact a basic requirement of the persistence of an identifier is to guaranteeing based on the assurance that the identifier string cannot be changed or deleted, but in the persistence is obviously related to the commitment of the RA to keep alive the service in the long term. Thus, if the first two criteria are quite simple to be assessed, the latter two can be used only as an input to argue about their long-term sustainability without being definitive. Thus, the PI string modification, allowed by PURL and Cool URI, and PI cancellation, allowed (under certain condition) by DOI and Handle, undermine "persistence" of their identifiers. Thus, according to our definition, the DOI and Handle systems can be considered Trusted systems if a proper policy is defined (similar to what done by DataCite) and they can be eligible for the IF.

Regarding the sustainability of the service, it is worth to notice that only the Cool URIs system requires that each institution has sustain the cost of the PI persistence, because, the system requires not only the stability of the resource, but also of its localizations (URL). Business and no profit initiatives instead, require a fee that covers the basic services, such as assignment, maintenance and resolution, but also advanced services such as citability, multiple resolutions and so forth, leaving free the content providers to move, organize, migrate, transform, etc, their contents.

The public initiatives provide a different level of PI service because they are not market-oriented but related to the public utility, thus they can provide a PI service free of costs for the content providers.

A concern is emerged in relation to the PURL sustainability model. In fact, the OCLC provides this service free of charge even if the OCLC initiative is a not for profit organization. This model assumes the participation of stable founders to cover the costs of services, and the decisions about dismissing a service depend exclusively on the financial commitment of the participants. Therefore, a change in the funding objectives may shutdown the service.

Regarding the guidelines and criteria that content providers have to adopt for their content before accessing to a PI service, it seems that there is a main distinction between the public initiatives such as NBN systems or ARK because they require a specific digital preservation and repository trustworthiness commitment, and the other systems because they require less formalities to the content

providers. This evidence might be linked to the different purposes of these initiatives, as they are business oriented and need to increase as much as possible the number of their users/costumers to survive.

In conclusion, according our assessment criteria, the systems that are eligible for our IF are: DOI but under certain policy conditions as implemented by DataCite, Handle (with the same recommendations defined for DOI), NBNs and ARK.

6.1.2 PI systems for Authors benchmarking

Scopus ID

- **RA**: Scopus system is managed by Elsevier.
- **R-URL**: The full Scopus id service is available only for registered users. The web service that "resolves" a Scopus id has this form: http://www.scopus.com/authid/detail.url?authorId=7003901227
- **RO**: It resolves to the list of papers authored by an author that has to be registered in the system.
- **UNI**: The uniqueness is not assured at global level since the string is a simple number and it is valid only within the Scopus system. (e.g. Author ID: 7003901227).
- **MOD**: There is no way to manage the ID by the user directly. A tool like Author Feedback Wizard is implemented to allow to suggest changes only to their profile[30].
- **DEL**: There is not enough information available to assess this criterion with precision.
- **INI**: This is a business oriented initiative.
- **FEE**: The user can access to the service through a regional, institutional or group accounts. Thus we assume that a fee is required to these open these accounts.
- **RR**: The reliability of the resolution is assured by the service itself.
- **UNCOUP**: As said above, the ID is strictly liked to the system thus it is not possible separate the ID to its resolution service because the ID is devoid of context information like the authority namespace.
- **RE-TO-RE**: The full resolution service is allowed only within the system, but the results are a profile description of the author and the list of associated articles.

ResearcherID

- **RA**: The service is managed by Thomson Reuters - www.researcherid.com
- **R-URL**: www.researcherid.com/rid/
- **RO**: The resolution extract the information about the author, its affiliation and the list of associated publications.
- **UNI:** The uniqueness is not assured at global level since the string is a simple number and it is valid only within the ResearcherID system.
- **MOD**: The ID string cannot be modified.
- **DEL**: The deletion of a ResearcherID profile is allowed. The User has to contact the Global Customer Support and requests the account be deleted[31].
- **INI**: It is a Business initiative (Thomson Reuters).
- **FEE**: The service is free of charge.
- **RR**: The reliability of the resolution is assured by the RA.

[30] http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authorprofile

[31] http://researchanalytics.thomsonreuters.com/solutions/researcherid/rid-ul-faq/

- **UNCOUP**: The ResearcherID is linked to the system. As example the use of the ResearcherID badge that can be place in almost any online environment and links people directly to profile and full publication list in ResearcherID, is built in this way: http://www.researcherid.com/rid/C-3402-2008 where the string C-3402-2008 is the ResearchId and the previous part is the "resolver" link. Thus, the ID is strictly liked to the system and it is not possible separate it to its resolution service since the ID is devoid of context information like authority namespace
- **RE-TO-RE**: The ResearchID is in relation with the paper authored by person identifiers. Moreover the author id is in relation with other ResearchIDs that identify the person.

AuthorClaim

- **RA**: Authorclaim.org
- **R-URL**: No resolution service is available. The System generate a permalink of the user profile like this: http://authorclaim.org/profile/pbe4/
- **RO**: The permalink points to the user information and to a list of the publications
- **UNI**: the uniqueness is assured since the result of the author registration is a permalink. Thus the system uses the DNS (like PURL) to address the uniqueness, but the service does not generate an identifier in strict sense.
- **MOD**: It is not possible to change the ID string
- **DEL**: The user can delete his own profile
- **INI**: This is a not for profit initiative
- **FEE**: No fee is required
- **RR**: The reliability of the resolution is not addressed since it is possible to change the entire profile associated to the ID. For instance it is possible to change the Name, Surname, etc.
- **UNCOUP**: Not applicable
- **RE-TO-RE**: The system allows the users to declare their authorship of articles present in the database collection [32]takes into account.

arXiv Author ID

- **RA**: Cornell University Library http://arxiv.org . The Author ID is assigned only if an article is submitted by an author to the arXiv repository.
- **R-URL**: http://arxiv.org
- **RO**: The resolution result is the list of articles written by the author identifier provided via Atom protocol.
- **UNI**: the arXiv Author ID does not address the global uniqueness because the identifier does not include any authority namespace definition. The form of arXiv Author ID is like this: http://arxiv.org/a/warner_s_1
- **MOD**: The ID string cannot be modified.
- **DEL**: There are not enough information available to assess this criterion with precision.
- **INI**: The initiative is not for profit.
- **FEE**: The service is free of charge.
- **RR**: The reliability of the resolution is based on the trustworthiness of the author. In fact it is possible to change information like the name, surname etc, once the user is logged in.

[32] http://authorclaim.org/collections

- **UNCOUP**: Similarly to other systems, the arXiv Author ID is strictly liked to the system thus it is not possible separate the ID to its resolution service because the ID is devoid of context information like authority namespace.
- **RE-TO-RE**: The resolution result is a list of the publication associated to an Author ID.

Digital Author ID

- **RA**: SURF is responsible for the governance and strategic developments of the overall identifier system[33].
- **R-URL**: There is not a specific resolution service devoted to resolve a DAI but this identifier can be used in several service such as NARCIS [34]to obtain information and the list publication done by a certain author.
- **RO**: As said, the DAI can be "resolved" to the author profile and his the list of publications
- **UNI**: the DAI addresses the uniqueness because the identifier is composed by the authority namespace:info:eu-repo/dai/nl/123456785
- **MOD**: The DAI string cannot be modified
- **DEL**: There is no information about the possibility of deleting a DAI
- **INI**: It is a service managed by a not for profit organizations (SURF)
- **FEE**: The system is not free to access. Existing users of the GGC can use the NTA and DAI functionality free of additional charges. For research institutions that do not have access to the GGC, a separate 'DAI contract' is available for the use of light version of the WebGGC for a limited fixed fee. The NTA is funded through license fees for the use of GGC.
- **RR**: The data in the NTA contains personal data. This data is protected by Dutch law. Users of the NTA have permission to use this data only for bibliographic purposes. This permission has been granted by the Dutch Data Protection Authority (Dutch DPA). As said, the system is based on the CRIS that have the responsibility of the author profile management.
- **UNCOUP**: A URI-field DAI looks like this: info:eu-repo/dai/nl/123456785. The DAI is the number after the string info:eu-repo/dai/nl/. A DAI is a number like 123456785. The string: info:eu-repo/dai/nl/ is the authority namespace, telling the user or machine that the number is a DAI originating from the Netherlands. At the moment the INFO-URI namespace is used as an authority namespace. The DAI is URI-field under the EU-REPO sub-namespace. This namespace defines components for compound objects in the Institutional Repositories. Thus the resolution service is separated by the identifiers itself.
- **RE-TO-RE**: The DAI links the publications with their authors

VIAF

- **RA**: OCLC
- **R-URL**: http://viaf.org/
- **RO**: The system resolves to the name and in conjunction of WorldCat disambiguates the IDs coming from Authority files of registered institutions.
- **UNI**: The uniqueness is not assured since the VIAF Id is only a numerical string like VIAF ID: 20398616.
- **MOD**: Changes on the VIAF ID string seems not allowed.
- **DEL**: The ID of each single Authority file can be deleted but the VIAF ID remains because it can point to a list of IDs of different Authority files that represent the same person. This is an association one-to-n where n is the number of IDs that refer to the same person.

[33] http://wiki.surffoundation.nl/display/standards/DAI#DAI-responsibilities

[34] http://www.narcis.nl/

- **INI**: The VIAF is a joint project of several national libraries plus selected regional and trans-national library agencies and the service is managed by a not for profit organization (OCLC). The project's goal is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web.
- **FEE**: The service is free of access.
- **RR**: The reliability of resolution is assured mainly by an agreement between the service and the agencies that wish to participate in VIAF. The application process requires submission of test files of bibliographic and authority files from the participating agency to the VIAF consortium. OCLC Research—on behalf of the VIAF consortium—processes the records to match name headings from the source files with VIAF records built from content previously supplied by other VIAF consortium members. If an agency's application is successful, the agency enters into agreement and sends its full bibliographic and authority files to OCLC Research with updates sent on a regular cycle.
- **UNCOUP**: The VIAF number is an ID like this: VIAF ID: 20398616 that can be resolved by the online VIAF resolution service or trough the Permalink: http://viaf.org/viaf/20398616. It is clear that the uniqueness of identifiers is given by the combination of the resolver URL and the ID string.
- **RE-TO-RE**: The Author ID is linked to the publications of the Author.

PI System	RA	R-URL	RO	UNI	MOD	DEL	INI	FEE	RR	UNCOUP	RETORE
SCOPUS Author ID	Y	Y	RO	N	N	?	B	Y	Y	N	Y
Researcher ID	Y	Y	RO	N	N	Y	B	N	Y	N	Y
AuthorClaim	Y	N	RO	Y	N	Y	NP	N	N	N	Y
arXiv Author ID	Y	Y	RO	N	N	?	NP	N	N	N	Y
DAI	Y	Y	RO	Y	N	N	NP	N	Y	Y	Y
VIAF	Y	Y	RO	N	N	N	NP	N	Y	N	Y

Table 2: Benchmark for Author Identifier Systems

The Author PI systems are quite different respect to those for digital objects.

In fact, these PIs behave more as internal identifiers because the Author identity is managed in strict relation to the publications stored in the systems, like happens in Scopus or arXiv. In particular, the arXiv system generates an Author ID only if a paper is submitted at least to the arXiv repository.

The AuthorClaim, instead, cannot be considered a PI system but rather an automatic tool for generating a Web page of an author with a permalink once the user has claimed his authorship respect of a list of papers harvested from the data providers linked to the service.

The unique PI system quite close to our Trusted PI definition seems to be the DAI system, because it is able to separate the PI to the resolution service, while preserving the meaning of the ID itself. This is possible because it includes in the PI definition the authority namespace that provides the contextual information necessary for addressing the global uniqueness of the PI. The problem of the DAI is that the service is not available on the Web for all users.

Finally, VIAF is a notable initiative because is based on trusted institutions such as national libraries but the system generates an ID that works only in conjunction with the VIAF Web service.

7 ARTICULATION WITH THE REST APARSEN WPs AND TASKS

Here we describe how this work is related with the other work packages and tasks of APARSEN

WP	Notes
WP11 Common Vision (M1-M18)	The IF matches the common vision interrelation view on the following aspects: • People: People support PI systems (and therefore organizational, political, economical and social aspects should be considered). • Data: Data may have persistent IDs associated with them. • Reputation: Reputation is associated with persistent IDs. Different PI systems may have different reputations. • Annotation: Annotations may be associated with PIs. • Policies: the use of PIs is regulated by agreed policies
WP13 Coordination of common standards (M4-M48)	The results of the Task 2210 are related to T1310 (Analysis of current standards) since we carried out a feature analysis on the present PI technologies in use to identify their trustworthiness against the IF requirements. The results has clearly outlined that the standardized solutions (IETF-URN, ISO-DOI, etc.) can be considered trustable. This result can be used as input for the WP13 analysis.
WP25 Interoperability and Intelligibility (M20 – M33)	The results of the WP22 Task 20 is the definition of a framework for enabling the interoperability among PIs. The definition of the IF and its dependant services is strongly related to the activities of the WP 25 that focuses on investigating and developing techniques to support syntactic and semantic interoperability of data between organizations and disciplines. The IF guaranteeing the persistent access to digital resources across systems, functions, metadata schema, semantic and linguistic barriers, should provide a first layer of interoperability on which more sophisticated models regarding the interoperability between metadata, standard protocols and ontologies can be implemented. For example, semantic interoperability of metadata - which express the relationships that someone claims to exists between entities – depends on the unique identification of all these entities, since otherwise expressing relationships between them and agree on the meaning of these relationships is of scarce utility. Moreover, the identification of the authority, which makes the claim is crucial as well. On this perspective, the IF (and its dependant services) can be at the core of the WP 25 activities related to semantic interoperability since it provides the key tokens to the management of identification for implementing effective solutions to support semantic interoperability.
WP31 Digital Rights and Access Managements (M27-M38)	The adoption of an PI Interoperability Framework can improve the DRM system on the resource identified. It is also able to allow the user to access to the appropriate copy according to the user and content rights

WP 24 Authenticity and Provenance	The IF support the Authenticity and Provenance issue for digital object. In fact the PI can support the provenance certification of a resource through the resolution service. Moreover, the Authenticity can be supported through the IF by linking the copies of "the same" object together. In this way, it should be possible to compare the copies to verify their compliance with the source. Another point of connection with the WP 24 activities deals with the life cycle tracking of a resource, The process of tracking a digital resource throughout its lifecycle can be managed through its PIs, which can be embedded into the logging files. Through the IF and the alternative PIs service, the system could trace all the phases and events which the resource may undergo (also in different systems, including changes in technical system and custody) without losing its authenticity and provenance evidence. The availability of mapping among PIs may also allow the implementation of services for exchanging and integrating provenance information. For example, associating the ontology/schema mapping of provenance models with the PIs mapping provided by the IF could improve provenance interoperability.

8 CONCLUSION AND NEXT STEPS

This work defines a framework to set up the conditions of the interoperability among PI systems. The research has analyzed and assessed the main PI initiatives for Digital Objects, Authors and Institutions and has identified which are the key entities of a PID in order to define the possible relations and interactions necessary for the construction of a concrete interoperability platform. The Framework is thought to support the design and development of new interoperability services. For instance, the IF should allow the citability of cross-domain stable resources and consequently to build a number of services tailored on specific community requirements. The success of an IF derives from the pervasiveness of its adoption that is, in fact, one of the key factors for an extensive consensus building and for the long term sustainability of the IF itself. To this end, we have introduced the concept of Trusted PID defining a set of criteria that a PID has to address to be eligible to the IF, and at the same time, we left to the individual PID the responsibility of guaranteeing suitable policies for any aspect of the Digital Preservation Plan underpinning that system, like for example, the content selection/granularity criteria, the Trusted Digital Repositories policies and certification, the trustworthiness of the PI management, and so on. In fact, within each PID there can be different approaches and policies but we assume that the user community is free to choose the best solution and we trust them for the correctness of the assertions.

8.1 INTEROPERABILITY FRAMEWORK FOLLOW UP

In the second part of the WP 22 some of the identified services will be designed, taking the IF as a reference. In particular, by addressing the citability issues, advanced services for resources identified by different PI systems, can be implemented, such as:

- powerful metrics for assessing impact of scientific production,
- cross citations allowing the user to move from one article to another at the citation level, regardless of journal or publisher,
- a European register of certified resources for an automatic citation indexing system,
- services based on unique identifiers for researchers, linked to alternative author identifiers and individual's research output, as those proposed by ORCID initiative.

These new advanced services will focus on some scenarios presented in this document and will be tailored on selecting use cases. The potential users of these new services can be the PIDs themselves, search engines, citizen services, social network, data providers, and so forth.

Following the reviewers recommendations, the WP22 team will implement a validation mechanism in order to evaluate the Interoperability Framework for PI by around 30 experts a part of them external to the APARSEEN consortium. Moreover an action plan to set up a demonstrator for WP22 Interoperability Framework and related services, is under preparation with the limited resources for software development in WP22 and some external possible synergies with other projects like SciDip-ES or other initiatives like ORCID and DOI or NBN large communities. Based on the infrastructure of that demonstrator some basic services will be tested and refined in order to implement the user requirements collected during the former work in the WP22 with the questionnaire and the use cases definition.

9 REFERENCES

[1] CARROLL, J. M. (1995). Introduction: the scenario perspective on system development. In J. M. Carroll (Ed.) *Scenario-based design: envisioning work and technology in system development* (pp. 1-18). New York: John Wiley & Sons, Inc

[2] KUUTTI, K. (1996). Activity theory as a potential framework for human-computer interaction research. In . A. Nardi (Ed.) *Context and consciousness: activity theory and human-computer interaction*, (pp17-44). Cambridge, Massachusetts: The MIT Press

[3] Stefan Gradmann, INTEROPERABILITY. A key concept for large scale, persistent digital libraries. Digitalpreservationeurope (DPE) project - Briefing Paper - September 2008 http://www.digitalpreservationeurope.eu/publications/briefs/interoperability.pdf

[4] Norman Paskin - Interoperability Identifiers - Briefing Paper - Digitalpreservationeurope.eu http://www.digitalpreservationeurope.eu/publications/briefs/identifier-interoperability.pdf

[5] E. Bellini, C. Cirinnà, M. Lunghi, E. Damiani, C. Fugazza PIs distributed system for Cultural Heritage digital objects. iPRES Conference 2008 http://www.rinascimento-digitale.it/documenti-ipress2008pi.phtml

[6] Warner, S., Bekaert, J., Lagoze, C., Liu, X., Payette, S., and Van de Sompel, H. (2007) Pathways: augmenting interoperability across scholarly repositories. International Journal on Digital Libraries. 7(1), pp. 35-52. doi:10.1007/s00799-007-0016-7 ; http://dx.doi.org/10.1007/s00799-007-0016-7

[7] Makx Dekkers and Thomas Baker CORES project, Standards Interoperability Forum Resolution on Metadata Element Identifiers http://www.cores-eu.net/interoperability/cores-resolution/

[8] PIs Linking Infrastrucure (PILIN) project - http://www.pilin.net.au/

[9] Resourcing Identifier Interoperability for Repositories (RIDIR) project http://www.jisc.ac.uk/whatwedo/programmes/reppres/ridir.aspx

[10] N. Nicholas, N. Ward and K. Blinco A Policy Checklist for Enabling Persistence of identifiers…D-Lib Magazine January/February 2009 http://www.dlib.org/dlib/january09/nicholas/01nicholas.html

[11] OAIS Reference Model for an Open Archival Information System OAIS http://public.ccsds.org/publications/archive/650x0b1.PDF

[12] PersID project – http://www.persid.org/

[13] PersID project report III: (IIIa) Current State and State of the Art & (IIIb) User Requirements. urn:nbn:nl:ui:13-9g4-i1s http://www.persid.org/downloads/finalreports/PersID_Report_Part_3_final.pdf

[14] Catalogue of criteria for assessing the trustworthiness of PI systems http://files.d-nb.de/nestor/materialien/nestor_mat_13_en.pdf
Bellini E, Cirinnà C. and Lunghi, M. PIs for Cultural Heritage Digitalpreservationeurope (DPE) EU project – Briefing Paper http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf

[15] DELOS Reference Model http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf

ANNEX I

SURVEY RESULTS AND ANALYSIS

Task leader
University of Trento (UNITN)

Authors:
Barbara Bazzanella (UNITN), Paolo Bouquet (UNITN),
Emanuele Bellini (FRD) Maurizio Lunghi (FRD) Chiara Cirinnà (FRD)

Questionnaire structure

In the questionnaire we considered three kinds of PI systems based on their referent type:

1. PIs for digital objects
2. PIs for authors and creators
3. PIs for organizations

The questionnaire was composed of five sections:
1. **PI for digital objects**: in this section we focused on identification practices for digital objects. We aimed to collect information about the adopted identifier systems for digital objects and their limits, the types of digital objects indentified by PIs, the level of granularity adopted by assigning PIs, the use of preservation practices (i.e. versioning), metadata schemas and repositories for digital objects, the access systems to digital objects and metadata.

2. **PI for authors/information creators**: in this section we collected data about the main producers of digital contents, the identification systems used to identify them, the obstacles in adopting these systems, the requirements that such systems should have in terms of discipline, governance and geographical boundaries, trust and search facilities practices.

3. **PI for organisations**: in this section we investigated the user practices in adopting PIs for organizations, adoption obstacles, granularity requirements and search facilities practices.

4. **Criteria for the adoption of a PI system for digital objects**: in the fourth section we deepened the criteria adopted by the users for the adoption of PI systems for digital objects. We considered aspects related to technology, organization of the service, scope, naming rules and expected services.

5. **Digital preservation strategies and practices**: in the final section we addressed issues concerning digital preservation strategies and practices with a special focus on the use of written guidelines, time span, funding and financial sustainability.

At the end of the questionnaire, we collected some general information about the respondent, like organization type and sector, job title and country. In particular, the organization type allowed us to investigate possible differences among the represented stakeholder groups.

The questionnaire was placed online using the Surveygizmo[35] service.

SECTION I – GENERAL INFORMATION ABOUT RESPONDENTS

1 Organization type and sector

The questionnaire received 103 full responses from participants of a variety of organizations types: mainly libraries, universities, archives and publishing organization, as shown in Table 3[36], across a

[35] www.surveygizmo.com/

number of sectors: primarily academic/research, government, public sector. It is worth to notice the low number of participants from the private sector. Therefore, we should be cautious about generalizing the survey results to this organization sector. Furthermore, the number of responses suggests of considering an "error bar" of about 10%, thus there are no substantial differences between results like 15% and 25%.

Organization Type	Freq	Percent %
Library	49	47
University	28	27
Research organization	19	18
Scientific data archive	16	15
Document archive	8	7
Publishing organization	8	7
Competence center	6	6
Other archives	5	5
Museum	3	3
Professional Association	3	3
Training center	8	2
Other	7	7
Total N. of respondents	103	

Table 3: Organization types

Organization sector	Freq	Percent %
Academic/research	72	70
Public sector	18	17
Business	17	16
Government	5	5
Private	2	2
Other	7	7

[36] It should be noted that many participants have categorized their organization in more than one type (e.g. Scientific data archive and research organization). This explains why the sum of the frequencies in the frequency table does not equal the 100%. This is also true for the other questions with the multiple selection option.

Total N. of respondents	103

Table 4: Organization sector

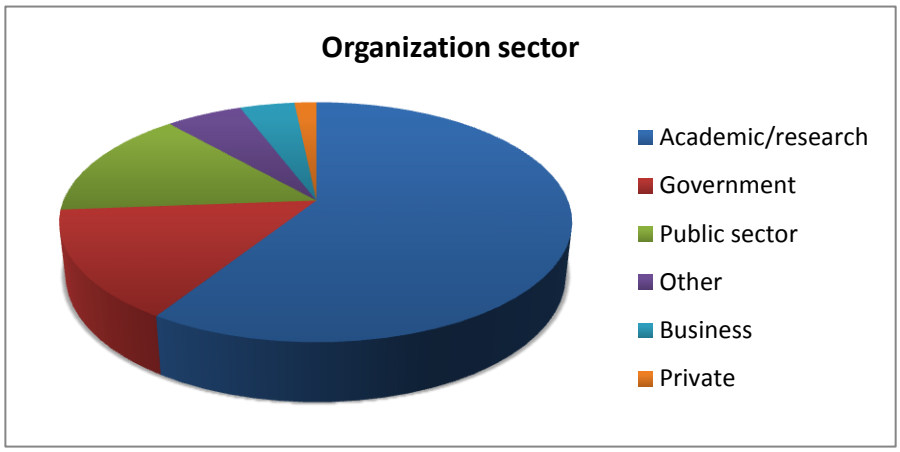


Figure 7: Participants' organization sectors

2 Provenance

A detailed geographic distribution of the respondents is reported in Table 5. In summary, as shown in Figure 8, the majority of the respondents were from European countries (85%). Only 15 % of the respondents were from extra European countries (mainly from USA). This is another important element to take into account for the generalization of the results. Our results can be used to understand important aspects about the current state of use of PIs in European countries, but they do not provide enough information about the use of PIs systems in countries outside the Europe. Collecting data from these countries (USA in particular) would allow very useful information for future analysis and comparisons.

Country	Freq	Percent %
Germany	20	19
USA	12	12
UK	11	11
Netherland	8	8
Italy	7	7
Sweden	5	5
Denmark	3	3
Finland	3	3

Greece	3	3
Hungary	3	3
France	3	3
Switzerland	2	2
Spain	3	2
Australia	2	2
Cyprus	2	2
Lituania	2	2
Slovenia	2	2
Others	14	14
Total N. of respondents	103	

Table 5: Geographic distribution of respondents

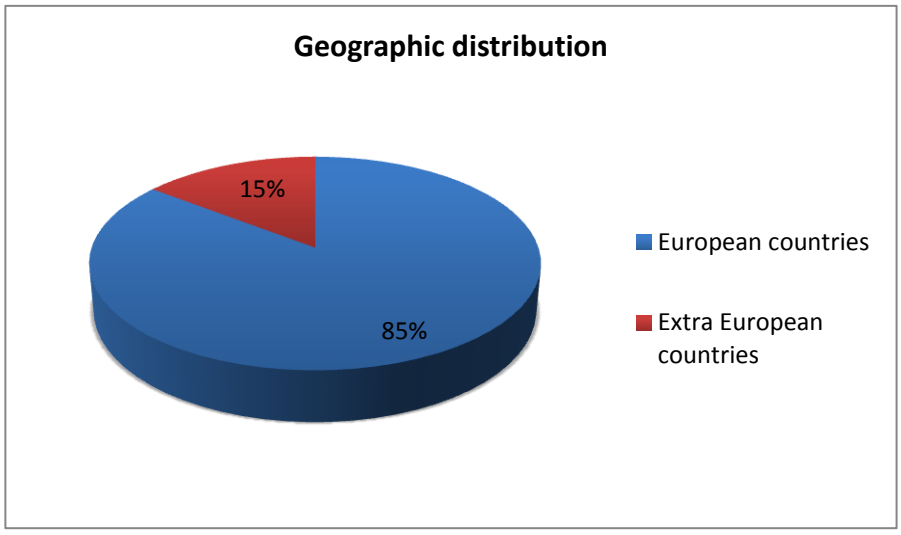


Figure 8: Geographic distribution of respondents

SECTION 2: PI FOR DIGITAL OBJECTS

Types of digital objects

Question 1.1: Which types of digital objects are managed by your organization? (Please tick all that apply)

As shown in Table 6, textual documents and images are the types of objects with the highest frequency (>86%). More than half of participants reported also video, websites, audio, databases and datasets. Processed data, multimedia, maps-geospatial data and raw data are reported to be managed less frequently (40-45%). Compound objects, software's and 3D objects present lower frequency (< 40%).

Object type	Freq	Percent %
Textual documents	**101**	**98**
Images	**89**	**86**
Video	65	63
Websites	64	62
Audio	62	60
Databases	55	53
Datasets	55	53
Processed data	46	45
Multimedia	46	45
Maps/Geospatial data	43	42
Raw data	42	41
Compound objects	39	38
Software	31	30
3D objects	20	19
Other	5	5
Total N. of respondents	103	

Table 6: Object types

Level of granularity

Question 1.2: Which level of granularity (the " level of detail at which PIs needs to be assigned", e.g. chapter – paragraph – subparagraph) do your digital objects require? (See glossary for definitions)

This question has been formulated in a very general way, without specifying the type of object to be considered. Therefore, respondents were free to express their opinion about the level of granularity considering a specific object (e.g. a book, an article) or considering a generic digital object maintaining a higher level of abstraction.

For this reason, we collected very heterogeneous answers. In order to present the results in a more understandable way, we have introduced some criteria to aggregate the answers.

First of all, we distinguished between respondents that suggested a level of granularity corresponding to the whole object, such as a book, an image, an article or more generically a digital object and respondents that reported a deeper level of granularity (i.e. high level of granularity), such as a chapter, a paragraph, a page and so on. Then we aggregated the very few subjects who mentioned a level of granularity above the object itself, such as collections or series. Finally, we created a different category for respondents, which reported that the level of granularity depends from the kind of digital object or specific needs. In Table 7 we report the results according to this categorization. We note that respondents preferred a high level of granularity, even though a substantial number of them reported granularity at the object level. Very few mentioned levels of granularity above the object.

Level of granularity	Freq	Percent %
High level granularity (parts of object)	**41**	**40**
Granularity at the level object level	35	34
Low level granularity (collections of objects)	5	5
Depending on the object type	15	14
Other	7	7
Total N. of respondents	103	

Table 7: Granularity

Content versioning

Question 1.3: 1.3 Do you manage the content versioning?

From Table 8 we can see that the most common approach for content versioning is linking a new version to the original version through metadata, followed by the practise of considering the new version as an autonomous object. The use of naming rules is less common among the participants.

Content versioning	Freq	Percent %
Yes - The new version is linked to the original version through metadata	**50**	**48**
No - We consider the new version as a autonomous new object	32	31
Yes - The new version is linked to the original version through naming rules	16	15
Other (please specify)	16	15

Total N. of respondents	103

Table 8: Content versioning

Metadata schema

Question 1.4: What metadata schema do you use to manage your digital objects? (Please tick all that apply)

About the metadata schema, Dublin Core is the most frequently reported schema (almost 71%). However, it is worth to note that a large number of participants (around 46%) reported to use other metadata schemas such as: AGLS and AGRkMS, CIDOC DC compatble, CrossRef, NLM, Darwin Core, MIX, GML, DataCite Metadata Scheme, DOI, DTD schema, dublin core extended, EADXML, EML, BDP, ISO19139, ESE, FGDC CSDGM, FOXML, PREMIS, ISO 19115, INSPIRE, LOM, MAB2, MAG, Midas, mpeg7, NLM DTDs, OAI-ORE, panFMP, PANGAEA_\xf6Framework for Metadata Portals, ISO19xxx, DIF, pica, dcx, onix, Ex Libris-DNX, RIF-CS, TEI 5, TEXTMD, MEPG7, DocumentMD, UKETD_DC, VRA Core, PBCore, EAD, ISBD/RDF, FRAD/FRSAD/RDF. This list suggests the heterogeneity of the current metadata schemas and shows the complexity of the issues related to the metadata interoperability.

Metadata schema	Freq	Percent %
Dublin Core	73	71
Other	47	45
My own database schema	37	36
METS	34	33
MODS	23	22
MARCXML	22	21
My own ontology	15	14
DDI	8	8
Naming conventions	8	8
DIDL	7	7
xMetaDiss	5	5
CIDOC CRM	2	2
Epicur	1	1

Total N. of respondents	103

Table 9: Metadata schema

Repository system

Question 1.5: Which repository system do you use to manage your digital objects? (Please tick all that apply)

We found a very similar pattern of results about the repository system. Fedora, Dspace and Database are the most frequently reported systems. However, also in this case we have a large number of participants who claimed to use other systems such as Access, Documentum, eXist, filemaker, Informix, Invenio, mongoDB, MySQL, POStGRESQL, Oracle, MSSQL, PANGAEA, CouchDB, SYBASE, xml information system.

Repository	Freq	Percent %
Other	54	52
Fedora	26	25
Dspace	25	24
Database	24	23
Eprint	6	6
Greenstone	4	4
CASTOR	1	1
Filestore	1	1
Safety Deposit Box	1	1
Total N. of respondents	103	

Table 10: Repository system

Access system

Question 1.6: Which access system to your metadata/digital objects have you implemented?

About the access systems to metadata or digital objects, the majority of participants reported to use OAI-PMH protocol. As in the previous two questions, we found a considerable number of participants (around 30%) reporting other access systems. Among these: Access and authentication with digital certificate, Science Direct (Elsevier), bi-directional DOI-linking, Blacklight/Solr, Canto cumulus, Custom UI, Custom Web Pages, CWS, storage backend, local systems, http, OAI-PMH, delivery applications, NESSTAR, ORE, Atom/RSS, OpenURL, PRIMO, Proprietary system based on

Cumulus, RSA token/ passwords, SOAP, SRU-protocol, Tivoli Access MAnager [TAM], Web services discoverable through our OPAC, XTF, Z39.50.

Access system	Freq	Percent %
OAI-PMH protocol	66	64
Other	30	29
DBMS access permission	23	22
Repository splash page	17	16
Linked Data	16	15
None	12	12
RDF/SPARQL	10	10
Total N. of respondents	103	

Table 11: Access system

PI systems

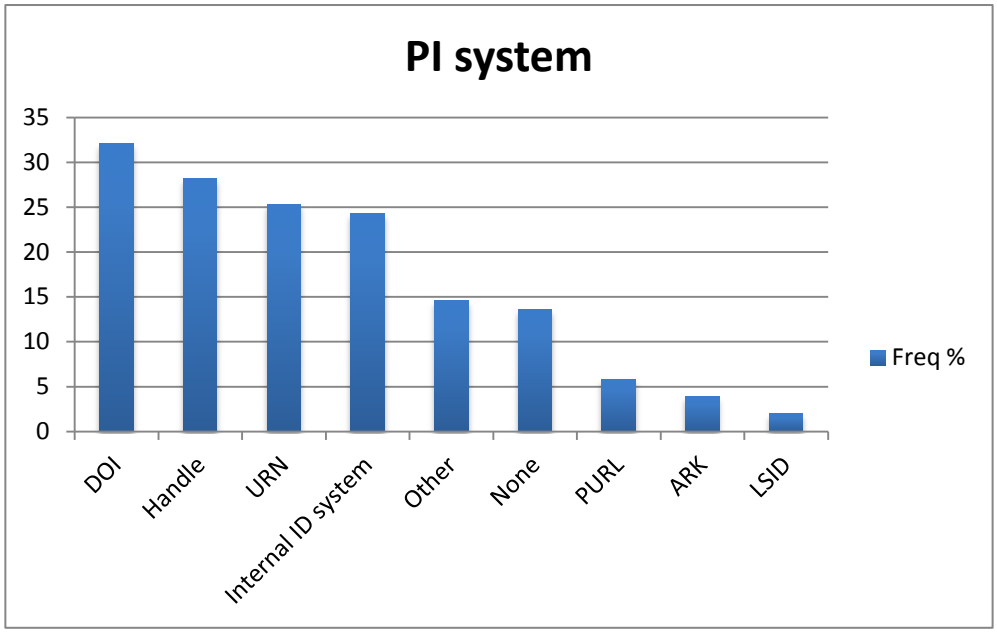
Question1.7: Which PI system do you use to identify your digital objects? (Please tick all that apply)

About the PI system for digital objects, DOI, Handle and URN are the most frequently adopted systems. It is important to notice that around 25% of respondents reported to use systems to generate internal identifiers. Since we assume that the use of internal IDs is much more widespread than this, our interpretation of the result is that probably people intended this option as the exclusive use of internal IDs to identify digital objects.

PI system	Freq.	Percent. %
DOI	**33**	**32**
Handle	29	28
URN	26	25
System-generated internal identifiers	25	24
Other	15	14
None	14	13

PURL	6	6
ARK	4	4
LSID	2	2
Total N. of respondents	103	

Table 12: PI systems



A further analysis about the use of PIs has been performed to investigate possible differences in the distribution of use of PI systems between the different groups of stakeholders. The idea is to show whether certain systems are more frequently used by certain groups of users.
We limited the analysis to the main groups of stakeholders in our sample, that is:

Stakeholder group	Freq.	Percent. %
Library	49	47
University	28	27
Research Organization	19	18
Archive	24	23
Publisher	8	8
Total N. of respondents	103	

Table 13: Stakeholder groups before filtering

Since a small number of participants selected more than one organization type (e.g. university and research organization), producing a moderate overlapping between the stakeholder groups, we decided to include in the analysis by stakeholder groups only the answers with a unique organization type (i.e. without overlapping) in order to reduce the noise in the data. After filtering the data according to this approach, we found the following frequency distribution of stakeholder groups:

Stakeholder group	Freq.	Percent. %
Library	34	33
University	12	11
Research Organization	7	7
Archive	24	23
Publisher	5	5
Total N. of respondents	103	

Table 14: Stakeholder groups after filtering

From the comparison between Table 13 and Table 14, it comes out that University and Research Organization are the two groups with the higher degree of overlapping in our sample.

The analysis by stakeholder groups shows that:

- DOI is the most common PI system used by universities, research organizations, archives and publishers.
- URN is the most widely used PI system by libraries.
- Libraries and archives show to use quite frequently internal PI systems.

	Libraries	Universities	Research organizations	Archives	Publishers	
DOI	7	**4**	**5**	**8**	**5**	29
Handle	10	3	2	7	0	22
URN	**16**	1	2	1	0	20
Internal	9	1	3	**8**	1	22
Other	5	2	0	6	0	13
None	4	2	0	4	0	10
PURL	1	0	2	1	0	4
ARK	1	0	0	2	0	3
LSID	0	0	0	2	0	2
	34	12	7	24	5	

Table 15: PI systems for digital objects by stakeholder groups

Limits

Question 1.8: Based on your experience, concerning the selected system, have you experienced any limits associated to the following issues?

Considering the most frequently used PI systems (DOI, Handle, URN, internal systems), we found that many users did not report any limit associated to the adopted system.
Costs and low adoption are the two main limits experienced by the DOI users, versioning by Handle users and low adoption by URN users. Not surprisingly, users of internal systems reported two limits: that they are locally defined and the lack of standardization.

Value	DOI	Handle	URN	Internal identifier system	ARK	LSID	PURL	Other
Low adoption	6	3	8	0	1	1	1	2
Locally defined	1	1	3	11	2	0	3	5
No standard associated	3	1	5	9	0		4	2
Sustainability	3	3		2	0	2	0	5
Performance	1	2	2	0	0	1	0	5
Granularity	3	2	4	4	0	1	0	0
Trust		1	2	2	0	1	1	0
Versioning	5	4	4	3	0	0	0	3
Governance structures	3	2	4	1	0	1	2	4
Ongoing cost	10	1	1	1	0	0	1	1
None	12	12	8	6	0	0	1	2
Other	3	3	1	0	1	2	1	
TOT	33	29	26	25	4	2	6	15

Table 16: Limits associated to PI systems

SECTION 2: PI FOR AUTHORS/INFORMATION CREATORS

Digital objects producers

Question 2.1: Which are the main producers of digital objects managed by your organization? (Please tick all that apply)

About the producers of digital objects, affiliated authors and public organizations are reported as the main producers in the organizations of the respondents.

Producers	Freq	Percent %
Affiliated authors	67	65
Public organizations	51	49
External authors	39	38
Private organizations	24	23
Other	16	15
Total N. of respondents	103	

Table 17: Digital objects producers

PI systems for authors and creators

Question 2.2: Which PI systems do you use to identify authors/information creators of digital objects? (Please tick all that apply)

The results indicate that the use of PIs for authors and creators is considerably less diffused than the use of PIs for digital objects. In particular, more than half of participants reported to not use any PI system for authors and creators of digital objects and more than 20% of them reported adopting internal ID systems. One quarter of the participants who indicated using other systems claimed to plan to adopt ORCID when available. Notably, we obtain the same pattern of results even if we analyze the data by stakeholder groups (see Table 19), in that, in all groups, around half of the participants reported to not use PI systems for authors (see Figure 10) .

PI system	Freq	Percent %
None	**54**	**52**
Internal ID	24	23
Other	21	20
ScopusID	5	5
DAI	4	4
ResearcherID	4	4
AuthorClaim	2	2
Open ID	1	1
Total N. of respondents	103	

Table 18: PI for authors and creators

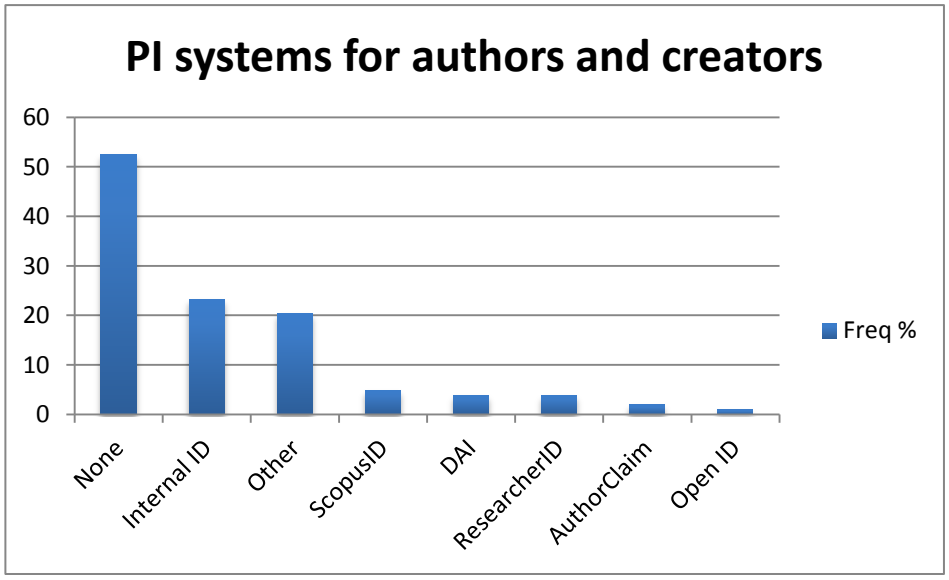


Figure 9: PI systems for authors and creators

	Libraries	Universities	Research organization	Archives	Publishers	
None	**18**	**6**	**4**	**12**	**3**	43
Internal ID	4	3	1	7	1	24
Other	9	0	1	5	2	17
ScopusID	2	2	0	0	0	4
DAI	2	0	0	1	0	3
ResearcherID	1	2	0	1	0	4
AuthorClaim	0	1	1	0	0	2
Open ID	0	0	0	0	1	1
	34	12	7	24	5	

Table 19: PIs for authors and creators by stakeholder groups

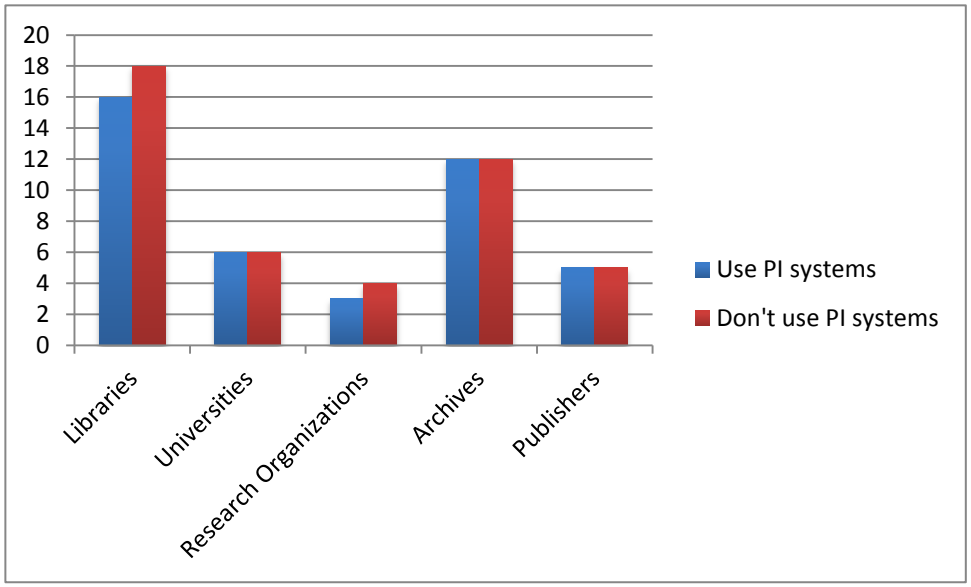


Figure 10: Use of PI systems for authors by stakeholder groups

Obstacles

Question 2.3: Which are the main obstacles in adopting a PI system for authors/information creators?

The results show that many organizations do not use PI systems for authors and creators mainly because they do not consider it as a key issue or because of lack of awareness. In general, looking also at the explanations reported by those who selected the other option, we can say that users perceive a certain level of immaturity for systems for author identification which concerns services, trust and authority.

Obstacles	Freq	Percent %
It is not a key issue for the organisation	23	22
Authors do not know about (or do not care for) this service	19	18
Low attractiveness of the service due to low level of adoption	13	13
Other (please specify)	11	10
National legislation with regard to privacy of personal data	8	8
Lack of trust and authority	7	7
Total N. of respondents	103	

Table 20: Obstacles in using PI systems for authors and creators

Requirements

Question 2.4: Which characteristics a PI system for authors/information creators should have? *(Please indicate the top three issues driving your organisation's interest)*

As we can see from the Table 21, the participants indicated three main requirements for a PI system for author and creators. The system should be cross-discipline, managed by public/government institutions and transnational. From these answers, we can argue that people are in favour of a global system but they need to trust in the authority, which manages it.

Requirements	Freq	Percent %
Cross-disciplinary	**83**	**80**
Discipline-specific	10	10
Managed by public/government institution	**74**	**72**
Privately managed	7	7
Nationally not limited	**57**	**55**
Nationally limited	5	5
Other	9	9

Table 21: Characteristics of a PI system for authors and creators

The same analysis conducted by stakeholder groups produced the following pattern of results:

	Libraries	Universities	Research organization	Archives	Publishers	
Cross-disciplinary	**26**	**10**	**6**	**19**	**4**	65
Discipline-specific	2	0	0	0	1	3
Managed by public/government institution	**27**	**8**	**5**	**16**	**1**	57
Privately managed	2	0	0	2	0	4
Nationally not limited	**22**	**10**	**3**	**17**	**3**	55
Nationally limited	4	1	1	0	0	6
	34	12	7	24	5	

We can note that there is a complete agreement between the stakeholder communities about the requirements for an authors/creators identifier system: it should be cross-discipline, managed by a public/government institution and trans-national. This is an important clue to propose co-ordination actions, which aim at creating consensus between the stakeholder communities.

Trust

Question 2.5: Which factors contribute to trust in a PI system for authors/information creators? (Please indicate the top three issues driving your organisation's interest)

About the trust of a PI system for authors and creators, the most frequently reported factors are the trust toward the organization running the system and the methods of verification. As shown in Table 23, we did not find differences between the stakeholder groups about ranking these factors.

Factors contributing to the trust	Freq	Percent %
Trusted organization running the system	74	72
Methods of verification	68	66
Supported by stable funders	32	31
Validation by publishers	31	30
Author self-curation	27	26
Other	8	8
Validation by educators	7	7
Total N. of respondents	103	

Table 22: Factors that contribute to the trust in a PI system

	Libraries	Universities	Research organization	Archives	Publishers	
Trusted organization running the system	**26**	**6**	**5**	**17**	**4**	65
Methods of verification	**21**	**6**	**4**	**18**	1	60
Supported by stable funders	11	3	1	0	2	17
Validation by publishers	9	5	3	4	3	24
Author self-curation	9	5	1	4	2	21
Other	4	0	1	0	0	5
Validation by educators	1	3	0	0	1	5
	34	12	7	24	5	

Table 23: Factors that contribute to the trust in a PI system by stakeholder groups

Question 2.6: Have you exposed your author identifiers to search facilities for information discovery?

As shown in Figure 11, not only the use of PIs systems for authors and creators is at a level of immaturity, but also the use of technologies based on them, (such as those for information discovery) are still scarcely diffused among the participants.

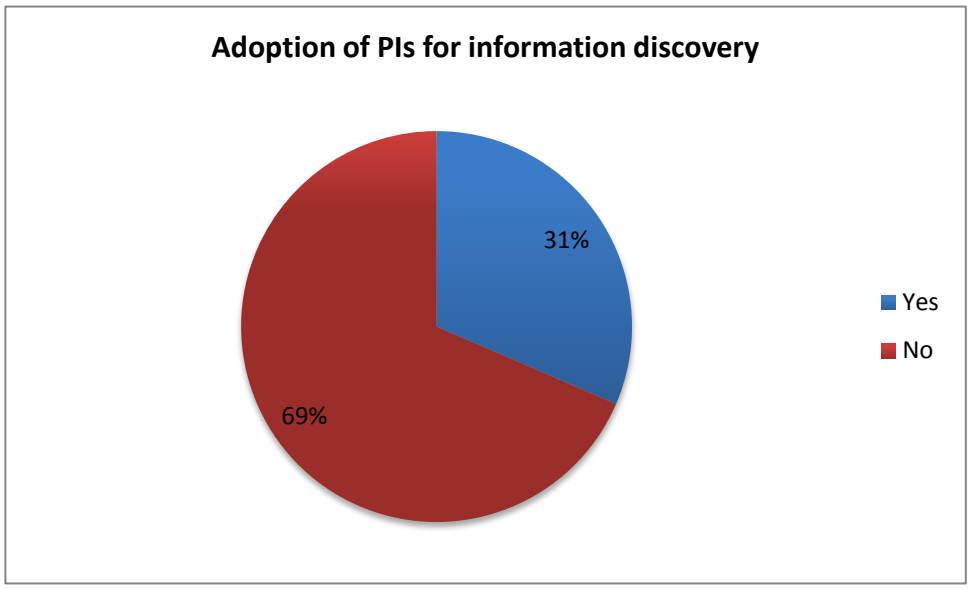


Figure 11: Use of PI for authors in search facilities for information discovery

SECTION 3: PI FOR ORGANISATIONS

Question 3.1: Which PI system do you use to identify your organisation? (Please tick all that apply)

As shown in Table 24, the use of identification systems for organizations is scarcely diffused among the participants and the level of granularity is low.
Almost 40% of the respondents declared to not use PI systems for organizations. Among the users, the most frequently adopted approach to identify organizations is by URLs. The analysis by stakeholder groups (see Table 25) shows that libraries are the only stakeholder group whose majority of respondents reported to use some PI systems for organizations. The most common PI system within this group is URL. For the other groups, the pattern of results is inverted, that is the majority of respondents claimed to not use PI systems for organizations.

About the level of granularity, Figure 14 shows that 65% of users of PI systems for organizations, report to not adopt identifiers for units of the organization, like departments, divisions and other subdivisions.

PI system	Freq	Percent %
None	40	39
URL	25	24
Other	13	13
MARC organization code	12	12
Handles for organization identifiers	10	10
URN	10	10
DOI	9	9
ISIL	7	7
OCLC	5	5
Fedora Identifiers	4	4
PURL	2	2
Total N. of respondents	103	

Table 24: Pi systems for organizations

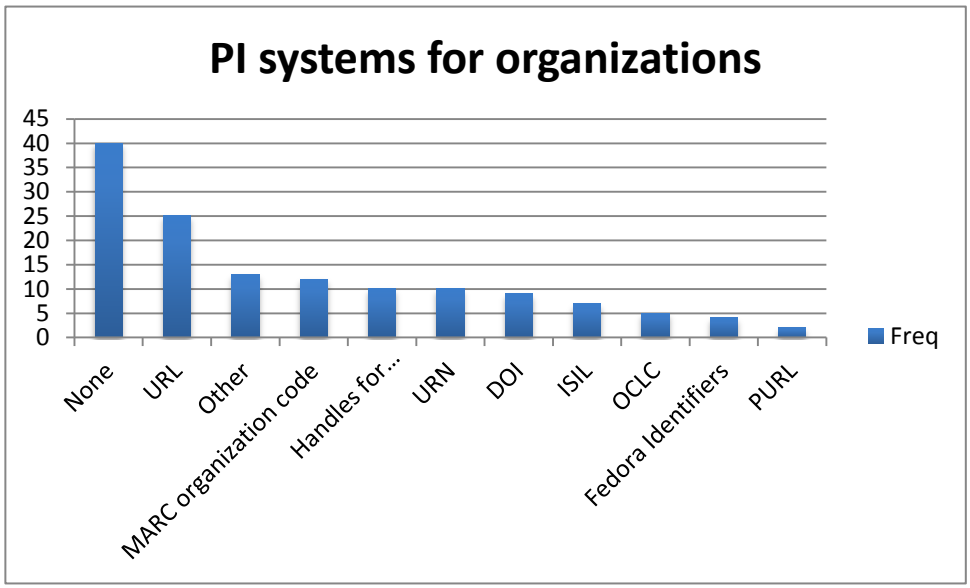


Figure 12: PI systems for organizations

	Libraries	Universities	Research organizations	Archives	Publishers	
None	5	**7**	**4**	**14**	**3**	33
URL	**8**	2	1	9	1	21
Other	7	0	0	3	0	10
MARC org. code	7	0	0	0	1	8
Handles	4	0	0	4	0	13
URN	7	0	1	1	0	9
DOI	0	3	2	2	0	7
ISIL	6	0	0	0	0	6
OCLC	2	0	0	1	0	3
Fedora Identifiers	1	0	1	1	0	3
PURL	0	0	0	0	0	0
	34	12	7	24	5	

Table 25: Pi systems for organizations by stakeholder groups

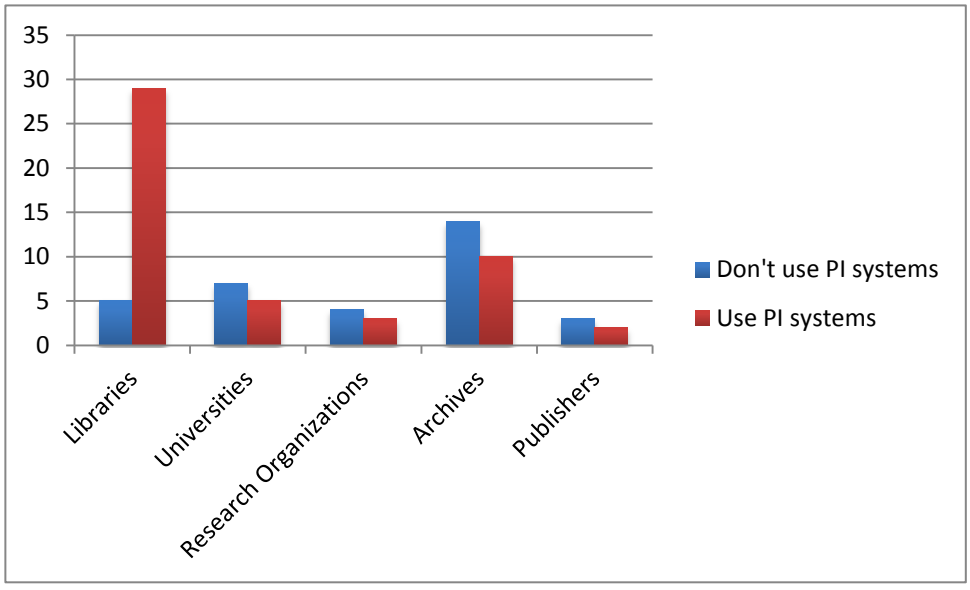


Figure 13: Use of PI systems for organizations by stakeholder groups

Question 3.2: Which are the main obstacles in adopting a PI system for organisations? (Please tick all that apply)

If we compare the obstacles that the respondents reported about the use of PI systems for authors with those about the use of PI systems for organizations, we can notice that the two most frequently selected obstacles are the same: the lack of awareness and the fact that the use of PI systems is not considered a key issue for the organization. This result confirms that one of the main action of intervention to promote agreement across the different stakeholder communities about the adoption of PI systems should start from increasing the level of awareness about the available systems and their potential positive effects. However, the users of PI systems for organizations seem to be slightly more aware of the potential beneficial impact of using PIs for information discovery compared to users of PIs systems for authors, as shown in Figure 15.

Value	Freq	Percent %
No enough information about this service	20	19
It is not a key issue for the organization	19	18
Low attractiveness of the service due to low level adoption	6	6
Lack of trust and authority	6	6
Other	4	4
Total N. of respondents	103	

Question 3.3: Does it also include identifiers for divisions, departments, or other subordinate units of the parent organization it serves?

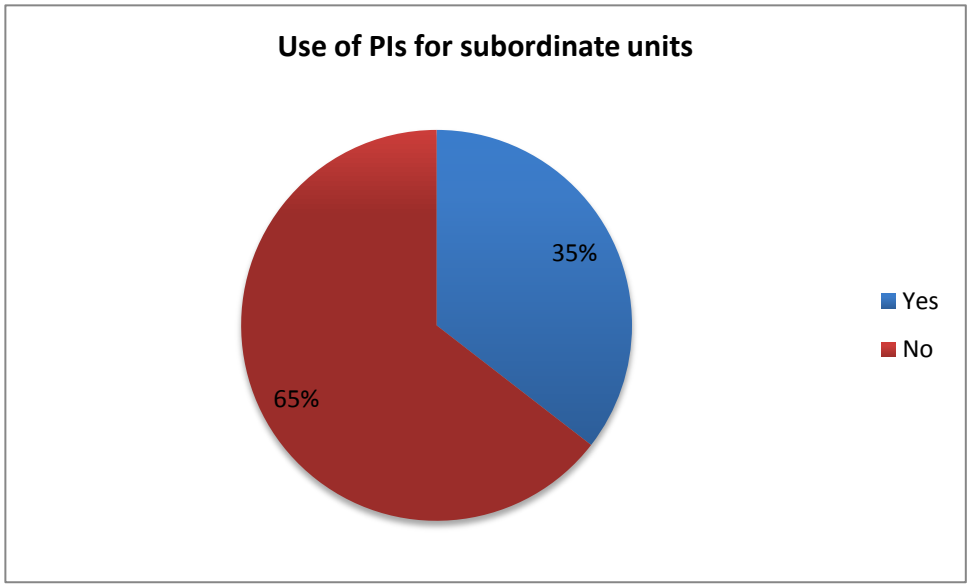


Figure 14: Use of PI for departments, divisions and other units

Question 3.4: Have you exposed your organisation identifiers to search facilities for information discovery?

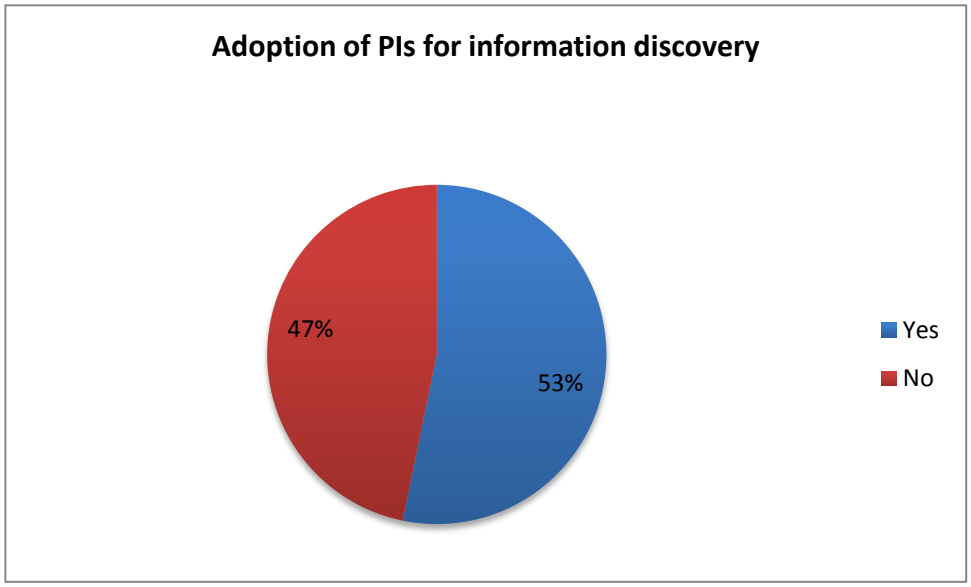


Figure 15: Use of PIs for organizations in search facilities for information discovery

SECTION 4: CRITERIA FOR THE ADOPTION OF A PI SYSTEM FOR DIGITAL OBJECTS

Question 4.1: Which of the following features are decisive to adopt a PI system for digital objects? Select one of the two coupled sentences.

We investigated the user requirements in 4 domains: technology, organization of the service, scope and naming rules.

In terms of technology, our results indicate that users prefer to adopt a system that represents a standard de facto, widely adopted and based on an open source infrastructure.

About the organization of the service, distributed naming authority and supported by an institution with a mandate were the preferred options.

In terms of scope, the respondents reported to prefer systems open to any digital objects and cross-community.

Finally, about naming rules opaque identifiers that support deep granularity are preferred than semantic identifiers supporting low-level granularity. No relevant differences were found between the stakeholder groups in the requirements for adopting a PI system for digital objects, as shown in Table 27.

Domain	Requirements	Freq	Percent %
TECHNOLOGY			
	Standard de facto	**55**	**53**
	Standard de jure	37	36
	Open source infrastructure	**91**	**88**
	Proprietary infrastructure	4	
	Widely adopted	**58**	**56**
	Established and mature	37	36
ORGANIZATION OF THE SERVICE			
	Distributed naming authority	**50**	**48**
	Centralized naming authority	39	38
	Supported by an institution with a mandate	**57**	**55**
	Supported by a stable funder	35	34
SCOPE			
	Openness (open to any digital object/actor)	**84**	**81**
	Closeness (aimed to a particular set of digital objects/actors)	11	11
	Cross-community identifier	**78**	**76**
	Community-oriented identifier	17	16
NAMING RULE			
	Opaque Identifier	**57**	**55**

	Semantic Identifier	36	35
	Deep granularity supported	**59**	**57**
	Low-level granularity	33	32

Table 26: Requirements for adopting PI system for digital objects

	Libraries	Universities	Research organizations	Archives	Publishers	
Standard de facto	**20**	**7**	**4**	10	**3**	45
Standard de jure	9	5	3	**11**	1	29
Open source infrastructure	**32**	**12**	**6**	**22**	1	73
Proprietary infrastructure	1	0	0	1	1	3
Widely adopted	**19**	**8**	**4**	11	**3**	35
Established and mature	11	4	3	11	1	30
Distributed naming authority	**16**	**8**	2	**13**	**3**	42
Centralized naming authority	14	3	**4**	7	1	29
Supported by an institution with a mandate	**4**	**6**	**4**	**11**	**3**	38
Supported by a stable funder	2	0	1	9	2	14
Openness (open to any digital object/actor)	**29**	**12**	**5**	**19**	**4**	69
Closeness (aimed to a particular set of digital objects/actors)	2	0	0	3	1	6
Cross-community identifier	**25**	**11**	**6**	**20**	2	63
Community-oriented identifier	5	1	0	3	2	10
Opaque Identifier	**20**	**7**	**4**	**17**	2	50
Semantic Identifier	9	4	2	3	2	20
Deep granularity supported	**16**	**8**	**5**	**16**	**3**	48
Low-level granularity	10	3	1	5	1	20
	34	12	7	24	5	

Table 27: Requirements by stakeholder groups

Services

Question 4.2: Which of the following services do you consider most important for your organisation? (Please indicate the top five issues driving your organisation's interest)

The majority of participants indicated citability as the most important service associated to the use of PIs, followed by services which support resolution (i.e. global resolution services, resolution to the resource or to metadata). More than half of participants reported services for digital object certification among the required services. The analysis by stakeholder groups shows that the three main services for libraries, archives and publishers are citability, global resolution services and resolution to the resource; those for universities are citability, global resolution, digital object certification and metrics, those for research organizations are citability, resolution to metadata and digital object certification. This trend seems to indicate that, if citabilty is a desired service for all the stakeholder groups, aspects related to the resolution mechanisms are more relevant for libraries archives and publishers, while aspects related to certification (and metrics) are more important for universities and research organizations.

Value	Freq	Percent %
Citability	76	74
Global resolution service	62	60
PI resolution service to the resource	57	55
Digital Object certification	55	53
PI resolution service to metadata	50	48
Association of PI to multiple location (URLs)	41	40
Metrics	31	30
Multiple association name	27	26
Link digital object to dynamic dataset	19	19
Others	3	3
TOT	103	

Table 28: Services

	Libraries	Universities	Research organizations	Archives	Publishers	
Citability	**25**	**11**	**5**	**16**	**4**	61

Global resolution service	**23**	**7**	3	**15**	**3**	51
PI resolution service to the resource	**20**	4	4	**13**	**3**	44
Digital Object certification	17	**7**	**5**	12	1	42
PI resolution service to metadata	18	5	**5**	8	1	37
Association of PI to multiple location (URLs)	16	4	1	11	1	33
Metrics	10	**7**	2	3	1	23
Multiple association name	9	3	1	4	1	18
Link digital object to dynamic dataset	6	3	1	5	0	15
	34	12	7	24	5	

SECTION 5: LONG-TERM (L. T.) DIGITAL PRESERVATION STRATEGIES AND PRACTICES

Question 5.1 Are l. t. digital preservation practices in use in your organization?

From our results, it stands out that the majority of the respondents, across all the stakeholder groups, use practices for digital preservation (see Figure 16 and Figure 17) and, among them, the Open Archival Information System (OAIS) is the most frequently used written reference model for digital preservation, as shown in Table 29. However a variety of alternative models appear to be used.

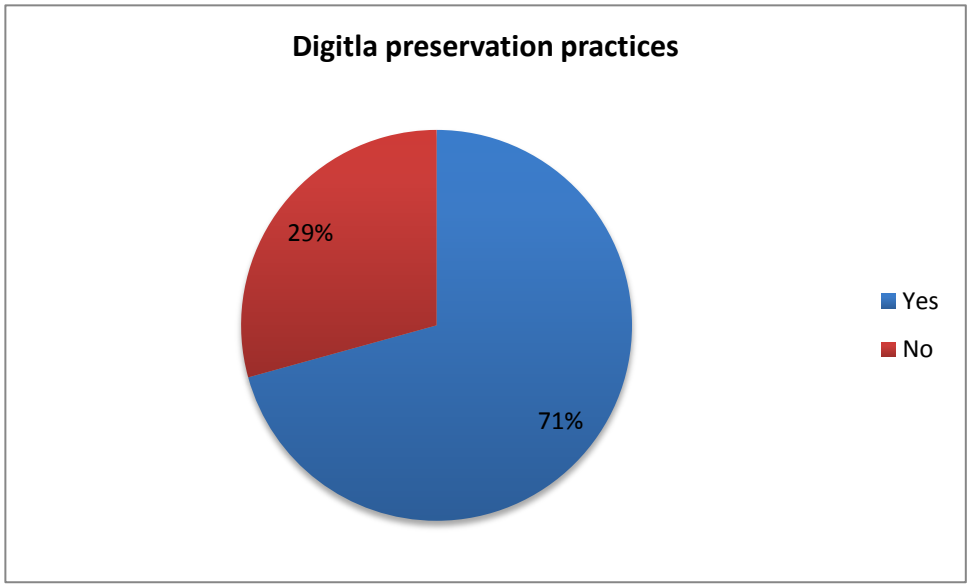


Figure 16: use of digital preservation practices

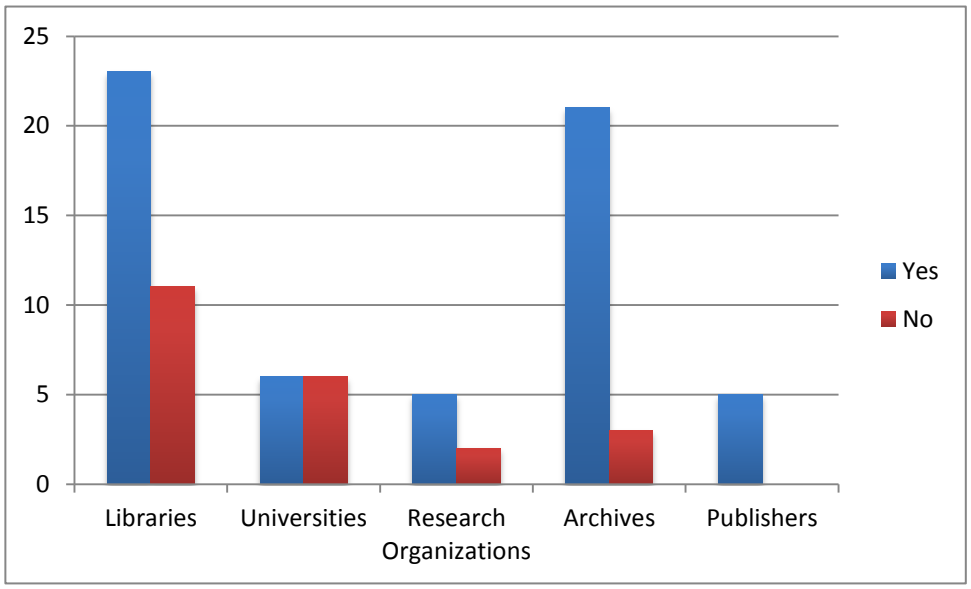


Figure 17: Digital preservation practices by stakeholder groups

Question 5.2: If yes, do you have any written guidelines for the long-term preservation of digital documents? (Please tick all that apply)

Value	Freq	Percent %
OAIS Reference Model	41	56
Long-term archive	25	34
Other	22	30
Collection of representation information	14	19
TRAC checklist	13	18
DRAMBORA	11	15
PLANETS suite	9	12
Dataseal of Approval	8	11
DINI certificate	6	8
European Framework for Audit and Certification of Digital Repository	4	5
LOCKSS	4	5
TOT	73	

Table 29: Guidelines for digital preservation

Duration

Question 5.3: How many years do you plan to preserve the digital objects? (Please tick all that apply)

In terms of duration of preservation, it is clearly indicated by the participants the requirement of preserving permanently their digital resources and this is valid across all the stakeholder groups and independently by the adopted PI system.

Value	Freq	Percent %
1-5 years	2	2
5-10 years	8	8
10-20 years	6	6
20-30 years	2	2
Permanently	**62**	**60**
Will hand over responsibility for preservation to another organisation at some point	5	5

Other	6	6
TOT	103	

Table 30: Temporal dimension of digital preservation practices

	Libraries	Universities	Research organizations	Archives	Publishers	
1-5 years	1	0	0	1	0	2
5-10 years	3	0	1	1	1	6
10-20 years	1	1	0	3	0	5
20-30 years	1	0	0	1	0	2
Permanently	**21**	**4**	**2**	**19**	**3**	**49**
Will hand over responsibility for preservation to another organisation at some point	0	1	1	1	1	4
Other	1	1	3	1	0	6
	23	6	5	21	5	

Table 31: Temporal dimension of digital preservation practices by stakeholders groups

	DOI	Handle	URN	Internal	None	
1-5 years	0	1	0	0	0	2
5-10 years	1	1	0	0	2	4
10-20 years	0	2	0	1	2	5
20-30 years	0	1	0	0	0	1
Permanently	**6**	**8**	**11**	**6**	**4**	**35**
Will hand over responsibility for preservation to another organisation at some point	1	0	1	1	0	3
Other	0	1	1	0	0	2
	12	13	15	7	13	

Table 32: Temporal dimension of digital preservation practices by PI systems

Digital preservation practices

Question 5.4: Which digital preservation practices are in use? (Please tick all that apply)

Among the respondents who reported to use digital preservation practices, transformation and multiple copies are the most common practices, followed by offsite storage used by more than half of them. In particular, the analysis by stakeholder groups shows that transformation is the preferred practice used by libraries and universities, offline storage by research organization and publishers, multiple copies by archives.

Value	Freq	Percent %
Transformation	**46**	**63**
Multiple copies	**45**	**62**
Offsite storage	39	53
Creation of Representation Information	27	37
Use third party service	17	23
Emulation	4	5
Other	4	5
TOT	73	

Table 33: Preservation practices

	Libraries	Universities	Research organizations	Archives	Publishers	
Transformation	**15**	**5**	2	15	2	39
Multiple copies	13	4	2	**16**	2	37
Offsite storage	12	3	**4**	12	**4**	35
Creation of Representation Information	8	1	0	10	0	19
Use third party service	5	2	3	4	3	17
Emulation	1	0	0	2	0	3
Other	1	0	0	0	0	1
	34	12	7	24	5	

In order to analyze whether there are differences between the adopted PI systems in terms of the preservation practices in use, we perform a further analysis filtering the data on the basis of the PI system. As in the case of the analysis by stakeholder groups, we considered only the answers without overlapping across the systems.

	DOI	Handle	URN	Internal	None
Transformation	3	**6**	**8**	4	4
Multiple copies	3	4	7	4	**6**
Offsite storage	**6**	5	**8**	3	1
Creation of Representation Information	0	1	0	6	2
Use third party service	2	2	5	0	1
Emulation	0	0	2	1	0
Other	1	0	0	0	0
	12	13	15	7	13

Offline storage seems to be a preservation practise largely adopted by DOI and URN users. The latter reported to adopt also transformation practices, as the Handle users. Participants who reported not using PI systems use multiple copies as the favourite preservation practise.

Funding and financial aspects

In terms of funding and financial sustainability, it is quite evident that there is a disparity between the reported need of preserving digital resources permanently and the scarce commitment in terms of financial sustainability. We found that around 50% of participants[37] reported a lack of a specific funding model for digital preservation practices. Around 33% of participants declared funding between 10K and 1M euros/year. As shown in Table 35, digital preservation practices are mainly funded from the organization budget (around 67%, aggregating the first and the second options in the table). But which are the stakeholders that have the highest budget for digital preservation?

From our analysis it stands out that archives are the stakeholder group investing more money in digital preservation.

Question 5.5: Has your organization specific funding for digital preservation activities?

[37] Option 1 (33%) + option 3 (17.5%) on Table 34.

Value	Freq	Percent %
Not at the moment but we have planned to do this	**34**	**33**
Yes between 100K and 1M euros/year	20	19
No – it is not a priority	**18**	**17**
Yes between 10K and 100K euros/year	14	13
Yes more than 1M euros per year	7	7
Yes between 1K and 10K euros/year	4	4
Yes, less than 1K euros/year	3	3
TOT	103	

Table 34: Funding for digital preservation

	Libraries	Universities	Research organizations	Archives	Publishers	
Not at the moment but we have planned to do this	**12**	**6**	**6**	5	1	39
Yes between 100K and 1M euros/year	8	1	1	**6**	2	37
No – it is not a priority	6	2	3	2	1	35
Yes between 10K and 100K euros/year	5	2	1	1	0	19
Yes more than 1M euros per year	1	0	0	**6**	0	17
Yes between 1K and 10K euros/year	2	0	1	0	1	3
Yes, less than 1K euros/year	1	0	0	2	0	1
	34	12	7	24	5	

Question 5.6: How is the long-term preservation of digital documents financed at you organisation?

Value	Freq	Percent %
Exclusively from the organization budget	**30**	**40**
Mainly from the organization budget	**20**	**27**
Mainly from third-party funds	10	13
Use third party service	6	8
Exclusively from third-party funds	5	7
Ratios are about equivalent	3	4
TOT	74	

Table 35: Source of funding

ANNEX II

PI SYSTEMS FEATURES ANALYSIS

Task leader
University of Trento (UNITN)

Authors:
Barbara Bazzanella (UNITN), Paolo Bouquet (UNITN),
Emanuele Bellini (FRD) Maurizio Lunghi (FRD) Chiara Cirinnà (FRD)

PI Systems Features Analysis

This section is divided into two parts. In the first part, the results of the feature analysis are reported distinguishing between PI systems for digital objects and PI systems for authors/creators. Since PI

systems for organizations are at a very immature stage of development and are still scarcely adopted, we limited the analysis to PI systems for digital objects and authors.

The second part reports the main results of the PI questionnaire.

Concepts

A fundamental characteristic of the digital era is that an increasing amount of digital assets are being produced and stored by research and education communities. There is an obvious need to discover, access and manage these resources over time, especially in e-science communities where rising tide of data are being generated. Therefore, it has become clear that identifiers are crucial keys for managing huge amounts of digital objects and related authors (and their institutions).

In e-Science, identifiers can be conveniently distinguished by the type of the referent entity in digital object identifiers, if the referent is a digital object, or author identifier, if the identifier is used to identify the author or contributor of a digital object. The reason for this distinction is that it is possible to locate and access a digital object while this cannot be done with authors. In other words, while in the case of digital objects, a digital identifier refers to a digital entity in the digital world, in the case of authors a digital identifier is used to refer to a representation of a real world entity. This distinction has important implications for the kinds of requirements in the two domains. Moreover, since the level of maturity of PI systems for digital objects and PI systems for authors is quite different, in this report the distinction has been used to frame the feature analysis.

PIs are conceptually composed of the following elements:

- **Resource**: The actual entity an identifier refers to. Entities of most concern to the study are originators (authors and contributors), scholarly works (publications to support scientific scholarship), and scientific data (e.g. measurement data and calculations). Entities can therefore be computation objects (i.e., information resources, digital in nature) or otherwise real world entities (i.e. non-digital entities).
- **Name:** A specialised name conforming to an understood syntax designed to at a minimum identify the Entity to which it refers, its Referent.
- **Protocol**: If part of an identifier specifies the location and modalities of the referent for subsequent access. This part is not mandatory.

There are various identifier schemes we have reviewed in the desk research activities. A digital identifier scheme can combine the following three elements:

a. a rule for identifier generation which is a mandatory element;
b. optional semantics related to the various strings composing the identifier;
c. optional, a protocol for accessing the referent object.

An **identifier management system** is a system that deals with identifying entities in a system by using identifiers and enables management of the full lifecycle of a resource by guaranteeing: 1) global uniqueness, 2) persistent reference to the resource and 3) reliable services for managing and using identifiers over time. In the system identifiers are used only as a way to make unambiguous reference to an entity and not as tokens to access the system (this allows to distinguish ID management systems from authentication services). Furthermore, "system" in this context refers to the governance, organisational and administrative procedures in addition to the technical implementations they use.

PIs system enabling technologies

In order to set up a PI system for online digital objects it is necessary a proper infrastructure. Presently there are three main infrastructure able to support a PI system implementation: Handle system, URI/URL, and URI/URN .

URI/URN based

URI is a universal way to name web resources which different granularity can be a file, a web page, a picture. URN is a type of URI defined by IETF Standard for the identification of web resources. URN describes the digital resource by naming it accordingly to a Namespace[38] Identifier (NID) which indicates the identifier system to apply and a Namespace Specific String (NSS) for naming the resource locally. The URN identifier is opaque in the sense that is not tied up with some changeable metadata about the referred resource. URN does not provide a way to locate the resource which can be located and so made it actionable by combining it with another type of URI called URL which also describes how to resolve a URN. So Interoperability between different URN lies on the ability to resolve them and for this reason there is not a build in feature in the URN scheme but there are many initiatives that are addressing this issue. The URN working group has announced to be restarted by IETF around middle 2011 for working on different issues such as interoperability between different implementations and new services.

HANDLE System

The Handle system is a well established infrastructure for providing resolution services for digital object identifiers. It covers a variarity of objects such as articles, books, documents, metadata, learning content and data sets. It does not demand the level of granularity an object should have to be identified. The most wide spread Handle implementation is DOI. Handle infrastructure charges a service fee. The system is composed of a set of protocols called handle which take care of storing and accessing a digital resource. Every identifier consists of two parts a prefix and suffix. The identifiers are opaque and not based on any changeable attributes such a location or ownership, for this reason provenance information is not encoded in the identifier itself. Scalability in the handle system is achieved by a hierarchical replicated model of two levels of services. There is a global service where namespaces are registered and local servers which manage identifiers registered under their own prefixes. Interoperability is ensured between different implementation based on the use of different protocols. Using the services of the network identifiers is resolvable.

URI/URL based

A URL (Uniform Resource Locator) is a URI that, in addition to identifying a resource, provides means to act or to obtain a representation of the resource by describing its primary access mechanism or the "location" ("location") in a the network. For example, the URL identifies http://www.rinascimento-digitale.it/ is a URI that identify an asset and at the same time provides the that a representation of that resource (the HTML for the current version of the home page) is obtainable via HTTP from a network host named www.rinascimento-digitale.it.Built upon these technologies several applications have been developed. In the section 5.2.2 a detailed description of these applications is provided. The table below shows the relation between these applications and their enabling technologies. In fact, the NBN applications family is based on the URN –based standards RFC 3188, the DOI and Datacite are based on the Handle infrastructure and PURL and ARK exploit the HTTP URL infrastructure.

Technology	Application

[38] A Namespace is an abstract container providing context for the items it holds and allows disambiguation of items having the same name (residing in different namespaces). An example is the URN namespace such as National Bibliography Number (RFC 3188- NBN).

URI/URN	NBN:DE, NBN:IT
Handle System	DOI, DataCite
URI/URL	PURL, ARK, XRI

PI systems for digital objects feature analysis

Before we describe the benchmarking analysis, we provide a short description of the main systems that have been assessed.

HANDLE System[39]

The Handle system is a well established infrastructure for providing resolution services for digital object identifiers. It covers a variety of objects such as articles, books, documents, metadata, learning content and data sets. It does not specify the level of granularity an object should have to be identified. The most wide spread Handle implementation is DOI. Handle infrastructure charges a service fee.

The system is composed of a set of protocols called handle, which take care of storing and accessing a digital resource. Every identifier consists of two parts: a prefix and suffix. The identifiers are opaque and not based on any changeable attributes such a location or ownership. For this reason provenance information is not encoded in the identifier itself.

Scalability in the handle system is achieved by a hierarchical replicated model of two levels of services. There is a global service where namespaces are registered and local servers which manage identifiers registered under their own prefixes. Interoperability is ensured between different implementations based on the use of different protocols. Using the services of the network identifiers is resolvable.

Digital Object Identifier (DOI)[40]

DOI is an implementation of the Handle system; it inherits scalability, decentralization and resolution capabilities and it applies a service fee. DOI has become an ISO standard (TC46/SC9) and through this process it will update some of its parts in order to be compliant with other standards such as the US National Standard ANSI/NISO Z39.84, Syntax for the Digital Object Identifier. DOI has been proposed as a system that fulfills this fundamental requirement, providing a generic framework applicable to any digital object. More specifically, any logical entity is a potential entity assigned a DOI, including tangible entities ("manifestations") or abstractions ("works"). A DOI identifier is an alphanumeric string that identifies digital content, such as an ebook, a journal article, a musical track and so on. The structure of a DOI is composed of two parts: the prefix and the suffix, separated by a forward slash / :

1. The Prefix or Publisher ID: is assigned by a DOI registration agency to the publisher. All prefixes begin with the number "10" and then a number designating the organization or publisher.

2. The Suffix or Item ID: is assigned by the publisher and can be made up of any alphanumeric sequences of characters as long as each object can be uniquely identified. It is important to note that a pre existing identifier (e.g. ISBN) can be incorporated into the suffix of a DOI.

[39] www.handle.net

[40] www.doi.org

Interoperability in DOI is managed at metadata level. The DOI approach assumes that managing resource interoperability requires not only DOI but also metadata that is a way to provide a description of what is identified in a structured way, allowing services about the object to be built for any purpose. Each DOI is associated with a series of metadata. All the metadata associated to a DOI are well formed if they are drawn from a controlled vocabulary of values, which are supported by a data dictionary in which those values are concisely defined. This allows that the values in one metadata scheme (or namespace) can be mapped to those in another scheme.

DOI system provides some tools and API for registering objects, associating metadata, searching and managing resolvers.

ARK

ARK is a URL created to allow persistence by naming digital objects. Every ARK identifier is formed by a sequence of characters, which starts with a protocol and host name called "Name Mapping Authority" (NMA) as optional information which makes the ARK actionable. In addition to the NMA there is what is considered the immutable part of the identifier, which starts with ark: label followed by a Name Assigning Authority Number (NAAN) identifying the naming organization and by the object name. The naming authority[41] is used to denote stewardship. ARK maintains a link to an object but also a descriptive record for the object to be used for verifying an object's identity after a brief inspection. ARK requests are based on http protocol and in addition there are tools for creating ARK identifiers such as NOID or "uuidgen".

Identity and managing of the digital resources are coordinated by a policy at the level of Naming Authority. The Naming Authorities are about thirty.

DATACITE

DataCite is a consortium which aims to support the access to research data and allow citability and recognition of contribution. The goal is to establish a not-for-profit agency for enabling organisations to register research datasets and assign PIs to them, so that research datasets can be handled as independent, citable, unique scientific objects. DataCite is a Global DOI Registration agency for scientific content, which offers dataset registration and cataloguing services. DataCite targets non-commercial information institutions and libraries instead of publishers. DataCite applies a fee for joining the system.

DataCite has chosen to assign DOI names as identifiers for scholarly publications, in order to cite and link datasets among themselves. Alongside DOI other identifier names could be used. DataCite provides a metadata[42] scheme for describing digital objects and the relationship with other resources; this is used for recommending a standard citation format based on a small number of required properties and for providing interoperability with other data management schemes.

The DataCite architecture from June 2011 consist in a central metadata registry, which will contain the descriptions from all objects registered by any DataCite member and it is freely browsable and searchable for any third party.

URN:NBN/DNB

[41] A naming authority is an independent authority that assigns names and guarantees their uniqueness and persistence. A naming resolution service corresponds to every naming authority and carries out the name resolution. In a PI distributed approach is foreseen that the responsibility of generation and resolution can be delegated to other institutions called sub-naming authorities who manage a portion of the name domain/space.

[42] Metadata: the term literally means "data about data". Metadata provide additional information about a certain digital object, such as its author, creation data (time and date), possible access restrictions or the application used to create the file.

The German national library (Deutsche National Bibliothek - DNB) has built a resolver for URN:NBN for Germany, Switzerland and Austria. DNB offers free of charge a URN resolver, and other services for assigning URN identifiers to resources and for managing links among resources. In this way the URN created using DNB services are also actionable.

URN:NBN/IT

The consortium of legal deposit in Italy (whose members are National Library of Florence, National Library of Rome, National Library of Venice and Fondazione Rinascimento Digitale) together with the Conference of Italian University Rectors (CRUI) are developing the URN:NBN service for the resources deposited at Libraries and consequently subjected to a digital preservation process. As a start the URN:NBN/IT will be focused on the resources coming from the university repositories such as PhD theses.

PURL

Purl is a persistent URL, which refers to a web resource. Purl aims to solve the broken reference problem: if the location of a resource changes it is possible to update the corresponding Purl.

A Purl identifier is composed of a protocol, a domain and a resource name. The protocol for resolving the identifiers is encoded in the Purl. A domain is encoded as a URL path refers to a maintainer. Domains have a hierarchical structure. The domain name together with the resource name constitutes the Purl's identifier, so provenance information can be encoded in the identifier itself. In order to create a PURL identifier it is necessary to create a domain. A set of metadata are associated with the identifier name and with the domain in order to allow maintenance.

Features

In order to perform the feature assessment of the described PI systems, we have identified a set of descriptive criteria for assessment. The definition of these criteria has been suggested by several studies [5] [16] [10] and aims to map some of the main features that will be useful for the design of the IF. In order to facilitate the reading and the comparison between the different systems, the results of the assessment analysis have been organized into a matrix.

Unfortunately, some features have not been discovered in the available system documentation. In these cases a question mark will appear in the matrix for the correspondent feature.

The following criteria have been considered in the feature matrix:

Actionability: This criterion is another propriety that can be provided by the system implementation. The actionability of a PI concerns the fact that the PI is linked to an infrastructure which is able to interact with the environment. Some PI systems are 'actionable' because are linked to protocols (e.g. HTTP) that are an integrated part of the identifier itself (e.g. PURL). In this sense, we can define this modality as "Direct" (D) while the opposite "Indirect" (I) is referred to a PI that needs the address of the related REST-ful resolution service as a "prefix" to be actionable.

Costs: This criterion indicates whether there is a fee for obtaining identifiers or for membership.

Deletable ID: Indicates if the identifier can be deleted once has been created and assigned

Policy for assignment: means that the registration authority follows well known policies and criteria for the PI assignment**.** The most common policy is the definition of Naming

rules. It is interesting to understand if the NAA defines a rule for generating the ID or leaves the final user free to decide a sequence of characters, which defines the identifier.

Policy for data provider: means that the NAA defines some requirements that the data provider has to meet to obtain an ID.

Resolution: This criterion describes the capability of a PI to be resolved to a single (S) object such as webpage or file, or to both object and metadata (B) or to multiple objects (M) such as different formats of the same objects, or different content types, through the same PI.

Right management: This criterion indicates the capability of a PI system of supporting DRM for the identified object.

Scalability: indicates the degree to which operationally the system can handle increasing numbers of referents.

In this section we describe some of the most important current digital object identifiers systems.

The following table provides an overview of the feature of the systems based on the features described above.

PI System	Action.	Costs	Delet.	Policy for assignment	Policy for data provider	Resol.	Right management	Scalability
DOI	D+I	yes	Yes	Yes	No	B+M	Yes	Yes
Handle		Yes	Yes	Yes	No	B+M	Yes	Yes
ARK	D	No	Yes	Yes	No	B	No	Yes
DataCite	D+I	yes	No	Yes	Yes	B+M	Yes	Yes
NBN/DNB	I	No	No	Yes	Yes	B	No	Yes
NBN/IT	I	No	No	Yes	Yes	B	No	Yes

PURL	D	No	Yes	No	No	S	No	Yes

Feature matrix for Digital Object Identifier Systems

The feature analysis results shows several differences among the systems that can drive the choices of institutions aimed in adopting a PI for their objects. It is clear that these systems address different requirements according to their business model, level of service that they aim to provide and their designated community. [43] For instance, if a community needs to manage directly their own PI and a requirements is the possibility to delete a PI, some systems like DOI, Handle and PURL allow this, others like NBN systems do not it. Again, if the an institution requires a multiple resolution service for their resources, they might adopt the DOI or Handle solutions. The feature analysis results suggest that there is not a PI service clearly better than another and even if some overlapping exist, they address different user requirements and in several cases different communities. The only exception is for the PURL solution. In fact, according to the benchmark result, PURL does not seem be a good solution for supporting a reliable persistent identification of resource that require high level of trust such scientific and cultural heritage objects. The level of service provided by this solutions, that was the first initiative on the field, is now too basic for addressing the identification requirements highlighted by the survey results (see Chapter 5).

Moreover, each PI tends to be linked to a specific phase of the resource lifecycle. For instance once a resource is published on a journal supported by a publisher usually receives a DOI, the same resource deposited in an Open Archive can receive also an Handle; if the resource is forwarded to the National Library for preservation will receive an NBN, and so on. All these PI associated to the same resource inform the user about the existence of several copies of the resource accessible from different resolution services that can be represent different reuse conditions. Therefore, the presence of more PI systems is something of desiderated, the challenge is to make these systems, interoperable.

PI systems for Authors benchmarking

As in the previous section, we first provide a brief description of the PI systems for authors, which have been included in the benchmarking analysis.

Scopus

The main goal of Scopus Author Identifiers system is to address identity author issues. Scopus is a commercial system promoted by Elsevier.

When an author's works are added to the Scopus database, an author profile is created. The profile includes the author's name, a Scopus author identification number (Author ID), citation statistics, subject areas in which the author is publishing, and the author's publication and affiliation history.

Scopus Author Identifier functionality distinguishes between names of authors by assigning each author in Scopus a unique number and grouping together all of the documents written by that author. It also groups the name variants together, so that authors, even if cited differently, are identified with their specific papers. To determine which author names should be grouped together under a single identifier number, the Scopus Author Identifier uses an algorithm that matches author names based on their affiliation, address, subject area, source title, dates of publication citations, and co-authors.

[43] OAIS Designated community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.

Researcher ID

ResearcherID is a free author identification system offered by Thomson Reuters. The system has been advertised as"a global, multi-disciplinary scholarly research community" where each author is assigned a unique identifier, name ResearcherID. By means of a personal profile, it is argued; an author can resolve identity issues but also manage and share professional information and add dynamic citation metrics and collaboration networks to the personal profile. This allows the author to increase the visibility of his/her work. In the profile the author can register information about institutions he has worked for, research areas of interest, descriptive text, keywords, role (e.g., academic researcher, student, or librarian), and contact information. Moreover, the author can import his/her publications list from Web of Knowledge, EndNote/EndNote Web, or the generic RIS citation format produced by many other personal citation managers. ResearcherID is a commercial system but it is announced as a free and open system. Open system means that anyone can search.

The registry and view public profile to find collaborators, review publication lists and explore how research is used around the world. The owner of a profile has full control over what information appears in your public profile. ResearcherId is a unique identifier that consists of alphanumeric characters. Each number contains the year of registration. There is no independent verification of authorship for articles in ResearcherID and there are no methods to avoid that an author registers multiple times.

AuthorClaim

AuthorClaim is an open source solution with the same objectives of Re-searcherID.com that is a free author registration system that links scholars with their publications. During registration, one provides an e-mail address, full name, and institutional affiliations. Name variations are generated by the system which can be edited by the registrant. After entering this information the system searches a set group of databases for candidate articles and associates them to the profile. Users of bibliographic databases that use AuthorClaim record can link right to the author's profile page or homepage. The system allows the author to get regular statistics about downloads and citations of his works. AuthorClaims aims to extend the RePEc model, using the same software infrastructure; to the entire academic domain. The system generates the author's profile automatically: this represents one of the main differences between AuthorClaim and Re-searcherId.

There is no way to browse or search profiles, even by name. Unless researchers have publicized their AuthorClaim URLs.There is no mechanism to manually add publications that are missed by the system. You can not submit documents to AuthorClaim. Identifiers are used for the identification of an author's profile not for publications.

ArXiv Author ID

ArXiv Author ID is an example of a discipline-specific initiative promoted by the Cornell University Library with the aim to accurately identify and disambiguate all authors of all articles in arXiv which is an archive of e-prints publications in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics. The system requires that users must opt in to have a public author identifier and to expose the record of their articles on arXiv for use in other services. However, a future improvement of the system will create public author identifiers for all authors of arXiv articles without needing to enlist the help of each author to check their record before opting in.

DAI Digital Author Identifier

The Digital Author Identifier (DAI) is a unique national number assigned to every author who has been appointed to a position at a Dutch university or research institute or has some other relevant connection with one of these organisations. The DAI brings together different ways of writing the author's name and distinguishes between authors with the same name. DAI is also a non commercial solution. SURF is responsible for the governance and strategic developments of the overall identifier

system. The system consists of a central component - the National Thesaurus for Author names (NTA), part of the Shared Cataloguing System (GGC) - and decentral components - the local Current Research Information System (CRIS) - located at each university. Currently, all universities have implemented the METIS system as their CRIS solution.The NTA contains authors who are registered by Public Libraries and Research Libraries who are member of OCLC (Previously OCLC-PICA).Population of the data is done by qualified cataloguers at libraries or specialist departments, and employees who administer the CRIS currently used by all Dutch universities, METIS.

Changes in the central system may be made by cataloguers at university libraries and research information departments with access to WinIBW or WebGGC and METIS administrators, through a special version of the WebGGC. Note that each research institute is only allowed to make changes to the record-part associated with their own individual institution.

Feature analysis

The problem of uniquely identifying authors of digital contents (especially scientific contents) has recently become an important issue within the e-Science environment. Many different stakeholders, such as researchers, institutions, universities, publishers, funding agencies more and more recognize the benefits of using unique identifiers not only for scholarly work but also for the creators of (and contributors to) that work. However, the different involved communities of stakeholders often have very different views about the features that a scheme or system for author identification should have. For example, an author identification system should play an important role in computing academic metrics. These metrics can be used by the universities to evaluate their scholars or by the funding agencies to make decisions on funding and can strongly influence the reputation of an author. Therefore, while an individual author could prefer to control the information associated to his/her identifier, universities or funding agencies could prefer a system that uses claims by external sources or that automatically integrate the information from other identification systems. Moreover, since in this context we are dealing with identifiers for people (i.e. non-digital resources) the problem of building an author identifier system is complicated by social challenges such as privacy, security, self-curation, reputation and many other important issues.

In this section we report an example of feature matrix for author identifier systems. Not all the features have been discovered in the available system documentation, so in this case a question mark will appear in the matrix for the correspondent characteristic. A more detailed analysis of the feature will be conducted during the face to face interview with the stakeholders whom hopefully will fill up the gaps of the documentation.

The following features have been considered in the next matrix:

- **Auto-claim**: offers services for claiming rights on the digital content, e.g. services for authors to claim authorships on scientific publications.
- **Centralization:** indicates that there is a single centre of authority.
- **Costs:** indicates whenever there is a fee for identifiers or membership required.
- **Level of Uniqueness:** indicates the scope of the identifier uniqueness. An identifier can be unique at repository level (R), in a federation or among a group of sources (F) or at global level (G).
- **Service provider:** indicates that the system also provides for additional services based on IDs and related contents, e.g. ranking and statistics about popularity and citations, searching for entities and metadata.

The following matrix summarizes a feature analysis of the systems

System	Auto-Claim	Centralization	Cost	Level of Uniqueness	Service Provider
SCOPUS Author ID	no	yes	Yes	R	yes
Researcher ID	yes	yes	No	R	yes
AutorClaim	yes	yes	Yes	R	yes
ArXiv Author ID	yes	yes	No	R	yes
DAI	no	yes	No	F	yes

Feature Matrix for Author Identifier Systems

ANNEX III

GLOSSARY

Task leader
University of Trento (UNITN)

Authors:
Barbara Bazzanella (UNITN), Paolo Bouquet (UNITN),
Emanuele Bellini (FRD) Maurizio Lunghi (FRD) Chiara Cirinnà (FRD)

Glossary

Access system	is the mechanism that provides the ability to interact with a system, to retrieve relevant information (e.g. digital objects) and use this information
Archive	An organization that intends to preserve information for access and use by a Designated Community
Author identifier	is a unique expression that makes it possible to disambiguate authors from each other. The use of these Ids has been recognized as a fundamental issue to establish the identity of authors and other contributors and reliably link them to their published works.
Authority	An authority for an identifier is a party that has a mandate for assigning and maintaining the naming service and it is responsible for it. This responsibility can be assigned by a community mandate or institutional mandate.
Centralized Naming authority	Identifier management for a range of authorities is centralised if all authorities manage their identifiers through a common identifier management system, hosted on their behalf by a central party.
Citability	An **entity** is cited if its representation is communicated to an audience through some medium. The entity is citable if it can be cited. For example, citing the **identifier** (("**Handle** server 102.100.272", "XYZ"), "PILIN policy on citation") means coming up with an appropriate representation of the identifier (e.g. hdl:102.100.272/XYZ), and embedding that representation in a PDF(PILIN).
Curation	Curation describes a range of activities and processes done to **create**, manage, maintain, and validate an object fpor long time period.
Distributed Naming authority	In a decentralized identifier management system, there is no single centralized authority that assign and manage the naming service on behalf of all the parties. Instead each party, also called a peer, make a local autonomous management according to a minimum shared rules. Peers directly interact with each other and share information or provide service to other peers.
Digital object	An object composed of a set of bit sequences (OAIS). Pragmatically, it is a unit of information that can be identified, such as anything that might be stored in a digital repository. Examples of Digital Objects include documents, articles, books, images, web pages, applications, audio files, raw data, database. A digital object is assumed here to belong to at least one digital repository.
Repository system	a system in which digital objects are stored for possible subsequent access, retrieval and management (e.g. DSpace, Fedora, etc.). Place where digital resources are held with or without a resource management system (file system).

Granularity	Granularity refers to the level of detail at which PIs will need to be or may be assigned. In some situations, it may be necessary to cite a web page which serves as access to a collection of web files, or to cite a journal article, an item, or a chapter or a subset of a data file or perhaps a result of a database query. However, due to rights management, some finer details may be required. Each institution would need to evaluate whether a PI service provides the right level of granularity for their type of resources.
Identifier (ID)	it is an expression composed by one or more characters, digits or codes, that uniquely identifies an object. Identifiers can be local or global. Local identifiers uniquely identify entities in a given context or system (e.g. the employee IDs used by a company), whereas global identifiers identify entities across systems and contexts (e.g. ISBN).
Identifier scheme	is a scheme that defines the characteristics of an identifier, such as, for example, the syntax used to create the ID, the information and the kinds of metadata that can be associated to it, if the ID is resolvable, if it is language-dependent, how it is assigned and so on.
ID management system:	is a system that deals with identifying entities in a system by using identifiers. In the system IDs are used only as a way to make unambiguous reference to an entity and not as tokens to access to the system (this allows to distinguish ID management systems from authentication services described below).
Long Term Preservation	The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term (OAIS).
Metadata	the term literally means "data about data". Metadata provide additional information about a certain digital object, such as its author, creation data (time and date), Representation Information, Preservation Description Information (PDI), including possible access restrictions or the application used to create the file. XML is a a standard to add metadata to documents and make them machine-readable.
Namespace	an abstract container providing context for the items it holds and allows disambiguation of items having the same name (residing in different namespaces). The namespace are registered by Internet Assigned Numbers Authority (IANA) and are defined by IETF-RFC where is identified also the naming authority. Examples is the URN namespace such as National Bibliography Number (RFC 3188-NBN) under the responsibilities of National Libraries.
Naming authority	Independent authority that assigns names and guarantees their uniqueness and persistence. A naming resolution service corresponds to every naming authority and carries out the name resolution. In a PI distributed approach is foreseen that the responsibility of generation and resolution can be delegated to other institutions called sub-naming authorities who manage a portion of the name domain/space

Opaque PI	A semantic PI is referred to the capability of extracting meaningfulness from the identifier. Examples are the mnemonic-based identifiers rather than those that contain a meaningless character sequence, although this has no relevance to machine processing.
Semantic PI:	A semantic PI is referred to the capability of extracting meaningfulness from the identifier. Examples are the mnemonic-based identifiers rather than those that contain a meaningless character sequence, although this has no relevance to machine processing.
Persistent	A component is persistent if it is managed and maintained for a defined timespan. Maintaining the component includes ensuring that its **published** content (such as its **association data**) is valid at all times. Normally when an identifier is called persistent, persistence of association is meant.
PI (PI	it is a maintainable identifier that allows to refer to and have reliable access to a **digital object** over long periods. A PI has to be always resolvable through a resolution system
Proprietary system	is a system which relies upon software and hardware which are licensed from a copyright holder.
Resolution service (dereference):	An identifier is resolved by providing information on how to access the **thing** it identifies. This information is the **resolution** of the identifier: it is the output of the Resolve action (PILIN) In other words it is the process in which an identifier is the input (a request) to a service to receive in return a specific output (resource, metadata, etc).
Trustworthy Digital Repository (TDR)	repository which has a current certification.(ISO 16919)
Versioning	A versioning of a **digital object** is an abstraction fixing the content but not the appearance of the digital object. Two instances belong to the same version if they have the same content; they belong to different version if they have different content, but are still seen to be underlying the same **thing**. Versions may include revisions, transformations, translations, and so forth. Expressions in the FRBR model are a type of version.
URI	A Uniform Resource Identifier is the generic set of all names/addresses that are short strings that refer to resources
URL	A Uniform Resource Locator is a URI that, in addition to identifying a resource, provides means of acting upon or obtaining a representation of the resource by describing its primary access mechanism or network "location"
URN	A Uniform Resource Name is a URI that uses the URN scheme, and does not imply availability of the identified resource. URNs are intended to serve as persistent, location-independent resource identifiers and are designed to make it easy to map other namespaces (that share the properties of URNs) into URN-space. Therefore, the URN syntax provides a means to encode character data in a form that can be sent in existing protocols, transcribed on most keyboards, etc. (IETF-RFC1737).