briefing paper

digital preservation europe

# Persistent Identifiers for Cultural Heritage

**It is well-known that Internet resources tend to have a short life; their identification and persistent location pose complex problems that affect many technological and organizational issues involving the citation, retrieval and preservation of cultural/scientific resources. This is by no means technical problem alone: persistent digital object identification, including texts, music, video, still images, scientific documents and the like, is still a major issue that prevents the use of today's Internet as a trustworthy platform for the research and dissemination of scientific and cultural content.**

## Why do we need a 'Persistent Identifier'?

Long term preservation, dissemination and access of cultural digital objects are now among the core missions of cultural institutions such as universities, archives, museums and libraries. The use of URLs can not be considered a reliable approach for addressing these issues due to the structural instability of links (ex. domains no longer available) and related resources (relocation or updating). The current use of the URL approach increases the risk of losing cultural documents or under-using available cultural collections. In the Cultural Heritage (CH) domain it is essential not only to identify a resource but also to guarantee continuous access to it.

A trustworthy solution is to associate a persistent identifier (PI) with a digital resource that will remain the same regardless of where the resource is located.
These are the main steps to be performed in order to implement a PI system:
1) Selection of resources that need a PI
2) Resource name assignment and register creation
3) Resolution of a PI with the associated URL
4) Maintenance of the register that associates PI-URL and guarantee of continuous access to the resources

The first step is the prerogative of each cultural institution whereas the steps thereafter can be delegated to other authorities in order to guarantee better economic and functional sustainability of the service.

## Requirements of a PI system

A CH institution should choose a PI infrastructure using the following system requirements as a guideline:

- Global uniqueness
- Persistence
- Resolvability
- Reliability
- Authority
- Flexibility
- Interoperability
- Costs

## Glossary

“Object” any entity of interest in an intellectual property transaction is defined by metadata and terminology from data dictionary (indecs framework et al) to ensure that “what you mean is what I mean’’ (interoperability) . Objects can be physical, digital, or abstract, e.g. people, organisations, agreements, etc.

Resolution service (dereference): The process in which an identifier is the input (a request) to a network service to receive in return a specific output (resource, metadata, etc)

Naming authority: Independent authority that assigns names and guarantees their uniqueness and persistence. A naming resolution service corresponds to every naming authority and carries out the name resolution. A PI distributed system foresees that the responsibility of generation and resolution can be delegated to other institutions called sub-naming authorities who manage a portion of the name domain/space.

Namespace: an abstract container providing context for the items it holds and allows disambiguation of items having the same name (residing in different namespaces).

Register: Name association table between URNs and one or more URL.

Repository: Place where digital resources are held (DSpace, Fedora, Codex, etc.) with or without a resource management system(file system)

URI: A Uniform Resource Identifier is the generic set of all names/addresses that are short strings that refer to resources

URL: A Uniform Resource Locator is a URI that, in addition to identifying a resource, provides means of acting upon or obtaining a representation of the resource by describing its primary access mechanism or network "location".

### Global uniqueness

We consider the identifier a label that is associated with an object in a certain context. “Context” is intended as both the kind of standard used for the name syntax (i.e. URN:NBN:IT:xxx-xxxx ) and the identification of the authority (sub-namespace) that assigns this label.

### Persistence

Persistence refers to the permanent lifetime of an identifier. It is not possible to reassign the PI to other resources or to delete it. That is, the PI will be globally unique forever, and may well be used as a resource’s reference far beyond the lifetime of the identified resource or the naming authority involved. Persistence is evidently a specific matter in a cultural institution’s service or policy. The only guarantee of the usefulness and persistence of identifier systems is the commitment shown by the organisations who assign, manage, and resolve the identifiers.

### Resolvability

Resolvability refers to the possibility of retrieving a resource only if it is published. It is important to distinguish the concept of identification from resolution. The choice of the identification namespace does not necessarily imply choosing a corresponding resolution architecture.

### Reliability

To assure reliability of a PI system, two aspects have to be assessed: the PI infrastructure must always be active (service redundancy, back-up deposit services, etc.) and the register updated (through automatic systems).

### Authority

The only guarantee of the usefulness and persistence of identifier systems is the commitment shown by the organisations who assign, manage and resolve the identifiers. In the CH domain the tendency is to make use of services provided by public institutions like national libraries, state archives etc. Requirements like the authority and credibility of a PI system should be carefully evaluated before adopting a solution.

### Flexibility

An identifier system will be more effective if it is able to accommodate the special requirements of different types of material or collections. For instance, an identifier system should be able to manage different levels of granularity because what an ‘identifier’ must point to is quite different in the user application fields.

### Interoperability

This aspect is fundamental for guaranteeing the possibility of diffusing and accessing cultural digital objects.

Many technologies and approaches are available and some of them are tailored for specific sector requirements. Among different systems interoperability must be realised at least at the service level offering common and easy user interfaces. System interoperability can be based on the adoption of open standards.

Costs

In the CH domain the PI systems adopted should be free of charge or at least cost-sustainable because the role of cultural institutions is to guarantee free access to resources over time and to avoid a digital divide.

## Other considerations

Granularity

Granularity refers to the level of detail at which persistent identifiers will need to be assigned. The granularity requirement will have a considerable impact on the identifier system an institution adopts.

In some situations, it may be necessary to cite a web page which serves as access to a collection of web files, or to cite a journal article, an item, or a chapter. However, due to rights management, some finer details may be required. Each institution should evaluate whether a PI service provides the right level of granularity for their type of resources.

Opaque or Semantics PIs

A persistent identifier may not contain any information about the object it identifies (opaque id) due to the fact that it is made up of random characters that have no associated semantics. An opaque identifier requires a resolution service in order to be identified, yet it may have some built-in meaning (semantic id).

It is generally easier to memorise and use mnemonic-based identifiers rather than those that contain a meaningless character sequence, although this has no relevance to machine processing.

Versioning

Each new version of a resource will require a separate persistent identifier. A new version can be considered as a different digital object because its content or physical format may have changed. Managing different versions can be achieved through naming rules or metadata fields.

## How can technologies help us?

The PI application requires a database that can keep track of the current location of a digital object, called a ‘resolver database’. A resolver database maps the resource location and redirects the user to the current location. The resolver database and its resolution service may be implemented in different ways: centralized or distributed, DNS-based or not.

Centralized: This architecture is based on a central point that generates the resource names and assures their resolvability and reliability over time. This solution implies a centralization of the responsibilities and costs management; therefore a centralized resolution service has a single point of failure.

Distributed: This architecture requires distributed registers and resolution services for each sub-naming authority committed to manage its own PI names; a “top level authority” manages the resolution redirection process to the appropriate resolution service.

## Current Technologies

PURLs (persistent URL): a PURL is a Persistent Uniform Resource Locator. Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service and by using the standard redirect capabilities of the web server which can redirect the requests for resources using a persistent identifier to the actual location of the document or resource.
www.purl.org

URN (Uniform Resource Name) : URN is a URI that uses the URN scheme, and does not imply availability of the identified resource. URNs are intended to serve as persistent, location-independent resource identifiers and are designed to make it easy to map other namespaces (that share the properties of URNs) into URN-space. Therefore, the URN syntax provides a means to encode character data in a form that can be sent in existing protocols, transcribed on most keyboards, etc.
www.ietf.org/rfc/rfc1737.txt

Handle System: The Handle System is a general-purpose global name service that allows secured name resolution and administration over networks such as the Internet. The Handle System manages handles, which are unique names for digital objects and other Internet resources. A naming authority is authorised to create and maintain Handles, the identifier for it must be unique to that authority but has no prescribed syntax.
www.handle.net

XRI (OASIS Extensible Resource Identifier): The purpose of XRI is to define a URI scheme and a corresponding URN namespace for distributed directory services that enable the identification of resources (including people and organizations) and the sharing of data across domains, enterprises and applications.
www.oasis- open.org

ARK (Archival Resource Key) (IETF Internet draft): The scheme intended to facilitate the persistent naming and retrieval of information objects. A founding principle of the ARK is that persistence is purely a matter of service and is neither inherent in an object nor conferred on it by a particular naming syntax. ARK identifier resolves to 3 different outputs: resource, metadata, preservation commitment..
www.cdlib.org/inside/diglib/ark

N2T (Name to thing): N2T is a consortium of cultural memory organizations and a small, ordinary web server, mirrored in several instances globally for reliability. This project intends to protect 200 organizations’ URLs from hostname instability with 200 rewrite rules by simple HTTP redirects for each organization.
www.n2t.info

DNS-based: The HTTP protocol is used to ‘activate’ the citation link on the web through an HTTP request to a resolution service. This DNS-based approach does not need specific clients or plug-ins for standard internet browsers.

Not DNS-based: Further implementations have helped develop a specific protocol for naming management and PI resolution (e.g. DOI). In this case a specific client (or a browser plug-in) is required to resolve a specific identifier and access the digital objects or their associated metadata. This solution can provide a proxy to extend the service to the HTTP protocol.

## Research opportunities

With the growth of Information technology companies, more and more attention is being given to the issue of URL stability when accessing resources on the Internet. Persistent identifier systems are a relatively new answer to this problem. The extremely dynamic context in which these systems operate is causing large research margins to emerge. Here are some interesting and currently unresolved aspects to study more in depth:

- the current tendency today is to adopt systems which relate to the use domain (eg. NBN in the library domain). However a resource can be part of more than one domain and can be identified by different systems. Thus it is necessary to guarantee interoperability between different identification systems and implementations based on the same namespace;

- Persistent Identifiers allow access to resources but also to their metadata, which are fundamental for enabling the user to identify content. Therefore, it is evermore important to develop advanced metadata management and user services, such as for research that extends to different repositories

- semantic relationships among multimedia objects can be taken in consideration in order to define ontologies and a better understanding of Internet resources.

## Resources

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

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

URN:NBN
http://www.ietf.org/rfc/rfc3188.txt
URN:NBN:DE
http://www.persistent-identifier.de

URN:NBN:IT
http://www.rinascimento-digitale.it

DOI
http://www.doi.org

ARK
http://www.cdlib.org/inside/diglib/ark/

PADI
http://www.nla.gov.au/padi/topics/36.html

PILIN
https://www.pilin.net.au/

OpenURL
http://www.niso.org/committees/committee_ax.html

Emanuele Bellini, Chiara Cirinnà, and Maurizio Lunghi, *Fondazione Rinascimento Digitale,*
{bellini,cirinna,lunghi}@rinascimento-digitale.it