

---

ICANN72 | 筹备周 — 根区标签生成规则更新

太平洋夏令时 (PDT) 2021 年 10 月 14 日星期四 — 11:00 至 12:00

皮提南 · 古阿莫帕塔纳

(PITINAN

KOOARMORNPATANA):

谢谢。大家早上好，下午好，晚上好。欢迎参加 ICANN 72 筹备周的根区 LGR 更新会议。今天，我们的议程是这样的。我首先占用五分钟时间，向大家简要介绍一下根区 LGR 项目的背景情况，然后将时间交给我们今天的主要发言人。今天的主要发言人有：拉丁文生成专家组的成员马茨 · 杜芙柏格 (Mats Dufberg)、来自日本的日文生成专家组主席堀田弘 (Hiro Hotta)，以及来自整合专家组的米歇尔 · 叙伊尼亚尔 (Michel Suignard)，他将向我们介绍根区 LGR 的下一个版本。

现在我们进入第一个议程：根区 LGR 简介。我们简单回顾一下这个项目的发展过程。大约是在 2010 年前后，ICANN 社群发现了变体字顶级域的需求。为此，社群成员一起编制了一份整合问题报告。当年那份报告的内容主要是分析如果在根区中实施变体，可能会引发什么问题。也就是说，那份报告的目的是研究问题，而不是探讨解决方案。

那份报告研究了六种文字，包括阿拉伯文、中文、西里尔文、梵文、希腊文和拉丁文，涵盖现代语言中使用的多种文字。

后来，那份报告在 2012 年编制完成并发布，报告中还指出，我们应该建立某种机制来定义各种文字的变体。这种识别机制应该成为定义变体的单一来源，因为最终所有这些 TLD 都将进入根区。

---

*注意：以下内容为针对音频文件的誊写文本。尽管文本誊写稿基本准确，但也可因音频不清晰和语法纠正而导致文本不完整或不准确。该文本仅为原始音频文件的补充文件，不应视作权威记录。*

社群发现可以使用根区 LGR 机制来定义变体，并且于 2013 年就利用 LGR [流程]开发根区一事获得了董事会的批准。此后，社群针对各种文字成立了生成专家组，开始制定根区 LGR。

与此同时，ICANN 组织还针对如何管理 IDN 变体字顶级域制定了相关建议。该建议于 2019 年获得董事会批准。决议要求 GNSO 和 ccNSO 在政策制定工作中采纳此建议，并且 [听不清] 整合根区 LGR 的运用。

后来，ICANN 董事会于 2020 年发布决议，要求 GNSO 和 ccNSO 考虑根区 LGR 技术应用建议；2021 年，ICANN 董事会再次发布决议，要求 GNSO 也发布 SubPro 报告，其中还包含有关根区 LGR 应用和下一轮工作的事项。

那么，LGR 是如何运作的呢？LGR 就像一个用于生成标签的公式，是一组标签生成规则将这些规则整合到根区后，机器和工具都可以读取它们，从而我们可以根据某种机制识别标签是否有效；对于有效的标签，我们还可以识别其变体，以便允许分配或加以阻止。

LGR 程序机制的工作分为两个步骤。第一步，生成专家组针对特定语言文字制定解决方案；第二步，各语言文字的生成专家组将敲定的解决方案提交给整合专家组。整合专家组 (IP) 将考虑生成专家组 (GP) 的提案，在此期间，GP 和 IP 之间可能会进行多轮次的协商。一旦解决方案成功通过审核并得到批准，便会被整合到根区中，以供用于验证 TLD。

GP 的工作基本就是这样。对于屏幕上的这 28 种语言文字，我们现在已经为其中的 26 种建立了 17 个生成专家组。如大家所见，这项工作是从 2014 年开始的，已经持续了多年。时至今日，许多 GP 已

经完成了他们的工作；日文和拉丁文目前正在进行公众意见征询。缅甸文 GP 仍在最终撰写提案，接下来也将进行公众意见征询。

情况就是这样。这是根区 LGR 的时间表。我们从 2013 年开始组建专家组，各专家组则是从 2014 年开始投入工作。随着时间的推移，部分语言文字的工作逐步完成，并被整合到了根区中。

现在，我们的根区 LGR 已更新到第 4 版，包含共计 18 种文字。预计到 2022 年，我们将再整合 8 种文字，届时将包含共计 26 种文字。以上就是我们迄今为止的工作进度。稍后，来自拉丁文 GP 和日文 GP 的成员将向大家介绍他们的解决方案，同时我在此邀请大家积极参与与他们提案的公众意见征询。

接下来，有请来自拉丁文 GP 的马茨·杜芙柏格。现在把时间交给你了。

马茨·杜芙柏格：

非常感谢。请播放下一张幻灯片。这就是我的那部分演示文稿。我将介绍一下我们已经开始征询公众意见的提案。这些就是我将要讨论的主要内容。请播放下一张幻灯片。

我们已经开始就拉丁文提案征询公众意见，截止时间为今年 11 月 23 日。这里提供有提案链接，我邀请大家都点进去看看我们的提案，我们一定会认真考虑所有大大小小的意见。欢迎大家提出任何意见。请播放下一张幻灯片。

这个演示文稿是我们提案的简介，我们希望通过这种方式来降低门槛，方便大家阅读并向我们提出意见。我将重点介绍主要文档及其依赖条件，LGR XML 文件是一个规范文档，我就不多说了。

其中有一些简短章节，我今天在这里就不作讨论了，那些章节包含的只是一般信息，不值得在这里讨论。提案中的第二章定义了提案针对拉丁文拟定的限制。第一个限制是，提案不得包含任何未纳入“最大启动字汇”中的字符。MSR 是 IDNA 协议有效码点的子集，后者则是统一码的子集。因此，MSR 也是统一码的一个子集，我们的提案不能超出这个范围。

整合专家组已经定义了 MSR。在此过程中，我们提出了一些字符，并建议 MSR 和整合专家组 (IP) 将这些字符添加到 MSR 的新修订版中。MSR 是统一码的一个活动子集，这意味着我们可以根据需要进行增减。

在此过程中，我们提出了一些想要添加的字符。第 4 章描述了我们的工作流程。我们得确定我们要在提案中使用哪些字符。我们决定使用世代失调分级表 (EGIDS)，下一张幻灯片对此进行了说明。从代表国际性语言的 0 级，到代表教育语言的 4 级。我们还针对至少拥有 100 万使用者的语言增加了第 5 级，不过这项工作仍在进行中。就是这些语言。请播放下一张幻灯片。

这里我们可以看到 EGIDS 量表，也就是我们使用的世代失调分级表。这个量表对世界上几乎所有语言进行了分级。我们根据它来选择想要研究的语言。我们选择了 0-4 级语言，稍后还会加上 5 级语言，那些都是至少拥有 100 万使用者的语言。表中还有其他一些级别，但我们将其排除掉了。

附录 6，这个列表中包含我们选择的所有语言。我不知道具体有多少种语言，但确实挺多。其中包含的每种语言都有对应的语言名

---

称。我们选择了 212 种语言。该列表中包含语言名称、ISO 语言代码，以及该语言的 EGIDS 级别。

对于列表中的每种语言，我们确定了其所有书写字符，因此我们可以看到人们用什么字符来书写这些语言。我们的意见是，人们应该能够使用这 212 种语言对应的字符所构成的单词来创建顶级域。报告中并未记录每种语言的字符集，但第九章中提供了引用。我们还确定了备选变体。这个问题在后面有更多介绍。请播放下一张幻灯片。

拉丁文字汇提案当然是基于统一码码点，因为这是一个硬性限制。在最简单的情况下，码点就是一个字符，例如字母 A。码点也可以是修饰符标记。它可以与另一个码点结合使用，形成一个字符。请看幻灯片上的示例：顶部带有波浪号的 g，它就是由两个码点组合而成。

我们有许多由基本字符和重音符组合而成的码点，例如幻灯片上显示的带尖音符的 a。它只是一个码点。第五章的引言部分详细说明了包含或排除某种语言中的某个字符的原则。请播放下一张幻灯片。

拉丁文提案中列出了 218 个字符。其中 197 个字符只是一个码点，它们很简单。另外 21 个字符则由两个或多个码点的序列组成，比如一个基本字符加一到两个标记。所以，每个字符都有对应的构成码点，它可能是单一码点，也可能是一个码点序列。

使用该字符书写的语言，不一定要是所有语言，只要语言数量足够证明所选语言在使用该字符。列表中也引用了那些语言的字母表。对于字符 a 到 Z，我们没有专门列出任何佐证语言，因为所有语言都默认包含这些字符。请播放下一张幻灯片。

---

字汇也是 XML 文件的主要部分之一。第五章的字汇是按码点数字排序的，但如果你想查看按字形分组的相同字符，比如所有带有不同重音符的 A，可以在附录 C 中找到。这两个列表的内容相同，只是排序方式不一样。请播放下一张幻灯片。

第五章的第四部分列出了排除的字符。这些字符会在至少一种选定语言中得到使用，但由于不属于 MSR，因此无法包括进来。MSR 是一种预处理机制，限制了可以包含的字符范围。所以拉丁文 GP 不能包含任何不在 MSR 中的字符。

因此，如果你现了某个你认为应该包括进来的字符，但它不在 MSR 中，那么拉丁文 GP 对于这种情况也无能为力。得先更新 MSR 才行。请播放下一张幻灯片。

我刚才介绍了字汇。接下来是变体字符集，这也是提案中的一个重要部分。第六章介绍的便是这些内容。那么，什么是变体？变体字符集由两个或以上的字符组成，这些字符在某种程度上被视为相同的字符。它们可能具有相同的字形，或者可以在特定语言文字社群的某些或所有领域互换使用。

对于一个变体字符集，我们不能单独使用其中包含的两个字符。在大多数情况下，如果使用其中一个字符，则另一个字符将无法出现在相同上下文中。因此，对于拉丁文，大多数变体字规则都会导致锁定变体。

文字内部变体字符集的字符来自相同的语言文字，例如拉丁文，跨文字变体字符集的字符则来自不同的语言文字，主要包括拉丁文、西里尔文和希腊文，也有一些字符集是组合而成的。请播放下一张幻灯片。

---

第六章以及附录 D1 至 D9 介绍了变体字符集的原理、变体字符集的数据分析，以及备选变体字符集。通过阅读这些章节，大家就会明白，我们是根据什么得出那些字符是变体的结论的。

其中有两个变体字符集较为特殊。它们与之前的 IDNA 版本 2003 有关，并且它们允许分配其他变体。这两个字符集涉及...一个是升音 S 和 SS，另一个是带点 I 和无点 I。这两个字符集在提案中显得较为特殊。请播放下一张幻灯片。

最后，附录 E 中提供了一个视觉上容易混淆的字符列表。那些字符不是变体，我们不建议将它们视为变体。我们没有将它们纳入正式的 LGR XML 文件中。之所以在这里单独提出来，是为了给任何分析语言文字之间的视觉相似性的人提供参考。谢谢。对，流程就是这样。我们现在正在征求公众意见，希望能在明年 1 月提交最终版本。欢迎大家阅读我们的提案并发表看法。我不知道你们是如何处理相关问题的。

皮提南·古阿莫帕塔纳：

聊天室中有没有人提出问题？如果有人想要发言，也可以举手示意。好的，好像没有人。那么我们继续下一项议程。如果大家对于马茨或拉丁文 GP 还有任何疑问，我们稍后也可以回头来讨论。接下来，有请堀田弘先生给大家简要介绍日文根区 LGR 的情况。堀田弘，时间交给你了。

堀田弘:

谢谢，皮提南。大家好。我的网络连接不是很稳定。如果我在发言过程中意外断开连接，还请见谅。不过我应该能很快重新连线，希望如此吧。好的。请播放下一张幻灯片。

这些是日文生成专家组的成员。与其他 GP 中的情况大致相同，我们的各位成员也有着不同的[看法]。请播放下一张幻灯片。

这是日文文字和语言的概述。我认为日文颇为独特，因为日文有三种文字，即汉字、日语平假名和语片假名，即便是 ASCII 和数字字符也不例外。这些字符可按任何顺序组合在一个单词中。随着 LGR 字符在日本工业标准 (JIS) 一级和二级字符中得到定义，我们选择了[它们]作为日文 LGR 的字汇。

由于汉字的字符数量很多，我们的字汇中包含大约 6000 个字符。当然，这些字符都在 MSR 中。我们有三种文字，其中的汉字也用于中文和韩文。这给定义变体带来了一些困难。我稍后再说这个问题。

除了汉字，我们还有日语平假名和日语片假名。日语平假名主要用在日语汉字的后面，用于帮助完整阅读单词，用作副词或连词，以及用于将难以读写的日语汉字改写为易于读写的形式，例如，它们让五岁左右的小朋友也能轻松进行读写。

日语片假名主要用作其他词类。日语片假名主要用于表示外语长词和拟声词。

基本上，我们可以认为所有日文字符都是独立的，但是，正如我刚才提到的，中文、韩文和日文之间会共用一些汉字，在中文和韩文中，某些汉字字符集的两个或多个字符若具有相同的名称和发音，它们将被视为变体。



---

日语社群中的一些人认为，某些中文或韩文变体字符在日文中也应视为变体。请播放下一张幻灯片。

这就是日语根区 LGR 的简要介绍。正如我刚才所说的，日文字汇包含 6000 个字符，而变体是指具有相同名称和发音的多个字符 — 日文本身不存在变体。但是就视觉相似性而言，我们定义了 12 个变体字符集，这一点较为独特，我们考虑了中文和韩文 LGR 中定义的汉字变体。我想我们可能从中文和韩文 LGR 中引入了数百个变体字符。

这是第二个重点内容：不得分配除原始标签之外的变体标签。这是日文中最难决断的问题，但我们最终还是决定不使用变体标签。

还有 WLE，这是一条非常简单的特定于日文的规则。不得以任何小型假名、叠字符或长音符作为标签的开头。该规则同样适用于普通日文单词。请播放下一张幻灯片。

特别注意事项与视觉相似性。我们定义了几组具有视觉相似性的日文字符。首先是连字符和[长音减号]这两个符号。另外，统一码联盟列出了不同文字之间容易混淆的字符。我们根据该列表挑选了以下十对备选字符，它们具有较高的视觉相似性。这些字符共有十对。右下角显示了包含这些字符的单词。请播放下一张幻灯片。

基于统一码的 10 对具有视觉相似性的字符的确认。我们通过实地研究调查了所有这些字符是否具有视觉相似性。我们通过 40 名测试者开展了这项调查。其中 20 人能轻松阅读测试使用的日文，另外 20 人则不能。我们就每对字符的视觉相似性进行了评分，结果显示，所有字符对的评分都低于 3.2。这意味着它们确实容易混淆。所以我们决定将它们归为变体字符。

在第二项实地研究中，我们调查除了上述十对字符之外，是否还有其他具有视觉相似性的字符对。我们向大约 176 人投放了调查问卷，其中 73 人做出了回复，结果显示没有任何一对字符会使超过 3% 的受访者产生混淆，因此我们认为只需要将基于统一码易混淆列表的那十对字符归为变体。请播放下一张幻灯片。

这是第二个特殊主题。正如我前面提到的，日文标签中基本上允许使用任何字符组合。这个特点可能会显著增加变体字符串的数量。另外，还要考虑到许多变体的定义是源自中文和韩文 LGR。例如，Keio Daigaku 是一个四字符的汉字字符串，意思是庆应大学。它有三种变体字符串。

实际上，庆应大学在 .jp 顶级域下注册并使用了所有四个变体字二级域，这意味着庆应大学希望将它们全部用作变体。但如果允许庆应大学使用所有这些变体，那么就要同时允许这四个 TLD 进入根区。虽然四个不算多，但是在考虑某些更长的标签时，这条规则可能使根区规模出现爆炸式增长。因此，我们需要减少可分配变体标签的数量，以免根区规模出现爆炸式增长。请播放下一张幻灯片。

日文 GP 尝试了很多方法来减少可分配标签的数量，但没有任何一种方法能够有效地将可分配标签的数量减少至例如五个以下。所以我们最终决定只允许使用有效的申请标签，而阻止所有变体标签。这样一来，就只存在一个可分配标签，即申请标签，而任何基于该标签的变体标签都将被禁止。请播放下一张幻灯片。

我们目前已启动公众意见征询，截止时间为 11 月 16 日。谢谢。欢迎提出任何意见。

---

皮提南·古阿莫帕塔纳： 谢谢堀田弘先生。聊天室中有人提出问题吗？

萨尔麦德·侯赛因

(SARMAD HUSSAIN):

我们有一个问题，但这个问题并非特定于日文。这是一个较为宽泛的问题，我们还是等一下再说吧，也许等米歇尔发言之后，我们再向整合专家组提出这个问题。

皮提南·古阿莫帕塔纳：

好的。谢谢萨尔麦德。谢谢堀田弘先生。我们继续下一项议程，有请整合专家组的米歇尔·叙伊尼亚尔介绍根区 LGR 5。米歇尔，时间交给你了。

米歇尔·叙伊尼亚尔：

好的。请播放下一张幻灯片。正如大家所见，我的发言将包括四个主题。第一个主题是关于整合：我们目前正在进行的中文、日文和韩文 LGR 整合工作进展如何？

[听不清]关于如何处理标签的幻灯片[听不清]，亚美尼亚文、西里尔文、希腊文和拉丁文的一些有趣的地方，以及根区 LGR 5 的范围。接下来，我将详细介绍每个主题。请播放下一张幻灯片。

这个问题实际上和堀田弘先生刚才说内容的有点关系。这是我们现在面临的一项主要任务，也就是通过中文、日文和韩文各自的 LGR 将这三种语言文字整合起来。我们现在已经完成了中文、日文和韩文 LGR 的相关工作。我们已经为 LGR 定义了规则集。事实上，只有日文 LGR 仍在进行公众意见征询。其他两种语言文字已经在前段时

---

间完成了公众意见征询。那是一项非常复杂的任务。那些 LGR 都很大，包含成千上万个码点。有关变体的问题非常复杂，我们可以现在或者在稍后的提问环节讨论这个问题。

自从获得这三种语言文字的 LGR 以来，我们就一直在努力整合它们。显然，[我们没有]完成这项工作，但我们一直在与各个 GP 合作，研究如何进行整合。

首先，我们发现了一个有趣的现象，这些 LGR 中的变体有些是单独创建到 LGR 中的，有些则是通过整合沿袭得来的。沿袭变体有两组。对于其中一组，在通过整合其他字汇得到的结果 LGR 中，这些变体至少有两个码点，或者它们本身就对很多韩文和日文字符有所影响，因为实际上，中文 LGR 的变体数量目前可以说是稳居第一。也许有些人会知道，这是由于中文字汇中同时包含简体和繁体字符。许多[听不清]字符都可以用简体或繁体形式来表示。这其中是有历史原因的。

一些简体字符也确实存在于日文或韩文中，它们的字形相同，而且在繁体字符中也有许多变体。我们可能经常发现，有许多繁体字符虽然字形相同，使用的却是过去使用的定义。

因此，到目前为止，中文 LGR 的变体数量最多，因为其中许多变体字符集在日文或韩文 LGR 中至少有两个字符。那些 LGR 也必须对那些字符进行定义。事实上，这意味着我们基本上必须要求韩文 LGR 和日文 LGR 纳入那些变体字符集，它们在中文 LGR 变体字符集的既有基础上添加了至少两个码点。

这就是第一部分，主旨基本上就是确保所有变体字符集，即从中文变体字符集中沿袭的至少包含两个字符的变体字符集，也都在日文

和韩文 LGR 中得到定义。日文和韩文 LGR 可能阻止这些字符，但在中文 LGR 中，它们有时是可分配的字符。这是该系统的另外一个要点。

除此之外，中文 LGR 实际上还从最初提交的中文 LGR 沿袭了某些未列出的变体。那些主要是由于日语假名和朝鲜文之间的混淆造成的，因为正如大家之前看到的，日文 LGR 创建的变体字符集中包含一些汉字和假名字符。那些字符显然也需要导入到中文 LGR 中，因为很明显，在很多情况下，中文字符也包括那些字符，可以说是相当于日语汉字的等效字符。韩文也是如此。但韩文又有点不同。它涉及朝鲜文和朝鲜汉字 — 朝鲜汉字基本上就是指韩文中的汉字。朝鲜文和一些中文字符之间存在混淆，这些字符也必须沿袭到中文 LGR 中。

至于韩文和日文，情况则要稍微简单一点。正如我之前所说的，我们只引入了一到两个码点，但很明显，对于那些包含与不存在于韩文和日文字符集中的其他中文字符相关联的某个码点的变体字符集，韩文和日文终归要将它们包括进来。因此，韩文和日文 LGR 中也会显示这些变体，我们称之为其他字汇。

显然，当通过合并的 LGR 文件确定标签冲突时，结果将会不出所料。在每个根区 LGR 中，我们定义了“要素 LGR”，它们实际上就是 GP 提议的 LGR 的整合；我们还编制了另一个文件，我们称之为“合并 LGR”或“通用 LGR”，我们基本上就是通过该文件来检测现有或授权标签下的申请标签冲突。因此，我们需要一种机制来确定人们实际上可以创建 — 新的申请标签仍然可供授权。

[在那三组中，实际上就只有中文 LGR 包含那些可分配的标签。]那些[有]变体。事实上，所有 LGR 都包含变体，但那些 LGR 阻止了变

体。需要明确的是，日文 LGR 中并不是真的没有变体，只是它们都被阻止了。日文 LGR 仍在进行公众意见征询，但我们预计该流程不会出现重大问题。让我们拭目以待。请播放下一张幻灯片。

这张幻灯片，或者说这张图片，也是根区 LGR 概述的一部分，上面实际上包含了许多详细信息，但它的重点是，我们在提交标签进行根区授权时确实会出现两个要素。第一个是原始标签，第二个是拟议的变体标签。那是我们进入该流程的两个路径。第一个是我们通常所说的申请标签，它是我们提出申请时的主要关注点。

申请人 — 显然，我们使用要素 LGR 来验证申请标签是否有效，确保它满足字汇要求、规则要求或者 LGR 中提出的任何特殊要求，然后我们相应地做出决定，或者是拒绝它，或者是认为它无效，在这种情况下，我们的流程将会就此打住。然后我们生成索引变体，这基本上就是我们创建该标签的唯一表达的一种机制，确保它可以与已经授权或进入根区的现有标签进行比较。如果它是唯一的，即尚未使用，那么我们会接受它。如果它不是唯一的，我们会拒绝它。这基本上就意味着我们已经授权了该标签的另一个变体。在这种情况下，我们基本上就会拒绝它。

同时，第二个路径基本上可以作为一个可选步骤，申请人也可以提出变体标签。在提出变体标签时，我们可以自行研究，也可以通过查看 LGR 来确定什么样的标签可能成为申请标签的变体。创建好变体标签后，我们显然首先必须确保它是有效的，这时候，我们将再次使用要素 LGR。然后我们还必须确保它是可分配的。如果它不可分配，那么系统就会阻止并拒绝它；如果它是可分配的，那么就会成为一个可分配变体，可以授权到根区中。

在这里需要注意的是，我们要小心选择将哪一个标签作为原始标签或变体标签，因为变体之间的映射并非对称的。对于变体字符集的每个字符，原始字符和变体字符之间可能有不同的映射类型。有时候，我们必须小心选择允许存在变体标签的原始标签，因为如果选择了一个变体标签，在某些情况下，原始标签可能会变成不可分配的标签。两者之间的映射并非完全对称。如果要想让两个标签都可分配，我们有时候在选择第一个标签时就必须小心。我们可能需要查看 LGR 定义和变体字符集映射，以及这些变体字符集的映射类型。请播放下一张幻灯片。

这基本上就是 — 我们有四种语言文字，它们彼此之间存在关联。它们是亚美尼亚文、西里尔文、希腊文和拉丁文。其中大部分语言文字的 LGR 已经完成。只有拉丁文 LGR 仍在进行公众意见征询。希腊文 LGR 不久前也进行了公众意见征询，但还有些收尾工作没有完成。亚美尼亚文和西里尔文 LGR 则在很久以前就完成了这些工作，所以实际上 [已经延期了。]

我现在快速解释一下我们的工作流程，系统中的所有变体都分别记录在四个 LGR 中。顺便说一下，这与我们之前针对 C/J/K 的工作流程有点不同，但是...基本上，与四种语言文字相互依存的变体都分别记录在各个 LGR 中。所以，我们基本上可以将这四个 LGR 视为一种具有一致性的集合，以便...如果你在一个 LGR 中发现了与另一个 LGR 相关的变体，那么就会在那个 LGR 中找到对应的变体。

也有例外情况，我们称之为“通用字形”。基本上，O、圆圈或 C 之类的字符就是这种情况。那些字符实际上存在于与希腊文无关的其他 LGR 中 [听不清]，例如亚美尼亚文、西里尔文、希腊文和拉丁文。某些已进入根区的南亚文字中会有这种字符，但它们与这四种

文字没有任何关系。所以我们决定，作为一种整合机制，我们在整合时只列出拉丁文 LGR 中包含的那些字符。其他字符显然 [听不清] 字汇，但我们不希望它们出现在元素 LGR 中以便减少干扰，并且我们不会针对它们编制变体字符集，否则它们可能导致变体字符集过于庞大，同时并不能带来更多有用的信息。

正如我之前所说的，在确定标签冲突时，我们会使用一个合并文件，我们称之为“通用 LGR”。我们始终使用各个要素 LGR 来执行有效性验证，这一点没有改变。有两个 LGR 的工作出现了这种延期情况，即亚美尼亚文和西里尔文。因为在最终确定四个 LGR 之前，我们并不能确定会发生什么情况，尤其是当我们加入文字内部变体时，基本上就是将一种语言文字引入到另一种语言文字中，某种语言文字的两个字符之间的变体...这种情况实际上很常见。我们曾经遇到过这种情况。例如，拉丁文 LGR 向西里尔文引入了一些文字内部变体，这些变体原本不在西里尔文提案（原始西里尔文 LGR）中。我们基本上就必须将这些文字内部变体添加到西里尔文 LGR 中。

因此，我们看到延期的 LGR 不得不做出一些技术更改，但我们确实曾与生成专家组或者至少与[正式]生成专家组的成员进行过沟通，确认那是可以接受的。整合将会纳入那些添加的映射，这将导致与亚美尼亚文和西里尔文的原始 LGR 有所出入。但实际上，GP 成员已经就此进行过磋商，确认它们是可以接受的。显然，完全整合的 LGR 集将随根区 LGR 5 再次接受公众评议。请播放下一张幻灯片。

根区 LGR 5 包括现有的 18 种语言文字，那是根区 LGR 4 的一部分。我们添加了两种语言文字，两个新的 C/J/K，因为我们之前已经有了中文，所以现在添加了日文和韩文。我们还添加了两种新的字母语言文字，即拉丁文和希腊文。就像我之前说的，有两种字母语言文字



的工作已延期，即西里尔文和亚美尼亚文。最后是缅甸文，其相关工作仍在进行中。缅甸文[听不清]有望很快开始进行公众意见征询。我不想 — 这更像是一则 ICANN 官方消息，而不是我的个人意见。

根区 LGR 5 的目标是包含 25 种语言文字。有两种语言文字的工作还没有完成，即塔安娜文和藏文。我之前好像看到哪张幻灯片中提到是 28 种。我不确定。我只数出了 27 种，少了一种。或者也许另一张幻灯片上提到说是 28 种。我不知道，但我只数出了 27 种。

皮提南·古阿莫帕塔纳： 你好，米歇尔。日文有两种，就是日语片假名和日语平假名。

米歇尔·叙伊尼亚尔： 原来如此。好的。是的，我心里还在嘀咕为什么数出来的数字会有出入。不管怎样，我的介绍就是这些了。我后面还有一张幻灯片，但那只是供大家参考的。所以我的发言结束了。我想现在应该进入提问环节了，欢迎大家提出问题。谢谢。

皮提南·古阿莫帕塔纳： 谢谢米歇尔。萨尔麦德，你要不要继续提你刚才说的那个问题？或者那个问题已经得到了部分解决？

萨尔麦德·侯赛因。 谢谢，皮提南。我这就读出我们的问题。从某种程度上说，米歇尔的发言可能已经解答了这个问题，但我们还是再次提出问题，看米歇尔是否想补充任何其他内容。这个问题是唐娜·奥斯汀 (Donna Austin) 提出来的。刚才提到日文 LGR 中没有变体，是否有其他语言

文字也得出了这个结论？就是这个问题。关于不允许使用变体，是否有其他语言文字也得出了这个结论？

米歇尔·叙伊尼亚尔：

重申一下，并非如此 — 变体是有的。许多其他 LGR 中都有变体。但是大部分都会阻止变体。分配变体会导致很复杂的情况，因为如果允许分配变体，从 DNS 的角度来看，情况就会很复杂。人们更多地... 我们从 IP 方面进行尝试，经过不断努力，我们发现如果定义可分配方案，那就必须最大限度地减少可分配变体的数量。以中文为例，我认为最多可以存在五个标签，包括原始标签在内。申请原始标签后，就只能有额外四个可分配变体标签。最多不能超过这个限制。

拉丁文 LGR 中也有相关缓解措施来限制可分配标签的数量。大多数语言文字的处理都相对简单。但日文的情况确实很复杂，我们曾与日本 GP 一起努力寻找合适的解决方案，设法为申请标签创建数量有限的可分配变体。但始终没有找到合适的解决方案。我们反复研究，尝试从不再角度寻找解决方案。

但我们始终找不到合适的解决方案 — 在最坏的情况下，变体规模甚至会出现爆炸式增长，达到成千上万个 — 如果你创建了一个糟糕的标签 — 所谓糟糕，我的意思是指可能导致发生最坏的情况，而这种现象很普遍 — 即使你有相应的工具 — 工具不会 — 枚举数量也会非常多，情况会非常复杂，以至于该工具就算持续运行数天也无法给出任何解决方案。

重申一下，基于这种情况，我们不能有太多可分配的变体。因此，经过与日文 GP 磋商，我们得出了这个结论：唯一的解决方案就是

阻止所有变体。这就是我们得出这个结论的原因。但再次重申一下，变体是存在的，只是都被阻止了。

顺便说一下，许多 LGR 现在都是这样处理的。许多 LGR 都会阻止变体。并非只有日文 LGR 如此。

皮提南·古阿莫帕塔纳： 谢谢米歇尔。还有什么问题吗？

萨尔麦德·侯赛因： 罗伯特·兰卡姆布威 (Robert Nkambwe) 提出了一个问题。“这项工作与 UA 是否存在包含与被包含的关系？很抱歉提出这个问题，我对 ICANN 还比较生疏。”他的问题是：根区 LGR 与 UA 是否存在包含与被包含的关系？

聊天室中已经有人给出了一些回答，皮提南你要不要回答一下，如果你愿意的话。

皮提南·古阿莫帕塔纳： 当然。谢谢。谢谢你提出这个问题，罗伯特。是的，二者实际上是存在关联的。从某种意义上说，IDN 是生成非 ASCII 或国际化域名的机制 — IDN 是在技术上可以使用的本地语言文字域名，而至于如何让互联网上的系统、网站或应用程序接受这些 IDN，那就是 UA 的范畴了。

另外，UA 不仅仅与 IDN 有关。它相当于一种互联机制。UA 一方面用于确保 IDN 被互联网上的应用程序所接受，另一方面[音频不清晰]，即新 gTLD，新的或长或短的 TLD。除了常见的点什么什么域名

之外，你可能已经看到过[点新内容]的域名，比如现在我们有更长的 .photography，以及新的较短的点域名，例如 [.run] 等等。因此，UA 涵盖 IDN 和非 IDN 域名。从这方面说，它更像是一种互联机制。所以，UA 和 IDN 之间的关系是相辅相成的。不知道这样回答是否解决了一

萨尔麦德·侯赛因：

谢谢。好的。我们实际上还收到了一条评论，我认为最好是由米歇尔来回答，如果可以的话。这条评论也是来自罗伯特·兰卡姆布威。他说，“我以为在整合到根区之前会进行严格的测试。但流程中似乎没有体现这一点。”我们的时间不多了，米歇尔，你能否介绍一下我们在整合之前会对根区 LGR 提案进行哪些测试？谢谢。

米歇尔·叙伊尼亚尔：

好的。我们会要求每个 GP 提供一组测试标签。我们希望通过这组标签来对 LGR 提案中的字汇执行这些规则。我们还会验证现有根区的现有 TLD 能否通过 LGR。如果无法通过，那就有点糟糕了。因此，我们确实会验证现有的已授权根区标签。

我们还会用数据来测试 LGR，基本上就是利用各种现有数据，我们会使用非常大的数据集副本来测试每个语言文字的 LGR。大家可以在网上找到那些信息。我们还会验证大多数常用词能否通过 LGR。当出现偏差时，我们会回头询问 GP 为什么会出现这种情况。有时是因为特定语言文字的 LGR 经过一些 [听不清] 修订或其他更新。有时则涉及一些历史原因。GP 每次都会给出令我们满意的回答。

---

所以，我们确实会通过多种机制进行测试。那些测试往往是在整合之前完成的，因为整合实际上主要是一个机械过程，基本上就是根据根区 LGR 将 LGR 调整至一个合理的状态。但并非 — 在此之前，GP 会通过 IP 协作来完成各项技术工作。

皮提南·古阿莫帕塔纳：

谢谢米歇尔。我觉得[这样回答已经足够详细了。我们已经超时了]，现在我要宣布会议结束了。谢谢大家参加本次会议，谢谢大家的互动。请大家积极参与那两个公众意见征询程序，希望大家多多提供反馈。不再啰嗦了，我要结束本次电话会议了。祝大家今天过得愉快，祝大家在后面的 ICANN 会议活动中一切顺利。希望我们下周能再见。再见。

[会议记录结束]