XML based application to ITU-T Recommendations

TSB

November 2009
Contents

1. XML project (Project Rx) in ITU-T
   1-1 background and objectives
   1-2 Scope and system components
   1-3 Plan

2. Conversion system from Word documents to ITU-T XML documents
   2-1 Overview
   2-2 Conversion method
   2-3 Conversion process
   2-4 Remediation process
   2-5 Experimental results

3. Application to terms and definitions processing
   3-1 Terms and definition processing
   3-2 Experimental results

4. Conclusion
1. XML project (Project Rx) in ITU-T
1-1 Background and objectives

Background

- ITU-T membership requested changes
- Need to apply right technology for 21st century publication
- Improve utility of ITU-T output, especially through the Internet
- Enable topic-focused (rather than printed image-focused) research / information delivery

Objectives

- Establish the framework for a long term effort to from static, Microsoft WORD-based ITU-T Recommendations to dynamic, XML-based documents
- Define the appropriate XML Schema
- Develop prototype system
  - Document conversion into ITU-T XML format
  - XML document element processing (extraction of logical elements, link to existing metadata)
  - Format conversion for various presentations (using different style sheet)
1-2 Scope and system components

**Scope of Project Rx**

- **WORD documents**
- **XML conversion**
- **XML documents**
- **Linking to existing metadata**
- **Existing metadata**
- **Extraction of logical elements**
  - table of contents
  - keywords
  - definitions
  - references
  - etc.
- **Various presentation styles on request**

**Editing/authoring**
1-3 Plan

**XML schema**
- Develop of basic XML schema
- Enhancement of XML schema

**XML conversion**
- Conversion for header part (definitions, references, etc.) of Rec.
- Conversion for body/annex Part (clause hierarchy) of Rec.

**System development**
- Extraction of logical elements
- Various presentations on demand
- Linking to existing metadata

2009 2010
2. Conversion system
from Word documents
to ITU-T XML documents
2-1 Overview

**Purpose**
- Basic experiment to convert existing Recommendation documents (.doc) into logically structured XML documents
- Preparation for extraction of typical Recommendation elements such as references, definitions and abbreviations

**Input document**
- Word 2007 XML format saved as XML file

**Output document**
- ITU-T XML; basically based on schema proposed by Japan to TSAG

**Diagram:**
- **Existing Rec. documents in Word format**
- **Save as Word 2007 format**
- **Conversion process**
- **Rec. documents in Word 2007 XML format**
- **Rec. documents in ITU-T XML format**
- **DB entry of XML Documents**
- **DB**
2-2 Conversion method

**Input document**
- Documents are supposed to conform to “Author’s Guide” (March, 2007 version) with some allowance
- Word XML as a Sequence of “paragraphs” with some “style” data

**Output document**
- ITU-T XML; includes metadata reflecting logically structured elements

**Conversion process**
- Automatic restructuring of document header elements utilizing “style” information
- Remediation by operator as supplementary process
2-2 Conversion method

-Example of Input vs Output-

2 References

The following ITU-T Recommendations and other references indicated were valid. All Recommendations and users of this Recommendation are therefore encouraged to use the most recent edition of the Recommendations and other currently valid ITU-T Recommendations is regularly published in this Recommendation does not give it, as a stand-alone document.


3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere.

3.1.1 policy decision physical entity (PD-PE) [ITU-T Q.3300]: instance of the policy decision functional entity (PD-FE) id...
2-3 Conversion process

- Eliminating irrelevant document elements in pre-processing
- Prototype focuses on five elements (Scope, Reference, Definition, Abbreviation and Bibliography) of header part
- Each element is analyzed and restructured into ITU-T XML form by the XSLT (XML Stylesheet Language Transformations)
2-3-1 Pre-processing

- Extraction of core XML document.xml from the docx files(package)
- Cleaning of the non essential MS Word markups in document.xml;
  - Word specific internal references (Bookmark etc.)
  - Soft-Hyphen
  - Unnecessary Spaces
  - Merging the neighboring same type of tags (</w:r><w:r>, </w:t><w:t>)
2-3-2 Document element detection

- To detect various document elements, “pStyle” and “t(ext)” information are utilized.
- If “pStyle”=‘Heading1’ and the second content of “t” =‘Scope’ / ‘References’ / ‘Definitions’ / ‘Abbreviations’ / ‘Bibliography, then each part is mapped into the respective element:<scope>, <references>, <definitions>, <abbreviations> and <bibliography>.
- Alternative expressions for “Heading” are allowed, e.g. “Normative references” for “References”, “Abbreviations and acronyms” for “Abbreviations”.

```xml
<w:p>
  <w:pPr>
    <w:pStyle w:val="Heading1"/>
  </w:pPr>
  <w:r>
    <w:t>1</w:t>
    <w:t>References</w:t>
    <w:t>Scope</w:t>
  </w:r>
</w:p>
<w:p>
  <w:r>
    <w:t>This Recommendation specifies</w:t>
  </w:r>
</w:p>
```

Move to “Scope” part processing
2-3-3 “Scope” element processing

- If the “scope” part contains a sub-clause structure, it is mapped into a hierarchal <clause> structure. “clauses” are identified by the “pStyle” of ‘Heading#’.
- “t(ext)” that has “pStyle” of ‘Normal’ is mapped into <p> element.
- “t” with that has certain “pStyle” characteristics such as ‘Note’, ‘Enumlevel’, ‘Figure’ and ‘Equation’ are mapped into <note>, <ol> or <ul>, <figure> and <equation>, respectively.

```
<w:p>
  <w:pPr>
    <w:pStyle w:val="Heading1"/>
  </w:pPr>
  <w:r>
    <w:t>1<tab/>
    <w:t>Scope</w:t>
  </w:r>
</w:p>
<w:p>
  <w:p>
    <w:r>
      <w:t>This Recommendation specifies high-
    </w:r>
  </w:p>
</w:p>

<scope>
  <p>This Recommendation specifies high-
  <p>The high-level requirements and rela-
  <p>More detailed requirements and serv-
  <p>It is recognized that a specific rea-
  <p>Administrations may require provid-
</scope>
```

ITU-T XML

WordXML
2-3-4 “Reference” element processing

- If the “references” part contains a sub-clause structure, it is mapped into a hierarchal <clause> structure.
- “t” that has “pStyle” of ‘Normal’ is mapped into <p>.
- “p” with ‘RefText’ is mapped into <referenced-document>.
- The first “t” is mapped into ‘id’ attribute.
- The second “t” is separated into two parts by ‘,’. The first part of the second “t” is mapped into <handle> and the second part of the second “t” is mapped into <title>.
- If the “reference” part contains a hyperlink, it is mapped into <url>.
2-3-4 “Reference” element processing (continued)

- ITU-T F.703 (2000), Multimedia conversations
  <http://www.itu.int/rec/T-REC-F.703>
  <http://www.itu.int/rec/T-REC-F.790>
- Recommendation ITU-T F.902 (1995), Interactive services definition
  <http://www.itu.int/rec/T-REC-F.902>

Standard case

The “pStyle is to be set to ‘enumlev1’


Common text with ISO

The format of reference for the common text with ISO is different from one of ITU-T standard format. But it is allowed.
2-3-5 “Definition” element processing

- If the “definition” part contains a sub-clause structure, it is mapped into a hierarchal <clause> structure.
- If the “p” contains more than one “t”, it is mapped into <definition>.
- The second “t” is separated into two parts by ‘:’.
- The first part is mapped into <term>, and if it contains part surrounded by ‘[]’, it is mapped into ‘xref’ attribute.
- The second part is mapped into <definition-text>.
- <definition-text> may includes <p>, <note>, <ol>/<ul>, <figure> and <equation> in accordance with the input WordXML.
2-3-5 “Definition” element processing (continued)
-Various format for “Terms defined elsewhere”-

3.1 Terms defined elsewhere
This Recommendation uses the following terms defined elsewhere:

3.1.1 application [b-ITU-T Y.101]: A structured set of capabilities, with functionality supported by one or more services.

3.1.2 content provider [ITU-T Y.1910]: The entity that owns or is licensed to distribute content assets.

(a) Standard case
2-3-5 “Definition” element processing (continued)

-Various format for “Terms defined elsewhere”-

3.1 Terms defined elsewhere
This Recommendation uses the following terms defined in [ITU-T G.661]:
- channel addition/removal (steady-state) gain response;
- channel gain;

(b) Case with only ‘term’ (without ‘id’ and ‘definition-text’)
2-3-5 "Definition" element processing (continued)

Various format for "Terms defined elsewhere"-

3.1 Terms defined elsewhere
This Recommendation uses the following terms defined elsewhere:

3.1.1 agent: [ITU-T X.701]
3.1.2 alarm reporting: [ITU-T M.3100]

(b) Case with ‘term’ and ‘definition-text’, but this ‘definition-text’ is just ‘reference’
2-3-6 “Abbreviation” element processing

- If the “Abbreviation” part contains a sub-clause structure, it is mapped into a hierarchical <clause> structure.
- If the “p” contains more than one “t”, it is mapped into <definition>.
- The first part is mapped into <term>.
- The second part is mapped into <definition-text>.

WordXML

ITU-T XML
2-3-7 “Bibliography” element processing

- “Bibliography” element is processed in the same way as the “Reference” element.
- Should there be multiple commas (‘,’) in the reference element, it is not possible to distinguish the <handle> from the <title>. Therefore, we treat the item as having a <<null>><handle>.

Word text

<referenced-document id="b-Liberty Authn">
   <url>http://www.projectliberty.org/liberty/specifications_1</url>
</referenced-document>
2-3-8 Post-processing

- Building of the ITU metadata from SQL Server
- Insertion of the ITU metadata block into document.xml
- Saving of this XML document as a custom XML part into the docx package

Example of metadata block in output XML document:
2-4 Remediation process

-Example requiring remediation at input level-

• In the case that “Definitions” and “Abbreviations” are mixed into one section.
2-4 Remediation process
-Example requiring remediation at output level-

- If the definition part include more than two ‘:’, it isn’t properly processed.

3 Definitions
This Recommendation defines the following terms:

3.1 4:4:4: A notation that defines the relative raster to be equal.
3.2 4:2:2: A notation that defines the relative raster to have twice the horizontal resolution on the three-component rasters.
3.3 4:2:0: A three-component raster to have twice the vertical resolution on the first channel.

Word text

ITU-T XML
2-5 Experimental results

- Applied to all published recommendations approved since April 2007 (about 270 Recommendations)
  - About 60%: Successfully processed
  - About 30%: Recovered with some “light-weight” remediation** by operator

**Format correction, Style correction, Spelling correction etc
2-5 Experimental results

- Remaining issues (the other 10% -

- Non-standard document structure
- Unexpected format
- Equation
- Figure
- Table
- Special font(Symbols)
- File size
2-5 Experimental results
-Examples of difficult cases(1)-

3.1.6 **Mathematical definitions**: PMD can be described in terms of Stokes or Jones vectors. The evolution of the output Jones vector with angular optical frequency $\omega = 2\pi \nu = 2\pi c / \lambda$, is the source of system impairment. All parameters, vectors and matrices in the following are functions of angular optical frequency.

For the following considerations it is assumed that the signal is fully polarized and that polarization dependent loss (PDL) is negligible.

The normalized Jones vector $\vec{j}$, with complex elements $j_\alpha$ and $j_\beta$, is defined as:

$$\vec{j} = \begin{bmatrix} \cos \theta \exp(-i \mu / 2) \\ \sin \theta \exp(i \mu / 2) \end{bmatrix}$$  \hspace{1cm} (2-4)

where:

- $\theta$ is the linear orientation of the Jones vector
- $\mu$ is the phase separation of the two elements of the Jones vector
- $i$ is $\sqrt{-1}$, the imaginary unit

• “Definition text” includes “Equations”.

ITU © 2009
2-5 Experimental results
-Examples of difficult cases(2)-

3 Definitions
This Recommendation defines the following terms as shown in Table 1:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AD$</td>
<td>Absolute audiovisual delay</td>
<td>–</td>
</tr>
<tr>
<td>$b_n$</td>
<td>Video bit rate $(n = 1, 2, \ldots, N)$</td>
<td>kbit/s</td>
</tr>
<tr>
<td>$B_{pl,n}$</td>
<td>Speech packet-loss robustness</td>
<td>–</td>
</tr>
<tr>
<td>$B_{br}$</td>
<td>Video bit rate</td>
<td>–</td>
</tr>
<tr>
<td>$D_{r,n,m}$</td>
<td>Degree of video quality robustness against packet loss $(n = 1, 2, \ldots, N,$ (m = 1, 2, \ldots, M))</td>
<td>–</td>
</tr>
<tr>
<td>$D_{r,fr}$</td>
<td>Degree of video quality robustness due to frame rate reduction</td>
<td>–</td>
</tr>
<tr>
<td>$D_{r,fr,n}$</td>
<td>Degree of video quality robustness due to frame rate reduction $(n = 1, 2, \ldots, N)$</td>
<td>–</td>
</tr>
</tbody>
</table>

- “Definitions” are represented as a “Table”.

ITU © 2009 29
2-5 Experimental results
-Examples of difficult cases(3)-

3.4 **D-value**: D-value is computed directly from measurements of the difference $\Delta S_{SW}$ between the send sensitivities for diffuse and direct sound, $S_{2i}(\text{diff})$ and $S_{2i}(\text{direct})$, respectively.

$$\Delta S_{SW} = S_{2i}(\text{diff}) - S_{2i}(\text{direct})$$

$D$ is computed as a weighted average of $\Delta S_{SW}$.

3.5 **ear-drum reference point (DRP)**: Point located at the end of the ear canal, corresponding to the ear-drum position.

3.6 **free-field equalization**: The transfer characteristics of the artificial head is equalized in such a way that, for frontal sound incidence in anechoic conditions, the frequency response of the artificial head is flat. This equalization is specific to the HATS used.

- “Definitions” includes some “special symbols”. " special symbols"
3. Application to terms and definitions processing
In this UI, a term specified by the user is retrieved and the results (with all the relevant Recs) are shown.
### 3-2 Experimental results

- From the output XML documents (about 200 Recommendations), about 2800 terms and definitions are newly extracted.

<table>
<thead>
<tr>
<th>Isn_def</th>
<th>IDREC</th>
<th>NOM</th>
<th>DATE_APP</th>
<th>term</th>
<th>abbrev</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>4510</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>accounting rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4511</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>settlement rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4512</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>termination charge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4513</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>collection charge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4514</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>lease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4515</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>rental</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4516</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>network (service) access component</td>
<td>service</td>
<td>A tariff component, normally intended to compensate Administrations for the facilities required for a customer to access a service or services, which is independent</td>
</tr>
<tr>
<td>4517</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>network (service) utilization component</td>
<td>service</td>
<td>A tariff component which is normally intended to cover the costs of a service that are dependent on the customer’s use of the network resources and any</td>
</tr>
<tr>
<td>4518</td>
<td>5702</td>
<td>D.000</td>
<td>2002-06-14</td>
<td>service invocation component</td>
<td>service</td>
<td>A tariff component which is normally intended to cover the per event cost of activating a service, already</td>
</tr>
</tbody>
</table>
4. Conclusion
4 Conclusion

- The prototype system realizes the conversion from the existing Recommendations in Word format to ITU-T XML documents, which have the ITU-T Recommendation specific logical structure.
- The more the documents are conforming to the standard format, the less the operator’s assistance is necessary. (→ Newly created Recommendations are strongly recommended to conform to the Guideline.)
- An example of application – Terms and Definitions processing – utilizing the output XML documents is shown. This indicates the potential usability of the XML documents. (→ This process would be introduced into the ordinary Editing / Publishing process.)
- The harmonization with the similar effort in ISO is continuously pursued.