

Ts & Cs

- "Free" content platforms e.g. <https://policies.google.com/terms?hl=en#toc-permission>
- Can't share data back to indigenous communities because of "copyright"

Right to Monetize

You grant to YouTube the right to monetize your Content on the Service (and such monetization may include displaying ads on or within Content or charging users a fee for access). This Agreement does not entitle you to any payments. Starting November 18, 2020, any payments you may be entitled to receive from YouTube

Scope

This license is:

- worldwide, which means it's valid anywhere in the world
- non-exclusive, which means you can license your content to others
- royalty-free, which means there are no monetary fees for this license

Rights

This license allows Google to:

- host, reproduce, distribute, communicate, and use your content — for example, to save your content on our systems and make it accessible from anywhere you go
- publish, publicly perform, or publicly display your content, if you've made it visible to others
- modify and create derivative works based on your content, such as reformatting or translating it

Language as a Service

Google ASR

Feature	Standard models (all models except enhanced video and phone call)		Enhanced models (video, phone call)	
	0-60 Minutes	Over 60 Mins up to 1 Million Mins	0-60 Minutes	Over 60 Mins up to 1 Million Mins
Speech Recognition (without Data Logging - default)	Free	\$0.006 / 15 seconds **	Free	\$0.009 / 15 seconds **
Speech Recognition (with Data Logging opt-in)	Free	\$0.004 / 15 seconds **	Free	\$0.006 / 15 seconds **

used to be other way around

*We won't take your data anymore
(now that we've got all we need).
Like 1993 Apology to Hawaiians.*

Azure Services

Instance	Category	Features	Price
Standard - Web/Container 100 concurrent requests for Base model 20 concurrent requests for Custom model ¹	Speech to Text	Standard ²	\$1 per audio hour
		Custom	\$1.40 per audio hour Endpoint hosting: \$0.0538 per model per hour
		Conversation Transcription Multichannel Audio ^{review}	\$2.10 per audio hour ³
	Text to Speech	Neural ⁴	Real-time synthesis: \$16 per 1M characters ⁴ Long audio creation: \$100 per 1M characters
		Custom Neural ^{4,5}	Training: \$52 per compute hour, up to \$4,992 per training Real-time synthesis: \$24 per 1M characters Endpoint hosting: \$4.04 per model per hour Long audio creation: \$100 per 1M characters
	Speech Translation	Standard	\$2.50 per audio hour
Speaker Recognition	Speaker Verification	Speaker Verification	\$5 per 1,000 transactions
		Speaker Identification	\$10 per 1,000 transactions
	Voice Storage		\$0.20 per 1,000 voice profiles (10,000 free voice profiles per month)

Dear Mr./Mrs.,

My name is Jovan and I work for Apple Maps, Data Partnership Team.

As you may have heard, in last couple of years Apple has been working on improvements of its own maps.

Apple Maps can be found on all iOS and macOS devices in Maps application.

This application makes it easy to get around by car, bike, on foot, or on transit. Discover places with Guides or explore with Look Around.

As we are working towards increasing the quality of data in New Zealand , we want to portray Māori naming on our map.

I came across on your service here: <https://github.com/TeHikuMedia/nga-kupu>

Is it possible to get the data in some digital format?

Can you explain us please, what are the next steps to get permission from you to use the naming?

I appreciate your response.

Kia ora Jovan,

We'd love to have a conversation about how we could work together. As a precursor, I suggest you read <https://papareo.nz> in its entirety as well as <https://www.wired.co.uk/article/maori-language-tech>

Once you and your team have familiarized yourselves with indigenous data sovereignty we can begin a conversation.

Why we are angry

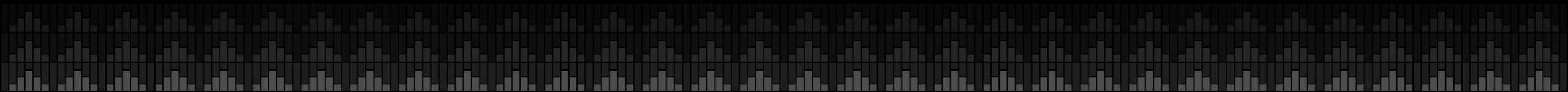
- USA bans our languages, literally beats it out of our ancestors
- US corporates now taking data from our communities
- Packaging data and reselling it back to us
- Indigenous people still landless, houseless, living in tents while Mark, Bezos, Elon, etc. buy up stolen indigenous land.

LREC 2022 Reflections

- Open data, data in public domain, and big tech "making things public for the benefit of all"
- Sharing
- Sovereignty for French language where USA & China are threats
- Desire/need for studying social media data
- Science, science, science \Rightarrow *our communities don't care about w.e.r. or BERT scores*

Practical advice

when working with indigenous peoples, their language, and their data



Outline

- Project Leadership
 - Stakeholder Engagement
 - Data Collection
 - Data Labelling
 - Data Storage
 - Security & Privacy
 - Application
- 

Essential Team Skills

- Know the language and the culture
- Speak the language
- Software/web development

Project Leadership

- Indigenous lead, otherwise shared leadership with >50% indigenous representation
- Funding - indigenous group should receive funds and distribute to others
- Own the responsibility



We can lead these projects!

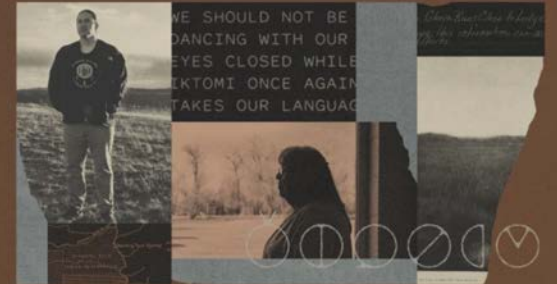
Stakeholder Engagement

- Collaborate with others from your tribe/group
 - Are you in a position to make decisions that will affect all peoples from the group?
- Have **real** representation, not tokenistic representation
- Accountability is held by the community from which the data/tools are derived

U.S. NEWS

Lakota elders helped a white man preserve their language. Then he tried to sell it back to them.

"No matter how it was collected, where it was collected, when it was collected, our language belongs to us," said Ray Taken Alive, a Lakota teacher.



— Ray Taken Alive and Gloria Runs Close To Lodge-Goggles have both accused the founder of The Language Conservancy of taking their family's data for his work.

J.D. Reeves for NBC News / Tara Rose Weston /

Data Collection

- Be transparent - why, who, how
- Options for participation - working with elders vs. millennials
- Worry about data related details later
- You can always label, you can't always collect



Data Labelling

- Adopt an iterative strategy – you won't get it right the first time
- Don't be afraid to develop your own labels
- Universal often means a euro-centric & the labels may not work for you

Te Ara Encyclopedia by Ministry for Culture and Heritage, New Zealand

Source name:
The title (this field can be left blank).

Author:
Author name or names (this field can be left blank).

Source type:
Source type is a single character (this field can be left blank). Valid source types include 'W' (Website), 'A' (Article), 'B' (Book), 'I' (Interview), 'S' (Self), 'D' (Document), 'M' (Machine).

Source url: Currently:
Change:
URL for the source (e.g. a website or API endpoint). This field can be left blank.

Description:

New Zealand's first recognisable encyclopedia was The Cyclopaedia of New Zealand, a commercial venture compiled and published between 1897 and 1908 in which businesses or people usually paid to be covered. In 1966 the New Zealand Government published An Encyclopaedia of New Zealand, its first official encyclopedia, in three volumes. Although now superseded by Te Ara, its historical importance led to its inclusion as a separate digital resource within the Te Ara website.

Te Ara was developed between 2001 and 2014 and edited by historian Jock Phillips, who oversaw a full-time staff of about 20 writers, editors, image and resource researchers and designers during its creation.[5] In 2010 during the development of the encyclopedia, the decision was made to integrate the Dictionary of New Zealand Biography into Te Ara. On completion of the work in 2014, Jock Phillips' contribution to the project was recognised with a Prime Minister's Award for Literary Achievement. The

Any extra information about the source (this field can be left blank).

Added by:

Te Ara Encyclopedia by Ministry for Culture and Heritage, New Zealand

Name:
A descriptive name for this document

Document url:
URL for the document (e.g. a website or API endpoint). This field can be left blank.

Source: [Te Ara Encyclopedia by Ministry for Culture and Heritage, New Zealand](#)

Notes:

Any miscellaneous observations about the text

Original file: Currently:
Change: No file selected.
The original document. This can be any type of file.

Added by:

Datetime updated:

Original file preview:

Data Labelling

- Prodigy – label with active learning, <https://prodi.gy/>
 - off the shelf, run on your own infrastructure, academic and non-profit license available (just email them)
- Django – python based open source web framework
 - run it anywhere and everywhere
 - automatic data migrations
 - Comes with feature rich admin & permissions


Django

- Build fast, fix later
- Iterative data development w/migrations & JSON fields
- Use APIs for data access, front end, etc
- Hosted in AWS
- Easily hook-in jupyter notebooks to DB

K Mahelona
Not you? [Logout](#)

0 0 3215 total
0 0 0
hēkona hēkona hēkona
today

Kua haere anō hoki te ahi kei waho.



Transcription: kua haere anō hoki te ahi kei waho
Word Error Rate: 0.00

[Follow Up](#) [Noise](#) [Delete](#) [Save](#)

[Auto](#) [Approve](#) [Ka Tika](#) [Kia Kaha](#) [Skip](#)

[Help on Reviewing](#)

Sort By: [Random](#) [Recent](#) [High word error rate](#)

K Mahelona
Not you? [Logout](#)

0 0
hēkona kapohanga reo

Nā, koirā te āhuetanga i mahue mokemoke mai



[Mahia anō](#) [Tiakina](#) [Skip](#)



Skip



Follow Up

Noise

Approve

Ka Tika

Kia Kaha

Delete

Incorrect



Correct

Move the slider based on how good you think the speaker's pronunciation is.

Notes

Dataset None

H	e	o	i	,		p	i	r	i		a	n	a		a		T	e		A	r	a	w	a		k	i		t	e	
0	1	2	3	4	5	7	8	9	10	11	13	14	15	16	18	19	21	22	23	25	26	27	28	29	30	32	33	34	36	37	38

K	ā	w	a	n	a	t	a	n	g	a	,		t	a	n	a		o	a	t	i	t	i	a	n	g	a		m	ō	
40	41	42	43	44	45	46	47	48	49	50	51	52	54	55	56	57	58	60	61	62	63	64	65	66	67	68	69	70	72	73	74

t	a	n	a		r	o	h	e
76	77	78	79	80	82	83	84	85

Data Storage

- Store in ***paid*** cloud services
- Store in local "clouds" or local IT infrastructure ***if practical***
- Don't store at Universities
 - Otherwise ensure indigenous stakeholders have "the keys" to the database and oversight of access protocols.
- Don't use "free" services e.g. SoundCloud, YouTube, Facebook

Data Storage

- Open Stack, openstack.org
 - Open source cloud. <https://catalystcloud.nz/>
- Local Network/Distributed Storage
 - synology.com - can backup to other locations or cloud
 - resilio.com - bittorrent tech for distributed storage
- Amazon S3

Security & Privacy

- Protocols around access
- Always use HTTPS
- Potential weak points
 - collaborator passwords ⇒ require MFA
- You can't talk data sovereignty without respecting user's privacy



Whare Kōrero App Privacy

Summary

Personal Information

We don't collect any personal information on you. Unless you contact us via email or telephone, we'll never know who you are or be able to collect any personal information.

Content Analytics

We collect anonymous statistics on the content that you view and play in our App. This helps us report back to stakeholders on the value of Māori and indigenous media, and it helps better inform what types of content our communities wish to access.

Application

- Demonstrate benefit back to community from which the data/tools were derived
- Go beyond another research project and another publication

Māori ASR Applications

Kaituhi

Automatic te reo Māori transcriptions.



All data uploaded to kaituhi falls under our [Kaitiakitanga License](#). If you're interested in using Kaituhi for your mahi, please [get in touch](#) with us.

REO MORA

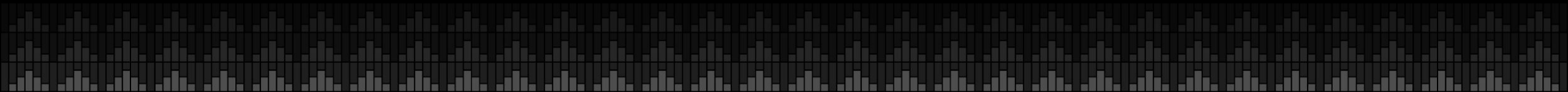
[Home](#) [Latest Resources](#) [About](#) [Testimonials](#) [Contact](#)



Learn to Speak Te Reo Māori in 12 Weeks

Real-time Pronunciation Feedback

- Read a target sentence
- Character level feedback
- On-device measurement
- 40ms per 1s of audio



Kaitiakitanga License

Kaitiakitanga License

- Uphold a set of principles that,
 - build Māori capability
 - ensure accountability and benefits back to the community from which the data/knowledge was derived
 - maintain sovereignty
- Prohibit use of data, software, tools etc. for applications that surveil, discriminate, persecute, etc.
- Affirmative action for data, software, etc.

Closing Remarks

- Sovereignty
 - Don't study us and write papers on us, empower us to build the solutions we need
 - Control/ownership/guardianship remains with community from which data is collected
- Open isn't a catchall for good
- Take an iterative approach
- Practice what you preach

tku.nz/p.G75

Links to articles and resources on data sovereignty

