16.5 国际化的URI

直到今天，URI 还没有为国际化提供足够的支持。除了少数（定义得很糟的）例外，URI 如今还是由 US-ASCII 字符的一个子集组成的。人们正在努力使主机名和 URL 的路径中能包含更丰富的集合中的字符，但直到现在，这些标准还没有被广泛接受和部署。现在让我们来回顾一下当前的一些尝试。

16.5.1 全球性的可转抄能力与有意义的字符的较量

URI 的设计者们希望世界上每个人都能通过电子邮件、电话、公告板，甚至无线电来共享 URI。他们还希望 URI 容易使用和记忆，但这两个目标是相互冲突的。

为了让世界各地的人们都能够便捷地输入、操控，以及共享 URI，设计者们为 URI 选择了常用字符的一个很有限的子集（基本的拉丁字母表中的字母、数字以及少数特殊符号）。世界上绝大多数软件和键盘都支持这个小的字符集合。

但不幸的是，限制了字符集的话，URI 就无法被全球的人们方便地使用和记忆。世
389 界上有很大一部分人甚至都不认识拉丁字母，他们几乎无法把 URI 当作抽象模式来记忆。

URI 的设计者们觉得确保资源标识符的*可转抄能力*（transcribability）和共享能力比让它们由最有意义的字符组成更加重要，因此（如今的）URI 基本上是由 ASCII 字符的受限子集构成的。

16.5.2 URI字符集合

URI 中允许出现的 US-ASCII 字符的子集，可以被分成*保留*、*未保留*以及*转义*字符这几类。未保留的字符可用于 URI 允许其出现的任何部分。保留的字符在很多 URI 中都有特殊的含义，因此一般来说不能使用它们。表 16-7 中列出了全部未保留、保留，以及转义字符。

表16-7 URI字符语法

字符类别	字符列表
未保留	[A-Za-z0-9] | "-" | "_" | "." | "!" | "~" | "*" | "'" | "(" | ")"
保留	";" | "/" | "?" | ":" | "@" | "&" | "=" | "" | "$" | ","
转义	"%" <HEX> <HEX>