

# Etapprapport

## - för PTS innovationstävling

---

<b>Projekttitel:</b>	Talresursinsamlaren – För ett tillgängligare Wikipedia genom Wikispeech
<b>Projektorganisation:</b>	Wikimedia Sverige, med stöd av STTS AB och KTH.
<b>Etapprapport för etapp nr:</b>	1
<b>Författare av detta dokument:</b>	André Costa
<b>Datum för skrivande av etapprapport:</b>	2020-03-11
<b>Version av detta dokument:</b>	2.0
<b>Bilaga:</b>	Bilaga 1: Budget och etapper

## Innehållsförteckning

<b>1. Instruktioner</b>	<b>1</b>
<b>2. Etappprogress</b>	<b>1</b>
2.1 Aktiviteter, utfall och avvikelser	1
2.2 Avvikelser från plan	3
2.3 Kostnader för etappen	4
2.4 Händelser under perioden som etapprapporten avser	4
<b>3. Uppdaterad bilaga – Budget och etapper</b>	<b>5</b>
3.1 Kommande händelser	5
3.2 Uppdaterad riskanalys	5

## 1. Instruktioner

Detta dokument fylls i vid rapportering av respektive etapp och avser den aktuella etappen.

## 2. Etappprogress

I detta kapitel beskrivs hur den redovisade etappen har gått och eventuella händelser/avvikelser samt tillhörande anledningar/orsaker.

### 2.1 Aktiviteter, utfall och avvikelser

Nedan beskrivs aktiviteter och utfall från etappen.

Hämta aktiviteter från senaste uppdaterade Projektplan. Redovisa progress/utfall och avvikelser i tabellen nedan. Utfallet ska redovisas i % av utförd aktivitet (OBS! inte % av mängden förbrukade timmar).

Om det finns "rester" av aktiviteter från föregående etapp/etapper så lägger ni dem i tabellen "Kvarvarande aktiviteter från föregående etapper" och redovisar på samma sätt.

Här följer ett exempel på hur detta kan se ut. **OBS! Texter skrivna kursivt ska tas bort från dokumentet och exempeltexter ska ersättas med era egna uppgifter.**

<i>Etapp 2: Utveckling och förberedelse av utvärdering</i>	Tidsperiod	Avikelser i tid
<b>Etappens tidsperiod (startdatum – slutdatum):</b>	190901-200304	Etappen förlängdes inledningsvis från 191201 till 200214 och sedan igen till 200304

<b>Aktiviteter under etappen:</b>	<b>Utfall i % (vad har ni utfört av det som var planerat för denna etapp)</b>	<b>Avvikelser (större ändringar som gjorts – avvikelser från plan och anledning )</b>
<ul style="list-style-type: none"> <li>Projektledning: Produktbeskrivning på mediawiki.org; Etapprapport 1 inskickad</li> </ul>	Produktbeskrivning på mediawiki.org: 100 % Etapprapport 1: 100 % klar	Produktbeskrivningen sker istället på meta.wikimedia.org
<ul style="list-style-type: none"> <li>Community konsultation: Ett uppstrukturerat ärendehanteringssystem; uppsatta wiki-sidor; mailutskick på samtliga relevanta maillistor</li> </ul>	Ett uppstrukturerat ärendehanteringssystem: 100 %	Mailutskicket har inväntat att de relevanta wikisidorna helt

## Etapprapportering

	Uppsatta wiki-sidor: 100 % Mailutskick på samtliga relevanta maillistor: 0 %	uppdaterats. Detta borde därmed kunna göras tidigt i etapp 2.
<ul style="list-style-type: none"> <li>Kommunikation: Blogginlägg publicerat; pressmeddelande utskickat</li> </ul>	Blogginlägg publicerat: 100 % klar  Pressmeddelande utskickat: 0 % klar	
<ul style="list-style-type: none"> <li>Kommunikation: Deltagande på konferenser</li> </ul>	Deltagande på konferenser: 100 % klar	
<ul style="list-style-type: none"> <li>Integrering med spelare (och Wikimediaplattformarna)</li> </ul>	100 % klart	
<ul style="list-style-type: none"> <li>Utveckling Inspelare: Definition av API, grundläggande funktionalitet</li> </ul>	75 % klart	
<ul style="list-style-type: none"> <li>Utveckling Lagring: Första grund för lagring finns</li> </ul>	75 % klart	
<ul style="list-style-type: none"> <li>Utveckling Manus: Definition av API, grundläggande funktionalitet</li> </ul>	50 % klart	
<ul style="list-style-type: none"> <li>Expertdiskussioner/rapportskrivande</li> </ul>	100 % klart	
<ul style="list-style-type: none"> <li>Resor och evenemang: Rapport från arbetet med WMDE i två veckor</li> </ul>	100 % klart	
<ul style="list-style-type: none"> <li>Arbetstid för referensgrupp, evenemang samt kommunikation</li> </ul>	0 % klart	Referensgruppen har ej engagerats under denna etapp

Kvarvarande aktiviteter från föregående etapper:	Utfall i % (vad har ni utfört av det som var planerat för denna etapp)	Avvikelser (större ändringar som gjorts - avvikelser från plan och anledning )
<ul style="list-style-type: none"> <li>N/A</li> </ul>		
<ul style="list-style-type: none"> <li>N/A</li> </ul>		

### 2.2 Avvikelser från plan

I detta kapitel beskrivs mer detaljerat om de avvikelser (orsaker och påföljder) som gjorts från plan och som kortfattat beskrivits i föregående kapitel (kapitel 2.1).

- Allmänt: Samtliga projektparter kom av olika anledningar igång med arbetet i projektet senare än planerat. Detta har till viss del berott på två andra stora projekt som blivit beviljade vilket det aktuella projektet kommer att kunna dra nytta av, men som krävt mer arbetstid nu inledningsvis. Det ena, Verktyg för Partnerskap innebär ett fördjupat samarbete med Wikimedia Foundation kring teknisk utveckling vilket har en direkt påverkan när det kommer till att lansera Wikispeech text-till-tal på projekten. Det andra projektet beskrivs under Utveckling Manus nedan. Detta resulterade i en revidering av etappdatumen i December 2019.  
På grund av ovanstående blev den planerade rekryteringen av ytterligare en utvecklare hos Wikimedia Sverige kraftigt försenad vilket inneburit att man slutligen inte kunnat lägga så mycket tid i projektet som planerat för etappen. Vi räknar med att återstoden av vad vi planerat göra under etapp 1 kommer att hinnas med under etapp 2.
- Aktiviteter före projektstart: För att dra nytta av den visibilitet Wikimedia Sverige projekt erhöll som en del av Wikimania 2019 har en del av leverablerna kring Community konsultation och Kommunikation genomförts, på egen bekostnad, före projektstart.
- Projektledning: Produktbeskrivning på mediawiki.org: Informationsstrukturen har uppdaterats så att projektinformation nu ligger på meta.wikimedia.org (meta) och enbart teknisk information återfinns på mediawiki.org.
- Kommunikation: pressmeddelande utskickat: De två naturliga tillfällena för ett pressmeddelande (projektstart samt beslut om stöd från PTS) nyttjades ej. Pressmeddelande kommer istället att invänta nästa lämpliga tillfälle. Notera att projektet omnämns i pressmeddelande från KTH<sup>1</sup>.
- Kommunikation: Deltagande på konferenser: Denna punkt kommer att spridas över etapperna där det bäst passar och bäst möjligheter ges. Utöver de konferenser som nämn senare presenterades projektet även på Wikimania 2019 precis före projektstart.
- Utveckling Inspelare/Lagring: På grund av de tidigare nämnda förseningarna samt den återkoppling vi fick från Wikimedia Deutschland har dessa komponenter inte kommit så långt som planerat. Vi räknar med att kunna sätta dem i fokus igen under inledningen av etapp 2.

---

1

<https://www.kth.se/forskning/artiklar/ett-av-varldens-storsta-mediarkiv-ska-bli-sokbart-1.937794>

- Utveckling Manus: Definition av API, grundläggande funktionalitet: Arbetet med manusverktyget kommer inledningsvis att gå långsammare än först planerat. Detta då utvecklingen måste synkas med ett annat projekt som blev godkänt under perioden. Detta projekt<sup>2</sup> kommer innebära att manusverktyget kan göras mer ambitiöst vilket kommer att ge bättre slutresultat, men kräver lite mer omsorgsfullt inledande arbete.
- Arbetstid för referensgrupp, evenemang samt kommunikation: Denna punkt kommer att spridas över etapperna där det bäst passar och återkopplingen är till störst nytta.

### 2.3 Kostnader för etappen

Nedan beskrivs finansiering och kostnader som projektet haft under etappen.

Finansiering	Belopp i kronor
Ej nyttjad finansiering från tidigare etapper (PTS samt egen finansiering)	N/A
Erhållen finansiering från PTS för aktuell etapp	614 760 kr
Budgeterad egen finansiering för aktuell etapp	212 609 kr
<b>Total finansiering för etappen</b>	<b>827 369 kr</b>
Kostnad	Belopp i kronor
Utfall egen finansiering för aktuell etapp	WMSE: 79 684 kr STTS: 18 342 kr KTH: 114 583 kr Dyslexif.: 0 kr Total: 212 609 kr
Utfall PTS finansiering för aktuell etapp	WMSE: 305 033 kr STTS: 186 658 kr KTH: 95 105 kr Dyslexif.: 0 kr Total: 586 796 kr
<b>Utfall total kostnad för aktuell etapp</b>	<b>799 405 kr</b>
<b>Outnyttjad finansiering vid etappens slut som flyttas till nästa etapp (Total finansiering för etappen – Utfall total kostnad för aktuell etapp)</b>	<b>27 964 kr</b>

### 2.4 Händelser under perioden som etapprapporten avser

I detta kapitel beskrivs viktiga händelser som skett under etappen.

Före etappen: Projektet presenterades under Wikimediarörelsens världskonferens, Wikimania, som i år hölls i Stockholm. Där presenterades projektet för de delar av rörelsen som brinner starkast för att göra kunskapen mer tillgänglig via tal samt de som experimenterat med insamlande av

<sup>2</sup>

<https://www.vinnova.se/p/djupinlard-talsyntes-for-upplasnig-av-lang-och-informationsrik-text-pa-svenska/>

talresurser. Som ett resultat av detta kontaktades vi av LinguaLibre som också är ett open source initiativ kring insamling av taldata.

Ärendehanteringssystemet: Strukturen för ärendehanteringssystemet (Phabricator) har uppdaterats för att åtgärda de brister vi upptäckt sedan förra strukturen sattes upp för Wikispeech (2016-2018). De öppna ärenden som kvarstod efter Wikispeech (2016-2018) har gått igenom och kategoriserats för att avgöra om de är relevanta för detta projektet eller ej.

Projektuppstart: Arbetsredskap för projektet har utvecklats däribland: mötes- och rapporteringsschema samt format för dessa; avtal har satts upp mellan samarbetsparterna; ett system har etablerats för delning av material.

Återkoppling från Wikimedia Foundation: En sammanställning har gjorts över de åtgärder som begärts och kommentarer som getts från Wikimedia Foundation vilka kan anses ha blockerat tidigare implementering av Wikispeech på Wikimedias plattformar.

Mjukvaruinventering: En inventering har påbörjats av existerande öppen mjukvara vilken skulle kunna återanvändas, eller ligga till grund för, de planerade komponenterna.

Nyrekrytering: Wikimedia Sverige har rekryterat ytterligare en utvecklare för att arbeta i projektet.

Wikimedia Deutschland: Under januari genomfördes ett längre besök hos Wikimedia Deutschland för att ta del av deras erfarenheter kring MediaWiki-utveckling samt för att etablera en gemensam stilguide och vokabulär som kommer ligga till grund för deras kommande kodgranskning och stöd i projektet. En genomlysning av den existerande kodbasen gjordes även med fokus på de delar Wikimedia Foundation kommer titta på inför aktivering. Besöket sammanfattades i en rapport<sup>3</sup>.

Koduppdateringar: Delar av den kodbas som byggdes för Wikispeech (2016-2018) har uppdaterats för att underlätta vidareutveckling inom ramen för detta projekt. Detta är även ett stöd för den säkerhetsgranskning som planeras av Wikimedia Foundation.

Deltagande på konferenser: KTH har deltagit med projektet vid ett antal konferenser däribland: SSW, Interspeech, CLARIN ERICs årskonferens, Swe-Clarins årskonferens. Utöver dessa har ett flertal seminarier hållits.

Alternativ syntes: Ett alternativ till den syntesmotor som användes i Wikispeech (2016-2018) har undersökts. Detta är en pågående utvärdering av en modernare form av talsyntes baserad på "deep learning". Fördelen är att den går snabbare att träna, och att kvaliteten på den syntetiserade rösten har potential att bli mycket bra. Nackdelen är att syntesen har sämre prestanda (den är

---

3

[https://se.wikimedia.org/wiki/Projekt:Wikispeech\\_-\\_Talresursinsamlaren\\_2019/WMDE\\_site\\_visit\\_2020-01-20-24](https://se.wikimedia.org/wiki/Projekt:Wikispeech_-_Talresursinsamlaren_2019/WMDE_site_visit_2020-01-20-24)

långsammare att köra). Detta är också en del i arbetet med att undersöka alternativ till MaryTTS-systemet som vi nu använder. MaryTTS främsta brist är att den är osmidig för detta projekts syften, och börjar bli lite ålderstigen, men dess fördel är att den har utvecklats under lång tid och är tämligen stabil.

Synergier: Planering har påbörjats för hur detta projektet och det projekt som beviljats medel från Vinnova bäst kan dra nytta av varandra.

### 3. Uppdaterad bilaga – budget och etapper

PTS vill att projektet har uppdaterat dokumentet Budget och etapper för nästkommande etapp i denna etappredovisning och skickar med denna som en separat bilaga. Detta ska göras även om det inte skett några förändringar sedan förra rapporteringen. Dokumentet ska dateras och versionshanteras.

**Observera att det endast är kommande etapp(er) som ska justeras - alltså den etapp som ni nu rapporterar ska inte justeras även om den inte blev som ni planerat.**

Den uppdaterade dokumentet kan t.ex. innehålla fler eller andra aktiviteter än vad som funnits med ursprungligen samt ev. andra uppdateringar som behövt göras som åtgärder för att hantera eventuella förändringar eller avvikelser i denna etapp. Lägg även in nya rader och döp dem "Kvarvarande aktiviteter från föregående etapper" om det är så att ni flyttar aktiviteter från etapper som redan är avslutade (se exempel i kapitel 2.1).

#### 3.1 Kommande händelser

I detta kapitel beskrivs kommande intressanta händelser av vikt för projektet. Nedan beskrivs även nya händelser som tillkommit sen ursprunglig version.

Omarkitekturering: Två större omarkitektureringar kommer att ske i text-till-tal komponenten som ett direkt resultat av den återkoppling Wikimedia Sverige fick från Wikimedia Deutschland.

Likasinnade organisationer: Under nästa etapp kommer vi på nytt ta kontakt med Mozilla Voice och LinguaLibre för att bättre förstå deras initiativ. Detta kan dels ge oss möjlighet att identifiera komponenter som kan återanvändas, dels se till att även dessa organisationer kan dra nytta av den taldata som samlas in och komponenter vi bygger.

#### 3.2 Uppdaterad riskanalys

I detta avsnitt beskrivs projektet risker vid etapprapporteringstillfället. Riskanalysen i tabellen nedan är uppdatera med nya risker som identifierats sen förra etapprapporteringen. Även de gamla riskerna bedöms och viktas på nytt.

Låt även gamla risker som bedöms vara avklarade stå kvar i tabellen.

<b>Version av riskanalysen (v.1 är den som var med i förstudieorganiet):</b>	v.2
--	-----

<b>Beskriv vad som är uppdaterat/förändrat från förra versionen:</b>	Ytterligare åtgärder har tagits för att förebygga risk nr. 4.
--	---

Så här vill PTS att man beskriver riskerna:

**Sannolikhet** – hur stor är sannolikheten att denna risk kommer att slå in, 1 är låg sannolikhet och 3 är hög sannolikhet.

**Effekt** – hur stor blir effekten om den identifierade risken blir aktuell, 1 är liten effekt och 3 är stor effekt.

Risker med riskfaktor 9 bör elimineras, risker med riskfaktor större eller lika med 4 ska elimineras, minimeras eller hanteras. Övriga risker ska bevakas.

Risk nr.	Risk	Sannolikhet (S)	Effekt (E)	Riskfaktor (S)*(E)
1	GDPR och annan lagstiftning omöjliggör helt fri delning av inspelat tal	2	2	4
2	Basteknologi er från Språkbanken Tal hinner inte bli färdiga i tid för att komma projektet till nytta under projektiden	2	1	2
3	Skyddet för användares integritet hos Wikimedia orsakar brister i den demografiska metadata vilket försämrar taldataets användbarhet	2	1	2

## Etapprapportering

4	Kod- och säkerhetsgranskning av nyutvecklad mjukvara sker inte och denna kan därför inte gå live	1	2	2
---	--	---	---	---

Förebyggande åtgärder ska beskrivas utförligt (för risker  $\geq 4$ ).

Risk nr.	Förebyggande åtgärd (Beskrivningen hur risken hanteras)
1	<p>Det råder ännu osäkerhet kring tolkningen av den nya lagen GDPR, så vi måste planera in en viss flexibilitet. En del av projektgenomförandet handlar därför om att utreda om anonymisering av röst och användardata behöver ske. Om lagstiftningen förhindrar lagring och delning av ren röstdata ökar vi fokus på anonymisering och tillför resurser (som tas från inspelat tal).</p> <p>Fokus under projekttiden kommer därför inte vara den aktiva insamlingen av inspelat tal utan att bygga verktygslådan (inklusive anonymisering) för att omgående kunna påbörja insamling när lagen förtydligas. Därtill kan verktyg som fokuserar på insamlande av annoteringar etc. nyttjas utan problem då de inte samlar in uppgifter som kan identifiera en person.</p>
2	Risikfaktor under 4. Risken ger marginella och tillfälliga effekter på projektgenomförandet.
3	Risikfaktor under 4. Verktyg och metoder tas fram så att de ger stöd för användandet av demografiska metadata, men fungerar även utan. Även med begränsad mängd demografisk data har det insamlade datat ett mycket stort värde för många områden.
4	4 Risikfaktor under 4. Risken har minimerats genom att projektet förutom med Wikimedia Foundation har etablerat samarbete med Wikimedia Deutschland som har ett stort utvecklarteam med nödvändiga erfarenheter och behörigheter. Sedan projektstart har Wikimedia Sverige även inlett ett fördjupat samarbete med Wikimedia Foundation kring bl.a. framtida mjukvaruutveckling vilket öppnar upp för ökat samarbete och dialog kring kod och säkerhetsgranskning.

