A Persistent Identifier e-Infrastructure

Barbara Bazzanella
University of Trento
Via Sommarive 5, I-38123
Trento, Italy
barbara.bazzanella@unitn.it

ABSTRACT

Persistent identifiers (PIDs) have been recognized as a crucial enabling component for 2020 e-science infrastructures[1], having the potential of providing global keys for information access, reuse and exchange and creating a complex network of links which connect all the relevant entities in the research data landscape (e.g. digital objects to authors and datasets, authors to institutions and projects, projects to research products and fundings). The creation and full exploitation of this valuable network of connections is currently hindered by the fragmentation and lack of coordination of the persistent identifier ecosystem. Several initiatives have emerged with the aim of offering global identifier repositories for digital and non-digital entities but they are still focused on the needs of specific communities and the lack of interoperability between them is one of the major hurdles for the development of a globally connected scholarly infrastructure. The aim of this paper is to propose a Persistent Identifier e-infrastructure (based on an identifier service called Entity Name System) which provides a technical layer of interoperability which allows current identifier systems to interoperate and be coordinated across geographical, temporal, disciplinary, organization and technological boundaries. The Persistent Identifier interoperability e-infrastructure is presented as a cross-cutting core service enabling the development of advanced added-value services tailored to the specific needs of different communities and stakeholders of the e-science environment.

General Terms

Infrastructure

Keywords

persistent identifier e-infrastructure, interoperability, e-science research infrastructures, Entity Name System

[1] http://ec.europa.eu/programmes/horizon2020/en/h2020-section/european-research-infrastructures-including-e-infrastructures

1. INTRODUCTION

Science is global in scope, but it is only recently with the development of advanced information and communication technologies, that science is becoming global in practice. ICT-based infrastructures for science (i.e. e-science infrastructures) are at the root of this process, promoting the realization of an integrated information space where researchers can cooperate and share resources independently from their geographical location, and the access to increasing volumes of data and their processing is facilitated and empowered, making science more efficient and innovative. These infrastructures provide tools and services to support the full life cycle of scientific data (to gather, capture, transfer and process data), the dissemination of data across the boundaries of nations and scientific disciplines, the cross-linking of data in the digital space, the integration between scientific data and publications. According to the framework proposed by the High Level Expert Group of Scientific data [11], e-science infrastructures can be seen layered systems where different actors, data types and services interrelate within a global space and community services specific to each community or discipline rest upon common low level services cutting across the global system. A solid infrastructure for managing unique identifiers for all the entities involved within the global scientific data infrastructure - including digital objects, authors, contributors, datasets, funding agencies, projects and many others - is a a critical low level service to provide the layer of interoperation and trust of data necessary to enable access, use, reuse and exchange of data (see Figure 1) in a collaborative integrated research environment [5].

However, since a number of different identifier systems with different scope, level of maturity and technical sophistication are already in use by different communities and no single integrating identifier system seems meet the needs of all the communities and provide a service to identify all the relevant entities which populate the articulated network of connections within the research arena, the identifier infrastructure should not only provide a layer for assigning identifiers to resources and managing them, but it should provide an interoperability infrastructure which makes existing identifier systems able to interoperate and be integrated without the need to introduce a further identification solution in addition to those already consolidated and adopted by the different communities. The development of an interoperable identifier infrastructure is an essential step for unlocking the value of research data and creating a digital globally connected

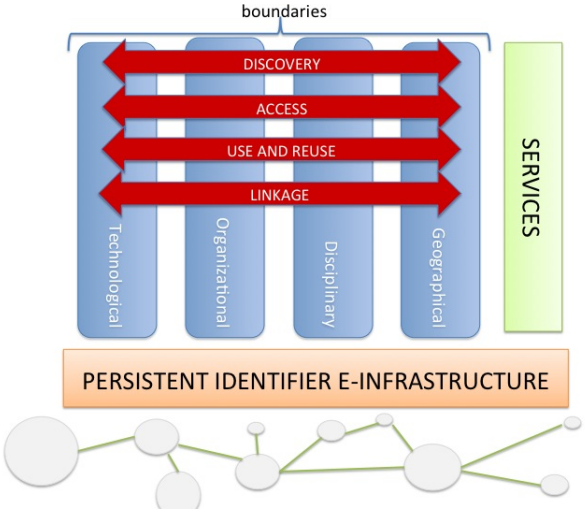


Figure 1: A Persistent Identifier e-Infrastructure

research environment in the near future. Even though, as pointed out in the DIGOIDUNA study [5], this is far from being a merely technical issue, opening a multidimensional spectrum of challenges dealing with economic, societal and policy aspects which need to be integrated into a coordinated model, the technical implementation of the agreed framework is an unavoidable step to secure the concrete and efficient operation of the infrastructure. This paper proposes a technical infrastructure exploiting an existing solution for managing global identifiers (called Entity Name System) which aims to provide a technical layer of interoperability allowing current identifier systems to interoperate and be coordinated across geographical, temporal, disciplinary, organization and technological boundaries. The Persistent Identifier interoperability e-infrastructure is designed as a cross-cutting core service enabling the development of advanced added-value services tailored to the specific needs of different communities and stakeholders of the e-science environment.

2. FROM URLS TO PERSISTENT IDENTIFIERS

The ability to reliably identify and locate digital information over time has become increasingly relevant in recent years in distributed digital environments. The Web infrastructure offers a very direct way to locate digital information based on the Uniform Resource Locator (URL). The URL specifies the physical location on a particular server from which to retrieve the digital resource (which could be a digital document, a dataset, an image, a video or any other digital resource on the Internet). However, since the Web is highly dynamic and resources are often moved to different locations during their lifecycle, the identification of digital content through URLs has proven to be a very fragile mechanism. When a digital object is transferred to a different destination or it goes off-line, the corresponding URL ceases to identify and locate the object and the link becomes "a broken link". Moreover, if the location where the object was initially stored, is subsequently occupied by a different object, the corresponding URL could be used to locate two different resources at two different moments of time. This explains why URLs are only temporary identifiers and cannot be used to provide ongoing access to digital resources.

Persistent Identifiers (PIDs) have been introduced as a solution to address this issue providing an identification mechanism in which the identifier is not strictly bound to a specific digital location. Unlike a URL, a persistent identifier is a permanent association between a unique name and an information object which can be the resource itself or a representation of it (i.e. metadata describing it). This association is maintained independently of the physical location of the information object. If the location changes, the persistent identifier still remains the same providing a different way to retrieve the resource (e.g. a different URL where the object is placed) or an appropriate representation of the resource. Indeed persistent identifiers can be used to identify both digital and non digital entities (e.g. people). Even though at first persistent identifiers were mainly used for identifying digital content (publications and scholarly works for example), it has become increasingly evident that many non-digital resources need to be uniquely identified in order to extract value from the representation of digital assets. In the scholarly domain, for example, the need to unambiguously represent authors and contributors and associate them with their scientific outputs (e.g. publications, datasets, software), has favored the development of several author identification systems. More recently, other initiatives like the I^2 (Institutional Identifiers) working group[2] have started to define a standard for an institutional identifier by proposing to leverage existing solutions like ISNI.

Many different persistent identifier solutions (e.g. URN, Handle, DOI, ARK, PURL, ISNI, ORCID) have been proposed in recent years which aim to reproduce in the digital environment the two main functions that traditional identifier systems provide in other cultural contexts (like identifiers for books in traditional libraries), i.e. **identification** and **access**. Identification means using a label to name an object and distinguish it from other similar objects. Persistent identifiers aim to identify resources in 1) unique, 2) location-independent, 3) persistent way. This means that 1) a persistent identifier is only assigned to a single object and never reused within the domain of creation, 2) a persistent identifier is not intrinsically bound to the location of the object; 3) the association between the identifier and the object should be maintained over time. Identifiers that are designed simply to identify resources have little utility in the digital world. The second requirement of persistent identifiers is that they operate as durable keys to access to digital content. As we have stated above, access to the identified resources (or information about them) should be guaranteed over time. This is usually realized through different strategies, like a layer of indirection within the HTTP protocol (e.g. PURL, ARK), a resolver mechanism dissociated from the HTTP protocol (e.g. Handle, DOI, URN) or conferring stability to Web identifiers (e.g. Cool URIs). More importantly persistent access is ensured thorough a complex social and organizational infrastructure of policies and rules involving registration agencies and content providers (see for example the social infrastructure of registration agencies coordinated by the International DOI Foundation which reg-

[2] `http://www.niso.org/workrooms/i2`

ulates the DOI system).

2.1 The current landscape of Persistent Identifiers in science

Identification and long-term accessibility are fundamental in most sectors of human activity, but are crucial for scientific information management especially in recent years due to the rising growth of scientific production, the digitization of content and the distribution of data and services across different systems and networked infrastructures.

The consistent adoption and use of persistent identifiers is a critical step for all the main phases of scientific production and fruition of its products on a global scale. Experimental data should be collected, discovered and shared within a global scientific community and across different science domains, data should be uniquely attributed to the people who contributed to their generation and connected with scientific works, projects and publications. Authors should be uniquely identified across disciplines and other boundaries and associated with their entire scientific production and linked to their professional activities (e.g. projects, events, teaching experiences) and membership institutions. Persistent identifiers have been recognized as fundamental building blocks for enabling accessibility, trustworthiness, provenance and quality assessment in e-science. This explains why assessing the impact of the use of different identifier solutions for digital objects, authors and other relevant entities has become a critical issue for policy makers and funding agencies especially when they aim for the realization of large-scale ICT infrastructures for e-science as the fundamental scientific production environment. This attention is confirmed by the recent EU Framework Program for Research and Innovation (Horizon 2020) in the area of Research Infrastructures[3], which envisions the development of a digital identifier infrastructure for digital objects and authors as a core service across e-infrastructures.

However, widespread adoption of persistent identifiers is far from being realized and the level of maturity and technical sophistication of the current identification solutions is widely diversified. While identification systems are well established in some specific domains and for certain kinds of resources (e.g. DOI for scholarly and scientific publications, URN for digital resources in many libraries and institutional repositories, ARK for digital objects in traditional and digital libraries), persistent identifiers are only recently (and quite slowly) emerging for other entities in the scientific domain. The introduction of non-ambiguous and persistent identifiers for authors and contributors is quite a recent practice, which have started to produce a number of local (sometimes national) ad hoc solutions in specific domains or systems (e.g. DAI in the Dutch Research System, author identifiers in arXiv, Scopus Author id developed by Elsevier, ResearcherID developed by Thomson Reuters). It's only recently that we are assisting to the development of more global integrating solutions for identifying authors and contributors across systems (e.g. ISNI, ORCID). Other identifier solutions (e.g. DOI through DataCite) have started to be adopted for identifying complex scientific entities, like datasets. Even more recent are persistent identifiers for institutions (e.g. Ringgold in the publisher domain). Another aspect of the current persistent identifier solutions is that resources can be part of different domains and can be identified by different identifiers in different systems. The same digital object which is assigned a DOI in the publishing domain can be assigned a URN within an institutional repository. Nowadays there is no overall integrating solution to map and retrieve different identifiers for the same resource and link a resource to all the entities (in turn identified by other persistent identifiers) with which it is interconnected. This makes hard to reuse identifiers across domains, integrate metadata from different sources and create integrating cross-boundary services based on different identification systems.

From this brief overview, two aspects of the persistent identifier landscape in e-science emerge: 1) **the fragmentation of the ecosystem** populated by a number of identifier solutions not equally diffused and consolidated 2) **a lack of an interoperability solution** for current persistent identifier systems which are nowadays difficult to integrate to offer interconnected services.

2.2 Toward Interoperability for Persistent Identifiers

In the last few years a number of initiatives and projects have started to create the ground for the realization of a global interoperable e-science framework based on the interoperability between identification systems. A study conducted on behalf of the European Commission, named DIGOIDUNA [5], has investigated the fundamental role of digital identifiers as enablers of value in e-science infrastructures and has performed a detailed analysis of strengths, weaknesses, opportunities and threats of the current digital identifier landscape in order to identify the main challenges and a set of recommendations which policy makers and relevant stakeholders should address to develop an open and sustainable persistent identifier infrastructure supporting information access and preservation. One of the main conclusions of the study is that to transform digital identifiers from simple means to manage data to keys for supplying knowledge and deliver value to the stakeholders within the research production, it is necessary to foster the development of an interoperable, cross-domain infrastructure for persistent identifiers supporting data access and sharing across national, organizational, disciplinary and technological boundaries. The implementation of this infrastructure poses several technical challenges but raises also a multidimensional spectrum of organizational, social and economical issues which should be addressed to ensure a coordinated ecosystem. Within the APARSEN project, the research on persistent identifiers has focused mainly on the definition of an interoperability framework for persistent identifier systems [1] which defines some key assumptions and requirements to identify the trustable candidate systems which can take part in the framework, an ontology which specifies the structure of data and the core set of relationships linking the identified entities within the framework and finally a small set of services which can be implemented on top of the framework. A demonstrator has also been developed to provide evidence of the potential applicability of the model and related basic services [2].

[3] http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-infrastructures_en.pdf

Other initiatives have started to define cooperation agreements and complementary architectures to ensure interoperability between independent systems or organizations. ORCID and ISNI for example have made a first advance in this direction by rendering ORCID compatible with the ISNI ISO standard and assigning a block of numbers for identifying ORCID entities which cannot be reassigned by ISNI to different people[4]. The integration between Researcher ID and ORCID is another example of a bi-directional integrating initiative aimed at making information on the two systems interoperable and complementing. Similarly, the ODIN project[5] aims to define a roadmap for the integration and scalability of the DataCite and ORCID identifiers solutions to create a layer of interoperability between persistent identifiers for researchers, research works and their outputs (publications and data) in order to address four main challenges concerning research data management: accessibility, discovery, interoperability and scalability. The proposed solution is based on a conceptual model of interoperability [3] for linking research data and their contributors (embedding the corresponding PIs into metadata) through the coordination and alignment of the information flow across data centers, DataCite, and ORCID. The RDA PID Interest Group[6]is another example of the recent effort of coordinating the use of persistent identifiers for supporting referencing and citation of research products and their authors and contributors and manage the lifecycle of research data production.

Finally, other initiatives have been started within specific communities. In the library domain, the BIBFRAME initiative[7] has defined a lightweight framework (metamodel) for bibliographic description based on linked data principles to improve the integration, discoverability and reuse of library resources and their descriptions in a networked distributed environment. At the core of the proposed data model, there is the concept of BIBFRAME authority which is a resource representing a person, organization, place, topic, temporal expression and other entities associated with a BIBFRAME Work, Instance, or Annotation (i.e. the remaining classes of the model). BIBFRAME authorities are used not only to identify (via URIs) the above mentioned entities within the description, but also to link to external resources (for example traditional authorities) referring to the same entities by including their corresponding IDs. In this way, the mechanism of BIBFRAME authorities should provide a common lightweight interoperability layer over different Web-based authority resources connceting a BIBFRAME resource, such as a Work or Instance, and one or more authorities for related entities, such as a person, organization, or place, identified by other identifiers systems like a ID.LOC.GOV, ISNI, VIAF and others.

All these initiatives have the merit of having increased the awareness and consensus among relevant stakeholders and communities about the crucial role of a coordinated ecosystem of persistent identifiers at the heart of a global infrastructure for e-science. A lot of work has been done to define common objectives and share conceptual models and strategies to solve the persistent identifier interoperability problem. However, a solid technological solution for interoperating identifiers for digital objects, contributors, authors and other relevant entities is still lacking in the effort to develop a sustainable infrastructure providing a core layer of interoperability on which cross-cutting advances services for science and education can be implemented to encourage openness and collaboration across disciplines, communities and geographical boundaries. Based on the valuable results of the above mentioned initiatives, but also exploiting the experience on persistent global identifiers gained in the course of the OKKAM FP7 project[8], this paper addresses the same problem from a slightly different perspective, proposing a technical solution to implement a persistent identifier interoperability core service for e-science infrastructures. In the next section we start to describe the three main functionalities which should be supported by this core service.

3. INTEROPERABLE PERSISTENT IDENTIFIERS AS VALUE ENABLERS OF E-SCIENCE INFRASTRUCTURES

Interoperable persistent identifiers are key building blocks in managing the complex information space of e-science infrastructures and extracting value from it. We have identified three main core functionalities which explain this crucial role.

1. Ensuring and enhancing the persistent access, use and reuse of resources or related information across different boundaries (e.g. technological, disciplinary, institutional).

2. Providing the means for explicitly representing the network of relationships among all the relevant entities in the research landscape (authors, contributors, publications, data, research projects, grants, intitutions) and creating an integrated information space which can be walked through starting from any of the links and from which new knowledge can be formed.

3. Enabling the development of added-value services on top of integrated digital information spaces.

The maintenance of a solid relationship between the identifier and the associated entity, digital (e.g. an electronic publication) or non-digital (e.g. the author of the publication) is the fundamental mechanism to ensure persistent access and reuse of the resource itself or information related to it. This stable association is what confers persistence to the identifier. In an interoperability infrastructure this means not only guaranteeing the persistent link between a given identifier and the identified resource, but also managing possible alternative links (implemented by other identifier systems) which may provide a continued alternative access to the resource in case the first connection is not accessible (e.g. broken link or denied access permission). This means that the infrastructure should be able to connect identifiers for the same entity across different systems. Such a requirement can

[4] http://orcid.org/blog/2013/04/22/orcid-and-isni-issue-joint-statement-interoperation-april-2013
[5] http://odin-project.eu/
[6] https://rd-alliance.org/internal-groups/pid-interest-group.html
[7] http://www.loc.gov/bibframe/
[8] http://project.okkam.org/

be addressed, for example, by managing matching functionalities with allow to identify "same-as" relationships between persistent identifiers, i.e. two identifiers refer to the same entity. For example, given a DOI for an article the identifier interoperability infrastructure could provide access to the identified publication through a redirection mechanism which involves the DOI resolver, but could also provide alternative persistent identifiers for the resource, if any, (for example an URN or an ARK), giving alternative ways to access the target information object.

The implementation of this coreference mechanism has been largely discussed within WP22 of the APARSEN project and has been included as one of the fundamentals of the framework. In the APARSEN framework, coreferences between persistent identifiers (and the identity between the referents) are not inferred based on matching on metadata information describing the identified entities, but are directly extracted from the information object. Since often resources are identified by more than one PID (e.g. a document can be identified by a DOI and by a URN) and the presence of alternative identifiers can be made explicit in the metadata provided by the persistent identifier management systems (e.g. in the DOI kernel metadata the "referentIdentifiers" element is used for this purpose), the framework, and the related demonstrator, rest on the idea that the co-existence of two or more identifiers in the metadata about the entity can be exploited to automatically generate trusted identity relationships between information objects, by transitivity.[9] In brief, these are the only trusted co-references according to the APARSEN approach and they can be reliably used to integrate information across PID domains. This cautious approach has the advantage to reduce the risk of generating false positive matches, due to the fact that the matching process is based on the coreference information directly provided by trusted PID domains, but has the disadvantage to exclude from the integration process all the objects not linked through the inferred coreference chains. Since, as we have stated above, the use of PID is largely fragmented and inadequate for many entities potentially relevant for the e-science domain, it is difficult to imagine a broad applicability of the proposed approach to include the entire spectrum of entity types of interest.

In order to exploit the value of e-infrastructure data, it is necessary to have stable access not only to the single resources but also the relationships among these resources [10], like an author and his/her research output or the publications related to a given dataset. According to this perspective, a second element of value of managing persistent identifiers deals with making explicit and reusable the relations between the relevant entities within the scientific data infrastructure[9]. Again this can be realized making interoperable identifier systems for different types of resources, like those for authors and contributors with those for digital objects. The persistent identifier interoperability infrastructure should be able to provide the identification capabilities necessary to represent structured knowledge that can be integrated across systems and used to discover new elements of knowledge by querying and navigating the information space. For example, data providers should be able to represent their data and metadata by reusing identifiers already assigned to the relevant entities instead of assigning new identifiers. A dataset should be not only identified by a unique ID but should also be related to its author as part of its metadata. If the author has already been assigned an author ID registered within the infrastructure identifier registry, it is crucial that the data center can reuse the same ID for uniquely identifying the dataset since through it many relevant relationships can be inferred (for example that among the author publications there is one article based on the experimental results on the dataset).

The interoperability infrastructure for persistent identifiers is also crucial for the development of community added-value services which can be build on top of the (now fully) accessible scientific data and network of relationships around them. Due to the interoperability layer not only the information is extracted and integrated across systems but also the higher level services based on this information can interoperate and produce additional value, for example by facilitating the sharing of research findings, improving accessibility to research products and identifying authors and contributors of scientific outputs. For instance, enabling automatic discoverable connections between relevant entities participating in the scientific production value chain, like funding agencies, grants, projects, contributors, institutions and many others, research administration services for assessing the impact of research programs can be developed and provide a valuable instrument for research funders and policy makers.

From this perspective, identifiers and metadata enriched by uniquely identified information are value enablers of e-science infrastructures, by increasing the interoperability of data, facilitating the access to relevant and trustable information, increasing the trustworthiness of sources, revealing links and dependencies between data and solving ambiguity issues.

4. THE ENTITY NAME SYSTEM

The aim of this paper is to propose a technical solution to implement the layer of interoperability for persistent identifiers in e-science infrastructures. This solution is based on the Entity Name System (ENS) prototype developed in the context of the EU-funded project OKKAM[10]. The ENS[11] is a scalable infrastructure for assigning and managing unique identifiers for entities in decentralized distributed information environments like the Web and foster their global reuse. The first prototype of this system has emerged as a solution to the entity identification problem in the Semantic Web [6] and in other distributed contexts, that is the problem of integrating information about entities which are assigned different identifiers in different systems or by different users [7]. In order to deal with this problem, the ENS provides a

[9]Assuming for example that an object, say o1, is identified by a DOI and another object, say o2, is identified by the same DOI as o1 and by an ARK, the ARK of o2 can be used to derive the identity relation between o1 and a third object, o3, identified by the same ARK as o2, by transitivity of the identity relation. In this way chains of coreferences can be automatically generated (provided that the metadata information from different PID domains is structured in a common way) by simply trusting the coreferences included in the information objects.

[10]`http://project.okkam.org/`

[11]`http://api.okkam.org/`

Figure 2: ENS Infrastructure

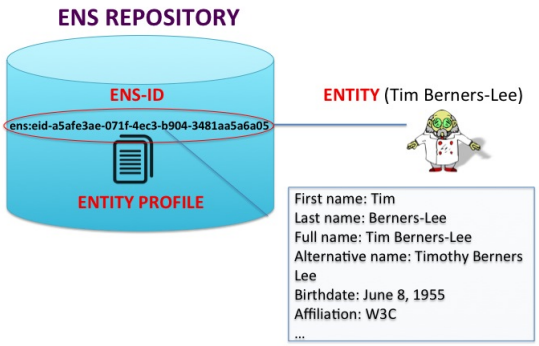


Figure 3: ENS Repository

service to assign global unique identifiers to entities named in information sources and reuse these identifiers across systems boundaries regardless of the place or domain where they have been first assigned. To this purpose the ENS has a repository for storing entity identifiers along with a short set of descriptive metadata, i.e. an entity profile, which is used with the aim to disambiguate each entity from the others. When a human user or an application searches the system for an identifier (for example by keywords), information in the entity profiles is used to establish (through advanced entity matching algorithms) if an identifier has been assigned and stored for that entity. Otherwise, a new identifier is minted and returned by the system. The systematic reuse of the identifiers created and maintained in the ENS would reduce the multiplication of identifiers for entities and enable a frictionless entity-centric integration of information spread and scattered on the Web. The ENS infrastructure is based on the following core basic functionalities, as shown in Figure 2:

- STORAGE: maintaining a large scale entity repository which can ensure the persistent association between a unique entity identifier (ENS-ID) and the corresponding entity.
- MATCHING: mapping any arbitrary description of an entity to its global ENS-ID.
- ACCESS: providing services (i.e. interfaces, APIs) to make ENS identifiers searchable and easily retrievable by humans and machines.
- RESOLUTION: given an ENS-ID in input providing a short description (i.e. entity profile) about the identified entity in output.
- LIFECYCLE MANAGEMENT: supporting few basic operations like entity creation, merging, splitting to ensure the lifecycle management of the ENS identifiers in the system.

By providing a technical infrastructure for the registration and management of global identifiers for use on digital networked environments, the ENS has many features common to existing persistent identifier systems. First of all, the main goal of the ENS is to store the persistent association between a string of characters (the ENS-ID) and an entity. Secondly, ENS identifiers are actionable identifiers but are not locators (URLs). Third, the ENS provides a resolver which allows to enter an ENS-ID and access a small set of metadata providing a short description of the corresponding entity. Fourth, the ENS stores identifiers along with a small set of metadata providing descriptive information about an identified referent. This information is returned by the resolution service. The relationships between the entity, the ENS-ID and the metadata description (entity profile) is shown if Figure 3.

In addition, the ENS has some distinguishing aspects. While many persistent identifier solutions have been developed to identify specific kinds of entities (e.g. DOI and URN for digital objects, ORCID and ISNI for authors and contributors), The ENS-IDs are digital identifiers for entities of any type (digital and non-digital entities) like people, institutions, publications, Web pages, events, locations and so on. Another difference concerns the scope of the identification system. The majority of the current persistent identifier solutions were introduced to solve the problem of changes in location or name of the resources on digital networks (i.e. the broken link issue) by maintaining a persistent binding between the identified resource and an online location where the object or a representation of it can be retrieved. The ENS has been developed as a service for enabling the fulfillment of entity-centric approaches for data integration in digital distributed environments, like the Semantic Web. The issue in this second case is distinctly related to global naming and reference rather than to persistent resolution. Finally, the ENS metadata model has not been developed to address semantic interoperability issues (like for example the DOI data model), that is enabling the automatic reuse of information originated in one context in another context, but has been created to enable disambiguation and entity matching within the ENS identifier repository. The ENS metadata model consists of a minimum set of metadata which should be sufficient to uniquely identify the entity and distinguish it from the other stored entities. The metadata are used for making the identifiers searchable and retrievable (search

queries are matched on metadata values) and to provide a short description of the identified referent to a user.

From the above comparison it emerges that the ENS has the potential to fill some of the interoperability gaps of the PID landscape even though an evolution of the system is required. As we have stated in the introduction of this paper, one of the main challenges of the modern research infrastructures is not only to allow persistent access and reuse of digital information, but to create a global interoperability environment where data and information can be seamlessly exchanged across disciplines, institutions and services and integrated knowledge can be extracted through an articulated network of connections linking all the relevant entities in the landscape, like for example data to authors, contributors and journal articles, authors to publications, co-authors and institutions, projects to institutions, authors and funding agencies and so on. The value of these connections can be used to provide added-value services like citability, tracking of research output, quality metrics, provenance and many others. One of the major gaps to exploit the value of this connectivity is the lack of interoperability between current PID systems which hinders the possibility of creating and navigating this valuable network and leads to the creation of information islands in a very similar way to what has been described for the Semantic Web. This is not surprising since tailored local PID solutions have been developed with the aim of addressing needs of specific communities without having interoperability purposes in mind. The ENS has been instead designed as an interoperability solution from the beginning. In the next section we will discuss how the ENS can realize the technological infrastructure for addressing the instance-level information integration problem at the core of e-science infrastructures. Some recent crucial modifications and additional functionalities are also presented as part of the evolution of the system toward a novel infrastructure capable of satisfying the three main requirements discussed in Section 3

5. THE EVOLUTION OF THE ENS TOWARD AN INTEROPERABLE INFRASTRUCTURE FOR PERSISTENT IDENTIFIERS

Up to this point, the ENS has been presented as an infrastructure supporting the identification of several types of entities and implementing a sophisticated matching mechanism to allow the reuse of identifiers across independently produced content. However, three additional features need to be addressed by the ENS in order to become a productive interoperability infrastructure for persistent identifiers in e-science.

First of all, the system should not operate as a centralized solution for global persistent identifiers but as an integrating infrastructure federating current persistent identifier solutions to ensure interoperability. It has become clear in the last few years [5] that a unique global identifier solution is not the right answer to the interoperability problem of identifiers. This is because many solutions have been consolidated in some domains (e.g. publishers or institutional repositories) and local tailored systems are difficult to be overcome since they provide services tuned to the specific

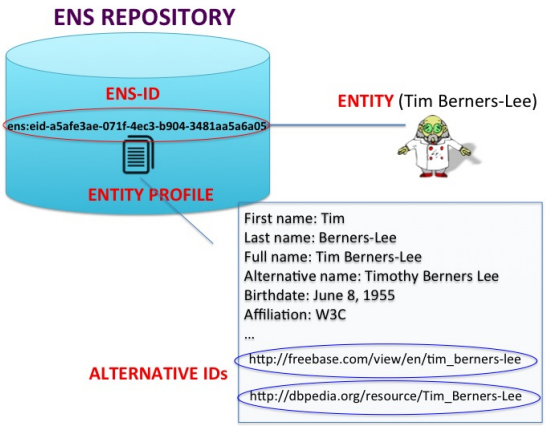
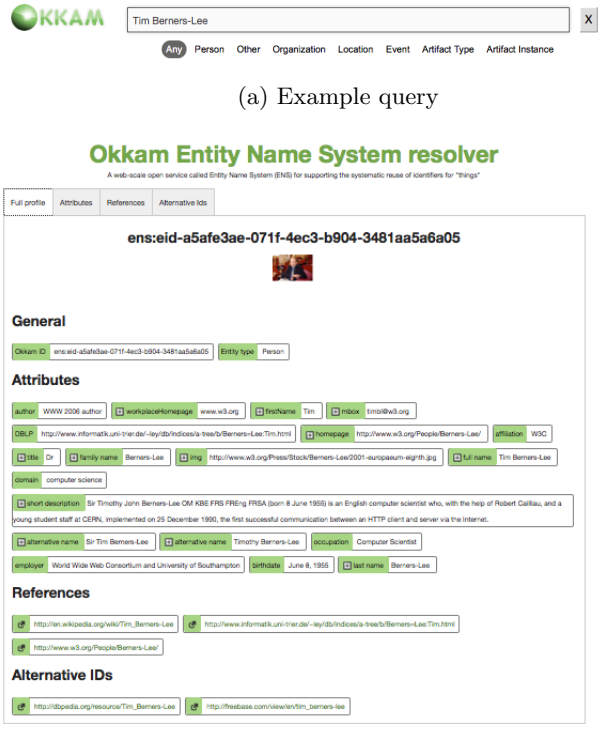


Figure 4: ENS Alternative ID Management Service

needs of specific stakeholders. To work as an integrating PID infrastructure the ENS needs to facilitate interoperability between systems already in use and support the development of added value services which can address both specific community needs and cross-boundary requirements. Technically this can be realized through an effective management of mappings between the ENS identifier assigned to a given entity and any other (persistent) identifier for the same entity (alternative ID management service). In this way, an ENS-ID can be viewed as unifying integration service providing a single entry point to multiple alternative identifiers for the same entity. The ENS infrastructure has the basic core service for registering and managing alternative identifiers. All the alternative IDs available for the entity are stored in the ENS registry as part of the entity profile (see Figure 4). The functioning of the alternative ID management service can be understood by performing a simple query for an entity through the search interface of the ENS[12]. For example by entering the keyword <Tim Berners-Lee>, the ENS (through its default resolver) returns a short description of the scientist through its core set of metadata of the entity profile and a list of the alternative identifiers for the searched entity. Figure 5 shows the screenshots for the example query. In the example the alternative identifiers for the target entity (i.e. Tim Berners-Lee) are URLs belonging respectively to dbpedia and freebase namespaces. The "alternative-id" relationship between them and the binding to the ENS-ID of the entity has been established through the matching functionality when structured information about the target entity has been imported from these knowledge bases into the ENS. The matching algorithms implemented in the ENS use the descriptive metadata in input to establish if an ENS-ID has already be assigned to the entity. If the entity has already registered, the import function updates the profile and imports the IDs used in the original sources as alternative IDs. Otherwise a new profile is created and the imported information is used to fill the core metadata of the profile (through vocabulary mapping) including the alternative ID field. The alternative ID management service could be used to map any kinds of alternative identifier including alternative persistent identifiers, like for example,

[12]The search interface is available at `http://api.okkam.org/search/`



(a) Example query

(b) Seach output

Figure 5: ENS search interface screenshots

referring to our previous example, the Scopus ID and the ORCID ID for Tim Berners-Lee (if available). This mapping would enable a first level of interoperability between the two identification systems allowing to identify (and access) two islands of information in the corresponding systems which refer to the same entity and create a bridge between them. Going back to our example, entering a Scopus ID one can find the alternative ORCID ID and by resolving this ID, access to information about the target entity. In the case of digital objects, the alternative identifiers can be used to get alternative access to the resource on different servers as well as related information. For this purpose, a redirect service, based on the alternative IDs associated to the ENS-ID, has been recently developed which allows users to resolve the ENS-ID into third-party data sources[13]. For a given ENS-ID, the service allows to get a list of resolvers and redirect to a selected resolver. It should be noted that the ENS approach for managing alternative identifiers differs from that proposed by the APARSEN Interoperability Framework. In the APARSEN framework the co-reference between alternative identifiers is provided directly by content providers and this mechanism allows to create a linkage between previously disconnected resources (see footnote 9). On the contrary, the ENS alternative ID management service connects the alternative identifiers to the profile of the identified entity and therefore links them to the unique ENS-ID for that entity. In this way, the ENS-ID works as the glue for bridging all the alternative IDs referring to the entity. Any of these IDs (in use in different systems) can be used to interrogate the ENS and retrieve the corresponding unique ENS-ID which in turn gives access to all the alternative IDs of the profile. Through the alternative IDs, alternative ways of access to the resource or information about the resource are enabled, empowering the cross-boundary integration and mash-up of data. Moreover, a profile can be updated with additional alternative identifiers across time as the entities named in different sources are matched and aligned with the ENS identifiers via a process of automatic entity matching.

A second aspect deals with persistence. In [4] we have discussed the evolution of the ENS to a persistent ENS through the separation of the ID (e.g. `peid?8af7c50f?f072?4384?905b?03875c341863`) from the resolver (`http://www.okkam.org`). This introduces a level of indirection between the identifier and its referent and ensures the persistent binding between them. By default, the ENS-ID is combined with the ENS default resolver and its resolution returns a small set of metadata (included in the ENS entity profile) related to the identified entity. The real potential of separating the token id from the resolver rests on the possibility of associating the same ID to multiple resolvers, enabling a mechanism of multiple resolution. Different actors can create or reuse persistent ENS-ID (PEID) for entities of interest using the ENS and through their local resolvers enable precise (and long-term) access to information they store (see Figure 6 extracted from [4]). While ID management is addressed by the ENS, information management, including persistence of the content, and reliable resolution (excluding the default resolution service provided by the ENS) is managed by content providers, in line with the main assumptions of the APARSEN interoperability framework for PIDs but also addressing the requirements of the linked data community. The ENS PEIDs can be reused as part of Cool URIs allowing Linked Data users to create URIs resolvable to any information source they like. At the same time, persistent identifiers users can reuse the same PEIDs to identify information objects and resources managed by trusted institutions which ensure their persistent access and association to a physical location. Due to this change of paradigm, the ENS differs from a centralized authoritative service for minting and resolving global identifiers, allowing to every one the reuse of the ENS-IDs to create persistent identifiers (through domain resolvers) or Cool persistent URIs (through the web service resolution mechanism). The last point is important since several initiatives[14] have highlighted the need to develop a co-ordinated solution to identifier issues across the PID and the Linked Data community (as stated for example in the Den Haag Manifesto[15]). The recent improvement of the ENS may offer such a solution, enabling data creators and curators to combine the technical strengths and opportunities of the (Semantic) Web vision with the organizational, economical and social requirements legitimately raised by the PID community and stakeholders. This has a strong impact on the development of services to support the integration of information across sources since it opens the door to new forms of interactions between open structured data published on the Web and content stored by more

[13]More information is available at `http://community.okkam.org/`

[14]For example, the Persistent Object Identifiers seminar at The Hague in June 2011 and the Links That Last workshop in Cambridge in July 2012

[15]available at `http://www.knowledge-exchange.info/Default.aspx?ID=462`

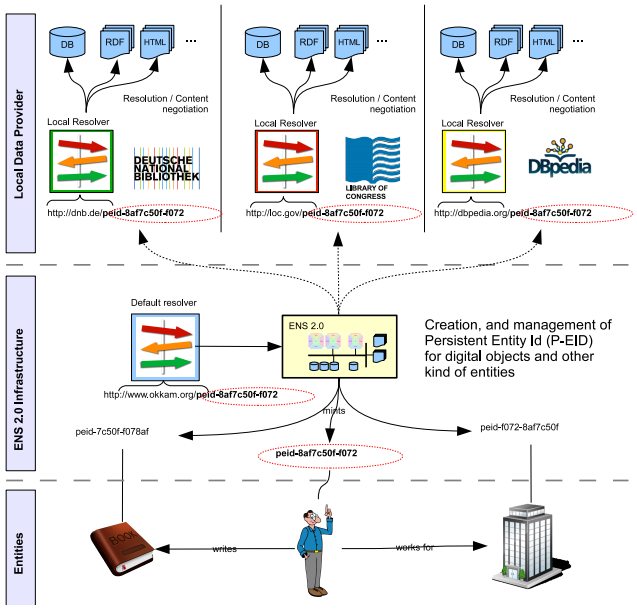


Figure 6: Multiple Resolution in the ENS

traditional cultural heritage institutions.

The third aspect deals with vocabulary mapping. Different persistent IDs may be associated with different vocabularies used to represent the identified resources. If a mapping among them is available, information structured according to a given schema and retrievable thanks to a given ID can be directly re-used to integrate or update the information of another source adopting a different schema to represent the same entity. Therefore, in order to support semantic interoperability across services and communities, the ENS should provide an extensive mapping of vocabularies and schemes adopted in different PID domains. A service, called OKKAM Synapsis[16], is currently under development to automatically compute the mappings between terms in controlled vocabularies and ontologies toward the ENS core set of metadata. Synapsis is designed as a Web application to support a community-driven effort in the collection and maintenance of mappings. Through the application, a user (human user or API user) can search mappings for a given property by using different filters (e.g. author, status, date), find clusters of mappings for all the registered properties, propose new mappings (which then can be accepted by the administrator of the service) and edit or rate existing mappings (i.e. add comments and manually evaluate mappings by classifying each mapping into one of different categories). While in the APARSEN Interoperability Framework semantic interoperability is addressed by proposing a common ontology which should be used by content providers to expose their data in a common way, the ENS approach focuses on the alignment of different vocabularies through ontology mapping. This has the advantage that users can maintain their own vocabularies and ontologies, without the need to restructure their content according to a new model. The mapping of vocabularies allows supporting the building of crosswalks between them and can be extended to include an indefinite number of vocabularies.

[16]http://api.okkam.org/synapsis/

6. BUILDING ADDED VALUE SERVICES ON TOP OF THE ENS INFRASTRUCTURE

A number of added value services can be built on top of the interoperability layer provided by the ENS infrastructure and usable by other systems or infrastructures. We describe some examples.

1. **GLOBAL RESOLUTION SERVICE:** Based on the ENS redirect service described above, a global resolution service can be implemented, which determines the appropriate resolver for a given PID. Moreover, if alternative IDs are associated with the searched PID, the service returns alternative resolvers to access the identified resource via alternative routes.

2. **METADATA ENTITY IDENTIFICATION SERVICE**: This service allows assigning unique identifiers to entities named within the metadata of other resources. For example, if the metadata of a journal publication include author information, the system allows assigning a unique ID to the author which can be an istantaneously generated ENS ID if the entity has not been registered in the repository before, or can be selected among the IDs available in the entity profile if the entity matches one already stored in the system.

3. **METADATA EXCHANGE SERVICE**: By linking a PID to alternative IDs, the ENS interoperability layer can be exploited to develop services for automatic exchange of metadata across systems using different identification solutions. For example, given a PID for an author (e.g. an ORCID ID), the service provides the link to external sources of information (e.g. Scopus, ResearcherID, arXiv) where information about the same author can be found and automatically imported into the original author profile. This can be done thanks to the mapping between the corresponding vocabularies provided by the ENS interoperability layer (via the Synapsis service).

4. **IDENTITY LINKAGE SERVICE**: When a PID for an entity (e.g. an author) is entered, the service returns all the entities related to that entity belonging to a certain entity type (like for example all the author's publications) and allows to navigate the entire chain of links connecting the identified entity to all the related entities (e.g. starting from the PID of a dataset it is possible to go back to the contributors, the related publications, the research projects and so on). Semantic Web technologies provide a possible solution to implement this service. Metadata from different sources can be represented as RDF assertions about resources identified by unique IDs. The ENS interoperability layer offers two unifying elements to integrate data from different sources of metadata: the unique global ENS IDs and their "same-as" relationships with alternative IDs and the vocabulary mappings.

7. CONCLUSIONS

Interoperability between persistent identifiers is a critical concept for enabling the development of fully-integrated services for research e-infrastructures in order to improve circulation, transfer and access to integrated scientific information and promote cross-boundary collaboration and competition. In this paper we propose a scalable infrastructure to allow current persistent identifier solutions to interoperate and provide integrated access to multiple heterogeneous sources. The proposed infrastructure is based on the OKKAM Entity Name System and implements three main technical core functionalities 1) the management of coreferences among PIDs (alternative id management service) ; 2) the assignment and management of global Persistent Cool identifiers; 3) the mapping of vocabularies across PID domains. Beyond the technical requirements, the implementation of the system will add value to the PID systems only if a governance layer is agreed among them. Therefore, effort is currently dedicated to create the social and organizational support among the relevant stakeholders to transform the ENS into a public open infrastructure for PID interoperability maintained (but not owned) by a Trustee monitored by a board of protectors according to a Trust agreement. As a first step to increase the trust and community support around the ENS infrastructure, we are currently working to propose the ENS interoperability services as part of the offerings of the APARSEN Virtual Centre of Excellence [8] that brings together a diverse set of stakeholders, researchers and practitioners in digital data and digital preservation.

8. REFERENCES

[1] APARSEN D22.1: Persistent identifiers interoperability framework. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D22__1-01-1__9.pdf, 2012.

[2] APARSEN D22.3: Demonstrator set up and definition of added value services. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/02/APARSEN-REP-D22_3-01-1_0.pdf, 2013.

[3] APARSEN D4.1: Conceptual model of interoperability. http://files.figshare.com/1239137/D4.1_Conceptual_Model_of_Interoperability.pdf, 2013.

[4] B. Bazzanella, S. Bortoli, and P. Bouquet. Can persistent identifiers be cool? *IJDC*, 8(1):14–28, 2013.

[5] P. Bouquet, B. Bazzanella, M. Dow, and R. Riestra. DIGOIDUNA FINAL REPORT: Digital Object Identifiers and Unique Author Identifiers to enable services for data quality assessment, provenance and access. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/digoiduna.pdf, 2011.

[6] P. Bouquet, H. Stoermer, and B. Bazzanella. An entity name system (ens) for the semantic web. In *ESWC*, pages 258–272, 2008.

[7] P. Bouquet, H. Stoermer, C. Niederée, and A. Mana. Entity name system: The back-bone of an open and scalable web of data. In *ICSC*, pages 554–561, 2008.

[8] D. Giaretta and all APARSEN partners. APARSEN D11.4: Virtual centre of excellence development. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/02/APARSEN-REP-D11_4-01-1_0.pdf, 2014.

[9] A. Hayes, S. Mann, A. Aryani, S. Sabine, L. Blackall, P. Waugh, and S. Ridgway. Identity awareness and re-use of research data in veillance and social computing. In *Proceedings of The IEEE International Symposium on Technology and Society (ISTAS)*, 2013.

[10] T. Weigel, M. Lautenschlager, F. Toussaint, and S. Kindermann. A framework for extended persistent identification of scientific assets. *Data Science Journal*, 12, March 2013.

[11] J. Wood, T. Andersson, A. Bachem, C. Best, F. Genova, D. R. Lopez, W. Los, M. Marinucci, L. Romary, H. V. de Sompel, J. Vigen, P. Wittenburg, D. Giaretta, and R. L. Hudson. Riding the wave - how europe can gain from the rising tide of scientific data. final report of the high level expert group on scientific data. a submission to the european commission. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf, October 2010.