

# XML based application to ITU-T Recommendations

TSB

November 2009

# Contents

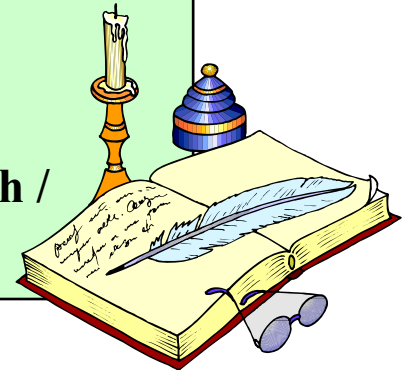
1. XML project (Project Rx) in ITU-T
  - 1-1 background and objectives
  - 1-2 Scope and system components
  - 1-3 Plan
2. Conversion system from Word documents to ITU-T XML documents
  - 2-1 Overview
  - 2-2 Conversion method
  - 2-3 Conversion process
  - 2-4 Remediation process
  - 2-5 Experimental 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

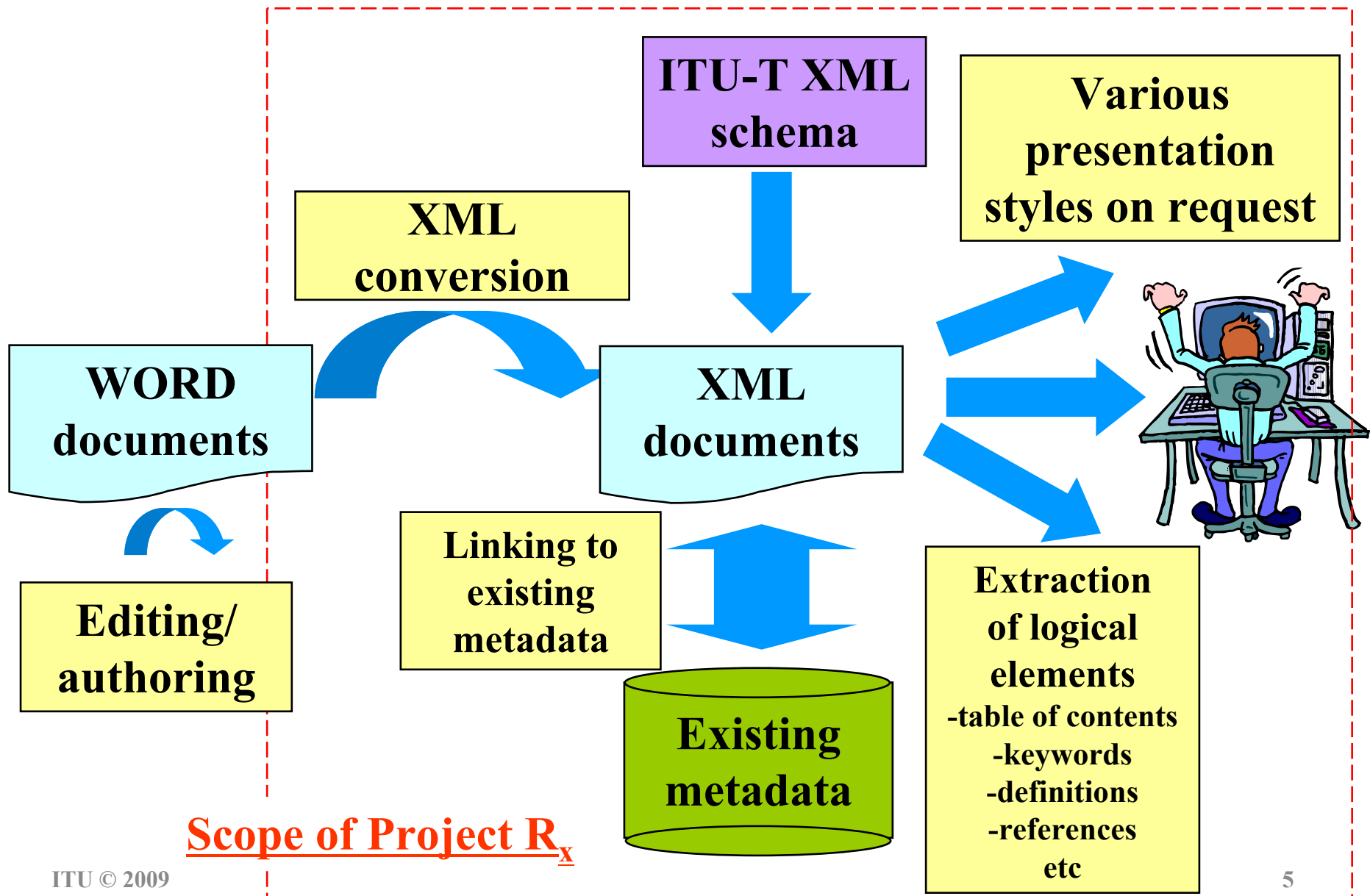
- **ITU-T membership requested changes**
- **Need to apply right technology for 21st century publication**
- **Improve utility of ITU-T output, especially through the Internet**
- **Enable topic-focused (rather than printed image-focused) research / information delivery**



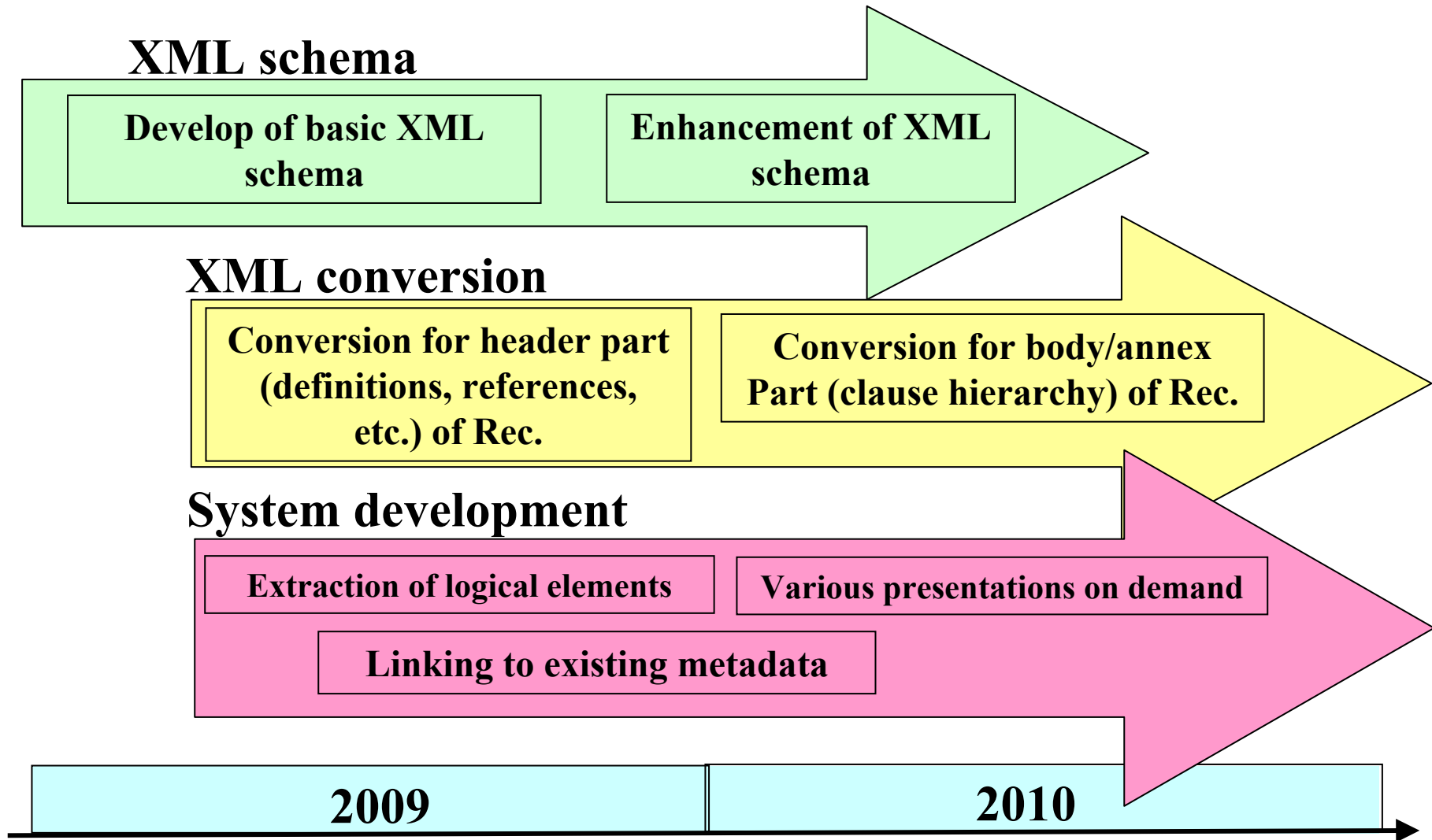
## Objectives

- **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



# 1-3 Plan



## 2. Conversion system from Word documents to ITU-T XML documents

# 2-1 Overview

## Purpose

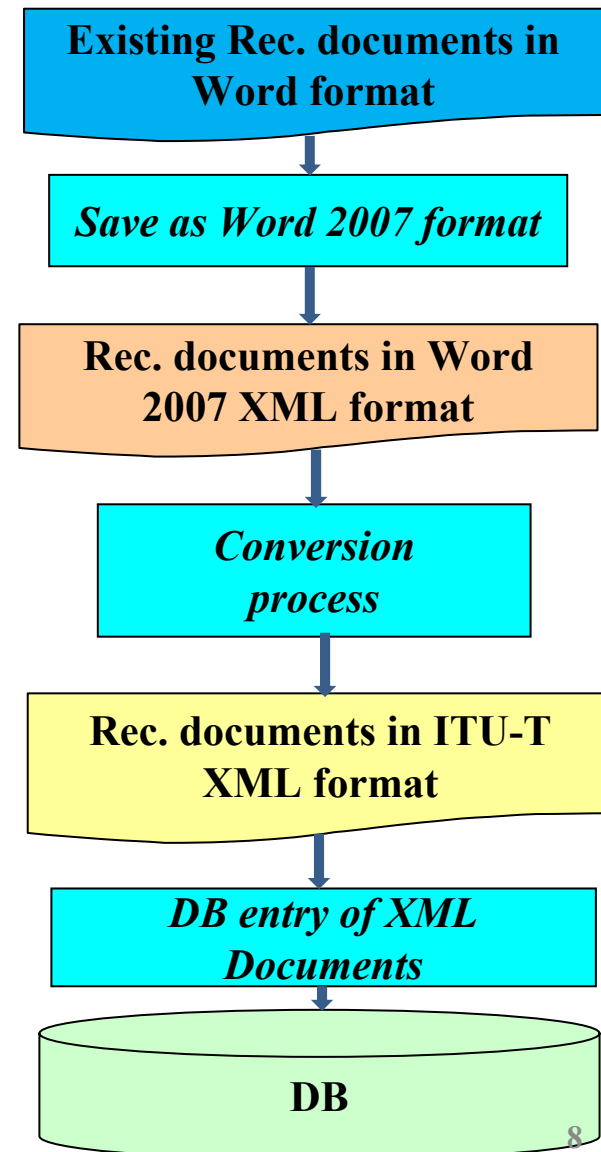
- Basic experiment to convert existing Recommendation documents (.doc) into logically structured XML documents
- Preparation for extraction of typical Recommendation elements such as references, definitions and abbreviations

## Input document

- Word 2007 XML format saved as XML file

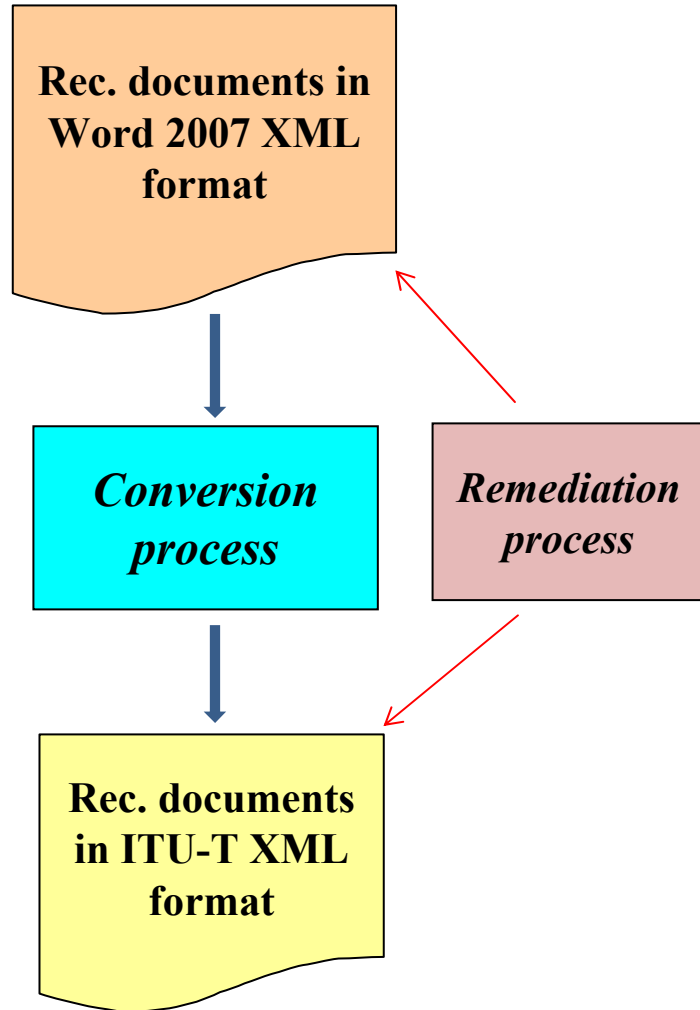
## Output document

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





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

[ITU-T Q.3300] ITU-T Recommendation Q.3300 (2008) *Q.33xx series of Recommendations.*

[ITU-T Y.2012] ITU-T Recommendation Y.2012 (2006) *architecture of the NGN release 1.*

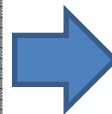
[ITU-T Y.2111] ITU-T Recommendation Y.2111 (2006) *functions in Next Generation Networks.*

**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) identified in the resource and admission control function (RACF) of the network element (NE) that is connected to the network.



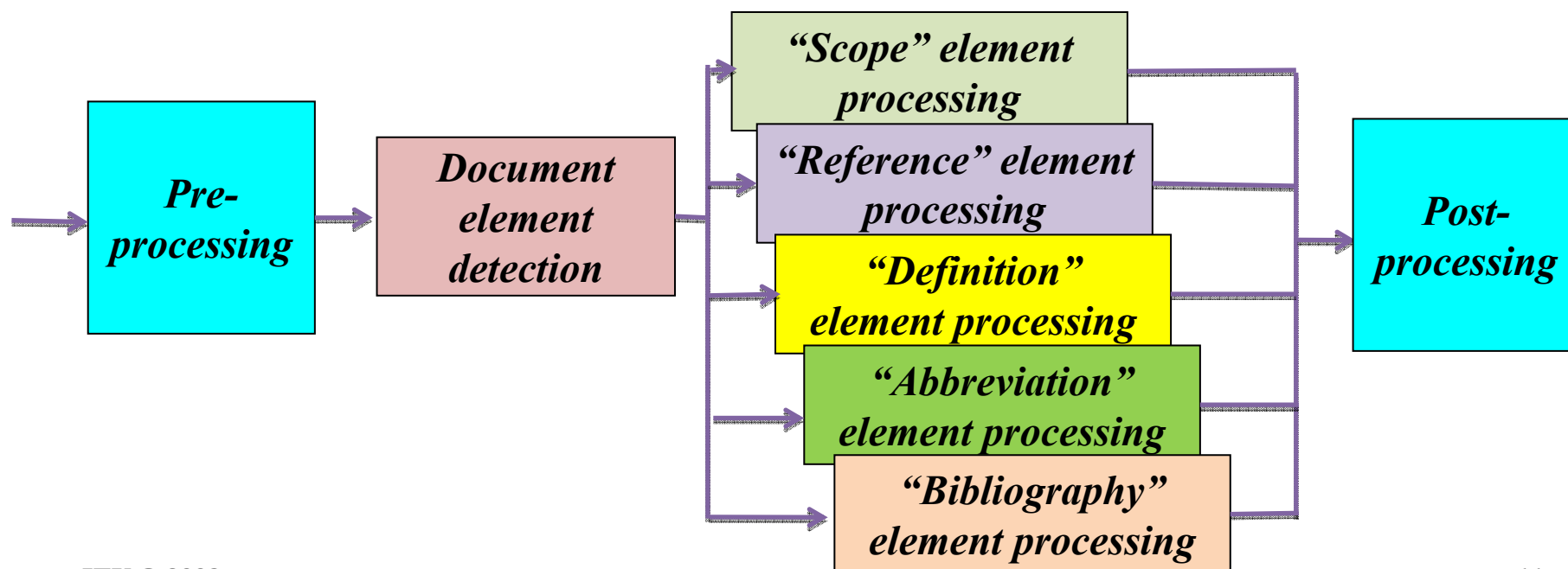
```
<scope>
  <p>Interface Rw defines an interface between a po
  <p>This Recommendation provides a protocol-indepe
  <li>-overview;</li>
  <li>-Rw reference model;</li>
  <li>-functional elements and capabilities.</li>
</scope>
<references>
  <p>The following ITU-T Recommendations and other
  <referenced-document id="ITU-T Q.3300">
    <handle>ITU-T Recommendation Q.3300 (2008)</h
    <title>Architectural framework for the Q.33xx
  </referenced-document>
  <referenced-document id="ITU-T Y.2012">
    <handle>ITU-T Recommendation Y.2012 (2006)</h
    <title>Functional requirements and architectu
  </referenced-document>
  <referenced-document id="ITU-T Y.2111">
    <handle>ITU-T Recommendation Y.2111 (2006)</h
    <title>Resource and admission control functio
  </referenced-document>
</references>
<definitions>
  <clause>
    <title>Terms defined elsewhere</title>
    <p>This Recommendation uses the following ter
    <definition xref="ITU-T Q.3300">
      <term>policy decision physical entity (PD
      <definition-text>
```

**Word document**

**ITU-T XML document**

## 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”.

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



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

**ITU-T XML**

## 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>.

```
<w:p>
  <w:r>
    <w:t>The following ITU-T Recommendations and other re
  </w:r>
</w:p>
<w:p>
  <w:pPr>
    <w:pStyle w:val="Reftext"/>
  </w:pPr>
  <w:r>
    <w:t>[ITU-T E.106] </w:t>
    <w:tab/>
    <w:t>ITU-T Recommendation E.106 (2003), International
  </w:r>
</w:p>
```

**WordXML**

```
<references>
  <p>The following ITU-T Recommendations and other references
  <referenced-document id="ITU-T E.106">
    <handle>ITU-T Recommendation E.106 (2003)</handle>
    <title>International Emergency Preference Scheme (IEPS)
  </referenced-document>
  <referenced-document id="ITU-T E.107">
    <handle>ITU-T Recommendation E.107 (2007)</handle>
    <title>Emergency Telecommunications Service (ETS) and in
  </referenced-document>
```

**ITU-T XML**

## 2-3-4 “Reference” element processing (continued)

[ITU-T F.703]	<u>Recommendation ITU-T F.703 (2000), <i>Multimedia conversat</i></u> < <a href="http://www.itu.int/rec/T-REC-F.703">http://www.itu.int/rec/T-REC-F.703</a> >
[ITU-T F.790]	<u>Recommendation ITU-T F.790 (2007), <i>Telecommunications a</i></u> <u><i>guidelines for older persons and persons with disabilities.</i></u> < <a href="http://www.itu.int/rec/T-REC-F.790">http://www.itu.int/rec/T-REC-F.790</a> >
[ITU-T F.902]	<u>Recommendation ITU-T F.902 (1995), <i>Interactive services de</i></u> < <a href="http://www.itu.int/rec/T-REC-F.902">http://www.itu.int/rec/T-REC-F.902</a> >

### Standard case

The “pStyle is to be set to ‘enumlev1’

<b>2.1</b>	<b>Identical Recommendations   International Standards</b>
-	<u>ITU-T Recommendation X.207 (1993)   ISO/IEC 9545:1994, <i>Information techn</i></u> <u><i>Interconnection – Application layer structure.</i></u>
-	ITU-T Recommendation X.500 (2008)   ISO/IEC 9594-1:2008, <i>Information techn</i> <i>Interconnection – The Directory: Overview of concepts, models and services.</i>
-	ITU-T Recommendation X.501 (2005)   ISO/IEC 9594-2:2005, <i>Information techn</i> <i>Interconnection – The Directory: Models.</i>

### 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

```
<w:t>Terms defined elsewhere</w:t>
</w:r>
</w:p>
<w:p>
  <w:r>
    <w:t>This Recommendation uses the following t
  </w:r>
</w:p>
<w:p>
  <w:r>
    <w:t>3.1.1</w:t>
    <w:tab/>
    <w:t>accounting [ITU-T X.462]: The action of
  </w:r>
</w:p>
```

### WordXML

```
<definitions>
  <clause>
    <title>Terms defined elsewhere</title>
    <p>This Recommendation uses the following
    <definition xref="ITU-T X.462">
      <term>accounting</term>
      <definition-text>
        <p>The action of collecting infor
      </definition-text>
    </definition>
    <definition xref="ITU-T Y.2091">
      <term>address</term>
      <definition-text>
        <p>An address is the identifier
```

- 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 include <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 provide content assets.



```
<definitions>
  <clause>
    <title>Terms defined elsewhere</title>
    <p>This Recommendation uses the following terms defined elsewhere:</p>
    <definition xref="b-ITU-T Y.101">
      <term>application</term>
      <definition-text>
        <p>A structured set of capabilities, which provide value-added
      </definition-text>
    </definition>
  </clause>
</definitions>
```

**(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;



```
<definitions>
  <clause>
    <title>Terms defined elsewhere</title>
    <p>This Recommendation uses the following terms defined in [ITU-T G.661]:</p>
    <definition>
      <term>channel addition/removal (steady-state) gain response;</term>
    </definition>
```

**(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]**



```
<definitions>
  <clause>
    <title>Terms defined elsewhere</title>
    <p>This Recommendation uses the following terms defined elsewhere:</p>
    <definition>
      <term>agent</term>
      <definition-text>
        <p>[ITU-T X.701]</p>
      </definition-text>
    </definition>
```

**(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 hierarchal <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>.

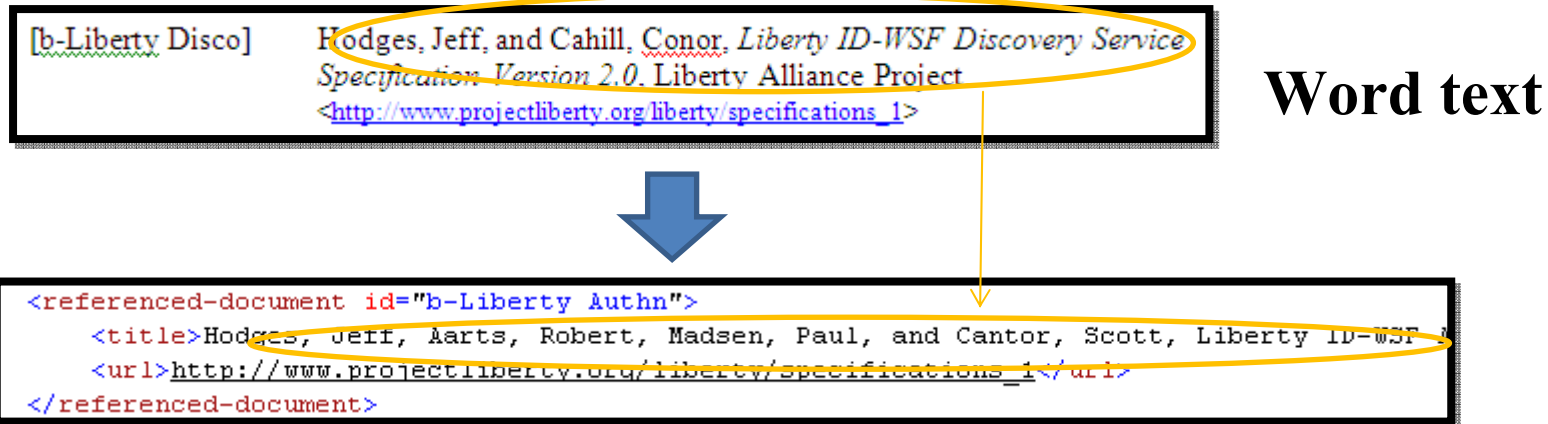
```
<w:p>
  <w:pPr>
    <w:pStyle w:val="Heading1"/>
  </w:pPr>
  <w:r>
    <w:t>4</w:t>
    <w:tab/>
    <w:t>Abbreviations and acronyms</w:t>
  </w:r>
</w:p>
<w:p>
  <w:r>
    <w:t>This Recommendation uses the following
  </w:r>
</w:p>
<w:p>
  <w:r>
    <w:t>ANI</w:t>
    <w:tab/>
    <w:tab/>
    <w:t>Application Network Interface</w:t>
  </w:r>
</w:p>
```

```
<abbreviations>
  <p>This Recommendation uses the following abbrev
  <definition>
    <term>ANI</term>
    <definition-text>Application Network Interfa
  </definition>
  <definition>
    <term>API</term>
    <definition-text>Application Programming Int
  </definition>
```

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>.



## 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

```
<document>
  <head>
    <organization>ITU</organization>
    <universal-id/>
    <language>en</language>
    <document-number>ITU-T H.720</document-number>
    <approval-date>2008-10-14</approval-date>
    <publication-date>2009-08-10</publication-date>
    <title>Overview of IPTV terminal devices and end systems</title>
    <section-level0>Audiovisual and multimedia systems</section-level0>
    <section-level1>IPTV multimedia services and applications for IPTV</section-level1>
    <section-level2>IPTV terminal devices</section-level2>
    <itu-metadata>
      <itu-sector>ITU-T</itu-sector>
      <doc-type>Recommendation</doc-type>
      <itu-id>9560</itu-id>
      <url>http://www.itu.int/itu-t/recommendations/rec.aspx?id=9560</url>
      <main-edition>1</main-edition>
      <sub_edition>0</sub_edition>
      <sg>16</sg>
      <approval-process>AAP</approval-process>
      <equivalent-standards/>
      <history>
        <edition>
          <itu-id>9560</itu-id>
          <main-edition>1</main-edition>
          <sub-edition>0</sub-edition>
          <name>H.720</name>
        </edition>
      </history>
    </itu-metadata>
  </head>
</document>
```

**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.

**3 Definitions and abbreviations**

**3.1 Terms defined elsewhere**

This Recommendation uses the following terms defined elsewhere:

**3.1.1 emergency telecommunications service** (priority communications to facilitate the work of emergency services (ITU-T Rec. E.107).)

**3.1.2 user:** A user includes end user (ITU-T Rec. E.107), equipment, terminal (e.g., FAX, PC), (functional) end user network.

**3.2 Terms defined in this Recommendation**

This Recommendation defines the following terms:

**3.2.1 asset:** Anything that has value to the organization for its continuity.

.....

**3.3 Abbreviations and acronyms**

This Recommendation uses the following abbreviations and acronyms:

3G ..... 3rd Generation

**3 Definitions and abbreviations**

**3.1 Terms defined elsewhere**

This Recommendation uses the following terms defined elsewhere:

**3.1.1 emergency telecommunications service** (priority communications to facilitate the work of emergency services (ITU-T Rec. E.107).)

**3.1.2 user:** A user includes end user (ITU-T Rec. E.107), equipment, terminal (e.g., FAX, PC), (functional) end user network.

**3.2 Terms defined in this Recommendation**

This Recommendation defines the following terms:

**3.2.1 asset:** Anything that has value to the organization for its continuity.

.....

**4. Abbreviations and acronyms**

This Recommendation uses the following abbreviations and acronyms:

3G ..... 3rd Generation

Set “Heading1” as style





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

**Word text**

```
<definitions>  
<p>This Recommendation defines the following terms:</p>  
<definition>  
<term>4</term>  
<definition-text>  
<p>A notation that defines the relative horizontal resolution of the raster to be equal.</p>  
</definition-text>  
</definition>
```



```
<definitions>  
<p>This Recommendation defines the following terms:</p>  
<definition>  
<term>4:4:4</term>  
<definition-text>  
<p>A notation that defines the relative horizontal resolution of the raster to be equal.</p>  
</definition-text>  
</definition>
```

## 2-5 Experimental results

- Applied to the 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,  $j_x$  and  $j_y$  is defined as:

$$\vec{j} = \begin{bmatrix} \cos \theta \exp(-i\mu/2) \\ \sin \theta \exp(i\mu/2) \end{bmatrix} \quad (3-4)$$

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”.

## 2-5 Experimental results

### -Examples of difficult cases(2)-

#### 3 Definitions

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

Table 1 – List of definitions

Name	Description	Unit
$AD$	Absolute audiovisual delay	–
$b_n$	Video bit rate ( $n = 1, 2, \dots, N$ )	kbit/s
$Bpl_s$	Speech packet-loss robustness	–
$Br_V$	Video bit rate	kbit/s
$D_{bmf}$	Degree of video quality robustness against packet loss ( $n = 1, 2, \dots, N$ , $m = 1, 2, \dots, M$ )	–
$D_{F+V}$	Degree of video quality robustness due to frame rate reduction	–
$D_n$	Degree of video quality robustness due to frame rate reduction ( $n = 1, 2, \dots, N$ )	–

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

## 2-5 Experimental results

### -Examples of difficult cases(3)-

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

$$\Delta_{S_{mi}} = S_{\tau i} (\text{diff}) - S_{\tau i} (\text{direct})$$

$D$  is computed as a weighted average of  $\Delta_{S_{mi}}$ .

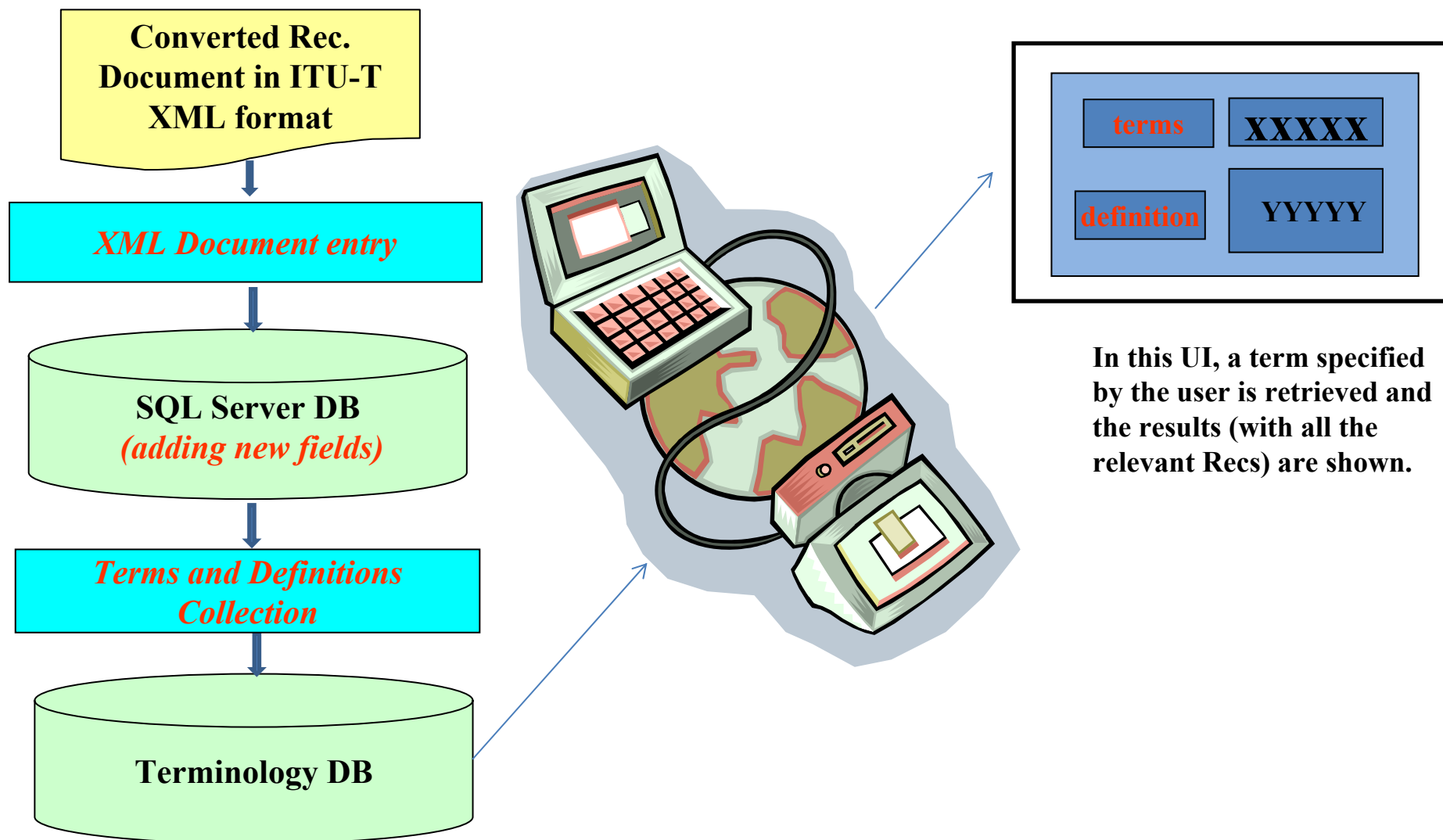
**3.5 ear-drum reference point (DRP):** Point located at the end of the ear canal, corresponding to the ear-drum position.

**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”.

### 3. Application to terms and definitions processing

# 3-1 terms and definitions processing





## 3-2 Experimental results

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

	A	B	C	D	E	F	G
1	isn_def	IDREC	NOM	DATE_APPR	term	abbrev	definition
2	4510	5702	D.000	2002-06-14	accounting rate		The rate agreed between Administrations in a given relation that is used for the establishment of
3	4511	5702	D.000	2002-06-14	settlement rate		A rate agreed between involved administrations/ROAs for terminating incoming traffic.
4	4512	5702	D.000	2002-06-14	termination charge		A charge set by the destination administration/ROA for terminating incoming traffic regardless of origin.
5	4513	5702	D.000	2002-06-14	collection charge		The charge established and collected by an Administration from its customers for the use of an
6	4514	5702	D.000	2002-06-14	lease		An agreement whereby a certain facility is made available by an Administration or Administrations to a customer or customers for his or their exclusive use.
7	4515	5702	D.000	2002-06-14	rental		Payment(s) due to Administrations for the provision of certain facilities or access to certain facilities/services
8	4516	5702	D.000	2002-06-14	network (service) access component	service	A tariff component, normally intended to compensate Administrations for the facilities required for a customer to access a service or services, which is independent
9	4517	5702	D.000	2002-06-14	network (service) utilization component	service	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
10	4518	5702	D.000	2002-06-14	service invocation component		A tariff component which is normally intended to cover the per event cost of activating a service, already

# 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.