

XML Internationalization

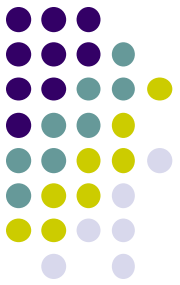
Presented by :

Swaran Lata

**Country Manager, W3C India Office
6, CGO complex, Electronics Niketan
New Delhi**

E-mail : slata@mit.gov.in

Table of Content

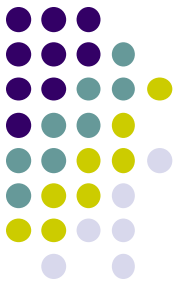


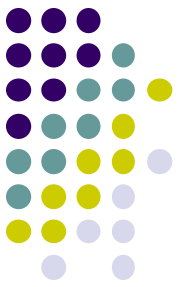
- What is Internationalization and Localization
- Localization Vs Internationalization
- Relation Between DCR and DCS
- Data Category Selection (DCS)
- Data Model
- Data Element
- Data Category Register
- Data Category Interchange Format (DCIF)
- Typical E-Gov
Land Record Codification
- XML
- What is Mark-up language
- How XML Works

Application:

Table of Content Contd..

- Advantage of XML over Database
- W3C XML Activity
- Main Attributes
- XML Internationalization
- XML Schema
- Need of XML Schema
- Design Principles for XML Schema
- What is Internationalization Tag Set (ITS)
- Metadata
- XML Metadata
- Benefits of Metadata
- What is POS tag
- POS TAG
- Using Three Levels Markups
- Working with multilingual documents





What is Internationalization and Localization

- **Localization-**

Localization refers to the adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific target market (a "locale").

- **Internationalization-**

Internationalization 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.



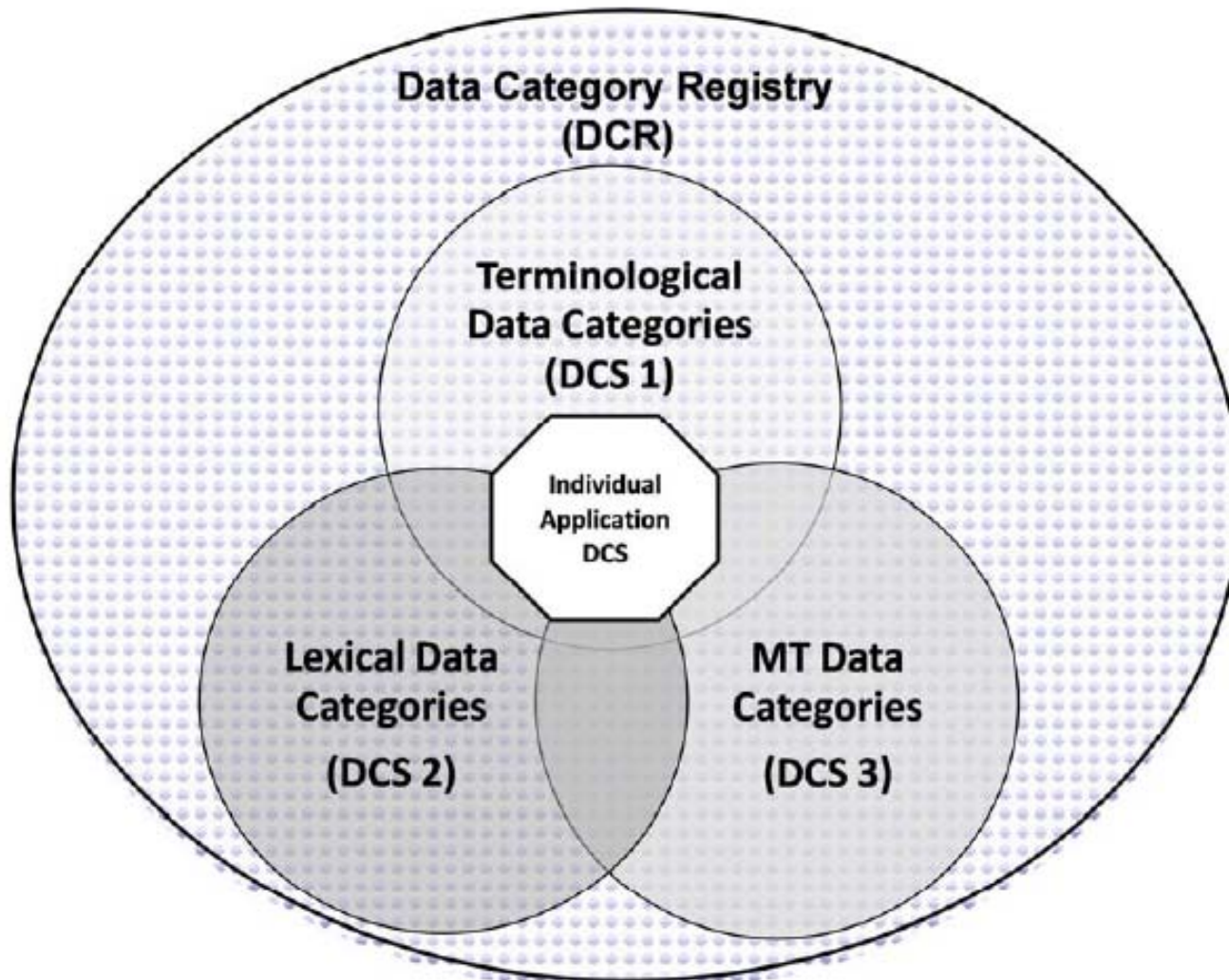
Ref: <http://www.w3.org/International/questions/qa-i18n>

Localization Vs Internationalization



- Designing and developing in a way that removes barriers to localization or international deployment.
- Providing support for features that may not be used until localization occurs.
- Enabling code to support local, regional, language, or culturally related preferences.
- Separating localizable elements from source code or content, such that localized alternatives can be loaded or selected based on the user's international preferences as needed.
- It can be localized quickly.

Relation Between DCR and DCS



Relation Between DCR and DCS.... Contd.



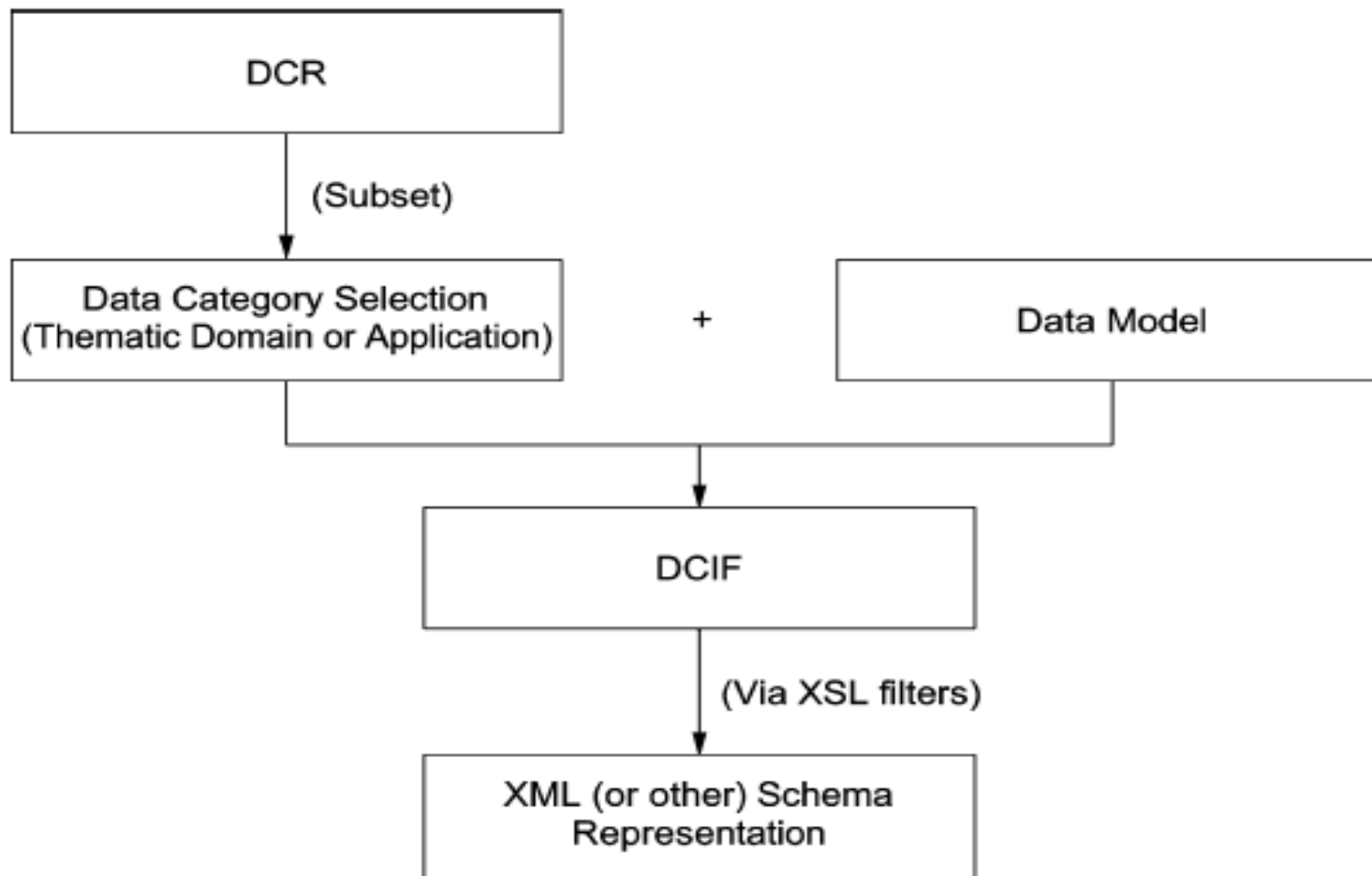
- The largest circle represents the entire collection of data category specifications included in the DCR.
- The smaller internal circles represent DCSs which are subsets of the DCR.
- The octagon represents an example of an application-specific subset.
- Terminological data collection explicitly refers to a set of reference data categories for terminology representation. e.g., */source/1, /responsibility/, /date/, etc.) as well as linguistically oriented ones (e.g., /part of speech/).*
- Lexicographical data: include data categories for the description of lexicographic data in order to ensure that the formats used for describing lexicographical, terminological and NLP-oriented data.
- MT Data categories includes machine translation (MT) dictionaries, etc.), or for specific applications such as metadata for language resources, multilingual data representation (e.g., translation memories, localization files, etc.)

Data Category Selection (DCS)



- DCS provides the formal representation of a data category, which shall comprise the specific attributes the document (e.g.-the data category name, definition, comments, etc).
- It shall also provide the context for its creation and management within the DCR.

Data Category Selection (DCS).....

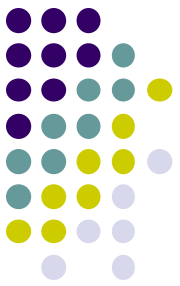


The role of data category selections in the context of the definition of linguistic annotation schemes.

Description ..



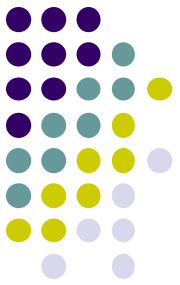
- Fig. shows the possible uses for a DCS. It may be merely a list of data categories that points back to the complete specification in the DCR.
- It can be represented by a complete subset or even superset of the DCR.
- It also presents the notion of a DCS, e.g., the choice of a specific set of data categories taken from the global DCR for use in a given thematic domain within the framework of language resources, or in a specific application.



Contd..

- The diagram exemplifies the various roles of a DCS in the process of defining and using any linguistic annotation scheme.
- The DCS can be seen as a documentary source for the linguistic annotation scheme because it contains the list of all data categories that can be used in conjunction with the annotation scheme.
- It is probably the best source of information for potential users or implementers who want to know whether a given data category corresponds to their needs.

Data Model

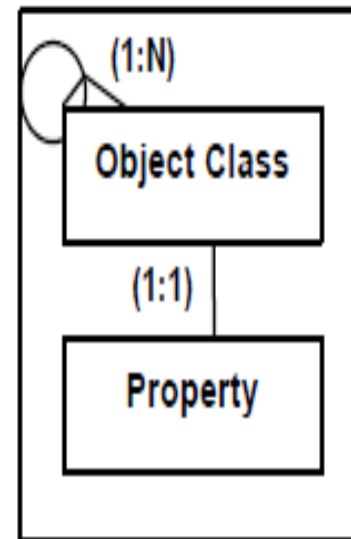


- Graphical and/or lexical representation of data, specifying their properties, structure and inter-relationships.
- Data model that represents an abstract view of the real world.

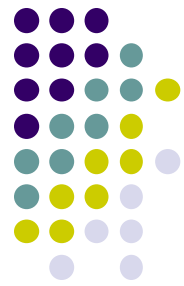
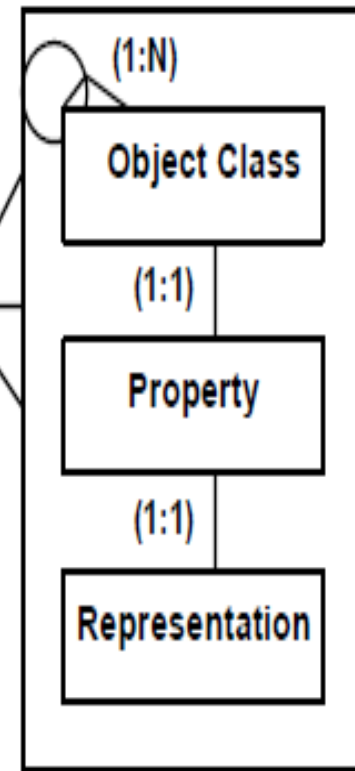
Data Element

- **Data element concept** – A DEC is concept that can be represented in the form of a data element, described independently of any particular representation.
- **Representation** – The representation is composed of a value domain, data type, units of measure (if necessary), and representation class (optionally).
- The **object class** is a set of things in the real world that can be identified with explicit boundaries and meaning and whose properties and behaviour follow the same rules.
- The **property** is a characteristic common to all members of an object class.

DATA ELEMENT CONCEPT



DATA ELEMENT



Data Category Register

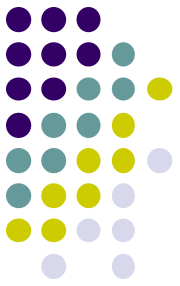


- Elementary descriptor used in a linguistic description or annotation scheme
- Example:

/Part of speech/, /Gender/, /Number/

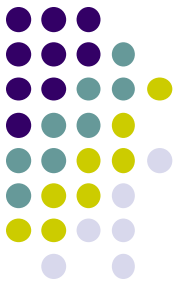
- The purpose of the DCR is to promote greater usability and reusability of annotated language resources.
- Increased semantic integrity for information in annotation documents by providing a set of formally-defined reference categories.

Data Category Interchange Format (DCIF)

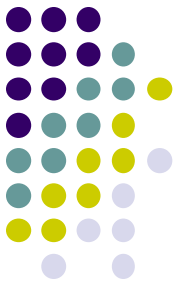


- DCIF used for archiving and exchanging all or part of the DCR within TC 37.
- For applications where individuals have to manipulate and transmit their own proprietary data categories in the field of language resources.
- DCIF is described as a hierarchical component model.
- The components relate to the major classes of the data model.

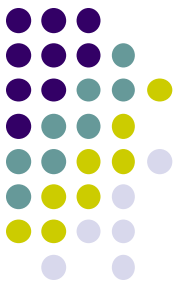
Cont..



- DCIF export of a DCS from the DCR will always result in a DCIF XML document that complies with both the DCR data model and the DCIF schema.
- Due to the looser DCIF schema, the DCR should always validate a DCIF XML document against the DCR data model upon import.



Typical E-Gov Application: Land Record Codification

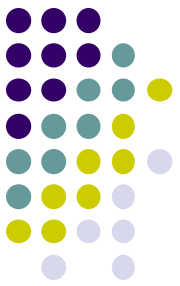


Metadata and Data Standards for Land Region Codification

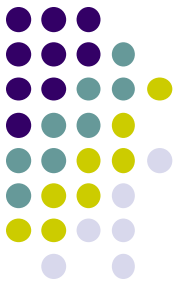
- To identify generic data elements associated with land region codification, and standardize their business formats and their metadata, to meet requirements of interoperability for vertical / horizontal data exchange between various domain applications of e-Governance.
- Revenue village has been taken as smallest unit of land region, for the purpose of standardization of land region codification

XML form ..

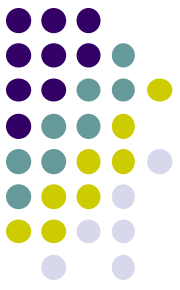
```
<?xml version="1.0" encoding="ISO-8859-1"?>  
<country> // ISO 3166-3 maintains Country code  
<state>  
<district>  
<sub-district>  
<urban-address>  
<town>  
<Language Code> // ISO 639-3 maintains Language code  
<measurement-area>  
<squaremeter>00.02-2b-0</squaremeter>  
</measurement-area>  
</Language Code>  
</town>  
</urban-address>  
</sub-district>  
</district>  
</state>  
</country>
```



XML



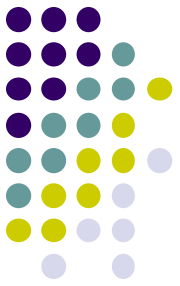
- XML is a mark-up language for documents containing structured information.
- Structured information contains both content (words, pictures, etc.).
- Some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table, etc.).
- A mark-up language is a mechanism to identify structures in a document.
- The XML specification defines a standard way to add mark-up to documents.



What is Mark-up language

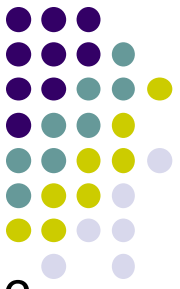
- Mark-up languages are designed for the processing, definition and presentation of text.
- The language specifies code for formatting, both the layout and style, within a text file.
- The code used to specify the formatting are called tags.
- A mark-up language is a set of words and symbols for describing the identity of pieces of a document.
- for example ‘this is a paragraph’, ‘this is a heading’, ‘this is a list’, ‘this is the caption of this figure’, etc.

How XML Works



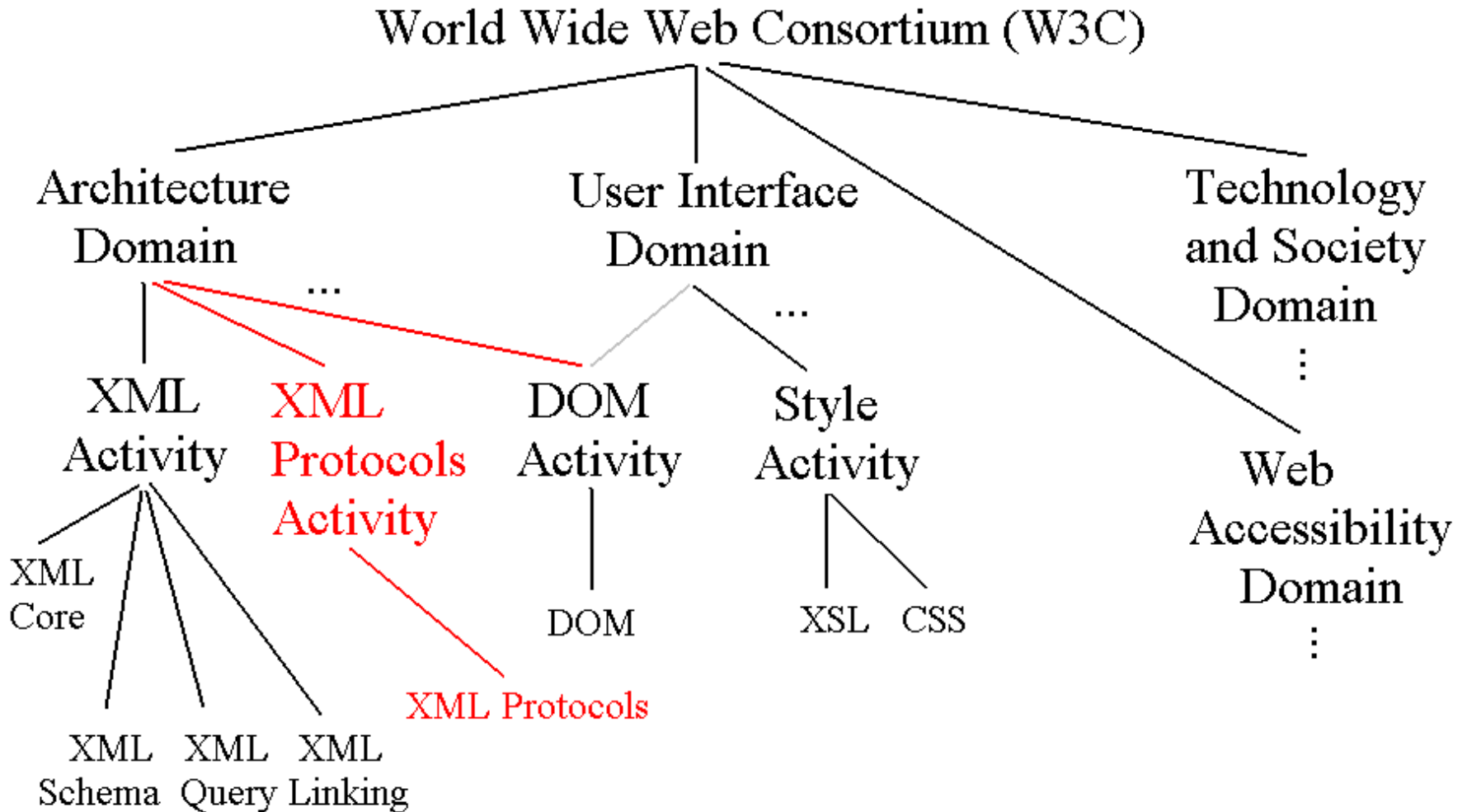
- XML makes it possible to pay attention to data structure, what it means, and what context it appears in.
- With the aid of rules set by XML Schemas, this aspect makes it easier for programs, such as search engines, to analyze an XML document and find the information it needs (and in its correct context).
- Plus, XML gives Website developers the flexibility to create their own set of customized tags for documents.

Advantage of XML over Database



- **Built in support**
 - for internationalization due to the fact that it utilizes Unicode.
- **Platform independence**
 - for instance, no need to worry about endianness.
- **Readable**
 - Human readable format makes it easier for developers to locate and fix errors than with previous data storage formats.
- **Extensibility**
 - XML allows developers to add extra information to a format without breaking applications.
- **Support off-the-shelf tools**
 - Large number of off-the-shelf tools for processing XML documents already exists.

W3C XML Activity



W3C XML Categorization



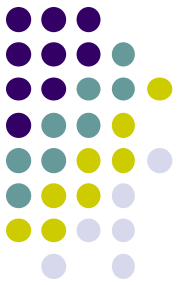
- **User Interface Domain-** The User Interface Domain seeks to improve all user/computer communications on the Web.
- **XML Protocol Activity-** It is develop technologies which allow two or more peers to communicate in a distributed environment, using XML as its encapsulation language.
- **DOM (Document Object Model)-** The DOM is a platform- and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents.
- **XSL-** XSL is a language for expressing style sheets.
Allowing developers to dictate the way Web pages are printed, and specifications allowing one to transfer XML documents across different applications.
- **CSS-** It is a way to design a website, or a group of websites, so that they have a consistent look and feel, and so that their look and feel is easy to change.
- **XML Core-** It is an application for processing Extensible Stylesheet Language Transformation (XSLT) in an XML file.
- **XML Schema-** An XML schema is a description of a type of XML document, provides a means for defining the structure, content and semantics of XML documents.
- **XML Query-** In XQL, a query returns XML document nodes from one or more XML documents.
- **XML Linking-** It allows elements to be inserted into XML documents in order to create and describe links between resources.

Main Attributes



- Defining mark-up for natural language labelling.
- Defining mark-up to specify text direction.
- Indicating which elements and attributes should be translated.
- Providing information related to text segmentation.
- Defining mark-up for unique identifiers.
- Defining mark-up for notes to localizers.
- Working with multilingual documents.

Ref: <http://www.w3.org/TR/2008/NOTE-xml-i18n-bp-20080213/>



XML Internationalization



| Best Practices | Tag | Example |
|--|--|---|
| Defining markup for natural language labelling | <p>Xml:lang</p> <p>defined for the root element of your document, and for any element where a change of language may occur.</p> | <pre><myRes> <messages> <msg id="1"> <langcode>en</langcode> <text>Cannot find file.</text> </msg> <msg id="2"> <langcode>hi</langcode> <text>राम अच्छा लड़का है</text> </msg> </messages> </myRes></pre> <p>Example's source code</p> <p>its:langRule element that specifies that the lang-code element holds the same values as the xml:lang attribute and applies to the text element.</p> <pre><its:rules xmlns:its="http://www.w3.org/2005/11/its" version="1.0"> <its:langRule selector="//text[../langcode]" langPointer="../langcode"/> </its:rules></pre> |



| Best Practices | Tag | Example |
|---|--|---|
| Defining markup to specify text direction | <p>its:dir attribute is defined for the root element of your document, and for any element that has text content.</p> | <pre><text xml:lang="en"> <body> <par>In Hindi, the title <quote xml:lang="hi" textdir="l2r">, राम जाता है </quote> means<quote>Internationalization Activity, W3C</quote>.</par> </body> </text></pre> <p>Example's souce code</p> <pre><its:rules xmlns:its="http://www.w3.org/2005/11/its" version="1.0"> <its:dirRule selector="//*[@textdir='l2r']" dir="ltr"/></pre> |

| Best Practices | Tag | Example |
|--|---|--|
| <p>Indicating which elements and attributes should be translated</p> | <p>its:translateRule</p> <p>element to address this requirement.</p> | <pre><myDoc xml:lang='en'> <head> <id xml:lang="zxx">H4-A3-F8-A1</id> <author> वाल्मीकि </author> <rev>v13 2007-10-27</rev> </head> <par>सर्च बटन पर क्लिक करो <ins>the <ui> सर्च </ui> बटन </ins> तरह -तरह के विकल्प और दिए गये विकल्प का फॉर्म भरो: <ref file="vat.png" alt=" आयकर फॉर्म " /></par> </myDoc></pre> <p>Example's souce code</p> <p>The following rules specify exceptions from the default ITS behavior for documents like the one above.</p> <pre><its:rules xmlns:its="http://www.w3.org/2005/11/its" version="1.0"> <its:translateRule selector="/myDoc/head" translate="no"/> <its:translateRule selector="//*/@alt" translate="yes"/> <its:translateRule selector="//del" translate="no" /> <its:translateRule selector="//*[lang('zxx')] //@*[lang('zxx')]" translate="no"/> </its:rules></pre> |



| Best Practices | Tag | Example |
|--|---|---|
| Providing information related to text segmentation | Ita:within Text Rule elements to indicate which elements should be treated as either part of their parents, or as a nested but independent run of text. | <pre><concept id="myConcept" xml:lang="hi"> <title>अलग अलग तरह के नृत्य </title> <conbody> भरतनाट्यम:<p><term> भरतनाट्यम </term><fn> भरतनाट्यम मिलती जुलती है कत्थक .</fn> भारतीय शास्त्रीय नृत्य. भारतीयों के शास्त्रीय नृत्य के इस प्रकार में महत्वपूर्ण गया है.</p> </conbody> </concept></pre> |

Example's source code

- The elements **term** and **b** should be treated as part of their parent.
- The element **fn** should be treated as an independent run of text.

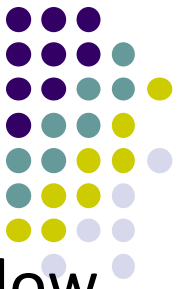
```
<its:rules
xmlns:its="http://www.w3.org/2005/11/its"
version="1.0">
<its:withinTextRule selector="//term | //b"
withinText="yes"/>
<its:withinTextRule selector="//fn"
withinText="nested"/>
</its:rules>
```

[Defining markup for unique identifiers](#)

xml:id

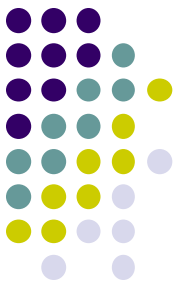
elements with translatable content can be associated with a unique identifier.

```
<xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.w3.org/XML/1998/namespace">
<xs:attribute name="id" type="xs:ID"/>
</xs:schema>
```



XML Schema

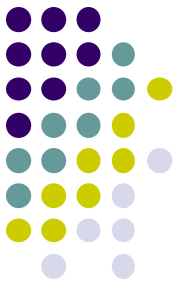
- XML Schemas express shared vocabularies and allow machines to carry out rules made by people.
- XML schema is to define a class of XML documents, and so the term "instance document" is often used to describe an XML document that conforms to a particular schema.
- It provides a means for defining the structure, content and semantics of XML documents.



Need of XML Schema

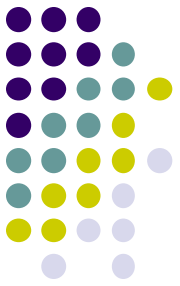
- To provide an inventory of XML mark-up constructs with which to write schemas.
- To define, describe and catalogue XML vocabularies for classes of XML documents.
- To express syntactic, structural and value constraints applicable to its document instances.

Design Principles for XML Schema



- More expressive than XML DTDs;
- expressed in XML;
- self-describing;
- usable by a wide variety of applications that employ XML;
- straightforwardly usable on the Internet;
- optimized for interoperability;
- simple enough to implement with modest design and runtime resources.

What is Internationalization Tag Set (ITS)

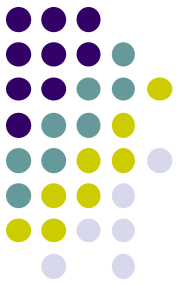


- ITS is a technology to easily create XML which is internationalized and can be localized effectively.

ITS for Schema developers

- User will find proposals for attribute and element names to be included in their new schema (also called "host vocabulary").
- It leads to easier recognition of the concepts represented by both schema users and processors.

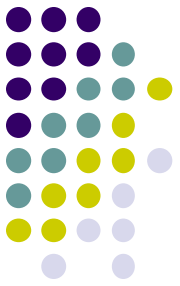
Ref: <http://www.w3.org/TR/2007/REC-its-20070403/>



Metadata

- Metadata describes how and when and by whom a particular set of data was collected, and how the data is formatted.
- It is essential for understanding information stored in data warehouses and has become increasingly important in XML-based Web applications.

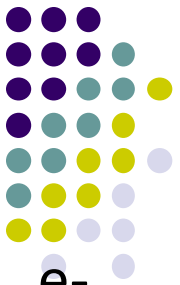
XML Metadata



Metadata built into the document

- Every element has a tag to tell you where the data is stored in the document.
- Descriptive tags give structure to the document and tell you what the data means (sort of).
- “Sort of” because it only tells the tag name, so this only has meaning to someone who already understands what the element or attribute means.

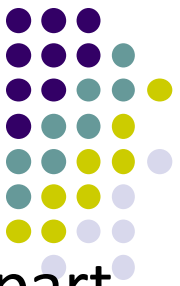
Benefits of Metadata



The fact that Metadata facilitates interoperability of e-Governance applications, is endorsed by the benefits which metadata provides:

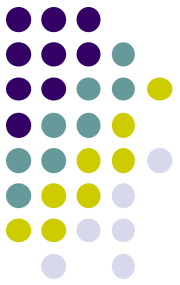
- Helps people find information
- Helps identify content
- Matches User and Content
- Exploits Government Information

What is POS tag



- Part-of-speech tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a sentence.
- The input to a tagging algorithm is a string of words of a natural language sentence and a specified tag set(a finite list of Part-of-speech tags). The output is a single best POS tag for each word.
- Tags plays an important role in Natural language applications like speech recognition, natural language parsing, information retrieval and information extraction.

POS TAG in XML

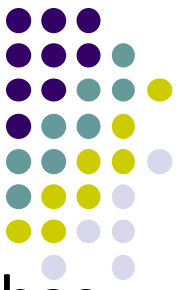


For english metadata is denoted by <en>

For hindi metadata is denoted by <hin>

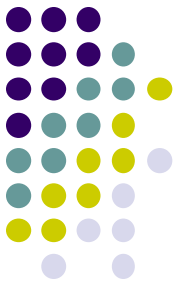
For bangla metadata is denoted by <ban>

| S.NO | English | Hindi | Bangla |
|------|--------------------|---------------------------|---------------------------|
| 1 | <en>Noun</en> | <hin>sangya</hin> | <ban>sangya</ban> |
| 2 | <en>Verb</en> | <hin>kriya</hin> | <ban>kriya</ban> |
| 3 | <en>Pronoun</en> | <hin>sarvnam</hin> | <ban>sarvnam</ban> |
| 4 | <en>Adjective</en> | <hin>vesheshan</hin> > | <ban>vesheshan</ban> > |
| 5 | <en>Adverb</en> | <hin>kriyavesheshan</hin> | <ban>kriyavesheshan</ban> |



Cont..

- In human languages, we often find that a word has several meanings (word sense ambiguity) at various content contexts (or content domain of a paragraph of a web page).
- Similarly, a word may have several linguistic parts of speech (POS ambiguity).
- For an example, the word "light" has several POS namely, verb, adjective, noun.
- Again, a metadata about a sentence helps in parsing during the machine translation of a web content.

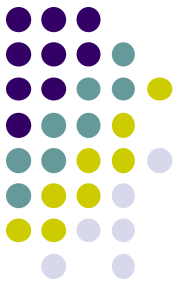


Cont..

The proposed 3-Layer XML Schema approach uses three schemas for a web content.

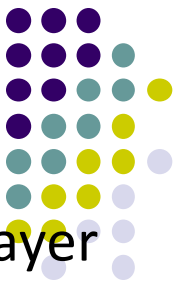
- The first schema is meant for content domain.
- The second schema is for sentence level metadata.
- The third one is meant for the word level metadata or markups.

Cont..



- The proposed 3-Layer XML Schema aims to markup both syntactic and semantic metadata information in the structure of an XML document.
- This approach is an excellent solution to yield meaningful translation.
- Such embedded information is very important to both the internationalization and localization processes.

Cont..



- We need to follow the following three basic steps of the 3-Layer Schemas to embed linguistic-related metadata information in the structure of an XML document in order to improve the translation process for obtaining more meaningful translation.
- This 3-layer schema scheme is also useful for the Translation Memory processes to keep context markups when Internationalization & Localization developers use this scheme for both source and target text.
- The 1st XML schema that contains various categories on content domain.
- The 2nd XML schema contains various categories on sentences.
- The 3rd XML schema contains various Parts-of-Speech categories on words.

Using Three Levels Markups

- The proposed scheme uses three XML elements namely, content domain, sentence category and POS category.
- The schematic block diagram of the proposed 3-Layered XML Schema approach is shown in fig.

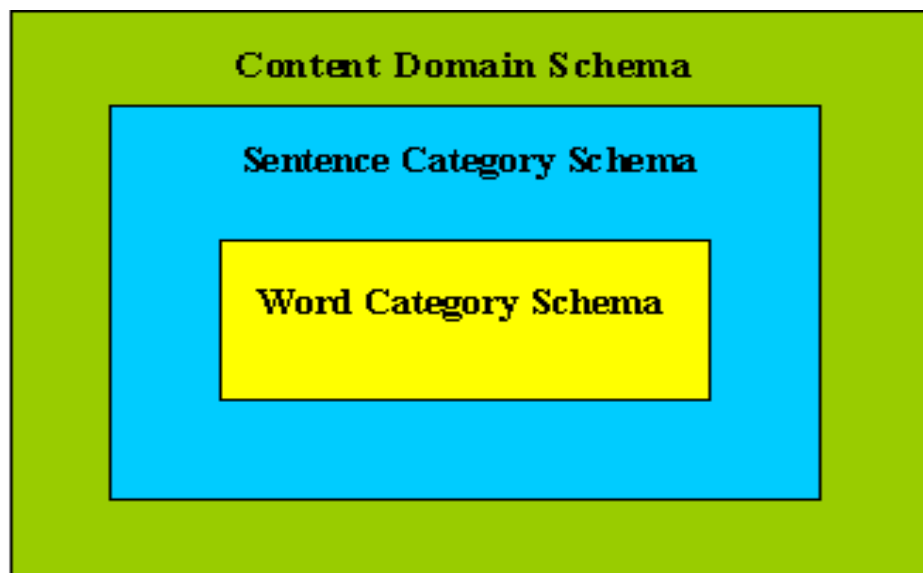
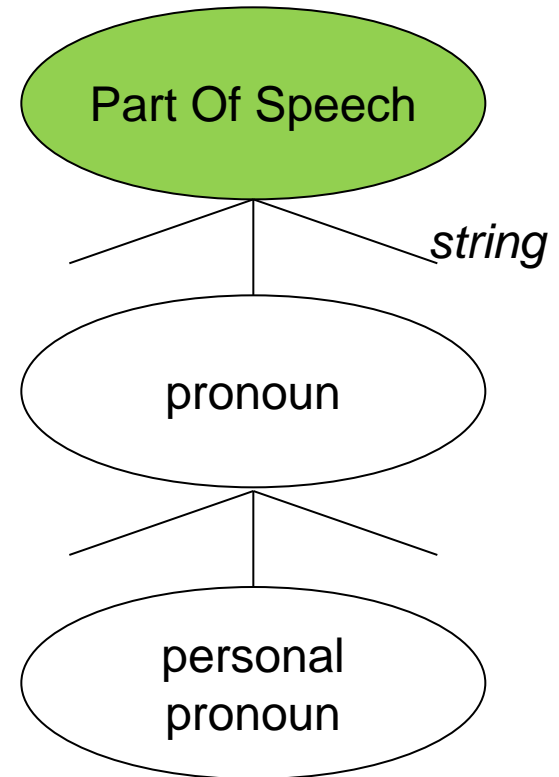


Fig. 1. The 3-Tier XML Schematic Block Diagram

Cont..

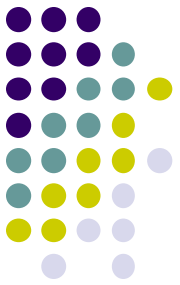
- Content domain includes various contexts namely, information technology, medicine, travel, personal, sports, mathematics, etc.
- Sentence categories include simple, compound, complex, proverbial, taunt, suspicion, active & passive voice, direct and indirect speech etc.
- Parts-of-Speech categories include noun, pronoun, verb, adjective, adverb, preposition, postposition, interjection, conjunction and indeclinable etc.



Example

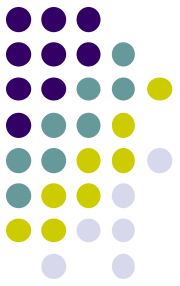
राम ने सभी बच्चों को किताबें बाँटी

```
<?xml version="1.0"?>
<text xml:lang="hi">
<content_domain name="general">
<sentence_cat id="7">
<pos_cat id="1" pos="N" type="NP" meaning="person name">
<xs:attribute value="sg">राम
</xs:attribute>
</pos_cat>
<pos_cat id="2" pos="PP">ने
</pos_cat>
<pos_cat id="3" pos="P">
<xs:attribute value="pl">सभी
</xs:attribute>
</pos_cat>
```



Contd..

```
<pos_cat id="4" pos="N" type="NC">  
<xs:attribute value="pl">बच्चों  
</xs:attribute>  
</pos_cat>  
<pos_cat id="5" pos="PP">को  
</pos_cat>  
<pos_cat id="6" pos="N" type="NC">  
<xs:attribute value="pl">किताबें  
</xs:attribute>  
</pos_cat>  
<pos_cat id="7" pos="A" type="AMN">  
<xs:attribute value="pl">बाँटी  
</xs:attribute>  
</pos_cat>  
</text>
```



Example in English and Bangla



```
<cont:content_domain name='agriculture'>
  <scat:sentence_cat name='demonstrative'>
    <pcat:pos_cat name='noun' meaning='farmer'> চাষি </pcat:pos_cat>
    <pcat:pos_cat name='verb' meaning='said'> বললেন </pcat:pos_cat>
    <pcat:pos_cat name='pronoun' meaning='I'> আমি </pcat:pos_cat>
    <pcat:pos_cat name='verb' type='missing_auxiliary' meaning='am'></pcat:pos_cat>
    <pcat:pos_cat name='adjective' meaning='a'> একজন </pcat:pos_cat>
    <pcat:pos_cat name='adjective' meaning='ordinary'> সামান্য </pcat:pos_cat>
    চাষি <pcat:pos_cat name='punctuation' type='sentence_final' meaning='.'> | </pcat:pos_cat>
  </scat:sentence_cat>
```

Example:

For the following Bengali or Bangla dialects sentence "Kaam (Kaaj in Bangla or Work in English) Saira Falo (Shesh Koro in Bangla or Complete in english)," we should markup the text with the three-layer metadata information in the following way:

```
<!-- Markup for Dialect -->
```

```
<text xml:lang="ben">
```

```
<content_domain name="dialect">
```

```
<!-- content domain metadata -->
```

```
.... other sentences
```

```
<sentence_cat name="imperative">
```

```
<!-- sentence level metadata is optional here -->
```

```
<pos_cat name="noun" meaning="work"> Kam </pos_cat>
```

```
<pos_cat name="verb" meaning="to complete"> Saira Falo </pos_cat>
```

```
<!-- word level parts-of-speech -->
```

```
</sentence_cat>
```

```
.....
```

```
</content_domain>
```

```
</text>
```

Working with multilingual documents



This refers specifically to situations where copies of the same content are stored in multiple languages in a single document.

Example :

This is an example of bad design. It shows a single document that contains multiple translations of the same content:

```
<messages>
```

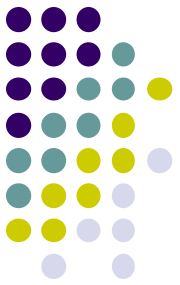
```
<msg xml:id='fileNotFound'>
```

```
<text xml:lang="en">File not found.</text>
```

```
<text xml:lang="fr">Fichier non trouvé.</text>
```

```
</msg>
```

```
</messages>
```



Contd..

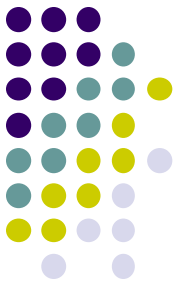
Instead, use one document for each language. Here one in English, and the other one in French. Other languages would go in similar separate documents.

Example :

```
<messages xml:lang="en">  
<msg xml:id='fileNotFound'>  
<text>File not found.</text>  
</msg>  
</messages>
```

OR

```
<messages xml:lang="fr">  
<msg xml:id='fileNotFound'>  
<text>Fichier non trouvé.</text>  
</msg>  
</messages>
```



THANK YOU