ROLE OF INDIAN LANGUAGE TAGS IN Web Standards

Swarn Lata , Director and Head , TDIL Programe , Dept. of Information technology
Somnath Chandra , Joint Director , TDIL Programe , Dept. of Information technology
Mudita Goswami , Project Engineer , C-DAC Pune , GIST Group

Abstract

This paper is based on BCP 47, ISO 15924, ISO 3166 and provides mechanism for identifying, selecting and using Language Tags and Language contents to process information on web by using web technologies. It describes how document formats, specifications, needs and implementations should handle language tags as specified by BCP 47, as well as data structures that extend these tags to describe international preferences. ISO 639 is the set of international standards that lists short codes for language names. This document defines the representation and Validation of the Language Tags

Need for Language Tags

Language Information can be used by the Applications to deliver to users the most appropriate information. The more content is tagged and tagged correctly through right Tags. It is useful for accessibility, authoring tools, translation tools, font selection, page rendering, search, and scripting and is used by screen readers and accessibility as these applications are interested in output produced by them in different language mode. For example, a browser can use language information. A page encoded in Unicode, Bengali and Assamese s may share the same code points for some characters.

Language tags also allow applying appropriate stylistic variations that are defined in a style sheet. It is also helpful for language-specific processing. Alternatively, the XSLT lang() function it is possible to extract language-specific text from a file.

Applications, protocols or specifications that use language tags are often faced with the problem of identifying sets of content that share certain language attributes.

- (XML- To spread data into web)
- Mobile Applications (To view applications in different languages)
- Web (To surf different sites of different languages)
- MAT Machine Aided Translation (To Translate data online)
- Morphological Parsing
- Speech Recognition/ synthesis
- Language-specific Typographic Behavior

We should maintain the languages tags in a way that system identifies these tags. Context Negotiation, OS based Negotiation, Server based Negotiation, Text to Speech (Spoken Form), Film Industry (Sub Titling of Films), and Hacker can also use it.

ISO 639-1, ISO 639-2 and ISO 639-3

ISO 639 is the set of international standards that lists short codes for language names. It provides two set of three-letter alphabetic codes for the representation of names of languages. ISO 639 consists of different parts, of which two parts have been approved and a third part that is in the final approval stage. The other parts are works in progress. The ISO 639 code is divided into two parts i.e. Two-character code (639-1) Example: **hi for Hindi, kn for kannada** and Three character code (639-2, 639-3,639-5) Example: **mar for Marathi, san for Sanskrit** Used in ISO 639-1.

Alpha-2: "Alpha-2" codes which are composed of 2 letters of the basic Latin alphabet. Alpha-2 is used by ISO 639-1. When code for wide range of Language required, it could cover maximum 26^2=676.To overcome from this problem, ALPHA 3 was developed.

Alpha-3: ISO 639-2 is the alpha-3 code. These two codes are provided (21 languages total), the bibliographic code is given first and the terminology code is given second. ISO 639-1 is the alpha-2 code. Multiple codes for the same language are to be considered synonyms. It is composed of Three-letters and is used in ISO 639-2, ISO 639-3, and ISO 639-5. It could cover

maximum 26^3=17,576. However, it requires some coordination between larger Systems.

In Alpha-3 the codes assigned to ISO 639-2 includes Four-special codes: Mis indicate Unicode Languages used in ISO 639, Mul indicates Multiple Language which should be applied when several languages are used & it is not practical to specify all the appropriate. Und indicates Undetermined, the Languages Exist but can't be identified yet. Zxx Languages that have no Linguistic Content, so they are not applicable and Qaa-qtz is reserved for local use (20X26= 520 codes) Including for Local treatment of dialects and it may not be interchangeable. The Bibliographic & Terminological code (B/T) in ISO 639-2 includes 23 double entries Like Dogri (doi, dgo). So, total codes is 520+23+4= 547 Codes. These codes are specifically used by ISO 639-2. Total codes available for ISO 639-2 and above are 17,576 codes. So, after deducting 547 codes from 17,576 codes, we only have 17, 029 codes, which can be used by ISO 639-3 and ISO 639-5. In World, 6 or 7 thousand Languages exists. 17,029 are adequate for the future use.

Alpha -4: "Alpha-4" codes which are composed of 4 letters of the basic Latin alphabet is proposed to be used in ISO 639-6. Mathematically, the upper limit for the number of languages and dialects that can be so represented is 26^4 = 456,976.

ISO 639-1 [2002]: Codes for the representation of names of languages: alpha-2 codes were developed by the ISO TC37/SC2 in 1988 for use in terminology, lexicography and linguistics.

ISO 639-2 [2006]: Alpha-3 code was devised primarily for use in bibliographic documentation and terminology. It includes identifiers for all of the languages represented in part 1, as well as for many other languages that have significant bodies of literature. It also provides identifiers for groups of languages, such as language families, that together indirectly cover most or all languages of the world. The maintenance agency for ISO 639-2 is the Library of Congress.

ISO 639-3 [working from 2007]: ISO 639-3 is an international standard for language codes. The standard describes three-letter codes for identifying languages. It extends the ISO 639-2 alpha-3 codes with an aim to cover all known natural languages. It is a superset of ISO 639-1 and of the individual languages. The Final standard contains 7589 entries. It includes individual languages contained in 639-2, modern languages, historic varieties, ancient languages and artificial languages.

Table of Language Tags in ISO 6391, 2 and 3

S.NO	639-1	639-2	639-3	Language Name
1.	as	asm	asm	**Assamese**
2.	-	sit	brx	**Bodo**
3.	bn	ben	ben	**Bengali**
4.	-	-	dgo	**Dogri**
5.	gu	guj	guj	**Gujarati**
6.	kn	kan	kan	**Kannada**
7.	ks	kas	kas	**Kashmiri**
8.	mr	mar	mar	**Marathi**
9.	ne	nep	nep	**Nepali**
10.	or	ori	ori	**Oriya**
11.	pa	pan	pan	**Panjabi**
12.	sd	snd	snd	**Sindhi**
13.	ta	tam	tam	**Tamil**
14.	te	tel	tel	**Telugu**
15.	ur	urd	urd	**Urdu**
16.	sa	san	san	**Sanskrit**
17.	-	kok	Gom (kok)	**Konkani**
18.	-	mai	mai	**Maithili**
19.	-	mni	mni	**Manipuri/ Meitei**
20.	-	sat	sat	**Santhali**
21.	hi	hin	hin	**Hindi**
22.	ml	mal	mal	**Malayalam**

Concept of Macro Languages and Collective Languages

Despite the changes in how language tags are formed and maintained, a few cases remain which the new design does not fully address. Problem is that of identifying variations of a language or within family of languages. While variant or region subtags are often useful for this purpose. Two new terms were introduced in ISO 639-2 called Macro languages and Collective languages. The term Collective language is not continued in ISO 639-3. Macro language is still working. There are about 1652 Languages

spoken in India, in which 350 are major languages. In which, there are 56 languages in total associated with ISO 639-2 which are considered to be macro languages in 639-3. 17 dialects converted into Indian individual language and some languages that were considered by ISO 639-2 to be dialects of one language are now in ISO 639-3 in certain contexts considered to be individual languages themselves. Examples: Haryanvi, Awadhi- which were dialect of Hindi now considered being individual language. Example of Macro language : **kok:** is the ISO 639-3 language code for Konkani macro language. There are 2 individual language codes assigned: **gom** — Goan Konkani. It has official recognition as an independent language and **knn** — Konkani language. It has no official recognition, treated by some linguists as a dialect of Marathi. Both languages are referred to as Konkani by their respective speakers. **Collective languages:** A collective language code element is an identifier that represents a group of individual languages that are not deemed to be one language in any usage context. Whereas ISO 639-2 includes three-letter identifiers for such collections of languages, this part of ISO 639 provides identifiers for individual languages and macro languages only. Collective languages and their ISO 639-2 codes are: **bih** - Bihari. This Language also has an ISO 639-1 code and **him** – Himachali.

BCP 47

A Best Current Practice (BCP) means that a certain manner of proceeding is in general the most logical choice a de facto standard of sorts. BCP is the name published by the Internet Engineering Task Force. It describes the structure, content, construction, and semantics of language tags for use in cases where it is desirable to indicate the language used in an information object. The language tag is composed of one or more parts, known as "subtags". Each subtag consists of a sequence of alphanumeric characters. BCP 47 can be used in no. of modern standards such as http, html, xml, png. Each language tag is composed of one or more “subtags” separated by hyphens. With the exception of private use language tags and grandfathered language tags, the subtags occur in the following order: a language subtag (potentially followed by up to three extended language subtags)- an optional script subtag, an optional region subtag, optional variant subtags, optional extension subtags, optional private use subtags. Language subtags are mainly derived from ISO 639-1 and ISO 639-2, script subtags from ISO 15924, and region subtags from ISO 3166-1 alpha-2 and UN M.49. Variant subtags are not derived from any standard. No extension subtags have yet been defined. The Language Subtag Registry, maintained by IANA, lists the current valid public subtags. Language tags consist of only language subtag, or a language subtag and a region subtag. For example, bn represents Bengali, and consists of a single language subtag (from ISO 639-1), while bn-IN represents Indian Bengali, and consists of the language subtag bn followed by the region subtag IN (from ISO 3166-1). Subtags are not case sensitive, but the specification recommends using the same case as in the Language Subtag Registry, where region subtags are uppercase, script subtags are titlecase and all other subtags are lowercase. Syntax will be:

Language-Tag = langtag
/ privateuse ; private use tag
/ grandfathered ; grandfathered registrations
langtag = (language ["-" script] ["-" region] *("-" variant) *("-" extension) ["-" privateuse])

For Example, bn language code is used for Bengali language; Beng Script code is used for Bengali Script and IN is used as Region Code for India. So, to represent the whole Language tag we can declare it as

bn-Beng-IN (Bengali language uses Bengali Script in India.)

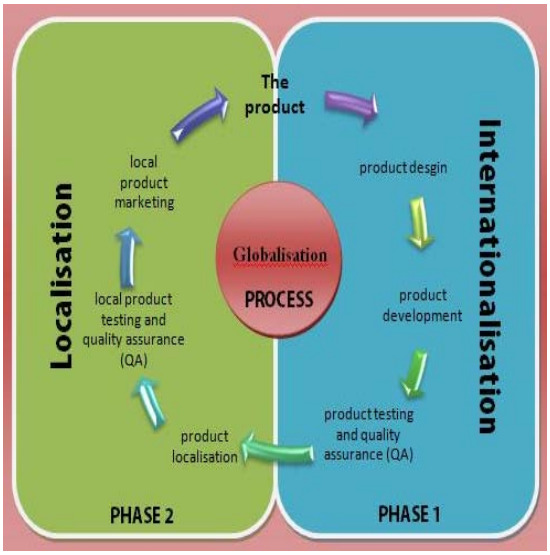
Language Tags and Localization

A locale is an id that refers to a set of user preferences that leads to globalization. Usually, the data associated with locale provides support for formatting and parsing of dates, times, numbers, and currencies; for measurement units, for sort-order, Translation for time zones, languages, countries, and scripts, also text boundaries (character, word, line, and sentence), text transformations (including transliterations), and supports other services. It’s abbreviated from L10N, being the 10 a way of replacing the middle 10 letters of the word) can be defined as the second phase of a larger process. There is very close relationship

between Localization and Language Tags as both are not complete without each other. Language Tags defines which language a user want to display or in which language a user want to see the output in the browser. Localization defines the Linguistic and cultural background of the Language including dates, times, numbers, and currencies, measurement units and time zones. The Locale data may change over time due to so many conditions like committees impose new standards, currency change.

Internationalisation

Languages Tags are the part of Internationalisation. It is the design and development of a product, application or document content that enables easy localization for target audiences that vary in culture, region, or language. Internationalisation in includes Characters, Language, Markup & text Text direction, Styling & layout, Forms, Navigation, Resource identifiers, Cultural issues, Troubleshooting and Miscellaneous. Internationalization is often written "i18n", where 18 is the number of letters between 'i' and 'n' in the English word.
The best practice for internationalization is to store and communicate language-neutral data, and format that data for the client. This formatting can take place on any of a number of the components in a system. Examples of Items to be formatted are: Text messages, GUI labels, Dates, Times, Numbers, Measurements, Icons, Currencies, Phone number, Honorifics, personal titles, Postal address, Page layouts, Morphological Meta data



The Internationalisation and Localization are the part of the Globalization Process. Both Internationalisation and Localization depend upon each other.

Relationship between ISO 639-1, 639-2 and 639-3

ISO 639 provides three language codes: one is a two-letter code (ISO 639-1) and two others are three-letter codes (ISO 639-2 and ISO 639-3) for the representation of names of languages. The three-letter codes in ISO 639-2 and ISO 639-3 are complementary and compatible. ISO 639-1 based on Alpha-2 and uses two character codes. ISO 639-2 covers a larger number of individual languages (due to its less-restrictive scope). It also includes identifiers for collections of languages. However, ISO 639-1 includes very less number of individual languages. ISO 639-3 cover almost all Individual languages of the world. Overall, the set of individual languages listed in ISO 639-3 is much larger than the set of individual languages listed in ISO 639-2. ISO 639-1 and ISO 639-2 are intended to focus on the major languages of the world that are most frequently represented in the total body of the world's literature, ISO 639-3 attempts to provide as complete an enumeration of languages as possible, including living, extinct, ancient and constructed languages, whether major or minor. As a result, ISO 639-3 lists a very large number of lesser-known languages.
ISO 639-1 was devised primarily for use in terminology, lexicography and linguistics. ISO

639-2 was devised primarily for use in terminology and bibliography. ISO 639-2 covers both (B/T codes), however ISO 639-1 covers only Terminology codes. ISO 639-3 is the superset of ISO 639-1. It enhances and provides more information originally based ISO 639-1. ISO 639-1 includes only 676 codes, which are very less in number. ISO 639-2 could cover maximum 26^3=17,576 only for ISO 639-2 and rest for ISO 639-3 and higher versions.

The alpha-3 codes for ISO 639-2 and ISO 639-3 overlap. The collective language code elements in ISO 639-2 have no counterpart in ISO 639-3, and there are many individual language code elements in ISO 639-3 that have no counterpart in ISO 639-2.

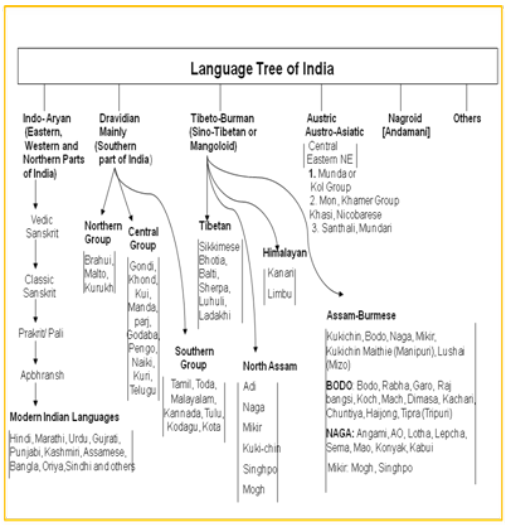
Some existing code elements in ISO 639-2, and the corresponding code elements in ISO 639-1, are designated in those parts of ISO 639 as individual language code elements, yet are in a one-to-many relationship with individual language code elements in this part of ISO 639. For purposes of this part of ISO 639, they are considered to be macro language code elements.

Relationship between ISO 639 and BCP 47

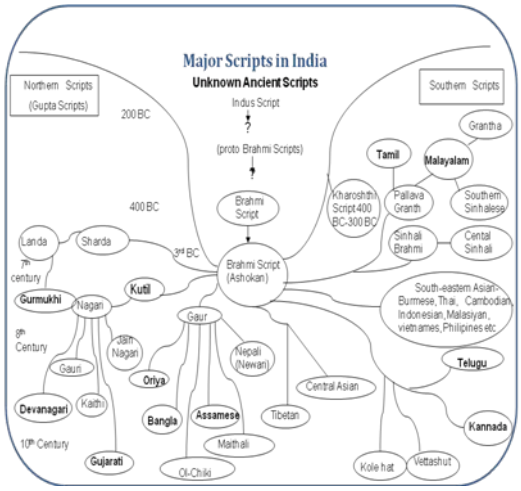
BCP 47, describes the structure, content, construction, and semantics of language tags for use in cases where it is desirable to indicate the language used in an information object. It also describes how to register values for use in language tags and the creation of user-defined extensions for private interchange.

It is considered an Internet Best Current Practices for the Internet Community and gives guidance for the use of ISO 639 codes. It specifies use of a 2-character code from ISO 639-1 when it exists; when a language does not have a 2-character code assigned the 3-character code is used.

Languages Position in India

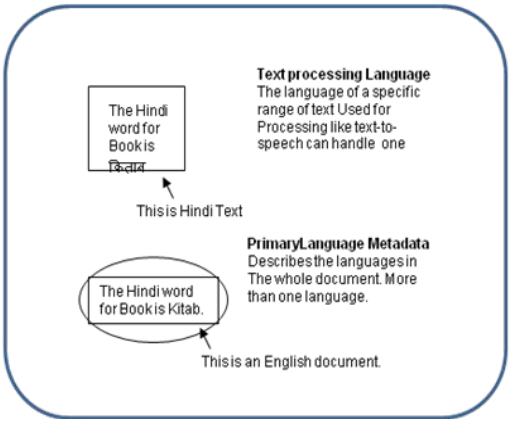


The dialects of Indian Scheduled Languages are: ASSAMESE; BENGALI: (i) Bengali, (ii) Chakma, (iii) Hajong, (iv) Rajbangsi; BODO:(i) Bodo; DOGRI:(i) Dogri; GUJARATI: (i) Gujarati,(ii) Gujrau,(iii) Saurashtri; HINDI: (i) Awadhi, (ii) Bagheli, (iii) Bagri Rajasthani, (iv) Banjari, (v) Bhadrawahi, (vi) Bharmauri, (vii) Bhojpuri, (viii) Brajbhasha, (ix) Bundeli, (x) Chambeali, (xi) Chhattisgarhi, (xii) Churahi, (xiii) Dhundhari, (xiv) Garhwali, (xv) Gojri, (xvi) Harauti, (xvii) Haryanvi, (xviii) Hindi, (xix) Jaunsari, (xx) Kangri, (xxi) Khairari, (xxii) Khari Boli, (xxiii) Khortha, (xxiv) Kulvi, (xxv) Kumauni, (xxvi) Kurmali Thar, (xxvii) Labani, (xxviii) Lamani, (xxix) Laria, (xxx) Lodhi, (xxxi) Magahi, (xxxii) Malvi, (xxxiii) Mandeali, (xxxiv) Marwari, (xxxv) Mewari, (xxxvi) Mewati, (xxxvii) Nagpuria, (xxxviii) Nimadi, (xxxix) Pahari, (xxxx) Panch Pargania, (xxxxi) Pangwali, (xxxxii) Powari, (xxxxiii) Rajasthani, (xxxxiv) Sadri, (xxxxv) Sirmauri, (xxxxvi) Sondwari, (xxxxvii) Sugali, (xxxxviii) Surgujia, (xxxxix) Surjapuri; KANNADA: (i) Badaga, (ii) Kannada, (iii) Kuruba; KASHMIRI: (i) Kashmiri, (ii) Kishtwari, (iii) Siraji; KONKANI: (i) Konkani, (ii) Kudumbi, (iii) Malwani; MAITHILI:(i) Maithili; MALAYALAM: (i) Malayalam, (ii) Yerava; MANIPURI: (i) Manipuri; MARATHI: (i) Marathi; NEPALI: (i) Nepali; ORIYA: (i) Bhatri, (ii) Oriya, (iii) Proja, (iv) Relli,(v) Sambalpuri; PUNJABI:(i) Bagri (Punjabi),(ii) Kahluri,(iii) Punjabi; SANSKRIT:(i) Sanskrit; SANTALI: (i) Karmali, (ii) Santali; SINDHI: (i) Kachchhi, (ii) Sindhi; TAMIL: (i) Kaikadi, (ii) Tamil, (iii) Yerukala; TELUGU: (i) Telugu, (ii) Vadari; URDU: (i) Urdu. Manipuri includes Meithei and Nepali includes Gorkhali.



Languages Tags in HTML, XHTML and XML

The web is a globe-spanning, borderless nation that speaks many languages and all of them. A web page created by an Indian, hosted in India, targeted at an Indian audience, can still easily be seen by people in other Countries. The primary language of the entire document can be set by including the lang and xml:lang attributes in the document's root html element. It is quite possible to differentiate individual phrases or passages written in another language by adding the attributes to their appropriate parent element; lang and xml:lang can be validly attached to almost any element. The lang attribute comes from HTML, while xml:lang is the XML equivalent to be used in XHTML documents. There are two ways in which one might speak about the language of content:



BIDI Embedding Controls

The Unicode Standard prescribes a *memory* representation order known as logical order. When text is presented in horizontal lines, most scripts display characters from left to right. However, there are several scripts (such as Arabic or Hebrew) where the natural ordering of horizontal text in display is from right to left. Ruby may be used for different reasons:

- Because the character is rare and the pronunciation unknown to many personal name characters often fall into this category;
- Because the character has more than one pronunciation, and the context is insufficient to determine which to use;
- Because the intended readers of the text are still learning the language and are not expected to always know the pronunciation and/or meaning of a term;
- Because the author is using a nonstandard pronunciation for the characters

Also, ruby may be used to show the meaning, rather than pronunciation, of a possibly-unfamiliar (usually foreign) or slang word. This is generally used with spoken dialogue and applies only to Japanese publications.

Registry and Maintenance

A Language-Tag consists of: langtag ; generated tag -or- private-use ; a private use tag -or- grandfathered ; grandfathered registrations
langtag = (language ["-" script] ["-" region] *("-" variant) *("-" extension) ["-" privateuse])
language= "bn", "hin", or a registered value, script = "Beng", "Deva", "Gujr" ISO 15924 codes, region = "IN", "CS", "FR" ISO 3166 codes or "142", "019", or UN M.49 codes, variant = "rozaj", "nedis", "1996", multiple subtags can be used in a tag, extension = single letter followed by additional subtags; more than one extension may be used in a language tag, private-use = "x-" followed by additional subtags, as many as are required, grandfathered = tags listed in the old registry that are not otherwise redundant (a closed list)

Language tag: It can then, optionally, be followed by various subtags. Today there are five kinds of subtags that follow the language identifier: scripts, regions, variants, extensions, and private use. The order, length, and content

of each subtag type is fixed, so a tag processor can always identify exactly which type of subtag it has, even if the processor doesn't have that subtag in its copy of the registry.

Script. These are based on ISO 15924 and indicate the writing system. The script subtag can occur at most once (it may be omitted) and must appear directly after the language. Some languages have a field in the registry indicating that a particular script code should be "suppressed".

Region: These are based mostly on ISO 3166-1 code and indicate the country or regional variation. The region code can also include selected UN M.49 region codes.

Variants: Variant subtags are not based on an external standard. They are all individually registered values, mostly indicating particular dialects or other language variations not covered by scripts or regions. Multiple variant subtags can be included in a tag. Each variant has fields in the registry, though, indicating which subtags it is intended for use with.

Extensions: Extensions are a mechanism whereby future additions to language tags can be standardized. Each extension has a single character subtag (a "singleton") that identifies it. Various restrictions apply to extensions and how they are formed, used, and administered. Extensions form the basis for future addition of features to language tags.

Private use: Private use subtags are not based on any standard at all. They are for use by individuals or groups that need to identify something language related that might not rise to the level of standardization. RFC 3066 included private use tags, but the whole tag was private use.

language-script- region-variant- extension-privateuse

hi-Deva-IN-bgc x-mudita diplays Hindi language is using Devanagari Script spoken in India have dialect called Haryanvi privately used by mudita.

Relationship between ISO 639 and ISO 15924

There's no definitive list associating scripts with languages (because most languages could be written with most scripts, and this actually happens when people are transliterating foreign languages).

ISO 639-2 also provides identifiers for groups of languages, such as language families, that together indirectly cover most or all languages of the world. There is a list in ISO 15924 used in Unicode 4.0 in which some Indian Scripts are included. Some more Scripts are added in Unicode 5.0 as well.

Maintenance of the Registry

Maintenance of the registry requires that as codes are assigned or withdrawn by ISO 639, ISO 15924, ISO 3166, and UN M.49, the Language Subtag Reviewer MUST evaluate each change, determine whether it conflicts with existing registry entries, and submit the information to IANA for inclusion in the registry. If a change takes place and the Language Subtag Reviewer does not do this in a timely manner, then any interested party MAY use the procedure to register the appropriate update.

Relationship between ISO 639 and ISO 3166

ISO 639 provides two and three-character codes for representing names of languages. ISO 3166 to denote the area in which a term, phrase, or language is used and provides two and three-character codes for representing names of countries. One should use codes from each list independently. However, the language code and country code may be used together to indicate a language variation spoken in a particular country as defined by BCP 47. For Example, Panjabi as spoken in the India may be indicated with the following: **pa-IN**

Relationship between Unicode and ISO 639

The Unicode Consortium does not own or maintain ISO 639 and neither bears responsibility for its contents nor has the ability to change them. Similarly, the Unicode is not responsible for the names or codes used by Microsoft or Macintosh.

INDIAN SCRIPTS IN UNICODE: Bengali, Brahmi, Devanagri, Gujarati, Gurumukhi, Kannadda, Malayalam, Oriya, Telegu, Tamil, Kharoshthi, Meitei/ Manipuri, Ol-Chiki, Lepcha, Saurashtra,Indus (Harappan), Signwritting.

INDIAN AND OTHER SCRIPTS: Sinhala – India/ SriLanka, Tibetan – India/ Tibat, Limbo -India/ Nepal. Kharoshthi is almost dead now and Brahmi has become the ancestor of all the native scripts of south Asia as well as Southeast Asian Countries. All the scripts in India maintained the same basic structure as original Brahmi.

Requested Changes to ISO 639-3 Language Code

The types of changes that are possible are to modify the reference information for an existing code element (reference name and additional names, language type. language scope or relationship to a macro language grouping), to retire a code element from use, including merging its scope of denotation into that of another code, to split an existing code element into two or more separate language code elements and to create a new code element for a previously unidentified language.

In case of Indian Language a single language code is normally provided for a language even though the language is written in more than one script. However, it is part of Prakrit. Same in case of Pali, which is also part of Prakrit, includes in all the series of ISO 639. However, Prakrit is removed in ISO 639-2. All these issues creates problem for the Indian Language user.

Since many of the Indian languages other than 22 Official languages are being encoded in the UNICODE standard such as Lepcha, Tibetan etc. the present effort of assigning Language Tags needs to be enhanced for such languages.

PROPOSED TABLE OF SCHEDULED LANGUAGE WITH TAGS

Lang. Name	Existing ISO 639-3	Pro. Lang Tags	Lang. Tags for Macro Lang	Lang. Status
BENGALI				**Scheduled (West Bengal)**
Bengali	ben	ben	bng	
Chakma	ccp	chk		
Hajong	haj	hjn		
Rajbangsi	-	rjb		
GUJARATI				**Scheduled (Gujarat)**
Gujarati	guj	gjr	gjt	
Gujrau	-	gjo		
Saurashtri	-	srs		
HINDI				**Scheduled (India)**
Awadhi	awa	awd		
Bagheli	bfy	bgh		
Bhojpuri	bho	bhj		
Brajbhasha	-	brj		
Bundeli	bns	bnd		
KASHMIRI				**Scheduled (Kashmir)**
Kashmiri	kas	kas	ksm	
Kishtwari	-	kst		
Siraji	-	srj		
MARATHI				**Scheduled (Maharashtra**
Marathi	mar	mar		
TAMIL				**Scheduled (Tamil Nadu)**
Kaikadi	kep	kkd		
Tamil	tam	tam	tml	
Yerukala	-	yrk		
TELUGU				**Scheduled (A.P)**
Telugu	tel	tlu	tlg	
Vadari	-	vdr		

Futuristic Approach for ISO 639-5 and ISO 639-6

ISO 639-5[2008]: provides a code consisting of language code elements comprising three-letter language identifiers for the representation of names of living and extinct language families and groups. Languages designed exclusively for machine use, such as computer-programming languages, are not included in this code. Alpha 4: A codes (composed of 4 letters of the basic Latin alphabet) is proposed to be used in ISO 639-6. It could cover maximum 264=456,976.

ISO 639-6[2008]: Codes for the representation of names of languages -Part 6: Alpha-4 representation for comprehensive coverage of language variation", is an international-standard being developed by ISO/TC 37/SC 2. The latest FDIS was issued in February, 2008. The data supporting ISO 639-6 is being researched and compiled by GeoLang Ltd on behalf of the British Standards Institute (BSI).

Conclusion

Language information is important for accessibility, styling, searching, and typically transmitted between the user agent and server can be used to help improve navigation for users and the localization of the content.

References

Language Localization November (2008) http://en.wikipedia.org/wiki/Language_localisation (last Access 05.01.09)
Tags for Identifying Languages http://www.rfc-editor.org/rfc/bcp/bcp47.txt (last Access 05.01.09)
Codes for the representation of names of scripts http://unicode.org/iso15924/iso15924-codes.html (last Access 05.01.09)
English country names and code elements http://www.iso.org/iso/country_codes/iso_3166_code_lists/english_country_names_and_code_elements.htm
(last Access 05.01.09)
Territory Containment (UN M.49) http://unicode.org/cldr/data/diff/supplemental/territory_containment_un_m_49.html (last Access 05.01.09)
Internationalization http://www.w3.org/International/questions/qa-when-lang-neg (last Access 05.01.09)
Brahmi http://www.ancientscripts.com/brahmi.html
(last Access 05.01.09)
Language codes http://std.dkuug.dk/jtc1/SC22/WG20/docs/n780-langcodes.txt (last Access 05.01.09)
The Unicode Character Code Charts By Script http://www.unicode.org/charts/(last Access 05.01.09)