

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?

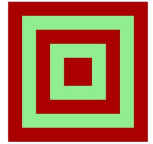
Wikimedia Sverige

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



WIKIMEDIA
S V E R I G E





STTS

Speech Technology Services



Dyslexiförbundet

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**



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?

Med Wikispeech kan vi
uppnå vår vision att ge
alla världens människor
tillgång till fri kunskap
(inte bara de läskunniga)



- 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>

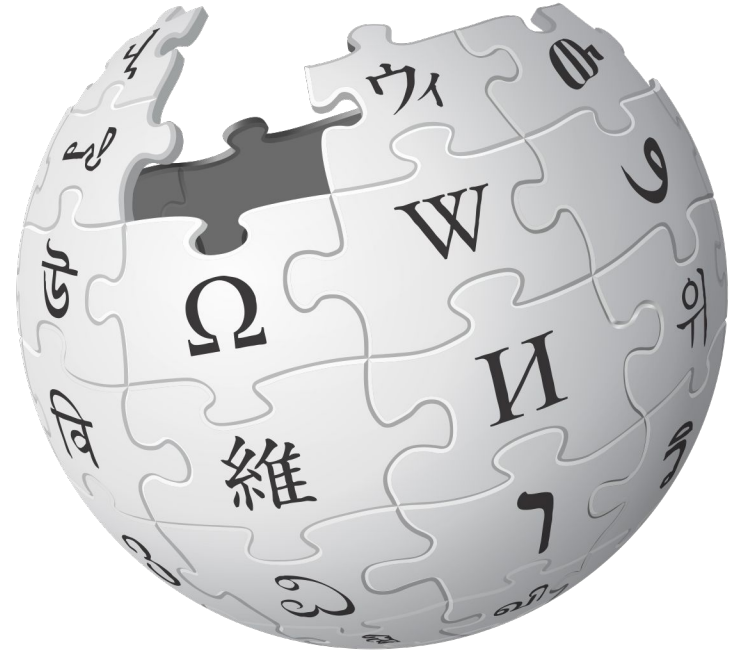
[2] <https://bit.ly/2qxKhMw>

[3] <https://bit.ly/1EFQDwP>



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/Sol2018>

[3] <https://bit.ly/2SH1zCR>

Säregnet 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.



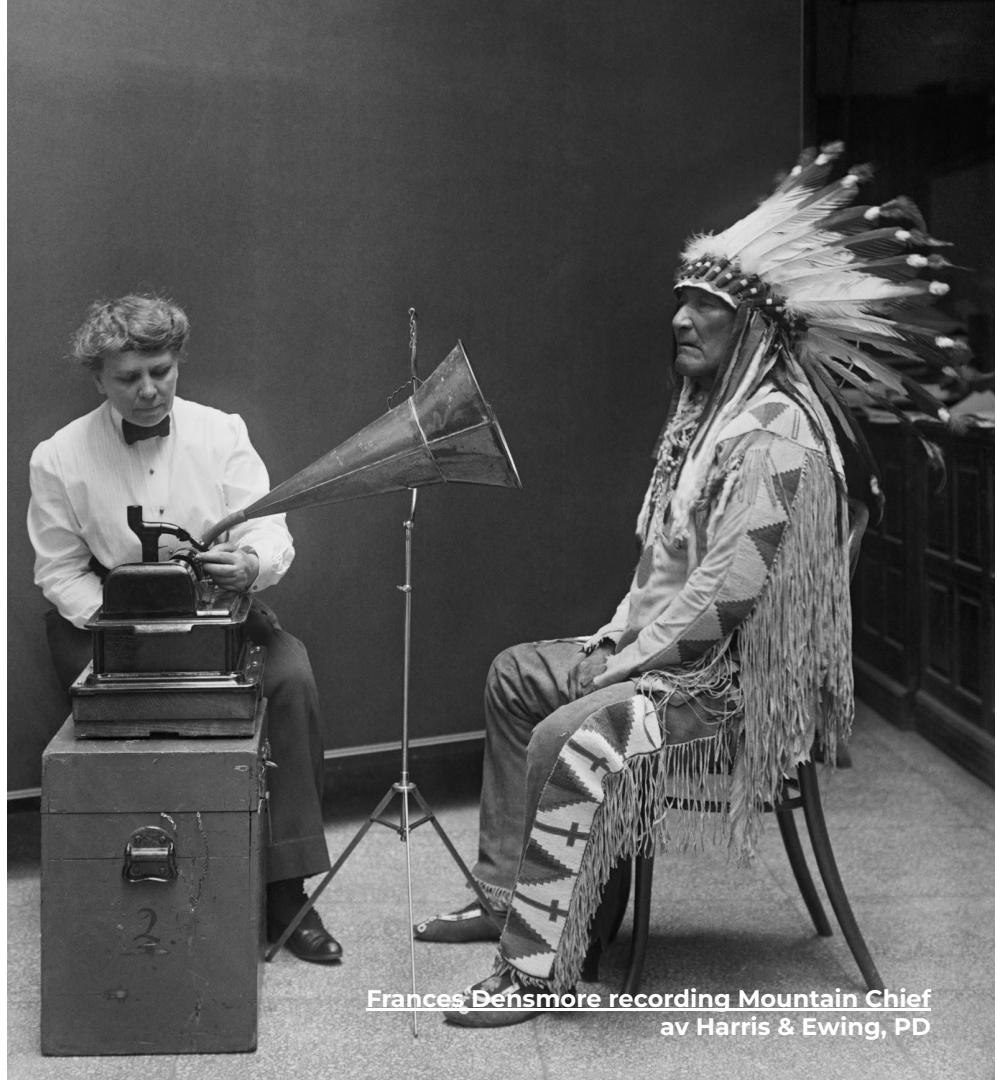


Bakgrund

Insamling av Taldata

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



Frances Denimore recording Mountain Chief
av Harris & Ewing, PD

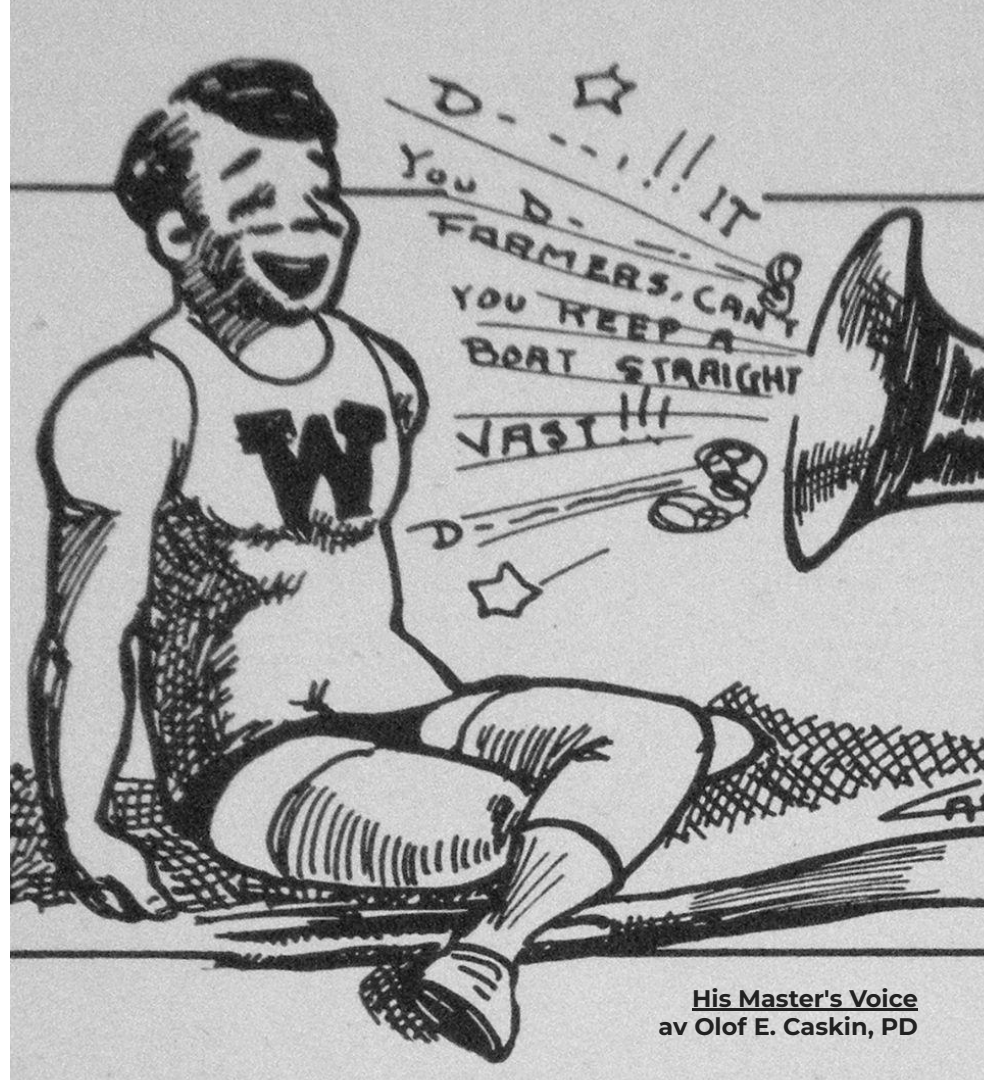
Kort Projektbeskrivning

- 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.



Talsyntesen motiverar gemenskapen

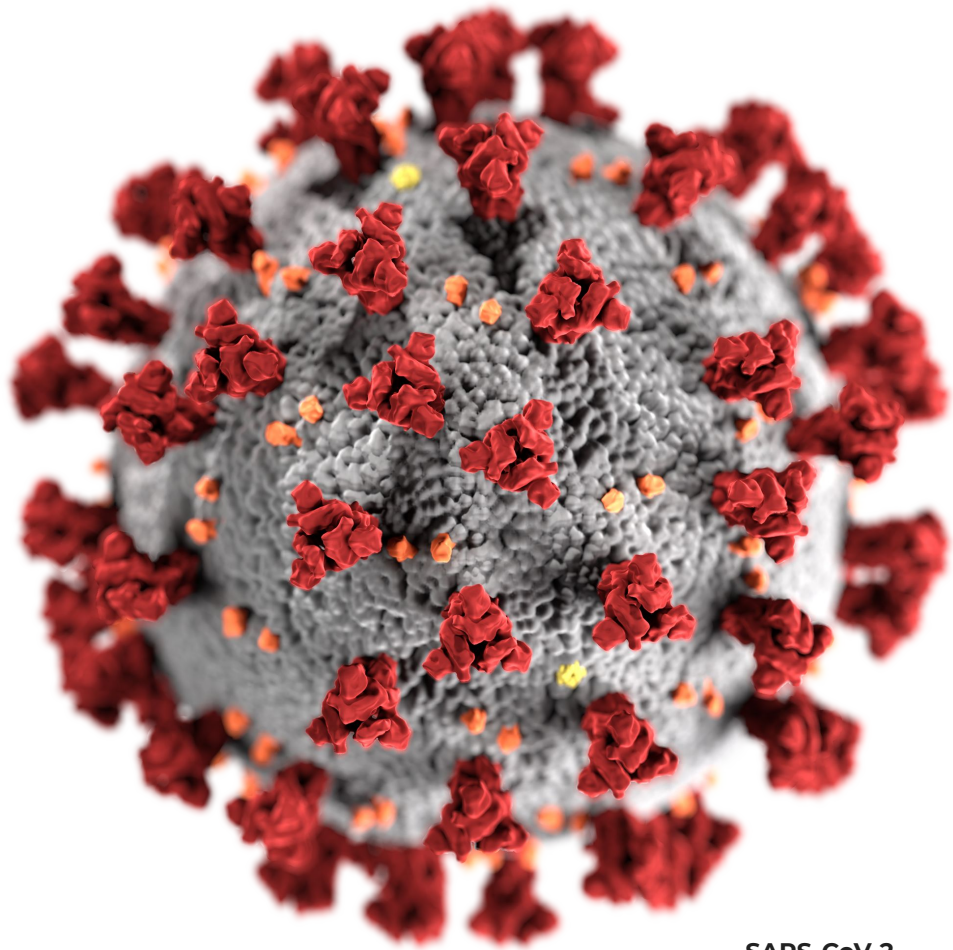
- Wikipediagemenskapens motivation till att bidra med taldata är Talsyntesen.
- Talsyntesen naturliga ingången till Talresursinsamlaren.
- Därför viktigt att få den tidigare byggda Talsyntesen aktiverad på Wikipedia.
- Reflekteras i senare revideringar av projektplanen.



Projektresultat

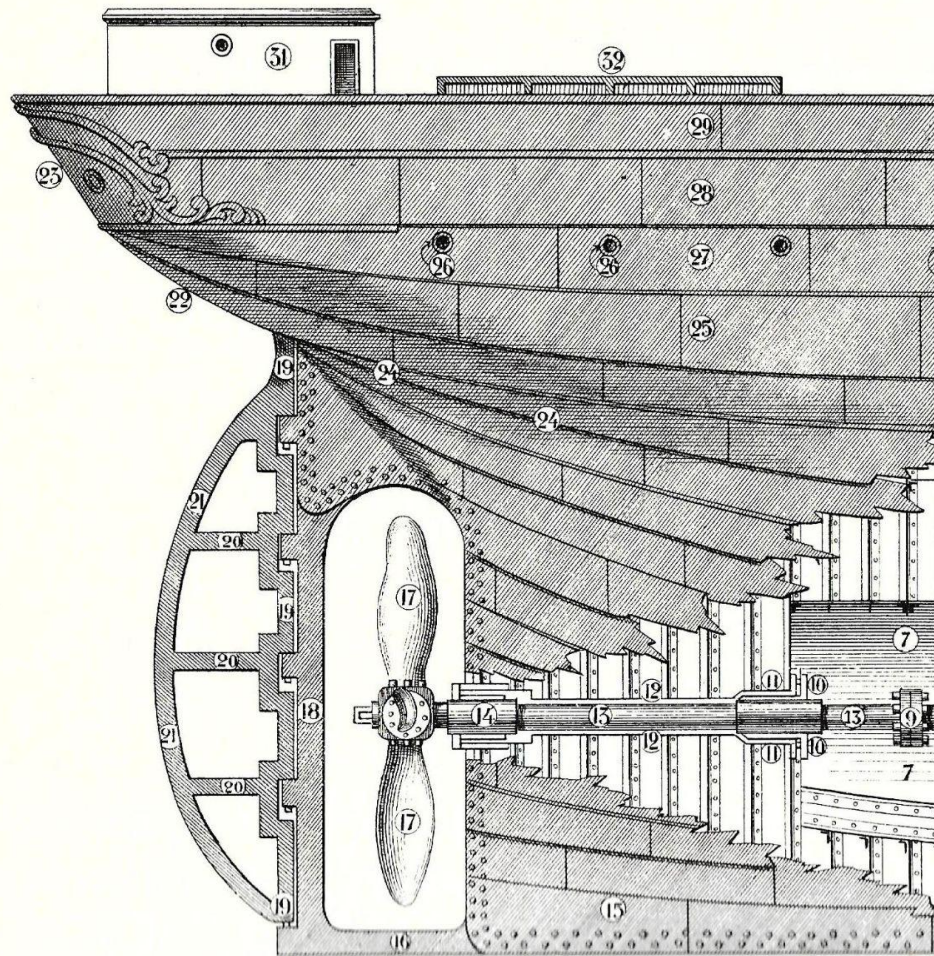
Projektförändringar

- COVID-19
- Tidigare relaterad 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



Talsyntesen

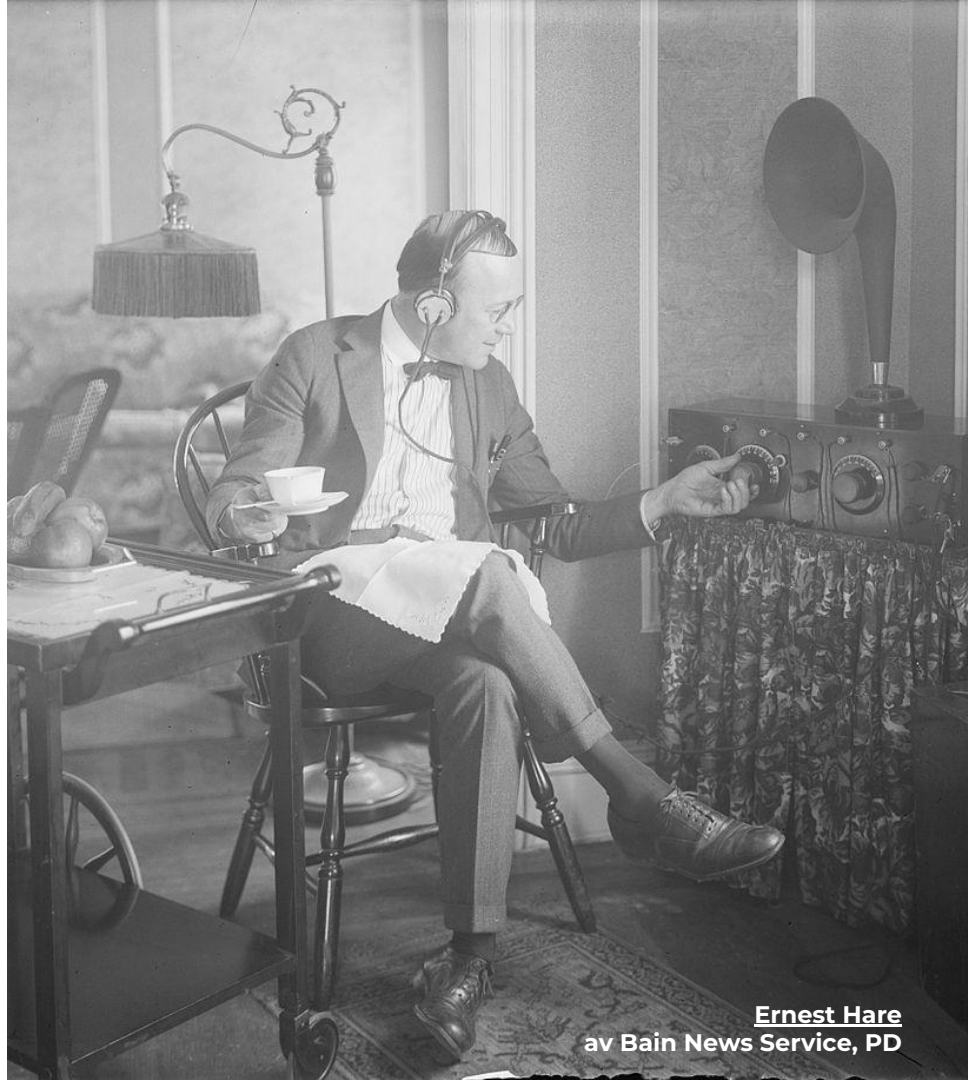
- Mjukvara anpassad för att uppfylla Wikimedia Foundations krav
- Arkitekturen ändrad i grunden för att anpassa till modernare krav.
- Genomgick kod- och prestandagranskning
- Ompaketerad för att göra den enklare att driftsätta för både stora och små installationer



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. rättningar



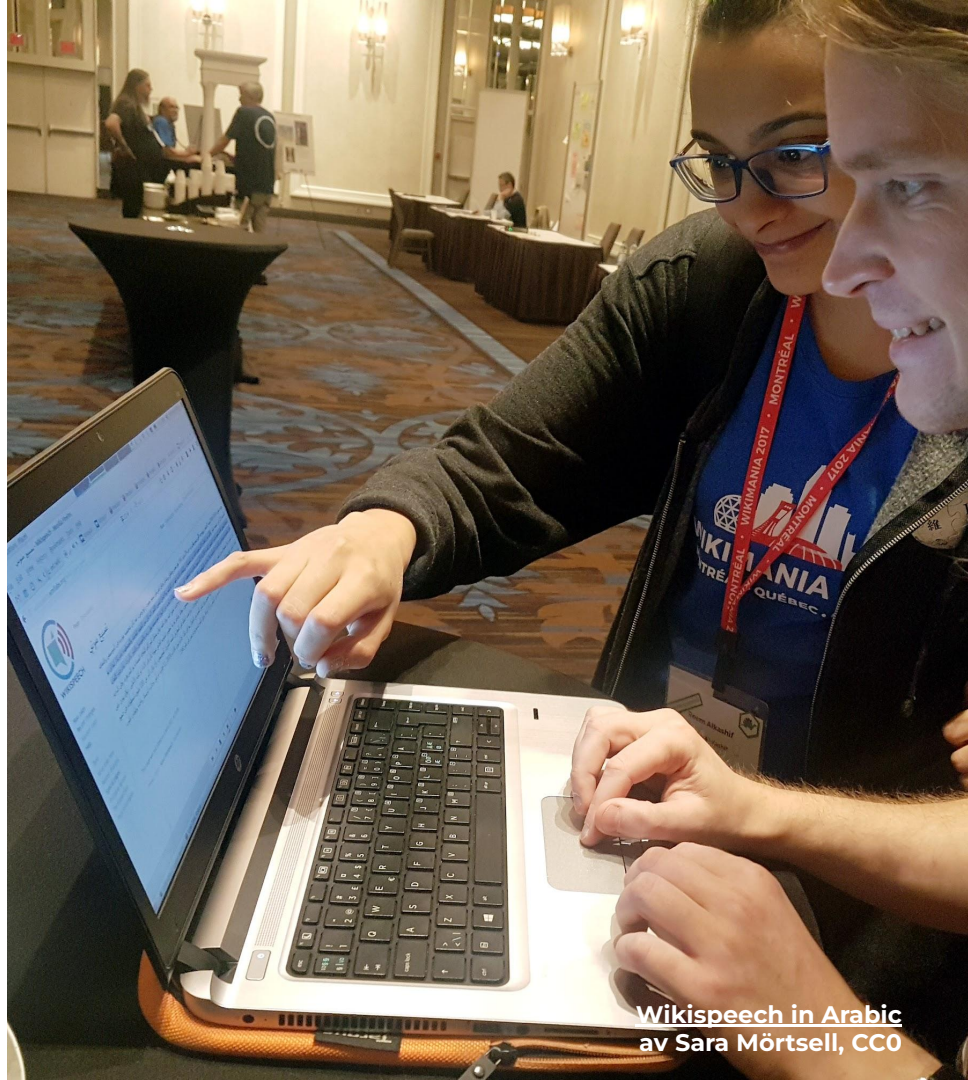
Talresursinsamlaren

- 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



Tester

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



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



hier beghint dat prologe op die ewaige

Mijn wi come toten werlt menschen en volgen hier na die ewangeliē. Hier om dat wi die ieste vande historie volgen willen soe sullen wi hier besaue die ewangeliē.

en die men hiet conuordanen. dats velen vier ewangelisten een ewangeliē gemaker bi oecode. En soe waer dat elck ewangeliē yet seint sonderlinges. oftē dat si alle alleens spreken. dat sullen wi tekenen mitter iester letteren van hoen name int wden ynet. Daer om salmen weete dat der ewangelisten vier sijn. dat is te weten. iharheus dien tekenmen na den menschen. om dat hi sijn ewangeliē beghint van xpc gheslachtet na den mensche. matheus is die ander. die tekenmen bi den leue. om dat hi sijn ewangeliē beghint vanden zopen inder iuders msten. die derde is iudas. die tekenmen na enen calue. om dat hi sijn ewangeliē beghint vander sacrificien diene in den tempel gode dede. Die vierde is sint ian. die tekenmen na enen adriē om dat hi hoerliker inder godliker bi standens ghesclommen was dan die andē. Doe dat hi in den beghint tivoert sach. dat was die soen in den bade. Want bouen alle voghelen soe vernuch die oren in de sonne sonder ymrogghen sijn. En alsoe saen als hi sijn ionghen ghekent heeft. soe hout hi se mit den

oghen inden schijn vander sonnen. en die dan ymrogghen die laet hi doet valle en die met en ymrogghen voedet hi op. En aldus sach sint ian sonder ymrogge in die sonne. Dat was hi schouwede son der comen mit gode wesen. soe dat hi sijn ewangeliē vander godheit beghint. so dat hi seide. In den beghint was tivoert. Want had hi yet hogher ghesproke. alle die woght en hadde en hadde hem niet mogghen verstaen. Want nu en was soe groot cler. die dat woert inden beghint y volmaectlic ontvinden coniste. Ende hier bi vloecht sint ian bouen dander. In desen ewangeliē sellen wi seite telken stede datter historie off den woerde toe behogen sal. alsoe wi voer gheuen heb ben. hier besuut sint ians ewangeliē. g.

Inden beghint was dat tivoert. ende tivoert was mit gode. ende god was tivoert. dat was inden beghint mit gode. Alle dinc warden mit dien woerden ghemact. en sonder dat woert en is niet ghemact. Dat ghemact is dat was bouen in hem. en dat leuen was der menschen licht. En dat licht scheide inden donckerheden. Ende die donckerheden beuigen niet. op sinte ians bapn ten auende. iii.



Ides conmes herodes dage in iuden was een pape. wos name was. Garthans van abias behoeten. Ende sijn wif was van aazons dochteren. Ende hoer name was Eliza beth. Want si waren beide gheuerlich voer gode. En si warden in alle he ven ghebode. ende sijn gheuerlicheden sonder beslaghen. Ende si en hadde geen kint. om dat elzaberch ondrachtich was en om dat si beide oter ghegheuen waren

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

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



S a m l a i n
r ö s t e r !



Att ordna arrangemang
för taldatainsamling

Toolkit Samla in röster
av Josefine Hellroth Larsson, CCO

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



Sedan projektslutet

Myndighetsinventering

Inventerat svenska myndigheters användning av MediaWiki och intresse av Wikispeech

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

Nästa steg, tekniskt event/workshop med efterföljande fallstudie.



Finess aktiverad

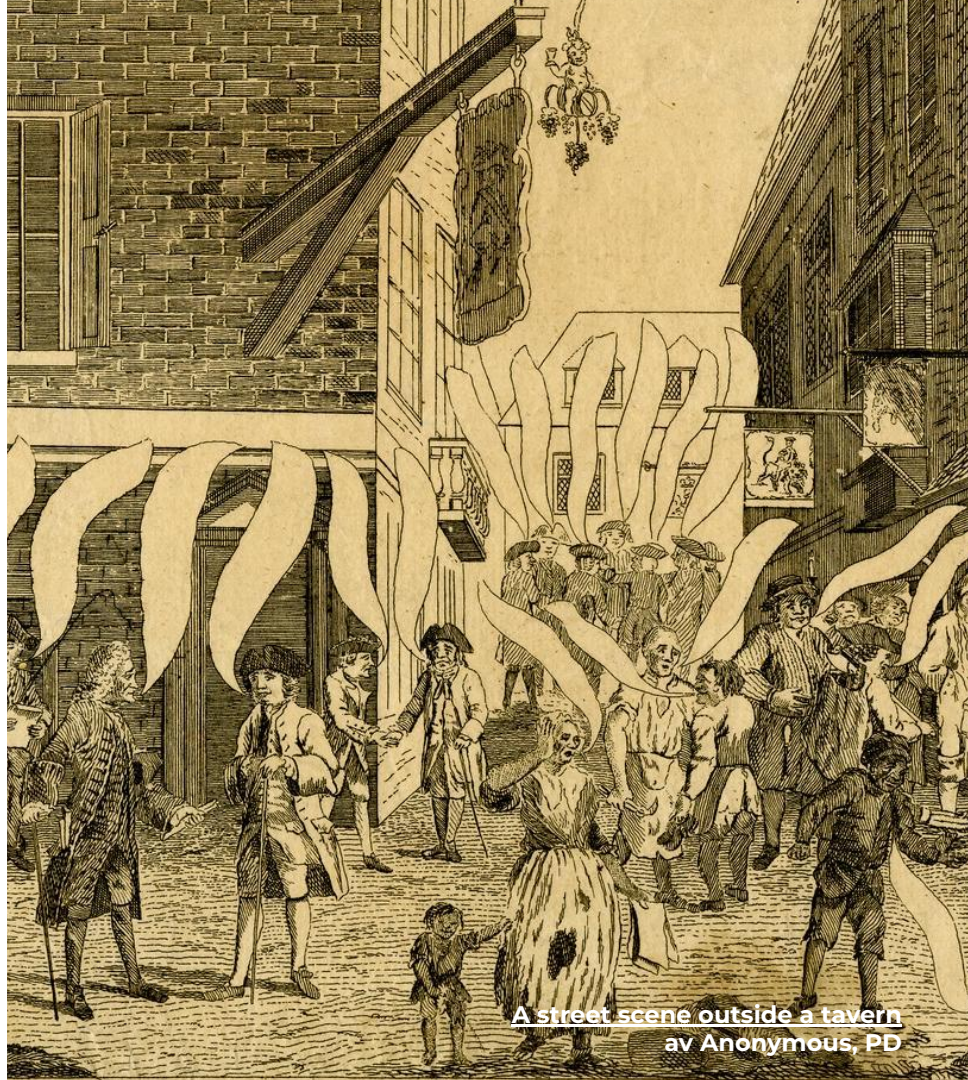
- Finess som alternativ till MediaWiki-tillägg
- Aktiverad juli 2021 på sv.wp
 - Alla med konto kan pröva, se [Wikipedia:Wikispeech](#)
- Återkoppling mottagen och implementerad
- Direkt lexikonförbättringar av gemenskapen innan sommaren



The screenshot shows the Swedish Wikipedia page for "Wikimedia Sverige". At the top, there is a navigation bar with the Wikipedia logo, the text "WIKIPEDIA Den fria encyklopedin", a search bar, and user information for "André Costa (...)". Below the navigation bar, there are tabs for "Artikel" and "Diskussion", and a "Läs" button. The main content area has a heading "Wikimedia Sverige" with a language selector "3 språk". The text describes Wikimedia Sverige (WMSE) as a Swedish ideell förening with its office in Stockholm, dedicated to spreading free knowledge. It mentions its affiliation with the Wikimedia Foundation and its focus on projects like Wikipedia, Wikimedia Commons, and Wikidata. A table of contents is visible on the left, listing sections like "Verksamhet", "Historik", "Rättsfall", "Funktionärer", "Utmärkelser", "Referenser", and "Externa länkar". On the right, there is a sidebar with a summary of the organization, including its type (ideell förening), purpose (sprida fri kunskap), address (Birger Jarlsgatan 57C, 113 56 Stockholm), founding year (2007), and membership (505 as of December 2019). The sidebar also lists the board members (Ordförande: Mattias Blomgren, Ledamöter: Johanna Berg, Sven-Erik Jonsson, Bengt Oberger, Ylva Pettersson, Brit Stakston, Per Hasselberg, Sofie Jansson).

Förredering av mest lästa artiklar

- För att bättre kunna hantera stora mängder användare
- Förrederar tal utifrån
 - centralt länkade artiklar
 - aktivt redigerade artiklar
- Kan användas för att förredera allt tal på en mindre wiki, t.ex. För en myndighet



Framtiden

Under 2022–2023

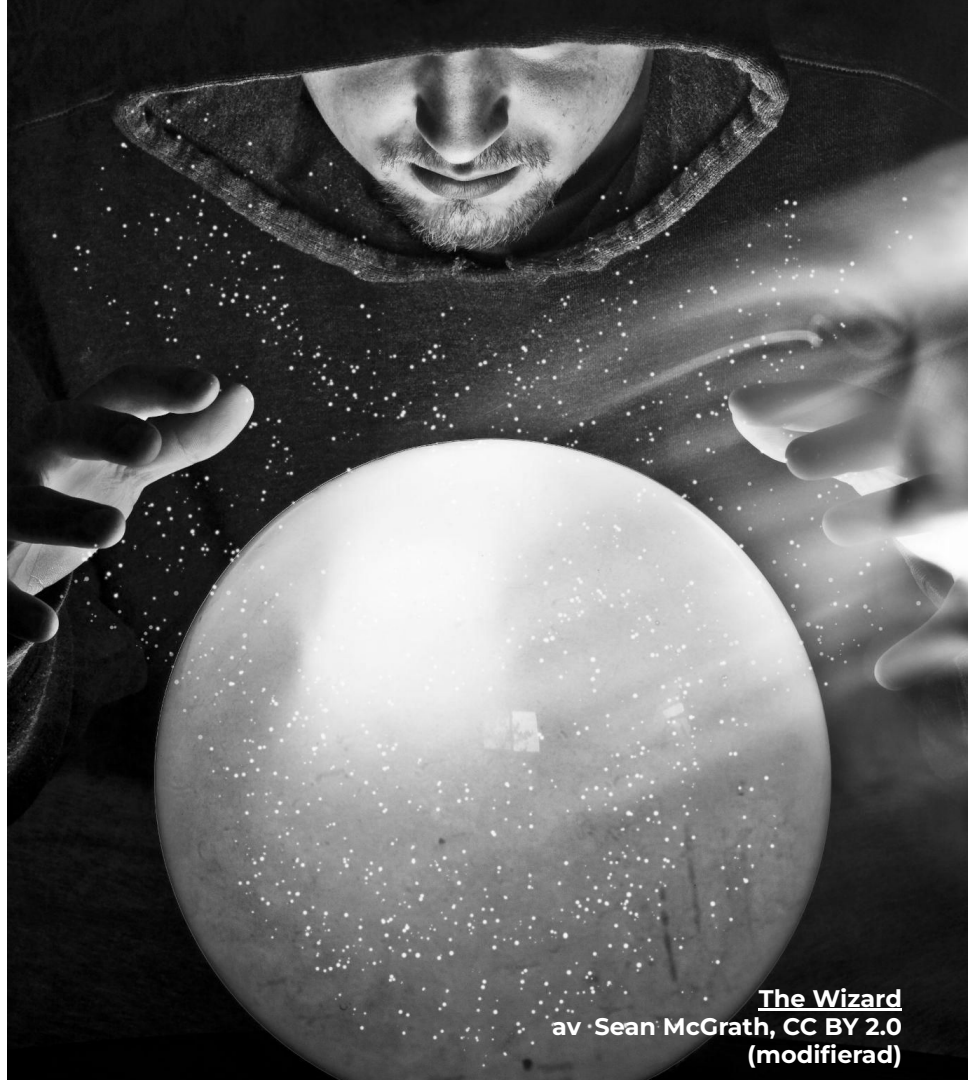
- Naturligare röster med KTH
- Söka projektmedel för vidareutveckling och färdigställa Talresursinsamlaren
- Wikimedia Foundation tar över driften
- Wikispeech lanseras på fler språk tack vare Talresursinsamlaren



Women Watching Stars
av Chōu Ota, PD

Framtiden...

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



**Tack!
Frågor?**



WIKIMEDIA
SVERIGE