

# ITS Widening Doors to Global XML-based Content

on this page: [introduction](#) - [challenges for global xml-based content](#) - [its users and usage scenarios](#) - [basic concepts](#) - [its applied to dita, docbook and open document](#) - [its used as bridgehead into translation](#) - [further reading](#)

**intended audience:** schema developers (DTDs, XML Schema, RelaxNG, etc.), content producers (localization engineers, authors, translators), vendors of content-related tools, and anyone who is new to XML internationalization and needs guidance on topics to consider

## introduction

### Worldwide Use of Content

XML has many built-in capabilities to support the worldwide use of content. Proper use of these capabilities for the purpose of internationalization and localization, however, sometimes requires considerable expertise. The ITS Working Group (ITS WG [ITS WG]) of the World Wide Web Consortium (W3C) is working on a standard which captures this expertise. The standard identifies concepts (such as "directionality") which are important for internationalization and localization, and defines their implementation as an Internationalization Tag Set (ITS [ITS]). The standard recommends special purpose meta-data and suggests a standard representation for this meta-data.

### Enhancing and Complementing

ITS can be used to enhance or complement new or existing XML-based content (for example content based on OASIS Standards such as DITA [DITA], DocBook [DocBook] and Open Document [Open Document]) with respect to internationalization and localization. Furthermore, ITS can serve as a bridgehead between native formats such as the ones mentioned, and XLIFF [XLIFF], a localization related format under development at OASIS.

## challenges for global xml-based content

### Localization

Content or software that is authored in one language (so-called source language) is often made available in additional languages or adapted with regard to other cultural aspects. This is done through a process called localization, where the original material is translated and otherwise adapted to the target audience.

### Internationalization

Document formats may be used by people in different parts of the world, and these people may need special markup to support the local language or script. For example, people authoring in languages such as Arabic, Hebrew, Persian or Urdu need special markup to demarcate directionality in mixed direction text. Considering this kind of need requires appropriate design and development. The corresponding engineering activity is part of a process called internationalization, where a format or other type of application is made ready for worldwide use.

From the viewpoints of feasibility, cost, and efficiency, it is important that the original material should be suitable for localization. For a detailed explanation of the terms "localization" and "internationalization", see [10n i18n].

#### XML-based Content

The increasing usage of XML as a medium for documentation-related content (e.g. DocBook, and DITA as formats for writing structured documentation, well suited to computer hardware and software manuals) and software-related content (e.g. the eXtensible User Interface Language [XUL]) creates challenges and opportunities in the domain of XML internationalization and localization.

Lack of meta-data and standard representation are two categories of issues that currently hinder efficient XML-related localization.

Without data about content, several important questions cannot be answered. Examples:

#### Lack of Meta-data

1. What is the language that is being used in the content?
2. Do specific parts of the content need special attention since they are terms (and thus may require a very specific translation)?
3. Are some parts of the content written in a specific script (and thus may for example require rendering from right-to-left rather than left-to-right)?
4. Which types of markup do not affect linguistic integrity (and thus should not affect text flow)?
5. Is there some content which does not need to be translated or even must not be translated?

Even if sufficient meta-data is available, efficiency still may be poor since no standard representation is being used. One of the reasons for this negative impact is the need to adapt tools and processes for each representation. Examples:

1. Language information may be represented by any combination of attributes such as "lang", "xml:lang", elements such as "locale", "lang", "language", values such as "E", "English", "en", "en-us"
2. Content which may not need to be translated (however, has to be part of the content in order to provide for example context information for a translator) may be represented by sundry mechanisms such as a "context" attribute, or even an XML comment

#### Lack of Standard Representation

## its users and usage scenarios

The Internationalization Tag Set (ITS [ITS]) can be used by many of the various players in comprehensive XML content creation scenarios. Examples:

- Schema Developers: A schema developer may integrate ITS markup declarations in his schema to allow users to indicate that specific parts of the content should not be translated.
- Content Producers: A content author or information architect uses ITS markup at the top of the document to identify a particular type of element or context in which a certain part of content should not be translated.
- Vendors of Content-related Tools: An editing application for translators may be skip content

which has been marked as "does not need to be translated" by means of ITS markup.

## basic concepts

ITS defines data categories as an abstract notion for meta-data for internationalization and localization of XML schemas and documents. ITS data categories target concepts such as:

### Data Categories

1. identifying language/locale
2. identifying terms
3. supporting bidirectional text
4. indicating of translatability
5. carrying localization notes
6. providing annotation/Ruby markup
7. indicating segmentation hints

Depending on the data category and its usage, there are additional attributes for adding information to the selected nodes, or for pointing to existing information in the document. For example, the data category for localization information can be used to add information to selected nodes, or to point at existing information in the document. For the former purpose, a `locNote` element can be used. For the latter purpose, a `locNotePointer` attribute can be used.

### Selection

For ITS markup which appears in an XML instance, it has to be clearly defined to which XML nodes the ITS information pertains. Thus, ITS defines selection mechanisms to specify to what parts of an XML document an ITS data category and its values should be applied.

Information (e.g. "translate this") captured by ITS markup (e.g. `its:translate='yes'`) always pertains to one or more element and attribute nodes. In a sense, ITS markup "selects" the XML node(s). Selection may be explicit or implicit. ITS distinguishes two approaches to selection: local, and with global rules.

### ITS and CSS

The mechanisms defined for ITS selection resemble those defined in [CSS2]. The local approach can be compared to the style attribute in CSS, and the approach with global rules is similar to the style element in CSS. In contrast to CSS, ITS uses XPath for identifying nodes. While the local approach puts ITS markup in the relevant element of the host vocabulary (e.g. the `author` element in DocBook) the rule-based, global approach puts the ITS markup in elements defined by ITS itself (namely the `rules` element) ITS markup can be used with XML documents (e.g. a DocBook article), or schemas (e.g. an XML Schema document for a proprietary document format).

Content authors need for example a simple way to work with the translatability data category in order to express whether the content of an element or attribute should be translated or not. Localization coordinators, on the other hand, need an efficient way for managing translations of large document sets based on the same schema. This could be realized by a specification of defaults for translatability and exceptions from the defaults (e.g. all `p` elements should be translated, but not `p` elements inside of an `index` element).

### Precedence and Inheritance

The power of the ITS selection mechanisms comes at a price: rules related to overriding/precedence, and inheritance, are needed.

## its applied to dita, docbook and open document

ITS can be used to enhance or complement some well-known OASIS XML vocabularies like DITA, DocBook and Open Document.

### General Questions

Four questions need to be addressed for each vocabulary (they will be exemplified using DITA):

1. What is the benefit of working with ITS? Example: Add missing data category such as Ruby
2. How should ITS be integrated? Example: Use ITS modules
3. How can existing markup be "reused"? Example: Associate with DITA's "translate" attribute
4. What caveats exist?

ITS Contributions to DITA

In the context of the OASIS Darwin Information Typing Architecture (DITA) , the use of ITS can add the following:

1. Addition of meta-data such as script directionality and Ruby annotations
2. More fine-grained indication of translatable content: translatable attributes and non-translatable elements can be designated easily

```
<!-- Translatable attribute (some are deprecated) -->
<its:translateRule selector="//@alt" translate="yes"/>
<its:translateRule selector="//topicgroup/@navtitle" translate="yes"/>

<!-- Non-translatable elements -->
<its:translateRule selector="//draft-comment/*" translate="no"/>
<its:translateRule selector="//draft-comment/descendant-or-self::*/*" translate="no"/>
```

3. Centralized meta-data such as which DITA elements mark terms, or which elements do not affect linguistic integrity

```
<!-- Terminology -->
<its:termRule selector="//term //dt"/>
<!-- Elements within text (inline) -->
<its:withinText withinText="yes"
  selector="... //keyword | //b | //i | //sub | //sup | ..." />
```

In the case of DITA, three directions for relating ITS and DITA come to mind.

### Relating ITS and DITA

- Since the DITA community is working on new versions, ITS might become part of core DITA (e.g. as a module)
- ITS markup can be associated with DITA markup. For example, the DITA "translate" attribute provides similar functionality as the "its:translate" attribute. ITS allows to associate the existing attribute to ITS using a global rule.

```
<its:translateRule selector="//*[@translate='no']" translate="no"/>
```

- DITA's specialization, customization, generalization mechanisms might be used

For certain types of meta-data DITA already provides markup equivalent to that recommended in ITS (see above).

```
<!-- Translatability flags -->
<its:translateRule selector="//*[@translate='no']" translate="no"/>
```

#### Caveats

Several features of DITA need special attention, when working with ITS

- specialization, customization, generalization mechanisms have to be compared carefully with the precedence, inheritance, and defaults defined in ITS
- inclusion in DITA is handled by means of the proprietary conref mechanism
- proliferation rules (e.g. for language information) are defined between DITA maps and other types of DITA objects

## its used as bridgehead into translation

Translation of XML-based content very often can either work on the native format, or on an intermediate/interchange format. In both cases, ITS works very well.

If used with a native format like DITA, ITS can enhance or complement the inherent translation-related features. A file with ITS rules can carry information which is needed in order to configure XML-aware localization tools. The "withinText" data category of ITS can for example capture the information which tags are "internal" in terms of the SDL/Trados TagEditor (an editor which translators use).

#### ITS and Native Format

If used with an interchange format like XLIFF, ITS information can be used in filters which extract content from the native file format. The "translate" data category of ITS can for example capture the information which elements and attributes contain material which needs to be placed in XLIFF's "source" element.

#### ITS and Interchange Format

---

Authors: Christian Lieske, SAP AG; Felix Sasaki, W3C; Yves Savourel, ENLASO Corp.

---