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)

COPYRIGHT PROTECTED DOCUMENT

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)