Internationalization Tag set 2.0 requirements

for

Indian Languages

Abstract

This document defines data categories and their implementation as a set of elements and attributes called the *Internationalization Tag Set (ITS)* 2.0. ITS 2.0 is the successor of ITS 1.0; it is designed to foster the creation of multilingual Web content, focusing on HTML, XML based formats in general, and to leverage localization workflows based on the XML Localization Interchange File Format (XLIFF).

Table of Contents

1. Introduction

ITS 2.0 is a technology to add metadata to Web content, for the benefit of localization, language technologies, and internationalization. The ITS 2.0 specification both identifies concepts (such as “Translate”) that are important for internationalization and localization, and defines implementations of these concepts (termed “ITS data categories”) as a set of elements and attributes called the *Internationalization Tag Set (ITS)*.This document provides Data categories introduced in ITS 2.0 version and also presents some of the issues with respect to Indian languages.

a. Purpose of this document

This document aims to realize many of the ideas formulated in the ITS 2.0 requirements with respect to Indian languages. This document also defines the variation of Named entities in Indian languages, its requirement in ITS 2.0 and suggestion for implementation in ITS 2.0.The aim of this document is to define proposed issues in some of the data categories define in ITS 2.0 and its possible suggestions w.r.t Indian languages.

2. Relationship with ITS 1.0 & New Principle

a. Relationship with ITS 1.0

ITS 2.0 has the following relations to ITS 1.0:

- It adopts and maintains the following principles from ITS 1.0:
    - It adopts the use of data categories to define discrete units of functionality
    - It adopts the separation of data category definition from the mapping of the data category to a given content format
    - It adopts the conformance principle of ITS1.0 that an implementation only needs to implement one data category to claim conformance to ITS 2.0
- ITS 2.0 supports all ITS 1.0 data category definitions and adds new definitions ITS 2.0 adds a number of new data categories not found in ITS 1.0.
- While ITS 1.0 addressed only XML, ITS 2.0 specifies implementations of data categories in *both* XML *and* HTML.

b. New Principles

ITS 2.0 also adds the following principles and features not found in ITS 1.0:

- ITS 2.0 data categories are intended to be format neutral, with support for XML, HTML, and NIF: a data category implementation only needs to support a single content format mapping in order to support a claim of ITS 2.0 conformance.
- ITS 2.0 provides algorithms to generate NIF out of HTML or XML with ITS 2.0 metadata.
- A global implementation of ITS 2.0 requires at least the XPath version 1.0. Other versions of XPath or other query languages (e.g., CSS selectors) can be expressed via a dedicated query Language attribute.

The new data categories included in ITS 2.0 are:

- Domain
- Disambiguation
- Locale Filter
- Provenance
- External Resource
- Target Pointer
- Id Value
- Preserve Space
- Localization Quality Issue
- Localization Quality Rating
- MT Confidence
- Allowed Characters
- Storage Size

3. Relationship with XLIFF

XLIFF is the XML Localisation Interchange File Format designed by a group of software providers, localization service providers, and localization tools providers. It is intended to give any software provider a single interchange file format that can be understood by any localization provider.

The core of XLIFF 2.0 consists of the minimum set of XML elements and attributes required to (a) prepare a document that contains text extracted from one or more files for localization, (b) allow it to be completed with the translation of the

extracted text, and (c) allow the generation of translated versions of the original document.

XLIFF is a bilingual document format designed for containing text that needs translation, its corresponding translation and auxiliary data that makes the translation proccess possible.
At creation time, an XLIFF file may contain only text in source language. Translations expressed in target language may be added at a later time.

ITS 2.0 is designed to foster the creation of multilingual Web content, focusing on HTML, XML based formats in general, and to leverage localization workflows based on the XML Localization Interchange File Format (XLIFF).

A user agent could use ITS rules for converting content into XLIFF

4. Issues in ITS 2.0 principles w.r.t Indian Languages

a. Translate Data Category

The Translate data category expresses information about whether the content of an element or attribute should be translated or not. The values of this data category are "yes" (translatable) or "no" (not translatable).

Issue:

Some words in Indian languages needs to transliterate instead of translation. The following few examples shows the requirements:

- **Rasgulla (Sweet Name)**
- **Bharatnatayam (Dancing style)**
- **City names**

In the above examples both Rasgulla & Bharatnatayam are Indian words and the localization of these words in Hindi are रसगुल्ला and भरतनाट्यम respectively. So for localization of such words there is a need to transliterate it instead of translate.

Suggestion :
So there is a need to introduce some mechanism for different Named entities in Indian languages .The detailed description of Named Entity is mentioned in section 5.

b. Localization Note

The Localization Note data category is used to communicate notes to localizers about a particular item of content.
This data category can be used for several purposes, including, but not limited to:

- Tell the translator how to translate parts of the content
- Expand on the meaning or contextual usage of a specific element, such as what a variable refers to or how a string will be used in the user interface
- Clarify ambiguity and show relationships between items sufficiently to allow correct translation (e.g., in many languages it is impossible to translate the word "enabled" in isolation without knowing the gender, number and case of the thing it refers to.)
- Indicate why a piece of text is emphasized (important, sarcastic, etc.)

Issue :

Localization note clarify ambiguity and show relationships between items to allow correct translation that depends on gender, number and case of the thing refer to. But in Indian languages some particular words depends on the part of speech also. The following examples shows the requirements in Bengali language :
Bengali word (“sarala”) is pronounced in two different ways when it is verb it is pronounced as /ʃɔrolo/ (moved) and /ʃɔrol / (easy) when it behaves as adjective.
When translated the target text may be equal or up to 1.5 to 2 times the source text.
Some examples in Hindi is given below :

- **सोना - Sleep(verb), Gold(Noun)**
- **आम - Common(Adjective) , Mango(Noun)**
- **भूल - Forget(verb), Fault(Noun)**

Some examples in for English is given below :

- **Fly - उड़ना(verb),मक्खी(noun)**
- **Bark-भौंकना(verb),पेड़ की छाल(noun)**

- **close-पास(adverb),बंद(verb)**
- **March-कदम ताल करना(verb),महीने का नाम(noun)**

Context dependent translation :

Many words in English depends on the context of the Sentences with same part-of-Speech for example :

Bank : किनारा(Noun), बैंक(Noun)

Need to introduce Part of Speech tagging in XML format for correctness of translation.

Proposed POS Tag Set for Indian Languages :

The proposed POS Tagging for Indian languages is attached as Annexure I.

c. Disambiguation

The Disambiguation data category is used to highlight (mark up) specific conceptual patterns that may require special treatment when localizing and translating content. This data category can be used for several purposes, including, but not limited to:
Informing a translation service that a certain fragment of text is subject to follow specific translation rules, e.g. for proper names, or officially regulated translations, as well as to conveying a very specific meaning of the fragment.

Issue :

Some personal name in India used village/town/city/profession in his/her personals name. Some examples are given below:

- कागैडु बैरअप्पा तिमप्पा नायर (Village name + fathers name + Given name + Last name)
- Rajesh Pilot (Given Name + Occupation n
- ame (Optional))
- Kagodu Bairappa Timmappa(Town name + Name + Caste name)

So in Disambiguation data category , personal names in Indian languages may be addressed such issues.

5. Named Entity Recognition

Named Entity Recognition is a subtask of Natural Language Processing. It involves classification of a word as person-name, location, organization, time, date, etc. The categories into which proper nouns are to be classified vary according to the applications, but common categories include, person names, location and organization names.

a. Types of Named Entities with examples

There are Eleven types of entities in Name as given below.

1. **Person:** Person entities are limited to humans. A person may be a single individual or a group.
2. **Organization:** Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure.
3. **Location:** Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.
4. **Facilities:** Facility entities are limited to buildings and other permanent man-made structures and real estate improvements.
5. **Locomotives:** A locomotive entity is a physical device primarily designed to move an object from one location to another, by (for example) carrying, pulling, or pushing the transported object. Vehicle entities may or may not have their own power source.
6. **Artifacts:** Artifact entities are objects or things, which are produced or shaped by human craft, such as tools, weapons/ammunition, art paintings, clothes, ornaments, medicines.
7. **Entertainment:** Entertainment entities denote activities, which are diverting and hold human attention or interest, giving pleasure, happiness, amusement especially performance of some kind such as dance, music, sports, events.
8. **Cuisine's:** This entity refers to various type of food, prepared in different manners such as Chinese food, South-Indian, North-Indian foods.
9. **Organisms:** Organism entities are living things and have the ability to act or function independently such as humans, viruses, bacteria etc.
10. **Plants:** These entities are living things having photosynthetic, eukaryotic, multicellular organisms of the kingdom Plantae, containing chloroplasts, having cellulose cell walls, and lacking the power of locomotion.
11. **Disease:** This entity refers to the state of a disordered or incorrectly functioning organ, part, structure, or system of the body resulting from the effect of genetic or developmental errors, infection, poisons, nutritional deficiency or imbalance, toxicity, or unfavorable environmental factors; illness; sickness; ailment such as fever, cancer etc.

b. Different Personal names in Indian languages-Examples

6. North Region

Titles: Shri, Shrimati(Married women),Kumari(Unmarried women/girls)

Pattern 1: Given name + Last name

Eq. Ravi Kumar

Pattern 2: Given name + Middle Name + Last name

Eq. Ramesh Kumar Sharma

Pattern 3: Last name + Given name

eg. Kumar Gaurav

Pattern 4: Given name + Family name + Last name (used by females)

Eq. Aishwarya Rai Bacchan

Pattern4 : Given Name + Occupation name (Optional)

Eq : Rajesh Pilot

Rajasthan :

Pattern 1 :Given name + Fathers name + Sir name + Caste

Eq: Aditya Pratap Singh Chauhan

7. **South Region**

Tamil:

Title: Thiru,Thrumati,Selvi

Pattern 1 : Given name + Father's Name

Eq : Ramesh Ramaiah

Pattern 2: Given name + Fathers/Husband name (used by females)

eq. Sunitha Gopalan

Pattern 2 : Village name + fathers name + Given name + Last name

Eq. कागैडु बैरअप्पा तिमप्पा नायर

Other Southern States

Title: Shri,Shrimati ,Kumari (Commonly used)

Sri,Srimati(optional)

Pattern 1: Initials(Family name) + Given name

eq. T.S Babu

Pattern2: Given name + caste

Pattern 3: Village/town name+ Fathers Name + Given name

Eq : Kagodu Bairappa Timmappa

Pattern 4: Town name + Name + Caste name

eq. Madurai Mani Iyer

Pattern 6: Given name + House name (used by Christian)

Eq : Josheph Allendery

Andhara Pradesh :

Pattern 1: Village name + Fathers name + Family Name + Given Name

Eq. Katravalapally Nishant Raghvendra Tejaswi

Kannada :

Pattern 1 : Place name + Given name

Eq: Kadidal Manjappa

Pattern 2 : Place name+ Fathers name + Given name

Eq : Kuppalli Venkatappa Puttappa

8. **West Region**

Pattern 1: Given name + Last name

Pattern 2: Given name + Fathers Name + Family Name

Eq. Sachin Ramesh Tendulkar

Pattern 3: Given name + Middle Name + Last name

East Region

Manipur:

Pattern1: Sirname + Given name + Middle name

Pattern 2: Chnu as a middle name used by women

Bengali:

Title: Shri Shri (Optional)

Eq : Shri Shri Ramanand Baba Thakur

Pattern 2: Given name (Nick Name) + Middle Name + Last name

Eq : Sunil Kumar pal

6. Proposed tagging for Named Entity in ITS 2.0

Named entities can be identified by using tagging within the XML Document.

For Example :

Personal names :

Pattern1

```
<article xmlns="http://docbook.org/ns /docbook"
      xmlns:its="http://www.w3.org/2005/11/its"
      its:version="2.0" version="5.0" xml:lang="hin">
 <info>
  <title>Peronal name</title>
  <author its:translate="no">
```

```
 < personname type="person" subtype1="Individual" subtype2="Familyname"> T.S
< personname type="person" subtype1="Individual" subtype2="Givenname">
Babu
</ personname ></ personname >

    <affiliation>
      <address><email>foo@example.com</email></address>
    </affiliation>
  </author>
 </info>
 <para>This is a short article.</para>
</article>
```

Titles in Personal names :

```
<article xmlns="http://docbook.org/ns /docbook"
       xmlns:its="http://www.w3.org/2005/11/its"
       its:version="2.0" version="5.0" xml:lang="hin">
 <info>
  <title>Peronal name</title>
  <author its:translate="no">

 < personname type="person" subtype1="Individual">
< personname type="person" subtype1="Individual" subtype2="Title">Dr.
</ personname > Abdul Kalam </ personname >

    <affiliation>
      <address><email>foo@example.com</email></address>
    </affiliation>
  </author>
 </info>
 <para>This is a short article.</para>
</article>
```

Location:

```
<article xmlns="http://docbook.org/ns /docbook"
       xmlns:its="http://www.w3.org/2005/11/its"
       its:version="2.0" version="5.0" xml:lang="hin">
 <info>
  <title>Peronal name</title>
  <author its:translate="no">
```

```
 < personname type="person" subtype1="Individual">
< personname type="person" subtype1="Individual" subtype2="Title">Dr.
</ personname > Abdul Kalam </ personname >

    <affiliation>
     <address>

<location type="location" subtype1="Nation">India<location type="location"
subtype1="Place">New Delhi</location>
</location>

<email>foo@example.com</email></address>
    </affiliation>
  </author>
 </info>
 <para>This is a short article.</para>
</article>
```

7. Summary

Sl. No.	Data category	Description	Comments w.r.t Indian Languages
1.	**Translate**	The Translate data category expresses information about whether the content of an element or attribute should be translated or not. The values of this data category are "yes" (translatable) or "no" (not translatable).	Some words in Indian languages needs to transliterate instead of translation. The following examples shows the requirements. • **Rasgulla(Sweet Name)** • **Bharatnatayam(Dancing style)** In the above examples both Rasgulla & Baharatnatyam are Indian words and the localization of these words in Hindi are रसगुल्ला and भरतनाट्यम respectively. So there is a need to introduce some mechanism for different identification of different Named entities in Indian languages .
2.	**Localization Note**	The Localization Note data category is used to communicate notes to localizers about a particular item of content.	Localization note clarify ambiguity and show relationships between items to allow correct translation that depends on gender, number and case of the thing refer to. But in Indian languages some particular words depends on the part of speech also. The following examples shows the requirements in Bengali language : Bengali word (“sarala”) is pronounced in two different ways when it is verb it is pronounced as /ʃɔrolo/ (moved) and /ʃɔrol / (easy) when it behaves as adjective. So there is need to consider part of speech also in Localize note for correct translation in Indian languages.
3.	**Disambiguation**	The Disambiguation data category is used to highlight (mark up) specific conceptual patterns that may require special treatment when localizing and translating content. This data category can	Named entities can be identified by using tagging within the XML Document.

		be used for several purposes, including, but not limited to: Informing a translation service that a certain fragment of text is subject to follow specific translation rules, e.g. for proper names, or officially regulated translations, as well as to conveying a very specific meaning of the fragment.	

8. References

i. http://www.w3.org/TR/2012/WD-its20-20121206/
ii. XLIFF and ITS: A Secret Marriage : 1st International XLIFF Symposium 2010
iii. https://www.oasis-open.org/committees/download.php/26820/XLIFF-2.0_requirements.pdf
iv. http://w3cindia.in/namepatterns.html

Annexure I

Sl. No	Category			Label	Annotation Convention**	Remarks
	Top level	Subtype (level 1)	Subtype (level 2)			
1	Noun			N	N	
1.1		Common		NN	N_NN	
1.2		Proper		NNP	N_NNP	
1.3		Verbal		NNV	N_NNV	The verbal noun sub type is only for languages such as Tamil and Malayalam)
1.4		Nloc		NST	N_NST	
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR_PRP	
2.2		Reflexive		PRF	PR_PRF	
2.3		Relative		PRL	PR_PRL	
2.4		Reciprocal		PRC	PR_PRC	
2.5		Wh-word		PRQ	PR_PRQ	
2.6		INDEFINITE		PRI	PR_PRI	
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM_DMD	
3.2		Relative		DMR	DM_DMR	
3.3		Wh-word		DMQ	DM_DMQ	
3.4		Indefinite		DMI	DM_DMI	
4	Verb			V	V	
4.1		Main		VM	V_VM	
4.1.1			Finite	VF	V_VM_VF	
4.1.2			Non-finite	VNF	V_VM_VNF	
4.1.3			Infinitive	VINF	V_VM_VINF	
4.1.4			Gerund	VNG	V_VM_VNG	
4.2		Verbal		VN	V__VN	*paTittam, naTattam, naTanam*

4.2		Auxiliary		VAUX	V_VAUX	
4.2.1			Finite	VAUX	V_VAUX_VF	
4.2.2			Non-finite	VNF	V_VAUX_VNF	
4.2.3			Infinitive	VINF	V_VAUX_VINF	
4.2.4			Gerund	VNG	V_VAUX_VNG	
4.2.5			PARTICIPLE NOUN	VNP	V_VAUX_VNP	
5	**Adjective**			**JJ**		
6	**Adverb**			**RB**		Only manner adverbs
7	**Postposition**			**PSP**		
8	**Conjunction**			**CC**	**CC**	
8.1		Co-ordinator		CCD	CC_CCD	
8.2		Subordinator		CCS	CC_CCS	
8.2.1			Quotative	UT	CC_CCS_UT	
9	**Particles**			**RP**	**RP**	
9.1		Default		RPD	RP_RPD	
9.2		Classifier		CL	RP_CL	
9.3		Interjection		INJ	RP_INJ	
9.4		Intensifier		INTF	RP_INTF	
9.5		Negation		NEG	RP_NEG	
10	**Quantifiers**			**QT**	**QT**	
10.1		General		QTF	QT_QTF	
10.2		Cardinals		QTC	QT_QTC	
10.3		Ordinals		QTO	QT_QTO	
11	**Residuals**			**RD**	**RD**	
11.1		Foreign word		RDF	RD_RDF	A word written in script other than the script of the original text
11.2		Symbol		SYM	RD_SYM	For symbols such as $, & etc

11.3		Punctuation		PUNC	RD_PUNC	Only for punctuations
11.4		Unknown		UNK	RD_UNK	
11.5		Echowords		ECH	RD_ECH	