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