〖ST™〗文本結構化標記語言

(a.k.a.,) Jing's Structured Text Markup Language

黄　京　（*RadioNoiseE*）

西历 2023 年 6 月 20 日

概要

本文档将介绍 JSTML，一种基于 C 语言构建的（极简易的）文本结构化标记语言；而它的设计目的是，用来写同学录。

其本质上是一个基于下推自动机[*1]的解析器，因设计用途的局限性，不允许出现嵌套等魔法。容错模型也较为简陋，而性能则没有进行任何优化（读入输出纯靠栈）。

将先介绍数据结构、语法，而后介绍实现细节、自动化 Lua 脚本、plain-TeX 输出样式文件等信息。

1　一些约定

- 下文中将会用「她」「其」等代词表示 JSTML 语言的独立解释器，即她的可执行文件。
- 将会用小型大写西文字母（Small Caps）表示脚本语言、宏语言、标记语言等，还用来表示某些特定的操作系统名称。
- 使用方全角引号（「」）表示被它们划定的特定字符；使用尖括号与斜体表示可选项描述或不需要显示输入的内容，不代表实际键入的字符。
- 将会使用脚注补充一些多馀信息，且响应国家倡议：正文中使用符合现行语言标准的简化字和两个全角空格宽的缩进。

2　字类型及数据类型、结构

2.1　字类型

在她眼中，所有的输入都属于字（token），及一个或一些字符（character）的集合。而字又被分为五种类别[*2]：

分界符　如其名，自然是作为两种数据结构的界定出现。其中，「<」被用作表示开始、而「>」表示结束。

标示符　标示其中一种数据结构的开始与结束。其中，「*」为开，而「/」为关。

汉字　主要的处理对象，也就是这门语言「标记」的东西。由几乎所以不属于其它三类的字符组成。

[*1] 一个很哲学的术语。

[*2] 致敬高德纳教授所创 TeX 的类别码（category code），由于大部分（如果不是全部的话）标记语言都有类别码的概念、而 TeX 实际上是宏语言，故特此说明。

空白　包括空格[*3]及横向制表符（tab）。

换行　指在 Macintosh 或 Unix、Posix 等系统下的回车换行符。

2.2　数据类型

上述的四种字组合便有了能够被她处理的，合法的唯二的数据类型：狗牌（tag）和八卦（text）。

其中狗牌表示对八卦的一个概述，故理论上应短小而精悍。也因此，她内部分配给狗牌的空间仅有 19 个字符长度[*4]。也就是说，如果你往狗牌里硬塞超过 19 个汉字，会导致分段错误或栈溢出[*5]，报错并继续运行（毕竟它只是狗牌呐）。

而另一种八卦所能容纳的字符就多多了，达 8192 个字符的长度。八卦与狗牌一一对应，是对狗牌的展开说明，也可以是扩充等一切合理（或不合理）的用途。

2.3　数据结构

在两种数据类型的基础上，又构建了两种数据结构（即，用来组织／表示数据类型的东西）：片段和累牍。一个片段只能出现在一行之中[*6]，而累牍则理论上横跨数行。

片段的狗牌被使用一对分界符界定，而八卦则被结束分界符和换行符界定，语法如下：

```
⟨Optional Space⟩<⟨Optional Space⟩狗牌⟨Optional Space⟩>⟨Optional Space⟩八卦⟨Carriage Return⟩
```

而 ⟨*Optional Space*⟩ 表示可选的被忽略的空白字类型、⟨*Carriage Return*⟩ 表示换行符（回车）。

累牍的狗牌需被括在分界符中，作为八卦的界定出现两次；同时需要使用累牍标示开、关来标识。语法如下：

```
⟨Optional Space⟩<⟨Optional Space⟩*⟨Optional Space⟩狗牌⟨Optional Space⟩>⟨Optional Space / Newline⟩
  ⟨Optional Space⟩八卦⟨Carriage Return⟩
  ⟨Ditto, Iteration / Recurse⟩
⟨Optional Space⟩<⟨Optional Space⟩/⟨Optional Space⟩狗牌⟨Optional Space⟩>
```

其中，⟨*Optional Space / Newline*⟩ 表示可选的空白类字符、⟨*Ditto, Iteration / Recurse*⟩ 代表对上一条语句的不限次数的重复[*7]。

3　参考范例

给出一个实际使用本标记语言的范例（仅供参考、雷同巧合）[*8]：

```
<姓名> 佚名
<性别> 不明确，TeX里的\empty、Lua里的nil、C里的0。
<政治面貌> 革命群众
<*教育背景>
```

[*3] 不包括中文的全角空格，其属于汉字类别。

[*4] 实现使用<uchar.h>的标准化头文件来支持万国码，所以一个字符是 8 字节长度。

[*5] 这些都是可以调整的，见 jstml.c 文件中对 MAXTOKEN、MAXTAG 等的宏定义。

[*6] 实际上，是由于它的末尾被且仅被换行符界定。这由于平台差异会出现事故：Windows 下换行符为<CR><LF>、而 Macintosh 和 Unix 系为<CR>、Posix 等不明确。故不一定支持使用 Windows 系统构建项目。

[*7] 当然是在输入的字符不超过那个栈的能力范围的情况下。

[*8] 其中部分文字来源于黄新刚的『雷太赫排版系统简介』，被 GNU 许可证保护。

巴灵顿大学：烈士（工商管理）——1927～1936
克莱登大学：勇士（比较文学）——1921～1927
卧龙岗大学：壮士（分子生物）——1919～1921
清华学堂：博士（有机化学）——1911～1919
京师大学堂：硕士（天体物理）——1898～1900
北洋大学：学士（应用数学）——1895～1898
</教育背景>
<*业余爱好>
搬砖砌墙，割草喂猪；挖坑灌水，淫湿作画。
研经修佛，以目窥密；布施洗礼，濯尘安卧。
</业余爱好>
<*信>
最大之乘，最正之宗；自如之理，乃见真实；修无为福，
胜于布施；受持此经，功德无量；应现设化，亦非真实。
</信>
<注> 午休时间谢绝来电。

4 plain-TeX 样式文件使用指北

本文件就是一个宏集兼驱动文件。使用它可以将本标记语言输出的中间文件转为 PDF 文件输出，通过修改它还可以有不同样式的结果。默认的标题名非常蠢，强烈建议修改标题后使用（见 jsdvr.tex 文件中的注释）。

另，请注意它是用 plain-TeX 而非 LaTeX 2_ε 写的，所以别乱改，会出问题。也请使用 LuaTeX 编译。

5 Lua 自动化脚本

该自动化脚本供处理大批量的使用本标记语言写成的文本文件使用。只需要创建一个名为 jsindex.ind 的文本文件，并将需要处理的文件名（带扩展名）使用西文逗号／换行符分隔后写在该文件里。

随后就可以使用 Lua 的独立解释器执行名为 jstml-auto.lua 的文件了。如果一切正常，幸运与你常在的话等一会就会凭空多出一个 PDF 格式的文件，这就是最终的输出了。如果报错，能自己解决最好，不行就开 issue。