〖ST™〗文本結構化標記語言

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

黄　京

西历 2023 年 6 月 18 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；使用方括号（[]）表示可选项描述，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。而字又被分为四种类别[*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）。

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，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，等等一切合理的用途。

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[*5]，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：

[opt space]<[opt space]狗牌[opt space]>[optspace]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）。

[*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] 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。故不支持使用 Windows 系统构建项目。