INTERNATIONAL STANDARD

ISO 639

Second edition
2023-11

Code for individual languages and language groups

Code pour les langues individuelles et les groupes de langues

STANDARDSISO.COM : Click to view the full PDF of ISO 639:2023

COPYRIGHT PROTECTED DOCUMENT

Contents

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared jointly by Technical Committee ISO/TC 37, *Language and terminology*, Subcommittee SC 2, *Terminology workflow and language coding*, and ISO/TC 46, *Information and documentation*, Subcommittee SC 4, *Technical interoperability*.

This second edition cancels and replaces ISO 639-1:2002, ISO 639-2:1998, ISO 639-3:2007, ISO 639-4:2010 and ISO 639-5:2008, which have been merged and technically revised.

The main changes are as follows:

— the wording of the text (with provisions, such as principles, rules and guidelines) has been harmonized to apply to all sets of language identifiers of the ISO 639 language code;

— the ISO 639 language code subsets contained in the former ISO 639-1, ISO 639-2, ISO 639-3 and ISO 639-5 are now represented by sets of language identifiers: Set 1, Set 2, Set 3 and Set 5, respectively, and are not presented in this document;

— new developments in the field of coding systems have been taken into account;

— the title has been modified;

— the wordings in Clause 1 and Clauses 4 to 7 have been coordinated and harmonized;

— the Normative references have been updated;

— Clause 3 has been revised to harmonize the terminological entries for covering all sets of language identifiers, coordinated with ISO 21636-1 and other pertinent standards;

— Clause 4 has been extended to cover additional guidelines;

— Clause 5 has been reformulated to harmonize previous texts;

v

— Clause 6 has been added;

— Clauses 7 and 8 have been revised to incorporate existing provisions, extending them to cover new requirements in the field of language coding;

— a new Annex A has been added to display the relationship between the sets of language identifiers and the language code elements of the ISO 639 language code;

— a new Annex B has been added to outline the role of the ISO 639 Maintenance Agency (ISO 639/MA) and its handling of Change Requests (CR).

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document provides comprehensive principles, rules and guidelines for the identification and specification of individual languages or language groups, as well as for their coding. The ISO 639 language code comprises language code elements each consisting of one to three language identifiers, one unique language reference name, zero or more language names in English and French, and a code element scope. Each language identifier designates a recognized individual language or language group.

For historical reasons, the language identifiers have been organized into several sets. These sets correspond to the code tables of the former ISO 639 series (all under the general title *Codes for the representation of individual languages and language groups*). This document merges these sets into a unified system of language code elements with one to three language identifiers each. Some language identifiers belong to more than one of the following sets (for more details, see 5.2):

— Set 1 comprises two-letter language identifiers (in accordance with the former ISO 639-1) for major, mostly national individual languages. This set has evolved from normative documents published in 1967.

— Set 2 comprises three-letter language identifiers (in accordance with the former ISO 639-2) for a larger number of widely known individual languages. Set 2 includes all individual languages covered by Set 1 and encompasses 20 language code elements for individual languages that each have two different three-letter language identifiers: one for bibliographic use (2B) and one for terminological use (2T). The latter distinction is based on documents widely used since the 1970s. It has emerged from two diverging language coding initiatives in the fields of librarianship and documentation on the one hand and terminology and other language applications on the other.

— Set 3 also comprises three-letter language identifiers (in accordance with the former ISO 639-3) and aims at covering comprehensively all individual languages (including all individual languages covered by Set 2).

— Set 5 also comprises three-letter language identifiers (in accordance with the former ISO 639-5), covering a larger set of language groups (including all language groups covered by Set 2). Set 5 was established in 2008. All language identifiers for language groups in Set 2 also belong to Set 5.

These four sets of language identifiers of the ISO 639 language code are used by very large user communities, which demands a high degree of coordinated code stability. Over the decades the stakeholders involved in developing the ISO 639 language code have coordinated their activities with the result that the currently valid language code elements constitute one single common code space. This implies that all language code elements, independently of which set or sets their language identifiers belong to, constitute the ISO 639 language code. Therefore, "code" occurs in the singular in the title of this document.

The above-listed sets of language identifiers are accessible at the Language Coding Agencies (LCA) hosting the data for the ISO 639 language code (see Annex B).

The language identifiers for individual languages or language groups can be used in a variety of applications. These include the specification of the language used in a text, the language of terms or words in a dictionary or terminology database, the language used in a spoken presentation, the identification of language proficiency, the capability of handling human language in software, the documentation of language resources, etc. The various sets of language identifiers are implemented in a broad range of applications, including normative documents, such as those for IETF BCP 47[12] language tags.

Code for individual languages and language groups

1   Scope

This document specifies the ISO 639 language code and establishes the harmonized terminology and general principles of language coding. It provides rules for the selection, formation, presentation and use of language identifiers as well as language reference names. It also gives provisions (i.e. principles, rules and guidelines) for the selection, formation and presentation of language names in English and French. Furthermore, it introduces provisions for the adoption of standardized language code elements using language names other than English or French.

NOTE   English, French and Russian are the official ISO languages.

In addition, this document gives guidance on the use of language identifiers and describes their possible combination with identifiers of other codes.

Specifically excluded from the ISO 639 language code are reconstructed languages or formal languages, such as computer programming languages and markup languages.

The ISO 639 language code is maintained by the ISO 639 Maintenance Agency (ISO 639/MA) (see Annex B).

2   Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements 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 3166-1, *Codes for the representation of names of countries and their subdivisions — Part 1: Country code*

ISO 15924, *Information and documentation — Codes for the representation of names of scripts*

3   Terms and definitions

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

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

—   ISO Online browsing platform: available at https://www.iso.org/obp

—   IEC Electropedia: available at https://www.electropedia.org/

3.1   Terms related to language and individual languages

3.1.1
human language
means of communication characterized by a systematic use of sounds, visual-spatial signs, characters or other written symbols or signs that can be combined to express or communicate meaning or a message between humans

Note 1 to entry: Human language was originally developed for, and mainly used in, direct communication between humans. Today its use is increasingly supported by information and communication technologies (ICTs).

Note 2 to entry: As the term "language" can represent different concepts, it is not listed as a synonym to the term "human language".

Note 3 to entry: Visual-spatial signs are indicated under *signed modality* (3.5.4).

[SOURCE: ISO 21636-1:—[1]), 3.1.1]

3.1.2
idiolect

comprehensive set of all expressions of *human language* (3.1.1) with their meaning, characterized by a coherent system of structural features, which is capable of coding complex facts and thoughts, potentially used by a given individual person, in a given type of situation, at a given time, and in a given medium

Note 1 to entry: Typically, a person has command of several idiolects of an *individual language* (3.1.3).

EXAMPLE 1   (hypothetical) The written informal American English used in the 1950s by a given person (an adult working-class person with no academic training from Philadelphia, USA, who is a native speaker of English and who does not show any particularly enhanced or constrained communicative functioning abilities).

EXAMPLE 2   (hypothetical) A given middle-aged renowned British academic in the field of linguistics in 2010: a) speaking with received pronunciation at panel discussions with colleagues in academic conferences, and b) discussing tennis tournaments with fellow tennis club members of the same town in the south-east of England.

[SOURCE: ISO 21636-1:—, 3.1.2, modified — Note 1 to entry shortened. Examples added to illustrate the concept.]

3.1.3
individual language

individual human language
largest set of *idiolects* (3.1.2), used by different *speakers* (3.1.5), which are all interconnected through high mutual intelligibility, or through a chain of high mutual intelligibility, or which are sociopolitically considered as a unit equivalent to such a largest set

Note 1 to entry: Individual languages also encompass *constructed languages* (3.3.2), but do not include formal languages (as defined in ISO 1087:2019, 3.1.10).

Note 2 to entry: Usually, in other contexts, individual languages are simply called "languages". However, the term "language" has multiple meanings and connotations, which can cause confusion in the context of this document. Still, when an attribute and possibly the plural clearly indicate that *individual languages* are meant, this document uses only "language(s)", as in "creole languages", "Asian languages" or *"living languages"* (3.3.3).

EXAMPLE   English, Guarani, LIBRAS (Língua Brasileira de Sinais/Brazilian Sign Language), Haitian Creole, Esperanto.

[SOURCE: ISO 21636-1:—, 3.1.3]

3.1.4
individual sign language

individual language (3.1.3) whose *basic modality* (3.5.5) is the *signed modality* (3.5.4)

Note 1 to entry: Usually "sign language" is part of the name of the respective individual sign language.

EXAMPLE   American Sign Language (ASL), langue des signes québéquoise/Quebec Sign Language (LSQ).

Note 2 to entry: Individual sign languages differ from the "signed modality", by which an individual language can be expressed which is normally expressed in another *language modality* (3.5.1), such as by "Signing Exact English" for expressing English. Therefore, the term "signed language" is not used as a synonym to the term "individual sign language".

[SOURCE: ISO 21636-1:—, 3.1.4]

---

1)   Under preparation. Stage at the time of publication: ISO/DIS 21636-1:2023.

3.1.5
speaker
person who is capable of making use of an *idiolect* (3.1.2)

Note 1 to entry: The term "speaker" covers the use of all *language modalities* (3.5.1) and is thus used to denote a generic concept "speaker", also covering all specific concepts such as "writer", "signer", etc., which can be introduced when needed.

[SOURCE: ISO 21636-1:—, 3.1.5, modified — Note 1 to entry shortened.]

3.1.6
language community
socially identifiable group that consists prototypically of the *speakers* (3.1.5) of an *individual language* (3.1.3) or *language variety* (3.2.4) and whose members identify with that individual language or language variety

Note 1 to entry: In the case of endangered languages, the speakers of the individual language can actually be a minority of a language community.

3.2 Terms related to linguistic variation and language varieties

3.2.1
linguistic variation
language variation
differences within and between *individual languages* (3.1.3)

[SOURCE: ISO 21636-1:—, 3.2.1]

3.2.2
external criterion for linguistic variation
set of properties of *idiolects* (3.1.2) that are based on factors external to the linguistic features of the idiolects' systems

Note 1 to entry: External criteria for linguistic variation contain properties of idiolects that pertain to the *speakers* (3.1.5) who use the idiolects, or to the *language use event* (ISO 21636-1:—, 3.1.6) in which the idiolects are used.

EXAMPLE   "Being characteristic of speakers from East Anglia" is a property which is the only element of an external criterion for linguistic variation [in this case, a criterion related to geographical space (see the example to ISO 21636-1:—, 3.2.4), defining a certain *dialect* (3.2.5) of English].

[SOURCE: ISO 21636-1:—, 3.2.2]

3.2.3
structural criterion for linguistic variation
set of properties of *idiolects* (3.1.2) that are based on the linguistic features of the idiolects' systems

Note 1 to entry: This set of properties includes in particular phonetic, phonological, morphological, syntactic, lexical, semantic or pragmatic properties.

Note 2 to entry: Elements of the structural criterion for linguistic variation are also called "markers", e.g. in ISO/TR 20694. The term "structural criterion for linguistic variation" is preferred because it integrates better with the framework for *linguistic variation* (3.2.1) developed in the ISO 21636 series.

[SOURCE: ISO 21636-1:—, 3.2.3]

3.2.4
language variety
variety
<language coding> largest subset of an *individual language* (3.1.3) that is internally consistent with regard to both an *external criterion for linguistic variation* (3.2.2) and a *structural criterion for linguistic variation* (3.2.3), and that can be identified and named

Note 1 to entry: Since terms such as "*linguistic variation*" (3.2.1), "language variation", "linguistic variant", "language variant" or "linguistic variety" are also used to represent other concepts, only the term "language variety" is used in this document.

[SOURCE: ISO 21636-1:—, 3.2.5]

3.2.5
dialect
<linguistics> *language variety* (3.2.4) specific to *speakers* (3.1.5) from a particular geographical location or region

[SOURCE: ISO 21636-1:—, 3.4.1, modified — Note 1 to entry deleted.]

3.2.6
standard variety
language variety (3.2.4) recognized as standard or official by most *speakers* (3.1.5) across the geographical area where the *individual language* (3.1.3) is spoken or used, or across a large part of that geographical area where several *dialects* (3.2.5) are used

Note 1 to entry: A standard variety of an individual language is typically used in official or public communication and in communication between users of different language varieties.

Note 2 to entry: A standard variety is often characterized by a high degree of standardization or normalization.

[SOURCE: ISO 21636-1:—, 3.4.2, modified — Note 1 to entry deleted.]

3.3 Terms related to types of individual languages subject to language coding

3.3.1
natural language
individual language (3.1.3) which is or was in active use in a *language community* (3.1.6), passed on from one generation of *speakers* (3.1.5) to the next

EXAMPLE   Bambara, English, Haitian Creole, Latin, LIBRAS (Língua Brasileira de Sinais/Brazilian Sign Language).

[SOURCE: ISO 21636-1:—, 3.1.9]

3.3.2
constructed language
individual language (3.1.3) whose rules are explicitly established prior to its use

EXAMPLE   Esperanto, Volapük, Quenya, Na'vi.

Note 1 to entry: Constructed languages do not include *reconstructed languages* (3.3.9), computer programming languages, mark-up languages or similar formal languages.

Note 2 to entry: Some constructed languages are based on one or several *natural languages* (3.3.1) and are therefore not artificial. Therefore, the term "artificial language", which is often used as a synonym, is not used in this document.

[SOURCE: ISO 21636-1:—, 3.1.10]

3.3.3
living language
individual language ([3.1.3](#)) used by a *language community* ([3.1.6](#)) at the present time

EXAMPLE   English, Bambara.

Note 1 to entry: Living languages include previously *extinct languages* ([3.3.4](#)) that have been revitalized, such as Hebrew or Manx.

3.3.4
extinct language
individual language ([3.1.3](#)) that is no longer in use and that has lost its last native *speaker* ([3.1.5](#)) in recent centuries

EXAMPLE   Old Prussian, Amanayé, Basay.

3.3.5
historical language
individual language ([3.1.3](#)) that went out of use before modern times

EXAMPLE   Old English, Middle English, Ancient Greek, Etruscan.

Note 1 to entry: Historical languages include known earlier epochs of a *living language* ([3.3.3](#)) or of an *extinct language* ([3.3.4](#)).

3.3.6
full language
individual language ([3.1.3](#)) that is used in a variety of domains, used to support communication across most or all social groups, such as genders and age groups of a *language community* ([3.1.6](#)), and stable enough to have at least one *language variety* ([3.2.4](#)) that is widely understood across the whole geographical area of the individual language

Note 1 to entry: Full languages are subject to gradual change as each new generation makes their individual language its own.

3.3.7
restricted language
language variety ([3.2.4](#)) of an *individual language* ([3.1.3](#)) or *constructed language* ([3.3.2](#)) that resembles in some contexts an individual language, but that is not a *full language* ([3.3.6](#)) in that it is used in limited circumstances

Note 1 to entry: Restricted languages are, for instance, avoidance languages, secret languages, ceremonial languages, play languages, urban languages or youth languages.

3.3.8
macrolanguage
individual language ([3.1.3](#)) that for the purpose of language coding can be subdivided into two or more other individual languages

EXAMPLE   In some contexts, Chinese is declared as the individual language of documents written in Han ideographic *script* ([3.6.2](#)), while the language of the document can be Mandarin, Yue, Wu or any other related individual language of China.

Note 1 to entry: In contrast to *language groups* ([3.4.1](#)), the individual languages that correspond to a macrolanguage shall be very closely related, and there shall be some context in which they are deemed recognizable as separate individual languages.

Note 2 to entry: Language development can change the relationship between a macrolanguage and related individual languages or *dialects* ([3.2.5](#)) over time.

3.3.9
reconstructed language
hypothetical *individual language* ([3.1.3](#)) that is not itself attested but postulated based on regular correspondences among a group of attested languages, of which the hypothetical language is held to be a common ancestral language

EXAMPLE   Proto-Indo-European, Proto-Germanic.

3.4   Terms related to language groups subject to language coding

3.4.1
language group
set of *individual languages* ([3.1.3](#)) based on a shared characteristic

EXAMPLE   "Caucasian languages" (geographical group of languages); "North Caucasian languages" [*language family* ([3.4.2](#))].

Note 1 to entry: Shared characteristics include, but are not limited to, being spoken in the same geographical region, being genetically related [i.e. being (a branch of) a language family] or sharing typological characteristics such as being a tonal language or being a morphologically isolating language, or combinations thereof.

Note 2 to entry: A language group can encompass several smaller language groups.

Note 3 to entry: An identified language group is a group of individual languages that does not correspond to a *macrolanguage* ([3.3.8](#)). It can contain macrolanguages among its members.

3.4.2
language family
language group ([3.4.1](#)) encompassing *individual languages* ([3.1.3](#)) based on a common ancestral language

EXAMPLE   North Caucasian languages.

Note 1 to entry: It is possible for a language family to consist of a single individual language as a member.

3.5   Terms related to modalities of language use

3.5.1
language modality
language variety ([3.2.4](#)) specific to a certain medium or channel used for communication by the *speaker* ([3.1.5](#))

Note 1 to entry: As specified in ISO 21636-1, language modalities encompass the *spoken modality* ([3.5.2](#)), the *written modality* ([3.5.3](#)), the *signed modality* ([3.5.4](#)), the drummed modality, the whistled modality, etc. This document refers only to the spoken modality, the written modality and the signed modality.

[SOURCE: ISO 21636-1:—, 3.4.7, modified — Notes 1 to 3 to entry replaced by a new Note 1 to entry.]

3.5.2
spoken modality
spoken language modality
language modality ([3.5.1](#)) used in oral communication, and that has as its primary mode of expression the articulated sounds of the *speaker's* ([3.1.5](#)) voice

Note 1 to entry: Events of language use in the spoken modality are usually also multimodal, as gestures, facial expressions and other similar phenomena almost always accompany speech.

[SOURCE: ISO 21636-1:—, 3.5.1, modified — Notes 1 to 3 to entry replaced by a new Note 1 to entry.]

3.5.3
written modality
written language modality
language modality (3.5.1) that makes use of a system of graphic symbols

Note 1 to entry: Some *constructed languages* (3.3.2) have the written modality as their *basic modality* (3.5.5), but in *natural languages* (3.3.1), the *spoken modality* (3.5.2) or *signed modality* (3.5.4) are basic modalities from which the written modality is derived.

[SOURCE: ISO 21636-1:—, 3.5.3, modified — Note 2 to entry deleted.]

3.5.4
signed modality
signed language modality
visual-spatial *language modality* (3.5.1) that uses a combination of hand shapes, palm orientation and movement of the hand, arm, mouth, head or body, and facial expression

Note 1 to entry: The language modality "signed modality" differs from *"individual sign language"* (3.1.4), which designates those *individual languages* (3.1.3) whose *basic modality* (3.5.5) is the signed modality.

Note 2 to entry: The signed modality is the basic modality for individual sign languages, but some other individual languages also have a signed modality.

EXAMPLE   Signing Exact English, the signed modality for expressing (spoken or written) English.

[SOURCE: ISO 21636-1:—, 3.5.4, modified — Last part of Note 1 to entry deleted.]

3.5.5
basic modality
basic language modality
language modality (3.5.1) in which a particular *individual language* (3.1.3) is most commonly used, in which it develops, and from which other language modalities (if any) of that individual language are derived

Note 1 to entry: For most *natural languages* (3.3.1), the basic modality is the *spoken modality* (3.5.2); in particular, the *written modality* (3.5.3) is derived from the spoken modality in the sense that in these natural languages writing is a representation of properties of oral language use [even in the case of non-alphabetic *writing systems* (3.6.1)].

Note 2 to entry: For *individual sign languages* (3.1.4), the basic modality is the *signed modality* (3.5.4). If these are written or otherwise graphically represented, the written modality is derived from the signed modality in the sense that writing is a representation of properties of visual-spatial language use (i.e. of the signed modality).

[SOURCE: ISO 21636-1:—, 3.5.11, modified — Note 1 to entry shortened and Note 3 to entry deleted.]

3.6   Terms related to written representation of individual languages

3.6.1
writing system
system for writing an *individual language* (3.1.3), including the *script* (3.6.2) and orthographic conventions used

3.6.2
script
<language coding> comprehensive set of graphic characters used for the written form of one or more *individual languages* (3.1.3)

EXAMPLE   Cyrillic, Hiragana.

Note 1 to entry: A script, as opposed to an arbitrary subset of characters, is defined in distinction to other scripts; in general, readers of one script can be unable to read the glyphs of another script easily, even where there is a historical relation between them.

3.6.3
script identifier
script symbol
<language coding> string of characters assigned to represent a *script* (3.6.2) unequivocally

3.7 Terms related to language coding and naming languages

3.7.1
code
collection of rules that maps the elements of a first set of values onto the elements of one or more different sets

Note 1 to entry: The first set of elements in the ISO 639 language code refers to *individual languages* (3.1.3) or *language groups* (3.4.1). The sets of elements mapped onto are code values, specifically *language identifiers* (3.7.10).

[SOURCE: ISO/IEC 2382:2015, 2121552, modified — Admitted term "coding scheme" deleted. "first set of values onto the elements of one or more different sets" replaced "first set onto the elements of a second set" in the definition. Notes 1 to 2 to entry replaced by a new Note 1 to entry. Notes 3 to 5 to entry deleted.]

3.7.2
code element
collection of structured information about an entity represented in a *code* (3.7.1)

3.7.3
language code
code (3.7.1) that maps *individual languages* (3.1.3) or *language groups* (3.4.1), represented by their unique *language reference name* (3.7.13), to *language identifiers* (3.7.10)

EXAMPLE   The ISO 639 language code.

Note 1 to entry: The activity and subject of developing a language code is called "language coding".

3.7.4
language code element
code element (3.7.2) of a *language code* (3.7.3) that represents an *individual language* (3.1.3) or a *language group* (3.4.1)

EXAMPLE   "Efik" [*language reference name* (3.7.13)], "Efik" [*language name* (3.7.12) in English], "efik" (language name in French), "efi" [*language identifier* (3.7.10)] and the information on the *code element scope* (3.7.5) "individual language" together constitute a language code element.

Note 1 to entry: A language code element is composed of a minimum of one unique language reference name and one or more language identifiers that belong to one or more sets.

Note 2 to entry: Language code elements cited in this document each represent an individual language or language group. In addition to the language code elements, the ISO 639 language code also contains four *special-purpose code elements* (3.7.9).

3.7.5
code element scope
attribute of a *code element* (3.7.2) that documents the way in which its denotation maps onto one or more *individual languages* (3.1.3)

Note 1 to entry: There are four code element scopes: *individual-language code element* (3.7.6), *macrolanguage code element* (3.7.7), *language-group code element* (3.7.8) and *special-purpose code element* (3.7.9).

3.7.6
individual-language code element
language code element ([3.7.4](#)) with a value of a *code element scope* ([3.7.5](#)) for an *individual language* ([3.1.3](#))

Note 1 to entry: The individual language represented by an individual-language code element is considered distinct from any other individual language represented by a different individual-language code element.

Note 2 to entry: An individual-language code element represents an individual language that is not a *macrolanguage* ([3.3.8](#)).

3.7.7
macrolanguage code element
language code element ([3.7.4](#)) with the value of a *code element scope* ([3.7.5](#)) for a *macrolanguage* ([3.3.8](#))

Note 1 to entry: A macrolanguage code element represents an *individual language* ([3.1.3](#)) that is a macrolanguage.

3.7.8
language-group code element
language code element ([3.7.4](#)) with the value of a *code element scope* ([3.7.5](#)) for a *language group* ([3.4.1](#))

Note 1 to entry: A language-group code element represents a group of *individual languages* ([3.1.3](#)) that does not correspond to a *macrolanguage* ([3.3.8](#)).

3.7.9
special-purpose code element
language code element ([3.7.4](#)) of a *language code* ([3.7.3](#)) with the value of a *code element scope* ([3.7.5](#)) for a special purpose

EXAMPLE   mul – multiple languages in the content

Note 1 to entry: A special-purpose code element does not refer to any *individual language* ([3.1.3](#)) or *language group* ([3.4.1](#)) identified in the ISO 639 language code but is useful for data management purposes.

3.7.10
language identifier
language symbol
string of characters assigned to an *individual language* ([3.1.3](#)) or a *language group* ([3.4.1](#)) for the purpose of identifying it unequivocally

Note 1 to entry: In the ISO 639 language code, the string of characters consists of a string of letters.

Note 2 to entry: When a single *language code element* ([3.7.4](#)) has multiple language identifiers, the language identifiers are synonymous, while each of them identifies the individual language or language group unequivocally.

EXAMPLE   The individual language "Dutch" is assigned the two-letter language identifier "nl", a three-letter identifier "nld" for use in the field of terminology and other language applications, and another three-letter identifier "dut" for use in the field of librarianship and documentation. The individual language "Polish" is assigned the two-letter language identifier "pl" and the three-letter identifier "pol". The language group "Khoisan languages" is assigned the three-letter language identifier "khi".

3.7.11
local-use language identifier
language identifier ([3.7.10](#)) reserved for local use

EXAMPLE   "qij" (designating "Akeu") temporarily used locally in place of a permanent language identifier.

Note 1 to entry: Local-use language identifiers are not part of any assigned *language code element* ([3.7.4](#)).

Note 2 to entry: All local-use language identifiers are composed of three letters.

9

3.7.12
language name
linguistic expression for naming an *individual language* (3.1.3) or a *language group* (3.4.1)

EXAMPLE   "anglais" (individual language "English" in French language), "Caucasian languages" (language group).

Note 1 to entry: The activity and subject of assigning language names is called "naming languages".

Note 2 to entry: Often an individual language or language group can be represented by more than one language name or variant of a language name in a given language.

3.7.13
language reference name
language name (3.7.12) for the purpose of designating an *individual language* (3.1.3) or *language group* (3.4.1) uniquely within a *language code* (3.7.3)

Note 1 to entry: When the common language names of two identified different individual languages or language groups are the same, disambiguating or differentiating information is added to construct a unique language reference name for each.

EXAMPLE   The language reference name "French" without disambiguating or differentiating information, the language reference names "Mor (Mor Islands)" and "Mor (Bomberai Peninsula)" with disambiguating or differentiating information.

4   Linguistic variation and fundamental concepts of identifying and categorizing individual languages

4.1   General

The objects of language coding according to this document are individual languages or language groups. Individual languages shall be full languages (see 4.2.2.2). Clause 4 describes the aspects of linguistic variation that shall be considered when identifying individual languages and introduces the different types of individual languages and language groups that are encompassed in the ISO 639 language code. According to this document, an individual language or language group is represented by one to three language identifiers, one unique language reference name, zero or more language names in English and French, and a code element scope, which together form a language code element (see 5.3).

4.2   Linguistic variation

4.2.1   Criteria and dimensions of linguistic variation

4.2.1.1   External criteria and structural criteria of linguistic variation

Individual languages and language groups are not static or homogeneous objects, but there is variation within them. Every language identifier in accordance with this document corresponds either to an individual language in its entirety with all its language varieties or to a language group.

In some conceptional aspects, this document follows the ISO 21636 series, which focuses on linguistic variation and according to which individual languages are conceived as sets of idiolects. Each idiolect is an internally homogeneous means of communication characterized by a coherent system of structural features as potentially used by a specific speaker in a specific situation. A specific situation comprises the type of situation, at a given time, and in a given medium.

Language varieties are subsets of individual languages; they are set apart one from another both by external criteria for linguistic variation and also by structural criteria for linguistic variation. All idiolects of an individual language are categorized into language varieties according to such criteria. External criteria refer in particular to either the speaker or the act of communication, or both, while

structural criteria identify, for instance, certain sounds, words or morphosyntactic properties which are characteristic of the language variety in question.

4.2.1.2 Dimensions of linguistic variation

Sets of external criteria of the same type (referring to similar properties of idiolects, e.g. properties related to time, to geographical area or to the social origin of the speaker) form a dimension of linguistic variation, resulting in different kinds of language varieties, including:

— the space dimension, resulting, in particular, in dialects;

— the time dimension, including, in particular, epochs of languages;

— the social group dimension, including sociolects and technolects;

— the medium dimension, referring to language modalities such as the spoken modality, the written modality and the signed modality, among others;

— the situation dimension, referring to registers;

— the person dimension, referring to personal varieties;

— the proficiency dimension, including learner varieties;

— the communicative functioning dimension.

Although this document is not about linguistic variation, some aspects of linguistic variation shall be considered when selecting an individual language or language group for coding. Therefore, this document focuses on:

— the medium dimension (limited to the spoken modality, written modality and signed modality);

— the time dimension (including, for instance, periods and epochs of languages);

— the space dimension (mainly referring to dialects) of linguistic variation.

4.2.1.3 Individual languages and dialects

Some individual languages are spoken in multiple regions, and there can be major differences in usage between regions, such as differences in dialect or script. Even if regional differences are regarded as significant, a single language code element is provided for an individual language encompassing all its internal spoken or written varieties, in particular all its dialects, including any linguistically normalized variety. On the other hand, there are borderline cases and some legacy language code elements where some users for some purposes can consider a language code element with a language identifier of Set 1 or Set 2, or in particular Set 3, to refer to a dialect rather than an individual language (see 6.2.1). Some of these cases are covered by the concept of macrolanguages (see 5.7.3).

For applications in which it is necessary to identify dialects or other language varieties, the framework of the ISO 21636 series can be used, possibly in combination with other standards such as ISO/TR 20694 or the codes of the ISO 3166 series.

In 6.2, criteria are provided for distinguishing between dialects and full individual languages for the purpose of language coding.

4.2.1.4 Linguistic norms and language status

An individual language or a certain language variety can be standardized by official or private bodies such as academies or language councils. The result can be a standard variety which is recognized across all or most regions where the language is used. For example, "Standard High German" is recognized (despite some regional differences) in all regions of Germany, Austria and Switzerland. Standardization

can be applied to orthography or any aspect of grammar, such as phonology, morphology, syntax, semantics, etc. The degree of standardization can vary greatly from one individual language to another.

An individual language or a certain language variety of that individual language can have an official status within a country or region. Although most of the roughly 7 000 individual languages identified in accordance with this document do not have a formally recognized official status, the existence of such status can be taken into account in language identifier assignment (in particular for the language identifiers in Set 1 and Set 2, see 6.2.6).

4.2.2 Identification and types of individual languages and language groups

4.2.2.1 Identification of individual languages

There is no one definition of "individual language" that is agreed upon by all domain experts and appropriate for all purposes. As a result, people can disagree with regard to their assessment of a given set of idiolects as forming or not forming an individual language, depending on their assessment of the mutual intelligibility and the sociopolitical situation. Therefore, there are cases where disputes arise whether the status of a given set of idiolects constitutes an individual language, or a variety of a language, or a language group of closely related languages.

Hence, both speakers of the individual language and domain experts can have different opinions as to whether two particular language varieties represent dialects of a single individual language or two distinct individual languages.

It is one of the goals of this document to provide language identifiers for every distinct full language (see 4.2.2.2) – whether a living language or an extinct language, and whether its primary modality is the spoken modality, the written modality or the signed modality – and for principal language groups that have been documented.

For this document, judgements regarding whether two language varieties are considered to be the same or different individual languages are based on a number of factors, including linguistic similarity, mutual intelligibility, a common literature, the views of speakers concerning the relationship of a language name and the concept for an identified individual language or language group, and other factors (see 6.1).

4.2.2.2 Full languages and restricted languages

Apart from language groups (see 4.2.2.6), language identifiers in the ISO 639 language code may only be assigned to full languages. Subclause 6.2.1 specifies the criteria used in this document for being considered a full language.

Endangered languages, extinct languages and attested historical languages which have met these criteria in the past are within the scope of this document.

By contrast, there are many kinds of restricted languages that have distinct names and are documented, but that do not qualify for being coded in accordance with this document. Restricted languages are, for instance, avoidance languages, secret languages, ceremonial languages, play languages, urban languages, youth languages, etc. They are very limited as to the domain of use or are specifically meant to exclude certain social groups of a larger community from communication or are so dynamic and unstable as to lack widespread linguistic norms of use. Thus, restricted languages are only used in a few thematic subjects, usually to support communication of only one gender or only selected social or age groups of a community. Furthermore, they usually are not stable enough to have at least one variety that is widely understood across the whole area of the respective individual language over the time span of a generation. Even though in some contexts a restricted language can resemble an individual language, it is not a full language.

4.2.2.3 Individual sign languages and signed modality of individual languages

Individual sign languages are distinct individual languages, not dependent upon or derived from spoken individual languages. There are hundreds of individual sign languages in use around the world today.

Common names of individual sign languages may make reference to a region in which they are used, e.g. "Quebec Sign Language" (LSQ).

Individual sign languages shall not be confused with the signed modality for mainly spoken individual languages, such as "Signing Exact English", a variety (signed modality) of English. For individual sign languages, the signed modality is the basic modality. If other modalities of an individual sign language exist, such as a written representation, they are derived from the signed modality.

4.2.2.4 Extinct languages and historical languages

In addition to living languages (currently used by language communities), the ISO 639 language code includes language identifiers that designate extinct languages, which went out of use in modern times, as well as historical languages, which went out of use before modern times. Some of the latter constitute an earlier epoch of a living language or of an extinct language, although for language coding they should be considered distinct from any modern individual language that can be descended from them.

A reconstructed language inferred from historical-comparative analysis but not attested from historical sources is not considered eligible for language coding as an extinct language or historical language.

4.2.2.5 Constructed languages

In addition to natural languages, there are many constructed languages that have been consciously created for certain purposes, such as to facilitate international communication, as a component of works of art (e.g. books, movies, TV series) or even as a hobby for the enjoyment of the creative act. A constructed language is eligible to receive a language identifier if the characteristics of a full language have been incorporated into the constructed language (see 6.2.3).

Constructed languages eligible for language coding in accordance with this document do not include historical comparative reconstructions of assumed but not documented protolanguages, as well as formal languages.

4.2.2.6 Language groups

Individual languages, whether present or past, can be combined into different sets based on shared characteristics. A well-known example of language groups is that of language families, which are based on a common ancestral language. Other language groups can be based on geographical vicinity, ethnic proximity, similar linguistic features, etc.

A large language group can encompass smaller language groups. While some language groups comprise many individual languages, others comprise a few or only one individual language. Each language group included in the ISO 639 language code encompasses all individual languages and smaller language groups associated with that group.

4.3 Individual languages and writing conventions

Many individual languages are written using more than one script, such as Kurdish, which is written either using Arabic script or Latin script. Also, even within the same script, there often are different orthographic conventions or rules. In some cases, such language-internal differences in script or orthographic conventions coincide with country borders.

Such differences, like other types of language-internal variation (see the ISO 21636 series), are outside the scope of this document, which assigns language identifiers to individual languages (and language groups) in their entirety. Therefore, individual languages written in one or more scripts or using more than one set of orthographic conventions are assigned only one language identifier in the respective

sets of language identifiers. If necessary, the name of the script or orthographic convention can be indicated. When indicating the script in combination with a language identifier, the script identifiers from ISO 15924 (see Clause 8) shall be used.

5 The ISO 639 language code

5.1 Introduction to the ISO 639 language code

The ISO 639 language code is conceived as a unified system of normative language code elements that each include one to three language identifiers (see 5.4), one unique language reference name (see 5.5), zero or more language names in English and French (see 5.6), and a code element scope (see 5.7).

The ISO 639 language code can be looked at from the perspective of the sets of language identifiers (see 5.2 and Annex A) as they emerged from their early use, or from the perspective of the structure of its language code elements (see 5.3), or from the perspective of the code element scopes (see 5.7).

5.2 Sets of language identifiers

From the perspective of the sets of language identifiers for individual languages and language groups, the ISO 639 language code is organized as follows (see also Annex A):

— Set 1 comprises two-letter language identifiers for a limited number of major, mostly national individual languages (including a few macrolanguages, see 5.7.3).

— Set 2 comprises three-letter language identifiers for a larger number of widely known individual languages (including all individual languages covered by Set 1) and a number of language groups. This set covers 20 individual languages that each have two different three-letter language identifiers: one for bibliographical use (2B) and one for terminological use (2T); see Table A.2.

— Set 3 also comprises three-letter language identifiers; it has been developed to cover all individual languages in the world comprehensively (including macrolanguages). All individual languages that have a language identifier in Set 2 are designated by the same language identifier in Set 3. In the case of the 20 individual languages designated by both a 2B and a 2T language identifier, only the 2T language identifiers are included in Set 3.

— Set 5 also comprises three-letter language identifiers covering a larger number of language groups (including language families) than Set 2 and is particularly important for librarianship and documentation. All language identifiers for language groups in Set 2 also belong to Set 5.

5.3 Language code elements and their structural components

Each language code element conceptually represents one identified individual language or language group. Any identified individual language or language group that is represented by a language code element in the ISO 639 language code is assigned at least one language identifier across the whole ISO 639 language code. Where there is more than one language identifier in a language code element, the language identifiers designate exactly the same individual language or language group.

The language code elements of the ISO 639 language code are normative and comprise the following structural components:

— usually one, sometimes two or rarely three language identifiers designating an identified individual language or language group as one of the core components of the language code element (see 5.4);

— one unique language reference name representing an identified individual language or language group as the second core component of the language code element (see 5.5);

— zero or more language names in English (see 5.6);

— zero or more language names in French (see 5.6);

— a code element scope (see 5.7).

For examples of language code elements with one, two and three language identifiers, see Table A.1 and Table A.2.

Other non-normative information can be provided, such as language type (see 4.2.2).

5.4 Language identifiers

Language identifiers are immutable across the whole ISO 639 language code. This implies that an assigned language identifier shall not be used or reused for any other identified individual language or language group.

Language identifiers are composed of the following 26 lowercase letters of the Latin alphabet: a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z. Other letters, symbols or signs shall not be used. Language identifiers of Set 1 are composed of two letters. Language identifiers of Set 2, Set 3 and Set 5 are composed of three letters. Any letter may be used more than once in a language identifier.

Language identifiers shall not be viewed as abbreviated forms of their respective language reference names, or any other language names for the respective individual language or language group. This applies to the two-letter language identifiers of Set 1 as well as to the three-letter language identifiers of the other sets.

Any language identifier in a new language code element is chosen from the number of still available three-letter combinations (not yet belonging to any of Set 2, Set 3 and Set 5, and not yet assigned in any language code element). Any assigned three-letter language identifier designates a single identified individual language (possibly a macrolanguage) or language group, regardless of the set to which it belongs.

While two-letter language identifiers in Set 1 are formally distinct from the three-letter language identifiers in Set 2, Set 3 and Set 5, the identified individual languages or language groups represented by the respective language code elements are identical. This implies that a three-letter language identifier designates exactly the same individual language or language group as a two-letter language identifier, if they belong to the same language code element with one single unique language reference name. This principle also applies to the 20 individual-language code elements that each have two different three-letter language identifiers in Set 2: one for bibliographic use (2B) and one for terminological use (2T); see Clause A.2.

5.5 Language reference names

The language reference name is the primary means by which the ISO 639 language code documents the individual language or language group that is designated by a language identifier. As such, the language reference name is the second core component of the language code element.

A language reference name is normative but not immutable. It can be modified, given exceptionally compelling reasons (see 5.8.2), provided it keeps representing the same individual language or language group designated by its language identifier(s).

Each language reference name in a language code element is normative insofar as it represents an identified individual language or language group. Language reference names in the ISO 639 language code are linguistic expressions assigned for the purpose of language coding. They are widely used language names taken from sources in English or recently widely used language names adapted to English. They are chosen to best serve the needs of uniquely designating individual languages or language groups. It is not intended or feasible to provide all existing names for any individual language or language group.

This document provides principles governing the assignment of language reference names (see 6.3) and rules for the formation and presentation of language reference names (see 6.4).

The use of language reference names within the ISO 639 language code does not imply that they have any special status within the respective language communities, in any other domain of usage or in any other context beyond the ISO 639 language code.

Applications may use other language names or spellings when language names are presented in user interfaces, provided that the language names in that context refer to the same individual language or language group as the language reference name used in the ISO 639 language code. When language code elements with language identifiers of any of the sets of language identifiers are documented within derivative specifications or other contexts, the language reference names used in the ISO 639 language code should be included.

5.6   Language names in English and French

Language code elements can contain zero or more language names in English and French. This document provides rules for the selection, formation and presentation of language names in English and French (see 6.5) as well as for language names in languages other than English or French (see 6.6).

The language names in English and French in the ISO 639 language code are normative, but not immutable. They can be modified or changed for highly compelling reasons following the rules and guidelines of this document (see 6.5).

5.7   Code element scopes and local-use language identifiers

5.7.1   Types of code element scopes

According to this document, a code element of the ISO 639 language code can have one of four different code element scopes: individual language, macrolanguage, language group and special-purpose code element.

Special-purpose language code elements are used for data management purposes and do not distinctly refer to any identified individual language or language group.

5.7.2   Code element scope of individual-language code elements

Individual languages represented by individual-language code elements are distinct full languages (see 4.2.2.2). They can be of various types (see 4.2.2), such as:

— living languages, including individual sign languages;

— extinct languages or historical languages;

— constructed languages.

For language code management purposes, macrolanguages (see 5.7.3), although they are a special case of individual languages, have their own code element scope. The code element scope of individual-language code elements is thus reserved for individual languages which are not a macrolanguage.

5.7.3   Code element scope of macrolanguage code elements

Some language code elements have the scope of a macrolanguage. The concept of macrolanguage was introduced to provide compatibility with existing applications and to provide coherence between the different sets of language identifiers, particularly between Set 1 and Set 2 in relation to Set 3. It provides a framework to allow for different practices in certain cases regarding what is considered an individual language or a language group.

While ambiguity is not, in general, desirable, it has been necessary for language coding in the case of macrolanguages due to actual practice in applications. Thus, a macrolanguage is deemed to be a single individual language in some contexts but is subdivided into two or more individual languages in other

contexts. This means that a macrolanguage code element corresponds in a one-to-many manner to at least two other individual-language code elements.

EXAMPLE   The macrolanguage "Kanuri" ("kau") encompasses the three individual languages "Central Kanuri" ("knc"), "Manga Kanuri" ("kby") and "Tumari Kanuri" ("krt").

"Chinese" has been used in some contexts and applications to mean specifically Mandarin Chinese "cmn", but it has also been used in other contexts as a group of languages that encompasses all Chinese languages. If the code element for Chinese "zh"/"zho"/"chi" were asserted to designate Mandarin Chinese only, it would break compatibility with some existing applications of language identifier Set 1 or Set 2 that have used these language identifiers to designate the Chinese language understood more broadly (including other language varieties that also happen to be assigned language identifiers of Set 3 as individual languages). The concept of macrolanguage intentionally allows for ambiguity in regard to the specific individual language designated by "zh"/"zho"/"chi".

A mapping from macrolanguages to their encompassed individual languages shall be defined and documented in the context of language identifier Set 3. This mapping shall be between code elements with a macrolanguage scope and code elements with an individual language scope; a macrolanguage cannot be mapped to other macrolanguages or to language groups. Macrolanguage mappings are normative for applications of any set of the language identifiers. In particular, a conforming application shall not use a macrolanguage code element as encompassing another macrolanguage or language group, or as encompassing an individual language that is not included in the mapping for that macrolanguage code element defined in the ISO 639 language code.

The relationships between macrolanguages and their encompassed individual languages can change over time. To minimize ambiguity in language coding, the creation of new macrolanguage code elements or a change of coverage of existing individual-language code elements to macrolanguage shall be based on evidence of application requirements for public interchange of language metadata, as well as appropriate linguistic relationships between the individual languages involved (see 6.2.4).

5.7.4   Code element scope of language-group code elements

Some language code elements provide language identifiers for language groups, also encompassing language families, as described in 4.2.2.6. Their language reference names include the word "languages" in plural form. Their language identifiers belong to Set 5, and some of them also to Set 2.

EXAMPLE 1   The language group "Caucasian languages" has the language identifier "cau" which belongs to Set 5 and Set 2.

EXAMPLE 2   The language family "North Caucasian languages" has the language identifier "cnn" which belongs to Set 5.

The exact coverage of language-group code elements is not determined in the ISO 639 language code. This implies that two different users can have divergent ideas and applications can reflect varying notions of the exact coverage of a given language-group code element.

5.7.5   Code element scope of special-purpose code elements

The ISO 639 language code includes four code elements with three-letter identifiers for special concepts that, unlike those in regular language code elements, do not designate identified individual languages or language groups. The names for special-purpose code elements shall not be considered as language reference names.

The identifiers for special-purpose code elements are included in order to allow for special purpose requirements of some applications. For example, an application can require an identifier to be applied as a document attribute even if the language of the document is unknown, or if the document includes content in multiple languages, or does not contain any linguistic content at all. Special-purpose code elements allow for these various cases.

The four special-purpose code elements have the following identifiers and meaning:

— "mis": miscellaneous language (no language code element is assigned for this individual language or language group or included in the subset of the language code used by a given application);

— "mul": multiple languages in the content;

— "und": language undetermined;

— "zxx": no linguistic content.

It is unlikely that additional special-purpose code elements will be created in the future.

5.7.6 Local-use language identifiers

The three-letter language identifiers "qaa" through "qtz" are reserved for local use and occur only in language identifier Set 2 and Set 3. No two-letter language identifiers are reserved for local use.

Local-use language identifiers shall not be assigned to a language code element or limited to any code element scope (see 5.7) in any set of the ISO 639 language code. Therefore, they are not subject to maintenance of the ISO 639 language code by the ISO 639 Maintenance Agency (ISO 639/MA).

Local-use language identifiers may be used locally, internal to an application or within a closed usage context. Applications may define local-use language code elements by assigning a language reference name and a code element scope to a local-use language identifier. They should not be used in data interchange, except by private agreement between parties.

5.8 Language code maintenance

5.8.1 Basic rules for language code maintenance

While all individual languages and language groups potentially can be subject to language coding, not all of them necessarily qualify for being assigned a language identifier in the ISO 639 language code. For future coordinated maintenance of the language code, the ISO 639 Maintenance Agency (ISO 639/MA) has been established to operate under approved Terms of Reference (MA-ToR), see Annex B.

Except for the special-purpose code elements, the ISO 639 language code requires that one language identifier always designate one specific individual language or language group, regardless of the language identifier set in which it is included. To ensure continuity and stability of the language code, a language identifier denoting an identified individual language or language group shall never be modified or changed in any way in the ISO 639 language code. This also implies that no three-letter language identifier of any set (namely Set 2, Set 3 or Set 5) once assigned to designate an identified individual language or language group shall be assigned to a different individual language or language group.

If an individual language or language group is no longer recommended for general usage, the respective language code element may be deprecated and marked as such. Justification of such decisions shall be documented by the ISO 639/MA.

Language identifiers of Set 1 and their language code elements can only be subject to changes for extraordinarily compelling reasons. Language identifiers of Set 3 may be admitted to Set 2, if they are found to be required for applications using language identifier Set 2 (see 6.2.6).

5.8.2 Language code maintenance activities

Language code maintenance comprises activities based on defined procedures, in particular:

— adding new language code elements, with language identifiers and language reference names for individual languages or language groups;

— deprecating language code elements and language identifiers from the respective set or sets where they appear;

— modifying a language reference name due to exceptionally compelling reasons;

— adding, modifying or deleting language names in English and French;

— changing a code element scope.

The above activities can be initiated by external Change Requests (CR) or by language code-inherent necessities for a modification.

A change or modification of a language reference name or language names in English or French shall not entail the change of the language identifier(s) in the language code element representing that individual language or language group.

EXAMPLE 1   The language reference name "Gavak" replaced "Dimir" which proved to be wrong, but the language identifier "dmc" remains.

In certain cases, given highly compelling reasons, the identity of the individual language or language group designated by the language identifier may be modified: broadened to encompass additional language varieties, or narrowed to exclude certain language varieties. For example, if a language code element exists for an individual language that is not well documented, but new research finds that the represented language variety is, in fact, a dialect of another individual language, then the former language code element may be deprecated. However, the language code element for the latter individual language may continue to be used, albeit with a modified denotation that now encompasses the additional dialect. Such modifications may only be made, however, if the meaning of the modified language code element as an individual language or language group is not fundamentally changed. Careful consideration shall be observed to maintain maximum stability for existing documents and applications.

EXAMPLE 2   Once "die" was used by accident for an endangered language (still in the process of being assigned a language identifier). It was deprecated and replaced by a more appropriate one; "die" was consequently blocked from being reassigned.

In complying with a request for a language code element for a new individual language or language group, the general principles and rules of this document apply.

5.9 Adoption of language code elements into other standards

Language code elements from the ISO 639 language code may be adopted into other standards, whether by adopting the ISO 639 language code as a whole, or by using language identifiers of a given set, or by using only language code elements with language identifiers from a given set. For this purpose, the following rules apply:

— Language identifiers and language reference names shall be adopted in unmodified form and always rendered in the Latin script.

— If language names in English or French are adopted, they shall be adopted in unmodified form.

— Language names in English or French may be replaced by language names in the target language of the adoption; the language reference name of the ISO 639 language code shall always be retained.

— Language names in languages other than English and French may be formed following the principles in 6.1.1, 6.1.2 and 6.3 and following the rules and conventions of the target language, including appropriate transcription (see 6.6).

— If a language code element of the ISO 639 language code is adopted, the code element scope shall not be changed.

If language code elements of the ISO 639 language code are adopted into other standards, other non-normative information can be provided as deemed necessary by the adopting ISO member body.

6 Criteria for language coding and naming languages

6.1 Information to be considered in language coding and naming languages

6.1.1 Pertinent reference sources

To determine the eligibility of an individual language or language group for coding and to find a suitable name for it, various reference sources shall be identified and consulted, in order to collect and analyse the information necessary to fulfil the tasks related to language coding and naming languages.

Given the diverse nature of individual languages and language groups, different kinds of reference sources can apply. For individual languages of wider distribution or large and well-researched language groups, the following reference sources should be consulted:

— linguistic and political (reference) works (e.g. almanacs, lexica, encyclopaedia);

— scientific articles;

— monolingual dictionaries;

— text corpora (including, for example, renowned newspapers or instruction manuals);

— texts from websites of higher education institutions, archives, language services or linguistic research institutes;

— contributions by domain experts specializing in the individual language or language group to be coded and named.

All reference sources should be of recent date and reliable content. For the collection of information needed to decide on the eligibility for coding, the reference sources can be in different languages. For the decision on adequate language names, reference sources shall be in the language in which an individual language or language group will be named. Detailed information on naming languages can be found in the respective clauses on language reference names (see 6.3), on language names in English and French (see 6.5), and on language names in languages other than English or French (see 6.6).

For individual languages or language groups that are not provided with many sources, consultation with domain experts can be the most reliable source of information in addition to other pertinent reference sources.

6.1.2 Information necessary to identify and name individual languages and language groups

Supporting information can be necessary to identify the individual language or language group in question and to ensure that the language names considered really do represent the individual language designated by the language identifier. Such information can include:

— range of distribution (e.g. geographical information, such as country or region of use) and size of communities of use (e.g. number of speakers) and similar statistics;

— linguistic information, such as mode of communication, writing system, status, linguistic features, etc.;

— temporal information, such as development phases and eras, or events in time, such as language reforms, etc.;

— cultural information, such as the use in songs, theatre, poetry, artistic objects, etc.;

— reference material information, such as number of sources, occurrence in mass communication media, etc.;

— existence of alternative communication modes, such as a signed modality (of an individual language which is not an individual sign language), braille, or augmentative and alternative communication (AAC).

Information necessary to determine appropriate language names can further include consideration of:

— the indication of official language names;

— the existence of alternate language names;

— names of related dialects;

— the existence of derogatory or outdated language names, which shall be avoided.

The language code does not aim to provide the above-listed kinds of information exhaustively, nor for all individual languages or language groups. Such information is important as background material for identifying and naming individual languages or language groups. It is collected and stored by the Language Coding Agencies (LCA) of the ISO 639/MA (see Annex B), as it is required for handling Change Requests for new or modifications to existing language code elements.

In some cases, a minimum of disambiguating information based on the above-listed background material shall be included in the language reference name where the same language name is used for different individual languages or language groups, or to specify other characteristics of an individual language or language group. Examples for such cases are given in 6.4.

6.2   Eligibility for language coding

6.2.1   Eligibility of an individual language for language coding

All individual languages that are full languages are candidates for language coding in the ISO 369 language code. However, judgements as to what is considered an individual language may differ. The distinction between an individual language and a dialect is based on the following criteria:

a)   Two related language varieties are normally considered to belong to the same individual language if speakers of each language variety have inherent understanding of the other language variety at a functional level (i.e. they can understand each other based on knowledge of their own language variety without needing to learn the other language variety). Where such mutual intelligibility does not exist, the two language varieties are generally seen to belong to different individual languages.

b)   Where spoken intelligibility between language varieties is marginal, the existence of a common literature or of a common ethnolinguistic identity with a central language variety that both speaker communities understand is a strong indicator that they should nevertheless be considered language varieties of the same individual language.

c)   Where there is enough intelligibility between language varieties to enable communication, they can nevertheless be treated as different individual languages when they have long-standing, distinctly named ethnolinguistic identities coupled with established linguistic normalization and literatures that are distinct.

Generally, this document takes a conservative approach, avoiding the proliferation of new language code elements. In particular, the establishment of additional language code elements for language varieties that are mutually intelligible with another already coded individual language requires significant justification, demonstrating that the criteria in c) above are clearly fulfilled.

To be considered a full language eligible for coding in this document, the individual language shall meet the following criteria:

— It is used in a broad range of domains.

— It is used to support communication across all social segments of a language community, including all genders and age groups.

— It is stable enough to have at least one variety that is widely understood across its whole geographical area over the time span of more than one generation.

— It is changing gradually as each new generation of speakers makes it their own.

The above differentiating criteria are intended to allow for the requirements of a broad range of applications. The criteria are worded to leave room for flexibility. Previous editions of ISO 639 and many language code elements were created before these criteria were established. Thus, some individual languages currently represented in the ISO 639 language code are not necessarily eligible as distinct, full individual languages under the criteria set forth in this document. Such legacy cases are maintained for reasons of language code stability.

6.2.2 Eligibility of an extinct language or historical language for language coding

Extinct languages or historical languages shall have an attested literature or be well-documented as an individual language known to have been used by some particular language community at some point in history in order to qualify for inclusion in the ISO 639 language code. Moreover, the individual language shall be treated as distinct from others by domain experts.

No reconstructed language inferred from historical-comparative analysis shall be categorized as an extinct language or historical language.

6.2.3 Eligibility of a constructed language for language coding

For constructed languages, the following guidelines are used to help determine whether a particular constructed language qualifies as a full language:

— The grammar and lexicon are complete enough for the constructed language to be fully functional (i.e. a grammar covering all areas of language structure and a lexicon of several thousand items).

— The constructed language has a substantial and growing text corpus (whether printed or electronic).

— The constructed language is "owned" (and further developed) by a vigorous user community rather than by its inventor.

— The significance of the constructed language is confirmed by non-trivial coverage in reliable, neutral sources (see also the list of pertinent reference sources in 6.1.1).

6.2.4 Eligibility of a macrolanguage for language coding

To qualify as a macrolanguage, a proposed macrolanguage code element shall meet the following criteria:

— The encompassed individual languages shall be closely related.

— Some applications require that a language be treated like an individual language in public data interchange.

— Other applications require that the same language be subdivided into two or more individual languages in public data interchange.

— The relationship between the macrolanguage code element and the encompassed individual-language code elements are explicitly defined, i.e. there is a defined mapping from the macrolanguage code element to two or more specific individual-language code elements.

— A macrolanguage code element shall not encompass another macrolanguage or language group.

— The creation of new macrolanguage code elements or the widening of an existing individual-language code element to the scope of a macrolanguage shall be based on evidence of application requirements for public data interchange, as well as appropriate linguistic relationships between the individual languages involved.

6.2.5 Eligibility of a language group for language coding

Not all language groups qualify for language coding. The creation of new language-group code elements shall be based on evidence of application requirements for public interchange of language metadata and checked for whether they are likely to be used for data interchange in a broad range of applications.

6.2.6 Criteria for the eligibility of individual languages for language coding in Set 2

6.2.6.1 General

The primary users of language identifier Set 2 include libraries, archives and other documentation organizations or entities. Consequently, the inclusion of a language identifier in Set 2 requires a significant body of literature in that individual language and about that individual language. In order to assess eligibility for Set 2, the criteria given in 6.2.6.2 to 6.2.6.6 are considered necessary.

Material supporting these criteria can be taken from the pertinent reference sources and from the information necessary to identify and name individual languages (see 6.1).

A newly identified individual language that meets the respective eligibility criteria in 6.2.1 to 6.2.4 is typically assigned a language identifier in Set 3. If the processing of a Change Request reveals that the individual language also meets the criteria for Set 2, it can additionally be added to Set 2.

6.2.6.2 Number of documents

The request for a new language identifier shall include evidence that there are at least 50 different documents in the individual language. These documents may be held by one or multiple organizations. Moreover, they shall not be variants or component parts of the same work. Documents include all forms of material and are not limited to text. This is a necessary requirement, but not sufficient in and of itself. In addition, the following requirements will be considered.

6.2.6.3 Size and variety of literature

The size and variety of the literature in the individual language (whether it has the written modality, spoken modality or signed modality as its basic modality) will be considered, and should be well-documented in the form of reference to library holdings or bibliographies or through more general statements quantifying the literature and its variation.

6.2.6.4 National or regional support

Assignment of a language identifier is explicitly supported by a national or regional language authority or standardization body. If such support for some reason is unobtainable, a recommendation from another authority or language organization can be taken into account.

6.2.6.5 Formal or official status

The individual language in question has some sort of "official" status. However, the assignment of formal status to individual languages is in no way consistently practised throughout the world, and the lack of such status is not a negative argument if other requirements are met.

6.2.6.6 Formal education

The individual language is used as a means of instruction in formal education on any level. Teaching of the individual language is also relevant, in particular if the teaching is extensive.

6.3 Principles governing the assignment of language reference names

Once a decision has been made on the eligibility of an individual language or language group for language coding, an appropriate language reference name shall be assigned. The assignment of a

language reference name for an individual language or a language group is a complex procedure and involves the selection of a language name from pertinent reference sources (see 6.1.1) and, if required, its adaptation according to the rules in 6.4.

The assignment of a language reference name is based on the following principles:

— A language reference name refers to an identified individual language or language group. It shall represent exactly the same identified individual language or language group as designated by the language identifier(s).

— No derogatory language name shall be used as a language reference name.

— Any outdated language name should be avoided as a language reference name.

— The assigned language reference name shall be unambiguous. If necessary, disambiguating information or differentiating information shall be integrated, as outlined in 6.4.2.

— If two or more suitable language names exist for an individual language or language group, the most widely used language name in pertinent reference sources should be selected.

— For amply documented individual languages, the most widely used language name in the reference sources is typically selected as the language reference name.

— For individual languages that are not provided with many sources, the language reference name should be selected upon consultation with domain experts.

— If two (or more) different language names are used to the same degree and one of them is primarily known in scientific communities and the other is more generally used, the generally used language name should be selected.

For the purpose of language coding, language reference names follow conventions in English reference sources (see 6.1.1).

The above similarly applies to individual sign languages, as well as to language groups.

6.4 Formation, disambiguation and presentation of language reference names

6.4.1 Rules for the formation of language reference names

The unique language reference name for each individual language or language group shall be formed to facilitate searchability, sortability and interoperability of language code elements in data exchange.

In addition to the above assignment principles (see 6.3), the following rules apply to the formation of language reference names:

— They shall start with an uppercase letter.

— They shall be composed using the 26 letters (a to z) of the Latin alphabet.

— They may contain SPACE (U+0020), HYPHEN-MINUS (U+002D), LEFT PARENTHESIS (U+0028) or RIGHT PARENTHESIS (U+0029); within parentheses, they can also include FULL STOP (U+002E) and digits (DIGIT ZERO U+0030 to DIGIT NINE U+0039).

— They shall not contain superscript, subscript or small capitals.

— They may contain APOSTROPHE (U+0027) and the following diacritics already used in established language reference names: á â à ä ã å ç é ê è ë í î ì ï ó ô ò ö õ ú û ù ü ñ in upper or lower case.

— They shall not be in inverted form.

EXAMPLE 1   "Middle English" and not "English, Middle".

Other elements in multiword language reference names can also start with an uppercase letter, if considered as proper names. This also applies to disambiguating information or differentiating information in parentheses as part of a language reference name.

EXAMPLE 2   North Caucasian languages, Nepalese Sign Language, Antigua and Barbuda Creole English, Middle Korean (10th-16th cent.), Official Aramaic (700–300 BCE), Konkani (macrolanguage), Mor (Bomberai Peninsula).

6.4.2   Rules for disambiguating or differentiating language reference names

6.4.2.1   General

Apart from these rules, there can be cases where some disambiguating information or differentiating information needs to be included in the name of an individual language or a language group in order to make it suitable as a language reference name. This kind of information is part of a language reference name and is expressed in English as a meta-language.

Disambiguating and differentiating are typically achieved through the addition of one of the following types of information:

— geographical information (see 6.4.2.2):

— compass direction or geographical area (prefixed in English);

— country or region name (suffixed in English in parentheses);

— time period information (see 6.4.2.3):

— epoch of an individual language (prefixed in English);

— time span or year range (suffixed in English in parentheses);

— ethnic community or language family information (suffixed in English in parentheses) (see 6.4.2.4);

— indication of the code element scope of a macrolanguage or an individual language (suffixed in English in parentheses) (see 6.4.2.5).

6.4.2.2   Geographical information

In many cases, a language name can be differentiated by prefixing the compass direction or the indication of a geographical area:

— compass direction (such as Northern/North, Eastern/East, Southern/South, Western/West and combinations thereof);

EXAMPLE 1   Southern Tiwa, Eastern Abnaki, Northwestern Dinka, West Central Banda.

— geographical area (such as town/city, region or country).

EXAMPLE 2   São Paulo Kaingáng, Scottish Gaelic, Egyptian Arabic.

If a language name designates two or more different individual languages that exist in the same or different geographical areas, the following geographical information drawn from the pertinent reference sources can be suffixed in parentheses as disambiguating information:

— Country names shall be taken from the list of English short names in ISO 3166-1 (suffixed in parentheses).

EXAMPLE 3   "Ainu (China)" for the Ainu language of the Turkic family that is spoken in China, and "Ainu (Japan)" for the unrelated Ainu language spoken in Japan.

— Region names to be suffixed in parentheses may be taken from the list of English first-level administrative subdivisions in ISO 3166-2 (only including letters and characters that are allowed

in language reference names according to 6.4.1). If the region name contains letters and characters that are not allowed in language reference names according to 6.4.1, these shall be replaced with letters and characters that are allowed.

EXAMPLE 4   "Gimi (Eastern Highlands)" and "Gimi (West New Britain)", "Kango (Bas-Uélé District)" and "Kango (Tshopo District)", "Mari (Madang Province)" and "Mari (East Sepik Province)".

Where the short name includes the governmental form of the country, e.g. "Republic of …", the governmental form may be omitted for shortness, provided the country name remains unambiguous. Thus, "Bolivarian Republic of Venezuela" may be reduced to "Venezuela", whereas the "Republic of the Congo" and the "Democratic Republic of the Congo" need to retain their full names.

6.4.2.3   Time period information

Where adequate, disambiguating information on a time period (such as an epoch of an individual language, or a year range or time span) can be suffixed in parentheses or prefixed:

— an indication of a year range or time span is suffixed in parentheses;

EXAMPLE 1   Middle French (ca. 1400 to 1600), Occitan (post 1500), Ancient Greek (to 1453), Official Aramaic (700–300 BCE), Middle Korean (10th-16th cent.).

— an indication of an epoch of an individual language (such as Ancient, Old, Middle, Modern, Ottoman) is prefixed.

EXAMPLE 2   Middle English, Ancient Greek, Ottoman Turkish.

6.4.2.4   Ethnic community or language family information

The following disambiguating information on an ethnic community or language family can be suffixed in parentheses:

— name of an ethnic community;

EXAMPLE 1   Tonga (Nyasa).

— indication of a language family.

EXAMPLE 2   Armenian (family).

6.4.2.5   Indication of the code element scope of a macrolanguage or an individual language

The following disambiguating information on macrolanguage or individual language can be suffixed in parentheses:

— indication of a macrolanguage;

EXAMPLE 1   "Konkani (macrolanguage)" ("kok").

— indication of an individual language.

EXAMPLE 2   "Konkani (individual language)" ("knn").

To keep language reference names as concise as possible, parentheses may include one unit of disambiguating or differentiating information only. The information unit selected depends on which of the above-listed information types is deemed to best disambiguate or differentiate the respective language name.

EXAMPLE 3   "Tonga (Nyasa)" and not "Tonga (Nyasa, individual language)".

6.4.3 Rules for the presentation of language reference names

Any presentation of the ISO 639 language code or subsets thereof shall show the language reference names in the language code elements in a form that makes them discernible from regular language names.

6.5 Rules for the selection, formation, and presentation of language names in English and French

The language code data accessible through the ISO 639/MA provide language names in English and French. While the language reference name fulfils a cross-lingual function in language coding and public data interchange, language names in English or French do not fulfil such a function. Nevertheless, the language names in English and French used in the ISO 639 language code are normative, but not immutable. Any modification requires completion of a process at the ISO 639/MA in reaction to Change Requests. The selection of language names in English or French for an already coded individual language or language group does not aim at completeness, but at the selection of the most appropriate language name(s) for the given individual language or language group.

More than one language name, in English or French respectively, can be deemed necessary if, for instance, one individual language is spoken across country borders and referred to with different language names by the respective parts of the language community.

EXAMPLE   Chewa or Nyanja, represented by the language identifier "nya", is referred to with a different name in Malawi and Zambia, partly with different historical and cultural development[2]. In Malawi, this individual language is called "Chewa" or "Chichewa" and is one of the two official languages. In Zambia, it is called "Nyanja" or "Chinyanja" ("language of the lake") in different areas and is one of the linguae francae with minority language status.

The selection, formation and presentation of language names in English or French follow the principles and rules for language reference names in 6.1, 6.3 and 6.4, with the following exceptions and amendments:

— Pertinent reference sources for English language names are in English. Pertinent reference sources for French language names are in French.

— Important disambiguating or differentiating information in language reference names does not need to be considered in the respective language names in English or French.

— Inverted forms of a language name are acceptable (common with language names in French).

— If a language name (even one that is widely used) is considered derogatory or offensive by parts of its language community, it shall not be presented.

— Information included in the French language names for disambiguating or differentiating is written in French as a meta-language.

The ISO 639/MA maintains documentation of changes and decisions concerning the selection of language names in English and French.

6.6 Recommendations for the selection, formation, and presentation of language names in languages other than English and French

If the language code or a subset thereof is adopted by an ISO member body using a language other than English or French, the language identifiers (see 5.4) and language reference names (see 5.5) of the ISO 639 language code shall be adopted in unmodified form. If language names in English or French are also adopted, they shall be adopted in unmodified form.

---

2)   This individual language is also spoken in other countries and is partly named differently there. To illustrate a use case for the selection and presentation of more than one English or French language names in a language code element, the example was simplified.

Additionally, one or more language names in each language of interest to the adopting ISO member body may be selected and assigned. The selection and assignment of these language names should follow the general principles and rules set out for English or French in this document (see 6.5) as much as possible.

Relevant aspects and reasons for the selection and assignment of particular language names shall be documented and maintained for future reference.

In addition to the above, it is recommended to apply the following principles and criteria:

— To determine suitable language names in the target language(s) of the adoption, a pool of pertinent reference sources in that language should be set up and investigated, as well as domain experts consulted, as outlined in 6.1.

— The principles for the assignment of language reference names given in 6.3 should be used as a guideline adapted to the target language of adoption. In particular, derogatory language names shall not be assigned or presented.

— The forms of the language names in languages other than English or French should follow the rules of the respective language(s). If applicable, the formation rules given in 6.4.1 should be applied correspondingly. If language names are provided in an individual language other than English or French, they should be written in the script(s) of that individual language. Language names provided in electronic form should be encoded using the UTF-8 encoding form of ISO/IEC 10646.

— The form of information for disambiguating or differentiating language reference names given in 6.4.2 can be modified, if necessary, in the target language of adoption. Information, whether prefixed or suffixed, to be included in the language name of the target language of adoption can be in that target language as a meta-language.

— Inverted forms of language names are acceptable.

— The number of language names assigned can differ from the number of language names in English or French. Adding new or alternate language names in the target language of adoption should be based on careful investigation and thoroughly justified.

— If language code elements are adopted into other standards, the rules in accordance with ISO 10241-2 for the adoption of terminological entries should be followed.

— ISO member bodies may decide to indicate language names by categories, such as "officially used" (possibly combined with the identifier of a country or geographical area), "preferred", "admitted" or "recognized" (possibly combined with the identifier of a country or geographical area or the identifier of the respective organization).

Language names in languages other than English and French are not immutable unless decided otherwise by the body responsible for the adoption of the language code or subsets thereof. Likewise, it can be decided that one or more language names will be considered normative in the target language of the adoption.

7 Fields of application

The language code may be implemented in a variety of applications. It is expected that few or none of these applications will utilize all sets of the language identifiers within the whole language code. Implementations that claim to conform to this document shall indicate which set of language identifiers is used. Implementations shall consistently use the chosen set of language identifiers. Mixing the two-letter language identifiers of Set 1 with the three-letter language identifiers of Set 2 or Set 3 in the same system or system environment is not recommended. If an application does mix two-letter and three-letter language identifiers, it should only use three-letter language identifiers either from Set 2 or from Set 3 for those identified individual languages or language groups that do not have a two-letter language identifier assigned to them in Set 1.