網路上語言使用的狀態報告

摘要報告 · 2022

網路上語言使用的狀態報告

摘要報告

為何需要這份報告？我們是誰?

字典顯示，語言是有結構的溝通方式，主要應用於人類之間。但不僅止於此，語言是我們與彼此分享的最基本傳統，多數時候由祖先留傳下來，幸運的話，我們也能傳承給後代。我們為自己與彼此，使用語言思考、說話、聆聽與想像……。語言便是是我們在這世上的身分及狀態的核心，語言有助我們說故事，分享我們對自己與彼此的認識。你說的是什麼語言？會用那個語言做夢嗎？你思考時使用的語言，與工作時不同嗎？你喜歡的音樂，是否使用你不一定能理解的語言？

每種語言都是在這個世界存在、處事與溝通的系統，更重要的是認識與想像的系統。每一種語言本身都是知識系統：我們的語言是我們理解自身世界並向他人解釋的基本方法。我們的語言 可以是口述、書面、視覺、手語及手勢，或是透過哨音或鼓聲。無論什麼形式，**語言便是知識的代理媒介**。換句話說，語言是表達我們所思所知最直接的方式。

再想想你用以說話、思考、作夢或寫作的語言，其中有多少是你能夠在數位空間裡用以完整分享與溝通的語言？你在網路上使用語言的經驗如何？你使用的硬體上有沒有你的語言的字元？你是否需要將鍵盤改造才能適用自己的語言？當你用搜尋引擎找資料時，搜尋結果是否以你希望的語言呈現？你是否得要學習不同於自己原本的語言，才能使用網路並在網上貢獻內容？如果以上有任何一個或數個答案為「否」，那麼你就是世上為數不多、能夠輕易以自己的語言使用網路的優勢族群。而且你的語言很可能是……英語。

網路及其各種數位空間，提供了現今最重要的知識、溝通與行動架構。然而，世界上有超過 7000 種口說語言（以及手語、點字等其他語言），**其中有多少語言我們能真正在網路上完整使用？真正的多元語言網路環境，看起來、感覺起來、聽起來又會是什麼樣貌呢?**

這份報告算是我們設法回答這個問題的其中一種方式。我們由以下三個組織合作：誰的知識？（Whose Knowledge?）、牛津大學網路研究所（Oxford Internet Institute）以及網路與社會中心（印度）（The Centre for Internet and Society (India)）。我們合作提供不同見解與經驗，分析網路上的語言使用狀態，並與其他同樣關心這些議題的人合作，希望能夠創造更多元語言的網路環境，創造數位科技與實踐。

這份報告期望做到以下三件事：

- **勾勒網路上的語言使用現況**：我們想嘗試了解，網路上目前有哪些語言以及語言重現的形式。透過量化資料（檢視不同數位平台、工具與空間的數據），以及質化資料（從人們自己對網路上語言的故事與經驗學習）進行。
- **提升意識，了解建立更多元語言網路環境的挑戰與機會**：創造並管理世上語言的科技、內容、社群的挑戰相當巨大，但也充滿刺激的可能性與機會。這份報告將會列出其中的挑戰與可能性。
- **提出行動議程：** 經由這些見解與意識，我們提出自己以及許多在世界上同樣在研究這些議題的其他人，為了確保更多元語言的網路環境，想要規劃並採取的行動。

這份報告代表什麼，又不代表什麼?

這份報告是半成品，或者該說仍在進行中。

有許多不同的個體、社群與機構長期從語言的諸多面向耕耘，近年來，更是致力於線上語言的諸多面向。我們深受啟發，但這份報告並非要完整全面地囊括他們所有人以及所有成果。此外，我們也不認識所有研究語言及網路的人，不過我們嘗試納入大多數我們已知且或多或少深受啟發的研究，並且列在我們的資源及致謝章節。

我們受限於所能夠收集到的資料，也在數據章節中討論了其中一些限制。歡迎任何指教與建議，讓我們精進並更新報告中所提供的資訊，也期待收到已經在研究相關議題者的來信，並希望未來能納入報告的更新版本。

我們盡可能以最容易理解的方式來撰寫這份報告，希望多個世代與社群的人能加入我們的研究，因此不願讓專業術語或「學術」語言成為閱讀與反思的障礙。我們也希望能盡可能翻譯成不同語言版本（譯者們：請與我們聯繫！），儘管這份報告有部分是先以英文撰寫，我們仍希望英文不會成為反思或行動的障礙。

以過去的諸多成果為基礎，我們希望這份報告能成為未來研究、討論並針對這類議題採取行動的「基準」。

我們是誰，為什麼要一起來做這份報告?

這份報告研究集合了三個組織的力量：網路與社會中心、牛津大學網路研究所、誰的知識？。我們都從不同的研究、政策與倡議面，對網路與數位科技的潛在影響產生興趣。

過去兩年，我們以自己的方式努力去了解網路上的知識不平等與不公平：線上內容的貢獻者是誰，貢獻方式又如何？我們很快便意識到，網路上關於不同語言版本知識的資料沒有多少。接著我們想要進一步了解這世界上的語言，此刻有多少存在網路上？網路的語言多元程度如何？我們僅能探索少數能使用公開資訊的領域，但仍希望對所有致力於多元語言網路環境的大家能有所貢獻。

簡要說明新冠肺炎與這份報告：這份報告在 2019 年新冠肺炎爆發前便開始動工，但大部分的分析、訪談與撰寫都在改變了我們個人與集體生活的全球疫情下進行。所有參與這份報告的人或多或少都受到影響，因此我們花了比預期更久的時間才能與世人分享成果。但新冠肺炎也提醒了我們彼此之間的相互關聯，能夠以不同語言表達複雜觀念對我們來說相當必要，擁有真正多元語言、充滿韌性與可及的（數位）基礎建設又是多麼重要。

如何閱讀這份報告?

這份報告採取我們所謂的「數位優先」模式，也就是透過網站可獲得最佳閱讀、聆聽及從中學習的經驗，因為這份報告有不同層次與層級。我們的報告匯聚了數據與故事。從統計觀點了解網路上的語言使用狀態，讓我們能概括地審視議題，有助於我們理解人們不同的體驗脈絡。而世界各地的人在不同情境下於網路上使用語言的體驗，有助我們更深入了解人們以自己的語言使用網路的難易度。透過故事與數據，我們得以開始面對潛在的脈絡、挑戰與機會。

這就是為什麼這份報告主要有三大層次：

- 網路上語言使用的狀態報告摘要內容，以及我們如何建立這份報告（也就是你此刻正在閱讀的內容！）
- 數據 分析某些我們每天使用的數位平台、應用程式及裝置上的幾個重要語言議題。這項工作由我們在牛津大學網路研究所的朋友主導，你能在這裡看到他們精彩的資料視覺化與分析。請注意，該分析僅限於我們能從公開資料集與素材取得的資料。其他方法論的限制會在文章中更詳細討論，但更重要的是：辨識語言很難找到單一一致的方式，要估計有多少人使用特定語言也同樣困難，特別是當這些語言和語言使用的方式都是隨著時間動態改變。
- 故事 讓我們更深入了解世界上不同的人與社群，怎麼用自己的語言體驗網路，以及現在要用他們自己的語言找到所需資訊有多困難。我們以書面及口說形式 邀稿這些故事，因此您會讀到文章，還有影音訪談。這項工作由我們在網路與社會中心的朋友主導，整合來自世界各地如此豐富交織的語言經驗貢獻，包括來自非洲、美洲與澳洲的清達里語（Chindali）、克里語（Cree）、奧吉布韋語（Ojibway）、馬普切語（Mapuzugun）、薩波特克語（Zapotec）以及亞倫特語（Arrernte）等原住民語言；歐洲的布列塔尼語（Breton）、巴斯克語（Basque）、薩丁尼亞語（Sardinian）以及卡累利阿語（Karelian）等少數語言；還有亞洲的孟加拉語（Bengali）、印尼語（Bahasa）以及僧伽羅語（Sinhala）等在區域及全球強勢的語言，以及全北非不同形式的阿拉伯語。

更重要的是，我們的貢獻者以自己的語言及英語書寫或口說，我們的摘要也以不同語言撰寫及口說錄音，因此希望您能以一種以上的語言享受閱讀及聆聽!

我們盡可能以視覺形式呈現這些貢獻，透過想像力豐富的插畫與動畫，謹慎拼出語言的不同社會與技術面向。同樣地，報告裡的其他資訊也是由貢獻者與插畫師討論並合作發展的成果。

網路的多元語言程度如何?

網路的多元語言程度還不如（且悲傷的是遠不及）我們的現實生活。我們透過研究世界上某些人的數據與故事來嘗試了解原因。這裡我們僅能簡單摘要並分析來自世界各地貢獻者的深度與豐富研究，想知道更多細節與啟發請務必閱讀他們的文章。

首先，我們檢視世界各地的人以不同語言使用網路的脈絡。我們探討資訊與知識在不同語言與地理環境下散佈（與否）的方式。接著進一步探討我們用來在網路上創造內容、溝通與分享資訊的主要平台與應用程式，以及這些平台與應用程式各自支援多少語言。我們仔細研究 Google 地圖與維基百科，日常生活中會使用的兩大多元語言內容空間，並探討兩個平台在不同語言下的運作方式。

過程中，我們會分享人們在網路上以自己的語言取得並貢獻知識的故事與經驗。正如我們所學到的，多數的貢獻者發現自己必須從第一語言切換為另一種語言，才能取得並貢獻與他們所重視議題有關的資訊。

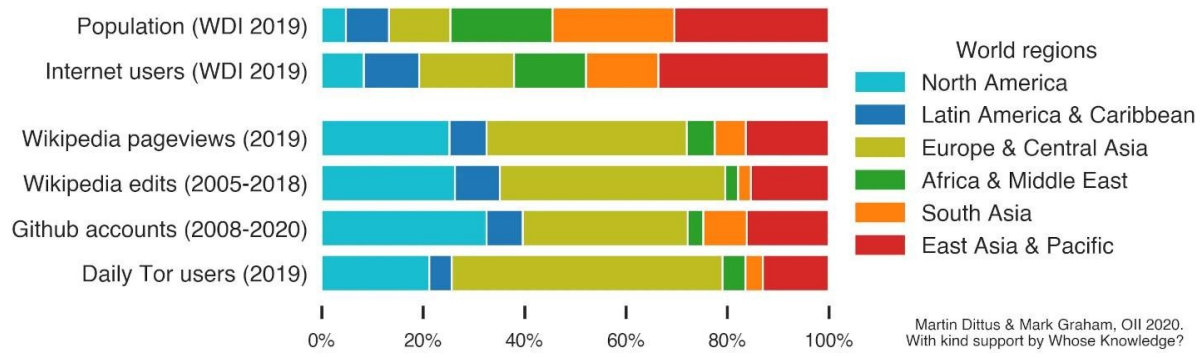
語言脈絡：地理與數位知識不平等

« 有口述傳統的語言不適合我們今日的網路世界。»
Ana Alonso

« 感覺上，這些平台普遍來說都延續殖民者的想法，認為某些語言有更高的溝通價值與能力，這對像馬普切語（Mapuzugun）這類少數語言來說是很不利的觀點。»
Kimeltuwe project

我們估計全世界 60% 以上 已在數位上彼此連結，多數人透過手機與行動裝置連結。上網人口中，四分之三來自全球南方（Global South）：亞洲、非洲、拉丁美洲、加勒比海與太平洋群島。但是這樣的網路近用性有什麼意義，又有多公平呢？我們在網路創造並生產公開知識的程度是否與我們消費知識的程度相當？

在 Martin 與 Mark 對全球人口與網路使用者數量比較的普查中，我們發現相較於其他人口，某些人口可以用網路做更多有意義的事，這種情況即使在已經廣為人知的數位空間裡也是如此。比方說，即使我們多數上網人口來自全球南方，也無法以知識創造者及生產者的角色使用網路，僅能以消費者的身分使用。維基百科的大部分編輯紀錄、 GitHub（程式碼託管網站）的大多數帳號、以及 Tor（一種安全的瀏覽器）的多數使用者，都來自歐洲與北美州。

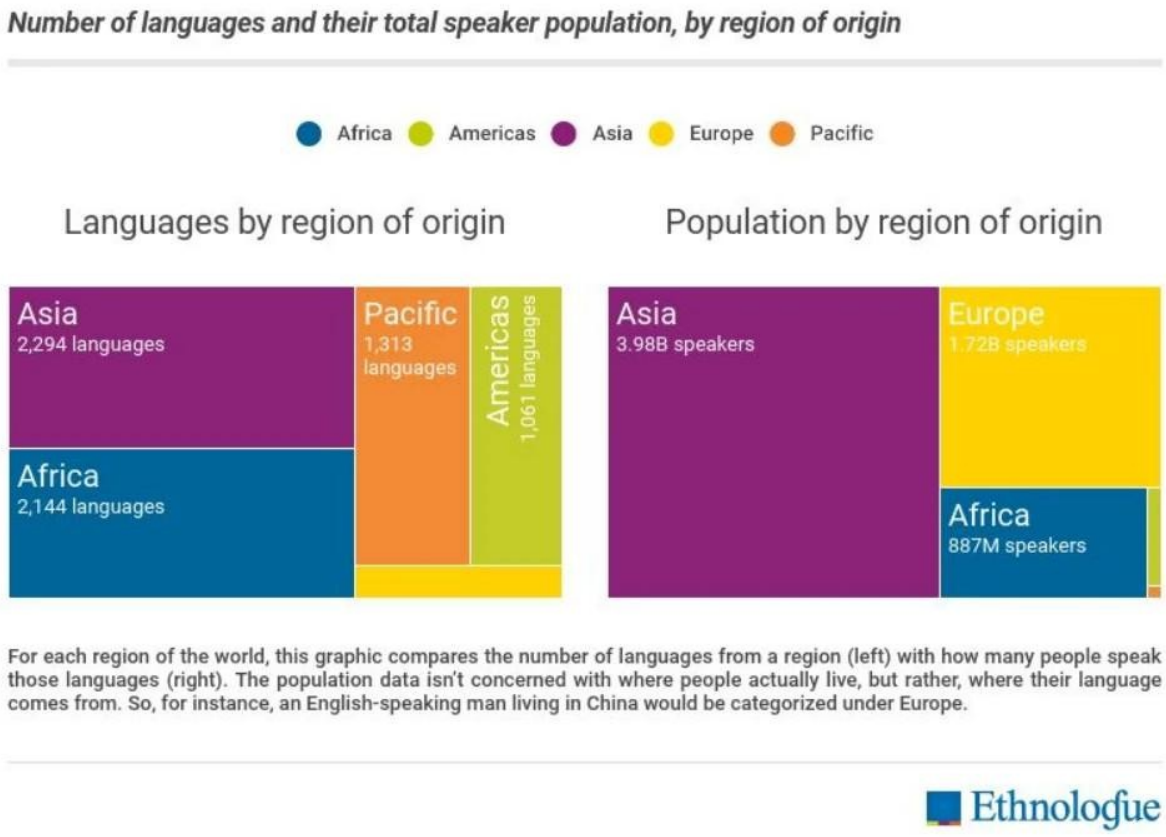


全球各地區數位參與方式。（來源：世界銀行 2019，Wikimedia 基金會 2019，維基百科 2018，Gitub 2020，Tor 2019）

這樣的近用性不均對語言來說代表什麼呢？我們是否都能夠以自己的語言使用網路？我們是否能夠以自己的語言創造內容與資訊？

正如其他預測所顯示，超過 75% 的網路使用者總共只以 10 種語言來上網 — 而這些語言多數都有歐洲殖民歷史（英語、法語、德語、葡萄牙語、西班牙語……），或是在特定地區當其他語言掙扎求生之際仍為強勢語言（漢語（Chinese）、阿拉伯語、俄羅斯語……）。2020 年，估計網路使用者總人口中有 25.9% 以英語上網，而 19.4% 的人以漢語使用網路。中國的上網人口為全球之冠，但要記得我們所謂的「漢語」指的不是單一語言，而是一個涵蓋許多不同語言的語系。

有趣的是，現今網路上使用的語言可能多源於歐洲，但相較於其他大陸，歐洲卻擁有世界最少的語言種類全球 7000 多種語言中，4000 種以上來自亞洲與非洲（各 2000 多種語言），太平洋群島及美洲則各 1000 多種語言。巴布亞紐幾內亞與印尼為擁有最多種語言的國家，巴國有超過 800 種，印尼則有超過 700 種語言。



全球各地區的語言數量及使用者總人口。來源: 民族語言網

根據以該語言為第一語言（或母語）的人數排名，許多南亞語言（印地語、孟加拉語、烏爾都語（Urdu）……）均排名全球前 10 大語言，卻都不是南亞人能夠用來上網的語言。此外，正如我們從 Ishan 身上所學到的，他的第一語言是孟加拉語，即使能夠以所選擇的語言取得數位知識，所尋求的資訊卻可能不存在這個語言的網路知識庫裡；以 Ishan 的情況來說，就是找不到身心障礙與性別權利相關的內容。東南亞的狀況也很相似，這裡有某種程度上全球最高的網路使用量以及最大的語言多樣性。但 Paska 發現，在印尼語內容中沒有性別權利相關內容，就跟 Ishan 發現在孟加拉語網路資訊中找不到一樣。

我們還知道，世界上有 7000 多種語言，其中只有 4000 種有書寫系統或文字，其中也不是所有語言都同樣廣泛使用文字。我們相信全球至少有半數語言主要透過口述傳統傳承，沒有書面文本。即使是有書寫文字的語言，出版品也多偏向歐洲殖民語言，和（比較沒那麼嚴重地）偏向以區域強勢語言出版。2010 年時 Google 估計歷史上有約 1.3 億本書籍出版，其中絕大部分的出版品以約 480 種語言出版。多數高知名度的科學或社會科學學術期刊為英語期刊。世上翻譯成最多種語言的書為聖經（譯入 3000 多種語言），世上譯入最多種語言的文件是聯合國世界人權宣言（Universal Declaration of Human Rights）（譯入 500 多種語言）。

這有什麼關係呢？因為數位語言科技仰賴自動處理已出版的素材，以精進他們不同語言的支援與內容。因此，當世界各地出版的文本本身已偏向某些語言（而且無法納入任何非書面語言）時，便會加深我們所體驗的語言不平等。那些在印刷出版業缺席，非文本為主的語言、以手語、聲音、手勢及動作為主要媒介的語言，因而也常常遭到數位語言科技所忽略。

比方說，如 Ana 所分享的：「網路的設計無法回應只有口述傳統語言的使用者。」在網路上書寫語言當道之際，很難找到來自口述與視覺語言傳統的內容。我們無法輕易搜尋如手勢、手語及哨音等內容。Joel 與 Caddie 在採訪中分享澳洲第一組原住民表情符號（emoji），創造於亞倫特人 （Arrernte）的領地班土哇（Mparntwe）／愛麗絲泉，並提到亞倫特語常結合身體手勢與口語來傳達意思。Emna 表示突尼西亞的狀況也相同，並提到國人所使用的不同語言：「說到語言保存，我們不該只注重書寫，我們也需要保存口述形式、手勢、手語、哨音等，這些都無法以書寫全面涵蓋。」

數位科技讓我們有機會重現以上提到的，基於文本、聲音、手勢等等的多重語言形式。數位科技同時也有助於我們保存並復振瀕危的語言：至少佔全世界所有語言的 40% 以上。每個月，會有兩個原住民語言及其所表達的知識死去，不復存在。

這些不同的語言脈絡為什麼沒能在網路上有更多重現呢?

Claudia 在她的文章裡提供三個面向，供我們了解語言與科技之間的關係：可用性、易用性，以及如何開發科技。正如我們在報告中通篇可見，Claudia 所謂的「多數語言」（我們也發現多為歐洲殖民語言或區域強勢語言）有廣泛的媒體、服務、介面與應用程式，包括鍵盤、機器翻譯或語音辨識等基礎建設，可用性遠高於其他語言。科技公司在這些多數語言的易用性上投入的時間與資源也相當多，因為這裡的可見利益最多。最後，她認為多數語言科技的發展過程要不是由上而下，就是很少和語言社群合作，就算偶爾嘗試和社群合作時，規劃及協調也都不太好。

這些關於內容的擔憂與挑戰也讓我們有方向能創造更多元的語言環境，我們稍後會回來探討這些可能性。

語言支援：平台與通訊應用程式

« 當你打出 “早安” 的英文，字還沒拼完，手機或電腦就已經跳出這個字的建議選項了。但我用清達里語要打出同樣意思的字（“mwalamusha”）時，卻必須完整輸入整個字，要花很多時間，而且打完畫面還會出現底線，因為電腦或手機不認識這個字。»

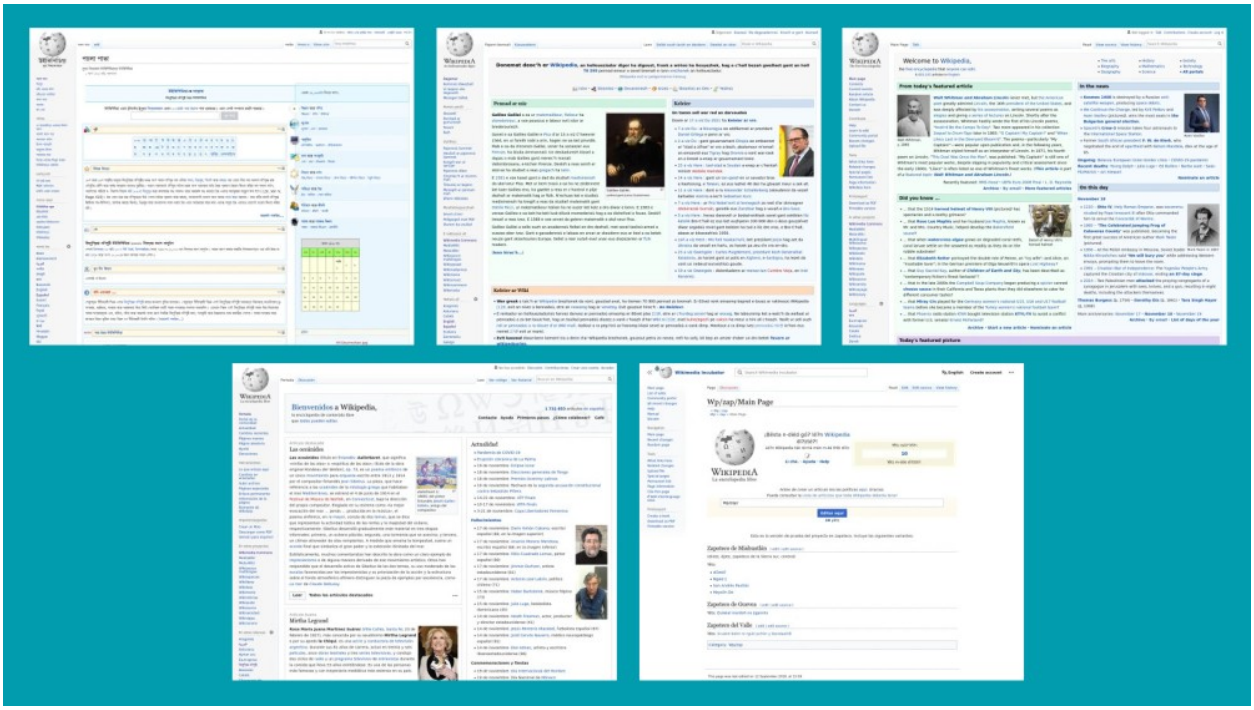
Donald Flywell Malanga

« 有僧伽羅語及坦米爾語字母的鍵盤很少見。我們父母會將僧伽羅字母縮小列印、裁減，然後黏在鍵帽上原本的英文字母旁邊。儘管已經開發出許多僧伽羅語字型，卻沒有一個像 Unicode （萬國碼）字型那麼通用。»

Uda Deshapriya

« 如果熱門應用程式與主要軟體介面不快點支援布列塔尼語（Breton），
在無法與法語應用程式競爭的情況下，這個語言最終會越來越無法吸引年輕世代。»
Claudia Soria

我們已深入了解網路上的語言仍不如我們真實世界那樣多元，並探討主要數位平台與應用程式提供了什麼樣的語言支援，例如透過不同語言的使用者介面，供我們以自己的語言溝通、創造並分享內容。

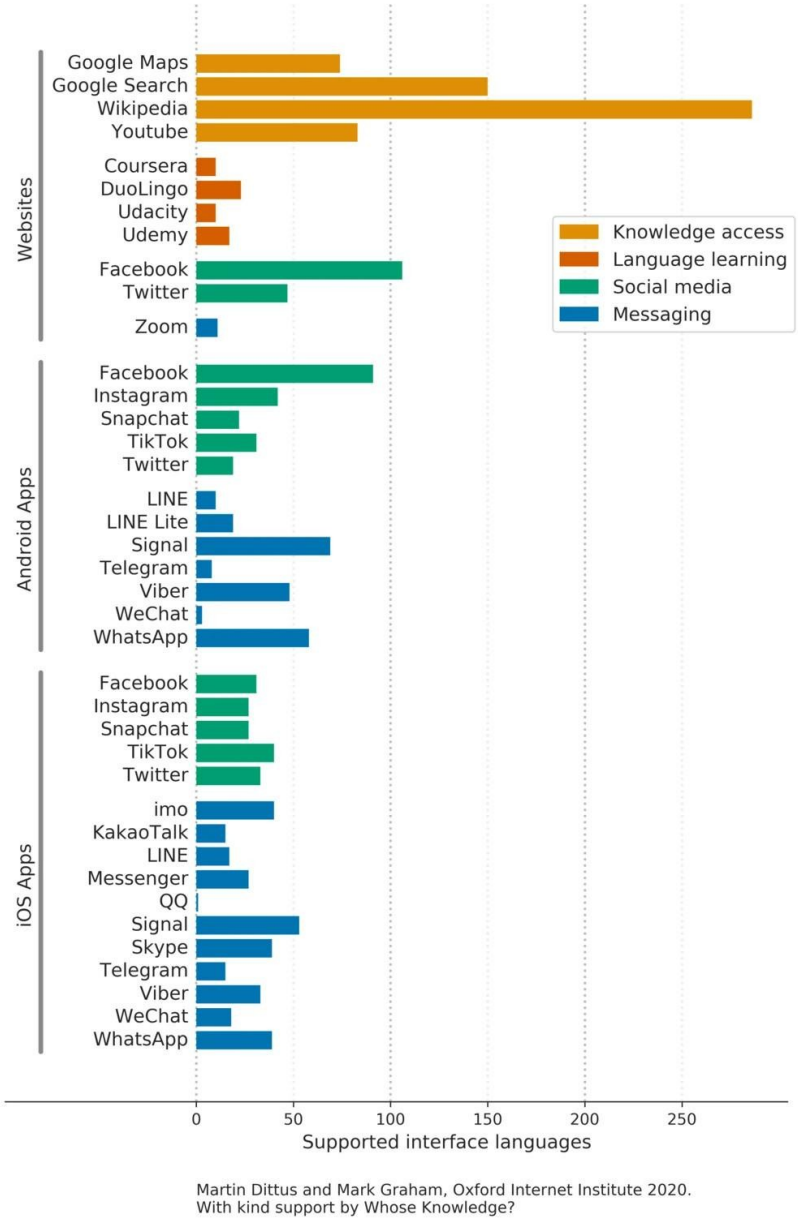


維基百科多種語言介面。

Martin 與 Mark 分析了 11 個網站、12 個 Android 應用程式以及 16 個 iOS 應用程式的語言支援。他們特別選擇專門用於收集、分享知識並受到廣泛使用的平台，尤其是那些努力布局全球、在各地都有使用者的平台。同樣地，他們仍得仰賴這些平台與應用程式可取得的公開資料。

這些平台分成四大類（已知有些會重疊）：

- **取得知識**（知識與資訊平台，包括搜尋引擎）：Google 地圖、Google 搜尋、維基百科、YouTube。
- **學習語言**（自主語言學習平台）：DuoLingo 和 Coursera 、Udacity、Udemy 等教育平台。
- **社群媒體**（面向公眾的社群媒體平台）：Facebook、 Instagram、Snapchat、TikTok、推特（Twitter）。
- **即時通訊**（私人及群組即時通訊）：imo 、 KakaoTalk 、 LINE 、 LINE Lite、Messenger、QQ、Signal、Skype、Telegram、Viber、微信、WhatsApp、Zoom。



依平台分類之各平台介面語言支援數量。

我們發現文本為主的語言支援在不同數位平台的分布極度不均。主要網路平台如維基百科、Google 搜尋及 Facebook，目前提供最多語言支援。有趣的是，維基百科（由全球各地志工編輯的非營利平台）是目前最全面翻譯的平台。維基百科支援將近 300 種語言，其次是 Google 搜尋支援 150 種語言，Facebook 則支援 70 - 100 種語言。即時通訊應用程式由 Signal 領先，在 Android 上支援將近 70 種語言、iOS 上 50 種語言。另一方面，多數平台在廣泛使用的口說語言中，僅專注於支援幾種，其餘大多數語言都沒有支援。以即時通訊應用程式 QQ 為例，僅支援漢語。

調查中多數平台通常會支援的少量語言包括：英語、西班牙語、葡萄牙語、法語，還有少數亞洲語言，如華語（Mandarin Chinese）、印尼語、日語及韓語。阿拉伯語及馬來語等主要語言受到的支援較少，其他

有數千萬人使用的語言也沒有多少介面支援。

如此缺乏語言支援，對世界上多數人來說代表什麼呢？2021 年，我們估計地球上約有 79 億人口，其中多數居住於亞洲（近 47 億）及非洲（近 14 億）。然而網路上的語言卻沒能服務世上大多數人口：

- *非洲語言使用者*：調查中的多數平台都沒有支援大部分非洲語言作為介面語言，90% 以上的非洲人因此需要切換為第二語言以使用平台，而對多數人來說，那就會是歐洲殖民語言或該區域更為強勢的語言。
- *南亞語言使用者*：在南亞，調查中的平台，約半數未提供任何區域語言的支援介面，即使是有數千萬人使用的印地語及孟加拉語等主要南亞語言，相較於其他語言，普遍仍較少受到支援。
- *東南亞語言使用者*：東南亞語言的支援狀況也很相似。儘管印尼語、越南語及泰語在調查的平台上普遍獲得語言支援，其他多數東南亞語言在我們調查的平台上則未有支援。

世界不同區域人民的真實日常生活，支持了 Martin 與 Mark 的研究結果。比方說，來自馬拉威的 Donald 發現，當他問瀕危班圖（Bantu）語系的清達里語使用者，他們如何利用手機溝通時，他們都形容使用清達里語是如何費時費力的過程，因為他們的手機多半僅內建英語、法語及阿拉伯語等語言支援，卻不支援清達里語。這些科技使用上的困難及社會經濟條件，限制了清達里語使用者選購智慧型手機或資費方案的能力。即使是對馬拉威官方語言齊切瓦語（Chichewa）使用者來說，缺乏語言支援也很很棘手：「如果得要使用我不會的英語，那我為什麼要買昂貴的手機或浪費時間設法上網？」

事實上，2018 年多數非洲語言缺乏語言支援的狀況便受到矚目，當時推特才初次承認斯瓦希里語（Swahili），那是東非及其他地方 5000 萬到 1.5 億人口所使用的語言（作為第一或第二語言）。在此之前，斯瓦希里語及多數其他非洲語言在該平台上都被當作印尼語。承認斯瓦希里語的文字及提供翻譯支援也不是由該科技公司發起，而是推特上的斯瓦希里語使用者請願帶來的結果。

拉丁美洲原住民族語言（不若西班牙語或葡萄牙語有殖民歷史），狀況也沒有比較好。馬普切語為現今智利與阿根廷馬普切人所使用的語言，他們在復振馬普切語的 Kimeltuwe 計畫訪談中提及：「要是能夠用馬普切語在 YouTube 或 Facebook 平台上發文該會有多好。還不是要求整個介面翻譯為馬普切語，只是想要能夠在現有選單中，標示該語言為馬普切語。比方說，上傳影片到 YouTube 或 Facebook 時，你無法添加馬普切語逐字稿，因為既定語言清單中沒有馬普切語。因此，你若想要上傳馬普切語逐字稿，只能選擇標示為西班牙語或英語。」

Martin 與 Mark 並未分析行動電話等特定裝置上的語言支援，但我們知道，數位鍵盤是語言學家與科技人員少數有最多進展的關鍵領域。比方說，Google 用於 Android 操作系統的智慧型手機鍵盤 Gboard，支援超過 900 多種語言，這是奠基於不同語言社群與學者的重大成果。然而，唯有買得起相對高階的智慧型手機，才能使用擁有這些功能的智慧型手機鍵盤。

同時，僧伽羅語在斯里蘭卡有超過 2000 萬人作為第一或第二語言，然而 Uda 使用僧伽羅語的經驗卻證明了，想要使用負責語言支援的科技人員無法輕易理解的語言來創造內容，仍然很困難，特別是如果文字形

式與西方歐洲語言所使用的拉丁文字相當不同。她說：「僧伽羅語 Unicode 的主要問題在於，拼出一個字母所需要插入的字元插入順序。按照順序會需要在子音符號後插入變音符號，這是依循拉丁文字建立的歐洲語言思維。然而，僧伽羅語有時候會是變音符號在先，子音在後。」

Unicode 是以語言書寫系統或文字表達文本的編碼技術標準。第 13 版有 143,859 個字元，適用現今超過 30 種書寫系統，因為同樣的書寫系統適用於不只一種語言（例如拉丁文字適用於多數西方歐洲語言，漢字適用於日語、漢語及韓語，天城體則適用於不同的南亞語言）。此外，Unicode 還有字母適用不再有人使用的語言之歷史文字。Unicode 聯盟（Unicode Consortium）（位於美國加州的非營利組織）對我們每天在不同介面上使用的表情符號有決定權。

想了解多數平台與應用程式有限甚至不均的語言技術支援，除了此篇簡要介紹之外，也可以到 Martin 與 Mark 關於語言支援的調查報告中，看更多世界各地其他貢獻者經驗的細節。請不要錯過！

語言內容：取得與生產

« 當地語言版本的女性主義內容特別難取得。婦女發展基金會（Women’s Development Foundation）為偏鄉婦女團體，從 1983 年開始提倡婦女權益。但直到 2019 年我們才開始在網路上以僧伽羅語分享女性主義相關的社會政治與經濟議題內容。»

Uda Deshapriya

« 可惜的是，從過去到現在都難以在網路上搜尋到印尼語版本有教育意義且正面的酷兒內容……若我們在最大且最熱門的搜尋引擎 Google 上搜尋『同志（LGBT）』或『homoseksualitas』（同性戀），會找到許多含有『penyimpangan』（偏差）、『dosa』（罪惡）與『penyakit』（疾病）等字眼的搜尋結果。»

Paska Darmawan

« 在網路上用孟加拉語找得到的（極少數的）酷兒與身心障礙同志相關資訊，大多出自恐同組織與身心健全主義團體之手，這些有心人士大幅地塑造進而推廣恐同主義和身心健全主義。»

Ishan Chakraborty

經由分析上網後所經歷的是什麼版本的世界以及誰的知識，我們得以了解網路上的內容。畢竟所有網站中，有 63% 以上內容以英語作為主要語言。

我們的貢獻者在文章與訪談中，探討以他們的語言有意義地使用網路時，會碰到的各種歷史、社會政治、經濟與科技挑戰樣態。更重要的是，他們都點出以自己的語言在網路上搜尋相關內容，還有以那些語言創造對他們有意義的內容時會遭遇的挑戰。換句話說，那些用其他語言為我們創造資訊與知識的人，根本不了解我們的脈絡與經驗，甚至還不尊重這些脈絡與經驗，這樣對我們來說根本不夠。我們必須能夠為自己及我們的社群產出有意義的知識，或是至少必須能夠支援以我們各種不同語言生產的內容，並進而擴充。

這對那些無法獲得網路服務，以及經歷著不同形式環環相扣的邊緣化與排擠的人來說更是如此。

正如 Joel 在原住民表情符號計畫（Indigemoji project）的訪談中所提及，一切都從他某天他感到挫敗而在車上發推文（tweet）開始，當時他就只是把車停到路邊，開始把亞倫特語跟表情符號配對以描述每個表情符號的意思。自表情符號在網路上問世以來的數十年間，第一民族或原住民族每每連署要用表情符號來表達他們的口述或視覺語言，例如亞倫特語，卻始終失敗。正如我們先前所述，Unicode 聯盟會研議公眾對於新增表情符號的要求，而要求新增代表澳洲原住民國旗的表情符號之類的連署則遭到駁回。對 Joel、Caddie 及其他人來說，原住民表情符號計畫已成為跨世代的工作，要努力突破這些實體與虛擬的多重形式邊緣化，以對自身原住民身分認同及語言有意義的方式來創造自己的內容。

列出表情圖案與相對應亞倫特語單字的推文。來源: 原住民族表情圖案。

我們必須記得，原住民族語言之所以成為現今的「少數」語言，是因為歷史上殖民中大規模種族屠殺導致原住民族遭到毀滅，或在特定區域或土地上從原本的主要居民變成少數人口。這些殖民過程同樣也影響了世界上千百萬人所使用的強勢語言。

Ishan 是有視覺障礙的酷兒學者，對他來說，光是要上網便得要先克服萬難。再來，他還要辛苦地搜尋與身心障礙、酷兒，甚至是與這類議題交集相關的孟加拉語版本資訊。這樣的情況導致了他口中的『邊緣化中的邊緣化』：「一邊是社會上的恐同與身心健全主義態度，另一邊是個人（酷兒與／或身心障礙者）內化的恐同與／或身心健全主義，這些條件互補之下便讓邊緣化機制永垂不朽。酷兒兼身心障礙的個體，在社會上的定位或許可以形容成是『邊緣化中的邊緣化』。」

換句話說，即使是孟加拉語這樣的強勢語言（全球有 3 億人口使用），網路上也缺乏其重要的近用過程及資訊。

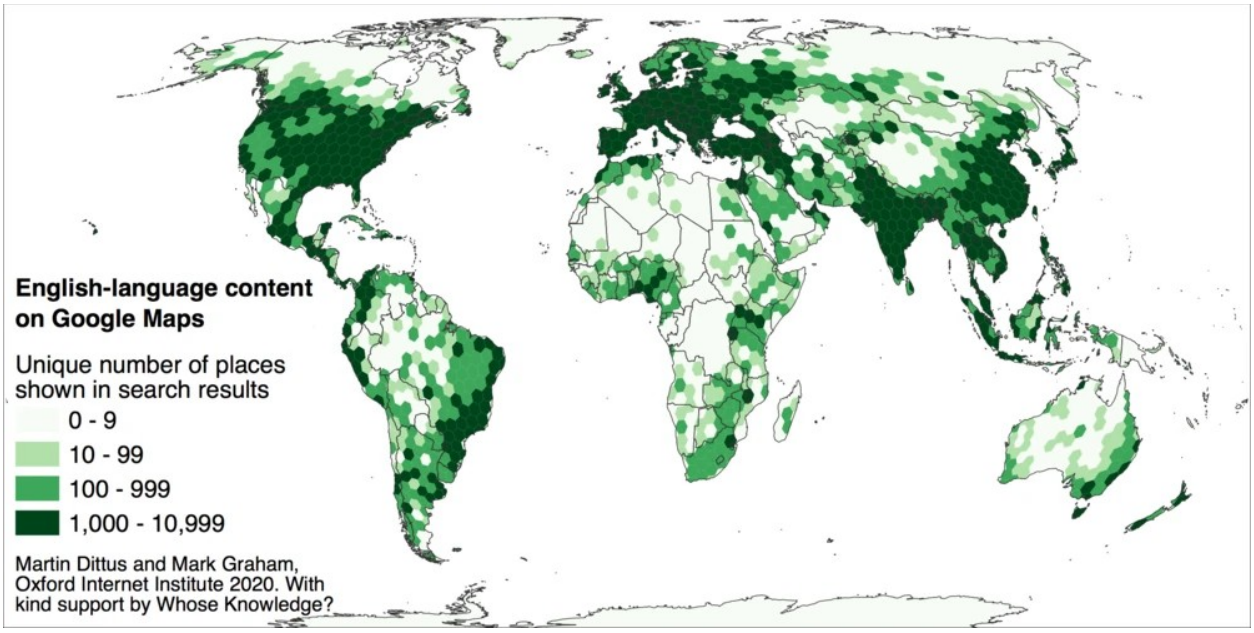
Martin 與 Mark 決定要深入分析 Google 地圖與維基百科，看不同語言在這兩種不同資訊與知識平台上的脈絡範圍與類別。

Google 地圖

我們能否以眾多不同語言使用 Google 地圖？根據所使用的語言，我們透過 Google 地圖所看見的世界是否會有不同版本？

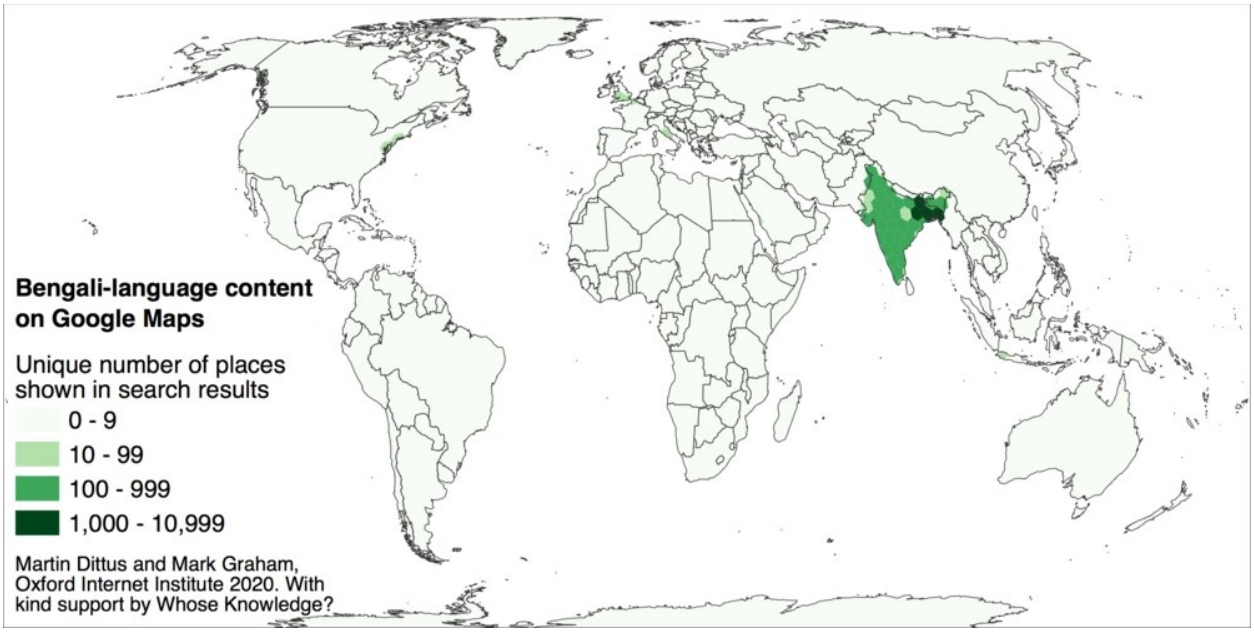
為了回答這些問題，Martin 與 Mark 收集十大最廣為使用的語言在 Google 地圖上涵蓋全球內容的資料，這些語言為：英語、華語、印地語、西班牙語、法語、阿拉伯語、孟加拉語、俄語及印尼語。他們收集了上千萬筆這些語言的個別搜尋結果，在這些資料當中辨識並劃出三百萬個不重複的位置（地點與其他定位）。

毫無意外，我們以英語使用 Google 地圖的時候，地圖內容最多。Google 的英語地圖涵蓋全世界，不過在全球北方（Global North）的密度較高（即有更多資訊），且聚焦於歐洲與北美。南亞、部分東南亞，以及大部分拉丁美洲的涵蓋率也相對不錯。不過，相較之下，非洲許多地方的內容就相對稀少。



英語使用者的 Google 地圖資訊密度。深色陰影區顯示搜尋結果包含更多地點的地方。

相較於涵蓋率相對好的英語地圖，我們發現孟加拉語（也就是 Ishan 的第一語言）則完全相反，涵蓋範圍僅限南亞，特別是印度與孟加拉， Google 地圖在世界其他地方則幾乎沒有孟加拉語的內容。為了有更多內容，並在印度與孟加拉以外的地方能夠導航，孟加拉語使用者必須切換為英文之類的第二語言。印地語（全球第三多人口使用的語言，僅次於英語及華語）的 Google 地圖也是相同狀況。



孟加拉語使用者的 Google 地圖資訊密度。深色陰影區顯示搜尋結果包含包含更多地點的地方。

更多關於 Google 地圖不同語言的情況請見 Martin 與 Mark 的詳細文章。

維基百科

正如 Martin 與 Mark 的平台調查所顯示，維基百科是網路上的語言支援前鋒，使用者介面翻譯的語言數量比我們探討過的任何商業平台還多，包括 Google 與 Facebook。

以維基百科條目的資訊與知識等實際內容而言，維基百科有超過 300 種語言版本。然而這些語言的使用者卻無法獲得相同內容或等量資訊。我們想要進一步探究並回答：維基百科內容在不同語言版本的涵蓋率究竟如何？是否有些語言重現的比例較其他語言高？某些語言社群可獲得的資訊是否比其他社群多？我們在 Martin 與 Mark 的維基百科分析中詳細回答了部分問題。

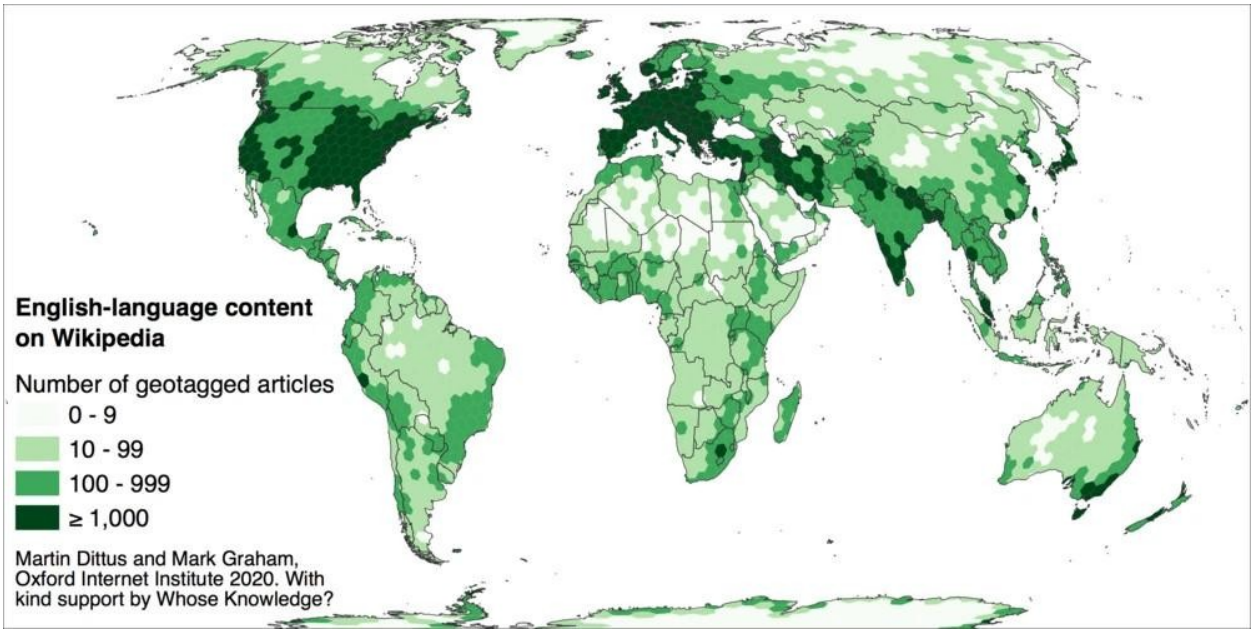
我們使用 2018 年加上地理標籤（在維基百科條目中嵌入地理位置資訊的方式，例如加上經緯度座標）的資料，分析不同語言的條目數量與內容增長。同時也根據「當地」語言進行分析，這些語言的定義為在 Unicode 通用區域資料庫（Unicode CLDR）（網際網路上的語言提供支援的程式碼）內分類為官方語言，或在任何國家至少有 30 % 的人口所使用的語言。

接著再辨識出最為盛行的當地語言，即各國最多人使用的語言。我們找到 73 種在至少一個國家最為盛行的語言。英語是最廣泛使用的語言，也是在 34 個國家為最盛行的語言。其次是阿拉伯語及西班牙語（18 國）、法語（ 13 國）、葡萄牙語（ 7 國）、德語（4 國）及荷蘭語（3 國）。漢語、義大利語、馬來語、羅馬尼亞語、希臘語及俄語在兩個國家為最盛行語言，其餘 60 個語言則僅在一個國家為最盛行語言。

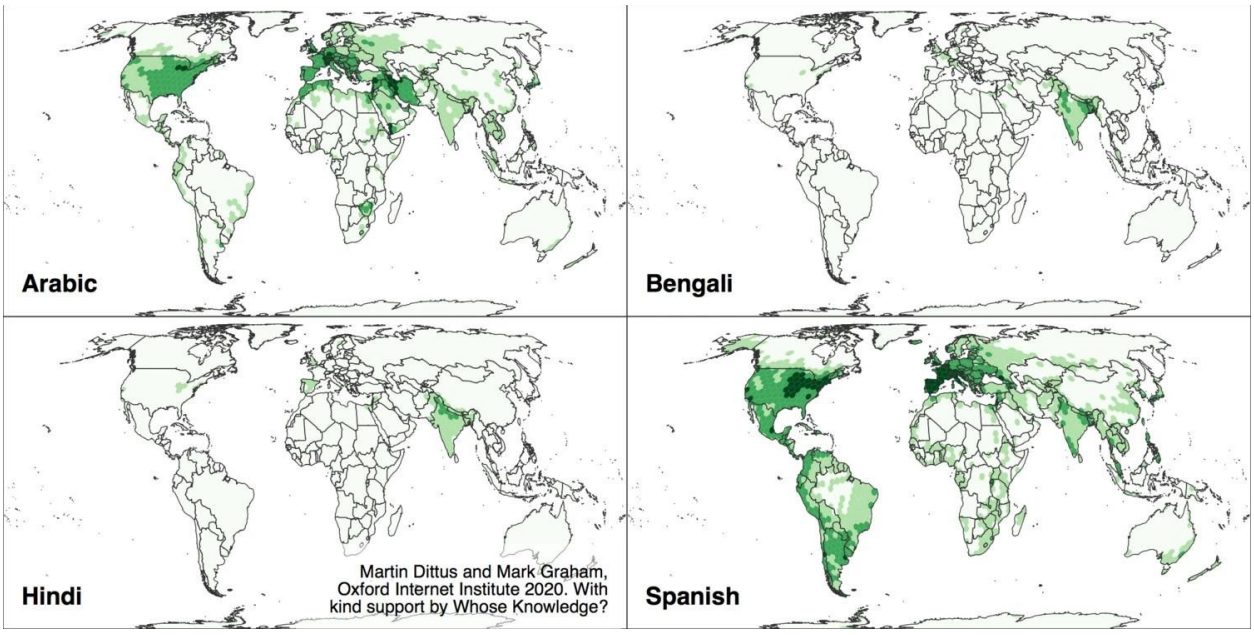
為了比較維基百科內容在每個國家當地語言的分布，我們找出該國維基百科擁有最大量條目的語言版本。

我們發現結果偏向英語內容。英語在 98 國為強勢的維基百科語言，其次是法語（9 國）、德語（ 8 國）、西班牙語（7 國）、加泰隆尼亞語及俄語（4 國）、義大利語及塞爾維亞語（3 國），以及荷蘭語、希臘語、阿拉伯語、塞爾維亞－克羅埃西亞語、瑞典語及羅馬尼亞語（2 國）。其餘 21 個維基百科語言則僅在一個國家最為盛行。

儘管每個語言版本的維基百科條目數量會變化且不斷增長，不同語言版本的維基百科在數量與規模上顯然都有很大差異，無論是條目數量或是編輯社群規模。英語版維基百科規模最大，有超過 600 萬篇條目與將近 4000 萬名註冊的貢獻者。第二大貢獻者社群為西班牙語、德語及法語版維基百科，各自擁有約 400－600 萬名貢獻者與大約 200 萬篇條目。其餘語言版本相較之下規模都偏小：僅約 20 個語言版本擁有超過 100 萬篇條目，70 個語言版本超過 10 萬篇條目。多數語言版本的維基百科內容，僅是英語版維基百科的一小部分。



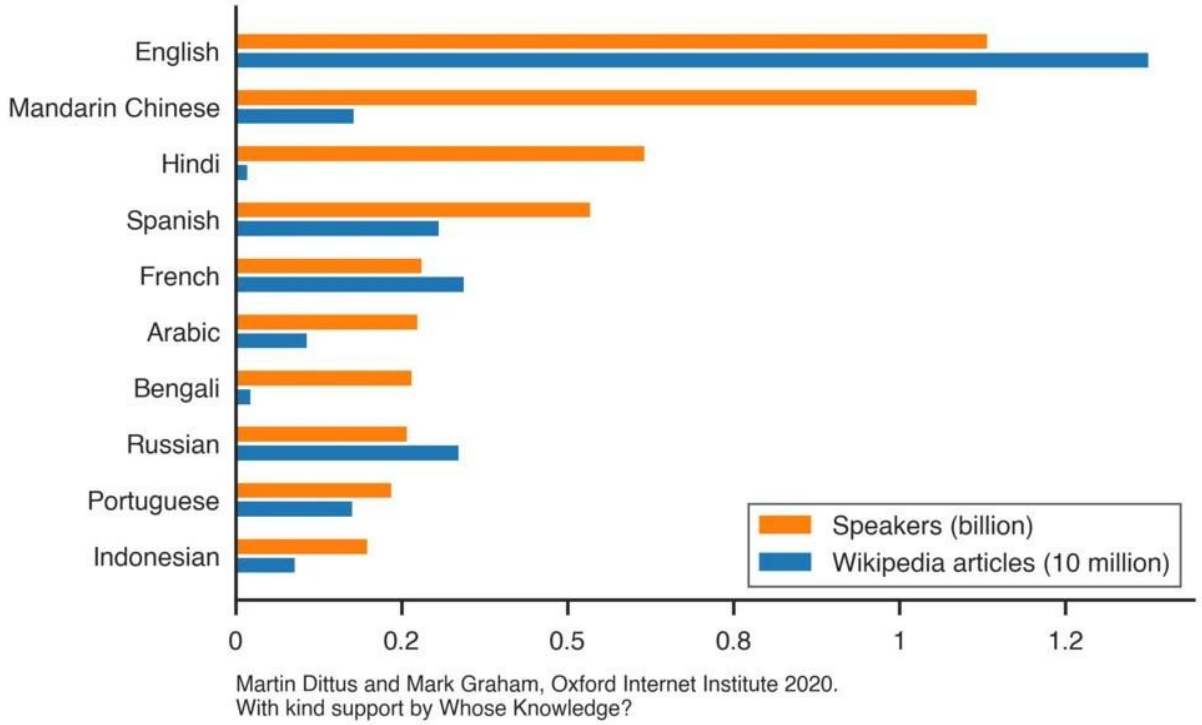
2018 年初期英語版維基百科的資訊密度。深色陰影區顯示含有更多地理標籤條目。



2018 年初期阿拉伯語、孟加拉語、印地語及西班牙語版維基百科的資訊密度。深色陰影區顯示含有更多地理標籤條目。

有趣的是，不同語言版本的維基百科內容分佈，呼應我們先前對 Google 地圖的研究。

若是看不同語言的維基百科條目數量與該語言使用者（包含第一與第二語言）數量相比，也是相同狀況。我們發現，英語、法語、西班牙語、俄語及葡萄牙語等歐洲語言，維基百科條目數量會與語言使用者數量成正比。但是其他廣為使用的語言卻非如此：華語、印地語、阿拉伯語、孟加拉語及印尼語都有上億使用者，然而他們的維基百科版本規模卻較小，相較於歐洲語言版本的條目數量也較少。儘管華語、印地語或阿拉伯語為世界前五大口說語言，使用者數量也勝過法語及葡萄牙語，但法語、西班牙語或葡萄牙語的維基百科條目數量仍比華語、印地語或阿拉伯語的版本要多。



世界十大最廣為使用語言之維基百科內容與使用者數量。（人口估計：2019 年民族語言網，包含第二語言使用者。）

Martin 與 Mark 的文章裡有更多不同語言版本的維基百科分析與資料視覺化，但是這些數據已經清楚呼應了我們來自各地貢獻者的生活經驗。

非主要歐洲殖民語言所經歷的邊緣化與排擠，深入真實與虛擬的世界，即使是阿拉伯語等相較之下仍屬強勢的全球語言。為了用自己的語言撰寫維基百科條目、使用依據該語言脈絡的參考內容，我們需要可靠且廣泛的已出版來源，但（正如我們先前所發現）這對世界上多數語言來說很稀少。身為維基人，Emna 分享在非洲以不同語言搜尋資源與參考資料有多困難：「以我身為維基人要用我們自己的語言尋找參考資料為例，當我說我們自己的語言，指的不僅是我們的突尼西亞阿拉伯語方言，或是阿拉伯語，而是我們橫越非洲時所發現的巨大資源與參考資料缺口。」

即使在歐洲，少數語言使用者以自己的語言使用或編輯維基百科時也會面臨困難。Claudia 發現，雖然她的許多布列塔尼語受訪者知道有布列塔尼語版的維基百科存在，「其中 19% 的人甚至會編輯現有條目或撰寫新條目（8%），」她認為大部分少數語言使用者會切換到強勢語言的原因是比較容易：「可用，不表示服務、介面、應用程式與維基百科真的有人使用。部分研究顯示，少數語言使用者在使用以語言為主的數位科技時，容易切換為他們會使用的強勢語言，可能是因為該語言版本的科技本來就比較好，或是可用服務選項更多。」

維基百科及其一系列免費且開源（程式碼皆可公開取得且集體打造）的知識計畫，是最有希望且有幫助的線上多元語言知識空間。比方說，其志工社群知道並理解英語、阿拉伯語或漢語並非以單一形式存在，不過要表達語言這種多重脈絡及內容卻往往不太容易。正如我們從分析與經驗中得知，維基百科同樣也苦於過去到現在的強權與優勢結構，這樣的結構會讓我們以不同語言、在語系之間創造並分享知識的方式與形式扭曲。

希望能有更多元語言網路環境的我們，作為個人、組織與社群，未來有什麼方向呢？在接下來的最後幾節我們將援引來自目前與大家分享的所有數據及故事的洞見，概要分享我們的心得，以及或許能帶領我們邁向真正多元語言網路環境的脈絡、理解與行動。

我們對多元語言網路環境有什麼樣的了解？

隨著這份報告的推進，我們對語言、網路，以及網路上的語言有了許多認識。以下概述我們一路走來最重要的見解。

心得： 語言不僅只是溝通工具，更是知識的代理媒介，也是存在於這世界的必要方式。這也是為什麼多元語言如此重要，因此我們更加尊重並肯定我們的諸多自我和不同世界的完整豐富與紋理。

脈絡： 人們用 7000 多種語言認識他們的世界並表達自我，若考量到點字、手語及音樂語言等，還不只 7000 種。

然而主要科技平台與應用程式的語言支援只占 7000 多種語言的一小部分，其中僅約 500 種語言以任何資訊或知識的形式重現於網路上。世上有些最廣泛使用的語言在網路上根本沒有什麼語言支援或資訊。網路上最豐富的語言支援、最淵博的資訊（包括在 Google 地圖與維基百科上），以及多數網站，都是英語版本。

反思： 網路環境遠不如我們想像或所需要的那樣有著多元語言。

分析： 多數人必須使用與他們最接近的歐洲殖民語言（英語、西班牙語、葡萄牙語、法語……）或是區域強勢語言（漢語、阿拉伯語……）來使用網路。過去到現在的強權與優勢結構對網路上語言使用（與否）的方式來說至關重要。

我們該如何做得更好？：打造多元語言網路環境的脈絡及行動

« 多數情況下，使用少數語言會需要龐大的毅力、意志與韌性，
因為少數語言的使用者經歷了許多缺點與困難。»
Claudia Soria

« 包容且能代表原住民族的多元語言網路環境目標，必須考量並面對殖民壓迫持續至今的社會遺產與經驗。
多元語言的網路環境不能只是想要具有代表性，而是要考量到殖民歷史，
也要同時能夠設法推廣由原住民為了原住民自身的語言存續及學習的環境。»
Jeffrey Ansloos and Ashley Caranto Morford

« 馬普切青年及小孩與科技及網路並肩成長。網路是這些人能夠接觸馬普切語的空間……我們族人的故事必須有人撰寫，必須以馬普切語寫下或口述……我們的故事不一定是什麼偉大英雄事蹟，不像那些殖民與後殖民國家所讚揚的故事。我們的歷史便是每位在困境與暴力中存活下來的馬普切人的故事。必須移居都市賺取薪水的婦女、從都市返鄉的男男女女，可是大社 (lof) 裡已經沒有他們的位子了，失根的他們只好重返都市，不再有地方可返鄉。每一位馬普切人的經驗與回憶，構成了我們族人的集體記憶。»
Kimeltuwe project

在網路上語言使用的狀態報告摘要的這部分，我們整合不同貢獻者對這份報告的不同見解，以及我們 2019 年的網路語言去殖民化大會，以了解世界上與網路上多元語言的不同脈絡、挑戰與機會。向我們自己與彼此提出四個主要問題，或許可以讓我們想像及設計出更多元語言的網路環境。

- **誰的權力與資源?**
- **誰的價值觀與知識?**
- **誰的科技與標準?**
- **誰的設計與想像?**

誰的權力與資源?

脈絡

我們的統計分析與人們的真實生活經驗都清楚顯示，語言在真實與虛擬世界的邊緣化不僅是基於使用語言的人數。

全球有超過 6000 個原住民族群的公民使用原住民族語言，他們原本是世界上大部分地區的主要居民，直到殖民與種族屠殺摧毀或削減他們的族人及語言。在某些語言最為多元的大陸上，例如亞洲與非洲的強勢語言，卻很少在網路上重現，有時甚至完全沒有。若考量使用諸如有許多變體的阿拉伯語、漢語、印地語、

孟加拉語、旁遮普語、坦米爾語、烏爾都語、印尼語、馬來語、斯瓦希里语、豪薩語等的人，離散在不同國家與大陸，即使這些語言在各自區域是比其他語言更為強勢的語言，在網路上顯然仍是遭到邊緣化的語言。

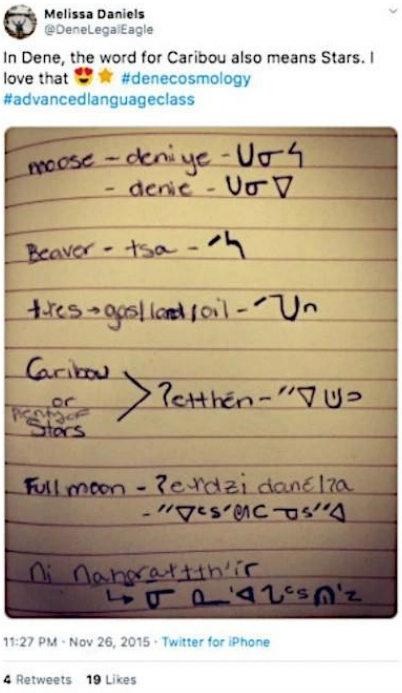
這些數位邊緣化與排擠的形式並非意外，而是由過去到現在的強權與優勢結構造成。這同時也代表投入語言基礎建設的資源，從出版與學界、政府到科技公司，都已經偏向某些區域（歐洲與北美）及某些語言（英語及其他西方歐洲語言）。即使在歐洲與北美，原住民、黑人及其他邊緣化社群都覺得很難跨世代保存他們的語言。

特別是當殖民與資本主義的力量，與其他歧視與壓迫制度環環相扣，諸如種族歧視、父權體制、恐同、身心健全主義、階級歧視及種性制度等。這表示，無論是否為世上最多人口所使用，某些語言就是網路上最為突出的語言，而且主要就是歐洲殖民語言。這也表示若有其他更為邊緣化的語言版本的資訊及知識，這些語言的內容會受到誰能夠及有權力創造而受限，或是妨礙他人生產替代資訊。比方說，網路上沒有僧伽羅語版本的女性主義內容，或是孟加拉語版本的和酷兒與身心障礙相關的正面內容。

由於語言為我們存在的核心，無法用自己的語言完整表達自己以及自己的諸多面向，這件事本身便是一種暴力。但是這些邊緣化的結果卻以其他的暴力形式呈現。正如 Uda 所說：「缺乏尊重、女性主義與人權友善的數位內容，導致以當地語言進行溝通的網路空間對婦女、酷兒及少數群體抱持惡意。顯然缺乏數位媒體能夠對抗始終沾染負面刻板印象的主流論述。這樣的情況會讓網路上的仇恨言論、性與性別暴力加劇。」這些暴力形式的魔爪會進一步伸向原住民、受壓迫階級與少數宗教社群、身心障礙與不同能力者，及其他邊緣化社群。

同時，這些社群也利用網路來對抗針對他們本身及語言的不同形式的跨世代暴力。Jeffrey 與 Ashley 分析了約 3800 則由 60 個加拿大聯邦政府承認的原住民語言團體所發出的推文，使用超過 35 組主題標籤（hashtag）與 57 個關鍵詞。他們發現，加拿大及世上其他地方的原住民族，透過推特上的主題標籤積極連結、參與並合作復振原住民族語言，使其蓬勃發展。

正如他們所說：「在加拿大原住民族語言遭殖民同化政策鎖定抹除的社會脈絡下，推特標籤網路形成對原住民來說獨特且有意義的脈絡，讓他們能分享關於原住民族語言的知識。在我們研究涵蓋的各種網路中，有語言恢復或重新連結跨世代殖民同化政策倖存者的實例。」



甸尼族（Dene）的 Melissa Daniels 同意引用這則推文，
並希望能認可提供這項教學的語言教師，甸尼族長老及教育者 Eileen Beaver。

行

- 承認強權與優勢在支持語言基礎建設的不同機制與過程下的結構與過程。
- 確保將資源賠償給邊緣化語言與社群，包括語言學習及語言規劃。
- 讓那些經歷多重交錯壓迫與暴力型態的社群，能有資源以他們所選擇的語言與形式，為自己社群創造並擴充來自社群的資訊與知識。

誰的價值觀與知識?

脈絡

網路的歷史與科技是根據西方知識論的世界觀建立，也就是西方認知與行事方法。更明確來說，網路從過去到現在都主要是為了白人（如今也有些棕色人）特權男性而設計，也主要由他們管理。這表示，最常處於網路架構與基礎建設核心的價值觀為科技決定論，即科技被視為任何社會變遷的主要（且有益）的肇因還有個人主義，即主要焦點與驅動因素為個人而非集體。

此外，這樣的世界觀扎根於啟蒙時代，是 18 世紀全球北方邁向某種理性為本的科學與科技運動。我們卻忘了，早在 18 世紀前，數學與科學便在全球南方遍地開花。比方說，第一套書寫與數字系統源於美索不達米亞，位於現今伊朗與伊拉克境內。更重要的是，我們忘了這樣的全球北方科學與科技啟蒙「黃金時代」的資源，都來自帝國時代於全球南方對亞洲、非洲、拉丁美洲、加勒比海及太平洋群島的大規模殖民、奴役種族屠殺以及榨取資源。現代資本主義的榨取性質，源於過去的殖民歷史，且至今屬於科技資本的一環。

不只物質資源，過程中遭到摧毀、忽略或破壞的還有其他形式的知識、作法與存在，也就是非西方知識論例如原住民知識，或世界上較不具優勢族群的知識。正如我們前面所提及，對語言如此重要的知識代理媒介來說，最具毀滅性的結果，便是非歐洲語言全然遭到貶值，且積極破壞或善意地忽略了口述與非書寫系統的語言。這種以少量優勢語言書寫內容的偏誤，持續將我們導向出版與學術的某種特定的書寫「知識」進而影響自然語言處理（natural language processing）所需的潛在紀錄與資料，或是某些算是網路基礎建設的自動語言處理系統，例如 Google 翻譯。

正如 Ana 所描述：「儘管全世界有將近一半的語言缺乏書寫系統且長期維持口述的傳統，擁有廣為人知且廣泛使用的字母系統的語言仍佔據了網路。網路強化了系統性的排擠，未來唯有那些書寫語言能受到保存。」

失去語言的同時，我們失去的未來比意識到的更多。我們同時失去不同語言的表達形式，以及這些語言內涵的所有世界觀與知識。在人類處於全球崩潰邊緣之際，原住民社群與其知識正守護著我們的生態系，保存生命本身。毫無意外地，我們知道語言流失與生態多樣性流失有直接關聯，也與全球生態系的毀壞有直接關聯。

網路可以是保存與擴充不同語言與知識形式的基礎建設，令人期待，因為其豐富的媒介形式能模仿並重現口述、視覺及超越文本的語言體現。但這樣根本的網路承諾，絕對不能再度奠基於殖民資本主義與父權價值觀。社群人民該如何保存與復振他們的語言與認同，按照自己選擇的方式分享他們的知識？舉例來說，在多數原住民部落，有些知識很神聖，不得公開分享。

Papa Reo 便是由部落主導的紐西蘭毛利語語音辨識技術。這個計畫由毛利社群創造並維護其科技與資料，他們相信這種形式的資料主權非常重要，如此才能確保透過毛利語分享的知識是為了毛利人且用於毛利人，而非用於追求營利的公司。有意思的是，儘管承認開源科技的價值，Papa Reo 團隊卻也決定不將他們的資料加到開源資料庫裡，因為毛利社群無法獲得多數開源社群擁有的資源與優勢。另一方面， Mukurtu 是與原住民社群一齊打造的開源平台實例，供他們管理自己的語言資料。

行動

- 創造、合作與分享為了公益、以集體及社群價值觀為設計核心的網路語言基礎建設，且聚焦女性主義者、原住民對主權及體現的概念。
- 繼續挑戰並批評本質為資本、專有、榨取及壓迫的語言科技基礎建設。
- 承認自由與開源語言科技也需要注意自身相對的優勢，並尊重邊緣化社群如何自主決定並定義「開放」，以及他們想要與世界分享什麼。

誰的科技與標準?

脈絡

世上大多數語言目前缺乏重現與支援，這不只是科技產業的責任。然而，來自全球北方的科技，要為持續維持並擴大網路上以語言為本的不平等及數位殖民主義負責。

因為這 7000 種口說語言中，大多不被認為或視為網路基礎建設的必要環節，負責設計並創造多數我們所使用的數位平台、工具、硬體及軟體的大型科技公司，會忽略創造真正多元語言網路環境的需求。畢竟，他們僅在有助生意時，才會知道需要提供語言支援：無論對象是歐洲殖民語言或他們口中「新興市場」的語言。事實上，大型科技平台正開始對南亞與東南亞的強勢語言提供更好的語言支援，因為這些語言正成為這些公司的最大客群。

同時，網路上有些最廣為分布的語言科技便是由這些公司創造並管控，因為他們才有這樣的資源與能力。維基百科因為開源且由全球志工社群支援，成為知名的例外。一般來說，邊緣化社群想要以各種邊緣化語言創造有深度且細緻的內容，其所需要的工具與科技，不會源自專有且利益導向的動機。更糟的是，目前這些公司所打造的語言科技，是大規模自動化系統，仰賴各種來源的大量語言資料，即使那些來源可能是充滿暴力及仇恨的語音，或排斥邊緣化團體的書寫。

正如 Jeffrey 與 Ashley 所形容：「在多數『倖存』的原住民族語言與學習社群中，種族歧視已被認為是來自推特生態系內最大的社會難題。更進一步來說，這些社群被真人與自動化機器人在內的各種使用者盯上以煽動性文本及多媒體內容等多種方式攻擊，有時候甚至達到仇恨言論的程度。以機器人使用者來說，這些帳號似乎會遵循相似的模式散播假訊息，且經常發表毫無意義的意見，反映出彙整和自動產生的內容。」

公司若是不夠在乎缺乏當地語言專家與內容會帶來的結果，便會導致極大的傷害與積極暴力。在緬甸 Facebook（或 Meta）基本上就等於網路，社運人士長年來提醒該公司要小心仇恨言論，後來公司終於成立緬甸語言團隊。2015 年，Facebook 有 4 位緬甸語使用者負責審查內容，但全緬甸有 730 萬活躍使用者。對語言與內容缺乏關心的結果是什麼呢？聯合國發現 Facebook 便是造成緬甸對羅興亞穆斯林種族屠殺的其中一個原因，這間公司在緬甸政府於國際法院遭控告一案中扮演關鍵角色。

同樣地，儘管印度為 Facebook 最大市場、還使用世界上最常見的口說語言，針對穆斯林、低階種性及其他邊緣化社群的仇恨言論在印度卻仍然未受到積極管理。事實上，該公司 84% 的「全球檢舉/語言涵蓋預算都用於對付美國的假訊息，但美國的使用者還不到總使用者的 10%。剩餘 16% 才用於世界其他地方。

科技公司必須承認，語言科技的擴充需要更深入廣泛的資源與社會政治脈絡，還要承諾打造安全親切的多元語言數位體驗。

我們的貢獻者探討了創造這些體驗所涉及的需求、挑戰與機會。特別是所有語言缺乏的基礎建設（從網路

存取到有效裝置）及科技，導致邊緣化語言在數位運用上很累人也很困難，緩慢且不切實際。我們提供幾個有力的實例。

語言科技往往不是為了邊緣化語言而設計。

Donald 談及他在馬拉威的社區，可輕易以他們的語言上網及溝通的裝置很罕見。他訪談了 20 位清達里語使用者，其中 10 位為學生，10 位為社區長輩。20 位受訪者中，僅有 5 位擁有智慧型手機或功能型手機，7 位沒有任何裝置。20 位中僅有 4 位擁有筆記型電腦，都是學生。至於上網，只有大學生能透過大學或個人資費方案上網。

上網後，世界上多數人也很少能以自己的語言使用鍵盤。多數社群必須拿主要為歐洲語言設計的鍵盤，再將自己語言的字元貼在鍵盤上。這樣做對個人電腦的鍵盤來說已經相當困難，手機的小鍵盤則根本不可能對主要為口述，沒有商定書寫系統的語言來說，更是困難。

如 Ana 對自己語言的描述：「鍵盤沒有正確的薩波特克語符號來代表我們語言的聲音與聲調。推動原住民族語言書寫的各方勢力長期協商設法達成共識，希望能使用如拉丁文字之類的統一格式，這種格式多少受到西方影響及強迫接受，也有一派的語言使用者接受並採用。」這對亞倫特語這類語言來說更加困難，因為他們的語言結合聲音與手勢，正如 Joel 與 Caddie 在原住民表情圖案計畫中所提及。

如果你來自原本在這個世界就感覺危險及不安的社群，而網路存取又不是以你自己的語言設計，那你就不太可能在網路上感到自在。Paska 指出：「印尼多數同志（LGBTQIA+）個人對網站的技術面仍舊不太熟悉，也不夠了解搜尋引擎如何運作。」

全球北方優勢地區的邊緣化社群也面臨了同樣挑戰，缺乏他們自己語言版本的正能量及有時可救人一命的內容。Jeffrey 與 Ashley 解釋：「以加拿大來說，主要挑戰之一……在於原住民族語言的現有翻譯科技有其技術限制。哈爾魁梅林語（Hul'qumi'num）、薩尼奇語（Sḵwx̱wú7mesh / Squamish）、肋筐恩語（Lewkungen）、克里語（Neheyawewin / Cree）等原住民族語言，被推特的翻譯科技誤判為德語、愛沙尼亞語、芬蘭語、越南語及法語。」

整體來說，Uda 證實：「以當地語言創造數位內容依然是挑戰，因為缺乏工具，以及現有工具太過複雜。以當地語言創造內容，會需要特殊的工具與技能。這項挑戰也是當地語言缺乏進步內容的原因。」

語言科技多數是由上而下的設計方法，優先考量為獲利而非平等與安全。Claudia 清楚描述了多數科技公司採取的方法，她說：「科技與媒體的供應都是大型公司由上往下傾倒，幾乎沒有什麼語言使用者社群的參與。在這樣的情況下，會發現這些方法很傲慢：由於可用的東西很少，不管提供什麼，從定義上來說都該是好的而且受歡迎的東西。這些公司提供的往往是現成解決方案，沒有考慮到少數語言使用者的真實需求願望與期待。彷彿就是認定，這些語言使用者無論獲得什麼產品或機會都該要心懷感激，不管是否真的有興趣或與他們的生活息息相關。van Esch et al. (2019) 則是知名的例外，他不斷強調規劃自然語言處理的應用發展時，要與語言使用者密切合作。」

這種方法很少去理解邊緣化語言社群生活的脈絡，還有如何以豐富且細緻的方式將他們的語言帶入網路世界所需要的不同設計。當 Emna 訪問來自蘇丹的 Gamil 時，他用蘇丹阿拉伯語回應：「蘇丹是有許多部落以及各種不同傳統與習俗的國家。因此，北部說的是阿拉伯方言（跟我說的一樣），那當然是因為殖民主義。東部跟西部的部落則使用不同的當地語言，只有他們會說而且聽得懂。對北部的人來說，很少能找到有人聽得懂，除非他／她曾經在那裡居住，與居民互動過。我們的語言源自庫希特（Cushitic）語族。庫希特與努比亞（Nubian）文明與我們的文化相關。高壩（High Dam）沉沒時，我們失去了庫希特主體意識，也沒有找來字典將庫希特語解碼翻譯為其他語言。但努比亞語為大家所知，而且存有翻譯。連在華為手機上都有努比亞語可選擇作為設定語言。東部跟西部的語言則沒有書寫文字（也可能有，我不確定，我得要確認）。我們在北部說的是阿拉伯語。我們是講阿拉伯語的非洲人，不是阿拉伯人。」

Emna 的訪談促使她說出：「網路需要納入所有使用者的需求，包括書寫以及不書寫的使用者。然而，改變不單只是使用者的責任，設計與開發軟體的公司也必須擔起未來網路設計的責任。感覺大家現在都在討論包容性、對抗種族歧視及其他形式的殖民主義。然而要在網路上推動改變，企業必須討論他們是如何持續導致排擠；網路設計師、網路科技工程師及科技公司老闆都應該有所貢獻，讓他們的軟體能夠為所有使用者所用。」

要批判性地分析這些不同的排擠形式，就要承認這些數位科技的根本，也就是程式碼本身，幾乎都是英文少有程式語言奠基於其他語言之上，亦即科技人員本身必須要英語夠好才能寫程式。Qalb 阿拉伯語程式語言是以阿拉伯語的語法及書法為本，是少數嘗試突破這種趨勢的實例。不過，一般來說，科技產業必須了解語言優勢是如何深植所有科技與科技人員心中。

正如 Jeffrey 與 Ashley 所說：「我們的研究清楚指出，在網路上推廣原住民族語之前，我們必須先對網路科技本身分析利弊，並且還關乎該語言在社會群體如何普及化的社會歷程。」

行動

- 認可由語言使用者自行設計與創造的多元語言科技與內容，是科技公司與標準組織必須列為優先的基本人權，且由聯合國教科文組織等全球機構支援。
- 為語言基礎建設打造由社群領導、全球治理的典範，能夠基於信任及尊重與科技公司及其他機構合作。
- 重點擺在創造語言資料與科技的社群同意與倫理，確保語言社群有權力並能安全地保障他們分享的方式與內容，特別是那些因為各種環環相扣的壓迫與歧視而更加邊緣化的社群。

誰的設計與想像?

從我們分享的數據與故事中，可得知唯有在我們將重點擺在語言社群本身的需求、設計與想像時，有意義且有效的語言基礎建設才可能成真。唯有被當成少數的世界多數也參與打造科技時，邊緣化語言才能蓬勃發展擴充。

我們的貢獻者建議了一系列作法來加以確保能這樣發展，其中包括要科技公司聘用來自邊緣化語言社群的科技人員與其他專家，還要以負責任的態度規劃社群資源。他們建議根據語言脈絡建立多重結構與過程，而非採用單一公式。如 Claudia 所說：「少數語言使用者不需要現成的解決方案：須聆聽他們的精確需求與要求，納入因應這些需求而生的產品中。不同的少數語言的社會語言學脈絡大相逕庭，因此解決方案也必須有所區別。」

我們的貢獻者也說明自己社群的責任，以及利用他們可用科技保存並擴充語言的方式。Ishan 表示：「身為身心障礙的酷兒，我深信我們必須善用可用的社會文化經濟資源，讓大家聽見並看見我們的經驗、期望與需求。我們這些網路使用者必須負起責任，打造包容且可用的網路。」

Ana 說明，相較於那些仰賴文本的平台，擁有較好口述及視覺基礎建設的平台在她的社群的使用率較高。「時至今日，以薩波特克語來說，網路上口述比書寫語言還要廣泛使用。YouTube、Facebook、WhatsApp 及 Instagram 等社交平台，對口述語言來說是較為可用且友善的資源。這些平台允許使用者上傳視覺內容讓訊息更豐富，由使用者自行決定傳遞訊息的方式，而不會落入書寫的窠臼。在全球擁有最多使用者的這些平台，正是原住民社群所使用的平台。以山脈薩波特克（Sierra Zapotec）社群來說，大多數使用者透過 Facebook 連上網路，他們透過該平台宣傳傳統祭典、舞蹈、音樂，講解重要事件，也向移民與當地社群發布公告。早在新冠肺炎之前，Facebook 已是移民家庭哀悼致意的重要工具，因為喪禮與儀式都透過該平台傳播。有些薩波特克社群也用同樣的平台來重新傳送廣播節目為廣泛使用類比溝通媒介如收音機的鄉村地區，與網路等通用且無所不在的空間搭起重要的橋梁。」

對那些可以使用書寫形式語言的人來說，主題標籤已經成為社群數位連結的有趣方式，能夠學習並傳播自己的語言，並鼓勵其他社群起而傚尤。如 Jeffrey 與 Ashley 所形容：「在加拿大原住民族語言遭殖民同化政策鎖定抹除的社會脈絡下，推特主題標籤網路形成對原住民來說獨特且有意義的脈絡，讓他們能分享關於原住民族語言的知識……比方說，哥威迅語（Gwichin）學習者的主題標籤網路鼓勵了奧吉布韋語學習者建立他們自己的主題標籤網路。同樣的，推特上的克里語網路鼓勵了哈爾魁梅林語學習者成立他們自己的推特每日一字。」

邊緣化社群對這些專有平台廣泛且有創意的用途，便是科技公司該與他們合作而非戒備的原因。

同時，Claudia 也提醒來自邊緣化社群的語言社運人士，要更見多識廣彼此整合，才不會在過程中流失能量與資源。「雖然值得讚揚，這些（社運人士的）計畫卻往往面臨缺乏整合、規劃不足，甚至能見度不足。這對資源有限的社群來說會造成很嚴重的問題：重複做別人做過的事。在這兩種情況下，最主要問題都在於不夠了解有什麼可用、又需要什麼。想要將少數語言科技去殖民化，就必須更清楚瞭解少數語言在數位媒體上的使用程度，以及頻率與用途又如何。了解少數語言使用者面臨的挑戰也一樣重要，當（若）他們嘗試使用這些語言：是否會面臨技術困難？是否因某種自己造成的妄想而遭受阻礙？由於以少數語言書寫等於曝光於外面世界，人們是否會避免這麼做以防遭到嘲弄或汙名化？同樣的，對於少數語言使用者想要什麼樣的數位機會所知也不多：他們想要或期待能用什麼呢？

這份報告便是要嘗試了解這些挑戰，以及改進的方向。不同語言一再複製同樣的事，不會有效果。比方說用英語打造應用程式，然後認為在印尼語環境下也能同樣運作，這種想法非常有問題。改善網路環境就要改變人們之間的權力動態，而非只是修正技術問題。網路上的多元語言環境是複雜的社會技術與政治問題我們必須將語言社群的需求而非網路技術列為優先，如此一來才能真正讓語言科技變得更有效實用。

更重要的是，我們知道要讓語言基礎建設朝好的方向改變，得要靠所謂「被當成少數的世界多數」的設計與想像。「原住民表情符號」出現的時機，正逢澳洲中部快速採用科技與更好的連結。這個機會讓請當地人得以想像他們能用這些新平台來做些什麼，如何不讓那些平台只是成為另一個殖民的力量？還有我們該如何將我們的語言與文化嵌入其中，把那些平台變成我們自己的？」

行動

- 讓基於在地但連結全球的語言社群的脈絡、需求、設計與想像成為語言科技的核心，而非認為單一語言可適用全體的科技。
- 有創意地全面使用網路科技，全面探索語言的體現（口述、視覺、手勢、文本……），讓不同形式的知識可輕易表達與分享。
- 向原住民族學習設計語言科技，尊重集體與社群回憶同時也為未來規劃。讓我們回顧以邁向未來。

最後，你能做什麼？

別人若英文講得沒有你那麼好，不表示他們是笨蛋。只表示他們更擅長世上其他 7000 種語言的其中一種。

我們大家擁有不同的技能與經驗，必須合作以創造真正多元語言的網路環境。我們也必須確保，以這些諸多語言分享的資訊與知識不會造成傷害，而是會為我們的世界帶來集體利益。我們需要所謂的「團結行動」。

若你身在科技界：

- 承認貴公司的政策對多元語言網路環境、以及深化共享的人類知識有何貢獻（與否）。
- 將（邊緣化）語言的國際化與在地化工作列為策略核心，而非視為周邊政策。與社群合作，而非由上到下毫無脈絡地推行。
- 接受經過完整研究而對大型語言模式與自動化語言科技提出的批評，以及缺乏思慮縝密的人為監督可能會造成的廣泛傷害。
- 在開發過程中，透過社群維護的較小型資料集，建立一套人為謹慎處理所有跟語言有關的規範，包括語言內容、資料集的建置以及審查管理工作。
- 以尊重的態度與社群合作，特別是那些最邊緣化且最有可能因為缺乏照護與關注而受傷的社群。
- 將資源用在那些花費時間及專業與你合作的語言社群。

若你身在科技標準組織：

- 承認語言標準需要豐富的脈絡。
- 與邊緣化語言社群建立更好的關係與過程，即使不是由社群主導也能夠與社群合作建立更多標準。
- 用心邀請更多來自邊緣化語言社群的成員參與治理，給予他們充分參與所需的資源。

若你身在政府：

- 認知到國民語言版本的內容必須要能為所有人可用，而非僅限少數優勢人口。
- 支持在你們區域內所有受到邊緣化或歧視的語言，擴充該語言版本的內容。
- 支持你們區域內所有邊緣化語言的保存與數位化，而非僅支持強勢語言。

若你身在自由開源科技與開放知識圈：

- 承認自由開源的科技與知識也有其權力不平衡與限制，即使本意是為了大眾利益。
- 尊重邊緣化社群在分享其知識時設下的界線，因為他們的知識在過去曾以不同方式遭到剝削與商品化。
- 與邊緣化語言社群合作，創造他們需要而非你認為他們需要的科技與知識。

若你身在館聯（GLAM；藝廊、圖書館、檔案庫、博物館與記憶）機構：

- 承認語言是你所策展、保存與展示的知識與文化核心。
- 與邊緣化語言社群合作，透過標示出處的方式（或是素材的擁有者或位置），確保他們的歷史及語言能以他們想要的方式獲得肯定、認可與強化。包括邊緣化社群選擇不要公開分享某些知識與素材的權利。這一點很重要，因為多數館聯機構都源於複雜的殖民與資本主義歷史，特別是全球北方的機構。
- 確保你收藏的語言素材可自由且輕易地供邊緣化社群及其同伴使用，讓我們能夠一起打造集體的語言基礎建設。

若你身在教育界：

- 承認我們的教育偏向以文本為主的來源以及某些語言。
- 擴大你教學及學習的方式，納入多種語言、不同的語言形式及其體現的知識。
- 盡可能地閱讀、聆聽、在翻譯中引用，並鼓勵其他人也效法。

若你身在出版界：

- 承認目前世界上多數出版偏向歐洲殖民語言。
- 擴大出版的語言數量，並將這些語言版本都數位化。
- 出版更多多元語言的書籍與素材。
- 實驗多模態的出版形式，讓口述、視覺、文本等不同語言形式能更易於同時分享。
- 尊重並認可你的譯者。

若你身在慈善界：

- 承認語言是各種人類專業、經驗與知識的核心，無論你資助的是什麼樣的議題。
- 為你支持的各種全球及區域活動、集會提供資源，進行多種語言口譯。
- 支援你所服務的社群，以他們的語言生產素材、保存與數位化，確保你自己的素材有你所服務的社群語言版本。

若你身在邊緣化語言社群：

- 承認自己並不孤單。
- 知道你的社群有權利決定想要用什麼方式與世界分享什麼知識。
- 與社群內的長輩、學者及年輕世代合作，也與其他社群的朋友合作，收集並分享這樣的知識。
- 若你想與其他從事同樣工作的人連結，請與我們聯繫！

若你單純只是熱愛語言，想知道能做什麼：

- 承認語言為我們存在及行事的核心，對各種知識及文化來說都很關鍵，包括你自己的！
- 與家人、朋友及社群討論英語及其他少數幾種語言如何以及為何稱霸網路近用及內容，還有該如何一起改變這種現象。
- 積極尋找、閱讀、聆聽並分享邊緣化語言社群的貢獻（包括這份報告！）。
- 若你願意收到我們計畫的更新通知，請在社群媒體上追蹤我們！

致謝

全球許多邊緣化社群（原住民及更多社群）將語言視為身份認同以及為人處世的核心，我們深愛也敬重大家，更與大家站在一起。這些社群以有意義的方式保存、活化並擴充這些語言及表達形式，激勵我們想像更多元語言及多種的網路環境，讓我們能夠成為最豐富全面的諸多面向。我們同時也深深感謝所有與我們同樣熱愛語言的社群、機構學者及科技人員，每天都努力讓網路環境如我們的真實世界般擁有多元語言。

致我們的諸多貢獻者、譯者，以及世界各地的社群（特別是那些加入我們 2019 網路語言去殖民化對話的人）：感謝你們對這個世界的貢獻以及扮演的角色，也謝謝你們耐心陪伴我們熬過這辛苦的兩年！特別感謝我們的插畫家如此有創意地將文章視覺化，謝謝我們的動畫師讓這些插畫動起來。

深深感謝所有從不同角度、以不同語言審閱我們工作成果的朋友與社群。所有的錯誤都歸我們，但你們的支持與團結，讓這份半成品有了更好的面貌。最後，致彼此以及我們有血緣與自行選擇的親友：即使只是虛擬的，若沒有大家彼此相互扶助，我們不可能撐得過 2019 與 2020 年。愛與信任是最棒的語言。

定義

我們所討論的不同語言及歷史面向，有許多不同方式可定義。但並非所有定義都能彼此認同！我們在這份報告中以特定方式使用了某些詞語。以下是我們對這些關鍵詞與用語的定義。

- **強勢語言**：在某個地區多數人口所使用的語言，或是透過特定過往強權與認可的形式、透過法律、政治或文化勢力稱霸的語言。比方說，相較於許多其他語言，印地語在南亞為強勢語言，特別是考量到印地語本身是一個語系，涵蓋許多語言或是有些人也稱為「方言」。同樣地，相較於其他形式的漢語及該區域的其他原住民族語言，華語因政府政策而成為中國的強勢語言。有些強勢語言在該區域或國家也是「官方」或「國家」語言。

- **歐洲殖民語言**：源於西歐，16 世紀開始，西歐國家與政府經由殖民過程擴散至非洲、亞洲、美洲、加勒比海與太平洋群島的語言。包括英語、西班牙語、法語、葡萄牙語、荷蘭語及德語。必須注意的是，這些語言不只對拉丁美洲（中美洲與南美洲）來說是「殖民者」語言，對北美原住民族來說也是。

- **全球南方與全球北方：**「全球南方」指的是遭到西歐國家殖民的非洲、亞洲、拉丁美洲、加勒比海及太平洋群島區域。不是地理名詞，而是意在反映這些國家與區域從過去到現在的社會經濟與政治條件特徵，並與歐洲及北美優勢國家有所區別，後者便是所謂的「全球北方」。該名詞由全球南方學者與社運人士創造及強化，以超越他們認為帶有貶意且令人討厭的用詞，例如「低度開發」、「開發中」國家，及「第三世界」。由於殖民也造成全球北方許多原住民族的種族屠殺或大規模毀滅，也由於全球南方有些個人及社群因參與殖民自己人而獲益，我們有時會說，全球南方裡存在全球北方，全球北方裡也存在全球南方。這些結構與過程也會影響語言在這些區域的地位。（見「被當成少數的世界多數」一詞。）

- **原住民族語言**：特定地區或地方的原住民族使用的語言便是原住民族語言。在全球許多地方遭到其他文化族群殖民及定居前，原住民族被視那些地方的「第一族群」或第一居民。世界上 7000 多種語言中，多數為原住民族部落所使用的語言。

- **語言及方言**：我們視任何人類之間有結構的表達系統為語言，無論是藉由語音、聲音、手語、手勢或書寫形式。有些語言學家將「方言」定義為用以描述聽起來像同一語言的變體，能夠「相互理解」，且不同變體的使用者能相互理解、對話。然而，由於「語言」及「方言」的差異往往取決於政治因素（而非語言學），根據過往的強權與優勢過程而定，我們在這份報告中甚少使用方言一詞。我們偏好使用「語系」，同一語系的語言可能擁有相似歷史背景但不同的特徵，例如阿拉伯語、漢語或印地語等語系。

- **當地語言**：在這份報告中，我們將當地語言定義為該國家或區域最多人口使用的語言。

- **邊緣化語言**：在這份報告中，邊緣化語言為網路上的語言支援或內容較為次要的語言，即該語言版本的資訊或知識較少。這些語言邊緣化的原因是過去到現在的強權與優勢結構及過程，包括殖民與資本主義，而非語言使用者人口數量。世界上有些邊緣化語言已經瀕危（例如許多原住民族語言）。但有些邊緣化語言卻在該區域或世界上有相當大量的使用人口，在網路上卻很少出現（舉一些強勢語言的幾個例子來說，像是亞洲的旁遮普語及坦米爾語，或非洲的豪薩語及祖魯語）。

- **少數與多數語言**：少數語言為少數（以數量而言）人口在任何指定領域或區域所使用的語言，多數語言則是該人口的多數（以數量而言）人所使用的語言。

- **被當成少數的世界多數**：過去到現在的強權與優勢結構，導致世上許多不同的社群與族群遭到歧視與壓迫。這些強權與優勢的形式環環相扣與交錯，因此有些社群因不同原因成為弱勢或受到壓迫：例如，因性別、性向、階級、種性、宗教、能力，當然還有語言。無論是網路上或真實世界，這些社群以數量來說構成世界多數人口，但往往因不具權力地位而被當成少數來對待。換句話說，他們是「被當成少數的世界多數」。

進一步了解如何引用及使用這份報告。

進一步了解我們的資源與啟發。