Recent Progress on Document Format Research in China

Ning LI,
Beijing Information Technology Institute

Jinghua ZHAO
China Electronics Standardization Institute

2004 2nd DocSII

Contents

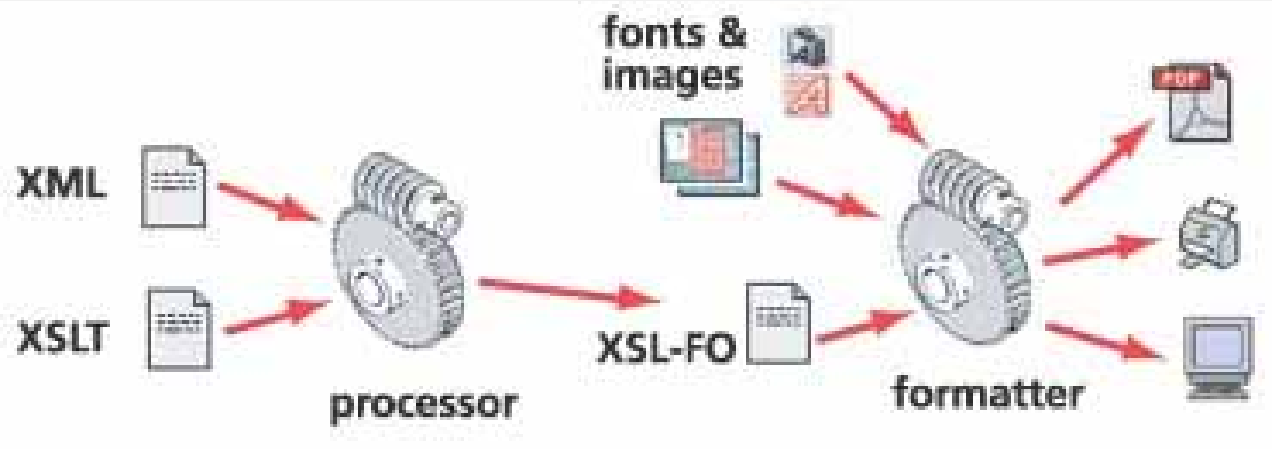
- **Standardization Approach**
  - Standard publications
  - XSL/XSLT practice
- Chinese Office Software Project – UOF
  - Background
  - Major Structures Introduction
- Suggestions and Comments

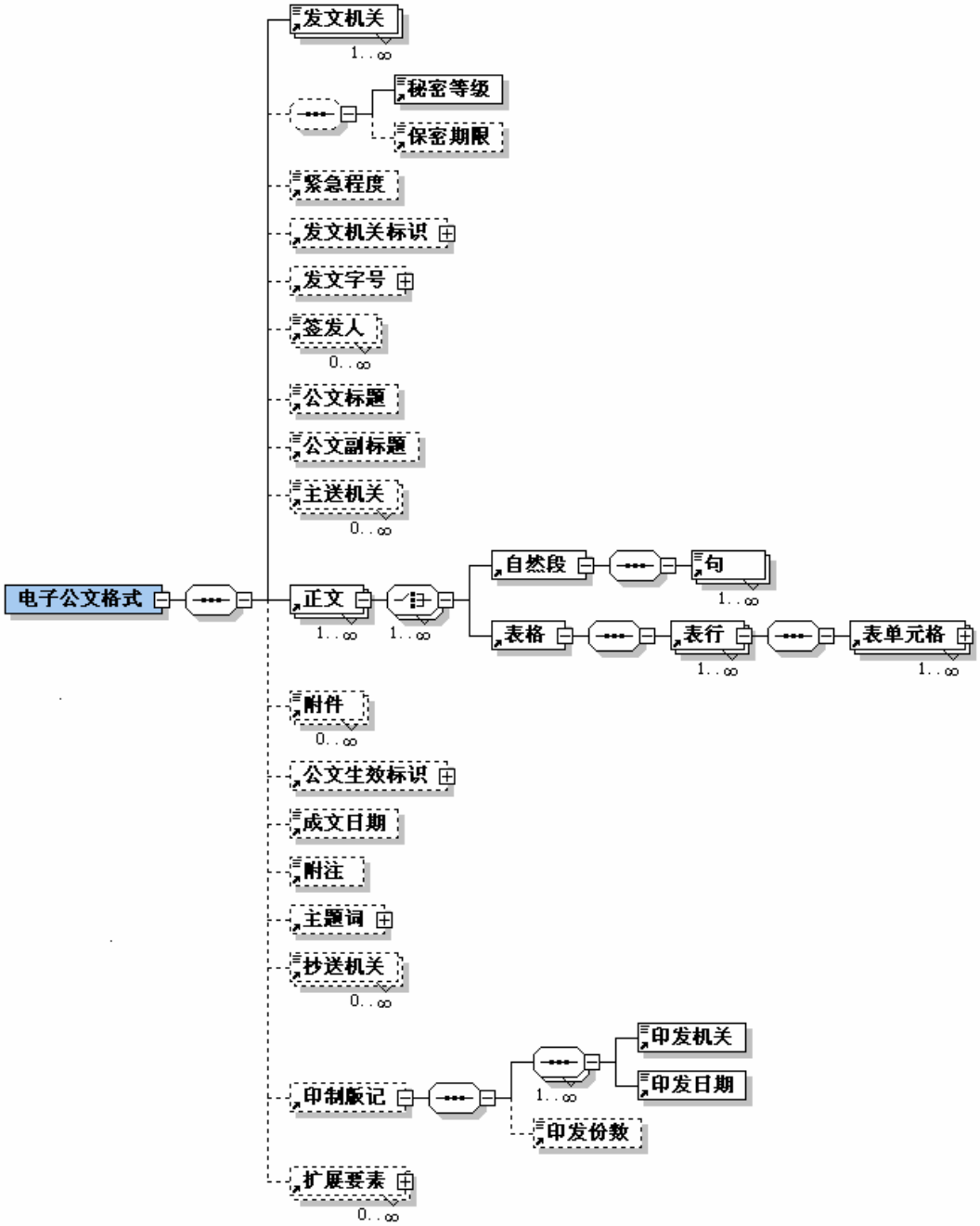
Standard publications: SGML/DSSSL/XML

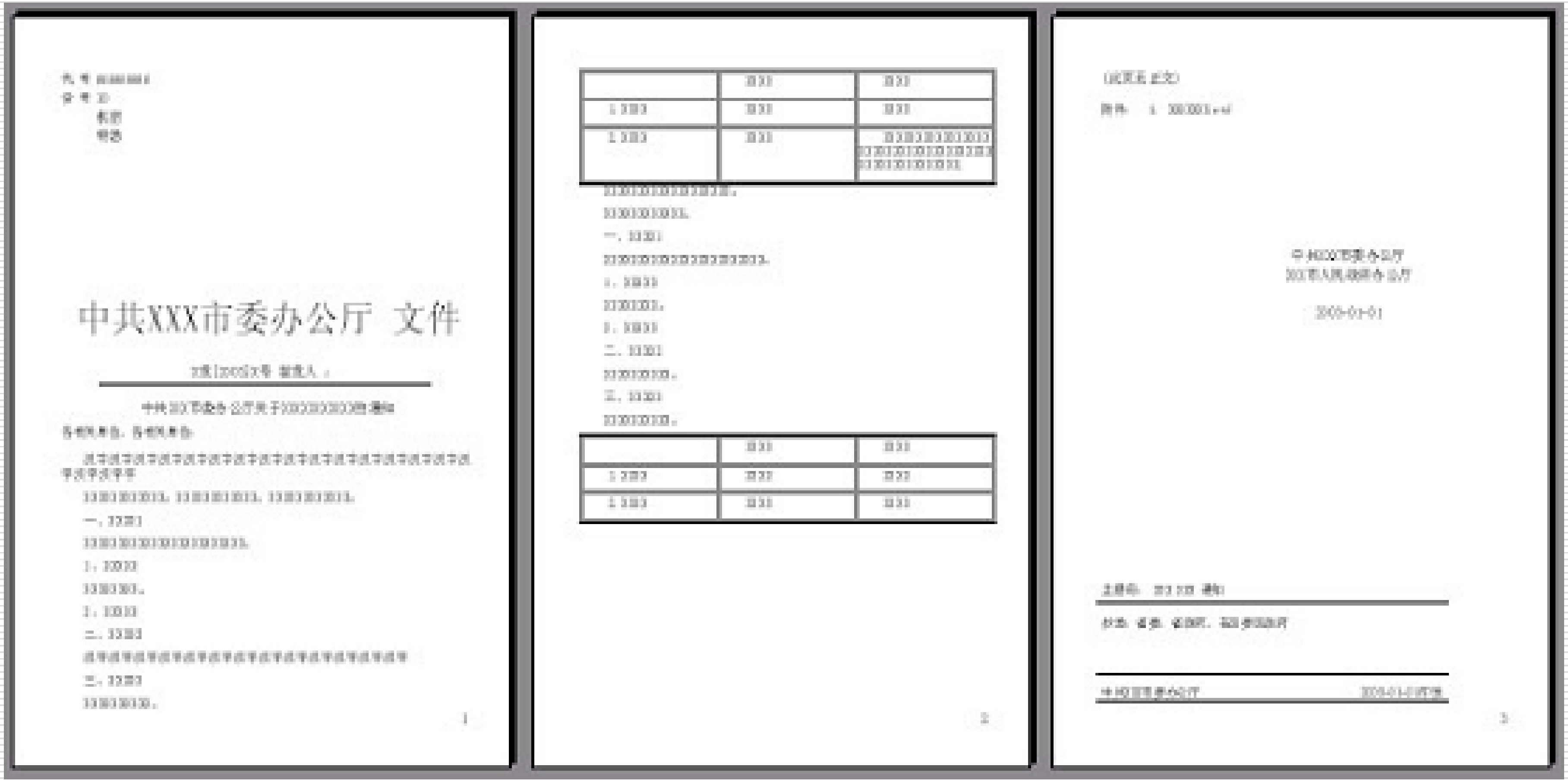
- GB/T 17970-2000信息技術　處理語言　文件式樣的語義及規格說明語言（DSSSL）2000-8-1
- GB/Z 17978-2000信息处理　SGML支持设施 SGML使用技术2000-8-1
- GB/T 18792-2002信息技术 文件描述和处理语言　超文本置标语言（HTML）2002-12-01
- GB/T 18793-2002信息技术 可扩展置标语言（XML）1.02002-12-01
- 信息技术 XML使用指南 (in drafting)

XSL/XSLT practice

- Transformation of ”*Specification for the structure of electronic official document based on XML*” (standard draft)

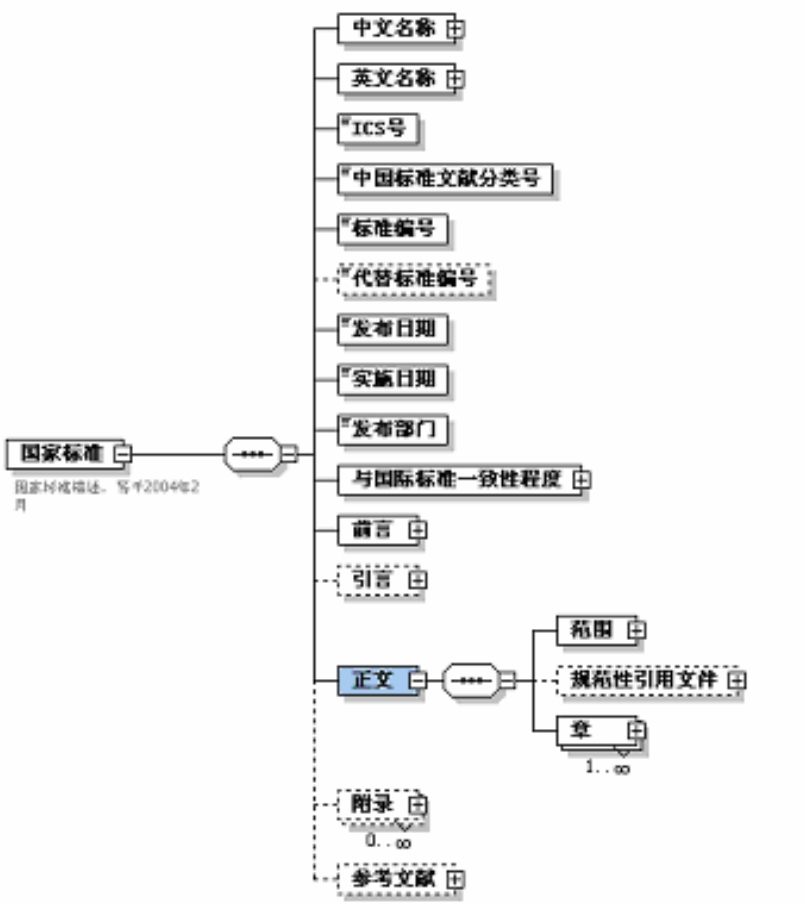



电子公文格式
发文机关
1..∞
秘密等级
保密期限
紧急程度
发文机关标识
发文字号
签发人
0..∞
公文标题
公文副标题
主送机关
0..∞
正文
1..∞
1..∞
自然段
句
1..∞
表格
表行
1..∞
表单元格
1..∞
附件
0..∞
公文生效标识
成文日期
附注
主题词
抄送机关
0..∞
印制版记
1..∞
印发机关
印发日期
印发份数
扩展要素
0..∞



中共XXX市委办公厅 文件
一、XXXX
二、XXXX
三、XXXX
1
一、XXXX
二、XXXX
三、XXXX
2
中共XXX市委办公厅
3

Transformation of national GB standard using XSLT



01.120
A00

GB

中华人民共和国国家标准

GB/T 1.1-2000

标准化工作导则
第一部分：标准的结构和编写规则

Directives for standardization
Part 1:Rules for the structure and
drafting of standards

2000-12-20 发布　　2001-08-01 实施

国家质量技术监督局 发布

GB/T 1.1-2000

1 范围

GB/T1的本部分规定了标准的结构和编写规则，给出了有关表述的一些样式，并提供了标准出版的格式和字体、字号。

本部分适用于国家标准、行业标准和地方标准的编写和出版，企业标准和标准化指导性技术文件的编写可参照使用。

2 规范性引用文件

下列文件中的条款通过GB/T1的本部分的引用而成为本部分的条款。凡是注日期的引用文件，其随后所有的修改单（不包括勘误的内容）或修订版均不适用于本部分。然而，鼓励根据本部分达成协议的各方研究是否可使用这些文件的最新版本。凡是不注日期的引用文件，其最新版本适用于本部分。

GB3101有关量、单位和符号的一般原则（eqvISO31-0）

GB/T788图书杂志开本及其幅面尺寸（eqvISO6716）

3 术语和定义

GB/T3935.1确定的以及下列术语和定义适用于GB/T1的本部分。

3.1

规范性要素 normative elements

要声明符合标准而应遵守的条款的要素，分为一般要素和技术要素。

3.2

资料性要素 informative elements

标识标准，介绍标准，提供标准的附加信息的要素，分为概述要素和补充要素。

3.2.1

概述要素 preliminary elements

标识标准，介绍其内容，背景，制定情况以及该标准与其他标准的关系的要素，即标准的封面、目次、前言和引言等。

3.2.2

补充要素 supplementary elements

提供附加信息，以帮助理解或使用标准的要素，即标准的资料性附录、参考文献和索引等。

3.2.2.1

测试第三层的条

4 总则

4.1 要求

标准所规定的条款应明确而无歧义，并且：

—— 在其范围所规定的界限内按需要力求完整

—— 清楚、准确、相互协调

—— 充分考虑最新技术水平（见3.6）

4.2 统一性

在每项标准或系列标准内，标准的结构、文体和术语应保持一致。系列标准的结构及其章、条的编号应尽可能相同。类似的条款应使用类似措辞来表述；相同的条款应使用相同的措辞来表述。

在每项标准或系列标准内，某一给定概念应使用相同的术语。对于已定义的概念应避免使用同义词。每个选用的术语应尽可能只有唯一的含义。

5 结构

5.1 内容划分

5.1.1 通则

表1给出了标准可能具有的层次名称、层次编号示例参见附录B。

表1 层次及其名称

名称	编号实例
部分	GBXX.1
章	1

在一般情况下，针对每个标准化对象应编制一项单独的标准，并作为整体出版。在诸如下列特殊情况下，可在相同的标准顺序号下将一项标准分成若干个单独的部分：

1） 标准篇幅过长

2） 后续部分的内容相互关联

3） 准的某些部分可能被法规引用

如果产品的不同方面会分别引起各方（例如生产者、认证机构、立法机关等）的关注，则这些不同方面可被编制成一项标准的若干部分或若干项单独的标准。例如，这些不同方面有：

a） 健康和安全要求

b） 性能要求

5.1.2 单项标准的内容划分

可按下列两种方式对一项标准的要素分类：

术语标准在内容划分上具有不同的要求，见附录C。

表2 标准中要素的类型编排序例

要素类型	要素的编排	要素所允许的内容
资料性概述要素	封面	名称

由要素的规范性或资料性的性质以及它们在标准中的位置来划分，可分为：

a） 资料性概述要素（见3.2.1）

b） 规范性一般和技术要素（见3.1）

5.2 层次的编排和编号

Contents

- Standardization Approach
  - Standard publications
  - XSL/XSLT practice
- **Chinese Office Software Project – UOF**
  - Background
  - Major Structures Introduction
- Suggestions and Comments

UOF (Unified Office Format) Background

- The proposed document format of Chinese office software
- A project supported by the Hi-Tech Research and Development Program of China (863 Project) since 2003
- XML-based Chinese Office Software Document Format Working Group was set up by China National IT Standardization Technical Committee in 2002. The Group is consisted of domestic IT standard body, key office software enterprises, research institutes, universities, and also some non-voting members

UOF Targets

- Its purpose is to develop an XML-based document format for Chinese office software in order to facilitate document interchange and information sharing
- The incoming format should be open, which means the document format is completely described in publicly accessible documents, and it could be distributed freely and implemented in programs without restrictions, royalty-free, and with no legal bindings

Three Basic components

- At the first stage, the project is focused on the requirements coming from the three major components of office software
  - word processor
  - spreadsheet processor
  - presentation processor

Three Specification Drafts

- As the result of first year's work, the following specification drafts are proposed:
  - UOF document format
  - UOF storage format
  - Application program Interface (API)
  - User interface of Chinese office software

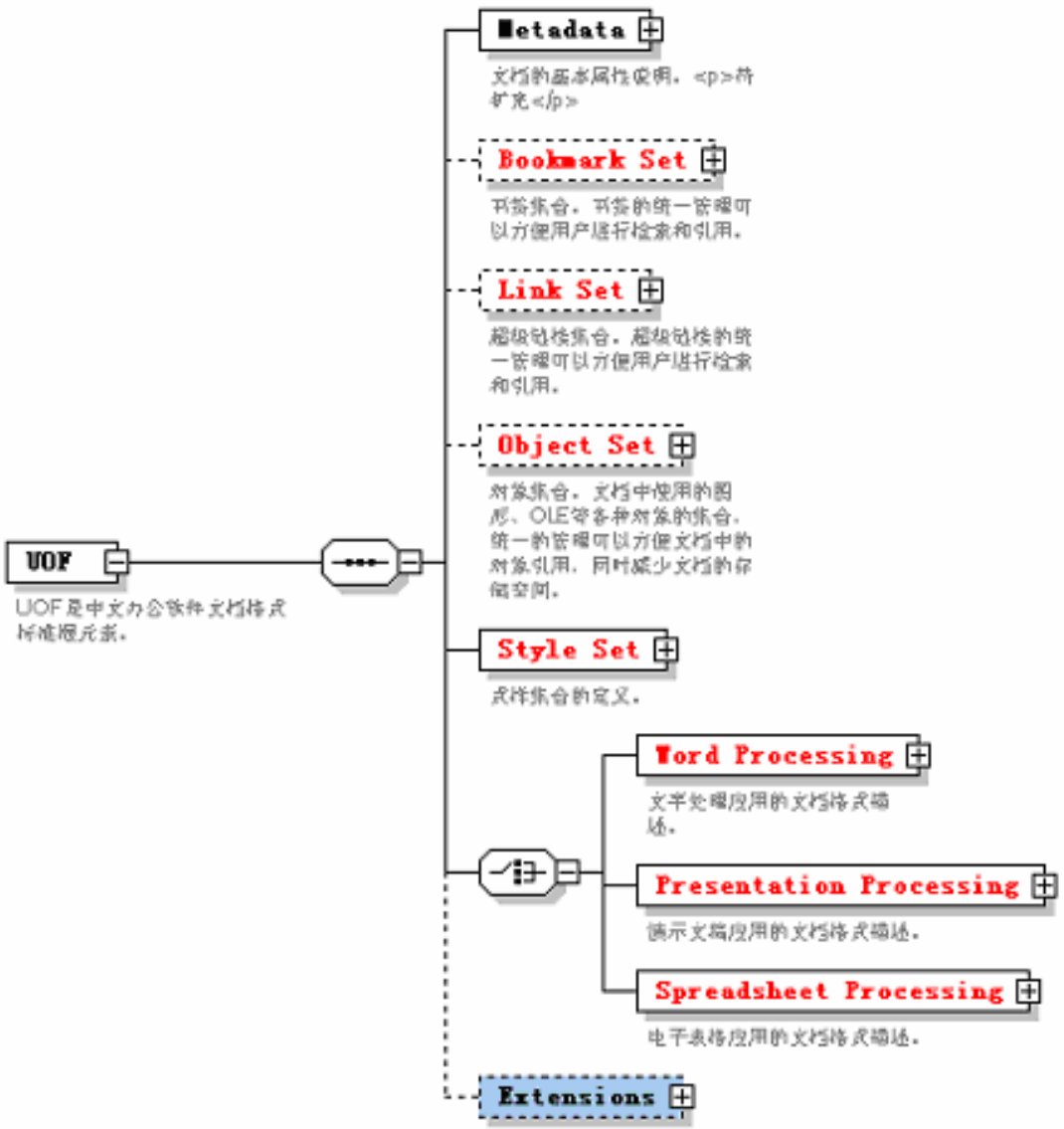
A Comparison of Various Formats

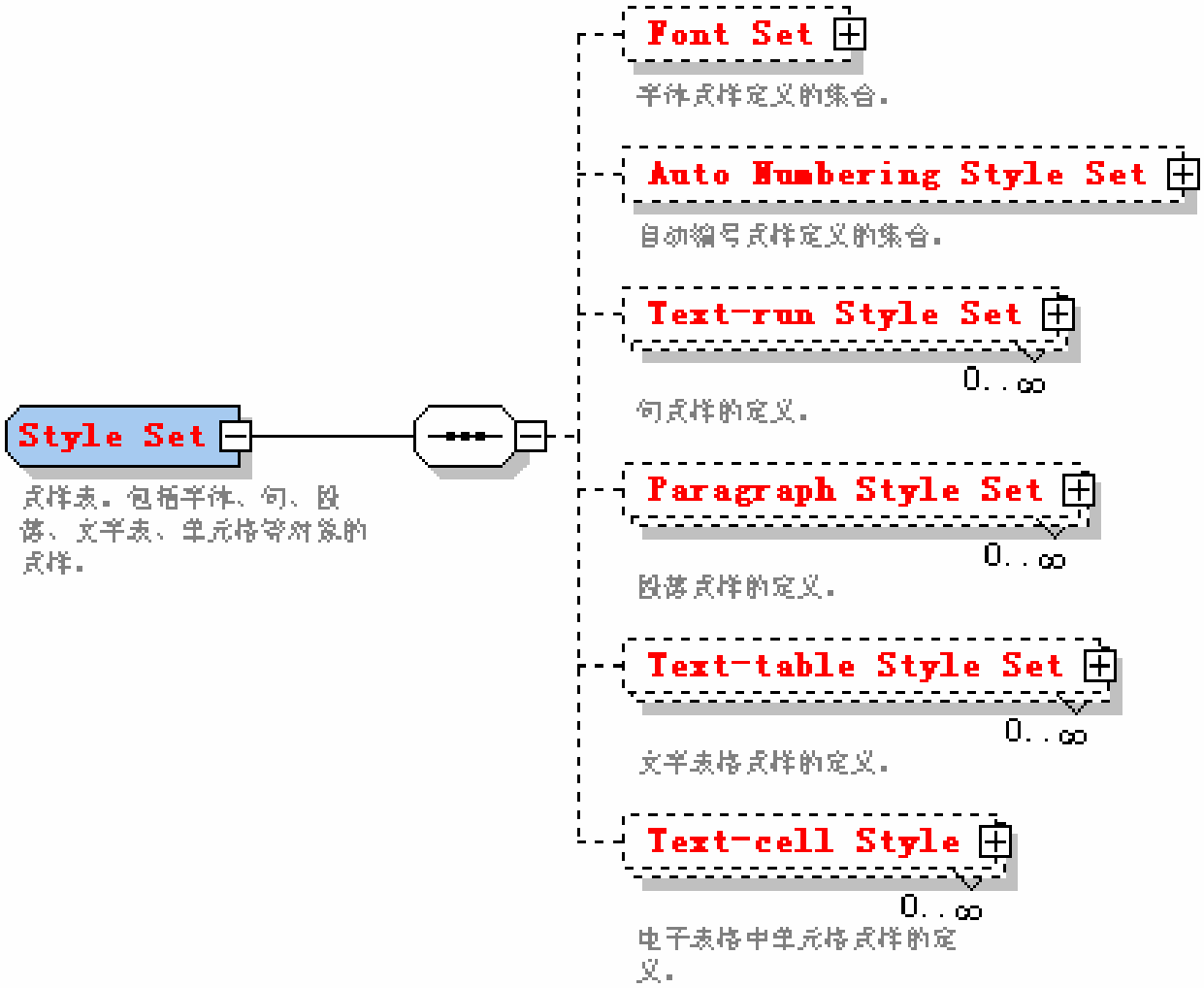
Document format	Open	Chinese support	Modifiable	PR constrains	Suitable for editing	Cross platform	Fidelity	Others
WordML	Y	Y	N	?	Y	N	M	Well designed
OpenOffice	Y	M	N	N	Y	Y	M	Too many attributes used
RedOffice	Y	Y	Y	N	Y	Y	M	
XML-FO	Y	N	Y	N	N	Y	M	
DSSSL	Y	N	Y	N	N	Y	M	Not widely adopted
PDF	N	Y	N	Y	N	Y	Y	

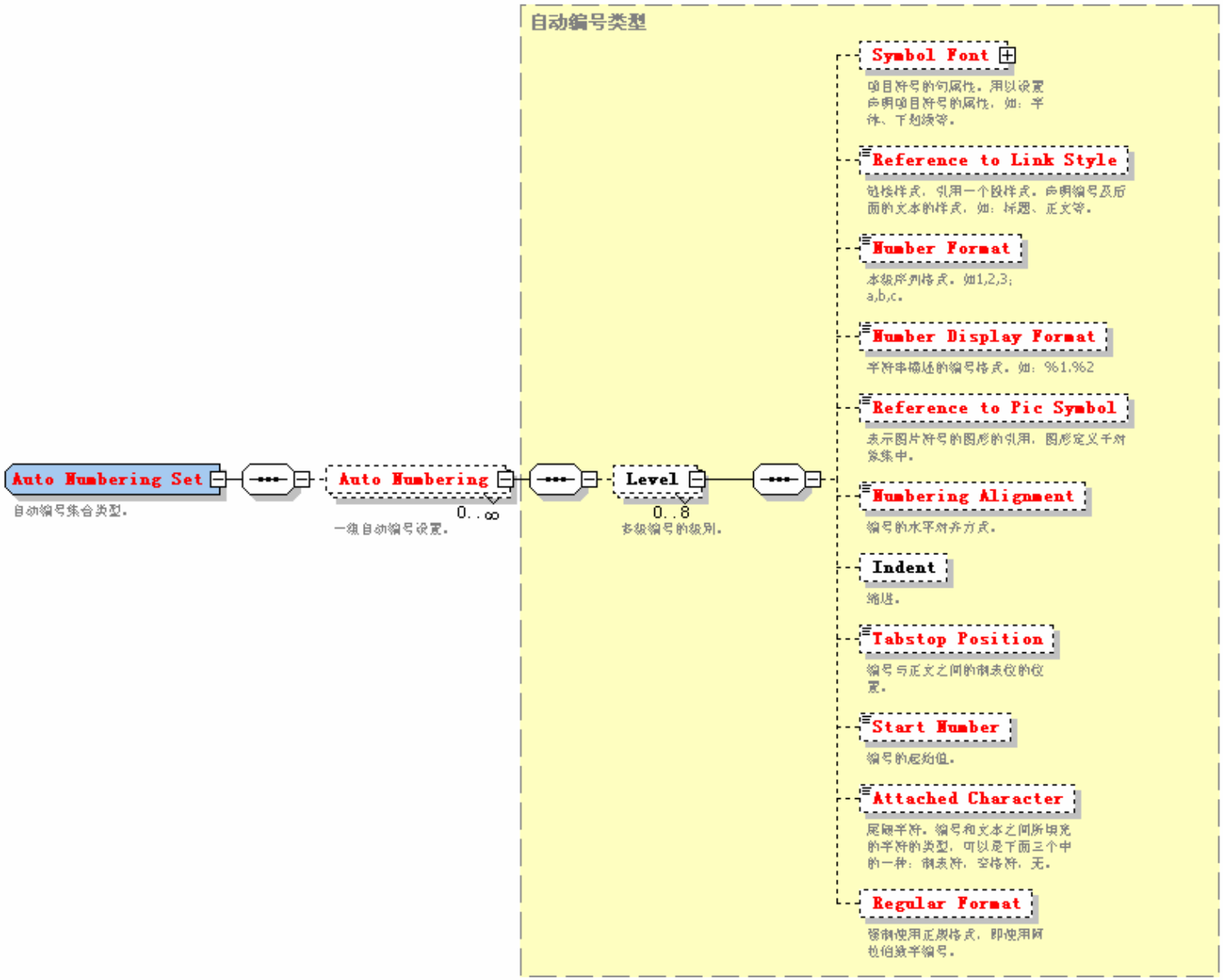
Key Points

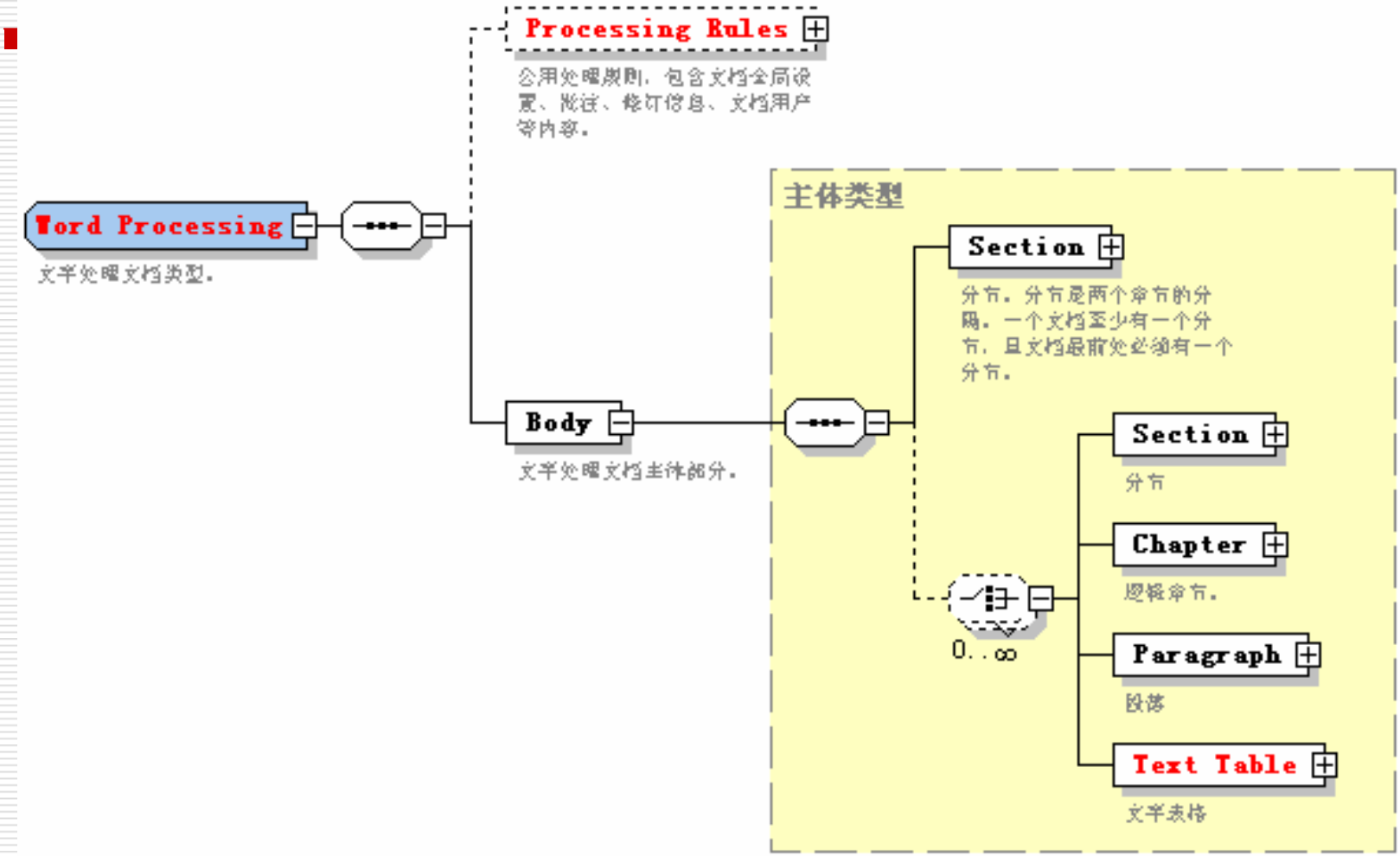
- Reserve editing semantics (e.g., how to express “emphasis”?)
- Support user XML data (how to incorporate User Defined Schema – USD?)
- Style referencing to reduce size and for better reusability
- Use more elements and less attributes
- Allow to use localized tag and attribute names
- Support future extensions

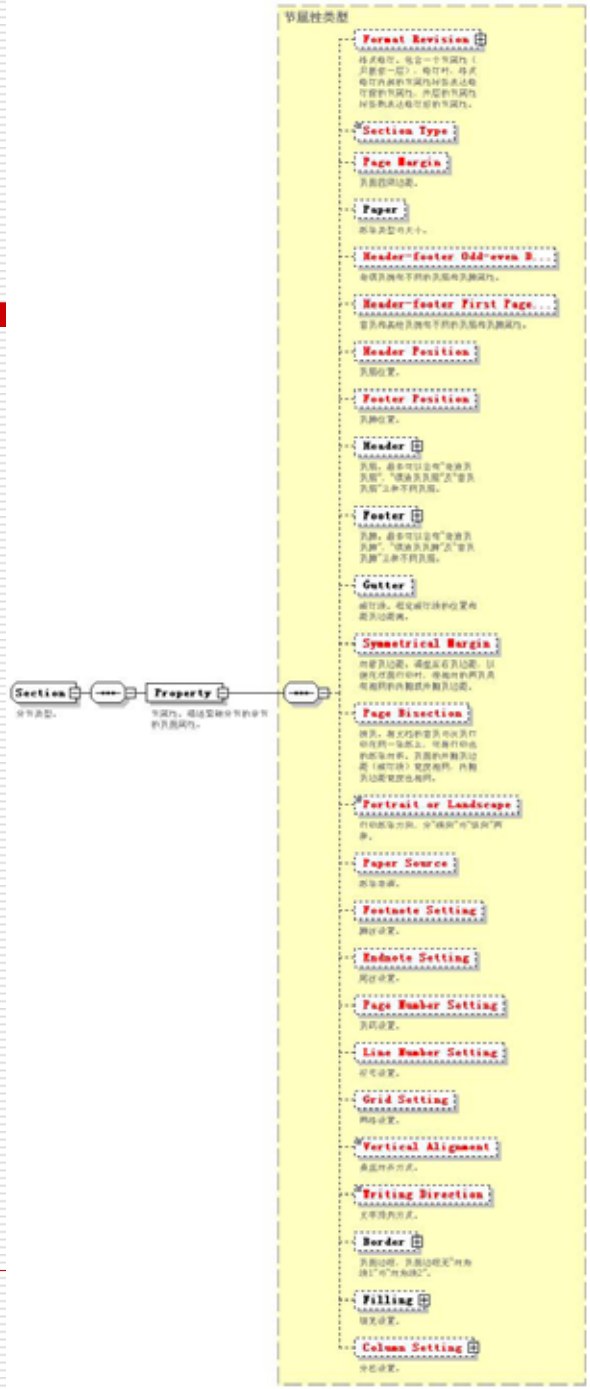
UOF Major Structures

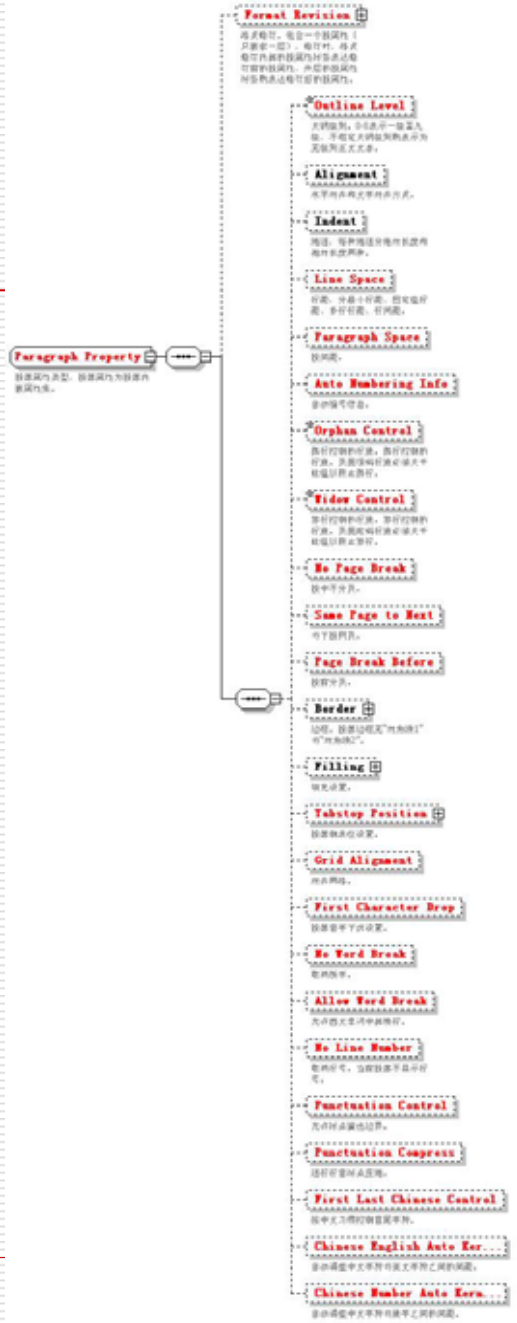

UOF
UOF是中文办公软件文档格式标准根元素。
Metadata
文档的基本属性说明。 并扩充
Bookmark Set
书签集合。书签的统一管理可以方便用户进行检索和引用。
Link Set
超级链接集合。超级链接的统一管理可以方便用户进行检索和引用。
Object Set
对象集合。文档中使用的图形、OLE等各种对象的集合。统一的管理可以方便文档中的对象引用，同时减少文档的存储空间。
Style Set
式样集合的定义。
Word Processing
文字处理应用的文档格式描述。
Presentation Processing
演示文稿应用的文档格式描述。
Spreadsheet Processing
电子表格应用的文档格式描述。
Extensions

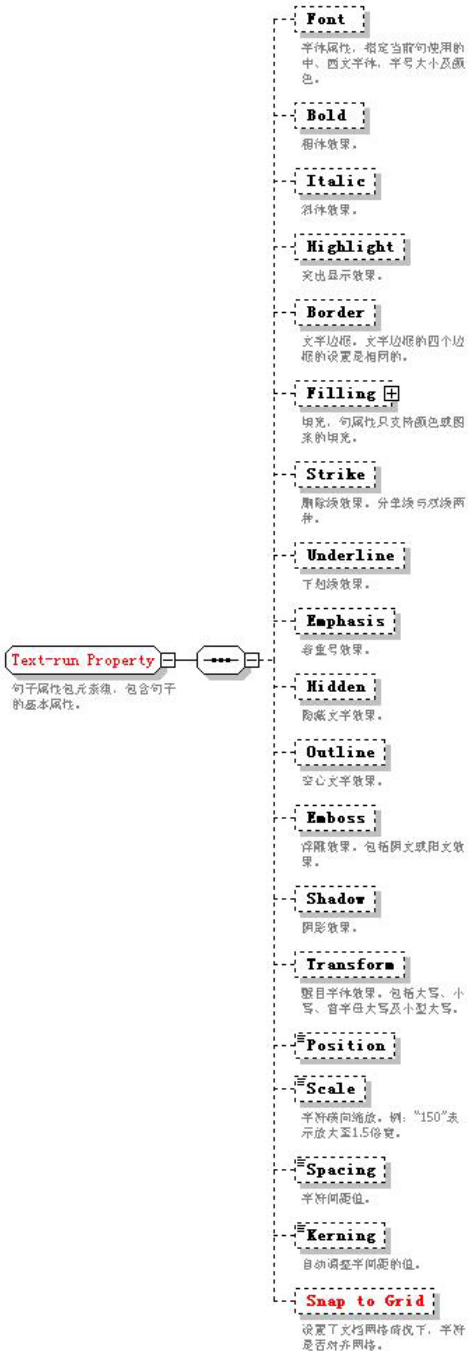

Style Set
式样表。包括字体、句、段落、文字表、单元格等对象的式样。
Font Set
字体式样定义的集合。
Auto Numbering Style Set
自动编号式样定义的集合。
Text-run Style Set
0..∞
句式样的定义。
Paragraph Style Set
0..∞
段落式样的定义。
Text-table Style Set
0..∞
文字表格式样的定义。
Text-cell Style
0..∞
电子表格中单元格式样的定义。

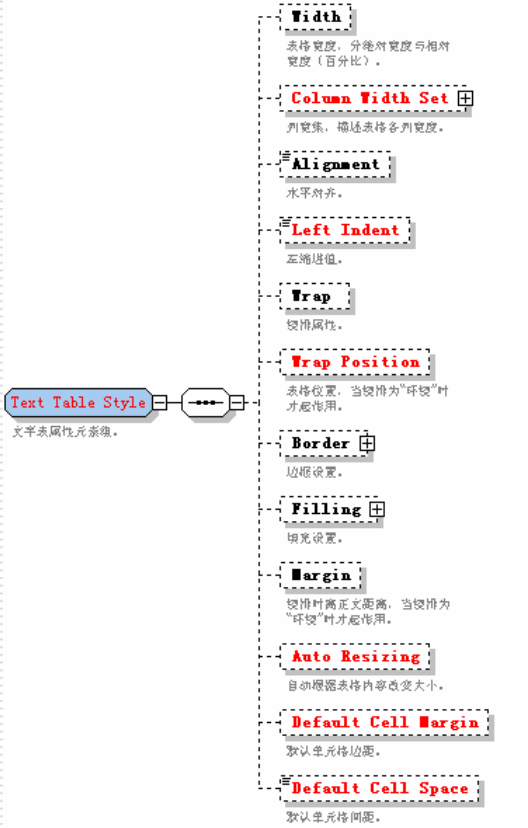

自动编号类型
Auto Numbering Set
自动编号集合类型。
Auto Numbering
0..∞
一级自动编号设置。
Level
0..8
多级编号的级别。
Symbol Font
项目符号的句属性，用以设置声明项目符号的属性，如：字体、下划线等。
Reference to Link Style
链接样式，引用一个段样式。声明编号及后面的文本的样式，如：标题、正文等。
Number Format
本级序列格式。如1,2,3;a,b,c。
Number Display Format
字符串描述的编号格式。如：%1.%2
Reference to Pic Symbol
表示图片符号的图形的引用，图形定义于对象集中。
Numbering Alignment
编号的水平对齐方式。
Indent
缩进。
Tabstop Position
编号与正文之间的制表位的位置。
Start Number
编号的起始值。
Attached Character
尾随字符，编号和文本之间所填充的字符的类型，可以是下面三个中的一种：制表符，空格符，无。
Regular Format
强制使用正规格式，即使用阿拉伯数字编号。


Processing Rules
公用处理规则，包含文档全局设置、批注、修订信息、文档用户等内容。
Word Processing
文字处理文档类型。
主体类型
Section
分节。分节是两个章节的分隔。一个文档至少有一个分节，且文档最前处必须有一个分节。
Body
文字处理文档主体部分。
Section
分节
Chapter
逻辑章节。
0..∞
Paragraph
段落
Text Table
文字表格


节属性类型
Section
Property
Format Revision
Section Type
Page Margin
Paper
Header-footer Odd-even H...
Header-footer First Page...
Header Position
Footer Position
Header
Footer
Gutter
Symmetrical Margin
Page Direction
Portrait or Landscape
Paper Source
Footnote Setting
Endnote Setting
Page Number Setting
Line Number Setting
Grid Setting
Vertical Alignment
Writing Direction
Border
Filling
Column Setting


Paragraph Property
Format Revision
Outline Level
Alignment
Indent
Line Space
Paragraph Space
Auto Numbering Info
Orphan Control
Widow Control
No Page Break
Same Page to Next
Page Break Before
Border
Filling
Tabstop Position
Grid Alignment
First Character Drop
No Word Break
Allow Word Break
No Line Number
Punctuation Control
Punctuation Compress
First Last Chinese Control
Chinese English Auto Ker...
Chinese Number Auto Kern...


Text-run Property
句子属性包元素组，包含句子的基本属性。
Font
字体属性，指定当前句使用的中、西文字体，字号大小及颜色。
Bold
粗体效果。
Italic
斜体效果。
Highlight
突出显示效果。
Border
文字边框，文字边框的四个边框的设置是相同的。
Filling
填充，句属性只支持颜色或图案的填充。
Strike
删除线效果，分单线与双线两种。
Underline
下划线效果。
Emphasis
着重号效果。
Hidden
隐藏文字效果。
Outline
空心文字效果。
Emboss
浮雕效果，包括阴文或阳文效果。
Shadow
阴影效果。
Transform
醒目字体效果，包括大写、小写、首字母大写及小型大写。
Position
Scale
字符横向缩放，例："150"表示放大至1.5倍宽。
Spacing
字符间距值。
Kerning
自动调整字间距的值。
Snap to Grid
设置了文档网格情况下，字符是否对齐网格。


Text Table Style
文字表属性元素组。
Width
表格宽度，分绝对宽度与相对宽度（百分比）。
Column Width Set
列宽集，描述表格各列宽度。
Alignment
水平对齐。
Left Indent
左缩进值。
Wrap
环绕属性。
Wrap Position
表格位置，当环绕为"环绕"时才起作用。
Border
边框设置。
Filling
填充设置。
Margin
环绕时离正文距离，当环绕为"环绕"时才起作用。
Auto Resizing
自动根据表格内容改变大小。
Default Cell Margin
默认单元格边距。
Default Cell Space
默认单元格间距。

Next to do

- Add in further facilities to support workflow, security control, etc., add in page description function to prevent page fidelity
- The three targets for compatibility we aimed are:
  - Content compatibility
  - Format compatibility
  - Layout compatibility

Chinese Specific Issues in UOF

Sequence Numbers

UOF name	Notation
decimal-full-width	全角数字：１，２，３，４，...
decimal-half-idth	半角数字：1，2，3，4，...
decimal-enclosed-circle	①，②，③，...
decimal-enclosed-fullstop	全角带句点：1.，2.，3.，...
decimal-enclosed-paren	(1)，(2)，(3)，...
decimal-enclosed-circle-chinese	中文：①，②，③，...
ideograph-enclosed-circle	㈠，㈡，㈢，...
ideograph-traditional	甲，乙，丙，...
ideograph-zodiac	子，丑，寅，...
chinese-counting	中文小写：一，二，三，...
chinese-legal-simplified	中文大写：壹，贰，叁，...

Date Formats

Date Format	Value	Display
[DBNum1]yyyy"年"m"月"d"日"	2002-8-9	二00二年八月九日
[DBNum1]yyyy"年"m"月"	2002-8-9	二00二年八月
[DBNum1]m"月"d"日"	2002-8-9	八月九日
yyyy"年"m"月"d"日"	2002-8-9	2002-8-9
yyyy"年"m"月"	2002-8-9	2002年8月
m"月"d"日"	2002-8-9	8月9日
'aaaa	2002-8-9	星期五
'aaa	2002-8-9	五

Time Formats

Time Format	Value	Display
h:mm	1.234	5:36
h:mm AM/PM	1.234	5:36 AM
h:mm:ss	1.234	5:36:58
h:mm:ss AM/PM	1.234	5:36:58 AM
h"时"mm"分"	1.234	5时36分
h"时"mm"分"ss"秒"	1.234	5时36分58秒
上午/下午h"时"mm"分"	1.234	上午5时36分
上午/下午h"时"mm"分"ss"秒"	1.234	上午5时36分57秒
[DBNum1]h"时"mm"分"	1.234	五时三十六分
[DBNum1]上午/下午h"时"mm"分"	1.234	上午五时三十六分

Digital Formats

Format	value	Display	Notation
000000	12345	0012345	邮编
[DBNum1][$-804]G/通用格式	12345	一万二千三百四十五	中文小写数字
[DBNum2][$-804]G/通用格式	12345	壹万贰仟叁佰肆拾伍	中文大写数字

- Graphics and shapes
- Math formulas
- ……

Contents

- Standardization Approach
  - Standard publications
  - XSL/XSLT practice
- Chinese Office Software Project – UOF
  - Background
  - Major Structures Introduction
- **Suggestions and Comments**

Suggestions and Comments 1

- Why Style is Complicated?
  - historical cultures
  - Multi-languages
  - …
- What is Content, what is Style?

—— All depend on **applications**, thus we should refine our application context before developing the function library

Suggestions and Comments 2

- Work Type C or Work Type B?
  - a) final form (non-revisable), e.g., PDF, PostScript, or SPDL
  - b) editing tool's form (revisable), e.g., Word form, RTF, or etc.
  - c) structuring language form with a style specification by style language (revisable), e.g., [SGML + DSSSL] or [XML + XSL]
  - Here we focus on the case c)…

  ——Yushi Komachi, The First Asian Forum for Information Technology

Suggestions and Comments 3

- Can Type C be WYSIWYG?
- Can Type C be Editable?
  - Editing semantics
- Content/Style Separated or Mixed?
  - Both required

——While we focus our eyes on “Work Type C”, we’d better take enough attention to “Work Type B” documents, as most documents we are using are edited by word processors like MS Word or Sun StarOffice

Suggestions and Comments 4

- DSSSL or XSL?
  - Disadvantages of DSSSL:
    - not widely implemented, especially at high end;
    - does not guarantee page fidelity
    - one-pass formatting and can’t always map from screen back to document
    - scheme syntax scares people away

XSLT Library

- EXSLT (http://www.exslt.org/)
- XSLT Stand Library (http://xsltsl.sourceforge.net/#id215188)

- The EXSLT project is creating a library to standardise extension functions. The XSLT Standard Library is complementary to the EXSLT project

EXSLT Modules

- **Common** covers common, basic extension elements and functions.
- **Math** covers extension elements and functions that provide facilities to do with maths.
- **Sets** covers those extension elements and functions that provide facilities to do with set manipulation.
- **Functions** are those extension elements and functions that allow users to define their own functions for use in expressions and patterns in XSLT.
- **Dates and Times** covers date and time-related extension elements and functions.
- **Strings** covers extension elements and functions that provide facilities to do with string manipulation.
- **Regular Expressions** covers extension elements and functions that provide facilities to do with regular expressions.
- **Dynamic** covers extension elements and functions that deal with the dynamic evaluation of strings containing XPath expressions.
- **Random** covers extension elements and functions that provide facilities to do with randomness.

Suggestions and Comments 5

- Moving the library work onto XSL side
- To enrich the XSL flow object set and build a template library could be the best place to start with
- Alternatively, why don’t we setup a new style language for our purpose? As some researchers have already pointed out that markup might not be the best way to implement styles

- Thanks for patient
- Welcome questions and comments