Indigenous Languages:

Zero To Digital

Bringing Your Language Online

TRANSLATION COMMONS

2022-2032 | INTERNATIONAL DECADE OF Indigenous Languages

1. INTRODUCTION

This document is one in a series of guidelines entitled *Zero to Digital* addressing language digitization practices, authored by Translation Commons. Readers can access the entire series at https://translationcommons.org/resources/

This document is an updated version of *A Guide to Bring Your Language Online*. It provides an overview of the steps to make it possible to read and write your language on digital devices. A digital device is any computer-based instrument. Examples include mobile phones, computer laptops and desktops, and even televisions, gaming devices, and more.

This process is called “language digitization”. This document will help you understand the requirements and the process for enabling your language on digital systems and introduce you to further resources and tools that can assist you.

In addition to using your language on digital devices, language digitization will also make it possible to share content in your language across the internet. Typical applications include text messaging, email, web sites, social media, and more. Key benefits of language digitization include fostering communication within local communities and their global diaspora, facilitating education, preservation, and revitalization of the language and culture. Language digitization can also improve participation with organizations that shape the policies and economies that affect the lives of community members and make efforts to address the community’s unique needs and perspectives more successful.

The intended audiences for this guideline include:

- Indigenous and minority communities seeking to make their language accessible on mobile devices and computers
- Technologists aiding in language digitization
- Organizations wanting to sustain language communities

Of course, to work on adding a language to digital systems, it will be useful to have both linguistic and technology skills. This guideline provides basic information that can help anyone understand the process. They can then involve others with the relevant skills to assist with the more detailed steps as needed.

To achieve this goal, we begin by discussing the basic elements of reading and writing. These elements also form a basis for digital systems to support text writing. After these high level descriptions, we step through the details of enabling a language on a digital system.

1.1 NON-WRITTEN LANGUAGES

Non-written languages, that is, languages that are used orally or are signed and have no established writing system, are beyond the scope of this document. However, you can still use your language on digital systems with audio and video resources. Guidance and tools are available for these languages from organizations and academic linguistics departments. There are also resources at Translation Commons.

2. LANGUAGE DIGITIZATION CONCEPTS

2.1 WRITING YOUR LANGUAGE

Before we get into the technology for digitization, consider the steps to teach someone to read and write a language. After all, language digitization relies on using written text, so these steps can provide insight into the kinds of information needed by a digital system to support your language.

Your first step is to list the letters (or ideograms) used in writing your language. It starts with letters, but eventually it will include numbers, punctuation, and other symbols important to the culture. We will use “letters” here to include all of these symbols.

As the letters are recognized by learners, it is important to teach the proper way to draw them. This can include the correct order of the strokes, and the specific shapes that are used and are important to distinguish letters from each other. Some creative people may want to vary the look of the letters. However, it is important that the writing be legible to all readers, so a standard form for each letter must be followed.

Continuing the analogy of teaching a person the written form of a language, this will include terminology, spelling, and grammar rules. Establishing orthography and writing conventions provides consistency and reduces ambiguity.

An important term is the name for the language itself. It may be used as the first step in communicating with outsiders to agree on the language to be used between them.

The teaching process is iterative, continually expanding the learner's vocabulary, phrasing, and expressiveness.

2.2 TEACHING VS. DIGITIZING

Digitizing a language follows a series of steps analogous to teaching a person to read and write. However, now consider that you will be describing your language and its requirements to technologists, linguists and others that can help you implement the language in digital systems.

The process begins with a list of all the symbols used in writing the language. Digital systems use the term “character” for each of these items, and the list is called a “character set”. Digital systems will need to support these characters, to input and display them, and to process them according to rules of the language. Sometimes a single character set is used by multiple languages, and so digital systems may already support your language’s character set. Examples include languages written in Latin, Cyrillic, Chinese, Devanagari, or one of the many other widely supported writing systems. More characters and character sets are added all the time to digital systems. If your character set is not already supported, then yours can be added too.

Analogous to writing letters by hand, digital systems require a way to enter and display each character. Digital systems will use keyboards to enter characters and use fonts to draw them.

In many languages, characters take on different shapes or positions depending on the surrounding characters or other context. The layout of text is also determined by various conventions such as those for hyphenation, word breaking, justification, and line breaking. For a digital system to handle this correctly, it needs this kind of language information. Often this is achieved by examining word lists and text examples, as well as explicit descriptions of spelling and grammar rules.

Digital systems also require terminology that is specific to their technology. Examples include “file”, “menu”, “keyboard”, “font”, etc. So it will be necessary, if these terms do not already exist in your language, to borrow or coin new terms.

We mentioned that it is important for a writing system to have rules for consistency and standardization for widespread mutual intelligibility. This also applies to languages used on digital systems, as we will explain.

As you may know, digital systems represent all information as numeric values. Text, images, videos, etc. are all encoded and processed as numeric values. For example, every character, in every language that is supported on digital systems, has been assigned a unique numeric value. When you type a character on a keyboard, the relevant numeric value is generated and used to represent the character in applications and when communicating with other devices. That same value is used by digital fonts to know which shape to draw.

It is important that all digital systems agree on the numeric values representing each character. This standardizes the processing of text. This standardization is also necessary to exchange text between apps, systems, and devices worldwide. Assigning numeric values to characters is called character encoding, and the universal agreement on these assignments is provided through the Unicode Standard.

Earlier, we mentioned the importance of the name of the language. Digital systems also use a standardized name for the language as a way to identify the correct keyboard, font, spelling, and grammar rules to use with text. It is also used by search engines to select content on the internet in the chosen language. This name that digital systems use to identify a language is called a “language tag”. To be efficient for digital systems and to be useful across the multilingual world, there are standards that define language tags. It will be important to have a standardized identifier for your language.

The background above is intended to be suggestive of the information and activities that are needed to establish your language on digital systems. In particular, you will use information about your language to ensure that digital systems support:

- All necessary characters
- Unicode encoding of those characters
- Keyboard to input characters
- Font to display characters
- Digital system and other terminology in your language
- Language tag to identify content in your language
- Rules to assist with spelling, grammar, and other aspects of text generation and correction.

In the following sections, we will provide more concrete steps to have digital systems support your language. The process you use can and likely will be iterative, and the individual steps can be developed in parallel and in the sequence that leverages the skills and resources that are available to you. As this guideline is intended to be an overview of the process, we will also recommend our other more detailed guidelines and other materials that will assist you.

3. COMMUNITY ENGAGEMENT

3.1 PERMISSIONS AND RIGHTS

An important first step is to ensure that the language community indeed wants their language digitized. Some communities have objections to the sharing of their culture, with concerns for the potential for erosion of their language, misrepresentation of their values, derogatory remarks, and cultural appropriation. There may also be concerns for how the information may be used by the organizations that shape policies and economies affecting their communities. Some concerns may be mitigated by the placement of safeguards around the sharing of their language and culture with outsiders.

In addition, in collecting language examples and other data, some personal or private data may be inadvertently included, which the authors, subjects, or owners may not

want published. It is important to secure appropriate permissions before sharing this data. For more information on this, see the *Language Data Gathering Guideline*.

If the community agrees, make the materials open-source to invite further development, as well as available for online use, to increase access.

3.2 ESTABLISHING LANGUAGE CONVENTIONS

Oral and written communications are possible even without defined orthography and grammar rules. However, non-standard spelling, grammar, vocabulary, or significant regional variations can hinder mutual intelligibility, content sharing, and search retrieval. In such cases, support on digital systems for many language-based features (spelling and grammar correction, search, predictive text, etc.) will be limited.

If this is the case for your language, your community can consider developing conventions for spelling, punctuation, and grammar, as well as dictionaries and teaching materials. Local educators, policymakers, community leaders, and linguists can help establish a consistent written form.

This standardization can improve written language use and community engagement. It is also important to digital systems, as it enables them to offer features such as correcting spelling and grammar, proposing words, and delivering more relevant search results.

To further increase community engagement, invite discussion of ways to make language support materials more widely available and useful. For example, one approach is to create online resources for immersion and primary education. Another is to share books, and other content with the communities that use the language.

4. PREPARATION

It will be useful to collect and bring together information about the language and its writing system, as a first step. Having documentation and examples will accelerate the process and provide answers to the many detailed questions that will be raised.

Information that will be useful includes the complete set of characters, and any formal documentation such as dictionaries, grammar books, and other linguistic or educational materials.

Text examples can be published information, but they can also be private letters, notes, pictures of signs on buildings, and image scans of handwriting. It is important for the examples to be as diverse as possible. People speak and write differently according to their age, gender, and role, and when addressing others that are of the same or different age, gender, seniority, etc. or addressing an individual, a small group or a large

gathering. Language usage may also be dependent on whether the discussion is informal, formal, part of a celebration or ritual, etc.

For more detailed instructions, see the *Language Data Gathering Guidelines*.

4.1 STANDARDS ORGANIZATIONS

The easiest and shortest path to have your language supported by digital systems is to ensure it is included in the relevant language and technology standards. Digital systems often require languages to be adopted by these standards before they will implement them.

There are several reasons for this. Many of these standards are often designed for efficiency on digital systems. For example, the standards are designed to be multilingual and work consistently across languages. They often are co-dependent, where one standard relies on definitions in other standards. Most standards also ensure a degree of stability, so that when key elements are defined, they will not change in a way that breaks existing systems or prevents use across different systems.

Therefore, it will be useful for you to become familiar with these standards organizations as part of your preparation. Their roles will be discussed in more detail in the implementation sections. However, reviewing their web sites or Wikipedia pages will help you understand their roles in creating standards used in digitization.

THE INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

The International Organization for Standardization (ISO) is composed of representatives from the national standards organizations of member countries to create international standards. In addition to reviewing the information in each ISO standard, it may be useful to connect with your national representative to the organization or to the technical committees.

- ISO 639 is a standard that lists languages and assigns them identifiers.
- A collection of letters and other symbols used to represent textual information in one or more writing systems is called a **script**. ISO 15924 lists scripts and assigns them identifiers.
- ISO 3166 lists countries and regions and assigns them identifiers.
- The ISO Technical Committee TC37 addresses language and terminology topics.

THE UNICODE CONSORTIUM

The Unicode Consortium has multiple roles. It is the organization that standardizes characters and assigns each a unique number, a process called “character encoding”.

Each character is also assigned numerous property values that digital systems reference to perform sorting, word and line breaking and many other functions needed for language processing. All of this information is contained in the Unicode Standard.

The Unicode Consortium is the current registration authority for the ISO 15924 standard, which assigns 4-letter codes to each script. The script names and their assigned codes are listed on the Unicode Consortium website.

The Unicode Consortium is also responsible for the Common Locale Data Repository (CLDR). CLDR contains many different types of data for each language. Digital systems use this data to have correct terminology, formatting, and presentation of information to users in their language. For example, this data includes the native names of days, months, countries, etc. and formats for dates, times, numbers, etc. and much more.

The Unicode Consortium also has a CLDR Keyboard Working Group that is maintaining an XML format for keyboard layouts. Keyboard layouts in this format are added to and maintained in the CLDR repository.

5. IMPLEMENTING LANGUAGE DIGITIZATION

This section will describe the steps to achieve digitization of your language for each of the technologies that are required. To communicate clearly with technologists and linguists that can assist you, we will introduce you to a few technical terms that you need to know.

There is also a pattern you will recognize. For each technology, you will first need to determine if it is capable of supporting your language. If your language is similar to other languages, it is possible the support already exists. Then, if your language is not enabled on a digital system, you can define the steps that are needed specific to your language and arrange to gain that support.

5.1 SCRIPT

As mentioned above, a script is a collection of letters and other symbols used to represent textual information in one or more writing systems. A single script can often be used to write many different languages and can include more characters than those used by your language. For example, the Latin script is used to write French, Swahili, and Finnish. This is why you may find your script is already supported.

It is also important to note that some languages can be written in more than one script. Each script is separate and requires its own standardization and implementation. Decide which script you intend to use on a digital system. Alternatively, you can use more than one script, but you will need to repeat the digitization process for each script.

Once you have identified all of the characters (including letters, digits, punctuation and other symbols) used in your language, you will want to evaluate if the script is included in the ISO and Unicode standards.

If you know the name of the script associated with your language, you can verify if that script is in the ISO 15924 standard.

If you don't know the name of the script, or if it isn't on the list, the Unicode Consortium has all of the characters in its Unicode Standard depicted in code charts on its website. There are more than 150,000 characters in the standard. You, or someone familiar with Unicode, can see if the characters in your set are already in the standard. The Unicode Consortium has a Where is my character? page that can help you search.

If the characters are not in the standard, then you will need to apply to have your script added to ISO 15924. Generally, you will first need to propose the addition of your characters to the Unicode Standard, and have them accepted.

5.2 UNICODE CHARACTER ENCODING

If the characters needed to write your language are not in the Unicode Standard then you will need to submit a proposal to the Unicode Consortium to add them. This is a lengthy and technical process. To accelerate the process and increase its chance of success, proposers should consider contacting organizations such as Translation Commons or the Script Encoding Initiative (sei@berkeley.edu) for assistance.

When the characters of your language are added to the Unicode Standard, they will each have been assigned a unique numeric value, called a ***code point***.

It is important to note, that a character, although thought of as a whole character by users, may be treated as consisting of several Unicode characters combined together. Take for example letters that are marked with accents or other diacritics, or tone marks. The Unicode Standard will have encoded base letters, and separately diacritic and tone marks, which can be combined with any base letter. These are called combining or non-spacing characters. In some cases, for legacy or other reasons, Unicode may have also encoded letters with their marks as whole or "***precomposed***" Unicode characters.

5.3 KEYBOARD

Keyboards are a primary method to enter text into digital systems. Keyboards can be physical devices or they can be on-screen software applications. Pressing the physical keys or selecting on-screen keys enters characters into the current application. Sequences or combinations of key selections can input additional characters, so that languages with more characters than there are available keys can be accommodated.

Digital systems often come with support for many keyboards, enabling input for different languages. Many devices support more than one keyboard simultaneously and you can switch back and forth between them, if you are multilingual. Additional keyboards can be installed if they are not provided by default.

For each digital system, you will want to research whether a keyboard for your language is available. The device manufacturer usually lists the available keyboards on their website as well as instructions for enabling or installing keyboards. In addition to the manufacturers, there are many third parties that offer keyboards as well as tools for creating custom keyboards.

To determine if keyboards exist for your language check manufacturer and third party websites and app stores (e.g. Apple App Store, Google Play, etc.):

- Gboard, the Google Keyboard for mobile
- Google Input Tools, virtual keyboards for Chrome on computers
- Keyman: a tool for creating and providing keyboards
- Microsoft Keyboard Layout Creator (MSKLC)
- SIL Ukelele a tool for creating custom keyboard for the Mac

If you intend to develop a keyboard for your language it may be helpful to first look at existing keyboard layouts published in the CLDR Keyboards Repository or in the Keyman Repository. If you want assistance with keyboard development, see the Translation Commons Keyboard Request Form.

Caution: Installing a keyboard from the web may pose security and privacy risks.

5.4 FONT

Fonts and software called rendering engines are the tools digital systems use to draw text. The software uses the font's data associated with each character's code point to display and position each character's image, called a ***glyph***. This can be a complex process as the shape and placement of each character can depend on the surrounding context. It is sometimes the case that this process is language-dependent. So a script can support multiple languages, but there might be slight differences in the rules for drawing glyphs specific to a language. You may need to have a font specific to your language, although often if a font supports your script, it will be usable for your language.

You will need to research whether there is a font that supports your script. If your characters are encoded in Unicode, then you will need a font that is Unicode based.

Some fonts require older or special purpose encodings and will not work with a Unicode character encoding.

Another consideration is that fonts can be specific to operating systems or devices. So, you may need separate fonts for Windows operating systems, Android and Apple phones, or other devices.

Some applications, for example word processors, may require configuration to enable the font.

Note that many mobile devices do not support font installation, while desktops and laptops allow it. If a suitable font is not available for mobile devices, you may need to urge the device vendors to make the required fonts available.

Web fonts allow websites to present text with a font provided by that website. This enables text display on web pages when the desired font is not available on the user's digital system. However, a web font only works on the web site that provides it and does not permanently install the font on the digital system.

Caution: Installing a web font on a mobile device may pose security and privacy risks and could void device warranties or support agreements. Some organizations block web fonts for security and privacy, which will limit or prevent using associated web pages on the organization's systems.

The Google Noto fonts are a collection of free Unicode-based fonts that support thousands of languages. This can be a good source to look for a font for your language. Also, many Noto fonts are available as web fonts.

There are many websites that offer fonts that can be downloaded, and you may find a Unicode font that supports your language. If not, you may need to reach out to font designers to create a font for your language. There are several organizations that may be of assistance including Translation Commons. To design a suitable font, you may be asked to provide information about all the required character shapes, and text presentation details such as joining and ligatures. It can also help to provide sample text that can be used for testing.

5.5 TERMINOLOGY

Digital systems, software applications, and the internet have numerous elements that you will want to refer to in your own language. However, as technology that is new to your language community, the terms may not exist, or literal translations may not be culturally appropriate. Translations to your language for terms such as file, menu, e-mail, web page, home page, cloud, keyboard, font, print, download, and many more may need to be agreed upon within your community.

Creating word lists, dictionaries, and bilingual dictionaries will accelerate adoption by digital systems, software applications, as well as users. Be sure to include everyday terms, as well as terms from various disciplines, in addition to the required user interface and technology terms. Documenting term frequency data, if available, is also useful.

The more information that you document, the better that spelling correction, predictive text, search, and other linguistic capabilities will be.

For recommendations on defining terms in your language, see *Terminology Guidelines*.

5.6 TEXT LAYOUT RULES

Writing involves more than entering and displaying text. Digital systems and software applications can layout text, format data (numbers, dates, etc.), and provide grammatical assistance, in conformance with the requirements of the language, as long as this information is documented.

For example, languages have rules for text segmentation, word breaking, hyphenation, justification, etc. Some orthographies lack explicit word boundaries, and require dictionary data or algorithms for text layout. The rules for line breaking also vary by language.

Documentation of grammatical rules for punctuation, sorting, plurals, gender, etc. can also be valuable. This kind of information can be added to the Common Locale Data Repository which will accelerate its adoption by digital systems.

5.7 LANGUAGE TAG

Digital systems use "**language tags**" to designate a language. They are used to declare the language used within files, documents, media, etc. They are also used in search queries to prescribe the language(s) that results should ideally contain.

There are standards that define the structure and values used in language tags. Digital systems rely on IETF BCP-47.

To have great precision, language tags can have multiple parts, called subtags. The primary subtag identifies the language. An optional subtag identifies the country or region. This is useful to distinguish dialects. Another optional subtag identifies the script. This is useful when a language can be written with more than one character set. There are languages, for example, that are written in Latin, Cyrillic, or Arabic character sets.

It will be important to have a standard BCP-47 tag for your language. The primary language tag will be based on ISO-639 codes. The region and script subtags will be

based on ISO 3166 and ISO 15924 codes respectively. Apply to the appropriate standards body if a needed code does not exist.

Once you have a tag for your language, you can associate it with content you have or produce. In multi-language texts, individual sections may be tagged with their language. In web documents, use the HTML lang attribute to declare the language of content.

5.8 COMMON LOCALE DATA REPOSITORY (CLDR)

The Unicode Common Locale Data Repository (CLDR) provides essential resources for software to support multiple languages, each with their unique conventions. Some of the data in CLDR that digital systems rely on include:

- Locale-specific formats and conventions for dates, times, calendars, numbers, currency, measurement units, and timezones.
- Translations for names of languages, scripts, countries, currencies, eras, time zones, cities, and emoji characters.
- Language and grammar details, including pluralization, gender, and sorting rules, writing direction, transliteration, number spelling, and text segmentation.
- Character set and keyboard layout data.

6. GAINING INDUSTRY ADOPTION

As you achieve a significant level of both language documentation and development of the technology basics (keyboard, font, language tag, locale data entered into CLDR), you will want to reach out to device manufacturers, and operating system, web platform, and software application developers to encourage them to add your language to their distribution. Without their support, your language community can still use the technology you implemented, but it can require extra steps for installation, and perhaps have limits on functionality.

Besides reaching out for support, the community can also consider developing their own applications specific to community needs. For example, educational software, or community messaging, calendars, and social media.

6.1 ADVANCED LANGUAGE TECHNOLOGY

The following technologies can be helpful to capture data for language digitization or to create and share content in your language. Developing these technologies can require sizable quantities of language data or otherwise be difficult to implement for your language.

Speech to Text: Converts spoken words into text, useful for transcribing documents or giving commands to applications.

Text to Speech: Generates natural speech from text, aiding hands-free interfaces and machine reading.

Transcription of Academic Audio Media: Important for language documentation. There are some open-source projects addressing this. For example: Accelerated Transcription for Linguists.

Machine Translation: Automates translation of text. There are organizations such as Apertium that provide free, open-source software enabling communities to develop machine translation for their language.

Optical Character Recognition (OCR): Digitizes text from images of handwriting or print. There are open-source projects available for various writing systems. A language model with frequently used words improves OCR accuracy.

7. NEXT STEPS

This document outlined the basic steps for digitizing your language and bringing it online. For more detailed instructions, visit the Translation Common website. There you will find additional guidelines in the *Zero to Digital* series. There are also many other documents and resources. Translation Commons also assists Indigenous and minority communities with language digitization. The website has forms to request assistance with keyboard development and other digitization steps.

8. ACKNOWLEDGEMENTS

This guideline is developed through Translation Commons' Language Digitization Initiative—an effort aligned with UNESCO's International Decade of Indigenous Languages. These resources aim to empower Indigenous and minority language communities to bring their languages online.

Translation Commons thanks its many volunteers that contributed their expertise and significant time to produce this guideline. Version 2.0 builds upon the work of prior versions.

Version 2.0 - 2025

Lead Authors: Tex Texin, Julie Anderson

Primary Editor: Deborah Anderson

Reviewing Editor: Rachel Moilliet

Version 1.0 - 2019

Lead Authors: Deborah W. Anderson, Lee Collins, Craig Cornelius, Craig Cummings

Design and Marketing: Mette Attar, Leonidas Pappas