Talresursinsamlaren

För ett tillgängligare Wikipedia genom Wikispeech

2019-09-01 – 2021-04-30

WIKIMEDIA SVERIGE

John Andersson & André Costa
2022-04-25, Stockholm

Vilka är vi?

WIKIMEDIA
SVERIGE

Wikimedia Sverige

- Idell förening
- Grundad 2007
- Lokalavdelning till Wikimedia Foundation
- 11 anställda
- Projektägare och kontaktpart mot PTS

WIKIMEDIA
SVERIGE

Speech Technology Services

KTH

- Avdelningen för Tal, Musik och Hörsel
- Tester och forskning

STTS

- Södermalms talteknologiservice
- Talsyntesutveckling, annotering samt manus

Dyslexiförbundet

- Referensgrupp samt utvärdering av inspelningsevent

Tidigare projekt för PTS

- Förstudie 2015
- Projektgenomförande 2016 – 2017
- Grunden lades för **Wikispeech**

WIKIMEDIA
SVERIGE

Wikispeech är en text-till-tal-lösning för Wikipedia

(och alla 1 000-tals sidor som använder MediaWiki-mjukvaran)

WIKIMEDIA
SVERIGE

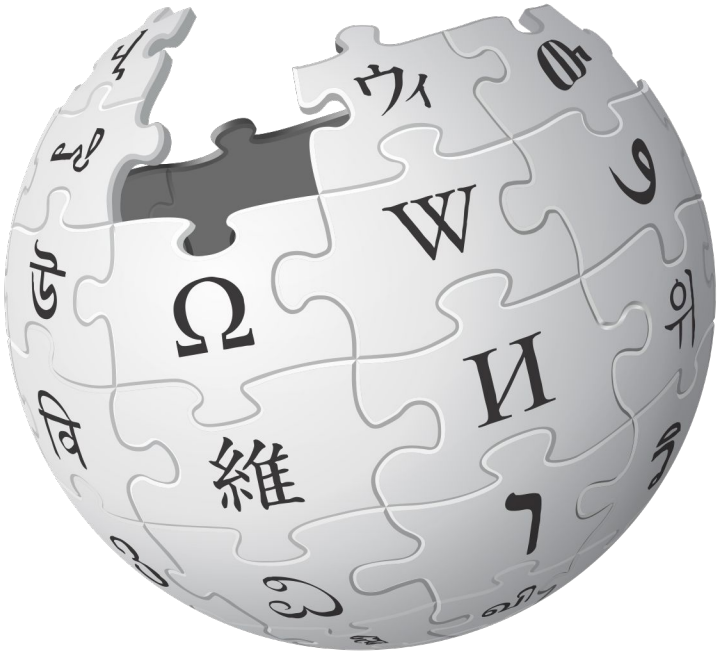
Bakgrund

Varför talsyntes på Wikipedia?

WIKIMEDIA
SVERIGE

Med Wikispeech kan vi uppnå vår vision att ge *alla världens människor* tillgång till fri kunskap

(inte bara de läskunniga)



WIKIMEDIA
SVERIGE

- 25-30 % av befolkningen lär sig bättre genom att lyssna än att läsa (s.k. *auditory learners*)[1][2]
- 13,7 % av världens befolkning är inte läskunniga, huvudsakligen i utvecklingsländer[3]
- Miljoner kan inte läsa p.g.a. funktionsvariationer
- Globalt har en bråkdel stöd från välfärdssamhället
- I Sverige är det många som har behov av ytterligare stöd

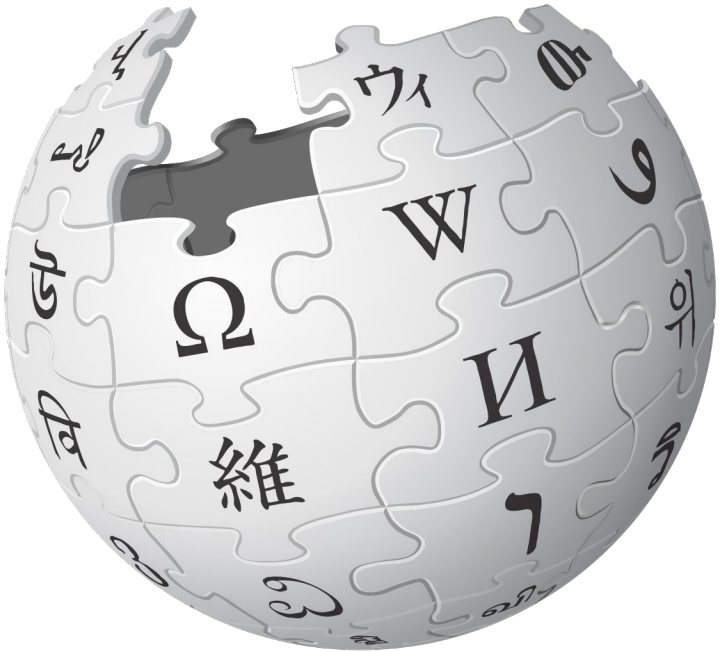
[1] https://bit.ly/2QLvo7P
[2] https://bit.ly/2qxKhMw
[3] https://bit.ly/1EFQDwP

WIKIMEDIA SVERIGE

Målning av Harold Copping, PD

Tillgänglighet på Wikipedia är viktigt:

- 310+ språk
- 265 miljarder sidvisningar 2021[1]
- 85 % av internetanvändarna i Sverige nyttjar Wikipedia[2]
- Snitt på ca 90 sidvisningar per person och år i Sverige på svenska[3] (det dubbla om alla språk räknas in)



[1] https://bit.ly/WikiStats2021
[2] https://bit.ly/SoI2018
[3] https://bit.ly/2SH1zCR

WIKIMEDIA SVERIGE

Säreget för Wikipedia

- Höga krav på integritet och öppenhet.
- Innehållet är i ständig förändring.
- Ej avgränsat ämnesområde och många special- eller lånord.
- En gemenskap som är van vid att kunna åtgärda brister.
- Alla bidrag är fria att återanvända.

WIKIMEDIA
SVERIGE

Snowflake
av Alexey Kljatov, CC BY 2.0

Bakgrund

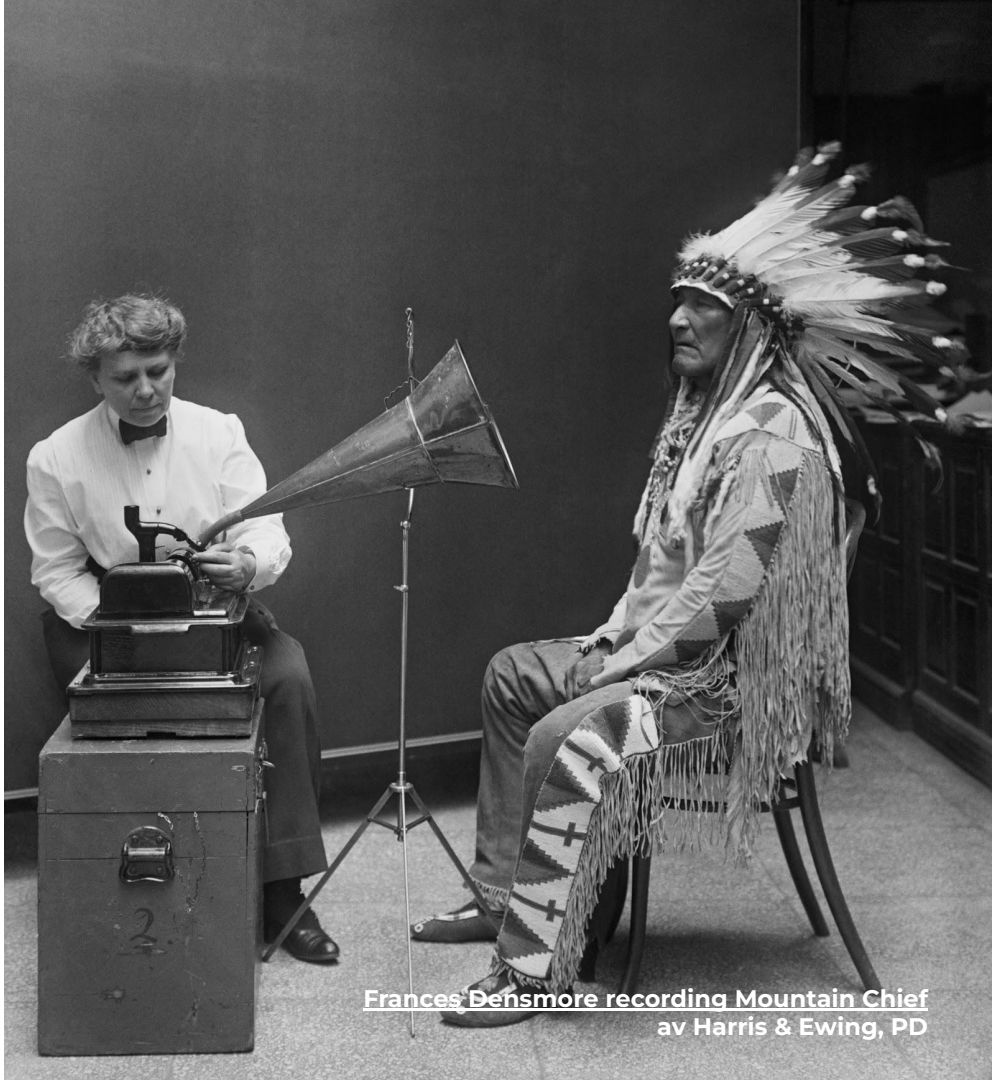
Insamling av Taldata

WIKIMEDIA
SVERIGE

Insamling av taldata

- För att svenska och andra språk ska fungera bättre som talsyntes
- För fler språk i Wikispeech – särskilt små språk av begränsat kommersiellt intresse. Viktigt för buy-in från gemenskapen.
- Fritt licensierade inspelningarna av värde för öppen programvara (FOSS) – inte bara ett verktyg utan en tjänst

WIKIMEDIA
SVERIGE



Frances Densmore recording Mountain Chief
av Harris & Ewing, PD

Kort Projektbeskrivning

WIKIMEDIA SVERIGE

- Vi bygger en värdefull resurs för talteknologi på svenska och gör Wikipedia mer tillgängligt.
- Vidareutveckla Talsyntesen för att lansera på Wikipedia.
- Utvecklar verktyg för att via crowdsourcing samla in stora mängder fritt licensierade inspelningar, samt tillhörande annoteringar, på svenska.
- Resultatet fritt återanvändbart.

WIKIMEDIA
SVERIGE

Don Adams
av General Artists Corporation, PD

Talsyntesen motiverar gemenskapen

- Wikipediagemenskapens motivation till att bidra med taldata är Talsyntesen.
- Talsyntesen naturliga ingången till Talresursinsamlaren.
- Därför viktigt att få den tidigare byggda Talsyntesen aktiverad på Wikipedia.
- Reflekteras i senare revideringar av projektplanen.

WIKIMEDIA
SVERIGE

His Master's Voice
av Olof E. Caskin, PD

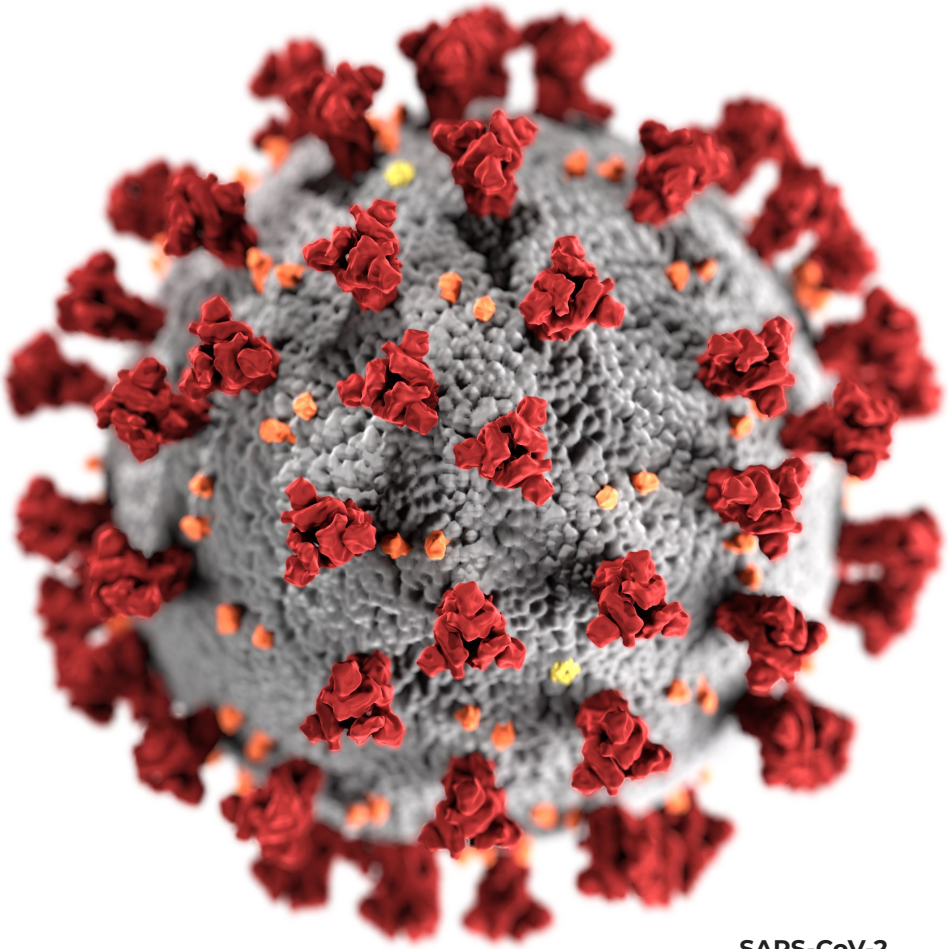
Projektresultat

WIKIMEDIA
SVERIGE

Projektförändringar

- COVID-19
- Tidigarelade utveckling av Talsyntesen
- Två centrala implementeringsval i första versionen av Talsyntesen problematiska
- Försenad kod-/säkerhetsgranskning
- Organisatoriska förändringar hos Wikimedia Foundation
- Förändrat budskap om aktivering av Talsyntesen

WIKIMEDIA SVERIGE

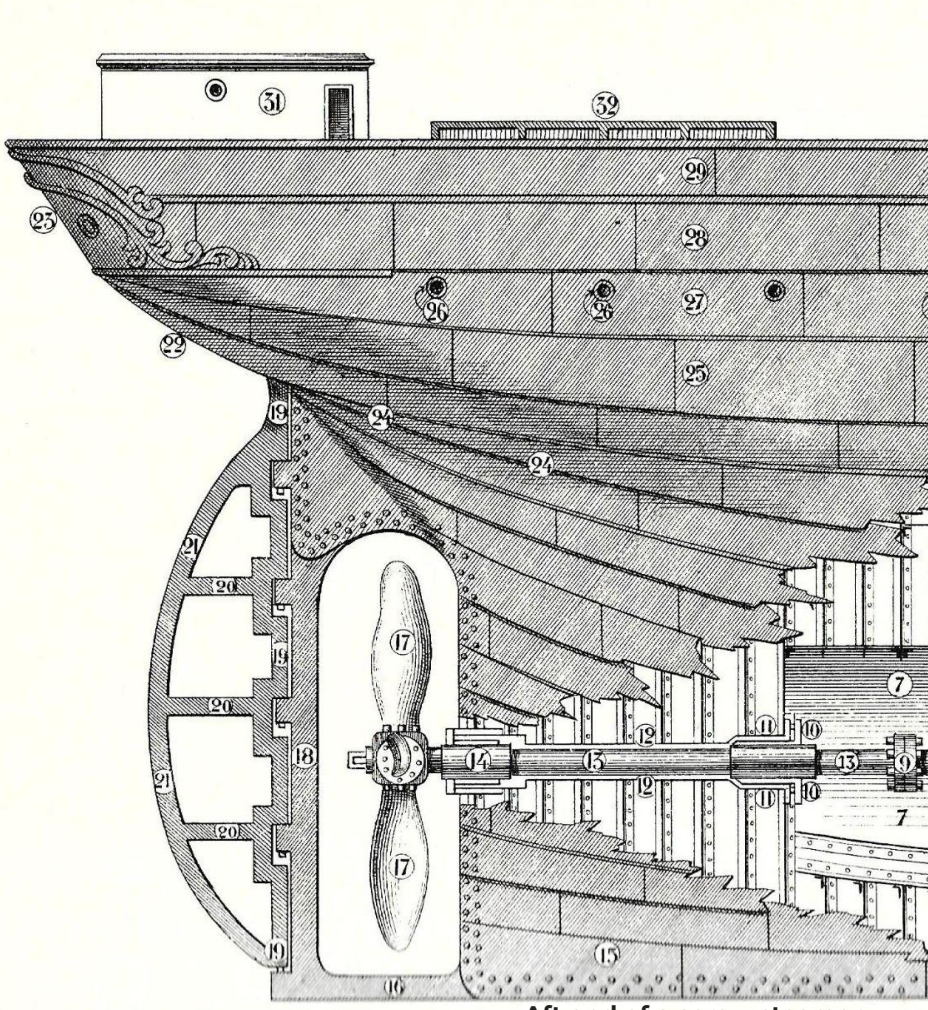


SARS-CoV-2
av Alissa Eckert & Dan Higgins (CDC), PD

Talsyntesen

- Mjukvara anpassad för att uppfylla Wikimedia Foundations krav
- Arkitekturen ändrad i grunden för att anpassa till modernare krav.
- Genomgick kod- och prestandagranskning
- Ompaketerad för att göra den enklare att driftsätta för både stora och små installationer

WIKIMEDIA
SVERIGE

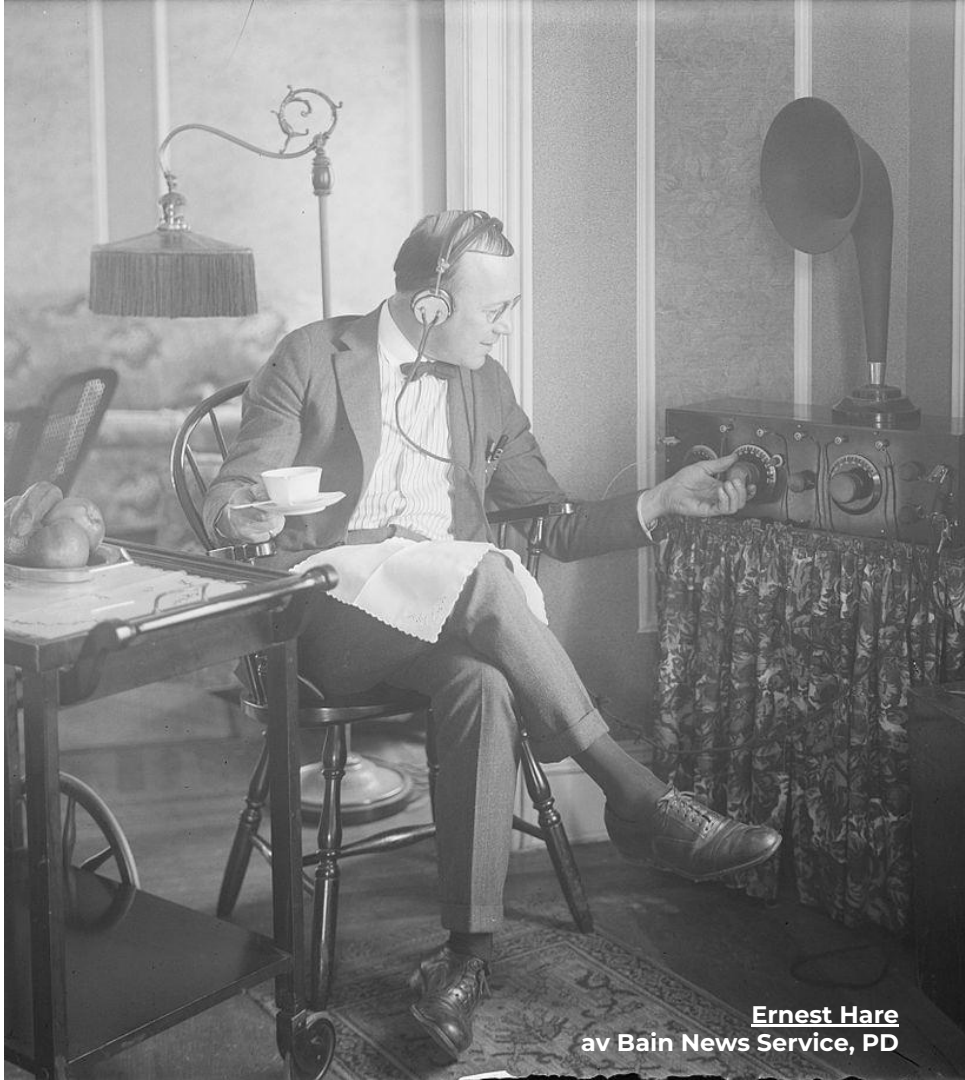


Aft end of a screw steamer
av Heinrich Paasch, PD

Talsyntesen 2

- Finns nu en färdig Talsyntes redo att aktiveras på Wikipedia
- Öppet redigerbart uttalslexikon som kan nyttjas av andra.
- Byggt så att Talresursinsamlaren kan kopplas in direkt därifrån för t.ex. rättnignar

WIKIMEDIA
SVERIGE



Ernest Hare
av Bain News Service, PD

Talresursinsamlaren

- Grundstomme för plattform på plats, byggd på MediaWiki
- Arkitektur byggd för att även ta höjd för funktionalitet som planeras under senare steg, t.ex. metadatasättning av dialekter
- Diskussioner med, och analys av, andra initiativ för att identifiera styrkor/svagheter
- Behöver ej WMFs godkännande för aktivering

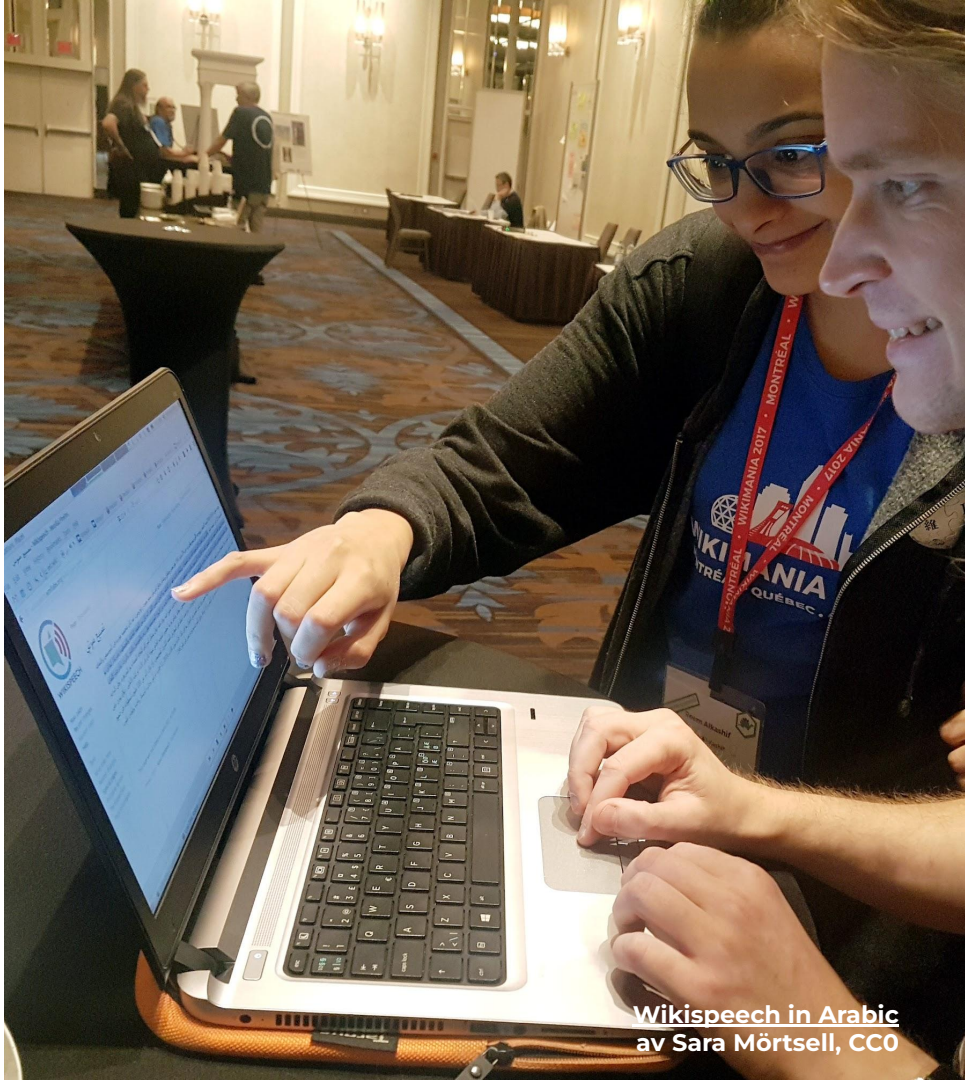
WIKIMEDIA
SVERIGE

Microphone
av Rick Harris, CC BY-SA 2.0

Tester

- Pandemin omöjliggjorde initiala testplaner
- KTH utvärderade tre alternativa webbaserade metoder
- Landade slutligen i en valideringsmetod som bygger på ARS (Audience Response System)

WIKIMEDIA
SVERIGE



Wikispeech in Arabic
av Sara Mörtsell, CC0

Manuskomponenten

- Kan automatiskt skapa effektiva manus som maximerar den nytta man får av en inspelningsinsats
- Utgår från all text på Wikipedia
- Komponenten är fristående och kan därmed fritt användas av andra i deras mjukvarulösningar

WIKIMEDIA
SVERIGE

Bibelsida från 1468
av Okänd, PD

Annotering av taldata

- STTS utvecklade flera komponenter för validering och annotering och av inspelat tal

- Talresursinsamlaren har stöd för att användare betygsätter varandras inspelningar

- Gränssnitt saknas i båda fallen

WIKIMEDIA
SVERIGE

Taldatainsamlingsevent

- Genomförde ett experimentellt evenemang med Dyslexiförbundet
- Utvecklade och publicerat gemensamt toolkit för event kring taldatainsamling
- Via en projektportal och direktkontakter har intresse skapats bland volontärer

WIKIMEDIA SVERIGE

Vidareanvändning och lagring

- Strukturering av ljudfiler på Wikimedia Commons
- Lexikala data på Wikidata

- Intern lagring av taldata
- KTH och Språkbanken Tal
- Mozilla Foundation och Common Voice

WIKIMEDIA SVERIGE

Tanked
av R. Miller, CC BY 2.0

Sedan projektslutet

WIKIMEDIA
SVERIGE

Myndighetsinventering

Inventerat svenska myndigheters användning av MediaWiki och intresse av Wikispeech

- Begränsad kännedom om MediaWiki
- Stort intresse för Wikispeech
- Wikispeech som tjänst, oaktat MediaWiki

Nästa steg, tekniskt event/workshop med efterföljande fallstudie.

WIKIMEDIA
SVERIGE

MediaWiki + Wikispeech
av André Costa, CC BY-SA 4.0

Finess aktiverad

- Finess som alternativ till MediaWiki-tillägg
- Aktiverad juli 2021 på sv.wp
  - Alla med konto kan pröva, se Wikipedia:Wikispeech
- Återkoppling mottagen och implementerad

- Direkt lexikonförbättringar av gemenskapen innan sommaren

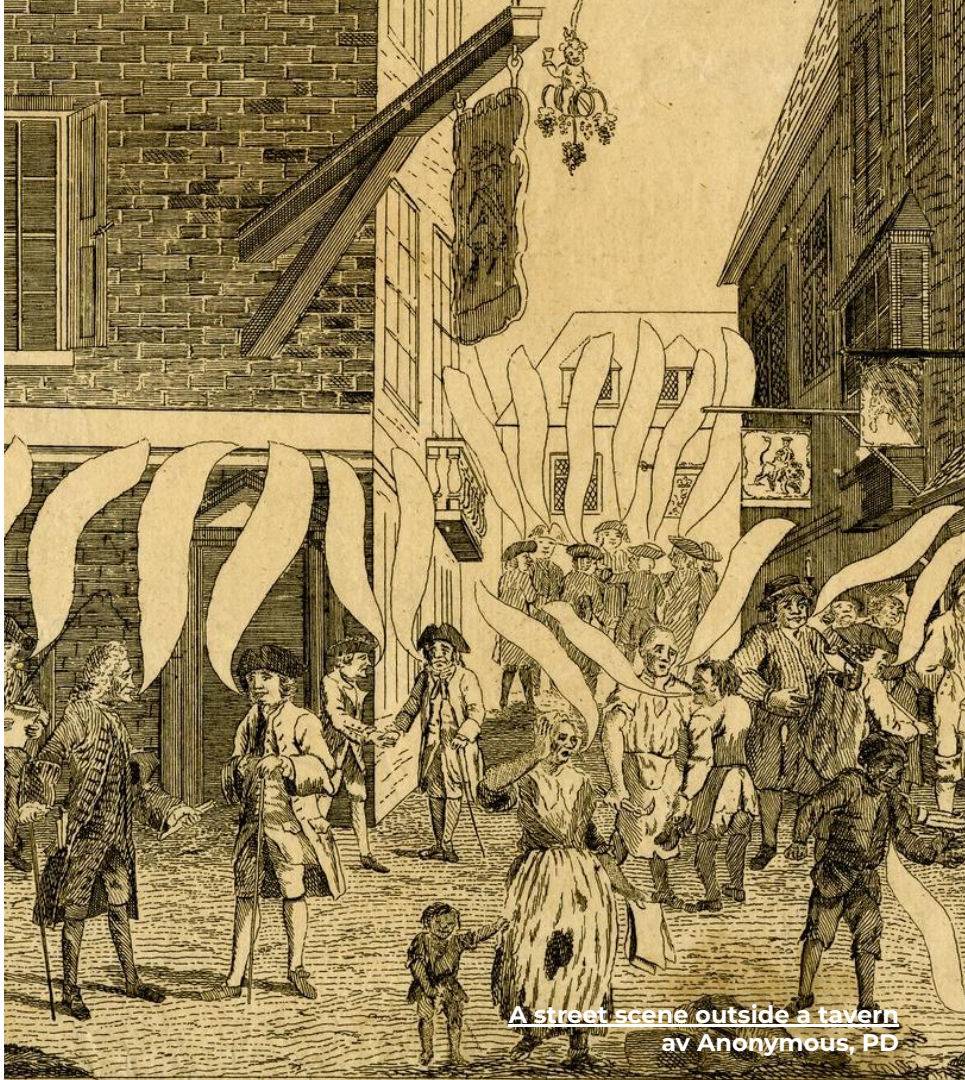
WIKIMEDIA SVERIGE

Förrendering av mest lästa artiklar

- För att bättre kunna hantera stora mängder användare
- Förrenderar tal utifrån
  - centralt länkade artiklar
  - aktivt redigerade artiklar

- Kan användas för att förrendera allt tal på en mindre wiki, t.ex. För en myndighet

WIKIMEDIA
SVERIGE



A street scene outside a tavern
av Anonymous, PD

Framtiden

WIKIMEDIA
SVERIGE

Under 2022–2023

- Naturligare röster med KTH
- Söka projektmedel för vidareutveckling och färdigställa Talresursinsamlaren
- Wikimedia Foundation tar över driften
- Wikispeech lanseras på fler språk tack vare Talresursinsamlaren

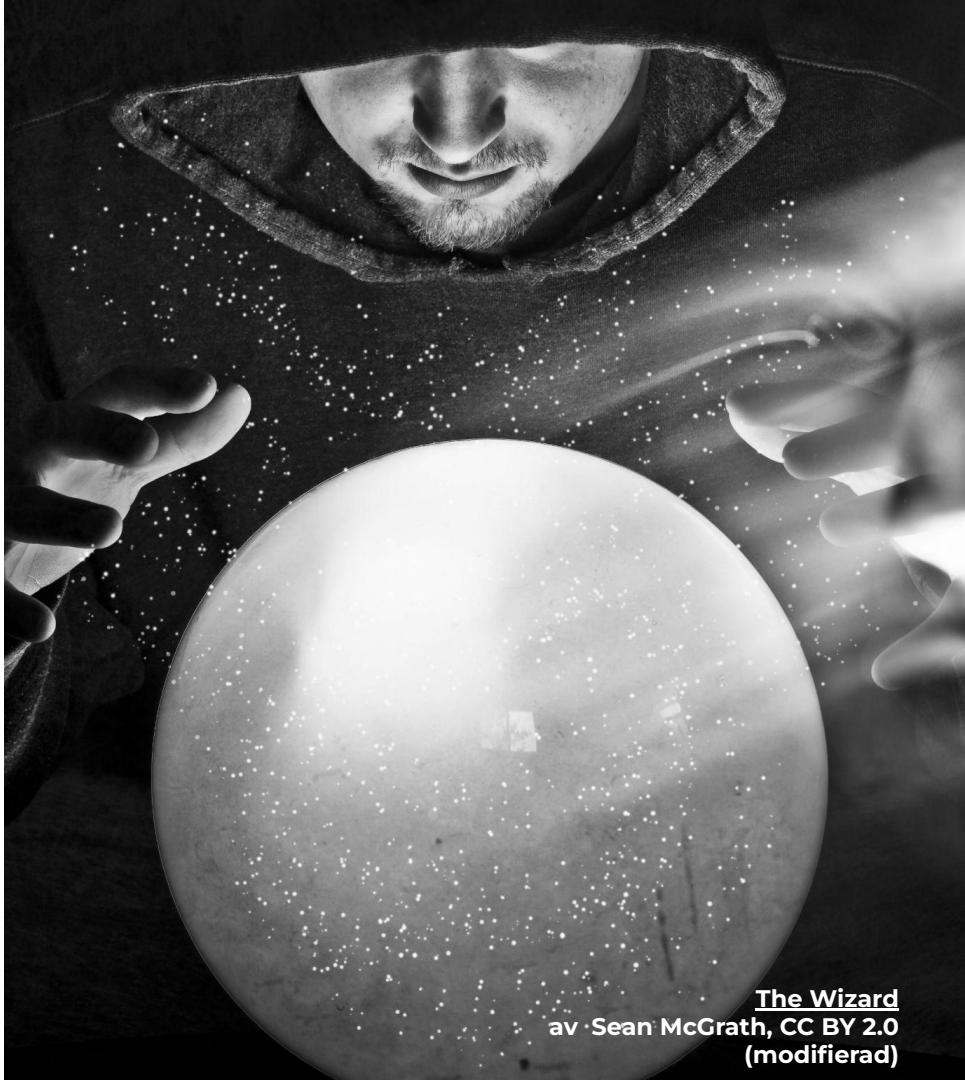
WIKIMEDIA
SVERIGE

Women Watching Stars
av Chōu Ōta, PD

Framtiden...

- Ytterligare projektmedel för insamling av värdefulla taldata
  - Dialekter
  - Information i krissituatoner
  - Taldata från minoriteter
- Stödja utveckling av neutral AI
- Integrering med Wikidatas lexikografiska data
- Användning och drift av Wikispeech inom svenska myndigheter

WIKIMEDIA
SVERIGE



The Wizard
av Sean McGrath, CC BY 2.0
(modifierad)

Tack!
Frågor?