

The International Journal of Digital Curation

Volume 8, Issue 1 | 2013

Can Persistent Identifiers Be Cool?

Barbara Bazzanella, Stefano Bortoli and Paolo Bouquet,
DISI, University of Trento

Abstract

The fast growth of scientific and non-scientific digital data, as well as the proliferation of new types of digital content, has led – among many other things – to a lot of innovative work on the concept of the identifier. Digital identifiers have become the key to preserving and accessing content, just as physical identifier tags have been the key to accessing paper-based content and other physical entities for millennia. Two main schools of thought have emerged: on the one hand, librarians and public repositories have pushed the concept of the *Persistent Identifier* (PI) as a way to guarantee long term identification and (sometimes) access; on the other hand, the extraordinary success of the web has led several researchers and web experts to push the concept of the *Cool URI* as the universal mechanism for identifying and accessing digital content. Both views have their pros and cons, but so far (with only a few exceptions) the two visions have developed in parallel, sometimes with a subtle underlying hostility.

In this paper, we present the evolution of the Entity Name System (ENS), an open service-based platform developed as part of the OKKAM EU co-funded project, which can reconcile these two approaches. The new system, called ENS2.0, is currently under development and will enable 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 PI community and stakeholders.





Introduction

In recent years, the rising growth of scientific and non-scientific digital data, as well as the proliferation of new types of digital content such as user generated content (e.g. product reviews, posts, pictures, blogs or tweets), has raised a whole range of new opportunities and challenges in the realm of digital curation and preservation (European Commission, [2010](#)). The possibility of finding and accessing a massive amount of content in a digital format and linking this content to its authors and other related entities, the availability of new ways of interpreting the meaning of data using scalable data analytics tools, and the development of new and much more powerful metrics for assessing the quality of data 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 concerning, for example, long-term accessibility and usability of digital content, quality assessment and provenance.

A system for managing digital identifiers for a variety of digital objects (e.g. articles, datasets, streams of data, posts, tweets) and non-digital objects (namely other real-world entities, such as authors, institutions, teams, geographic locations, and runs of experiments) becomes a key component to capitalize on the opportunities and address the challenges of global digital assets.

Currently, the eco-system of initiatives and systems for digital identifiers is quite fragmented and different solutions (most often local, ad hoc solutions) have been adopted by different stakeholders and communities. However, two main approaches can be distinguished within this complex landscape. The most consolidated approach, called the **Persistent Identifiers (PI) approach**, is currently adopted and endorsed by the vast majority of significant stakeholders in the production value chain of scholarly content, including national libraries, archives, publishers, research institutions and data centres. The strongest motivation behind this paradigm is the requirement of enabling trusted naming authorities, thus ensuring the long term preservation of digital content, providing guarantees on quality and integrity of data and content, and implementing access control policies which are compliant with the protection of intellectual property. The second approach, called the **Linked Data approach**¹, has more recently emerged from the Semantic Web and “open data” communities to enable new ways of publication and interaction with distributed structured data on the Web. The strongest argument here is that the Web architecture allows for the creation and management of resource identifiers (called URIs), which can be used to identify digital and non-digital resources without the need for third-party naming authorities; and that URIs provide a simple way for enabling a web of interconnected data (in analogy to the web of documents) in a fully decentralized and open way.

So far, these two visions have proceeded in parallel, or sometimes even with a subtle reciprocal hostility. On the one hand, the Linked Data approach definitely challenges the social and technical model behind the PI vision, as it works with no need for naming authorities and is designed on top of the standard web architecture. In particular, there is a concern within the PI community about the possibility that the web itself may become the *de facto* platform for digital preservation and the current

¹ See: <http://w3c.org/DesignIssues/LinkedData.html>

practices around HTTP URIs the *de facto* way of managing persistent identifiers, without taking into account the organizational and policy issues involved in long term infrastructure for data curation and preservation. On the other hand, the PI requirements and technical platforms challenge the purely decentralized and “fractal” nature of the web, as institutional curators and content managers need to rely on a fully trustable social organizations which can guarantee unique identification, authenticity, provenance and persistence in the long run.

However, some recent initiatives² show an increasing awareness in part of the PI community that the Linked Data practices and tools may offer an opportunity for increasing the value of data (in particular, by providing shared metadata vocabularies and through cross-linking) and even cover use cases which traditional solutions were not designed to address (for example, identifying dynamic resources).³ The problem is that there is not a common social and technical platform that the two communities can use to build a common solution which can bridge the gap between the two worlds. The aim of this paper is to show how an evolution of the Entity Name System (ENS), a distributed infrastructure for managing identifiers for different kinds of entities developed as part of the OKKAM EU co-funded project, can become such a common platform and can be used to reconcile the two alternative views. This platform, which we call ENS 2.0, is on the one hand fully compliant with the Linked Data technical requirements (use of HTTP URIs, useful lookup services, cross linking – plus Cool URIs support), while on the other hand takes very seriously the need for a social entity which can guarantee a basic coordination level among relevant stakeholders, and the uniqueness and persistence of the binding between digital identifiers and entities.

The ENS 2.0 is not intended to provide an alternative identification approach, but a solution to subsume schemes currently in use, making the other identification systems able to communicate and connect their contents across the boundaries of their systems. This can be obtained by:

1. Making explicit the mapping across PIs referring to the same entity and the relationships between related entities;
2. Making the PIs interoperable with other identification schemes on the Web, such as Linked Data;
3. Making PIs discoverable and reusable across the boundaries of local systems;
4. Supporting the full decoupling of a PI for an entity from its resolver, and maintaining the persistent link between each registered PI and its local resolver.

To this purpose, the ENS 2.0 is conceived as a logically centralized, but physically distributed infrastructure for managing persistent and global identifiers for digital and non-digital resources and also linking these identifiers to alternative IDs (e.g. DOI, URN, and URI) for the same named entities. The ENS 2.0 provides a single entry

² For example, the Persistent Object Identifiers seminar at The Hague in June 2011 and the Links That Last workshop in Cambridge in July 2012.

³ An interesting example of this collaboration is offered by the British Library, which is developing a version of the British National Bibliography (<http://www.bl.uk/bibliographic/natbib.html>), which it is making available as Linked Open Data via a Talis platform.

point to a public registry of local resolvers, which can resolve the same global ID to different local data or content referring to the same entity. This is realized through an identification schema, which decouples the unique ID from the location of the resolver(s) and allows the assigning of the ID to multiple resolution locations. This way, the resolution is distributed across the boundaries of different systems, but a trusted naming authority guarantees the uniqueness and the persistence of the IDs and their links to the related local resolver authorities. These authorities then should ensure that the identified digital materials are permanently managed and located over time. Furthermore, the same ID resolved by local PI authorities can be used to access HTML documents and RDF resources on the (Semantic) Web.


State of the Art: Comparing PIs and Cool URIs

In this section we compare the two main current identifier approaches aiming to provide an integrating solution to the identification challenges of global distributed digital assets.

It is widely recognized that the use of URLs (which have been adopted from the birth of the Web to identify network resources) cannot be considered a reliable approach to address these challenges, due to the fact that URLs have the double 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 (i.e., broken link problem). For this reason, two main approaches have been proposed to preserve access to a digital resource, regardless of its physical location: the **Persistent Identifier (PI)** approach and the **Cool URI** approach. In Table 1 we summarize the main fundamental differences between the two approaches. The aim of this comparison is to show how the alternative ENS 2.0 approach presented in this paper integrates elements of each infrastructure to reconcile them into an integrated, cross-boundary approach to digital identifiers.

Feature	Persistent Identifiers	Cool URIs
<i>Resolver</i>	<p>YES</p> <p>A resolver creates the link between a PI and the current location of the associated object.</p>	<p>NO</p> <p>The connection between the Cool URI and the Web document is regulated by the HTTP protocol.</p>
<i>Authority</i>	<p>YES</p> <p>There is an institutional commitment and defined roles and responsibilities to manage the PIs and guarantee their persistence, locatability or actionability in the long or short-term.</p>	<p>NO</p> <p>There is no authority that guarantees the Cool URI management or the lifecycle of the identified resources.</p>

Feature	Persistent Identifiers	Cool URIs
<i>Naming authorities</i>	YES The assignment of PIs is under the control of trusted naming authorities.	NO Anyone can assign a Cool URI to anything. There is no control over the Cool URI assignment.
<i>Level of trust</i>	HIGH Resource reliability is guaranteed by trusted institutions.	LOW Resource reliability is not guaranteed through the Web architecture.
<i>Policies</i>	YES Rules that govern the operation of the system are agreed by the management authorities.	NO Only general principles and best practices. For example, the linked data assumptions are based on HTTP URIs being immutable but this is not the case.
<i>Persistence</i>	YES PIs resolve to contents in a manner that persist over changes in location, ownership, description methods and other changeable features. Explicit policies for persistence are agreed within the reference community (e.g. DOI).	NO (or at least, to be proven) There is no guarantee that URIs persist as functional link over time. They will change. Some URI are poorly constructed, including components that people may want to change over time (e.g. brand names).
<i>Actionability of IDs</i>	PARTIALLY “Good” PIs should always be exposed in a form that is actionable, but some PIs have not been implemented (at least originally) as actionable IDs (e.g. URN). PIs have been made actionable by URLifying them.	YES Cool URIs are HTTP URIs that are actionable names.
<i>Uniqueness</i>	YES	NO The same resource may be available from many URIs.
<i>Content change</i>	NO The identified resource does not change over time.	YES Over time, different resources or variant versions of the same resource may be available in the same URI.



Feature	Persistent Identifiers	Cool URIs
<i>Content negotiation</i>	NO	YES The HTTP protocol allows the association of different versions or formats of a document (or more generally a digital resource) to the same URI so that users can specify which version they need.
<i>Cross linkage</i>	NO	YES Linking data to other data. Different kinds of data about the same asset can be produced in a decentralized way by different actors, then aggregated into a single graph.
<i>Effort for implementation</i>	HIGH	LOW
<i>Costs for users</i>	Potentially HIGH (e.g. DOIs)	LOW
<i>Sustainability issues</i>	MANY	FEW
<i>Identified entities</i>	Mainly digital objects. Emerging solutions to identify authors.	Everything.
<i>Bridge metadata</i>	NO	YES

Table 1. Comparing Persistent Identifiers and Cool URIs.

The first approach is based on the use of Persistent Identifiers (PIs). Unlike URLs, PIs are identifiers referring persistently to the resources to which they have been assigned, regardless of their physical location or current ownership. This kind of mapping is realized by introducing a layer of indirection between the PI and the target object, which combines two main aspects:

1. Decoupling the identifier of the resource from its location,
2. Providing a reliable mechanism for maintaining the mapping between the identifier and the resource.

This allows the location of an entity to be changed while maintaining the PI of the entity as an actionable identifier by actively managing the link between the PI and the URL to which it is resolved. Technically, the redirection mechanism is undertaken via

←—————→

a resolution service, named **resolver**, that provides the mapping from the persistent identifier to the location of the corresponding digital object. As the location of the object changes, the resolver database is updated to map the identifier onto the new location, guaranteeing the persistent location and access to the appropriate or current copy of the object (for an example in the domain of NBN, see Bellini et al. [2008](#)).

The PI approach presupposes some guarantee that resolvers are themselves persistent. Indeed, without such a guarantee, the disappearance of the resolver would break the binding between the PI and the digital object, and this of course would make the entire idea of persistence very weak. Therefore, the PI paradigm presupposes the existence of a social model of **registration** and **naming authorities**, which have the responsibility of overseeing the successful operation of the system by ensuring the persistence of resolvers and the respect for agreed naming policies. In this sense, keeping an identifier persistent is a matter of roles, policies and responsibilities, which contribute to create a trustworthy infrastructure based on the coordination between different parties, such as the authority that assigns the identifier, the resolution service and the content provider that manages the content. This makes it clear that a discussion on persistent identifiers 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 agreements on critical aspects, which underlie the development and maintenance of an identifier system. One of these aspects deals with the allocation of the **costs involved**, which has led to different business models with important implications and financial commitments for the final users of the PI communities.⁴ Note that cost represents another element which differentiates the PI approach from the Cool URIs approach, the latter being presented as an affordable (cost-free) solution by its sustainers.

The need for the commitment of many stakeholders on each requirement may also explain the fragmentation of the current landscape of PI systems, and the difficulty of making them interoperable across technological, social, economical, national and disciplinary boundaries. The social infrastructure underlying the PI approach represents one of the fundamental elements of differentiation with regard to the Cool URI approach, and it is often embraced as one of the guarantees of reliability and trust of the PI paradigm. For example, the assignment of a PI is under the strict control of a trusted naming authority, which guarantees that the ID, once assigned, will never be assigned to another entity and the identified entity will never change over time. A different version of the same entity will be identified with another ID and the user can be directed to the exact version they are interested in. In contrast, anyone can identify a digital resource by assigning a Cool URI to it and there is no guarantee that the associated content will never change through time. This means that multiple copies of the same resource (even with very slight differences) will be available from different locations on the Web.

The main stakeholders of the e-Science community are currently converging toward a restricted number of systems and initiatives for managing persistent

⁴ The case of the DOI is emblematic in this sense.

identifiers. Consider for example Handle⁵, DOI⁶, URN⁷-NBN, Purl⁸ and ARK⁹. Around these many value-added services are being built, such as DataCite¹⁰ and CrossRef¹¹. In this context, there is a different level of maturity between the more advanced systems for digital objects and the gradually emerging solutions for authors and other kinds of entities, such as institutions. The larger diffusion of PI solutions for digital objects compared to other kinds of non-digital objects is another feature which differentiates the PI approach from the Cool URI approach, which is intended to provide an identification solution for any kind of entity (including entities that are not on the Web) one may wish to name.

The second approach to managing identifiers for digital resources and other kinds of non-digital entities (e.g. authors) is the **Linked Data initiative**.¹² Linked Data was not started as an approach to persistent identifiers, but rather as a way of enabling the creation a **web of data**: a global space of interlinked datasets, which has the potential to reproduce the network effect that the web of documents had on hypertexts. This is important, as it aids our understanding some of the assumptions underlying the Linked Data principles, practices and tools, but also why the PI community is concerned about the possibility that, through Linked Data, the web itself may be taken as the platform for e-Science and the current practices around HTTP URIs as a way of managing persistent identifiers.

One of the basic principles underlying the Linked Data vision is that URIs, and more specifically HTTP URIs, should be used to name and describe **any resource**. The key advantage of HTTP URIs is that they can be looked up directly through a pervasive protocol: HTTP. The resolution in this case is based on a domain name resolution service and the resource is accessed via a Web service mechanism. In this respect, while PIs are made actionable by URL-ifying them, HTTP URIs are **purely actionable identifiers**. One important concept for understanding the Linked Data vision is that of **Cool URIs**.¹³ A Cool URI is a URI that represents “things which are not web pages, such as people, products, places, ideas and concepts such as ontology classes.” Two things are important about Cool URIs. First, there must be different URIs for the “thing” itself (for example, a person) and for the web document describing the thing (for example, the person’s web page). The intended relationship between a resource and its representing documents can be implemented using two different technical solutions, known as 303 URIs and hash URIs. These enable **content negotiation**, allowing us to vary the content provided according to the subject making the request (i.e. a Web browser or an RDF application).

⁵ Handle: <http://www.handle.net/>

⁶ The DOI System: <http://www.doi.org/>

⁷ Uniform Resource Names.

⁸ PURL: <http://purl.oclc.org/docs/index.html>

⁹ ARK: <https://confluence.ucop.edu/display/Curation/ARK>

¹⁰ DataCite: <http://www.datacite.org/>

¹¹ CrossRef: <http://www.crossref.org/>

¹² See: <http://linkeddata.org/>

¹³ See: <http://www.w3.org/TR/cooluris/>

The second important aspect about Cool URIs is that Cool URIs should be as stable as possible. This point was made very strongly by Tim Berners-Lee in his note entitled “Cool URIs don’t Change.” The document starts with the statement that:

“There are no reasons at all in theory for people to change URIs (or stop maintaining documents), but millions of reasons in practice...”

It concludes:

“The message here is, however, that many, many things can change and your URIs can and should stay the same. They only can if you think about how you design them.” (Berners-Lee, [1998](#)).

The last point shows some level of awareness in the Linked Data community about the persistence of identifiers. And indeed this approach is sometimes perceived as an easier and faster solution to providing an infrastructure for persistent identifiers. However, from the standpoint of the PI community, this approach has some weaknesses. First of all, resources identified through HTTP URIs may be dynamic, which means that if one resolves the same URI at different times the result can be different, causing problems in the case of citations. Secondly, resources – in particular non-informational ones like people – are typically identified by different HTTP URIs in each dataset where they are named, as this depends on the fact that HTTP URIs encode the domain name as part of the string composing the identifier. The lack of reliable and trustworthy services for mapping these different identifiers onto each other makes the use of HTTP URIs very difficult to address the PI community requirements. Thirdly, despite the fact that Cool URIs should not change, there is no guarantee about the persistence of a HTTP URI. The entire Linked Data approach is not based on naming authorities, and in some cases there is an explicit opposition to their introduction. This means that no one has the formal responsibility of avoiding changes in URIs beyond the commitment of each web master to put in place solutions that may prevent this from happening. Again, it is easy to see why the PI community and its key stakeholders can’t see this as a realistic solution to their needs.

However, it is important to mention that a more positive attitude towards the Linked Data vision is slowly emerging in the PI community, which tries to go beyond the vision of Cool URIs as a threat and begins to identify potential lines of collaboration and reciprocal improvement. In particular, themes which appear to be suitable for such a collaboration include methods for making sure PIs can be referred to as HTTP URIs (including content negotiation)¹⁴, the use of Linked Data vocabularies for naming elements in metadata schemas, and the use of owl:SameAs (identity) relations to help identifiers interoperability across PI systems/schemas. In return, there is an expressed will to work with the Linked Data community on defining simple policies/procedures to improve the persistence of HTTP URIs.¹⁵ Our work

¹⁴ For example, CrossRef has published metadata for 46 million DOIs as Linked Data, (i.e. Linked Data friendly DOIs). See: http://www.crossref.org/crweblog/2011/04/crossref_and_international_doi.html.

¹⁵ A concrete example of this bidirectional effort is represented by the Den Haag Manifesto (<http://www.knowledge-exchange.info/Default.aspx?ID=462>) which is intended as a first step toward a co-ordinated approach to identifier issues across the persistent identifier and linked data communities.

aims to contribute towards this convergence process by showing that it is possible create and maintain persistent identifiers suitable for adoption by both communities.

An Entity Name System for Persistent Cool URIs

In this paper we describe the evolution of the Entity Name System prototype (ENS 1.0) developed in the context of the EU-funded project OKKAM into a revised infrastructure called ENS 2.0, which is presented as a solution to reconcile the Persistent Identifier and the Cool URI approaches described in the previous section. As we will see, a simple but crucial modification of the ENS approach in dealing with the creation and maintenance of identifiers will support the evolution of the original ENS entity identifiers into Persistent Entity identifiers (PEIDs) and will enable the definition of a novel interoperability layer by creating a bridge between Persistent Identifier authorities and the Linked Data community.

The Identification Approach of the ENS 1.0

The expected role of the Entity Name System here described is to provide a solution for enabling the systematic reuse of identifiers for different kinds of entities and improving the linkage of data about these entities across system boundaries (e.g. on the Web). To achieve this goal it is necessary to design a highly scalable architecture supporting the creation, storage and retrieval of identifiers based on descriptions (i.e. sets of key-value pairs attributes).

The first version of the Entity Name System¹⁶ (ENS 1.0) was conceived to mint and maintain identifiers that could be adopted as reusable and global names for digital and non-digital resources (Bouquet et al., [2008a](#); [2008b](#)). To support this goal, the ENS offers the following services:

- **Entity ID creation**, allowing a registered user to create a new identifier for an entity if it does not exist in the system;
- **Entity ID resolution**, allowing to access the description associated with the identifier, supporting content negotiation;
- **Entity ID search**, allowing the retrieval of a ranked list of candidate matches given a description, according to an extensible set of sophisticated matching methods (see for example Ioannou et al., [2010](#); Stoermer et al., [2010](#));
- **Entity profile editing**, allowing a registered user to edit the descriptions associated with the identifier;
- **Entity ID lifecycle**, allowing the merger of possible duplicates due to inherent imperfection of the matching process (see more in Chaudhry et al., [2008](#));
- **Get alternative IDs**, allowing the retrieval of an extensible list of identifiers defined outside the ENS and referring to the same entity.

¹⁶ The service is currently available at: <http://api.okkam.org>

←—————→

The systematic reuse of the identifiers created and maintained in the ENS would enable a frictionless entity-centric integration of information spread and scattered on the Web. It is important to notice that by proposing a logically centralized service dealing with scalability, policies, sustainability and trust issues, the ENS 1.0 faced (and provided a solution to) many of the issues common to PI approach. However, the ENS 1.0 neglected some of the requirements of persistent identifiers, such as the separation between resolver and identifier, and the organization of the social infrastructure that must guarantee persistence and sustainability through time.

On the other side, in full Web 2.0 spirit, the ENS was conceived to be suitable for a community effort,¹⁷ support the identification of several types of entities, and implement a sophisticated matching layer to support the reuse of identifiers across independently produced (Semantic) Web content. Furthermore, the ENS offers a sophisticated security system based on an innovative combination of the use of personal certificates and role-based access control system, suitable to adapt to any organizational constraints. Unfortunately, the centralized and authority-oriented approach to dealing with identifiers management did not appeal the Semantic Web community. In a ‘Cool URI’ world, where everyone deliberately creates URIs for non-web resources, the main drawback of the ENS 1.0 ID was its being perceived as an authoritative Cool URI, always resolved by the ENS and not by the data owners.

For the reasons above, an evolution of the Entity Name System must be carried out to better meet the requirements of persistent identification and to comply with the goal of supporting the creation and maintenance of persistent cool identifiers.

From ENS to Persistent ENS (ENS 2.0)

The vision we propose in this paper can be realized in a novel infrastructure capable of satisfying persistency requirements and suitable for use when creating Cool URIs. Such infrastructure, depicted in Figure 1, is centered around an Entity Name System supporting the creation of identifiers for digital and non-digital entities (e.g. documents, authors, organizations, etc.) of interest for both traditional digital preservation (e.g. national libraries) and web stakeholders (e.g. dbpedia). Technically, in order to realize such a vision, it is necessary to upgrade the ENS to remove a persistency weakness, and to introduce the separation of the resolver from the identifier. This step is not particularly complex, as the OKKAM ENS ID is structured in the following way:

http://www.okkam.org/ens/id8af7c50f-f072-4384-905b-03875c341863

The domain name (in bold) must be separated to identify the *default resolver* (or any other local resolver using the same local part, as shown in Figure 1), whereas the path and the local part of the URL must be used to mint a new identifier:

peid-8af7c50f-f072-4384-905b-03875c341863

The initial part of the ID has been modified with the prefix *peid* which stands for **Persistent Entity Identifier**. It is important to notice that this evolution does not affect the usability of the identifiers already stored in the ENS, as they would be

¹⁷ The OKKAM Community Portal is available at: <http://community.okkam.org>

maintained as equivalent Cool URIs of the newly generated PEIDs. This simple evolution of the ENS identifier schema introduces the important compliance with the Persistent Identifier requirement and, at the same time, allows its adoption as part of decentralized and independent Cool URIs. Indeed, the *persistent entity identifier* can now be included in a RDF graph as a local fragment of the identifier of a resource.

Different actors can now create or reuse PEIDs for entities of interest using the ENS, as shown in the lower center part of Figure 1, and through their local resolvers enable precise access to information they store (top part of Figure 1). A consistent adoption of PEIDs would support the definition of a powerful information model, giving access to a huge and rich information space by means of global identifiers.

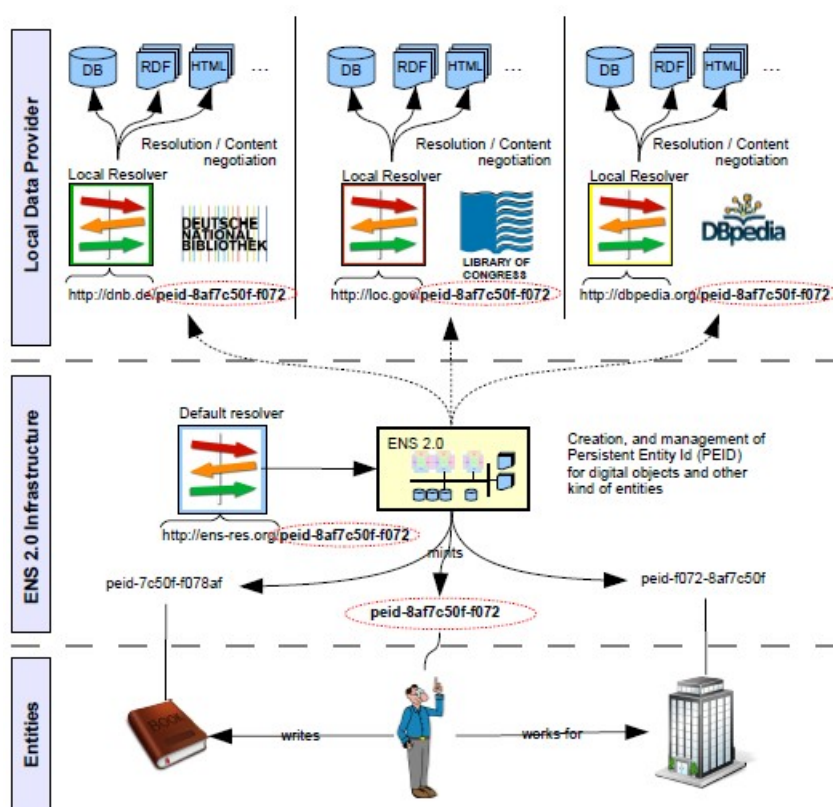


Figure 1. ENS 2.0 vision and application.

The reuse of PEIDs as part of Cool URIs allows the Linked Data practitioners to create URIs resolvable to the information sources they trust the most, whilst still presenting an explicit and persistent reference to the real world entity they are mentioning in their RDF graph and other Linked Open Data. It is important to notice that nothing prevents practitioners from using the default resolver, and thus relying on the ENS as a provider of complete Cool URIs.

Unlike other PI management systems (e.g DOI), we do not plan to allocate a definite range of identifiers for any of the PI local authorities that decides to rely on PEID, as leverage to enable the integration with other PIs authorities and Linked Data. We rather mint each identifier on-the-fly, relying on Universally Unique ID¹⁸ method

¹⁸ Universally Unique Identifier: http://en.wikipedia.org/wiki/Universally_unique_identifier

(roughly a random string generation with near-zero clash probability) and verify at each creation that the identifier was not previously created. Each local resolver will provide a resource for a subset of the PEIDs stored in the ENS.

A new structure for the identifiers is required to support the storage, retrieval and possible categorization of resolvers that give access to information by means of PEIDs. The list of resolvers for Persistent Entity Identifiers extends the set of auxiliary information, such alternative IDs and external references already associated with the current OKKAM IDs.

Enabling the ENS Interoperability Layer Infrastructure

The evolution of the ENS to comply with persistence identifier requirement, together with the storage and categorization of the resolvers, allows us to define a new interoperability layer between persistent identifiers authorities and also a sort of bridge between the two communities. Indeed, the ENS 2.0 can support:

- **Persistent Identifier mapping:** Discovering and storing identity mappings across PIs created and managed in different PI management systems, and making these mappings available through suitable APIs to authorized external applications.
- **PIs-Linked Open data integration:** Enabling full integration of PI management systems with Cool URIs for the same entities mentioned on the Web.
- **Data and service integration:** Using mappings across PIs to support the interoperability of data and services related to a given entity.
- **Identifier search:** Making PIs discoverable and therefore reusable across technical, social and organizational boundaries.

As final remark, it is important to notice that through the ENS search and creation services, it is possible for different stakeholders to incrementally reuse PEIDs unobtrusively, according to their identification scheme and without requiring any relevant changes on their local data representation and storage solutions. Furthermore, this allows the frictionless exchange of information about some specific entities without passing through the ENS, as the PEIDs would provide an unambiguous and precise means of reference to the entities of interest while the related contents are distributed across systems.

Conclusions and Future Work

In this article we propose a novel solution for creating persistent entity identifiers (PEID), which combine the strengths of Persistent Identifiers and at the same time are fully compatible with the Cool URIs approach. Our proposal is based on a fairly simple but essential evolution of the OKKAM Entity Name System regarding the creation and maintenance of global unique identifiers. The simple evolution also allows the definition of a possible interoperability layer enabling the frictionless integration of information managed by existing and future PI authorities, and to bridge the gap between the Persistent Identifier and Linked Data communities. To further

←—————→

enhance the persistence of the proposed approach, we are working on the social/organizational persistence aspects of the ENS, beyond the technical solution adopted. In particular, we are trying to define a organizational layer based on two main entities:

1. A Trust, governed by an international board, responsible for defining rules and guaranteeing rule enforcement, neutrality and availability of the ENS as an open and free (for non profit organization) utility; and
2. A Trustee as a real organization responsible for the practical maintenance and evolution of the system, on behalf of the Trust.

The details of this organizational body are still under discussion. We believe that the combined technical and organizational architectures will provide the groundwork to support the definition of innovative applications, thus bridging the gap between the Web and the PI communities.

Acknowledgements

This work was partially funded by the “Deep Relations” project (Provincia Autonoma di Trento, Legge 6/99, grant number 259812), by the European Commission under FP7 Project No. 296448 “Data Supply Chains for Pools, Services and Analytics in Economics and Finance”, and under FP7 Project No. 269977 “The Alliance Permanent Access to the Records of Science in Europe Network” (APARSEN).

References

- Bellini, E., Cirinnà C., Lunghi, M., Damiani, E. & Fugazza, C. (2008). *Persistent Identifiers distributed system for Cultural Heritage digital objects*. iPRES Conference 2008. Retrieved from <http://www.rinascimento-digitale.it/documenti-ipress2008pi.phtml>
- Berners-Lee, T. (1998). *Cool URIs don't change*. W3C. Retrieved from <http://www.w3.org/Provider/Style/URI>
- Bouquet, P., Stoermer, H. & Bazzanella, B. (2008a). *An entity name system (ENS) for the semantic web*. Proceedings of the European Semantic Web Conference. Springer: Berlin/Heidelberg.
- Bouquet, P., Stoermer, H., Niederee, C. & Mana, A. (2008b). *Entity Name System: The backbone of an open and scalable web of data*. Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008 554-561 IEEE Computer Society.
- Chaudhry, J., Palpanas, T., Andritsos, P., Mana, A. (2008). *Entity lifecycle management for OKKAM*. Proceedings of the Identity and Reference in the Semantic Web Workshop, Tenerife, Spain.



- European Commission. (2010). *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. Retrieved from <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>
- Ioannou, E., Nejdl, W., Niederee, C. & Velegrakis, Y. (2010). *On-the-fly entity-aware query processing in the presence of linkage*. Proceedings of the 36th International Conference on Very Large Data Bases. Singapore.
- Stoermer, H., Rassadko, N. & Vaidya, N. (2010). *Feature-based entity matching: The FBEM model, implementation, evaluation*. Proceedings of the 22nd International Conference on Advanced Information Systems Engineering. Springer: Berlin/Heidelberg. [doi:10.1007/978-3-642-13094-6_15](https://doi.org/10.1007/978-3-642-13094-6_15)