16.3 多语言字符编码入门

前一节描述了客户端和服务器是如何分别在 HTTP 的 Accept-Charset 首部和 Content-Type 首部的 charset 参数中携带字符编码信息的。对于工作中要开发大量国际化应用的 HTTP 程序员来说，为了能理解技术规范和相应的实现软件，需要深入地理解多语言字符系统。

由于术语很复杂且不一致，学习多语言字符系统不太容易。常常需要阅读标准文档，而且我们可能对工作涉及的那些语言还不太熟悉。本节是对字符系统及其标准的概览。如果读者对字符编码很熟悉，或者对这部分细节不感兴趣，可以直接跳到 16.4 节。

16.3.1 字符集术语

以下是应当了解的电子化字符系统的 8 个术语。

- *字符*
  字符是指字母、数字、标点、表意文字（比如汉语）、符号，或其他文本形式的书写“原子”。由统一字符集（Universal Character Set，UCS，它的非正式的名字是 Unicode[3]）首创，为多种语言中的很多字符开发了一系列标准化的文本名称，它们常用来便捷地命名字符，而且不会与其他字符冲突。[4]

- *字形*
  描述字符的笔画图案或唯一的图形化形状。如果一个字符有多种不同的写法，就有多个字形（参见图 16-3）。

- *编码后的字符*
  分配给字符的唯一数字编号，这样我们就可以操作它了。

- *代码空间*
  计划用于字符代码值的整数范围。

- *代码宽度*
  每个（固定大小的）字符代码所用的位数。

- *字符库*
  特定的工作字符集（全体字符的一个子集）。

376

注 3：Unicode 是一个以 UCS 为基础而成立的商业化联合组织，致力推广商业产品。
注 4：字符的名称看起来类似 LATIN CAPITAL LETTER S 和 ARABIC LETTER QAF 的形式。