ICT-PSP Project no. 270905

LINKED HERITAGE

Coordination of standard and technologies

for the enrichment of Europeana

Starting date: 1st April 2011

Ending date: 31st October 2013

Deliverable Number:	D2.2
Title of the Deliverable:	*State of the art report on persistent identifier standards and management tools*
Dissemination Level:	Public

Contractual Date of Delivery to EC:	Month 06
Actual Date of Delivery to EC:	August 2012

Project Co-ordinator

Company name: Istituto Centrale per il Catalogo Unico (ICCU)
Name of representative: Rosa Caffo
Address: Viale Castro Pretorio 105, I-00185 Roma
Phone number: +39.06.49210427
Fax number: +39.06. 06 4959302
E-mail: rcaffo@beniculturali.it
Project Web site address: http://www.linkedheritage.org

Context

WP	2
WP Leader	Collections Trust
Task	2.1
Task Leader	Collections Trust
Dependencies	None

Author(s)	Gordon McKenna (Collections Trust, UK) Carolien Fokke (Collections Trust, UK)
Reviewers	KMKG; EDITEUR; MEDRA
Approved by:	

History

Version	**Date**	**Author**	**Comments**
0.1	June 2012	As above.	
1.0	August 2012	As above	Final (subject to final review approval)
2.0	31 August 2012	As above	Added section 1.5 in the Introduction about the state of the deliverable

TABLE OF CONTENTS

1 INTRODUCTION 4

1.1 THE PURPOSE OF WORK PACKAGE 2 4
1.2 ROLE OF THIS DELIVERABLE IN THE PROJECT 4
1.3 APPROACH 5
1.4 STRUCTURE OF THE DELIVERABLE 5
1.5 STATUS OF THIS DOCUMENT 6

2 OVERVIEW OF PERSISTENT IDENTIFIERS 7

2.1 DEFINITION OF PERSISTENT IDENTIFIERS 7
2.2 WHY PERSISTENT IDENTIFIERS? 8
2.3 CONNECTING ENTITIES 11

3 GENERAL DIGITAL IDENTIFIER STANDARDS 12

4 SERVICE-ASSOCIATED DIGITAL IDENTIFIER STANDARDS 16

5 REQUIREMENTS FOR PERSISTENT IDENTIFICATION 20

5.1 CULTURAL HERITAGE INSTITUTION REQUIREMENTS 25
5.2 PERSISTENT IDENTIFIER SERVICE REQUIREMENTS 26

6 LINKED DATA AND PERSISTENT IDENTIFIERS 28

6.1 CREATING COOL URIS FROM NON-URI IDENTIFIERS 28
6.2 IMPLEMENTING URIS 31
6.3 CASE STUDY – BRITISH MUSEUM 33

7 EMBEDDING POLICY FOR PERSISTENT IDENTIFIERS 34

7.1 WHERE POLICY FITS IN 34
7.2 PROMOTING THE BENEFITS OF PERSISTENT IDENTIFIERS 35
7.3 THE ROLE OF THE MISSION STATEMENT 36
7.4 AVOIDING PERSISTENT IDENTIFIER DUPLICATION 37

8 BEST PRACTICE RECOMMENDATIONS 38

8.1 IDENTIFIER STANDARDS 38
8.2 CULTURAL HERITAGE INSTITUTION REQUIREMENTS FOR PIDS 38
8.3 PERSISTENT IDENTIFIER SERVICE REQUIREMENTS FOR PIDS 38
8.4 LINKED DATA AND PERSISTENT IDENTIFIERS 38
8.5 PID POLICY 39

9 CONCLUSIONS 40

9.1 WORK CARRIED OUT 40
9.2 FURTHER WORK IN THE *LINKED HERITAGE* PROJECT 40

APPENDIX: SURVEY OF INSTITUTIONAL MISSION STATEMENTS 42

1 INTRODUCTION

1.1 THE PURPOSE OF WORK PACKAGE 2

Work package 2 of the *Linked Heritage* project (WP 2) is tasked with:

1. Exploring the state of the art in linked data and its applications and potential;
2. Identifying the most appropriate models, processes and technologies for the deployment of cultural heritage information repositories as linked data;
3. Considering how linked data practices can be applied to cultural heritage information repositories, to enrich them and to allow them to align with other linked data stores and applications;
4. Exploring the state of the art in persistent identifiers (both standards and management tools);
5. Identifying the most appropriate approach to persistent identification, e.g. a unique standard or a set of different standards;
6. Designing a feasibility model and realising a demonstrator of a flexible, scalable, secure and reliable infrastructure for a network of 'linked data enabled' cultural heritage information repositories;
7. Exploring the state of the art in cultural metadata models, and in particular their interoperability across libraries, museums, archives, publishers, content industries, and the Europeana models (ESE and EDM);
8. Outlining the potential benefits that richer cultural heritage metadata could bring to Europeana, and to the other services which will use it.

1.2 ROLE OF THIS DELIVERABLE IN THE PROJECT

This deliverable is first of two deliverables which are the outcomes of Task 2.2 – *Resource identification*. This task looks at issues concerning persistent identifiers (PIDs) in cultural heritage information repositories with respect to standards, management best practices and software and hardware architectures for PID assignment and management. Its deliverables are:

- D2.2 – *State of the art report on persistent identifier standards and management tools*;
- D2.4 – *Specification of a management infrastructure for persistent identifiers*.

This deliverable (D2.2) has three roles in the project:

- Educate the partners, and the wider cultural heritage community, about persistent identifiers;
- Give best practice advice based on the use of persistent identifiers in the cultural heritage community, and in particular their use in the context of linked data;
- Inform the subsequent work of WP 2 in the rest of the project:
    - **Task 2.3** – *Technical specifications*: Deliverable: D2.4 – *Specification of a management infrastructure for persistent identifiers*;
    - **Task 2.4** – *Enabling linked cultural heritage data*.

1.3 APPROACH

This deliverable was created based on a process for creating similar deliverables that was developed, and successfully used, during the *ATHENA* project, and earlier, and used in the first deliverable (D2.1). Its steps are to:

1. ***Carry out research*** – Look at what already exists in the environment under discussion. Perhaps survey the project partners on what they are using and or their opinions;
2. ***Make an analysis of the research*** – Look for patterns and trends which can be explained;
3. ***Give simple advice*** – This should be practical and implementable by the partners in the project, and beyond;
4. ***Reuse or create tools*** – Tools should be: easy to use; relevant to the cultural sector audience; and be adaptable, with an open licence, which allows for derivatives to be created (e.g. multilingual versions);
5. ***Identify further needs*** – Leading to further work in the project, and later.

In addition the work undertaken in the ATHENA project has formed a part of the input for the project. The aim in this deliverable is “not reinvent the wheel”.

1.4 STRUCTURE OF THE DELIVERABLE

Section 2 of the deliverable gives an overview of persistent identifiers (PIDs): A definition; why there are important; and their role in connecting entities together in an interconnected network

Section 3 gives outlines the standards for PIDs, while *Section 4* describes the PID services providers that are available.

The requirements for PIDs are explored in *Section 5*, dividing between: Issues for the cultural heritage institution; issues to be assessed against a PID service provider; and a case study of the DOI (Digital Object Identifier) service.

Section 6 looks at PIDs in the linked data environment. The analysis is based on published best practice documents and deals with: The creation of ‘cool URIs’; their generic implementation in a linked data system; and the case study of the British Museum work in this area.

Embedding PIDs as part of the policy of an institution is the subject of *Section 7.* It contains sub-sections on:

- Where policy fits in;
- Promoting the benefits of persistent identifiers;
- The role of the mission statement;
- General collections management policy;
- Sustaining an information system;
- Avoiding persistent identifier duplication.

Best practice advice is given throughout the deliverable, and is brought together in *Section 8*.

Finally there is an *Appendix: Survey of institutional mission statements*

1.5 STATUS OF THIS DOCUMENT

This deliverable is an initial version developed by Collections Trust and its WP2 partners. It has been subjected to extensive review by other project partners, including KMKG, and by EDItEUR and mEDRA who have significant experience of the management of persistent identifier and services[1], but it has not yet been fully revised to incorporate their comments into the content of the deliverable.

Further work and a significantly revised version of the deliverable are already under way by Collections Trust and WP2, to take into account the issues raised during the review process.

[1] mEDRA is a Registration Agency for DOI and manages the ISBN-A service. EDItEUR manages the International ISBN Agency, coordinating the activities of around 160 national ISBN Registration Agencies, and also provides management support to the International ISTC Agency and International ISNI Agency. Note that project partner TIB is also the DOI Registration Agency associated with the DataCite service. These persistent identifiers are of course used across both commercial and cultural heritage sectors.

2 OVERVIEW OF PERSISTENT IDENTIFIERS

> "*The single most important part of the Linked Data approach is the adoption of web-scale identifiers (URIs) to identify things of interest: people, events, places, statistical observations, colours. Anything that we want to publish data about on the web needs to have a URI, allowing it to be referenced, browsed and linked using existing web tools. The existing tools of the web of documents are already designed to work well with things that have URIs. We can "like" them, discuss them, and refer to them in documents.*"[2]

2.1 DEFINITION OF PERSISTENT IDENTIFIERS

Although the subject of persistent identifiers (PIDs) can seem like a technical area of an institution's work, it is actually fairly straightforward. It is about:

- ***Identification*** – Using agreed strings of alphanumeric text (identifiers) to provide access, like a key, to descriptive information in a system. They also provide access to physical items using attached marks or labels.
- ***Persistence*** – Managing the identifiers in order to maintain the access.

Cultural heritage institutions should use persistent identifiers internally for two areas of their work:

Cultural entity identification

This is about the persistent identification of physical items[3], the information describing those items (metadata), their associated cultural entities (e.g. people, places and events), and their surrogates (both physical and digital).

Physical items in managed by cultural heritage institutions include artworks, documents, historical objects, and natural science specimens. Associated cultural entities include the creators and users of the items, the places where they were made or used, and events (e.g. wars) connected to the item.

Physical items are:

- Limited to being in one place at any one time;
- In a difficult to get to storage places;
- In poor physical condition where physical access is dangerous to an item.

To improve access in the above cases 'surrogates' (i.e. substitutes) for physical items are created. Surrogates include: photographs; digital images; 3D models; and physical copies. Digital surrogates, in particular, are used in web services like portals (e.g. Europeana)

A physical item and an institution's own metadata about the item often have the same identifier; with no separate identifier for the metadata record. Surrogates for an item should have different, but perhaps related identifiers.

[2] **Dodds, Leigh and Davis, Ian**. 'Identifier Patterns' in *Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data*. 2012. p4. Accessible from: **http://patterns.dataincubator.org/book/linked-data-patterns.pdf**

[3] The term 'item' is used in the deliverable to refer to all kinds of cultural heritage entity, physical and digital. It is used instead of more domain specific terms like 'object'.

The figure below gives a simple example of this cultural heritage metadata environment:

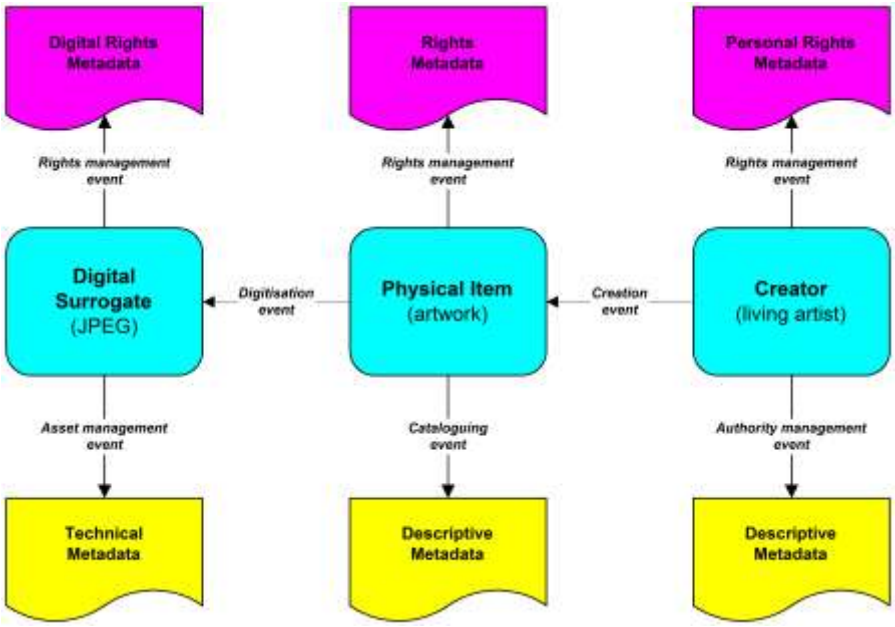

Figure 1: Simplified cultural metadata heritage environment

Institutions often wish to use the in-house identifiers for access to in-house created information, including that about associated cultural entities. However some of these entities already may have published recognised identifiers (e.g. ISBN for books) which can be used. Institutions may use the published information as well.

Collections management identification

This usually covers the identification of three things:

- ***Collection management events*** usually associated with physical items (e.g. acquisition, conservation, movement, IPR licensing, disposal, and exhibition). Surrogates for physical items will have similar kinds of events associated with them.
- ***Display and storage locations*** within an institution. This is used especially for the movement of physical items, but can also be applied to locations in a computer file system for digital surrogates (e.g. the licensing of a photograph of an item).
- ***The institution itself***. Used to externally identify the institution.

We do not cover this area in this deliverable, but limit ourselves to the exploring cultural entity identification. We focus on persistent identification in the context of linked data and its publication.

2.2 WHY PERSISTENT IDENTIFIERS?

Throughout its 'life-cycle' a cultural heritage item (e.g. physical object, archive component, or document) may have had many different 'identifiers' associated with it. Looking at a non-complete set of milestone events in the life-cycle it is possible to see where these identifiers were created, and to which entities they are associated. Here we identify the following milestones:

1. Before item creation;
2. At (or just after) item creation;
3. After creation, but before acquisition by the cultural heritage institution;
4. After acquisition by the institution.

Looking at each milestone in turn it is possible to list possible entities where identifiers are possibly associated.

Before item creation

For some types of material there are items which were used in the creation event or describe it. These items can be directly or indirectly associated with the created item. These include:

- Designs, sometimes registered;
- Patents;
- Preparatory materials, e.g. sketches, drawings, models, photographs;
- Other documentation, e.g. letters or notes about the item;
- Other works that the work might be based, e.g. the statue being painted, the play that the film is an adaption of, an original that this is a copy of).

If the data about these is known to the cultural heritage institution it is usually kept in a 'history file'. This may be paper-based, but is increasingly in digital format. Surrogates (e.g. copy of a patent) for the item might also appear here too.

If the pre-creation items are owned by the institution, or it has a surrogate for them, then it can be referenced by its own managed identifiers. However some of them may be in the collections of other institutions, or in the hands of private individuals, or just in mentioned in publications. These may have identifiers that provide access. Ideally an institution's surrogates will be related to the originals. However an institution would have to create its own identifiers if there is no access to the original.

At (or just after) item creation

This refers to identifiers given to an item, usually by the creator. This identifier may be part of the creation event itself, or take place soon after creation. The identifier will be associated with, and in some cases physically marked on the item. Some of these identifiers will not be unique enough on their own, but can be made unique by the addition of strings to make them so.

Examples include

- Titles;
- Numbers, e.g. print number or edition number;
- Page numbers (i.e. the object is described in a document with that page number).

It is rare that these are globally unique, but some are, so they are usually recorded as information (metadata).

After creation, but before acquisition by the cultural heritage institution

After its creation an item can undergo a series of events which may lead to it being assigned an identifier for that event:

- Being part of a collection (that no longer exists);
- Being in an exhibition or other temporary display (e.g. the Paris Salon);
- Sale catalogues (of dealers or other types of sales body);
- Field collection in an archaeological excavation or other similar event (e.g. find numbers);
- Part of a legal process (e.g. controlled by cultural protection legislation);
- Surrogates – physical and digital (e.g. photographs, copies, 3D model);
- Research;

- Publications;
- Procedures of the institution used during the acquiring the item, e.g.:
    - Pre-entry procedure of the institution;
    - Entry procedure;
    - Acquisition procedure.

The last are part of the process of the item becoming part of the institution's collections, and it is possible that the 'final' identifier will be assigned during them. Earlier identifiers will have a 'life' of their own. They could be referenced, mentioned in publications and other documents both private and public.

Again these identifiers should be recorded in an institution's information system.

After acquisition by the institution

It is also possible that activities, especially those not being managed by the institution, will lead to another 'unofficial' identifier being assigned item. Activities include:

- Exhibition (by another institution);
- Publication;
- Restoration;
- Conservation;
- Creating surrogates (including digital);
- Research;
- IPR licensing;
- Inclusion in an aggregation (thematic, regional, national, and international (e.g. Europeana)).

Any new identifiers created during these events should be recorded in the metadata of the object.

It can be seen that potentially any item will have many identifiers associated with it. This situation ought to be managed if an institution aims to tell the full 'story' of an object. The first step in management of these identifiers is to set up policies to direct it. *Section 7* below explores this later in the deliverable.

2.3 CONNECTING ENTITIES

Managing cultural heritage collections and their metadata, both descriptive and technical, can be viewed as managing links between different 'entities' which form a network of interconnections. Below is a very simplified diagram showing some of these entities viewed from the point of view of physical item (e.g. objects in a museum, items in an archive, documents in a library, and historic site, building, or environment):

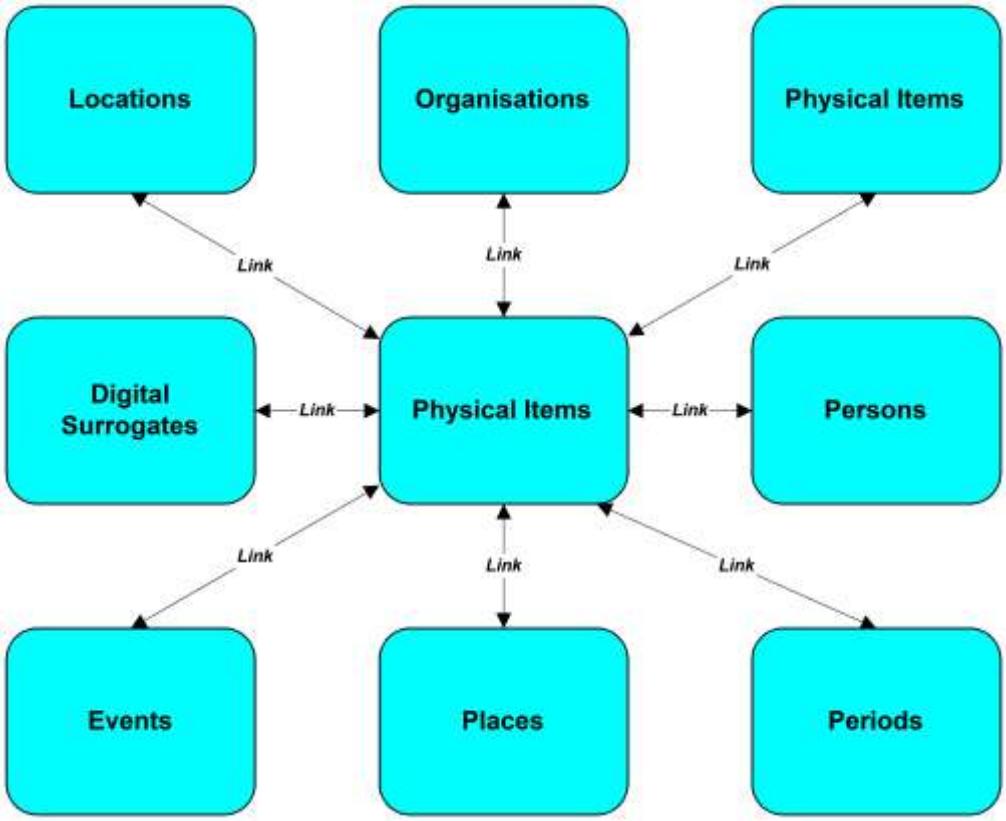

Figure 2: A network of entities

Note that the links work both ways so a collections management system should be able to answer questions like "*Which physical objects (in our collection) are associated with a place*" in addition to "*Which place is a physical item (in our collection) connected with*".

Potentially, at least, all entities could be connected to all the others, e.g. "*Which person(s) are associated with a period?*", "*Which event(s) are associated with a place*", and so on. To answer these questions, even within an internal system will require all the entities to have appropriate separate identifiers (and sensible descriptive metadata). To avoid confusion it should be clear what kind of entity is being referenced by the identifier.

3 GENERAL DIGITAL IDENTIFIER STANDARDS

Employing the same metadata scheme used in deliverable 2.1, we will describe an identifier standard in a Dublin Core (DC) derived format. 9 out of the 15 DC elements are used in the descriptions.

These elements are:

Title	The name (or names) under which the standard is known. Where there is an abbreviated and full name both are given.
Creator	The name of the institution which originally created the service or standard.
Publisher	The name of the institution that makes the standard publicly available.
Date	The date on which the standard was *originally* published.
Identifier	A number or other identifier under which a standard is published or a URL which points to the definition of the standard. Also include is a URL to a service's website.
Rights	Whether rights restrictions apply. An Open Standard, in this context is one where there is no requirement for a licence to use it.
Description	A textual description explaining the standard and its usage.
Subject	Keywords that identify the nature, or scope, of the standard.
Relation	Other standards that this one relates to, and associated websites.

The descriptions are aimed at a general reader, and are generally derived from the document gotten by following the *Identifier* link. More technical details for the standards given can be found in the various references and links given in the records. The purpose of this section is to allow the reader to have an easy reference to the range of relevant identifier standards in one place.

There are three, interrelated, standards for digital identifiers:

- URI (Uniform Resource Identifier);
- URL (Uniform Resource Locator);
- URN (Uniform Resource Name).

URI (Uniform Resource Identifier)

Title	***URI (Uniform Resource Identifier)***
Creator	Berners-Lee, T (W3C/MIT); Fielding, R (Day Software); Masinter, L (Adobe Systems)
Publisher	The Internet Society
Date	2005 (current standard) [original concepts in 1990]
Identifier	**http://www.rfc-editor.org/rfc/rfc3986.txt** (generic syntax)
Rights	[Open Standard]
Description	String of characters used to identify a name or a resource on the Internet. ***Form:*** The syntax of a URI is: `[scheme name]:[scheme-specific part]` • `scheme name` – includes examples as "`http`", "`ftp`", "`mailto`", file, or "`urn`" followed by a colon character, and then by a scheme-specific part • `scheme-specific part` – these are specified in the rules of the scheme. However they must conform to the general requirements for URIs. These include the rules on the use of particular characters. URLs and URNs are URIs.
Subject	identifier (digital)
Relation	URL (Uniform Resource Location); URN (Uniform Resource Name)

URL (Uniform Resource Locator)

Title	URL (Uniform Resource Locator)
Creator	T Berners-Lee (CERN), L Masinter (Xerox Corporation) & M McCahill (University of Minnesota) (Editors)
Publisher	Internet Engineering Task Force (IETF)
Date	1994 [original]
Identifier	**http://tools.ietf.org/html/rfc1738**
Rights	[Open Standard]
Description	A URI (i.e. a string) that specifies: • Where a resource is available; • The mechanism for retrieving it.

Description [continued]	***Form***: `scheme://domain:port/path?query_string#fragment_id` • `scheme` – defines the namespace, purpose, and the syntax of the remaining part, examples: http, https, gopher, wais, ftp. • `domain:port` – gives the destination location for the resource (domain name or IP address). Port is optional, if absent the default is used (for http default port = 80). • `path` – used to specify and find the resource. • `?query_string` – used to pass data to a piece of software to enable retrieval. • `fragment_id` – used to specify a part or a position within the overall resource. E.g. `http://www.linkedheritage.eu/index.php?en/138/about` (the 'About' page on Linked Heritage project website
Subject	identifier (digital)
Relation	URI (Uniform Resource Identifier); URN (Uniform Resource Name)

URN (Uniform Resource Name)

Title	URN (Uniform Resource Name)
Creator	Network Working Group (ed. R Moats, AT&T)
Publisher	Internet Engineering Task Force (IETF) (syntax); IANA, the Internet Assigned Numbers Authority (namespace assignment).
Date	1997
Identifier	**http://tools.ietf.org/html/rfc2141** (syntax)
Rights	[Open Standard]
Description	String acting as persistent, location-independent, resource identifiers, designed to make it easy to map other namespaces. Note that they do not point to a location and therefore might not be resolvable. ***Form***: `urn:<NID>:<NSS>` `<NID>` is the Namespace Identifier, and `<NSS>` is the Namespace Specific String. The Namespace ID determines the syntactic interpretation of the Namespace Specific String.

Description [continued]	E.g. `urn:isbn:0451450523` is URN for *The Last Unicorn*, identified by its book number. Example namespaces: ISBN; ISSN; ISAN; NBN[4]
Subject	identifier (digital)
Relation	URI (Uniform Resource Identifier); URL (Uniform Resource Locator)

In the context of digital identifiers, these three related general standards are the ones that are relevant for identifiers in general. This is supported by the advice for the standards that can be found in the Minerva Project's:

Technical Guidelines for Digital Cultural Content Creation Programmes:

http://www.minervaeurope.org/interoperability/technicalguidelines.htm

[with links to various versions]

It is says[5].

> *"Digitised resources **should** be unambiguously identified and uniquely addressable directly from a user's Web browser. It is important, for example, that the end user has the capability to directly and reliably cite an individual resource, rather than having to link to the Web site of a whole project.* ***Projects should make use of the Uniform Resource Identifier (URI) for this purpose, and should ensure that the URI is reasonably persistent.*** *Such URIs **should not** embed information about file format, server technology, institution structure of the provider service or any other information that is likely to change within the lifetime of the resource.*
>
> *Where appropriate, projects **should** consider the use of OpenURLs, Digital Object Identifiers or of persistent identifiers based on another identifier scheme."*

[4] National Bibliography Number. These are identifiers used by national libraries for those documents (e.g. web pages) where there is no identifier given by the publisher (e.g. an ISBN). The URN namespace for NBNs is described in RFC 3188 (**http://tools.ietf.org/html/rfc3188**). Some national libraries have resolution services for these URNs.

[5] p73 of the current (2008) English Language version.

4 SERVICE-ASSOCIATED DIGITAL IDENTIFIER STANDARDS

The following services support the persistent management of digital identifiers:

- PURL (Persistent URL) & Handle System;
- DOI (Digital Object Identifier);
- OpenURL;
- ARK (Archival Resource Key).

To do this they also define digital identifier standards:

PURL (Persistent URL) & Handle System

Title	PURL (Persistent URL) & Handle System
Creator	OCLC (Online Computer Library Center)
Publisher	OCLC (Online Computer Library Center)
Date	1995
Identifier	**http://purl.oclc.org/docs/help.html#overview**
Rights	OCLC (Online Computer Library Center)

Description	A URL pointing to a resolver (e.g. Handle) which redirects to current URL; Resolver software (OCLC free).
	Form: Has 3 parts –
	1. `Protocol` - used to access the PURL resolver (Handle System).
	2. `Resolver's address` – an IP address or domain name. (Resolved by the Domain Name Server (DNS)).
	3. `Name` – assigned by the user
	E.g.
	`http://purl.oclc.org/oclc/oluc/32127398/1`
	`----  ------------ --------------------`
	`  |  |  |`
	`protocol resolver address  name`

Subject	identifier (digital)
Relation	**http://purl.oclc.org** (PURL website); **http://www.ietf.org/rfc/rfc3986.txt** (Uniform Resource Identifier (URI): Generic Syntax); **http://www.handle.net** (Handle System website) [implementation];

Title	Handle System
Creator	Network Working Group
Publisher	Internet Engineering Task Force (IETF) [specifications]
Date	1994-2003
Identifier	**http://www.ietf.org/rfc/rfc3650.txt** (*Handle System Overview*)
	http://www.ietf.org/rfc/rfc3651.txt (*Handle System Namespace and Service Definition*)
	http://www.ietf.org/rfc/rfc3652.txt (*Handle System Protocol (ver 2.1) Specification*)
Rights	Internet Engineering Task Force (IETF) [specifications]
Description	Specification for a distributed computer system which assigns, manages, and resolves URLs. 'Handles' are the identifiers for digital objects. They are resolved into the information needed to locate and access the objects. Users are redirected to the current location. The information stored in the system has to be maintained with up-to-date information for the service to continue to work.
Subject	persistent identifier resolution
Relation	**http://purl.oclc.org/docs/help.html#overview** (PURL); ISO 26324:2012 (DOI); **http://www.handle.net** (Handle System website) [resolution service]

DOI (Digital Object Identifier)

Title	DOI (Digital Object Identifier)
Creator	International DOI Foundation
Publisher	International DOI Foundation
Date	1998 (creation of International DOI Foundation)
Identifier	ISO 26324:2012 [Information and documentation -- Digital object identifier system]
Rights	[Open standard] (definition); International DOI Foundation (implementation)

Description	A stored and maintained character string used to uniquely identify any kind of entity, physical, digital or abstract. Associated with the DOI is metadata. This can include a location (e.g. a URL) where the referenced document can be found. The metadata is maintained to reflect changes in physical changes in the documents location. ***Form***: Divided into two parts: 1. `Prefix` – identifies the registrant of name; 2. `Suffix` – chosen by the registrant to identify the object associated with the DOI. E.g. `doi:10.345/document.identifier12345` The system is implemented by a federation of registration agencies, coordinated by International DOI Foundation. These pay to be a member of the federation and must agree to meet the contractual obligations associated with the system. A DOI 'name' may be resolved by inputting it to a DOI resolver (e.g. at the International DOI Foundation) or may be represented as an HTTP string by preceding the DOI name by the string 'http://dx.doi.org/' and omitting 'doi:' DOI uses the Handle System resolution service. DOI was approved as an ISO standard in 2010.
Subject	identifier (digital)
Relation	**http://www.doi.org** (DOI website); **http://www.doi.org/hb.html** (handbook) **http://www.doi.org/factsheets.html** (factsheets) **http://www.handle.net** (Handle System) [resolution service]

OpenURL

Title	OpenURL
Creator	Herbert Van de Sompel [original]
Publisher	OCLC (Online Computer Library Center) [standard maintainer]
Date	2000 (original); 2010 (standard)
Identifier	**http://alcme.oclc.org/openurl/docs/pdf/openurl-01.pdf** [original]; ANSI/NISO Z39.88 (*The OpenURL Framework for Context-Sensitive Services*)
Rights	[Open standard]

Description	A URL, with embedded metadata, which enables users to more easily find a copy of a resource. The metadata is used by the resolver service. It is often bibliographic in nature, and OpenURLs are commonly used by libraries. ***Form***: In two parts: 1. `Base URL` for the resolver service; 2. `Query string`. E.g. [original version] `http://www.springerlink.com/openurl.asp?genre=journal&issn=0942-4962` The new standard version is slightly more complicated in form.
Subject	identifier (digital)
Relation	**http://www.oclc.org/research/activities/openurl/default.htm** (webpage)

ARK (Archival Resource Key)

Title	ARK (Archival Resource Key)
Creator	US National Library of Medicine (developer)
Publisher	California Digital Library (maintainer)
Date	2001
Identifier	**https://confluence.ucop.edu/download/attachments/16744455/arkspec.pdf?version=1**
Rights	[Open standard]
Description	A URL scheme which can identify both physical and digital objects. Form: `[http://NMAH/]ark:/NAAN/Name[Qualifier]` `NAAN` = Name Assigning Authority Number - mandatory unique identifier of the organization that originally named the object `NMAH` = Name Mapping Authority Host - optional and replaceable hostname of an organization that currently provides service for the object `Qualifier` = optional string that extends the base ARK to support access to subcomponents of an object or its variants (e.g. version, language).
Subject	identifier (digital); identifier (archival)
Relation	**https://confluence.ucop.edu/display/Curation/ARK** (webpage)

5 REQUIREMENTS FOR PERSISTENT IDENTIFICATION

A review of the literature on PIDs shows that many authors give a set of requirements for their successful implementation. Here are the ones we found:

Reference	Requirements
Bellini, Emanuele; Cirinnà, Chiara; and Lunghi, Maurizio. [IT & EUR] *Briefing Paper: Persistent Identifiers for Cultural Heritage.* Digital Preservation Europe. (2009). **http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf** [EN] **http://www.digitalpreservationeurope.eu/publications/briefs/it_persistent_identifiers_for_cultural.pdf** [IT] **http://www.digitalpreservationeurope.eu/publications/briefs/6_FRENCH.pdf** [FR] **http://www.digitalpreservationeurope.eu/publications/briefs/pt_identificadores%20permanentes.pdf** [PT] **http://www.digitalpreservationeurope.eu/publications/briefs/cz_trvale_identifikatory.pdf** [CZ]	*"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"*

Reference	Requirements
Davidson, Joy [UK] *Persistent Identifiers, 5. Issues to be Considered.* Digital Curation Centre (DCC). (2006) **http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/persistent-identifiers#5**	• *What should the identifier be identifying — the resource, the location, the metadata, or all of the above?* • *Does the identifier need to be globally or locally unique?* • *What basic functionality is required of the identifier scheme (identification vs retrievability)?* • *What level of granularity is required?* • *Are there legacy naming systems that need to be incorporated? If so, how will interoperability between naming systems be handled?* • *Will opaque or semantic identifiers be used?* • *Versioning can be problematic. When does a resource change significantly enough to warrant the application of a new identifier?* • *How will metadata be stored and bound to the identified resource?* • *Can the identification strategy scale to meet future needs?* • *At what stage in the workflow will identifiers be applied to a resource?* • *Who will be responsible for managing the identifiers over time?* • *How will the assignment and long-term management of identifiers be financed?*

Reference	Requirements
Hilse, Hans-Werner and Kothe, Jochen. [NL] *Implementing Persistent Identifiers*. Consortium of European Research Libraries (CERL). (2006). **http://xml.coverpages.org/ECPA-PersistentIdentifiers.pdf**	"... that documents can be identified unambiguously and located by those who need them."
Nicholas, Nick; Ward, Nigel; and Blinco, Kerry. [US] 'A Policy Checklist for Enabling Persistence of Identifiers' in *D-Lib Magazine*. January/February 2009. Volume 15 Number 1/2. Corporation for National Research Initiatives. **http://www.dlib.org/dlib/january09/nicholas/01nicholas.html**	*"... that well-managed resources remain available and accessible over the long term."* *"Persistence involves a guarantee to the user that the identifiers will be kept up to date, and this requires an ongoing commitment of resources. For that guarantee to be meaningful, identifier managers cannot undertake to identify everything in their domain: they need to decide on the resources for which they will provide persistent identifiers."*
Preserving Access to Digital Information (PADI). [AU] *Persistent identifiers*. National Library of Australia. (2002). **http://www.nla.gov.au/padi/topics/36.html**	[on ARKs]: "The scheme is underpinned by three requirements: " • A link from the object to a promise for stewardship; • A link from the object to metadata which describes it; • A link to the object itself (or appropriate substitute).

Reference	Requirements
Sollins, Karen R *Pervasive Persistent Identification for Information Centric Networking* (paper at ACM SIGCOMM 2012) Association for Computer Machinery (2012) **http://conferences.sigcomm.org/sigcomm/2012/paper/icn/p1.pdf**	• Identification assignment; • Management • Access control; • Reachability across space, time (e.g. persistence), scale, and changing conditions.
Tonkin, Emma. [UK] 'Persistent Identifiers: Considering the Options' in *Ariadne*, Issue 56. UKOLN. (July 2008). **http://www.ariadne.ac.uk/issue56/tonkin/**	Looks at: • Opacity • Authority and Centrality • Semantics, Flexibility and Complexity • Present-day Availability and Viability • Technical Solution versus Social Commitment
Wittenburg, Peter (responsible) (EU) *Persistent and Unique Identifiers*. CLARIN Project (2008) **www.clarin.eu/files/wg2-2-pid-doc-v4.pdf**	Gives criteria of: • Contexts of References • Resources and Granularity • Copies • Compatibility and Standards • Additional Information • [No] Semantics • Fragment Addressing

	• Performance/Robustness/Availability • Security • Independence/Openness • Costs

If we analyse the above in terms of requirements:

Requirement mentioned	Found in
Authority	• Bellini (et al); • Tonkin; • Wittenburg
Costs	• Bellini (et al); • Davidson, Joy; • Wittenburg.
Flexible/granular	• Bellini (et al) • Davidson, Joy; • Tonkin; • Wittenburg
Interoperable	• Bellini (et al); • Wittenburg
Management/policy	• Davidson, Joy; • Nicholas (et al); • PADI; • Sollins; • Tonkin; • Wittenburg
Persistentance	• Bellini (et al); • Nicholas (et al); • Sollins • Tonkin • Wittenburg
Reliable	• Bellini (et al); • Wittenburg
Resolvable	• Bellini (et al); • Davidson, Joy; • PADI
Uniqueness	• Bellini (et al); • Davidson, Joy; • Hilse and Kothe; • Sollins • Wittenburg

From this analysis we find that the most useful were those created by Digital Preservation Europe. These we have adapted, and added to, to give 10 requirements that have to be considered when planning to implement PIDs. Some of these should be considered by the cultural heritage institution itself, while the others should be put to a PID service provider under consideration.

5.1 CULTURAL HERITAGE INSTITUTION REQUIREMENTS

Some requirements cover the operations of the institution[6] which is considering using PIDs:

1. Uniqueness environment

A PID is label that is associated with something in a particular environment. On the Internet is should be globally unique, but may only be unique in combination with a limited name space. In the 'worse' case it may only be unique within an institution's own systems.

- **Institutions should be clear, and make public, in which environments its PIDs are unique.**

[6] These requirements should also be considered by institutions setting up services for the creation and management of persistent identifiers.

2. Persistent

Persistence refers to lifetime of an identifier. During this lifetime it should not possible to reassign it another item or to delete it. If an institution can guarantee that a PID will be managed so that it will survive changes to ownership, or PID service, then an external user can be confident of its persistent.

Therefore:

- **Institutions should commit themselves to the persistence of their PIDs. They should make it clear to others what they mean by 'persistent', and how this will be implemented.**

3. Resolvable

The choice to use PIDs does not imply that an external human user will be able to access anything that they can use effectively. Therefore:

- **Institutions should be clear, and make public, information about which, if any, their PIDs resolve to an available resource.**

4. Cost effective

Resources, particularly financial resources, are scarce in the cultural heritage sector. In addition institutions have a general mission to provide access to their items free of charge for non-commercial use. Therefore:

- **Cultural heritage institutions should use PID systems that are free of charge or at very low cost in relationship to their available resources.**

5. Supported by policy

Collections management, which includes access to collections and collections access, is a balance between the competing needs of the institution and its users. Also for anything to be successful it must be supported by the senior management who decide policy. Therefore:

- **The use of PIDs should be part of the written policy of the institution.**

6. Managed by embedded processes and procedures

Having policies on PIDs is only the start in the implementation of a PID system (though an important part). The policy mandate must be made real by how an institution operates. Therefore:

- **The management of an institution's PID system should be part of the written processes and procedures of the institution.**

These last two will be explored further in *Section 7*.

5.2 PERSISTENT IDENTIFIER SERVICE REQUIREMENTS

Other requirements are regarding the operations of the PID service being considered:

7. Reliable

For a PIDs service to function reliably these issues have to be assessed:

1. It should always be active (e.g. backed up, with redundant technology).
2. The register of PIDs should be updated (preferably automatically).

Therefore:

- **Institutions should evaluate and be assured of the technical reliability of a PID service (including their own) before adopting it.**

8. Authoritative

Some PID services are dependent on responsible institutions who: manage the system; assign the identifier; and resolve the identifiers to resources. Some services are provided by public institutions like national libraries and archives. For a service to be effectively supported a responsible institution must be able to demonstrate its commitment. Therefore:

- **Institutions should evaluate and be assured of the authority and credibility of a PIDs service's provider before adopting that system.**

9. Flexible

A PID system will work more effectively if it can handle the requirements of different types of collections. Parts of collections may be managed at different levels of 'granularity', from parts of an item, to individual items, to sets of items. The latter has an unbounded number of individual elements. Therefore:

- **Institutions should use PIDs services that are flexible enough to represent the granularity their collections.**

10. Interoperable

This is vital to ensure that cultural content can be shared and used by as a large a number of users as possible. Many PID solutions were designed for specific domains. Therefore:

- **Institutions should use intellectually open standards for the implementation of PIDs**

These criteria can form the basis of methodology for the testing of an institution's internal PID system, or the suitability of a prospective PID service provider. We will use it in the creation of deliverable D2.4 – *Specification of a management infrastructure for persistent identifiers*.

6 LINKED DATA AND PERSISTENT IDENTIFIERS

Tim Berners-Lee[7] gives four 'rules' or 'principles' for linked data (our emphasis):

1. ***Use URIs*** *as names for things*
2. ***Use HTTP URIs*** *so that people can look up those names.*
3. *When someone looks up a URI,* ***provide useful information****, using the standards (RDF*, SPARQL)*
4. *Include* ***links to other URIs****, so that they can discover more things.*

The URI is mentioned in all of the four. Therefore it is obvious that that persistent identifier URIs form a key component in linked data. So best practice advice[8] is:

> **To use persistent identifiers for things in the form of persistent URIs, which provide information to the user.**

Here:

- ***Things*** are the full range of entities identified above (e.g. physical items, digital surrogates, people, institutions, places, events, and periods);
- ***Provide information*** via the Internet, specifically the Web;
- ***User*** is can be a human being or a machine capable of using the information.

Such a persistent URI has become known as a 'cool URI', and are part of the Semantic Web[9].

The persistent identifiers created and managed by the service providers mentioned in section 6 above can be used for the linked data environment. For example in April 2011[10] the Registration Agency CrossRef announced that the DOIs assigned by them had been enabled for linked data.

6.1 CREATING COOL URIS FROM NON-URI IDENTIFIERS

An issue that needs to be considered is how to design a URI based on existing non-URI identifiers. This is particularly the case where institutions want to manage their identifiers. Work by Dodds and Davis[11] has investigated this and they say:

> "*Successful publishing of Linked Data requires the careful selection of good, clean, stable URIs for the resources in a dataset. This means that the most important first step in any Linked Data project is deciding on an appropriate identifier scheme: the conventions for how URIs will be assigned to resources*."

The URI design patterns they give are "*in wide use today and so are tried and tested in the field*."

They identify 8 patterns:

- ***Hierarchical URIs*** – where a set of items are arranged in a natural hierarchy (e.g. a book with chapters, or record release with tracks);
- ***Natural Keys*** – where data already contains a unique identifier (e.g. ISBN);
- ***Patterned URIs*** – for the creation of more predictable, human-readable URIs;
- ***Literal Keys*** – for non-global identifiers;
- ***Proxy URIs*** – dealing with the lack of standard identifiers for third-party resources;
- ***URL Slug*** – creating URLs from arbitrary text or keywords;

[7] **Berners-Lee, Tim**. *Linked Data - Design Issues*. 2009. Accessible from: **http://www.w3.org/DesignIssues/LinkedData.html**

[8] As suggested by **Minerva Project**. *Technical Guidelines for Digital Cultural Content Creation Programmes*. (2008), p73 Accessible at: **http://www.minervaeurope.org/interoperability/technicalguidelines.htm** [with links to various versions].

[9] The current Web is a 'web of documents'. The Semantic Web is a 'web of data'.

[10] For further details see: **http://www.crossref.org/CrossTech/2011/04/content_negotiation_for_crossr.html**

[11] **Dodds, Leigh and Davis, Ian**. *Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data*. 2012. pp4-11. Accessible from: **http://patterns.dataincubator.org/book/linked-data-patterns.pdf**. We also use their examples.

- ***Rebased URI*** – constructing one URI based on another;
- ***Shared Keys*** – simplifying the inter-linking of datasets.

Here are the patterns in more detail:

Hierarchical URIs

That collections and items can be arranged in a hierarchy is common. Therefore making the URIs for these will come naturally. The other advantage of doing this is that they are 'hackable' by users and developers. This means that is possible for them to 'navigate' up the hierarchy of entities by progressively removing the end sections of a URI.

The suggested solution is to create a URI which follows this pattern:

`:collection/:item/:sub-collection/:item`

An example would be where there is a collection of newspaper titles, with issues, and pages:

`titles/abc/issues/date/pages/1`

Note the use of descriptive labels for the elements of the hierarchy.

Here a user or machine could 'hack' to create a URI that gives them information about an issue of a newspaper, and hack again to give information about the newspaper title in the collection.

In can be envisaged that this pattern works best where the sub-items are always associated with one parent item in the collection, e.g. a page is always part of an issue. If this is not true the authors suggest using *Patterned URIs*.

The *discogs* dataset in *dataincubator* uses the form:

`http://discogs.dataincubator.org/release/22530/track/1-01`

Natural Keys

This pattern addresses the situation where a group of items already has a unique identifier, e.g. a database key field, or a URI which cannot be directly used on the web, e.g. ISBN. The existing identifier is used, via an algorithm, to create a usable URI. A simple way to do this is to concatenate the identifier to a suitable base URI.

The advantage of doing this is to avoid a situation where the same set of items has two identification systems which the need to map between them.

This pattern is often used in conjunction with *Patterned URIs*.

An example of use is the URIs at *BBC Programmes* which are derived from existing `programme id`s.

Patterned URIs

The aim here is to have URIs which are more predictable and make sense to human readers. They should be easier to remember, and for system developers to work with. The latter also allow other URIs to be constructed or hacked based on knowledge about a given example URI.

This functionality is enabled by following a simple naming pattern, for example based on the pluralised class name of the item:

`/objects/1234`

`/objects` indicate the collection of books objects respectively, and `1234` is the identifier of a particular object.

Using this technique ensures that the URI scheme has is the same as that for the underlying data, so providing a clear relation between the URI and the type of thing that it describes.

The BBC website uses `/programmes` to group together URIs that relate to series, brands and episodes.

Literal Keys

The Natural Keys technique, described above, enables the creation of URIs from existing identifiers. However it does not address the issue of how to publish these identifiers in RDF, nor the situation where natural keys change (e.g. ISBN-10 moving to ISBN-13).

This need is met by publishing the natural identifier as a literal value within a sub-class of an existing RDF vocabulary property. The suggestion is to publish within a sub-class of Dublin Core, `dc:identifier`. This has the additional advantage of the system being able to look up an associated resource, using a SPARQL query, and to support multiple identifiers for the same resource.

The *nasa* dataset in *dataincubator* uses Patterned URIs based on the NSSDC international designator, but includes these as literal values associated with each spacecraft using a custom property.

Proxy URIs

A linked data system generally needs to be able to deal with third-party resources, which often do not have URIs. To deal with this situation linked data publishers will have to create URIs from within their own domain, thus treating them identically to their own data.

When the third-party resources do have published URIs some alignment will have to take place. One way of achieving this is to publish equivalence links, using, for example, `owl:sameAs` or `skos:exactMatch`.

For example: There is still no agreed standard way of generating URIs for Internet Media Types. IANA have adopted RDF for publishing descriptions of registered media types. A data set containing descriptions of images may therefore use locally minted URIs for those media types:

```
ex:anImage a foaf:Image;
      dc:format <http://www.example.org/media-types/image/jpeg>
```

URL Slug

It is likely that a linked data publisher will have to deal with the situation where there is no numeric natural key for an entity; there is only a title or keyword. To get round this the publisher should generate a URL ‘slug’ from the available text. This text must be normalised for use on the Web by:

- Transforming the string to lowercase;
- Removing any special characters or punctuation that would require encoding a URL;
- Replacing spaces with dashes.

The original text should then be ‘preserved’ by using it to label the resource, e.g.:

```
#Generate a patterned URI with a simple URL slug from "Late Antiquity"

<http:www.institution.org/time-periods/late-antiquity>

rdfs:label "Late Antiquity"
```

However this technique has the issue of it being possible, even within the same system, of the same slug being generated for different resources. To alleviate this use the *Patterned URIs* technique to create different URI ‘sections’ for different types of item, e.g.

```
http:www.institution.org/plays/same-title

http:www.institution.org/films/same-title

http:www.institution.org/books/same-title
```

Rebased URI

This pattern is more technical, and is used where a linked data system or service needs to carry URL rewriting based on an existing URI. This is typically to support URI resolution for remote (RDF) resources.

One option is to prefix the resolving service URI onto the original URI with a parameter indicator.

`http://service.company.com/resolve?uri=http://institution.org/object/1234`

There are number of other options for rebasing, which we will not go into here, but they can be found in the **Dodds and Davis**, p9.

Shared Keys

One of the things which will simplify the publication of linked data is the convergence of Web identifiers. To encourage this process linked data publishers should create URIs based on a set of common non-web identifier, that are recognised and used by a sector or domain.

Linked data publishers should create *Patterned URIs* by applying the *Natural Keys* pattern, but prefer public, standard identifiers rather than those from internal systems. For example:

- MusicBrainz URIs are *Patterned URIs* built from a '`MusicBrainz ID`' (e.g. a74b1b7f-71a5-4011-9441-d0b5e4122711). This gives a URI in the form:

  `http://musicbrainz.org/artist/a74b1b7f-71a5-4011-9441-d0b5e4122711`

- The BBC decided to create URIs for artists that are algorithmically related to the MusicBrainz URIs using a common *Shared Key*. This gives a URI in the form:

  `http://www.bbc.co.uk/music/artists/a74b1b7f-71a5-4011-9441-d0b5e4122711`

Constructing URIs in this way from public identifiers means that they already can potentially be used outside of the immediate application. They ease inter-linking. For example, the pattern may avoid the need to look-up URIs in a SPARQL endpoint, allowing a developer to simplify use a URI template.

This pattern is best suited to situations where the shared identifiers are stable and rarely change.

Creating URIs – Best practice advice

Use the URI creation patterns and techniques given by Dodds and Davis[12]

- **Hierarchical URIs;**
- **Natural Keys;**
- **Patterned URIs;**
- **Literal Keys;**
- **Proxy URIs;**
- **URL Slug;**
- **Rebased URI;**
- **Shared Keys.**

to create a linked data system.

6.2 IMPLEMENTING URIS

The Web, with its clients (e.g. web browser and Semantic Web-enabled user agents) and servers, uses the HTTP protocol[13]. In general clients make requests for web documents (via a URI) and servers meet those requests. In addition HTTP supports a mechanism called 'content negotiation' which allows servers deliver different types of document, and different language versions. Different types can include HTML pages, XML, and RDF. Therefore a server can deliver HTML to a 'traditional' browser and RDF to

[12] **Dodds, Leigh and Davis, Ian**. *Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data*. 2012. pp4-11. Accessible from: **http://patterns.dataincubator.org/book/linked-data-patterns.pdf**
[13] See: **http://www.ietf.org/rfc/rfc2616.txt**

Semantic Web-enabled user agent. This functionality can be used to implement cool URIs in one of two ways[14]:

Hash URIs

A URI can have an additional part at its end which is separated by hash ("#") symbol. However this part is stripped off before making the request to the server. Therefore a URI that has a hash cannot be retrieved directly, and need not represent a Web document. So they can be used to identify non-document things unambiguously.

In an example where we wish to represent a set of items held by an institution we would have URIs like:

`http://www.institution.org/item#example1`

`http://www.institution.org/item#example2`

`http://www.institution.org/item#example3`

A Web client strips off the hash fragment to give a URI request to the server of:

`http://www.institution.org/item`

At this URI the institution's server could serve the client with an RDF document that contains descriptions of all the items, using the original hash URIs to identify them.

Content negotiation could also be used to redirect from the `item` URI to HTML or RDF representation. Which is returned would be based on the client and server configuration.

303 URIs

Here the HTTP status code, `303 See Other`, is used to the show that the requested item is not a Web document. What happens is that the request is redirected to a URI of a document which is about the item. So:

`http://www.institution.org/item/example1`

Would be redirected to:

`http://www.institution.org/doc/example1`

Content negotiation is then used to decide whether to return HTML or RDF (or more alternative forms). So an RDF document might have a URI of:

`http://www.institution.org/data/example1`

This RDF document would contain statements about the item, using the original URI, `http://www.institution.org/data/example1`, to identify it.

Implementing URIs – Best practice advice

Which of these two should be used? Sauermann and Cyganiak suggest that a publisher should use:

- **Hash URIs – for rather small and stable sets of resources that evolve together;**
- **303 URIs – where a publisher wants to have the flexibility of being able to deliver data about single items, groups of items, or all the items in repository. Using this method has the downside of slower response times.**

It is possible to combine both approaches to come up with the best solution under particular circumstances.

[14] **Sauermann, Leo and Cyganiak, Richard** (Eds.). *Cool URIs for the Semantic Web*. 2008. Accessible at: **http://www.w3.org/TR/cooluris/#solutions**

6.3 CASE STUDY – BRITISH MUSEUM

In September 2011 the British Museum implemented published its own linked data repository, with its in-house created PIDs[15]. The decision to create their own PIDs rather than using a service was partly driven by the cost implications of a using a service, and partly by a typical 'museum view' of PIDs. This is encapsulated in the *Statement on Cultural Resources in Digital Environments* by Light and Stein[16] which is a statement of principles which may be proposed to ICOM (International Committee of Museums) by its international documentation committee (CIDOC) at its triennial in 2013.

Its proposed principles are:

- Museums are the sole authority with responsibility for establishing globally unique and persistent identities (URIs) for each of the objects in their collections;
- Each museum should establish and publish on the internet such a unique and persistent identity – preferably as http URI (=URL) – for each of its objects;
- This URL should resolve to a human-readable description of the object, which is sufficiently detailed to identify it unambiguously;
- Ideally, this URL should additionally resolve to a comparable description in a machine-processible format, using best practice Linked Data principles;
- When describing the relationship of the collection object to its cultural context (people, places, events, etc.), the museum should where possible use URLs from common frameworks, rather than minting its own URLs for these concepts;
- A museum can choose to delegate this responsibility;
- The museum should encourage other institutions to use this set of URLs, by publishing metadata such as VoID descriptions (see http://www.w3.org/TR/void) of its collection resources;

Technical implementation

The museum decided to use the *303 URIs* method of implementation described above They also used the following pattern suggestion based on its domain name (`britishmuseum.org`):

`http://collection.britishmuseum.org/id/object/[PRN number]`

The PRN number is a museum-internal unique identifier for an object in its collections. This is the *Patterned URIs* technique suggested by Dodds and Davis.

E.g. The Rosetta Stone has the PRN number: YCA62958. Therefore its URI would be:

`http://collection.britishmuseum.org/id/object/YCA62958`

By default this will redirect to an html representation of the data:

`http://collection.britishmuseum.org/description/object/YCA62958.html`

One also has the option of receiving data in other formats like RDF/XML, Notation-3, Turtle, and JSON.

The British Museum is using standards and best practice techniques to implement PIDs in its linked data system.

[15] See: **http://collection.britishmuseum.org**

[16] Unpublished (2012).

7 EMBEDDING POLICY FOR PERSISTENT IDENTIFIERS

7.1 WHERE POLICY FITS IN

The management of its collections, and the associated information, is a major activity of any cultural heritage institution. Success in this a challenge to reach a balance between:

- Giving **access** to collections and ensuring the **preservation** of collections;
- The needs of the **collections** and the needs of the **people** who want to use them;
- **Institutional priorities**, which range from short, to medium, to long terms.

In order to meet the challenge the British Standards Institute developed, with the help and sponsorship of cross-domain set of cultural heritage institutions in the UK, a **Publicly Available Specification** (PAS 197) on a *Code of practice for cultural collections management*[17]. This was published in 2009, and will be reviewed soon with the aim of it forming the basis for national, and possibly international, standard in this area. The *Code* is not specific to any of the cultural heritage domains but was designed to be applicable to all.

The *Code* aims to:

- Enable a cultural heritage institution's top management to take a strategic and integrated approach to collections management.
- Provide a blueprint for creating strategies that are sustainable.
- Take into account the legal environment within which an institution operates.

At the top level of the hierarchy is the:

- ***Mission statement*** – A strategic statement giving a cultural heritage institution's fundamental purpose, especially with regard to its collection.

The mission statement **informs** the different areas of collections management policy which are based on three different strands of activity:

- ***Collections information***;
- ***Collections development***;
- ***Collections access***;
- ***Collections care and conservation***.

These policies are **met by** (implemented) processes and procedures which the institution uses. These may be based on a standard, like *SPECTRUM* for museums. Similar processes and procedure should be met by equivalents in the other domains. In all cases they must be documented in the form of a written manual adapted for the institution.

It should be noted that the creation of the mission statement, policies, processes and procedures is not a one-time only process. There must be a commitment to continual review and change of the framework.

[17] **British Standards Institute**. 2009. *PAS 197, Code of practice for cultural collections management*.

7.2 PROMOTING THE BENEFITS OF PERSISTENT IDENTIFIERS

One barrier for the institution to using persistent identifiers (PIDs is the experience of top management. An institution may be lucky, and they understand that PIDs are important, but this is unlikely. The subject of will probably be seen as a technical area of the institution's work and at too low a level. To alleviate this difficulty it is perhaps a good idea to 'sell' PIDs to top management in terms they will understand.

Implementing and maintaining the use of PIDs in an institution requires both investment, in time and other resources, and commitment from staff and management. In order for this to happen the first task that needs to be carried out is the creation of a document that outlines why managing PIDs will benefit the institution and its users – a 'business case'.

One way of making a business case is to set out the benefits. Benefits can be both 'direct' and 'indirect'. Direct benefits are those which are a result in doing the activity being considered. Indirect benefits are the beneficial 'by-products' of carrying of the activity. For PIDs the direct and indirect benefits are:

Direct benefits

- Ability to retrieve information and physical items, quickly and simply. Everything, both physical and digital, will be associated with an identifier which will point to it or to information about it. Access will be through a central index of some kind.
- Cost savings in staff time spent handling items or re-identifying information. Very important at all times, but especially at the time of the writing of this deliverable!
- Greater confidence in managing information and items. Managers will be able to demonstrate that they are managing the institution according to best practice.
- Improved access to information for all areas of curatorial expertise and other departments. Communications are improved. There are no 'information silos' that need mediation to enter.
- Using a standards-based approach will support applications for funded projects (e.g. from the EC). Funders are now demanding that beneficiaries conform to well know standards and best practices. Showing to them that an institution is using standards will contribute to success

Indirect benefits

- Greater clarity to funders about the extent and content of the institution's collections. Using PIDs will ensure that this can be demonstrated.
- Better-managed intellectual property leading to greater opportunities for use and commercial activity. Being able to link, via PIDs, IPR licenses to the works they cover will contribute to efficient management. Better management will also lead to better commercial exploitation.
- Enhanced ability to publish information and to make your collections visible online. Using PIDs that always point to online information will mean that the content will be there 24/7. Some aggregators, e.g. Europeana, will reward this by offering persistent content more prominence.
- Ability to share information through portals: local; regional; national; thematic; and international. PIDs form the necessary link to the content online, and therefore are the key to open the door to participation.
- Ensuring that information and knowledge is used effectively in the future even if local staff changes. It is always a stressful time, for an institution, when staff leaves. Always working in a way which ensures that their knowledge is always available to others will reduce that stress. PIDs are a key to that knowledge, pointing to shared information.

These benefits should appeal to top management and, hopefully, will gain their 'buy in'.

7.3 THE ROLE OF THE MISSION STATEMENT

In order to get an overview of the mission statements that institutions providing content to Europeana through the Linked Heritage project have a survey of their websites was carried out. This was supplemented by personal communication to key institutions in the project where additional information was thought necessary. It was assumed that given the range in different kinds of institutions represented by the Linked Heritage project the sample looked at is typical for the cultural heritage sector in general.

In looking at an institution's website it became obvious that not all had a section called '*Mission Statement*'. However it was possible to find 'mission-like' statements embedded in other parts of the website. These were found in sections like "*About us*", "*History of the museum*", and "*Legal basis*". They could also be found in documents accessible through the website like the "*Annual Report*".

The *Appendix* at the end of the deliverable shows the results of the survey. Note that some texts have been edited to make the statements for readable.

The lengths of statements vary in size. In one case the '*Mission Statement'* of the institution was a full colour, multi-page, booklet covering the full range of its activities. However most were of a more manageable size consisting of a few sentences, paragraphs or bullet points. At the other extreme a few were just a handful of words.

Looking at the mission statements it is possible to see four themes that they contain:

- ***Collection*** – what the institution has to offer. The types are:
    - Geographic reach (e.g. building, locality, region, nation, continent, worldwide)***;***
    - Temporal range (e.g. middle ages, contemporary);
    - Thematic basis:
        - Object type (e.g. fine art, visual art, natural science);
        - Human activity (e.g. dance);
        - Event (e.g. World War I);
        - Person, people, institution (e.g. writers, artists, politicians);
        - Subject (e.g. agriculture, music, archaeology, science, history, ethnography, folklore, forest culture, sport, oral tradition).
- ***Activities*** – that take place with the collection and with audiences. Examples are:
    - Acknowledge;
    - Advance knowledge;
- ***Audience*** – Who the collection and activities serve:
    - [Everyone – by implication];
    - Public;
    - Visitors;
- ***Quality*** – The standard of service provided. This is the least likely element to be in the statement and is usually an aspiration:
    - Achieve a balance;
    - Advanced;
    - Agent of change;

Obviously there is no mention of identifiers (persistent or not) in any of the mission statements. However it is possible to see the potential for implementing PIDs at a lower level of policy. In general they can be seen as helping to meet the parts of a mission statement that deals with audience, activities and especially quality:

- PIDs will allow links to relevant items, and other entities to be seamlessly followed. The users' online experience of collections will be significantly enhanced.

- PIDs are essential where information is shared between institutions and aggregated into services like Europeana.
- Activities where the links between objects need to be made, e.g. research, and the online experience, will benefit from PIDs.
- Using PIDs will make it simple to demonstrate the quality of an institution's service.
- PIDs will ensure that there will be no broken links between items and information.

Therefore best practice advice is that:

> **An institution's mission statement should include elements on audience, activities, sustainability, and quality that give a general environment for the implementation and management of persistent identifiers.**

7.4 AVOIDING PERSISTENT IDENTIFIER DUPLICATION

One important aspect of PID management is ensuring that the institution does not assign multiple PIDs to the same thing – physical or digital objects and collections. The consequence of assigning multiple PIDs to the same thing will be to cause confusion, incorrect links, and partial network of information.

Internally the issue should be:

- Mandated by appropriate policy;
- Managed by the roles identified in the last section;
- Implemented using the instructions in the relevant sections of an institution's procedural manual;
- Enabled in an institution's collection's management system.

The last requirement will probably be enabled, in a computer-based system, by maintaining a 'registry' of assigned PIDs and not allowing a change of PID without appropriate authority.

All four of these requirements should be in place for this need to be met. There is a danger that if any is missing that the others will not work properly.

Externally, particularly in the online environment of the Internet, the issue of multiple PIDs is mitigated by publishing PIDs, with appropriate descriptive and technical metadata for the things they are identifying. It is important to make clear the thing being identified by the PID. This will avoid confusion between the physical and its digital surrogate(s). Links, using PIDs, from digital surrogates to physical objects, and vice versa, should also be included in the metadata.

The PID systems discussed above can manage the external publication of PIDs. Management of PIDs is similar to that internally, with similar management controls and a maintained registry. However institutions may choose not to use them and instead publish PIDs and metadata themselves. One way of doing this would be to publish this information as 'linked open data'.

8 BEST PRACTICE RECOMMENDATIONS

This section summarises the best practice for digital identifiers which is given throughout the deliverable.

8.1 IDENTIFIER STANDARDS

For the creation of digital identifiers, for the entities they intend to manage, institutions should:

Make use of the Uniform Resource Identifier (URI) for this purpose, and should ensure that the URI is reasonably persistent.

8.2 CULTURAL HERITAGE INSTITUTION REQUIREMENTS FOR PIDS

The following are a set of institutional requirements which need to be considered when starting to use PIDs. Institutions should:

- **Be clear, and make public, in which environments its PIDs are unique;**
- **Commit themselves to the persistence of their PIDs. They should make it clear to others what they mean by 'persistent', and how this will be implemented;**
- **Be clear, and make public, information about which, if any, their PIDs resolve to an available resource;**
- **Use PID systems that are free of charge or at very low cost in relationship to their available resources;**
- **Make sure that the uses of PIDs are part of the written policy of the institution;**
- **Make sure that the management of an institution's PID system is part of the written processes and procedures of the institution.**

8.3 PERSISTENT IDENTIFIER SERVICE REQUIREMENTS FOR PIDS

If an institution is considering using a PID service it should:

- **Evaluate and be assured of the technical reliability of a PID system (including their own) before adopting it;**
- **Evaluate and be assured of the authority and credibility of a PIDs service's provider before adopting that system;**
- **Make sure that the service it uses is flexible enough to represent the granularity their collections;**
- **Make sure that the service uses intellectually open standards for the implementation of PIDs.**

8.4 LINKED DATA AND PERSISTENT IDENTIFIERS

Based on Berners-Lee's four principles of linked data institutions should:

Use persistent identifiers for things in the form of persistent URIs, which provide information to the user.

When creating URIs, from non-URI identifiers, institutions should:

Use the URI creation patterns and techniques given by Dodds and Davis[18]

[18] **Dodds, Leigh and Davis, Ian**. *Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data*. 2012. pp4-11. Accessible from: **http://patterns.dataincubator.org/book/linked-data-patterns.pdf**

- **Hierarchical URIs;**
- **Natural Keys;**
- **Patterned URIs;**
- **Literal Keys;**
- **Proxy URIs;**
- **URL Slug;**
- **Rebased URI;**
- **Shared Keys.**

to create a linked data system.

When implementing URIs institution should use:

- **Hash URIs – for rather small and stable sets of resources that evolve together;**
- **303 URIs – where a publisher wants to have the flexibility of being able to deliver data about single items, groups of items, or all the items in repository. Using this method has the downside of slower response times.**

It is possible to combine both approaches to come up with the best solution under particular circumstances.

8.5 PID POLICY

An institution's mission statement:

An institution's mission statement should include elements on audience, activities, sustainability, and quality that give a general environment for the implementation and management of persistent identifiers.

9 CONCLUSIONS

9.1 WORK CARRIED OUT

In this deliverable we have:

- Given an overview of persistent identifiers (PIDs):
    - Definition of PIDs
    - Why persistent identifiers?
    - Connecting entities?
- Looked at general and service-related identifier standards.
- Examined the requirements for persistent identification in terms of:
    - The cultural heritage institution;
    - PID service provider.
- Explored linked data and persistent identifiers:
    - Creating cool URIs;
    - Implementing URIs;
    - Case study – British Museum.
- How to embed policy for PIDs:
    - Where policy fits in;
    - Promoting the benefits of PIDs;
    - The role of the mission statement;
    - Avoiding PID duplication. .
- Gave best practice advice on:
    - What PID standards to use;
    - Cultural heritage institution requirements for PIDs;
    - Persistent identifier service requirements for PIDs;
    - Linked data and PIDs;
    - PID policy.

The deliverable, as a whole, represents an introduction to the topic, as well as acting as guidance for the rest of the project.

9.2 FURTHER WORK IN THE *LINKED HERITAGE* PROJECT

In the next 12 months work package 2 must create the following deliverables:

- D2.3 – *Specification of the technologies for large-scale implementation of cultural heritage linked data:* Technical specifications and demonstrator. The demonstrator will show an example of application of the solution proposed by Linked Heritage, applies to a range of content used for the investigation on enrichment processes. (Month 18).

  This deliverable will be based on a series of use cases for the publication and consumption for the linked data. The creation of these case studies will be the responsibility of CT, but will involve the thematic working group, who will input their experience.

- D2.4 – *Specification of a management infrastructure for persistent identifiers*: Technical and organisational specifications (Month 24). This will test the various PID services against the requirements set out in section 5 of this deliverable.

APPENDIX: SURVEY OF INSTITUTIONAL MISSION STATEMENTS

This appendix gives the mission statement (or mission-like statement), where available, of the cultural heritage institutions in the Linked Heritage project:

Belgium

Institution	**Mission statement (or mission-like statement)**
Packed – Platform voor de Archivering en Conservering van Audiovisuale Kunsten	*"The region of Flanders is striving to hold a top position in the information society. To be a European top region, our community requires a trustworthy, qualitative and sustainable digital memory. As a knowledge hub, PACKED vzw means to contribute to the creation of this memory."*
Royal Museums of Art and History	• "*Collecting and preserving cultural objects with a museological or scientific value in the museums' domain (Art, Archaeology, Musicology, History, Ethnology, Folklore - collections from all continents)* • *Keeping of a general inventory, archival information and documentation centre on its collections* • *Performing scientific research in connection to its collections* • *Valorisation and diffusion of the results of this scientific research on national and international level* • *Active participation in scientific projects and conferences on national and international level* • *Providing information on its collection to the general public and scientific audience* • *Providing public access to a collection database on museum objects, archival and library information* • *Providing publications of both scientific a nature and for the more general public*"

Bulgaria

Institution	**Mission statement (or mission-like statement)**
Bulgarian Academy of Science - Central Library	*"... dedicated to the development of science in conformity with the universal human values and with the country's national interests and promotes the enhancement of the intellectual and material wealth of the Bulgarian people."*

Cyprus

Institution	**Mission statement (or mission-like statement)**
The Cyprus Institute - The Cyprus Research and Educational Foundation	*"... the advancement of knowledge and its humane and benevolent application and the establishment of a new research and educational public-benefit organization which shall generally promote research and education in Cyprus and abroad and shall aim primarily at benefitting the public interest at large."*

Czech Republic

Institution	**Mission statement (or mission-like statement)**
Institutu Umeni - Divadelniho Ustavu [Arts and Theatre Institute]	*"... to provide the Czech and international public with a comprehensive range of services in the field of theatre and individual services connected to other branches of the arts (music, literature, dance and visual arts). The ATI collects objects and work relating to the theatre, processes and provides access to them, pursues research, initiates and participates in international projects, and publishes scholarly work."*

Estonia

Institution	**Mission statement (or mission-like statement)**
Eesti Vabariigi Kultuuriministeerium [Estonian Ministry of Culture]	"*To support the maintaining of the Estonian national identity by valuing, preserving, developing, acknowledging and spreading Estonian fine arts, cultural heritage and sport in Estonia and abroad supporting both the professional and amateur activities in creativity and sport.*"

France

Institution	**Mission statement (or mission-like statement)**
Ministry of Culture and Communication (MCC)	"*Defines, coordinates and evaluates the state policy on architecture, archives, museums, monuments and archaeological sites.*"

Germany

Institution	Mission statement (or mission-like statement)
Philipps-Universität Marburg - Bildarchiv Foto Marburg	*"... to secure, maintain and hand down to future generations documentary photographs of significance to cultural history."*
Prussian Cultural Heritage Foundation	*"The preservation and care of the collections, their structure and development, and the continuation of academic and scientific research form the basis for a mediation of cultures with a mission to encourage learning and understanding between different peoples. The Foundation embodies the shared governmental responsibility for culture in Germany. The Federal Government and the sixteen individual states share the legal and financial responsibility, a living manifestation of constitutional reality."*

Greece

Institution	Mission statement (or mission-like statement)
Hellenic Ministry of Culture	*"...entrusted with the preservation of the country's cultural heritage, the arts, as well as sports..."*

Hungary

Institution	Mission statement (or mission-like statement)
National Szechenyi Library	*"... to collect, process and preserve all the written heritage of Hungary and all documents pertaining to it."*

Ireland

Institution	Mission statement (or mission-like statement)
An Chomhairle Leabharlanna	*"… provision of advice, assistance and services… making of … recommendations … promote and facilitate library co-operation."*

Italy

Institution	Mission statement (or mission-like statement)
Archivio Fo Rame	*"... everything related to the artistic, political and personal* [of] *Dario Fo and Franca Rame."*
Istituto Centrale per il Catalogo Unico	*"... cataloguing the entire national bibliographic heritage. ... improve the knowledge of bibliographic collections ... simplify user access."*

Latvia

Institution	**Mission statement (or mission-like statement)**
Valsts Aģentūra "Kultūras informācijas sistēmas" (State agency "Culture Information Systems")	"*To help memory institutions - archives, libraries and museums to preserve and make accessible cultural heritage for future generations, using advanced information technology solutions.*"

Russia

Institution	**Mission statement (or mission-like statement)**
University of Kazan	*"Guided by the ideals of scientific truth, creative liberty and good citizenship, the University contributes to the formation of the knowledge society and to sustainable development of individuals, the region, the country and the world."*

Slovenia

Institution	**Mission statement (or mission-like statement)**
Institute for the Protection of Cultural Heritage of Slovenia	*"The goal of the institute is to preserve and protect the cultural heritage of Slovenia, to raise the broader public's interest in cultural heritage, as well as to achieve a balance of cultural monuments of the past with the existing natural and cultural environment and new architectural achievements."*

Spain

Institution	**Mission statement (or mission-like statement)**
Generalitat de Catalunya - Departament de Cultura i Mitjans de Comunicació	*"gencat IS CULTURE."*

Sweden

Institution	**Mission statement (or mission-like statement)**
Riksarkivet (National Archives of Sweden)	"*To provide the public with the means of accessing public records, to secure information for judicial and administrative purposes, and to provide documentation for purposes of research.*"

United Kingdom

Institution	**Mission statement (or mission-like statement)**
Collections Trust	"*To improve the quality of life by ensuring that cultural collections are available for use and enjoyment by everyone, now and for the future.*"
Digital Heritage	*"... stimulating, and invigorating digital cultural heritage."*