IT Standard Technology Exchange Seminar in Beijing

Today's Topic on Document Processing Technology -- Modularization of Document Descriptions

Yushi Komachi

Panasonic Communications, Shimomeguro, Tokyo Japan
email: komachi@y-adagio.com

2006-01-19

Abstract

Background and user requirements for today's electronic document descriptions are shown. Some related activities taken by the Japanese Government are introduced. They lead to a new topic of document processing technology: Modularization of Document Descriptions.

---

1. Background

1.1 XML data and the style specification

The XML is an internationally approved standard for structured data description. It is actually employed to describe web documents, newspapers (NewsML), multimedia contents (SMIL), etc.

For visible presentation of the XML data, some rendering systems such as formatters are required. User specification for the rendering system is a style specification, which can be interchanged as well as XML data. Style specification can be carried out by style specification languages such as DSSSL (ISO/IEC 10179) and XSL.

1.2 DocSII project (by CICC, 2002 through 2004)

The DocSII project focused on the style specification and made it easy to describe style specification of complicated Asian documents by developing style specification libraries (DSSSL libraries, Amds to ISO/IEC TR 19758). Those have been published by ISO[1].

1.3 XML in e-Government

The XML is actually used for describing a number of government doscuments[2]; existing and new documents. It is an extremely large work and description efficiency becomes essential.

---

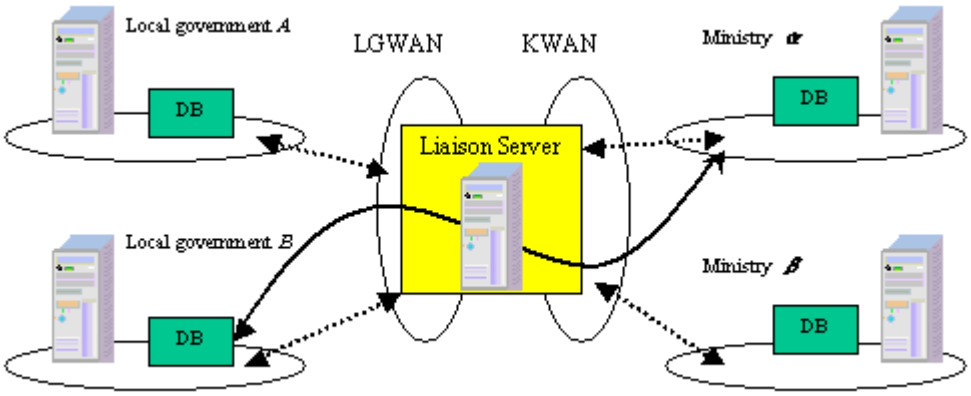
2. Activities in e-Government in Japan

2.1 Oberview

From "Periodical Reports from Local Governments to the Japanese Central Government", by M. Murata (XTech 2005, 2005-05)

Introduction

Ministry of Internal Affairs and Communications (MIC) in Japan created *E-Local Governments System Development Panel*. The goal of this panel is to allow e-local governments in Japan to be easily and consistently developed. One of the working groups reporting to this panel is *Working Group for liaison between the central government and local governments*. Under the auspices of this working group, a number of developers and officials conduct a project for building such a liaison system.



This project is initially concerned with periodical reports from local governments to the Ministry of Internal Affairs and Communications, but is expected to cover other information interchange between local governments and ministries. The focus in 2004 is design and analysis of periodical reports and workflow around them. In 2005, we plan to continue the design and analysis and further implement a pilot system for liaison between the Central Government and Local Governments.

Simply put, local governments in Japan are Prefectures, Government Ordinance Cities, or Municipalities. Municipalities report to prefectures, while government ordinance cities report to prefectures as well as the central government. Local governments depend on the central government for most funding.

Data standardization

Document analysis

There are 293 periodical reports from local governments to the Ministry of Internal Affairs and Communications. Moreover, each of them has one version for prefectures and another for municipalities.

We analyzed two periodical reports. They are dominated by numerical data and have mostly tabular structures, although other periodical reports contain prose as well as numerical data. As usual, we eliminate layout information and focus on structural or semantic information.

We find that periodical reports for municipalities and those for prefectures share common structures, but they have many differences. Some of the differences are fundamental and required, but others are caused by mistakes or layout constraints. We have attempted to make the two types of reports more similar by adopting logical decompositions of reports.

We also find that some pieces of information repeatedly appear in several periodical reports. Common schema components are appropriate for such common pieces of information.

Schemas and schema languages

Three grammar-based schema language have come to be widely recognized. They are DTD, W3C XML Schema, and RELAX NG. We have adopted RELAX NG, since (1) it is simple and powerful, and (2) RELAX NG schemas can be automatically converted to W3C XML Schema and DTD by . We created schemas by first creating XML documents by hand and then generating RELAX NG schemas by trang. Then, we converted such RELAX NG schemas to DTD and W3C XML Schema by trang.

Schematron

Periodical reports contain a large number of numerical data, and some of the data are computationally dependent on others. To capture such dependencies, we adopted . Schematron rules use XPath expressions for referencing to elements and attributes in XML documents.

Datatype library

We have developed a set of common datatypes. These datatypes handle the Japanese currency (yen), the Japanese calendar, and so forth. These datatypes are defined in RELAX NG but are derived from those in W3C XML Schema Part 2.

EGIX

We plan to use for representing glyphs that are not available in Unicode.

Future works

Although our schemas use common schema components, we feel that our schemas are not fully modularized yet. For example, our schemas do not directly capture similarities between those periodical reports from prefectures and those from municipalities. Moreover, some groups of periodical reports (e.g., those about governmental subsidy) appear to be similar. We plan to modularize our schemas for directly capturing such similarities.

We also plan to implement user interfaces for editing periodical reports. We are strongly interested in the use of XForms. However, as a short-term solution, we are also interested in using Open Office by converting our XML documents to the Open Document format and vice versa.

Bibliography

[Prefectures] Prefectures of Japan
[Government Ordinance Cities] Government Ordinance Cities
[Municipalities] Municipality of Japan
[RELAX NG] Document Schema Definition Languages -- Regular-grammar-based validation - RELAX NG,
[W3C XML Schema] W3C XML Schema,
Document Schema Definition Languages -- Rule-based validation - Schematron,
Embedding Glyph Identifiers in XML Documents, 20 December 2002
Trang,

2. Activities in e-Government in Japan

2.2 Working efficiency

Works [Structures, XML documents (XML instances)] for 1 month per 1 person

NOTE: Hard copy documents (sheets) are converted into XML documents defining the structures by RELAX NG, W3C XML schema and DTD.

- **RELAX NG schema: 30 hours**
- **Validation of schema and XML documents: 5 hours**
- **Comments description: 10 hours**
- **Document analysis and work report 10 hours**
- **W3C XML schema and DTD (conversion from RELAX NG): 3 hours**
- **XML document: 40 hours**

- **Review: 30 hours**
- **Review meetings: 22 hours**
- **Inclusion of review: 5 hours**

Efficiency

- **Creation of structures and XML documents: 4 sheets/0.6 month*person, 100 tags/0.6 month*person**
- **Review: 12 sheets/0.2 month*person**

  NOTE: Style specification are actually being developed but not yet estimated for the work efficiency.

As a conclusion of those works in 2004, we found that:

- **there are a number of common patterns in the document structure**
- **some modularizations are indispensable for the higher efficiency of works.**

---

3. Modularization support in document description languages

Today's description languages can support modularization functionalities.

3.1 Structure description

RELAX NG[3], [4], for example, can support modularity.

Referencing external patterns:
The external pattern can be used to reference a pattern defined in a separate file. The external keyword is followed by a quoted string specifying the URL of a file containing the pattern. The external pattern matches if the pattern contained in the specified URL matches.

In addition, RELAX NG has the features of
- Combining definitions
- Merging grammars
- Replacing definitions

3.2 Style specification

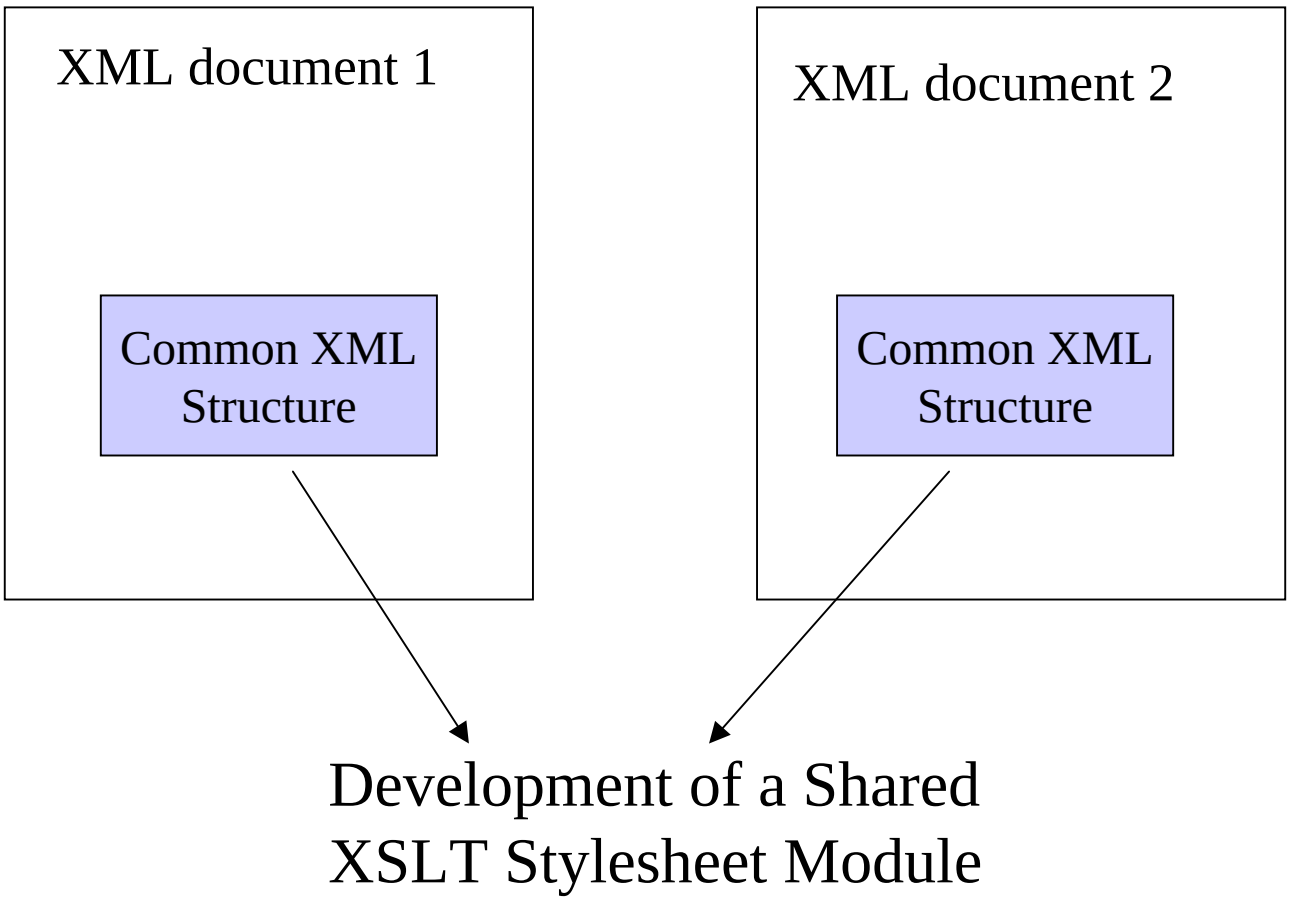
XSLT provides two mechanisms to combine stylesheets:
- <xsl:include href = uri-reference /> an inclusion mechanism that allows stylesheets to be combined without changing the semantics of the stylesheets being combined, and
- <xsl:import href = uri-reference /> an import mechanism that allows stylesheets to override each other.

---

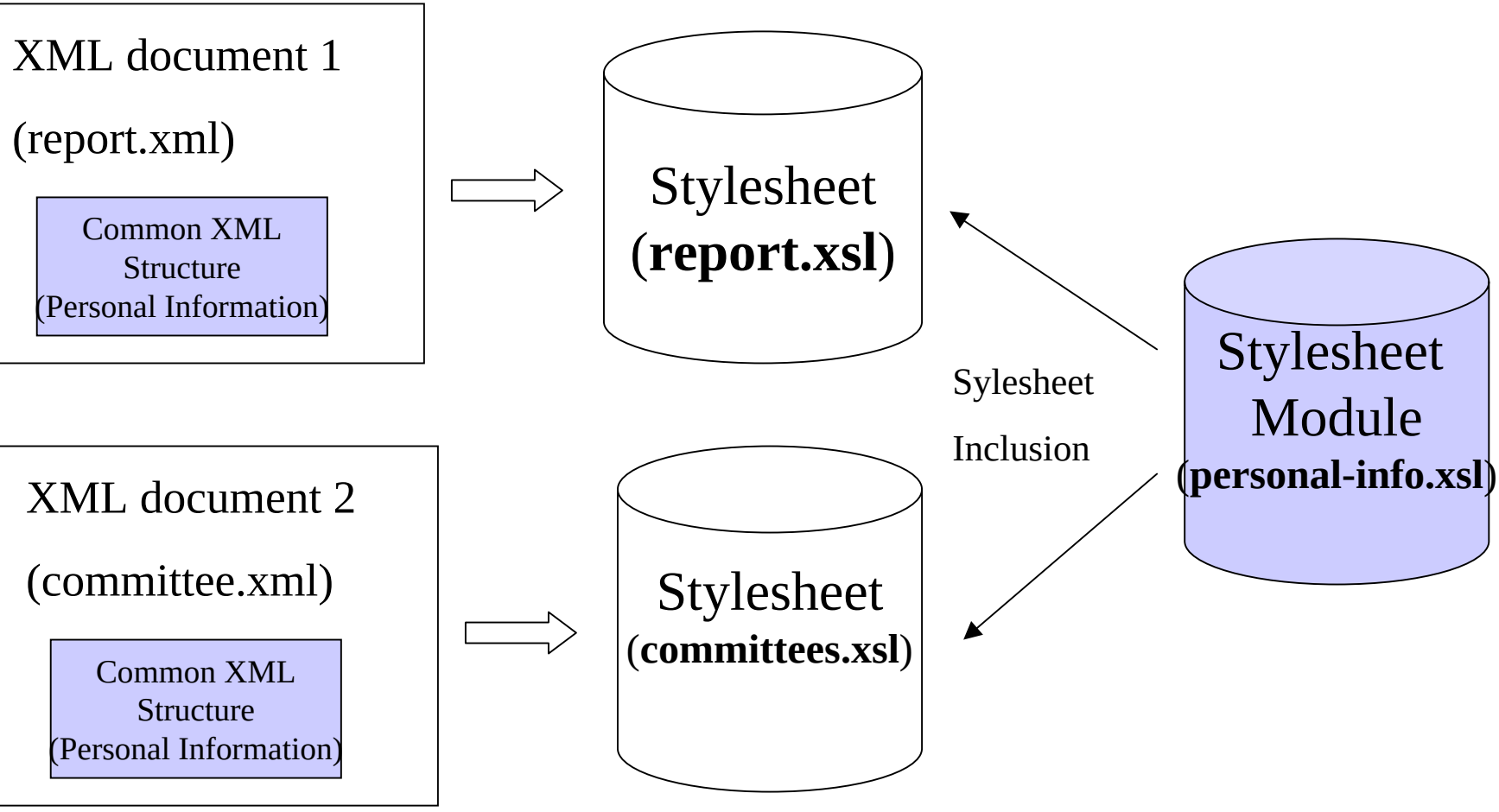
4. Examples for modularization 1

Schematics

Development of Modularized XLST Stylesheet



An Example of Modularized XLST Stylesheet



5. Examples for modularization 2

Files

Report

- **XML source**
- **Structure description (XML Schema description)**
- **Style specification (XSLT Stylesheet)**
- **Rendered output**

Committee

- **XML source**
- **Structure description (XML Schema description)**
- **Style specification (XSLT Stylesheet)**
- **Rendered output**

Modules

- **Structure module**
  which is imported in the Structure description files.
- **Style specification module**
  which is included in the Style specification files.


```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="report.xsl"?>
<rp:report xmlns:rp="http://www.utj.co.jp/namespaces/report"
xmlns:pi="http://www.utj.co.jp/namespaces/personal-info"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.utj.co.jp/namespaces/report report.xsd"
>
<rp:title>Conference Report</rp:title>
<rp:p>The "Asian IT Standardization Workshop" was successful. It was so fruitful.
Atendees from Japan were as follows.</rp:p>
<pi:personal-info>
<pi:person>
<pi:name>Dr. Yushi Komachi</pi:name>
<pi:organization>Panasonic Communications</pi:organization>
</pi:person>
<pi:person>
<pi:name>Keisuke Kamimura</pi:name>
<pi:organization>GLOCOM</pi:organization>
</pi:person>
<pi:person>
<pi:name>Takayuki Sato</pi:name>
<pi:organization>CICC</pi:organization>
</pi:person>
<pi:person>
<pi:name>Hiroko Shirakura</pi:name>
<pi:organization>CICC</pi:organization>
</pi:person>
</pi:personal-info>
</rp:report>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.utj.co.jp/namespaces/report"
xmlns="http://www.utj.co.jp/namespaces/report"
xmlns:pi="http://www.utj.co.jp/namespaces/personal-info">

  <xs:import namespace="http://www.utj.co.jp/namespaces/personal-info" schemaLocation="personal-info.xsd"/>

  <xs:element name="report">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="title"/>
                <xs:choice minOccurs="0" maxOccurs="unbounded">
          <xs:element ref="p"/>
          <xs:element ref="pi:personal-info"/>
        </xs:choice>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="title" type="xs:string"/>

  <xs:element name="p" type="xs:string"/>

</xs:schema>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:rp="http://www.utj.co.jp/namespaces/report"
>

<xsl:include href="personal-info.xsl"/>

<xsl:template match="/">
          <html>
                    <xsl:apply-templates/>
          </html>
</xsl:template>

<xsl:template match="rp:report">
          <body>
                    <xsl:apply-templates/>
          </body>
</xsl:template>

<xsl:template match="rp:title">
          <h1>
                    <xsl:value-of select="."/>
          </h1>
</xsl:template>

<xsl:template match="rp:p">
          <p>
                    <xsl:value-of select="."/>
          </p>
</xsl:template>

</xsl:stylesheet>
```

Conference Report

The "Asian IT Standardization Workshop" was successful. It was so fruitful. Atendees from Japan were as follows.

Name	Organization
Dr. Yushi Komachi	Panasonic Communications
Keisuke Kamimura	GLOCOM
Takayuki Sato	CICC
Hiroko Shirakura	CICC

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="committees.xsl"?>
<cm:committees xmlns:cm="http://www.utj.co.jp/namespaces/committees"
xmlns:pi="http://www.utj.co.jp/namespaces/personal-info"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.utj.co.jp/namespaces/committees committees.xsd"
>
<cm:committee name="CICC Asian IT Standardization Research Committee">
<cm:chair>
<pi:personal-info>
<pi:person>
<pi:name>Dr. Yushi Komachi</pi:name>
<pi:organization>Panasonic Communications</pi:organization>
</pi:person>
</pi:personal-info>
</cm:chair>
<cm:members>
<pi:personal-info>
<pi:person>
<pi:name>Tatsuo Kobayashi</pi:name>
<pi:organization>Just System</pi:organization>
</pi:person>
<pi:person>
<pi:name>Keisuke Kamimura</pi:name>
<pi:organization>GLOCOM</pi:organization>
</pi:person>
<pi:person>
<pi:name>Tateo Koike</pi:name>
<pi:organization>RICOH Printing Systems</pi:organization>
</pi:person>
<pi:person>
<pi:name>Yasuhiro Okui</pi:name>
<pi:organization>Nihon Unitec</pi:organization>
</pi:person>
</pi:personal-info>
</cm:members>
<cm:secretariat>
<pi:personal-info>
<pi:person>
<pi:name>Takayuki Sato</pi:name>
<pi:organization>CICC</pi:organization>
</pi:person>
<pi:person>
<pi:name>Hiroko Shirakura</pi:name>
<pi:organization>CICC</pi:organization>
</pi:person>
</pi:personal-info>
</cm:secretariat>
</cm:committee>
</cm:committees>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.utj.co.jp/namespaces/committees"
xmlns="http://www.utj.co.jp/namespaces/committees"
xmlns:pi="http://www.utj.co.jp/namespaces/personal-info">

  <xs:import namespace="http://www.utj.co.jp/namespaces/personal-info" schemaLocation="personal-info.xsd"/>

  <xs:element name="committees">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="committee" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="committee">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="chair"/>
        <xs:element ref="members"/>
        <xs:element ref="secretariat"/>
      </xs:sequence>
      <xs:attribute name="name" type="xs:string"/>
    </xs:complexType>
  </xs:element>

  <xs:element name="chair">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="pi:personal-info"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="members">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="pi:personal-info"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="secretariat">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="pi:personal-info"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

</xs:schema>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:cm="http://www.utj.co.jp/namespaces/committees"
>

<xsl:include href="personal-info.xsl"/>

<xsl:template match="/">
          <html>
                    <xsl:apply-templates/>
          </html>
</xsl:template>

<xsl:template match="cm:committees">
          <body>
                    <xsl:apply-templates/>
          </body>
</xsl:template>

<xsl:template match="cm:committee">
          <h1>Committee Name: <xsl:value-of select="@name"/></h1>
          <xsl:apply-templates/>
</xsl:template>

<xsl:template match="cm:chair">
          <h1>Chair</h1>
          <xsl:apply-templates/>
</xsl:template>

<xsl:template match="cm:members">
          <h1>Members List</h1>
          <xsl:apply-templates/>
</xsl:template>

<xsl:template match="cm:secretariat">
          <h1>Secretariat</h1>
          <xsl:apply-templates/>
</xsl:template>

</xsl:stylesheet>
```

Committee Name: CICC Asian IT Standardization Research Committee

Chair

Name	Organization
Dr. Yushi Komachi	Panasonic Communications

Members List

Name	Organization
Tatsuo Kobayashi	Just System
Keisuke Kamimura	GLOCOM
Tateo Koike	RICOH Printing Systems
Yasuhiro Okui	Nihon Unitec

Secretariat

Name	Organization
Takayuki Sato	CICC
Hiroko Shirakura	CICC

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.utj.co.jp/namespaces/personal-info"
xmlns="http://www.utj.co.jp/namespaces/personal-info">

  <xs:element name="personal-info">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="person" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="person">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="name"/>
        <xs:element ref="organization"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="name" type="xs:string"/>

  <xs:element name="organization" type="xs:string"/>

</xs:schema>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:pi="http://www.utj.co.jp/namespaces/personal-info"
>

<xsl:template match="pi:personal-info">
          <table frame="hsides" width="100%">
                    <tr>
                              <th width="20%">Name</th>
                              <th width="80%" align="left">Organization</th>
                    </tr>
                    <xsl:apply-templates/>
          </table>
</xsl:template>

<xsl:template match="pi:person">
          <tr>
                    <xsl:apply-templates/>
          </tr>
</xsl:template>

<xsl:template match="pi:name">
          <td>
                    <xsl:apply-templates/>
          </td>
</xsl:template>

<xsl:template match="pi:organization">
          <td>
                    <xsl:apply-templates/>
          </td>
</xsl:template>

</xsl:stylesheet>
```

6. Conclusion

As a conclusion, the following works are proposed:

- **to scan a number of documents (XMLized or to be XMLized) to find common patters.**
- **to create modules of Structure descriptions and corresponding Style specifications.**
- **to publish sets of the modules to contribute document interchange preserving document styles.**

References

[1] Amendments to ISO/IEC TR 19757, 2005-07(Amd.1 and 2)/08(Amd.3)
[2] M. Murata, Periodical Reports from Local Governments to the Japanese Central Government, XTech 2005, 2005-05
[3] ISO/IEC 19757-2, DSDL Part 2: Regular-grammar-based validation — RELAX NG, 2003-12
[4] FDAM1 to ISO/IEC 19757-2, RELAX NG — Amendment 1: Compact Syntax, 2005-09