INTERNATIONAL STANDARD

ISO 24619

First edition
2011-05-15

Language resource management — Persistent identification and sustainable access (PISA)

Gestion des ressources langagières — Identification et accès pérennes

Reference number
ISO 24619:2011(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 24619 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

Introduction

References and citations are an important part of documents and papers. Traditionally authors use them to provide proper acknowledgment to the author(s) of other papers as a source for their work or use them to support their argumentation. Citations usually contain information that enables a reader to establish the possible relevance of the cited paper and to identify it unambiguously. Any librarian or knowledgeable person is able to retrieve the document using well-established procedures based on the information in the citation.

The availability of directly accessible documents on the web has inspired the practice of adding a web location (URI [4]) to the citation information. This practice has made it possible to access referenced documents directly in web browsers as well as in other document viewers. This practice is already recommended in standards like ISO 690, although the emphasis there is more on identifying published resources and parts than on providing sustainable access to them. Increasingly often, such references need to be exploited by machines and software applications as well as by people, requiring reliable availability of the referenced resources. Problems with access that occur when resources are relocated have led to the use of persistent identifier (PID) frameworks [23], [24]. Current approaches [18], [19], [24] address the resource relocation problem by introducing resolver services that translate a resource identifier to its actual current location. These resolver services have an added advantage of permitting the association of additional metadata with the identifier. Elaborate frameworks such as the Digital Object Identifier (DOI) [14], use this feature to manage extra services, for instance copyright information.

The practice of using persistent identifiers to cite and reference scientific data, along with individual resources as well as data sets, is less well developed. It is no less powerful, however, in that it allows readers of a paper, or users of a knowledge resource, direct access to the primary scientific data to which the resource refers. When using references to access scientific data, including language resources, it becomes important to be able also to refer to and access parts of resources. This is especially true in the domain of language resources, where several layers of granularity are usually superimposed on the same data set or resource collection. Therefore, discussions in this International Standard concerning the use and requirements for PID frameworks extensively explore how these frameworks can deal efficiently with identifying and accessing parts of resources. Special recommendations indicate how to approach the granularity issue when issuing PIDs for resources and resource collections.

The need to apply PID frameworks for identifying resources contained in scientific data sets has also increased since modern archives and repositories have begun to weave a network of related complex resources that may be distributed over several locations. In these cases, permanent linkage is a prerequisite. In a multimedia lexicon for instance, a lexical item can refer to images not necessarily physically in the lexicon, or that are even referenced at a different site under control of a different organization. However, the link between the lexicon item and the image must remain valid, even if some servers or files are subject to relocation over time. Emerging e-Science scenarios, which make use of distributed services processing distributed resources, are also completely dependent on having transparent access from any processing service, irrespective of where it is located or what organization may operate it. This implies that resolving resource references should not be hampered in any way by unnecessary dependencies involving reliance on unsustainable or unpredictable services, whether they are technical or organizational.

The requirement that services like PID frameworks be accessible to the whole community of language resource and technology providers is further complicated by the need to provide resolvable PIDs without imposing commercial dependencies on resource providers other than the fundamental and well-established requirements for maintaining resources on the Internet.

Language resource management — Persistent identification and sustainable access (PISA)

1 Scope

This International Standard specifies requirements for the persistent identifier (PID) framework and for using PIDs as references and citations of language resources in documents as well as in language resources themselves. In this context, examples of language resources include such works as digital dictionaries, language-purposed terminological resources, machine-translation lexica, annotated multimedia/multimodal corpora, text corpora that have been annotated with, for example, morpho-syntactic information, and the like. Computational and applied linguists and information specialists create such resources.

This International Standard also addresses issues of persistence and granularity of references to resources, first by requiring that persistent references be implemented by using a PID framework and further by imposing requirements on any PID frameworks used for this purpose.

PID frameworks also allow the association of general metadata with the identifier, which can also contain citation information. This International Standard specifies minimum requirements for effective use of PIDs in language resources and cites the use of several possible existing standards and *de-facto* standards, such as: ISO 690 [16], APA [3], MLA [9] for citation information, ISO/IEC 21000-17, IETF RFC 5147, Annotea [2], temporal-fragment [22], XPointer for part identifier syntax and PURL [23], ARK [18], Handle System [24] and DOI [14].

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 12620:2009, *Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources*

ISO/IEC 21000-17:2006, *Information technology — Multimedia framework (MPEG-21) — Part 17: Fragment Identification of MPEG Resources*

W3C 2003, *XPointer Framework*: [online] W3C Recommendation 25 March 2003 [viewed 2010-08-04]. Available from: http://www.w3.org/TR/xptr-framework/

WILDE, E. and DUERST, M. *URI Fragment Identifiers for the text/plain Media Type*, IETF RFC 5147, April 2008 [viewed 2010-12-22]. Available from: http://www.rfc-editor.org/rfc/rfc5147.txt

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1 Resources

3.1.1

resource

digital object on the web with a specific identity that can be addressed with a **URI** (3.2.2)

NOTE 1 Adapted from IETF RFC 3986.

NOTE 2 In the context of this International Standard, a resource can also be a language resource that has an online representation.

NOTE 3 A resource can have several representations. Depending on the **PID framework** (3.2.5), identification of a specific representation can be encoded in the identifier (ARK, see B.3) or be left to the content negotiating process [8] between the **web client** (3.3.8) that uses the resolved PID to fetch the **resource** (3.1.1) and the **resource server** (3.3.6).

3.1.2

language resource

digital resource that provides information about one or more languages

NOTE Language resources cover lexicographical, terminological, morpho-syntactical, corpus-related, or semantic resources or digital resources used to study linguistic phenomena like texts and multimedia/multimodal recordings. They are created and used by linguists, information specialists, lexicographers and terminologists, among others. They frequently comprise many small records compiled within a larger work, and are often authoritative in nature, such as standardized terminologies and glossaries issued by standards bodies such as ISO, IETF, W3C, etc.

3.1.3

complex resource

resource (3.1.1) consisting of multiple constituent parts, each of which can be accessed individually

NOTE A complex resource can be a federated resource if its constituent parts are distributed over different **repositories** (3.1.6).

3.1.4

collection

grouping of any number of **resources** (3.1.1) that need to be referenced as a whole

3.1.5

published collection

purposefully built collection of resources that is maintained as an independent entity by an **archive** (3.1.7) or **repository** (3.1.6) and for which adequate **citation** (3.1.16) information is available

3.1.6

digital repository

repository

facility that provides reliable access to managed digital **resources** (3.1.1)

3.1.7

archive

digital archive

repository (3.1.6) dedicated to the long-term preservation of its associated data

NOTE Often the data in digital archives are also available online, which highlights the need for reliable **persistent identifiers** (3.2.4).

3.1.8
resource collection incarnation
incarnation
virtual embodiment of a disparate, otherwise non-aggregated **collection** (3.1.4) assembled for a specific purpose that is referenced by a single **PID** (3.2.4) concatenated with a **part identifier** (3.2.7) in order to access the components of the collection

NOTE A bibliography or index can use a single PID together with extensions to provide access to components in a set of **resources** (3.1.1) used in the production of a monograph or project without actually collecting the physical files in one location, which is to say that the individual items remain in their original locations, but are referenced as parts of a virtual whole.

3.1.9
version
particular form or variation of a **resource** (3.1.1) that differs from other instantiations of the resource in at least one aspect or item of information

NOTE Versions are often identified in sequential order (e.g. Version 1, 2, etc.), but version identification of dynamic resources subject to frequent change is often achieved by assigning a date-time stamp.

3.1.10
snapshot
instantaneous copy of a **resource** (3.1.1) representing the status of the resource or collection at a single point in time

3.1.11
abstract resource
non-network-retrievable resource identified by a **URI** (3.2.2), usually a concept such as a class or property

NOTE It is practice, for example in RDFS (RDF Schema) or OWL (web ontology language) ontologies, to identify abstract resources using URIs. Web architecture does not require any information resource to be retrievable with this kind of URI. If an identifier for an abstract resource is not meant to be **dereferenced** (3.4.1), such as can be the case with an XML namespace URI, it is not meaningful to issue a **PID** (3.2.4) for this resource.

3.1.12
resource part
part
identifiable, accessible entity embedded in an independent **resource** (3.1.1) or in a larger part thereof

NOTE Parts can be embedded in other parts. In dynamic web environments, subsetting into parts is subject to change and interpretation, which requires a certain level of user decision-making to designate and identify such sub-entities.

3.1.13
fragment
some portion or subset of a primary **resource** (3.1.1), some view on representations of the primary resource, or some other resource defined or described as a component of the resource defined or described by those representations

NOTE 1 Adapted from IETF RFC 3986.

NOTE 2 In this International Standard, the term *fragment* is used only in the IETF RFC 3986 sense, when in a web context a **client application** (3.3.5) retrieves the fragment from a containing resource.

3.1.14
terminal part
part (3.1.12) of a **resource** (3.1.1) that is not subdivided into smaller parts

3.1.15
internal part
part (3.1.12) of a **resource** (3.1.1) that is both embedded in the resource and subdivided into smaller parts

3.1.16
citation
information object containing information that directs a reader's or user's attention from one **resource** (3.1.1) to another

3.1.17
reference
digital object that links to data stored elsewhere

NOTE Although **citation** (3.1.16) and reference are commonly used as near-synonyms, for purposes of this International Standard, citations provide information for human readers and users, while references include the precise location where the referenced **resource** (3.1.1) can be found. References can be machine-readable, and can be configured as actionable given the required criteria.

3.1.18
annotation tier
separate information layer containing comments, notes, explanations, or other types of external remarks that can be attached to a **resource** (3.1.1)

NOTE For instance, maps or images can be annotated with supplemental information, or text corpora can be annotated in either in-line or standoff mode.

3.1.19
standoff annotation
annotations held outside the document that is being annotated

3.2 Identifiers

3.2.1
identifier
digital identifier
sequence of characters associated with digital, non-digital, or abstract entities, such as books, images, reports, metadata records or events

3.2.2
URI
Uniform Resource Identifier
string of characters used to identify or name a **resource** (3.1.1) with a syntax as defined in IETF RFC 3986

3.2.3
URI naming scheme
top level of the URI naming structure

NOTE 1 Every scheme specifies its own syntax conventions for **URIs** (3.2.2).

NOTE 2 Typical URI schemes include http, https, ftp, mailto, etc. and are registered with IANA.

3.2.4
PID
persistent identifier
unique **identifier** (3.2.1) that ensures permanent access for a digital object by providing access to it independently of its physical location or current ownership

NOTE Unique in this context means that the PID will not be issued again for other resources. However, the same PID can reference different representations or **incarnations** (3.1.8) of the resource at the discretion of the resource provider.

3.2.5
PID framework
scheme for specifying identifier strings [**PID** (3.2.4) scheme] for web-accessible digital objects together with a mechanism that enables the resolution of these identifiers into the object's current **URI** (3.1.1)

NOTE 1 A PID framework in the sense of this International Standard facilitates access to both individual objects and to **parts** (3.1.12) and **fragments** (3.1.13) contained in such objects. A PID framework can be solely dependent on existing web resolution protocols or it can entail the interaction of proxy-based resolvers.

NOTE 2 A PID framework in the sense of this International Standard also allows resolution of other information associated with the PID.

3.2.6
actionable identifier
URI (3.2.2) that has a resource-associated **identifier** (3.2.1) that is suitably encoded, such that when the URI is embedded in a web document and “clicked” on, the browser will be redirected to the **resource** (3.1.1), and possibly supplementary services related to the resource

NOTE 1 This functionality implies that the URI points to a suitable **resolver proxy** (3.3.7).

NOTE 2 In some **PID frameworks** (3.2.5), the **PIDs** (3.2.4) are URIs and are automatically actionable.

3.2.7
resource part identifier
part identifier
string of characters that refers to a **resource part** (3.1.12) that can be identified by some means within a given resource type (time in media, area in an image, record in a data stream, etc.)

NOTE Part identifiers in the sense of this International Standard are intended for server-side resolution in contrast to client-side resolution, which is characteristic of **fragment identifiers** (3.2.8).

3.2.8
fragment identifier
identifier (3.2.1) used to reference a **part** (3.1.12) of a **resource** (3.1.1) in a web context

NOTE 1 Adapted from IETF RFC 3986.

NOTE 2 A fragment identifier component as defined in IETF RFC 3986 is indicated by the presence of a number sign (“#”) character and terminated by the end of the **URI** (3.2.2). **Fragments** (3.1.13) in the sense of this RFC are resolved and retrieved from the resource by the local **client application** (3.3.5).

NOTE 3 There is a W3C draft proposal to change this handling of fragments [27].

3.3 Roles, institutions and services

3.3.1
archiving institution
institution responsible for maintaining a **digital archive** (3.1.7)

3.3.2
resource provider
organization that makes a **resource** (3.1.1) available online

NOTE A resource can also be a service.

3.3.3
resolver
PID resolver
software application that translates an **identifier** (3.2.1) into another more suitable identifier, specifically that translates a resource **PID** (3.2.4) into its **URI** (3.2.2) and in this way points a client application to the location of the **resource** (3.1.1)

3.3.4
resolution system
system designed to support the submission of a **persistent identifier** (3.2.4) to a network service in order to receive in return one or more pieces of current information related to the identified object, e.g. a location (**URI**) (3.2.2) of the object or metadata

NOTE The complete resolution system can be viewed as “the **PID resolver**” (3.3.3) but is often implemented as different resolvers or resolver services.

3.3.5
client application
software application that accesses a remote service usually on another computer system

3.3.6
resource server
computer that ultimately provides access to the object referenced by a specific client application request

3.3.7
resolver proxy
HTTP resolver proxy
application that implements a service supporting the use of **urlified** (3.4.3) **PIDs** (3.2.4) to access resources or other PID-related information, or both

3.3.8
web client
client application capable of accessing resources on the web using the HTTP protocol

3.4 Actions

3.4.1
dereference
to access the value referred to by a **reference** (3.1.17)

NOTE When used within the context of dereferencing a **URI** (3.2.4), it means obtaining a representation of the resource to which the URI points.

3.4.2
resolve
to translate an **identifier** (3.2.1) into another name or address suitable for accessing a resource

NOTE The resolution process may require multiple steps in order to obtain a suitable address for a resource.

3.4.3
urlify an identifier
to encode an **identifier** (3.2.1) as a suitable **URI** (3.2.4)

NOTE For example, this might be done with the purpose of creating an **actionable identifier** (3.2.6).

4 Background

PIDs can exist in all kinds of electronic resources and this International Standard does not make explicit statements about them, but the type of resource targeted by a PID has consequences for the requirements imposed on the individual PID. Resources can be characterized into three major types:

— independent resources as shown in Figure 1;

— any part of such an individual resource that requires further specification;

— a collection of resources that is referred to as a whole.

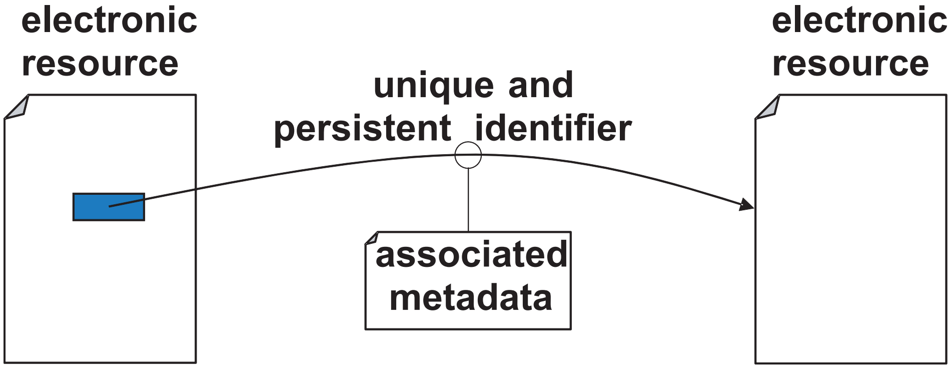


Figure 1 — Using unique PIDs to point from a source resource to a target resource

This International Standard concerns how to uniquely reference an electronic resource in a machine-readable way. In Figure 1, a unique and persistent identifier (PID) included in a source resource points to a target resource. The PID can be associated with metadata of different sorts.

The nature of a resource in this context is very broad and the means of referring to it is subject to context. An image, for instance, either can be an independent resource associated with its own unique PID and can be referenced as such, or can be embedded in a document where it lacks an identity of its own, in which case it is a part of that document. In addition, a reference can point to a part of this image. An individual resource can stand alone in one environment and be treated as part of a complex resource in another environment. An internal part of a resource may be viewed as a terminal part, but further processing in a dynamic environment may result in an entity that itself comes to contain accessible sub-parts. This International Standard is designed to support all these cases.

In the case of complex language resources, some resources should be assigned their own individual persistent identifiers. Other resources act as containing resources that have many constituent parts, in which case the containing resource should be assigned a PID, while its parts can be referenced by appending part identifiers to this PID. This International Standard provides guidelines for determining the appropriate approach to take with respect to any given resource.

This International Standard utilizes existing standards and practices for resource part and fragment identifier formats, where available, and provides guidelines for situations where current standards are inadequate or do not apply. A further discussion of resource types targeted by this International Standard may be found in Annex A.

With respect to collections of language resources, the standard takes two types of collections into account:

— Collections of resources that are maintained as complex resources in a more or less published static form so that the definition of the collection as such is maintained as an independent entity by an archive or repository, which then also provides a persistent identifier for such a collection. The archiving institution is responsible for maintaining the connection between the PID and the collection represented as a metadata entry in a catalogue, for example.

— A different type of collection that was not preconceived as a collection by its creators or the archiving institution(s) but achieves its status as a complex resource based on some research or other work that needs to be verifiable, such as the preparation of a monograph or the conduct of a scholarly or scientific project. Such collections, although purposefully constructed by the creator, may not have any significance outside the context of the original work for which they were created. Referring from the research documents to the collection may become tedious when the collection contains hundreds of individual resources. As a consequence, there is a need to refer to these types of collections with a PID that is associated with all its constituent resources and appropriate metadata. Of course this kind of reference is only possible if there is an incarnation of the collection.

5 Requirements for PID frameworks and PID use

5.1 General

Current standards and practices for using references and citations, especially in the domain of language resources, can be found in Annex A. This section focuses initially on requirements for the PID framework itself and thereafter on requirements for using PIDs as references and citations of language resources.

5.2 PID framework requirements

5.2.1 General

A PID framework in the sense of this International Standard shall support the following:

a) resolution of a single PID to multiple URIs or services;

b) association and access to related metadata;

c) adequate security to prevent malicious or accidental modification of PID/URI mappings and PID/metadata associations;

d) addressing of parts of a resource (part or fragment identifiers, or both);

e) encoding of the PID as a URI to render identifiers actionable in web documents without requiring client modifications.

5.2.2 Accommodating duplicate resources

It is common to provide duplicate or mirror resources or copies residing on different resource servers for data preservation purposes and to provide high-speed access. The PID framework should support this kind of duplication by allowing multiple URIs to be associated with a single PID.

5.2.3 Accessing resource metadata

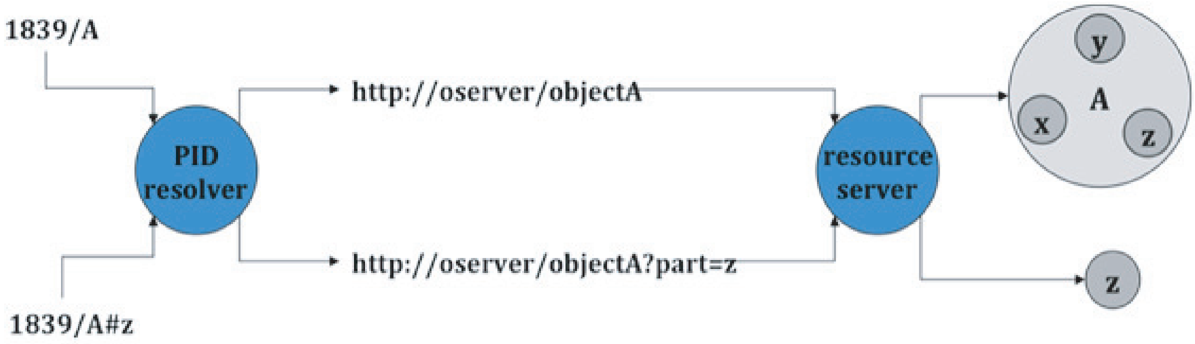
Next to providing a reliable URI to the resource, PID frameworks are also used to associate metadata with the resource in a secure and reliable way. Although this International Standard does not require metadata of any particular type to be available, it does require the possibility of resolving the PID to an associated metadata record encoded in XML format, so other services may be built on this feature.

5.2.4 Secure and reliable administration

The PID framework shall provide adequate security so that only the owner or caretaker of the resource can change the PID/URI mapping or the associated metadata.

5.2.5 Resource part identifiers

It is impossible to provide a PID for every identifiable part of a resource, or even to identify globally all possible options for segmenting resources into parts. Consequently, PID frameworks shall provide a system for assigning part or fragment identifiers in combination with the resource PID. Since the objective is to use a single string for identifying the resource part, PID syntax should support the concatenation of the PID and the part identifier. For example, the PID resolving system should be configurable such that a PID part identifier combination will resolve into a URI that can be correctly interpreted by the resource server in order to deliver the requested resource part [see A.2 a)].



A complex resource "A" with constituents x, y and z is identified by PID "1839/A".

The PID resolver translates the identifier "1839/A" into the URI http://oserver/objectA, which can be understood by the resource server to deliver object "A". The part "z" of resource "A" is identified by PID "1839/A#z". To enable easy dereferencing by the resource server, the PID resolver should be able to translate the identifier "1839/A#z" into the URI http://oserver/objectA?part=z or some similar query string that can be understood by the resource server for object A to deliver part "z".

Figure 2 — Processing part identifiers by the PID resolver (using the handle server as an implementation)

5.2.6 Urlified PIDs

The PID framework should provide a proxy resolver implementation that is able to resolve urlified PIDs. This allows web clients to resolve such an identifier using the HTTP scheme without needing special browser plug-ins or other special software. In some PID frameworks, the PID is already an HTTP URI, in which case a separate proxy resolver is superfluous.

5.3 PID usage

Citations of web-accessible digital language resources should be accompanied by a PID that resolves either to a URI for the resource itself or to a metadata record describing the resource. The latter option can be used if the resource itself cannot be made (immediately) available or in case of collections, in which case the metadata record should contain the identifiers for its constituents if available.

If the PID resolves to the URI of the resource, a metadata record in XML format and compliant with a declared schema pertaining to the resource should be associated separately with the identifier and made available via the resolving system. Citation information should be included in the metadata record. No other requirements are specified for the metadata record associated with the resource identifier.

In web documents the identifier (for instance, a handle) embedded in citations should in addition to a PID conforming to the PID scheme syntax also be presented in a urlified form that is encoded as a URI so that it becomes actionable in a web browser or other document viewer application. For example, using the Handle System (HS) as an implementation:

1839/00-0000-0000-0000-4 -> http://hdl.handle.net/1839/00-0000-0000-0000-0000-4

The resolver address http://hdl.handle.net identifies an HTTP resolver proxy, which is an application that can receive HTTP requests in the form of urlified PIDs and uses the "real" PID resolver to redirect the client to the resource. With the Handle System as an example, the HTTP resolver proxy is either the central resolver proxy of the Handle System or a Handle System resolver proxy guaranteed to be available by the resource provider.

Resource part specifications, which are analogous to page or chapter specifications, should be appended where relevant to the PID using an appropriate delimiter character, for instance using a scheduled extension of the HS:

1839/00-0000-0000-0000-4@time(100s,200s)

This extension refers to a segment of an audio file (the part identifier is non-standard).

If a compatible fragment identifier exists for the resource type, this element can be added to the encoded URI in order to create a composite actionable identifier. An HS example is:

http://handle.net/1839/00-0000-0000-0000-4?urlappend=#ffp(track_ID=101)*mp(/~time('npt','50'))

This example uses the special function “urlappend” of the handle proxy resolver to append a fragment identifier to the resolved handle.

5.4 Citation information and persistent identifiers

Existing practices can be maintained provided that URIs are replaced by PIDs and document part specifications are converted into part identifiers or fragment identifiers. References from web documents shall also provide a urlified PID if PID syntax does not comply with an IANA-registered URI scheme [13].

5.5 Referencing resource parts

5.5.1 General

Any applicable existing ISO or IETF standard for identifying a part of a resource can be used as a part identifier, unless otherwise specified. For resources where no such standard exists, it is permissible to use human-readable text in the citation, for instance: 10 s to 120 s for a time segment. This approach will not, however, work for software clients.

When using part or fragment identifiers for retrieving or dereferencing part of a resource, it is important to clarify the difference between using fragment identifiers such as those defined in IETF RFC 3986 and using the functionality of a suitable resource server. If the PID resolving process delivers a URI including a fragment identifier, for instance, using the following URI (with fragment specification according to the IETF RFC 5147 proposed standard for plain text media):

http://myserver/myresource#line=10,20

this URI will cause a web browser to fetch the whole document, and the browser itself will then isolate the document part from lines 10 to 20 and present it to the user. When the document is small, this is acceptable behaviour. However, when there is a need to present a fragment from a large 2 GB media file, it is necessary to use a special resource server and a URI such as with a part specification from Annodex [22]:

http://videoserver.com/videoA.anx?t=15.0/30.0

This notation will cause the resource server to transfer only the video segment from 15,0 s to 30,0 s.

Updates to this International Standard are expected to enlarge the list of applicable resource part/fragment identifier formats.

5.5.2 Media (time series)

ISO/IEC 21000-17 for MPEG-21 resources shall apply for media and time series. Annodex syntax can be used for specifying time intervals in URI queries and fragments [22] for all applicable media. For other formats, the part identifier will depend on the format used.

5.5.3 Textual resources

For XML-encoded textual resources, XPointer shall apply. IETF RFC 5147 shall apply for plain text documents. For other formats, the part identifier will depend on the format used.

5.5.4 Metadata registries, terminologies, ontologies

In accordance with ISO 12620:2009, each data category specification in the ISO/TC 37 Data Category Registry (DCR) shall have its own PID, which is formed as a concatenation of the PID for the DCR as a whole. These PIDs are configured as “cool” URIs [30] as per current practice of the World Wide Web Consortium.

Specific parts of an RDF graph can be addressed using one of the proposed RDF query languages, but no definitive specification is currently available. For other formats, the part identifier will depend on the format used.

5.6 Collections

Existing “published” collections shall be assigned an associated PID that is maintained by the collection sponsor. The PID shall refer to a description representing the collection, which can be a catalogue entry or an individual metadata description. For virtual collections, the PID should refer to a machine-readable metadata description which provides access to the relevant information, in particular to the resources that are included, referenced by their own individual PIDs.

6 Complementary requirements

6.1 Granularity of identifiers

With respect to granularity, this International Standard distinguishes between the identification of parts and of fragments, as indicated in 5.5. A fragment identifier is defined by IETF RFC 3986 as an optional component of a URI reference. In conformance with IETF RFC 3986, a URI can be assigned an optional fragment identifier, whereby the identifier is separated from the rest of the URI reference by a # (number sign) character. The separator is not considered part of the fragment identifier.

The URI with the fragment identifier can be used by an application to identify and usually to access a specific resource that is part of or embedded in a primary resource. The format of the fragment identifier depends on the resource type.

Interpreting and dereferencing the fragment identifier is a web client function, and therefore requires the complete primary resource to be downloaded, after which a client application can extract the required fragment. Furthermore, since the fragment identifier is not passed to other systems during the process of retrieval, some intermediaries in the web architecture (such as proxies) have no interaction with fragment identifiers, and HTTP redirection does not account for fragments. As an exception, fragment identifiers for RDF documents do not refer to parts of the document, but rather to the object in that document being described as having that fragment identifier. As a consequence, the use of fragment identifiers in combination with a URI allows the client application to isolate part of a resource based on knowledge specific to the client application.

In contrast to fragment resolution procedures, using a part identifier such as “z” in

http://myserver/myObjectService?part=z

relies on the remote server to isolate the part and send it to the client application[1].

[1] This strict separation of roles between a web client and server when dereferencing fragments may change. The media fragments working group has produced a W3C draft where it is proposed that web clients negotiate the transportation of only part of the resource from the server [27].

6.2 Recommendations

This International Standard supports different levels of granularity. The following recommendations are designed to encourage efficiency and promote interoperability with other naming schemes.

- If there is an existing identifier scheme for a type of resources, for instance, ISBN, this level of granularity should be retained, which is to say that no new PIDs should be issued without very good reasons, such as for chapters. Chapters would preferably be addressed using part identifiers in conjunction with the PID of the book.
- If the resource is associated with the complete content of a digital file, an individual PID should probably be assigned for this resource.
- If the resource is autonomous and exists outside a larger context, an individual PID should probably be assigned for this resource.
- If a resource should be citable apart from any containing resource, an individual PID should probably be assigned for this resource.

These recommendations are, however, subject to the needs of resource creators with respect to the level of granularity they deem suitable to the specific resource environment.

Annex A
(informative)

Independent resources, aggregated resources, and parts of resources

A.1 Overview

A.1.1 General

There is increasing demand in science and industry for options to reference digital language resources, resource parts or collections of language resources in an unambiguous persistent way. Not only is it desirable to be able to retrieve and validate references from scientific papers, but it is also increasingly necessary to maintain various types of references between language resources or parts of such resources.

A.1.2 Resources

A resource is anything that has an identity [7]. Despite the indefinite nature of this characterization, it is important for researchers to be able to identify linguistically meaningful units as coherent objects in a repository – both for human and, increasingly often, for machine readability. Such objects will be subject to separate manipulations and be used in different scenarios, which means that they can have an autonomous existence in larger contexts. Frequently such an object can be identified as a “single file” in a file system, but it can also be retrievable as a contained object (for instance, a data record) from a database system. Resources like this can have very different types or formats, such as

- a digitized video recording of an interview,
- a sound recording of a song,
- a complex annotation of a communication act,
- a photo documenting a speech event,
- a lexicon for a certain language,
- a grammar description,
- an eye tracking recording during a reading study,
- a metadata description of a resource or a resource collection,
- an integrated document containing texts and photos, etc.

The scope of such a resource is left to the discretion of its creator. Some resources are composed of a group of annotation tiers, while other originators may create a separate unit for each annotation tier, based on scientific and management considerations. Metadata for many resources may, for instance, be stored in a single large relational database. In such cases, a “metadata description” only exists as the result of a query. Such cases do not involve an identifiable resource, since the referenced resource is the entire database. This scenario indicates a need to identify options for referencing parts in such a resource.

When unique and persistent identifiers are associated with resources, it is necessary that the resources themselves also be persistent. The repository in question provides a resource with a unique identifier in order to ensure that when a reference is resolved, the “original content” is serviced (this discussion ignores user interface aspects that may change over time and offer the same content in a different way, although this may influence interpretation). Whenever the content of a resource is changed based on some manipulation, it now constitutes a new resource because it has taken on a new identity, which is separate from the previous version. In the case of dynamic databases that change their identities constantly in this way, they comprise snapshots that will be the units of reference. Any relatedness among such snapshots can be expressed in the metadata, for instance by qualified links, or in the citation information, or both.

A.1.3 Resource bundles

Sometimes repositories indicate that some of their resources are closely related due to a formal criterion. Examples include:

- sound and video recordings that were made at exactly the same moment for the same duration, which may include other recordings created in parallel, along with the annotations made concerning these recordings; all these resources share the same time and place and are therefore closely related, and users often want to refer to them as a whole;
- a lexicon that also includes photos or other media resources as extensions of lexical entries, in which case it may also be desirable to refer to closely related objects as a bundle;
- a text corpus and corresponding standoff annotation;
- a corpus of historical texts and corresponding facsimiles.

In general, these bundles should be configured as an incarnation of their grouped status, which is typically implemented by a joint metadata description that provides pointers to the constituent resources.

A.1.4 Resource collections

Collections are arbitrary groupings of any number of resources that are identified by some individual criteria, but are nonetheless referenced as a whole. Examples for such collections include:

- the Dutch National Spoken Corpus or the British National Corpus as published on a certain date;
- the collection of all resources describing a certain language that are stored in a given repository;
- the collection of otherwise disparate resources used as the basis for a dissertation study or monograph;
- a collection of the resources that have been bundled to store them as an archival package[2)] in a digital archive;
- collections of parts of a resource, for instance references to several segments of a media file, etc.

The underlying relationships of the identified resources can be arbitrary. Again, such collections have a formal status and it must be possible to reference their incarnation, which is in general achieved using a metadata description. Resource bundles as in A.1.3 are resource collections that share linguistically meaningful parameters.

A.1.5 Metadata descriptions

Each resource and resource collection should be associated with a metadata description as an incarnation of its existence. This is for performance optimization and long-term preservation reasons; it differentiates between the object as such and the many copies that can exist in different repositories. Metadata descriptions

[2)] The OAIS model discusses various types of packages that may be used to store collected resources as identifiable objects in an archive. It is the creator who decides on the content of the package based on some content or management criteria. The METS [20] standard, for example, supports the notion of “packages” in this sense.

not only stand for resources or collections, but also store additional information about them. If it is necessary to reference the metadata descriptions themselves, they must also be treated as objects just like other resources. Often, however, metadata records are dynamically generated as the result of a query on a complex database, which means that a metadata description is not a resource, but a part of a resource. This International Standard accounts for both scenarios, since PIDs can point to either resources or to various kinds of parts of resources. The resource server should be configured to generate the same information in either case.

In the case of collections and bundles, the metadata description itself references the included resources, which can be a recursive process. It is not the purpose of this International Standard to specify how a metadata description shall list the resources. It does, however, specify the mechanism that these metadata descriptions will use to reference the resources in the form of PIDs.

There is no single procedure for handling the versioning of metadata. Some repositories do not change the version number at all; some change it when added information changes; some also change the version number when a single object to which it refers is changed. This International Standard does not specify a single strict approach to versioning.

A.1.6 Granularity aspects

The distinction between a part of a resource, a resource itself and a resource collection has much to do with the level of granularity that an archiving institution is prepared to maintain with respect to a repository. Different approaches exist, which can reflect different needs and environments. As a consequence, this International Standard does not specify rules for granularity. It does elaborate guidelines for the granularity level of PIDs assigned to collections, resources and parts inasmuch as these issues are appropriate to performance, interoperability, efficiency and manageability.

Many projects give rise to questions concerning which objects are eligible to be issued their own individual PIDs and which can be handled by using a part identifier in combination with the PID of a “larger” containing resource. Performance and manageability are an issue here. Although PID frameworks are designed to handle many identifiers, every identifier still demands some maintenance and processing, so best practices suggest reserving individual PIDs for situations where they are actually useful. Coherence and interoperability are also relevant issues, such that objects of the same type should be treated at the same level of granularity as much as possible.

For some types of resources there are existing and proposed standards that indicate how to address resource parts on the web using a fragment identifier appended to the URI. As explained in Clause 6, this is not efficient for large resources; nevertheless, this option provides a powerful mechanism and, when available, it can be used to provide actionable resource part identifiers in documents.

A.2 Common practices for referencing resources on the Internet

The Internet architecture defines the Uniform Resource Identifier (URI) as a sequence of characters used to refer to resources on the Internet. The URI has many subclasses or schemes: many URI schemes specify an access mechanism and are therefore also (informally) called Uniform Resource Locators (URL [5]), because they specify a network location. These URIs are the most popular way to refer to resources, if only because, when embedded in a web document, the URI becomes immediately actionable. The URI syntax depends on the specific URI scheme used (http, ftp, file, gopher, etc.). The most widely used scheme, HTTP (IETF RFC 2616 [8]), supports the following features that are important for this International Standard:

a) The option to add parameters to the URI in a query in order to transfer extra information to the remote web server. For example: in the case of http://server/service?part=12, the parameters are interpreted by the remote resource server.

b) The option to add a fragment identifier to specify the part of the resource that is required. For example: in the case of http://server/document1#part1, the fragment identifier is not sent to the remote server; instead it is used by the client application to isolate the required document fragment from the complete transferred resource.

Too often the network location and local path information are embedded in the URIs of resources[3], which leads to problems when the location of resources is changed. The mixing of temporary location semantics into the identifier is harmful. For example, the protected resource

http://myhost.ourdomain/protected/R1.wav

may be unprotected tomorrow, and even the semantics of a resource name may change over time.

The URN URI scheme defined in IETF RFC 2141 [21], in contrast, provides names for resources instead of addressing them. URNs carry a namespace identifier from a list that is maintained by IANA [13], enabling integration without conflicts from other naming schemes (for example, ISBN, ISSN) as subsets of the URN. URNs should provide a globally unique persistent identifier usable for identification and access.

The syntax of the URN is as follows:

urn:<Namespace identifier (NID)>:<Namespace specific string (NSS)>

The namespace-specific string can take any form specified by the naming authority provided that it is unique within that namespace and avoids the use of a small number of restricted characters as specified in IETF RFC 2141. Although the URN is very suitable for naming resources, it has not been provided with a widely accepted way of resolving the name to the resource location, which has given rise to the different persistent identifier systems treated in Annex B.

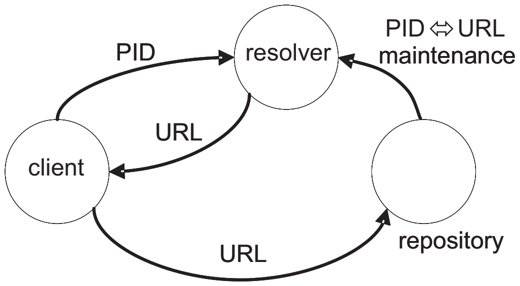
W3C is conscious of the problems of embedded location information in URIs and has published a TAG finding [25],[30] about this issue. The TAG finding, which contains important references such as the URIs document from Berners-Lee, claims that the construction of persistent URIs is possible and should be preferred to other solutions, such as PID frameworks.

A.3 Persistent identifier resolving systems

Persistent identifiers associated with a resolver service are designed to solve the common problem of broken links that occurs when resources on the web are moved to a new location or completely removed. Many approaches to persistent identification have been proposed in order to provide both consistent naming schemes for online resources and a resolver service to redirect users to the current location of a resource based on its persistent identifier.

The permanent aspect of a PID means that it will never be reassigned to any other resource and will not change regardless of where the resource is located or whatever protocol is used to access it. There are many formal identifier or naming schemes which have been discussed in the context of the naming of digital resources (e.g. URIs, URNs, handles, DOIs, ARKs, ISBNs, ISSNs, SICIs, BICIs, PIIs), although very few of these schemes will be fully effective in facilitating access to online resources in a distributed system unless they are either registered as URI naming schemes and supported by a resolution system or they are incorporated into another naming scheme which has some form of resolution system associated with it.

[3] This can be considered bad practice since most http servers allow for management of the resource URI separate from the resource location. In reality, however, it is very difficult to maintain location independence unless one uses redirection mechanisms with a similar administrative load as for instance the PURL system treated in B.1.



A client application needs to access a resource for which it knows the PID. It queries the resolver for the PID associated with the URL, which it uses to access the resource. The resolver's PID ⇔ URL mappings have to be maintained (by the archiving institution associated with the repository).

Figure A.1 — PID resolution

Without a resolution system, requests using the identifier cannot be routed to the appropriate server and used to retrieve the resource, or a reasonable substitute for it in the form of metadata. (See Figure A.1.) Short introductions to the most widely used PID resolving systems can be found in Annex B.

A.4 Referencing resource parts

A.4.1 General

This section describes current practices and standards for referencing parts of resources. Some standards and practices are used in existing citation conventions and in the web architecture. For the purpose of this International Standard, machine interpretable identifiers that unambiguously identify the required part or segment within the containing resource are the primary concern. Such part identifiers can trigger remote resource servers to deliver a representation of the required part to the client application or they can cause the client application itself to isolate the part from the resource. For any resource types where no current standard or practice recommendations exist, amendments can be made to the current standards and recommendations.

A.4.2 Media (time series)

Currently the following practices and standards exist for media:

ISO/IEC 21000-17, *Fragment Identification for MPEG Resources* specifies a normative syntax for URI Fragment Identifiers to be used for addressing parts of any resource whose Internet Media Type is one of the following: audio/mpeg, video/mpeg, video/mp4, audio/mp4, application/mp4, video/MPEG4-visual, application/mp21.

The Annodex digital media format developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) also includes a specification for temporal addressing through a special URI syntax [22]. This URI syntax allows for referencing segments in audio and video files. The URI syntax is meant to be interpreted by the remote server, which isolates the segment from the media file and sends it to the client. An example is:

http://example.com/video.anx?t=15.2/18.7

which will cause a segment of *video.anx* starting from 15,2 s and ending at 18,7 s to be transferred.

A.4.3 Textual resources

In the case of textual resources, individual parts can be referenced by part identifiers as long as they are qualified by the resource provider. In addition, if a text corpus resource consists of previously published texts and corresponding metadata are available, these texts and their parts can, if appropriate, also be referenced using customary conventions for indirect citation.

Practices with respect to referencing text file fragments include the following:

- the use of byte or character offsets to point to positions or “spans” in text. Relevant technology: Tipster [26], GATE [10] and all kinds of GATE derivatives;
- IETF RFC 5147, *URI Fragment Identifiers for the text/plain Media Type*, which is based on line or character offsets, or both;
- for XML and HTML documents, XPointer. For example, Annotea [2] is a project sponsored by W3C to enhance document-based collaboration through shared document metadata based on tags, bookmarks, and other annotations.

A.4.4 Knowledge sources

Knowledge sources include lexica, terminologies, concept registries, ontologies, etc. This area is the subject of dynamic developments, also with respect to representation formats. It is not the purpose of this section to provide an exhaustive picture, but rather to indicate some of the problems involved. It is also not the purpose of this International Standard to suggest specific solutions. It is rather the responsibility of the respective sub-communities to decide how to refer to parts of their respective resources and how they want to ensure the persistence of these references. Communities may, for instance, choose to abide by existing practices, such as to use PURLs, as in the case of Dublin Core.

For lexica, ISO 24613:2008 [15], *LMF (Lexical Markup Framework)* provides a metamodel for representing data in lexical databases. These data need to be referenceable at various levels by external resources or tools allowing references to individual lexical entries, senses or even more fine-grained information, such as the morphological division of a word stored in a lexicon. When lexica are encoded in an XML-based format, XPointer can be used to describe these references. Lexica themselves may also contain references to external resources, such as sound or video resources to exemplify pronunciation or example usage. These references typically refer not only to the entire resource, but rather to a part of it, such as a video fragment. Here, the resource identifier should be amended with a resource-specific part identifier designed to access the relevant information fragment. Finally, LMF lexica, as all other models within the ISO/TC 37 family of standards, reference data categories from the Data Category Registry (DCR) specified by ISO 12620.

Tools such as LEXUS [17] (which implements LMF), work in an environment with multimedia lexica where resources of different types become interwoven with each other. In such cases, a coherent referencing mechanism is necessary to guarantee interoperability.

It should be possible to reference individual entries in terminologies and concept-oriented registries, such as the ISO DCR. A basic level of interoperability can be achieved by referencing entries that are deemed to be identical and declaring them as such.

While individual resources can be referenced using a uniform mechanism such as PIDs, part identifiers are strongly dependent upon the type of resource being referenced. While standards that address this issue appear to be emerging for some resource types, for others the situation remains largely unclear, as the following example illustrates.

EXAMPLE Increasingly often in knowledge sources that include relations in addition to concept definitions, such as in ontologies, RDF is used as a representation format. A part identifier in these cases would be used to refer to a graph or sub-graph. Different solutions are being discussed, including RDF queries formulated in an RDF query language [12].

A.5 Referencing resource collections

Referencing collections of resources that belong to the class of “published collections” is widely known. Yet there is no clear common practice and the mechanisms used are not always machine-readable or interpretable. Generally users refer to publications about a given corpus as follows:

> John W. Du Bois, Santa Barbara Corpus of Spoken American English. Parts 1, 2 and 3, Linguistic Data Consortium, University of Pennsylvania, Philadelphia (2000).

or they refer to the origin in some form such as:

> Santa Barbara Corpus of Spoken American English. Part I. 2000. Collected by University of California, Santa Barbara Center of the Study of Discourse, directed by John W. Du Bois.

These references are feasible in cases where there is a clear responsibility for the creation process in terms of creators, creating institutions, archiving institutions, etc. Even in these cases, it is desirable to immediately access a machine-readable PID that can be exploited, for example, to automatically extract the size of a corpus from a formal metadata description, or to access other information.

No standard practice exists as yet for virtual collections that are created by researchers to carry out a research project and that can include a large number of resources created by different teams at different institutions, although there is a need to document the work so that other researchers can test the validity of the statements made. The reason for this lack may be that only in a few cases, such as in IMDI-based repositories [28], can researchers create metadata descriptions for these kinds of virtual collections which function as their incarnations, for example to initiate search operations.

A.6 Citation information

A.6.1 General

It is necessary to distinguish between different referencing contexts. In some contexts, such as an image in a lexicon, issuers are not usually interested in seeing information about the image in addition to what they have already found in the lexical attributes. They will only be interested in activating the reference and seeing the image itself immediately. In other contexts that are designed for a human to read a reference, it may be more appropriate to present the user with the complete citation information as one might see it in traditional print environments. In the case of virtual collections, the user may want to see more details about the resources included and therefore may reference the metadata description. In the first and last case the references are just the PIDs with a fragment indication attached, if necessary. The second case requires a closer look at the citation information.

In scholarly works, citations acknowledge and give credit to previous works. Readers are able to put claims to a test by consulting these works. The same holds true for primary scientific data on which authors base their research and where they want to give proper acknowledgment to the original creator or collector. So when citing electronic resources in a document, a reference or link to that resource is accompanied by citation information: human-readable text (metadata) that identifies the resource stake-holders (creator, collector, publisher, etc.).

More than one standard and many practices exist for specifying this citation information, and it would not be practical to list them all. Some widely used applicable ones in this context are given in A.6.2 to A.6.6.

A.6.2 ISO 690:2010

ISO 690:2010 [16] provides an overview of many electronic document types and specifies the elements to be included in bibliographic references. It is an important inspiration source for many citation practices. ISO 690 shows examples of using DOIs but allows other types of PIDs to be used. ISO 690 makes no requirements about PID frameworks; this International Standard does. Other PID frameworks allow, for example, organizations more control over the association of extra information with the PID (e.g. metadata) and provide other business models than DOI. All recommendations given in ISO 690 can be taken over with the provision of allowing PIDs from all frameworks that fulfil the requirements stated in this International Standard.

The following special elements are important for online resources:

- the element “Availability and access”, as in: “Available from: URL” and “Also available in PDF from: URL”;
- the element “Standard number” as in: “ISBN 0-7710-1932-7” or “ISSN 1045-1064”;
- the element(s) “Numeration and pagination” as in: “Accession number 01209277”, “pp. 5-21” or “lines 30-40”.

The value of the "Availability and access" element is usually a URL.

Using ISO 690 in the context of this International Standard would require replacing the URL by either the PID itself, while making the resolver framework type unambiguously clear, or using the urlified identifier. However, unless the PID syntax is officially accepted by W3C as a bona-fide manner to identify web resources, the urlified version should also be present. For example, again using the Handle System as a possible implementation:

Available from: hdl:4263537/4069, http://hdl.handle.net/4263537/4086.

or

Available from: http://hdl.handle.net/4263537/4086.

For the "Standard number" element, it is possible to use the PID itself or, if a suitable URN namespace is available, to use the complete qualified URN as in: urn:doi:10.1392/BC1.0. (The request for this URN name space is pending.)

The elements "Numeration and pagination" can be given the value of a part identifier or fragment identifier.

A.6.3 APA Style Guide

American Psychological Association (APA) style specifies many manuscript and documentation features as well as the organization of citations and references. The APA Style Guide to Electronic References [3] defines two appropriate elements for online resources: "retrieved" and "from". The intent of the first is to specify the date of retrieval when the referenced material was "used" and that of the second to specify the source (URL) of the document and may be preceded by a description of the URL.

EXAMPLE

Rogers, B. (2078). Faster-than-light travel: What we've learned in the first twenty years. Retrieved August 24, 2079, from Mars University, Institute for Martian Studies web site, http://www.eg.spacecentraltoday.mars/university/dept.html

The APA Style Guide explicitly states that if a persistent identifier such as a DOI is available, the PID should be used as the source instead of a URL.

A.6.4 MLA Handbook

The MLA Handbook [9], published by the Modern Language Association of America, is an academic style guide that provides guidelines for writing and documentation of research in the humanities.

In the 7th edition of the MLA Handbook [9], the use of URLs for on-line resources is optional. However, if it would be required or the resource could not be otherwise located, a URL can be added to the citation after author, title(s), publisher and date of access.

EXAMPLE For instance, a specific web page:

Cornell University Library. "Introduction to Research." *Cornell University Library*. Cornell University, 2009. Web. 19 June 2009 <http://www.library.cornell.edu/resrch/intro>.

Replacing the URL with a urlified persistent identifier is consistent with the MLA style.

A.6.5 STD-DOI and DataCite

Publication and Citation of Scientific Primary Data (STD-DOI) [6] was a project funded by the German Science Foundation. Its aim was to make primary scientific data citable as publications. In this system, a data set was attributed to its investigators as authors, as would be the practice for a work cited in the conventional scientific literature. Thus, scientific primary data would not be exclusively understood as part of a scientific publication, but could have its own identity.

EXAMPLE

> Kamm, H; Machon, L; Donner, S (2004): Gas Chromatography (KTB Field Lab), GFZ Potsdam.
> doi:10.1594/GFZ/ICDP/KTB/ktb-geoch-gaschr-p

The STD-DOI project has now been replaced by DataCite (http://www.datacite.org), a consortium of mostly technical libraries and information centres that pursues the same goals as STD-DOI.

A.6.6 A Proposed Standard for the Scholarly Citation of Quantitative Data

A recent proposal [1] for a standard for citing quantitative data from the social-sciences domain requires the use of a persistent identifier in citations. The proposal calls for the use of at a minimum six required (metadata) citation components. The first three are traditional (author, date and title), followed by the persistent identifier, a universal numeric fingerprint that allows for verification that the data have not changed and a so-called bridge service, which allows the urlified PID to have an actionable identifier.

EXAMPLE

> Micah Altman; Karin MacDonald; Michael P. McDonald, 2005, “Replication data for: From Crayons to Computers: The Evolution of Computer Use in Redistricting”, hdl:1902.1/AMXGCNKCLU UNF:3:J0PkMygLPflyT1E/8xO/EA==
> http://id.thedata.org/hdl%3A1902.1%2FAMXGCNKCLU

Annex B
(informative)

Persistent identifier system implementations

B.1 Persistent URL (PURL)

A PURL (Persistent URL) [23] is a Uniform Resource Locator (URL) (i.e. location-based Uniform Resource Identifier or URI) that does not directly describe the location of the resource to be retrieved, but instead describes an intermediate (more persistent) location which, when retrieved, results in redirection to the current location of the final resource. This is a standard Hypertext Transfer Protocol (HTTP) redirect, thus no acceptance of new protocols or modifications to client software are required.

PURLs were developed by the Online Computer Library Centre (OCLC) in the mid-1990s, primarily to reduce the maintenance burden of the URLs contained in catalogue records created for Internet resources. OCLC was an active participant in the IETF working groups on URNs and was fully aware of how far the groups were from consensus on a standard solution for persistent identification. They therefore developed PURLs as an interim solution to address the lack of progress in persistent naming for Internet resources.

The PURL framework only allows the association of one single location with the identifier and provides for no additional metadata information. The resolver and management software are publicly available.

B.2 Handle System (HS)

The Handle System (HS) [24] is a distributed persistent naming system developed for digital library applications. It was developed by the Corporation for National Research Initiatives (CNRI) and had its origin in a computer science technical reports project, Networked Computer Science Technical Reports Library (NCSTRL), funded by the Defense Advanced Research Projects Agency (DARPA) in the US. Part of this project was to develop an architecture for the underlying infrastructure of an open distributed digital library.

The Handle System is well known in the digital library world. It provides a PID syntax specification and also a resolving system implementation. The syntax of a PID or handle in the Handle System is very simple:

<prefix>/<suffix>   example: 15.12345/abcd6789

The Handle System top authority assigns the prefix on request to an institution or organization and will be able to resolve any such handle, which means that its Global Handle Registry will be able to identify the prefixes, the prefixes will point to local handle services and these will know how to interpret the suffixes. The specific syntax of the suffixes is left to the local handle system owner as long as the syntax complies with URI specifications. The following examples are taken from the DOI webpages:

10.1000/123456, 10.1000/ISBN1-900512-44-0

10.2345/S1384107697000225

10.4567/0361-9230(1997)42:<OaEoSR>2.0.TX;2-B

The Handle System is more than a simple naming scheme; it is supported by a resolution system consisting of a distributed system of global, local, and caching servers. A Global Handle Registry maintained by the Corporation for National Research Initiatives (CNRI) registers the top-level naming authorities, both to ensure the uniqueness of the names and to route requests for handle resolution. This procedure is unique among handle services only in that it provides the service used to manage naming authorities, all of which are managed as handles. The naming authority handle is a special handle that provides information to be used by clients to access and utilize the local handle service for handles maintained by the naming authority in question. Local handle services are operated by organizations. They resolve the requests routed to them and return the current address(es) or other information about the resource sought. They therefore hold handles that provide information about resources as registered by a respective naming authority. A local handle service can itself be composed of a number of servers. Finally, caching servers associated with local servers allow frequently accessed handles to be resolved without the need to request the address from the Global Registry.

Although the HS does not define a URN namespace, it can be used to implement resolving functionality for a URN schema such as is requested for DOI [14].

The HS software is publicly available and can be downloaded from the CNRI Handle site. CNRI makes available local service software, client software and simple management tools, a caching handle server, tools for the creation and administration of handles and naming authorities, and a proxy server to enable web clients to resolve handles. To enable web browsers to resolve handles without the use of a proxy, CNRI has developed a Handle Resolver plug-in, which is available for download.

The HS allows the association of a handle with multiple records of type URL, specifically, UTF8-encoded URIs that specify the location of the object identified by a handle. The HS also supports user-defined data types, which can be used to associate metadata with the resource. The HS, although it can make use of HTTP, for instance by using resolver proxies, does not intrinsically depend on the HTTP protocol, but defines its own.

Currently, the HS does not support direct use of part identifiers or fragments. At the moment the only possibility is to encode a part identifier as part of the urlified handle, which, of course, makes it dependent on the HTTP protocol.

B.3 Archival Resource Key (ARK)

One of the most recently proposed identifier schemes is ARK (Archival Resource Key), which has been developed by John Kunze in his work for the US NLM (National Library of Medicine). The ARK proposal is still an Internet draft, of which the latest version was issued in July 2008 [19]. It is currently being tested and implemented by the California Digital Library (CDL) for collections that it manages. The ARK is a scheme intended to facilitate the persistent naming and retrieval of information objects and is being developed specifically to meet the needs of those who maintain archival digital objects.

The ARK system has a number of primary features.

- An ARK identifier is associated with three services:
  - providing a link to the resource;
  - providing a link to the resource's metadata; and
  - providing a link the resource provider's promise about its persistence.
- Its naming scheme is constrained to discourage semantics in the identifier.
- It accommodates resource part identifiers and possible different resource representations.
- It does not have to be urlified to become actionable, which means an ARK can be a URI.

An ARK has five components:

[http://NMAH/]ark:/NAAN/Name[Qualifier]

The ARK consists of an optional and mutable Name Mapping Authority Hostport (NMAH), the “ark:” label, the Name Assigning Authority Number (NAAN), the assigned Name, and an optional and possibly mutable Qualifier supported by the NMA. The NAAN and Name together form the immutable persistent identifier for the object. The Qualifier can be used to refer to the object parts or to a specific representation, or both. The NMAH points to the resolver service.

Although there are ARK resolvers in operation, there is currently no publicly available resolver software for the ARK system. Some reports describe it as work in progress.

B.4 Additional PID systems

PURL, the Handle System and ARK system are by no means an exhaustive list of persistent identifier frameworks, but they are a good representation of existing approaches.

One recent (2005) PID scheme and resolving protocol is the eXtensible Resource Identifier (XRI) [29], which was created by the OASIS industry consortium. In contrast to the ARK system, XRI encourages, for instance, the use of semantics in the identifier by constructing it as a series of self-describing tags. It also uses special global context symbols in the identifier to indicate persistency and semantic context of the identifier parts. Furthermore, it allows for the merging of identifiers from other persistent identifier schemes within an XRI identifier.

W3C itself does not encourage the use of any PID resolving system other than HTTP redirects such as used in PURL. There is a W3C TAG [25] draft indicating the W3C point of view, that careful use of URIs (avoiding the use of physical path information) will at least make the need to use PIDs in order to avoid broken links superfluous.

Annex C
(informative)

Abbreviated terms

See Table C.1.

Table C.1 — Abbreviated terms

APA style	American Psychological Association	Style guide for citations
ARK	Archival Resource Key	A PID framework
BICI	Book Item and Component Identifier	Unique identifier for components within an ISBN-referenced publication
DARPA	Defense Advanced Research Projects Agency	Military research agency of the USA
DataCite	Publication and Citation of Scientific Primary Data	Citation mechanism for primary scientific data
DCR	Data Category Registry	Formally defined set of linguistic categories for reference and use in annotated resources
DNS	Domain Name System	
DOI	Digital Object Identifier	Persistent identifier of an on-line object; A PID framework built on the HS
FTP	File Transfer Protocol	Protocol for file transmission over TCP/IP
GATE	General Architecture for Text Engineering	Language technology toolkit
GOPHER	[not an acronym]	Predecessor of the HTTP protocol
HTTP	Hypertext Transfer Protocol	Internet transmission protocol for HTML pages
HS	Handle System	A PID framework
IANA	Internet Assigned Numbers Authority	Entity that manages IP domains and DNS zones
IETF	Internet Engineering Task Force	Entity that develops Internet standards
IMDI	ISLE Metadata Initiative	XML-based metadata standard for language resources
IP	Internet Protocol	
ISBN	International Standard Book Number	Unique ten-digit number used to identify books as per ISO 2108[31]
ISSN	International Standard Serial Number	Unique eight-digit number used to identify a periodical publication as per ISO 3297[32]
LEXUS	[not an acronym]	Web-based lexicon tool
LMF	Lexical Markup Framework	Standardized framework for electronic lexica
METS	Metadata Encoding and Transmission Standard	XML-based metadata standard for digital libraries
MLA Style Guide	Modern Language Association of America	A style guide for citations
MPEG	Moving Picture Experts Group	ISO/IEC JTC1/SC29 WG11, which develops audio and video encoding standards
NAAN	Name Assigning Authority Number	Organization Identification within the ARK PID system

Table C.1 (*continued*)

NCSTRL	Networked Computer Science Technical Reference Library	Distributed network of Computer Science technical reports in Cornell University's Computer Science Department
NID	Namespace Identifier	Unique repository identification code
NMA	Network Management Application	Software that forms a part of the Open Systems Interconnection reference model
OAIS	Open Archival Information System	ISO reference model for long-term data preservation
OCLC	Online Computer Library Center	Computer library service and research organization responsible for Dublin Core
OWL	Web Ontology Language	W3C family of RDF knowledge representation languages for authoring ontologies
PDF	Portable Document Format	Open standard for electronic documents
PID	Persistent Identifier	Sustainable reference to an online resource
PII	Publisher Item Identifier	Unique identifier used by a number of scientific journal publishers
PURL	Persistent URL	Sustainable reference to a URL
RDF	Resource Description Framework	Description language for the Semantic Web
RFC	Request For Comments	IETF standard proposal intended for peer review
SICI	Serial Item and Contribution Identifier	Unique identifier for specific volumes, articles or other identifiable parts of a periodical
STD-DOI	Publication and Citation of Scientific Primary Data	Citation mechanism for primary scientific data
TBX	TermBase eXchange	Standard for the exchange of translation memories
TCP	Transmission Control Protocol	
TMF	Terminological Markup Framework	ISO 16642:2003[33]
URI	Uniform Resource Identifier	A compact string of characters used to identify or name a resource
URL	Uniform Resource Locator	URI that refers to the location of an online resource
URN	Uniform Resource Name	URI that names an online resource
XML	Extensible Markup Language	Formal meta language used for the creation of custom languages intended for the sharing of structured data
XRI	eXtensible Resource Identifier	Scheme and resolution protocol for abstract identifiers

Bibliography

[1] Altman, M. and King, G. A Proposed Standard for the Scholarly Citation of Quantitative Data [online], *D-Lib Magazine*, **13**(3/4), 2007, ISSN 1082-9873 [viewed 2010-08-04]. Available from: http://www.dlib.org/dlib/march07/altman/03altman.html

[2] Annotea wikipage [online] [viewed 2010-08-02]. Available from: http://en.wikipedia.org/wiki/Annotea

[3] APA. *Style Guide to Electronic References*, June 2007, ISBN: 1-4338-0309-7

[4] Berners-Lee, T., *et al*. *Uniform Resource Identifier (URI): Generic Syntax*, IETF RFC 3986, January 2005

[5] Berners-Lee, T., Masinter, L. and McCahill, M. *Uniform Resource Locators*, IETF RFC 1738, December 1994

[6] Brase, J. *Using Digital Library Techniques — Registration of Scientific Primary Data* [online]. Lecture Notes in Computer Science 3232, 488-494, 2004, ISSN 0302-9743

[7] Dublin Core Metadata Initiative (DCMI), *Terminology* [online] [viewed 2010-08-04]. Available from: http://www.ukoln.ac.uk/metadata/dcmi/abstract-model/2004-12-08/#sect-7

[8] Fielding. R., *et al*. *Hypertext Transfer Protocol — HTTP/1.1*, IETF RFC 2616, June 1999

[9] Modern Language Association of America. *MLA Handbook for Writers of Research Papers*. New York: MLA, 7th ed. New York: MLA, 2009

[10] GATE. [viewed 2010-08-04], Available from: http://www.gate.ac.uk

[11] González, R. and Suarez Araújo, C.P., eds. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Paris: European Language Resource Association. pp. 1321-1326, 2002

[12] Haase, P., Broekstra, J., Eberhart, A. and Volz, R. A comparison of RDF query languages. *Proceedings of the Third International Semantic Web Conference*, Hiroshima, Japan, 2004

[13] IANA. *Approved URI schemes* [online], [viewed 2010-08-04]. Available from: http://www.iana.org/assignments/uri-schemes.html

[14] International DOI Foundation, *The Digital Object Identifier (DOI) System* [online], February 2001 [viewed 2010-08-04], Available from: http://dx.doi.org/10.1000/203

[15] ISO 24613:2008, *Language resource management — Lexical markup framework (LMF)*

[16] ISO 690:2010, *Information and documentation — Guidelines for bibliographic references and citations to information resources*

[17] Kemps-Snijders, M., Nederhof, M.-J. and Wittenburg, P. LEXUS, a web based tool for manipulating lexical resources. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC2006) (pp. 1862–1865) [CD–ROM]

[18] Kunze, J. Towards Electronic Persistence Using ARK Identifiers. *Proceedings of the 3rd ECDL Workshop on Web Archives* [online], August 2003 (PDF) [viewed 2010-08-04]. Available from: http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=kunze

[19] Kunze, J. and Rodgers, R.P.C. *ARK Persistent Identifier Scheme* (Internet Draft [online], updated: May 2008) [viewed 2010-08-04]. Available from: http://tools.ietf.org/id/draft-kunze-ark-15.txt

[20] METS. *Metadata Encoding and Transmission Standard* [online] [viewed 2010-08-04]. Available from: http://www.loc.gov/standards/mets/

[21] MOATS, R. *URN Syntax*, IETF RFC 2141, May 1997

[22] PFEIFFER, S., *et al*. *Specifying time intervals in URI queries and fragments of time-based Web resources* [online], 2005 [viewed 2010-08-04]. Available from: http://www.annodex.net/TR/draft-pfeiffer-temporal-fragments-03.html

[23] SHAFER, K., *et al. Introduction to Persistent Uniform Resource Locators (PURL)* [online], 1996, [viewed 2010-08-04]. Available from: http://www.isoc.org/inet96/proceedings/a4/a4_1.htm

[24] SUN, S., LANNOM, L. and BOESCH, B. *Handle System Overview*. IETF RFC 3650, November 2003

[25] THOMPSON, H. and ORCHARD, D. *URNs, Namespaces and Registries*, W3C TAG Editor draft [online], 2006-08-17 [viewed 2010-08-04]. Available from: http://www.w3.org/2001/tag/doc/URNsAndRegistries-50

[26] Tipster Text Program [online] [viewed 2010-08-04]. Available from: http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/

[27] TRONCY, R., *et al. Use cases and requirements for Media Fragments* [online], *W3C Working Draft*, 2009-04-30 [viewed 2010-08-04]. Available from: http://www.w3.org/TR/2009/WD-media-frags-reqs-20090430/

[28] WITTENBURG, P., PETERS, W. and BROEDER, D. Metadata Proposals for Corpora and Lexica. In: R. GONZÁLEZ and C.P. SUAREZ ARAUJO, eds. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Paris: European Language Resource Association, 2002, pp. 1321-1326

[29] XRI. *Extensible Resource Identifier* [online] [viewed 2010-08-04]. Available from: http://www.oasis-open.org/committees/xri

[30] W3C. *Cool URIs for the Semantic Web*: W3C Interest Group Note 31 [online] March 2008. SAUERMANN, L. and CYGANIAK, R., eds., 2008 [viewed 2010-08-04]. Available from: http://www.w3.org/TR/cooluris/

[31] ISO 2108, *Information and documentation — International standard book number (ISBN)*

[32] ISO 3297, *Information and documentation — International standard serial number (ISSN)*

[33] ISO 16642:2003, *Computer applications in terminology — Terminological markup framework*

Alphabetical Index

ICS 01.140.20

Price based on 29 pages