普遍适用性
快速指南

“普遍适用性”是指什么?

有些软件无法识别或正确处理所有的域名和电子邮件地址。这是因为域名中含有的顶级字符串可能比过去常见的字符串要长，而且域名和电子邮件地址现在可以使用从基于统一码 (Unicode) 的字符集中提取的字符，而该字符集的涵盖范围要远远大于传统的美国信息交换标准码 (ASCII)[1]。鉴于此，**“普遍适用性”** (UA) 是指以统一方式正确**接受**、**验证**、**存储**、**处理**并**显示**所有有效域名和电子邮件地址的状态。

普遍适用性指导小组 (UASG) 是由社群主导的组织，其职责是提高意识、明确并解决与所有域名和电子邮件地址普遍适用性相关的问题。UASG 的宗旨是帮助确保全球范围内的互联网用户享有统一的良好体验。UASG 受 ICANN（互联网名称与数字地址分配机构）支持，其成员来自全球 200 多个组织，包括： Afilias、Apple、CNNIC、GoDaddy、Google、Microsoft 和 Verisign。如需了解更多有关 UASG 及其最新进展的信息，请访问： **www.uasg.tech**。

当软件和在线服务针对全部域名和电子邮件提供上述所有功能时，即表明它们支持普遍适用性。

本快速指南介绍了 UASG 为互联网系统处理域名和电子邮件地址的五个领域——“接受”、“验证”、“存储”、“处理”及“显示”——实现普遍适用性而拟定的建议。本指南适用于负责信息技术或软件产品工程事宜的高级管理人员及经理。本指南仅概要介绍了相应的 UASG 建议，而未提供软件架构师或软件工程师所需了解的一些重要细节。如需了解相关的细节信息，请参阅 UASG 007 “普遍适用性简介”。

[1] ASCII 是传统的互联网字符编码系统，相关定义，请参阅互联网标准 RFC 20 (https://tools.ietf.org/html/rfc20)。Unicode 是由统一码联盟 (Unicode Consortium) (http://unicode.org) 定义的。

接受

“接受”是指从用户界面、文件或 API（应用程序编程接口）接收域名或电子邮件地址以将其用于软件应用程序或在线服务的过程。

UASG 建议

- 输入栏应足够大，能够接受任何有效输入。根据其编码方式，域名可能需要多达 670 个字节的空间。除了域名，电子邮件地址中还包含本地部分（即 @ 符号前面的部分），这一部分可能长达 64 个字节，因此电子邮件地址的总长度可达 735 个字节。

- 应用程序和服务应该可以接受由“统一码转换格式 8 位字节”(UTF-8)[2] 编码的域名和电子邮件地址，且应该可以识别出 UTF-8 编码所占的字节数可能大于显示的字符数。

- 国际化域名 (IDN) 既能够以原始文字的形式输入和显示，也能够以 ASCII 形式输入和显示以实现向后兼容，例如“测试”和“xn--0zwm56d”这两种形式。对原始文字进行的 Unicode 编码称作 U-标签；而相应的 ASCII 兼容编码称作 A-标签[3]。软件应该既可以接受 A-标签，也可以接受 U-标签，但是若要进行显示和进行任何不需要 A-标签的处理操作，则应将 A-标签转换为 U-标签。

- 在绝大多数情况下，输入的域名或电子邮件地址应先转换为 Unicode 规范化形式 C (NFC)[4]，然后再执行进一步处理。但是，由于 NFC 也不是完全不会造成损失，因此在少数情况下，可能需要延迟规范化处理，直到执行进一步处理后确定域名或电子邮件地址所适用的确切环境为止。

[2] UTF-8 会将每个 Unicode 码点编码为由 1 到 4 个字节构成的序列。相关定义，请参阅 RFC 3629。

[3] U-标签与 A-标签之间的转换需采用 RFC 3492 和 RFC 5891 中定义的“国际化域名编码 (Punycode)”算法。RFC 3492 и RFC 5891.

[4] 请参阅《统一码标准》附录 15 中的“Unicode 规范化形式”(https://www.unicode.org/reports/tr15/tr15-47.html)。

验证

“验证”是指检查电子邮件地址或域名的语法是否正确无误，并在适当时检查域名是否按预期真实地存在于 DNS 中的过程。为了能够验证当今的域名和电子邮件地址，验证技术可能需要更新。

UASG 建议

- 对输入内容进行验证的方式应该符合其预期用途。所有域名均应按照《国际化域名应用》(IDNA) 标准（当前为 IDNA2008）进行验证[5]。 这样可以确保域名采用的语法是有效的。
- 如果输入的字符串按预期应属于 DNS 中的现有条目，则需通过 DNS 查询对其进行验证。
- 如果输入的字符串按预期应属于当前可能尚未纳入 DNS 的有效域名，则仍然可以对其部分内容进行验证。例如，可以按照由互联网号码分配机构 (IANA) 维护的有效顶级域 (TLD) 名官方列表对 TLD 进行验证[6]。
- 为验证电子邮件地址，需按上述说明对其域名部分进行验证。由于电子邮件地址中的本地部分完全是由接收邮件的邮件系统定义的，因此通常无法对这一部分进行验证。请求用户输入电子邮件地址两次可有助于检测拼写错误。
- 在大多数情况下，域名或电子邮件地址的所有组成部分（不属于 IDN 的 TLD 名称除外）均应采用一种文字（例如：阿拉伯文或汉字）或几种密切相关的文字（例如：日语汉字、日语片假名、日语平假名和罗马字）。应按照《统一码技术标准 39》中的“Unicode 安全机制”（https://unicode.org/reports/tr39/#Restriction_Level_Detection）来检查采用 Unicode 序列的文字是否符合最佳实践。

[5]有关 IDNA2008 的定义，请参阅 RFC 5890、5891、5892、5893 及 5894。

[6]请参阅“顶级域列表”（https://www.icann.org/resources/pages/tlds-2012-02-25-en）。

存储

“存储”是指暂时或长期存储域名和电子邮件地址的过程，不论预期的存储期限为多长，都应按明确定义的格式存储域名和电子邮件地址。

UASG 建议

- 在绝大多数情况下，在存储之前，都应先按照 Unicode 规范化形式 C (NFC) 对域名和电子邮件地址进行规范化处理。但是，由于 NFC 也不是完全不会造成损失，因此在少数情况下，可能需要延迟规范化处理，直到执行进一步处理后确定域名或电子邮件地址所适用的确切环境为止。
- 在大多数应用程序中，域名和电子邮件地址应存储在经过 UTF-8 编码（最常用且最受支持的 Unicode 编码方式）的文件和数据库中。有些情况下，如果软件必须与旧版数据库配合使用，则采用与数据库相同的编码方式可能会更容易。
- 在应用程序代码中，Unicode 的最佳表示形式取决于编程环境。许多常见的编程语言（包括 python 和 perl 脚本语言）都内置了对 Unicode 的支持，并会在输入和输出时自动转换为 UTF-8 或从 UTF-8 进行转换。
- 应用程序应该为 IDN 选择一种统一的内部表示形式，即统一采用 U-标签或 A-标签。由于每个 U-标签均可以转换为唯一的 A-标签，反之亦然，因此可采用其中任一形式。

处理

“处理”是指电子邮件地址或域名被应用程序或服务用于执行活动（例如搜索或对列表排序）或者被转换为另一种格式（例如由传统编码转换为 UTF-8）的过程。在处理过程中，可能需要额外进行验证。

UASG 建议

- 随着 Unicode 的不断发展，应在可行的情况下升级软件，以便使用最新版本的标准以及任何可用的图形和字体。由于用户设备、软件库和 Web 标准可能不支持最新版本，因此可能会错误地显示新分配的字符，例如将其显示为通用框 (□)，或者根本不显示。
- 如果推出了支持 UTF-8 输入或输出的 API，则需使用此类 API，而避免使用不支持 UTF-8 输入或输出的 API。应使用经过精心调试的标准库（如 GNU libidn2（https://www.gnu.org/software/libidn/#libidn2））来处理和验证 IDN；请勿使用自己创建的库进行处理和验证。
- 在域名和电子邮件地址中使用按从右到左顺序书写的文字时，有一些需要特别注意的事项。 有关其中部分事项的说明，请参阅 IDNA[7]（对于域名）和《统一码标准》附录[8]（对于电子邮件地址）。
- 在创建包含文字或语言信息的注册表或其他数据结构时，可包含尽可能多的文字或语言，最好是包含《统一码标准》支持的所有文字或语言[9]。请注意，有些语言可以使用不同的文字来书写，而有些文字则可用于书写多种不同的语言。

[7]请参阅 RFC 5893“国际化域名应用 (IDNA) 的从右到左书写文字”（https://tools.ietf.org/html/rfc5893）。

[8]请参阅 UAX#9“Unicode 双向算法”（http://unicode.org/reports/tr9）。

[9]请参阅 Unicode“支持的文字”（http://unicode.org/standard/supported.html）。

显示

“显示”是指用户界面以可视方式呈现电子邮件地址或域名的过程。如果底层操作系统支持所使用的文字和任何所需的呈现机制，且字符串是采用由《统一码标准》定义的编码方式存储的，那么域名和电子邮件的显示过程一般会非常简单。否则，可能需要进行特定于应用程序的转换操作。

UASG 建议

- 虽然当今的软件和设备能够显示几乎所有的 Unicode 码点，但旧版系统的支持范围可能有限，故而需要应用程序管理一些旧字体。此外，如果 Unicode 增加了新码点，设备和应用程序将无法显示这些新增码点，除非更新其字体库。
- 应以本机字符形式显示 IDN，除非明确要求将其显示为 A-标签。
- 域名和电子邮件地址既可显示为从左到右 (LTR) 书写的文本，如英文或俄文，也可显示为从右到左 (RTL) 书写的文本，例如：阿拉伯文或希伯来文。由于 Unicode 为单个码点（而非码点序列）指定了方向属性，因此有些混合使用 LTR 和 RTL 的文本（即“双向”文本）会对用户具有一定意义，而有些则没有意义。应使用 Unicode 限制级别标准 来标记可能具有误导性的字符串。
- 互联网用户使用多种不同的语言进行读和说。因此，在一些情况下，可能有必要专门为不同的语言或语言组分别设计不同的应用程序。

[10]请参阅《统一码技术标准 39》中的“Unicode 安全机制”（https://www.unicode.org/reports/tr39/#Restriction_Level_Detection），了解 Unicode 中等限制级别和高限制级别，进而检查采用 Unicode 序列的文字是否符合最佳实践。

实现普遍适用性准备就绪

源代码审核和单元测试

对于已开发或已升级的能够支持普遍适用性的软件和系统，应对其进行检查和测试以查找并修复错误，进而确保其能够正常运行。为了增强普遍适用性方面的意识，UASG 正在尽力与应用程序开发商和在线服务提供商接洽，以鼓励他们执行普遍适用性源代码审核和测试，并共享可用于开发标准化测试案例的一系列标准。

测试

U与此同时，UASG 也在努力制定可用于进行测试的网站、应用程序、电子邮件地址和域名的列表。 在一些情况下，测试可以自动运行，而无需人为干预。亚太互联网络信息中心 (APNIC) 实验室最近代表 ICANN 执行的通用顶级域 (gTLD) 调查便是自动测试的一个真实示例：https://tinyurl.com/new-gtld-ua。UASG 正在研究自动进行普遍适用性测试的方法，在适当的时候，UASG 会共享其研究成果。

更多阅读资料

以下文档提供了更多有关普遍适用性、Unicode 和国际化域名的信息。

- UASG 007“普遍适用性简介”（https://uasg.tech/documents）。
- RFC 5894“国际化域名应用 (IDNA)：背景、说明和理由”（https://www.rfc-editor.org/info/rfc5894）。
- “网上国际排印”，该概要图表展示了在网上处理不同语言所存在的种种问题（https://w3c.github.io/typography/gap-analysis/language-matrix.html）。

术语注释

实现普遍适用性的任务十分艰巨，这是由于人们已经习惯使用由少量独特的“字母”字符构成的简单文字，例如：拉丁文，因此如果将人们所熟悉的许多术语和概念应用于采用了不同规则的书写系统，这可能会造成很大的困惑。为将各种不同的书写系统整合到国际化域名 (IDN) 领域中，需要创造新术语和以新的特殊且特定的方式来使用之前的常见术语（例如“字符”）。本快速指南尽量避免使用此类术语，且在需要使用此类术语之处也会尽量提供术语定义，但是如需研究其他材料，包括本指南中引用的一些文档，可能需要更深入地了解相关术语。