土著语言：

从零开始走向语言数字化

如何让您使用的语言走进互联网

TRANSLATION COMMONS

2022-2032 | INTERNATIONAL DECADE OF Indigenous Languages

土著语言：从零开始走向语言数字化

作者：Deborah W. Anderson, Lee Collins, Craig Cornelius, Craig Cummings
其余作者和审校：Andrew Owen, Julia Nee, Lawrence Wolf-Sonkin, Anna Luisa Daigneault, Julie Anderson, Daniel Bogre Udell, Julie Anderson, Daniel Bogre Udell
插图与宣传：Mette Attar, Johanna Behm, Leonidas Pappas
项目协调：Ester Perez, Jeannette Stewart

1. 前言

Translation Commons是一个非盈利性的志愿者组织，我们致力于为各类语言的数字化提供帮助、为语言专业人士提供指导，并为语言行业提供课程及资源。

语言数字化倡议（LDI）我们进行的主要项目之一，该项目旨在帮助那些急需提高数字化能力的语言群体。全球有近六千种语言都没有被数字化，或者只被数字化了一小部分。而语言数字化倡议为这些语言群体提供了语言数字化流程指导。

我们与联合国教科文组织下属的2019国际土著语言年（2019 International Year of Indigenous Languages）行动一起，聚焦土著群体，探寻土著语言的数字化之路。土著语言使用者人数较少，因而在数字化世界中很难找寻到用土著语言所记录的内容，而这对于土著语言群体来讲是不公平的。语言数字化倡议的目标之一就是要保障这些人用土著语言获取网络信息的权利，从而确保他们能用母语参与全球网络活动，能够使用其母语版本的电脑软件，享受到现代电脑软件所带来的便利。本指南能够为这些语言群体提供数字化工具，提高其对语言数字化的理解，将土著语言带入到数字化世界，从而帮助他们加速语言数字化的进程，与此同时也保证了语言群体的自主权。除了制定本指南之外，我们还向这些语言群体提供教程以及组织开展研讨会，同时我们会向其介绍行业专家来提供标准化指导，以帮助这些语言群体完成语言的数字化。

1.1《从零开始走向语言数字化》系列

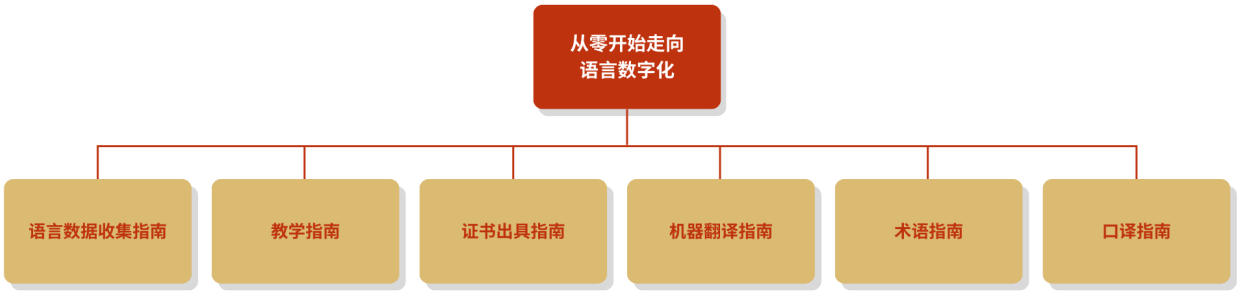
本指南是*《从零开始走向语言数字化》*系列指南中的一本，该系列为语言的数字化实践提供了全方位的指导。本系列指南由语言技术和语言学方面的专家共同编写而成。目标受众是所有希望能在数字化系统中使用自己母语的语言群体。

语言的数字化能够扩展语言群体的交流渠道。*语言数字化的益处*中详细说明了语言数字化能如何造福土著社群及世界。

欲了解语言数字化详细流程，Translation Commons网站中资源栏目发布了更多指南、演示和视频等语言数字化倡议的相关信息。

下方图1列举了所有为帮助土著群体而撰写的指南。

图1：《从零开始走向语言数字化》系列指南



本指南将主要介绍如何使手机及电脑应用支持您所用语言的书面语形式。指南中推荐的方案可以帮助以土著语言为母语的群体进行网络交流，在网络上分享知识和文件，以及使用各种之前无法使用的应用程序与设备。

本指南主要面向的群体有：

- 希望能够在移动设备及电脑上使用自己语言的土著群体
- 希望在语言数字化过程中提供技术支持的专家
- 希望帮助土著语言群体的各类组织

本指南旨在帮助您确定所需的工具，以及其使用方法。也会帮助您寻找到适宜的工具，使您能够在网络上使用自己的语言。

很多人问怎么才能在网络上使用自己的语言，而这个问题可以有很多种解决方法。想要在网络上使用一门语言，人们首先应该考虑到几个不同层次的技术（包括基于网络的技术以及移动设备技术）。

本指南还会讲述如何确保一门语言的使用者、读者以及书写者能够在网络上运用其书面语。网络上使用书面语的形式丰富多样，包括：日常对话、短信、邮件、社交媒体以及博客。本指南旨在帮助人们建立网站，创作内容，并促进人们与地方语言群体之间，以及与散居各地语言使用者之间的交流。对于某门语言来说，可能存在多种文字形式或书写习惯。本指南的目的不在于限定人们使用数字技术的方式，而是为了帮助人们更好地使用文内介绍的这些技术，提高土著语言在全世界的声望，增进公众认知，让人们能更方便地使用这些语言。虽然正式的文献研究、语法学和字典学在语言学研究和规范语言使用方面很有价值，但如果只是想在网络上使用一门语言，没有这些也没有关系。

本指南在指导委员会监督下制作，由特设小组提供建议，并在各方合作伙伴的帮助下协作完成。

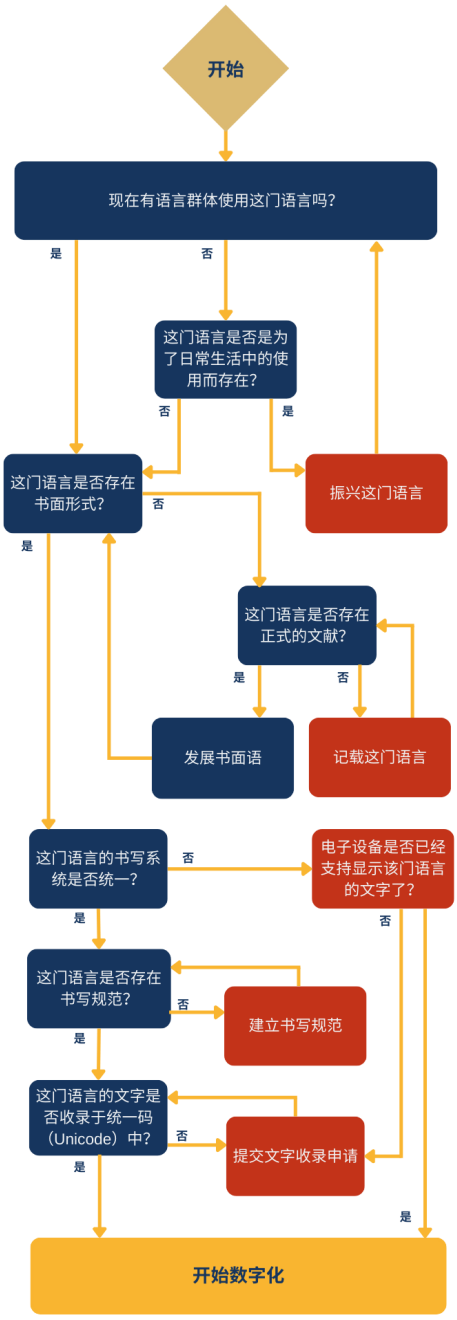
2. 流程概述

本指南提供了两套工作流程，第一套用于确定一门语言的现状，另一套用于制定技术方案，以实现语言的数字化。流程图中的每一步在本指南中都有对应章节详细说明。流程仅供参考，一些步骤也可以同时进行。

2.1. 语言现状工作流程

在网上开始使用一门语言前需要先确定它的现状，本流程（见图2）描述了这一过程。

图2：如何确定一门语言的现状

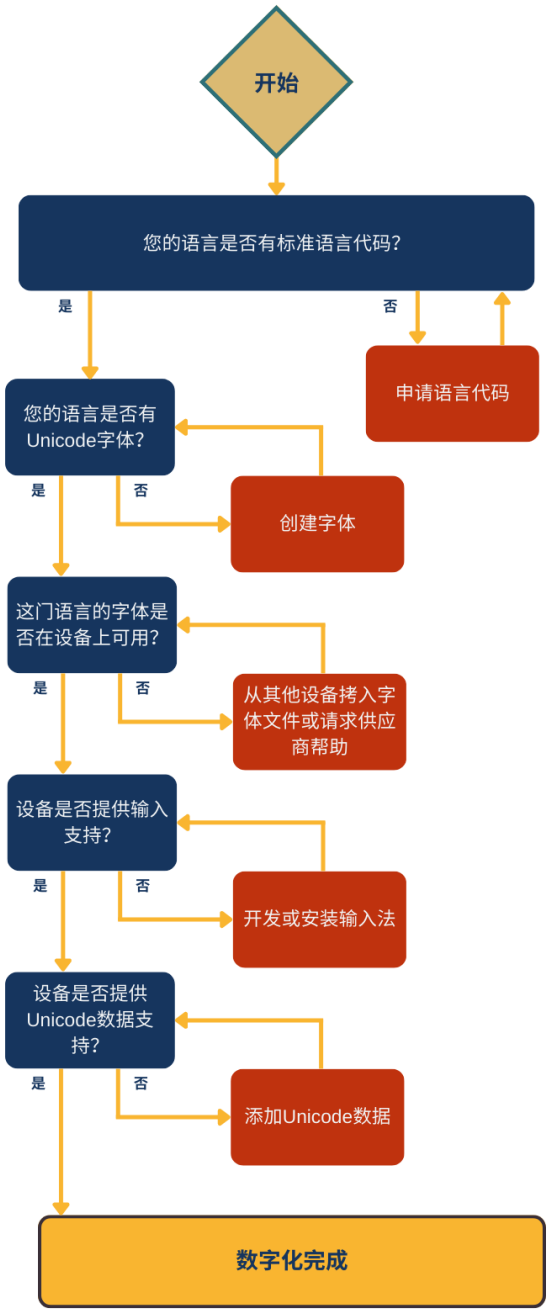


2.2. 语言数字化技术工作流程

在确定了一门语言的现状后，请根据以下工作流程（见图3），并运用已有的技术数字化您的语言。

图3：语言数字化技术工作流程

语言数字化技术工作流程图



3. 语言现状

本节将讲解确定语言现状工作流程中的各个步骤。这一流程主要是为了确认目前的计算机及移动设备对您所用语言的支持程度。同时还会针对整体方案的下一步提出建议。完成以下这些问题无需您本人有任何技术背景。

3.1. 现在有语言群体使用您所用的语言吗？

是：这门语言是否存在书面形式？
否：人们是不是在日常生活中使用这门语言？

3.2. 人们是不是在日常生活中使用这门语言？

是：振兴这门语言
否：这门语言是否存在书面形式？

3.2.1. 振兴这门语言

一门语言就算目前没有人会说，但只要将使用这门语言的群体组织起来并投入足够的精力，同样能够得到振兴。这个过程需要参考大量的语言文献资料，也需要教材和能用这些教材授课的教师，以及整个群体共同付出巨大的努力，进行数年甚至数十年的苦心耕耘。这样您才可能振兴这门语言。

新西兰在振兴毛利语（Māori，也叫做Te Reo）时就参考了一部分振兴威尔士语的方法。新西兰采用的方法包括：在各种媒介上使用土著语言进行广播、以及开设语言巢学校（使用毛利语授课的学校）等等。爱尔兰语的振兴也采取了相似的方法。世界各地的其他例子还有：

- **希伯来语**：即使在连续几个世纪没有人会说希伯来语的情况下，人们也成功振兴了这门语言。
- **切罗基语**：社群及政府部门为儿童打造了沉浸式语言学校（全天候使用目的语进行教学的全封闭学校），并开设了小学到高中都可选修的切罗基语课程，同时也为成人提供了相关培训。
- **穆特松语**：美国加州的阿玛·穆特松（Amah Mutsun）部落设立了语言振兴项目。
- **查克马语**：孟加拉国和印度东部的一门语言。当地群体已经开始重新使用查克马语进行教育和扫盲工作。
- **图尼卡语（Tunica）**：图尼卡语原先是美国图尼卡-比洛克希部落使用的语言，1948年最后一位母语使用者去世，图尼卡语进入休眠状态。此后，一批将图尼卡语作为继承语使用的人开始培养能流利使用这门语言的人才，并鼓励大家走进完全以图尼卡语教授图尼卡语的课堂，最终重新唤醒了这门语言。

- **康沃尔语**：康沃尔语于二十世纪初开始走向振兴，到了二十一世纪，说康沃尔语的人充分利用网络论坛，相互取得联系，从而得以每天使用这门语言，发扬康沃尔语的运动也借此取得了巨大进展。

3.3. 公共资料库是否记录了您所用的语言？

如果这门语言在公共资料库中或是语言库中。

是：这门语言是否存在书面形式？
否：这门语言是否存在标准语言代码？

3.4. 这门语言是否存在书面形式？

这门语言是否有书面语？

是：记录这门语言
否：发展书面语

3.4.1. 发展书面语

本指南不适用于没有书面语的语言，不过您依然可以通过音频和视频的方式在网络上使用自己的语言。对于那些主要以口语形式存在的语言，有其他组织可以提供相关指导和工具。某些学术机构下属的语言学部门也可以提供相关帮助。

当语言使用者和研究人员能够非常方便地找到该语言的音频及视频资源时，这些资源便能发挥出更大的作用。录音和录像中需要置入语言标签或语言代码，以便人们使用自动索引找到这些数据。可以采用例如像IETF BCP-47这样的标准语言代码。

3.4.2. 记载这门语言

是否有正式文献记载了您的语言？例如有语法学、词典学，或语言学等的相关研究。

是：语言已得到记载
否：语言未得到记载

3.4.2.1. 语言已得到记载

如果存在关于这门语言的字典和语法讲解等相关语言学资料，使用这门语言的人能否获取到这些资料？应想办法让相关群体更容易获取到这些资料，从而给他们带来更多的好处。比如可以在网络建立资料库，并为小学及沉浸式学校研发适宜的教材，并将这些资料的版

权转授给相关语言群体。还可以想办法取得或者制作电子版资料，这样各语言群体就可以根据情况和需要将它们发布到网上。

如果无法找到上述资源，则可以想办法与相关语言群体共享书籍、数字文库和其他有关这门语言的信息。

3.4.2.2. 语言未得到记载

就算这门语言尚未得到记载，人们仍可通过口语或者书面语的形式使用这门语言。然而拼写纠错、文本搜索和预测文本等功能就会受到限制。

3.5. 您所用的语言是否有一个统一的书写系统？

书写系统是指在书写一门语言的一种或多种文字时需要遵循的规范。很多语言的书面语中都存在多种不同类型的文字，比如很多人书写塞尔维亚-克罗地亚语时，会同时用到西里尔字母和拉丁字母两种形式的字母。很多情况下，一套书写系统也会由多种语言共用。比如缅甸语、掸语、孟语等好几种语言都可以用缅文书写。

这门语言是否至少有一套统一的书写系统，同时拼写规则也遵循这套标准？

是：这门语言是否存在书写规范？
否：电子设备能否显示这门语言的文字？

3.5.1. 电子设备是否已经支持显示该门语言的文字了？

如果这门语言没有统一的书写系统，那其书写可能并不规范，且没有遵循统一的拼写或语法规则（甚至使用不止一种文字。）尽管人们也可以通过现有的工具在电子设备中使用或阅读到这些文字，但拼写及语法检测工具在这种情况下就只能发挥有限的作用了。如果一门语言同时存在多套书写系统，我们一定要明白一点：不同的正字法（关于文字使用的规范性法则，包括拼写、连字符、大小写、断字、重音符号和标点符号的规范）可能代表了不同语言群体的利益。

您可以向语言学家或技术人员寻求以下帮助：

- 不同群体在书写时会使用不同的正字法与字符集，其中还会涉及不同的方言。键盘等语言工具的开发者应当明白这一点，并在工具中囊括多套字符集、附加符号及拼写检查方案，以满足不同人群的需求。
- 技术人员需要制定方案以识别同一语言的不同方言，同时根据不同群体的需要，尽可能简化不同方言之间的转换。
- 如果电子设备已经能以某种方式支持显示一门语言的文字时，技术人员则可以和该语言的使用群体一起确定字体、开发键盘。

- 如果书写系统不一致，妨碍到了人们使用这门语言，当地的教育工作者、政策制定者、语言群体代表和语言学家等相关方可以共同制定一套更加统一的书面语系统，以便人们在网上使用这门语言。

3.6. 这门语言是否存在书写规范？

这门语言的书写系统是否有一套人们普遍接受的正字法和语法规则？（不论其是正式的规则还是非正式的）

就算没有规范，人们依然可以进行非正式的书写与交流。不过，若使用非正式的单词拼写、语法和词汇，没有一套标准的语言规范，亦或是不同地区间的书写习惯存在很大差异的情况下，想要进行下列高级操作时就可能遇到困难，例如：在社交媒体上使用这门语言、共享文件、进行在线检索或通过其他工具搜索网站及信息等。

是：提交文字收录申请
否：建立书写规范

3.6.1. 提交文字收录申请

一套完善的语言规则（包括拼写、语法、标点等）是开发拼写纠错、文本预测、在线搜索和写作辅助等功能的基础，所以拥有一套语言规则将对人们使用语言产生巨大的帮助。

Unicode标准（统一码标准，是计算机科学领域里的一项业界标准。它为每种语言中的每个字符设定了统一并且唯一的二进制编码，以满足跨语言、跨平台进行文本转换、处理的要求。）是用于规定文字的标准，而非用于规定语言本身。一套文字系统可以被多种语言使用，比如英语、斯瓦希里语、印度尼西亚语等众多种语言中都使用了拉丁字母。如果Unicode支持您的语言所用的文字，请参照2.2章语言技术工作流程进行下一步。

如果一门语言已有书写系统而且已投入使用，但其文字尚未收录于Unicode标准中，对文字进行标准化便十分重要，因为这有助于人们在移动设备、笔记本电脑和台式机等可以联网或相互传输数据的设备上使用这门语言。如果文字未经标准化，但只要用户使用的是同一种字体和输入方式，他们仍然可以互相传输文件。但用户在使用网络工具和服务时就会受到限制。

为确保您所使用的文字达到Unicode标准：

1. 请检查这门语言使用的文字字符是否收录于Unicode中，或是否受Unicode支持。
2. 如果Unicode没有收录或不支持您所使用的文字，请提交收录申请。

3.6.2. 建立书写规范

虽然这一步并不是必须要做的，而且不同语言也可能有不同的书写方式，但如果您所在的语言群体准备规范语言文字，那么则需制定拼写、标点和语法使用规范，以确保人们能够使用网络进行交流。这一步也可以为优化输入法的词语联想和拼写纠错功能打下基础。此外，使用规范的语言文字也更容易在搜索引擎和其他网络服务中找到有用的信息。

3.7. 进行语言技术工作

当Unicode已支持您使用的文字，且您语言书写系统中涉及的文字字符已收录于Unicode中时，您便可以参照语言技术工作流程开始进行下一步。

4. 语言技术工作流程

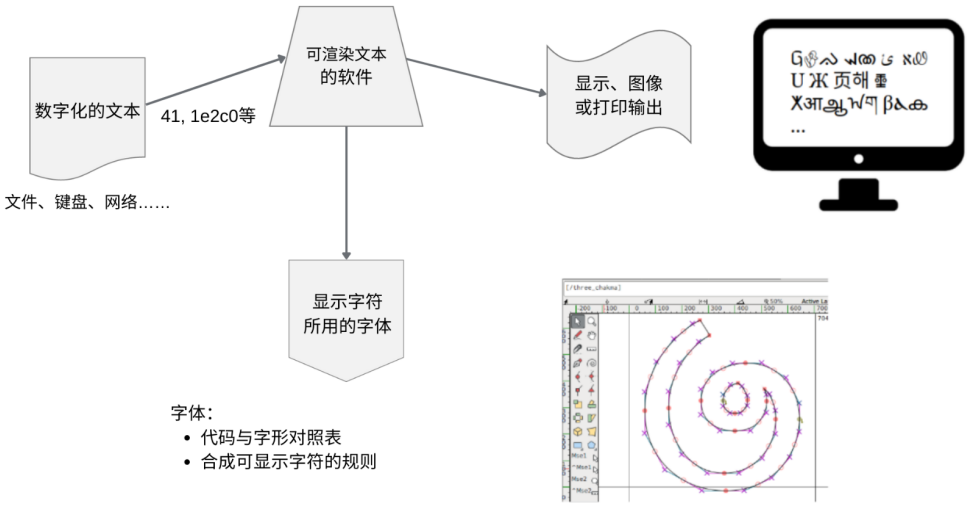
4.1关于文本数字化技术的说明

电子设备通过二进制数列储存字符，这样的数列称作*码点*。比如U+0041是字符“A”的码点。人们可以通过键盘输入得出码点，其可以存储在文件中，或在不同程序间传输，而某些处理字符编码的系统也可将码点*渲染*为文字。

*编码标准*为每串码点赋予了各自的含义。比如ISO-8859-1标准下编码的范围是0到255，这些个码点各自代表一个字符（文字或字母）。Unicode字符编码标准囊括了多种书写系统，为150多种文字提供了编码支持，其中包括拉丁字母、西里尔字母、中国文字、阿拉伯字母、希伯来字母、天城体文字、泰米尔字母和缅甸文字等等。而Unicode为这些语言的每一个文字或字母提供了一个码点，从而将它们数字化。

将码点与编码标准结合起来后，人们就可以通过与编码标准兼容的输入方式打出由编码组成的文本。因为文本的字符已收录在预先设置好的编码标准中，因此可以通过与上述标准兼容的字体重新显示为可读的文本。下图（图4）描述了上述的整体过程，由图可得Unicode等编码体系能够支持多种语言的文字。

图4：Unicode编码系统支持显示多种语言文字的全过程



很多语言的书写系统会在一些基本字母上加上附加或修饰符号，比如有的语言在使用中会给“e”加上重音符号，写作“é”。因此，一套编码规则可能包含另一功能：在基本字母后键入附加符号代码，就可以打出组合字符（类似上文中的é）。有时编码规则内包含了预组合字符，也就是将基本码点和组合码点合并成了一个码点。

因此，任何在电子设备中呈现的书面语都需要有：

- 像Unicode这样标准的字符编码系统
- 适用于这门语言所用文字的字体和渲染系统
- 能将码点渲染为文字并呈现到所需媒体设备上的方法或应用程序

本指南主要讲述了Unicode编码标准的相关知识，因为所有现代电子设备都在使用Unicode标准，同时它也是许多源文本采用的默认编码标准。Unicode为语言的数字化提供了极大的便利。

4.2. 语言数字化过程中所涉及到的技术性术语

本节旨在为语言数字化提供技术上的支持。很多语言使用的不是标准的字体和文字。这虽然并不妨碍人们在网络上使用一门语言，但是会导致使用者无法使用一些高级功能。以下步骤涵盖了语言数字化过程中所需要的各种技术支持。

要想使一门语言的书写系统数字化，至少需要以下几个基本要素：

- **字符（文字）**：首先要记录这门语言所有的字母或文字，包括：附加符号、合字、标点、数字、表意符号以及其他正字法涉及的符号。

- **ISO标识符**：用来标识语言和文字，其中也包括一些地域性代码，比如*es-MX*代表在墨西哥使用的西班牙语。
- **编码标准**：这个标准可以是正式的标准（如能将字符分类、定序或组合起来的Unicode标准），也可以是非正式的标准（如字体编码标准）。
- **字体支持**：若想设计一款包含所有字符的字体，则需要下列两项信息：
    - 与文本（如连体字、合字和组合字符等）的呈现效果有关的详细信息（Unicode文档中或已收录相关信息）
    - 供字体设计师和开发人员用于测试的示例文本
      在提供字体支持时，还可能需要更新渲染引擎或其他软件
- **输入方式**：向电子设备中输入文字的方法，最常见的输入方式是通过实体或虚拟键盘进行文字的输入。

一些补充资源包括：

- 通用语言环境数据存储库（CLDR），内含和一门语言相关的补充信息，如日历系统等；
- IETF BCP-47等语言代码标准；
- 支持分词功能的应用程序；
    - 无明显缺失的词典；
    - 空格，标点等符号的使用规则。

4.3. 您的语言是否有标准语言代码？

您的语言是否有标准语言代码，且该代码已得到认可？

是：这门语言是否有Unicode字体？
否：为这门语言申请语言代码

4.3.1. 申请语言代码

如果您所用的语言没有语言代码，您则需创建标准语言代码。其中需包含该语言中存在的方言或方言字。非标准语言代码或标签不仅会给资料的检索过程带来混乱，而且程序在识别语言资源的标识符时也可能会报错。在创建语言代码的过程中请遵循ISO-639和IETF BCP-47标准。

4.4. 您的语言是否有Unicode字体？

使用相关语言的人士能否在台式或笔记本电脑上使用与Unicode兼容的字体？

是：电子设备是否支持这门语言的字体？

否：创建字体

4.4.1. 创建字体

Unicode在收录一种新的文字后，工作人员可能还未开发出支持这种文字所用字符的Unicode字体。出现这种情况时，请联系字体设计师及技术人员，共同创建支持您所用文字的Unicode字体。

4.5. 电子设备是否支持您所用语言的字体？

相关群体所使用的各种设备是否能支持显示这种字体？如果可以，网页、社交媒体及其他应用程序上的文字都可以正常显示。

这里还需指明一点，对大部分电脑及移动设备来说，若能使用符合Unicode标准的文本、和Unicode兼容的字体，以及和Unicode兼容的输入法，其显示效果便能实现最大化。请检查Noto字族（此字族是一个谷歌推出的免费字体集合）或其他提供Unicode字体下载的网站有没有您所用语言的字体。请注意，您可能在很多移动设备上都无法直接安装或下载字体，不过您可以在台式及笔记本电脑上直接安装已下载的字体。

若您想让文档和网页上的字体正常显示的话，您可能需要先进行文字处理软件和浏览器等应用的配置。

各个网页可以用其自带的Web字体（互联网中网页所用的字体）显示文本，这样网页开发者就可以选择他们想用的字体，而且即便某个设备上未曾安装所需字体，网页内容也可以正常显示。请注意，Web字体只能在与其兼容的网页中使用，且此字体并没有永久安装在相关设备中。最新的Noto字族中的所有字体都可以用作Web字体，这也能为没有对应字体或不支持其安装的设备提供帮助。

是：相关电子设备上是否安装了对应的输入法？
否：手动安装字体或向供应商寻求帮助

4.5.1. 手动安装字体或向供应商寻求帮助

有些语言的书写系统刚刚完成标准化，虽然人们已经可以在某些电子产品中使用这些字体了，但它们还是和很多移动设备不兼容。请敦促设备供应商提升相关字体的兼容性。

您也可以在移动设备上手动安装Unicode字体。此操作同样可以让您在移动设备中使用您想使用的字体，但这也需要相关的专业知识。

注意：在移动设备上安装从网上下载的字体可能带来安全及隐私风险，也可能导致设备的保修或技术支持服务失效。

4.6. 相关设备是否安装了对应的输入法？

语言的使用者需要一种有效的输入法帮助他们高效地撰写信息，邮件，博客或基于该语言的其他内容。相关设备（移动设备或其他设备等）自带的输入法是否支持输入您所用的语言？

是：第三方输入法或设备是否支持这门语言？
否：开发输入法

4.7. 第三方输入法或设备是否支持这门语言？

系统的标准键盘是否内置了该语言的输入法?如果支持，请学习如何启用或安装适用于这门语言的输入法。具体操作每个设备可能不太相同，不过很多在线资源都可以提供帮助，可以上网查找。

移动设备供应商在iOS及安卓系统上已提供很多现成的输入法。例如：

- Gboard，谷歌为移动设备开发的输入法（https://support.google.com/gboard/answer/6380730?hl=en&co=GENIE.Platform=Android），支持iOS及安卓系统上的多种语言；
- Google输入工具（https://www.google.com/inputtools/）支持在电脑端Chrome浏览器的标签页中使用虚拟键盘（手机端暂不支持）；
- 苹果和谷歌应用商店（App store和Google Play）内也有很多第三方输入法。

值得一提的是，很多语言的键盘布局已经公开在各大数据库当中，例如通用语言环境数据存储库（CLDR）下的键盘数据库。

4.7.1. 开发输入法

有很多工具可以帮助我们安装不同语言的输入法或开发新键盘，比如：

- 苹果、谷歌应用商店或其他来源的键盘软件；
- Keyman：虚拟键盘软件（https://keyman.com/）；
- 微软键盘布局设计软件（https://www.microsoft.com/en-us/download/details.aspx?id=22339）；
- SIL Ukelele（https://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=ukelele）。

注意：安装下载的键盘或其他输入法可能带来安全及隐私风险。

4.8. 设备是否提供Unicode语言数据支持？

该语言是否已有Unicode语言数据？

是：可以开始在网络上使用您的语言。
否：请按需向通用语言环境数据存储库（CLDR）中添加语言标签、语言名称等必要的语言数据。关于通用语言环境数据存储库（CLDR）的更多内容请见《更多语言帮助》。

5. 更多语言帮助

Unicode标准涵盖了书写系统和语言范围内文本处理的很多方面。如果字体和输入法的问题得到解决，大部分文字处理、工作表和邮件方面的困难，还有很多其他方面的问题都会迎刃而解。

这里列举的步骤仅仅是开始。通用语言环境数据存储库 （CLDR） 通过收集不同语言环境（包括不同的语言和国家）下的数据，来“为各种应用程序提供关键信息以支持全世界的语言”。语言数据包括不同语言、国家、月份和工作日的名称，以及其他各类信息。也保证了各种程序可以针对不同区域，以特定的格式呈现日期、时间、数字等信息。

尽管用土著语言进行基础的文字交流时不会用到CLDR数据，但CLDR的存在使得这些语言的功能更加丰富完善。字体和键盘功能完善后，发邮件、信息，使用社交媒体等几乎所有操作也都会更加便利。

程序员可以利用这些信息，根据不同需要，在网络应用程序中呈现本地化后的日历、工作表、数值和菜单等UI信息。

通用语言环境数据存储库（CLDR）中还存储了关于语言的更多补充信息，例如书写系统会使用到的字符，以及用于键入的键盘布局等等，更多详情请见通用语言环境数据存储库键盘（CLDR Keyboards）页面。

不过，如果想在移动以及在线应用中获得更好的语言使用体验，可能需要更多语言数据及文字工具的支持。相比技术支持完备的语言，土著语言还需要下述几项功能：

- **文本分割与换行处断词**：想要让文本以正确格式呈现，并且用户在移动光标时能选择上正确的字素簇、字词及句子，文本分割功能必不可少。很多语言中词与词之间没有空格或标点等明显的标记，这时候就需要调用字典或相关算法来辨别什么时候可以断词。详情请见：userguide.icu-project.org/boundaryanalysis。
- **换行**：关于什么时候文本需要另起一行，不同语言有不同的规范。比如引用文本时，不同语言在引用内容前后会使用不同的字符，如何用这些字符隔开词与词、句与句，取决于不同语言的格式要求。而且，换行的具体位置也取决于书写系统中不同Uniccode字符的属性。有时候数字和货币符号也不能分开，比如像$10这种情况，否则就容易造成误解。不同语言，或者使用同一语言的不同区域之间，标点符号的

使用规则都有所差异。换行就是Unicode字词边界分析（userguide.icu-project.org/boundaryanalysis#TOC-Line-break-Boundary）的一个具体应用。

- **在文本中识别自然语言**：每篇文档都需要用语言代码等标识符明确标记，来表明文档使用的是哪种自然语言。文档得到标记后，各种工具就可以借此更快速地为用户查找到需要的信息。对于以多种语言写就的文档，每一节甚至每一段都可以用不同的标签标记，请务必遵循IETF BCP-47（https://tools.ietf.org/html/bcp47）等标签使用标准。
    - 不同应用程序识别标识符的机制不同，用户需要自行学习。值得一提的是，也不是任何标签都可以被识别，标签符需要从特定的列表中选取。
    - 针对在线文档及网站，HTML的*lang*属性（https://www.w3schools.com/tags/att_global_lang.asp）明确指定了不同HTML组件的语言标记。lang属性的值是从标准标识符中选取的语言代码，而不能是用户自定义的字符串。比如“*rs*”代表塞尔维亚语，“*de*”代表德语，“*zh-Hans*”或“*zh-CN*”（有时简写成“*zh*”）代表简体中文。

- **语言检测**：如果能明确文本使用的语言，在线服务和应用往往可以给出更精准有效的信息。cld2（https://github.com/optimaize/language-detector）等语言检测软件就是针对没有明确标记语言的文档研发的。这类工具通常会对文本内的字符进行统计分析，并据此提供一个可能的鉴别结果。很多情况下，多种语言会使用同一个书写系统，比如斯瓦希里语、拉科塔语、瓦尔皮利语和芬兰语都用拉丁字母；俄罗斯语、乌克兰语和哈萨克语都用西里尔字母；缅甸语、掸语、孟语都用缅文等等。
- **字词处理字典**：大部分字词处理软件支持基础的输入、编辑、分享及打印功能。文本预测、拼写纠错，语法建议等工具会分析单词列表，列表内还包含使用频率和字典等语言学相关数据。同义词和常用俗语方面的数据对在线搜索等工具也很有帮助。
- **非ASCII系统下的数字**：很多语言用到的数字与西方书写系统中的数字不同，比如缅文、阿德拉姆字母（Adlam）、阿拉伯字母和波斯字母等。但像制作工作表的很多软件不会把这些字符按照数值处理，而是按照文本处理。这类软件的程序员有时会参考这些字符的Unicode属性，将它们处理成数字，但并不是所有人都会这么做（https://en.wikipedia.org/wiki/Numerals_in_Unicode）。
- **译后UI界面**：对于很多软件（尤其是教育类软件和一些需要以用户的语言呈现信息的软件），UI界面里出现的文本都应该翻译，比如操作系统中“开始”和“打开文件”等功能键。但在很多情况下，想要让软件开发者为一些小众语言提供翻译服务并不现实。在UI提供的所有语言选择中，如果用户认识至少其中一种，对于翻译的需求就不会那么紧迫。
- **文字识别（OCR）**：很多语言都有大量的书面文献，比如书籍。OCR技术可以将其中的文字扫描转换成数字版本。现在有很多开源的OCR程序（https://pdf.iskysoft.com/ocr-pdf/open-source-ocr.html），可以训练它们学习新的书写系统。请务必留意，带有常用词语清单的语言模型可以极大地提升OCR程序的精确度。

5.1. 公开的语言资源

有很多免费或订阅制资源供语言群体使用：

5.1.1. 语言资源

- Panlex（https://panlex.org/）
- Wikitongues（https://wikitongues.org/）

5.1.2. 工具资源

- SIL - Keyman
- 字体工具
- 用于字典的开源工具

5.2. 高级语言技术

以下功能需要大量数据以训练机器学习模型。尽管开源的公共软件越来越多，但目前该领域内最多的还是学术研究和商用产品。对于大部分语言来说，以下这些功能在近期内仍然无法实现。

- **语音转文字**：识别出人类的语音并将其转化为文字。转化的文字可用于输入文字，或用于语音控制应用程序和设备。
- **文字转语音**：按照文本合成人类的语音。这样用户无需动手即可实现人机交互，机器也能为人类朗读文字。
- **学术音频转录**：这项技术对于语言档案编制工作，特别是对语言学研究的档案编制来说十分重要。一些开源软件正在着手解决这个需求，比如：
- **语言学家音频转录助手（Accelerated Transcription for Linguists）**：https://github.com/CoEDL/elpis
- **机器翻译**：用于将一种语言转换为另一种语言，这对计算机来说是最困难的任务之一。现在的系统只能在为数不多的语言间翻译，而且机器翻译无法理解语境，质量也比不上人工翻译。不过在新的机器学习技术下，机器翻译借助庞大的语料库，翻译质量得到飞速提升。现在机器还没有办法翻译大部分土著语言，不过很多高校已投入研究，很多翻译系统也进行了开源，成果正不断涌现。比如www.apertium.org就是为帮助机器翻译小众语言而开发的在线工具。

6. 本文术语表

ASCII：美国信息交换标准代码，电子通信的字符编码标准。

BCP-47：用于识别语言的*IETF*标签。

CLDR：通用语言环境数据存储库，用于提供更多语言信息。

字符：书写时具有语义价值的最小单元，用于指抽象的含义或字形，而不是特定的字形。另见：Unicode术语表（Unicode Glossary）

码点：代指特定字符或格式的数字。

附加符号：在字母或基础字形上加上的字符，通常为了改变发音或区别语义。

休眠语言：现在没有人会说的语言。

字体：若干字形的集合，用于以视觉形式呈现字符。

字形：单个用以指代字符的可辨认符号，书写时使用。

语法：使用*自然语言*时各语法单位结合的规律。

IETF：互联网工程任务组。

土著语言：某一特定地区中人们的母语。

IYIL2019：2019国际土著语言年。

语言振兴：扭转语言的颓势，或复苏*休眠语言*。

合字：由两个或多个字形结合而成的单个字形。

自然语言：人类社会中自然演化出来的语言，与编程语言等形式语言相对。自然语言包括口头语言、视觉语言、视觉与手势结合的语言（手语）以及书面语言。

Noto：包含了100多种字体的字族，旨在收录所有有Unicode编码的字符（目前收录了Unicode6.0及之前版本中涵盖的字符）。

正字法：某一语言文字的传统书写规则。

私人使用区：英语为缩写为PUA，指*Unicode*标准中不指定用途的一系列*码点*。

标点：空格和其他不发音符号，作用是帮助人们理解文本。

文字：字母、文字和其他书面符号的集合，文字通过一个或多个书写系统以文本形式呈现信息。

Translation Commons：自由分享语言方面知识的在线社区和平台。

UNESCO：联合国教科文组织。

Unicode：意为“统一码”，最常用的对各个书写系统中使用的字符进行编码的标准。

书写系统：使用一门语言的一种或多种文字时需要遵循的规范。

7. 参考文献

7.1. 语言振兴

《劳特利奇语言复兴手册》（Routledge Handbook of Language Revitalization），（Hinton, Huss, & Roche 2018）
《实践中的语言复兴绿皮书》（The Green Book of Language Revitalization in Practice），（Hinton & Hale 2001）
《拉美地区语言文献记载及振兴》（Language Documentation and Revitalization in Latin American Context），（Perez-Baez, Rogers, & Roses Labrada 2016）
《没有书面形式的语言如何发展正字法》（Developing Orthographies for Unwritten Languages），（Cahill & Rice 2014）
http://cherokeepreservation.org/what-we-do/cultural-preservation/cherokee-language/
https://language.cherokee.org/
http://amahmutsun.org/language
https://rising.globalvoices.org/blog/2011/11/29/languages-online-activism-to-save-chakma-language/
https://www.languageconservancy.org/programs/indigenous-language-program-support/

7.2. 语言登记册

https://www.ethnologue.com/
https://glottolog.org/

7.3. Unicode及字体编码

https://unicode.org/main.html
https://unicode.org/standard/supported.html
https://unicode.org/standard/where/
https://unicode.org/pending/proposals.htm
https://unicode.org/glossary/
https://linguistics.berkeley.edu/sei/

专用字体可以通过一些非常规的方法来处理，如使用私人使用区域（PUA）或其他字符编码标准，如ASCII或用定制的符号作为码点的阿拉伯语。这个过程称为字体编码。使用这种非常规方式编码后，用户可以看到自己想输入的字符，但其他人如果没有安装同一种字体，看到的就不是同一种字符。因为在线服务及工具的底层编码中没有包含对于该字符的解码信息，或者包含解码信息，但却和原始解码信息有所不同，因而无法正确解析文本。

相比起用字体来对文字进行编码，用Unicode对文本编码更有优势，但在该语言收录进Unicode之前，可能还是需要先使用专用字体对文字进行编码。这种情况下，字体应使用

Unicode私人使用区中的码点进行编码，而不应重复利用那些已经明确分配给其他字符的码点。这样代码的使用就不会混乱，给用户理解语言造成困扰，人们使用现有的字符时也不会受到影响。使用私人使用区域（PUA）进行编码的字体在编码时保持了较高程度的一致性，所以当一门语言的字符完成标准化后，如果之前使用的是私人使用区域（PUA），在将其转换为Unicode字体时就更加容易。

7.4. 语言代码

https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

7.5. 字体

https://www.google.com/get/noto/

设计字体所需工具包括：

- FontForge
- FontLab
- Glyphs
- BirdFont（https://birdfont.org/）

雷丁大学（The University of Reading）开设有字体设计硕士专业（typefacedesign.net/），那里的学生或许可以帮助创建新的Unicode字体。

商业字体开发人员也可以创建新的Unicode字体。

8. 备注

本指南只涉及了语言文字，但交流还可以有其他载体，比如：

- Emoji（表情符号）
- 口头语言
- 视觉与手势结合的语言（手语）
- 视觉语言

有了自己文字的语言也可以用其他文字书写，例如：

- 土耳其语原来使用的是阿拉伯字母，但现在使用的是拉丁字母；
- 中文也可以用拉丁字母书写（拼音）。

本指南也不涉及方言，但在提交语言标准时也有必要考虑方言。

想得到高质量的语言检索结果，需要做到以下几点：

- （从文本中）识别语言
- 文本分割：将文本以词语为单位进行分割
- 一门语言中词语可能有不同的变体，因此需要明确词语的*词干*，比如*housing*和*houses*的词干就是*house*。

本指南不涉及口头语言，但以下资源可做参考：

- Google地球（https://docs.google.com/forms/d/e/1FAIpQLSdphaDaz33syPoUDyTOTwwkaLWZx90zopUklha4uadfkUKG8A/viewform）
- https://www.blog.google/products/earth/indigenous-speakers-share-their-languages-google-earth/
- https://www.gerlingo.com/
- XTrans（https://www.ldc.upenn.edu/language-resources/tools/xtrans）是一个支持多平台，多语言，多渠道的先进转录工具，辅助人工转录过程，并可以对音频添加注释。