Intresseanmälan att kandidera som EDIH

Aktörer

Wikimedia Sverige (WMSE) är en s.k. tematisk hubb med fokus på institutionella innehållspartnerskap – "thematic hub for institutional content partnerships" – under uppbyggnad. Det är en ny roll för WMSE och del av en ny strategi för hela Wikimediarörelsen globala tillväxt – flera tematiska hubbar kommer att följa på andra platser i världen. Samarbetspartner och grundfinansiär är Wikimedia Foundation (WMF).

Som tematisk hubb erbjuder vi tjänster och bedriver produktutveckling i syfte att underlätta publicering och utbyte av metadata och mediefiler från och mellan institutioner och Wikimediaplattformarna (och dess mycket stora antal volontärer). Hubben omfattar även utveckling av språkrelaterade teknologier och resurser (t.ex. tal, lexem, OCR:ad text) och i även i den utveckling ingår institutionella partnerskap. Institutionerna omfattar arkiv, bibliotek, museer, forskningsinstitutioner, och andra typer av organisationer med digitala samlingar relevanta för Wikimediarörelsen och dess projekt: Wikipedia, Wikimedia Commons, Wikidata, Wikisource, Wiktionary, m.fl.

Produkt- och tjänsteutveckling inom hubben sker i projektform. Samarbetspartners varierar från projekt till projekt. Vi utvecklar enbart öppen mjukvara och använder/publicerar enbart öppen data och öppet innehåll. Den infrastruktur vi vanligen utvecklar på är den som används av den globala Wikimediarörelsen och som drivs och bekostas av WMF. Infrastrukturen stödjer alla steg i produktutvecklingsprocessen. Att vi kan använda WMF:s tekniska infrastruktur gör att vi kan sänka, nästan helt eliminera, egna tekniska driftskostnader.

Samarbetsavtal, hubben: WMF

Samarbeten, Språkteknologi: WMF, Wikimedia Deutschland (WMDE), Kungliga tekniska högskolan (KTH), STTS, Mozilla Foundation

Samarbeten, Institutionella partnerskap: Centralmuseernas samarbetsråd (och flera enskilda centralmuseer), Kungliga biblioteket (KB), Riksantikvarieämbetet (RAÄ), Institutet för Språk- och Folkminnen (ISOF), Arbetsam, UNESCO Archives, m.fl.

Informella partnerskap och nätverksmedverkan: De 173 Wikimediaorganisationer som finns etablerade världen över, Digisam, OpenGLAM, Europeana, Digital Public Library of America (DPLA).

Ledning: WMSE:s verksamhetschef, John Andersson, är också ansvarig för hubben. Utvecklingsprojekt inom hubben har olika projektledare. I relation till WMF sorterar hubben under dess avdelning för produktutveckling.

John Andersson, 073-39 65 189
john.andersson@wikimedia.se
2020-04-27

Vakanta roller/positioner, hubben: Mjukvaruutvecklare, dataspecialister, designers, och community managers är roller som vi kommer att rekrytera till för att utöka bemanningen i hubben.

Lokaler: WMSE:s kontor är på Goto10 som drivs av Internetstiftelsen, men kan komma att behöva byta adress om hubben expanderar.

Finansiering

Finansiering från Wikimedia Foundation (WMF)

WMSE erhåller 2019-20 finansiering från WMF för att bygga upp verksamheten som en "thematic hub for institutional content partnerships". Ett avtal för 2020-23 arbetas på i dagsläget. Målgruppen är till att börja med framförallt arkiv, bibliotek och museer samt volontärer inom Wikimediarörelsen med specifikt intresse för att samarbeta med kulturarvsinstitutioner och dra nytta av deras samlingar och kunskap i sitt wikiarbete. Alla som läser och använder Wikipedia, Wikimedia Commons, etc. är slutanvändarna.

Finansiering från Post- och Telestyrelsen (PTS)

WMSE erhåller finansiering från PTS för att utveckla öppna språkresurser och öppen mjukvara för talsyntes och insamling av talresurser (se också nedan).

Framtida finansiering

Med hubbens grundfinansiering säkrad av WMF avser vi att söka flera finansieringskällor för specifika utvecklingsinsatser och även möjliggöra mer R&D- och innovationsinriktad utveckling.

Finansiering från Postkodslotteriet

Vi har tidigare haft projektfinansiering från Postkodslotteriet, för projektet Kopplat öppet kulturarv samt FindingGLAMs, och har 2020 ansökt om att bli en av deras få stående förmånstagare. Som stående förmånstagare skulle vi ha stor frihet att själva välja till vilka av våra aktiviteter medlen används där arbetet kopplat till hubben skulle prioriteras.

Finansiering från Europeiska Unionen (EU)

Vi har nyligen ansökt om medel inom utlysningen Horizon 2020. Ansökan är inriktad mot vidareutveckling av talsyntes och öppna språkteknologiresurser, bl.a. att utöka vilka språk mjukvaran Wikispeech stödjer med hjälp av crowdsourcing.

John Andersson, 073-39 65 189
john.andersson@wikimedia.se

2020-04-27

Finansiering från Vinnova

Vi kommer möjligen att söka finansiering inom utlysningen Civic Tech. Denna kommer att vara inriktad mot institutionellt datautbyte och/eller språkresurser.

Finansiering från enskilda institutioner

Vi har ett aktivt större samarbetsprojekt med tillhörande finansiering i bidragsform från KB. Vi har också genomfört flera mindre projekt åt bl.a. RAÄ där vi formellt agerat konsult åt dem. Denna typen av finansieringskällor för mindre projekt kommer att fortsätta att komplettera större och mer långsiktiga finansieringskällor.

Aktiviteter, tjänster, och målgrupper

Mjukvara för talsyntes och annan språkteknologi

Vi utvecklar för närvarande Mediawikitillägget Wikispeech och en understödjande men fristående applikation för insamling av talresurser. Mjukvaran är i första hand avsedd för användning i Wikimediaprojekten (Wikipedia, Wikisource, etc.) men då de understödjande språkresursen och källkoden är öppen kan de återanvändas och anpassas av tredjepartsutvecklare. För närvarande stödjer Wikispeech svenska, engelska, och arabiska – givet tillgång till medel utvecklar vi stöd för ytterligare språk.

Den primära målgruppen är alla brukare av mjukvaran Mediawiki, som alla Wikimediaplattformar använder, och som vill öka tillgängligheten till innehållet i sina Mediawikiinstanser. Många tusentals wikis, som inte drivs av Wikimedia Foundation, men som använder Mediawiki, kommer att kunna använda Wikispeech. Sekundära målgrupper är bl.a. tredjepartsutvecklare och forskare (inom t.ex. talsyntes, språk, AI).

Den långsiktiga visionen för Wikispeech är att alla personer som behöver – för att de inte kan läsa, har lässvårigheter eller helt enkelt föredrar att ta in information genom att lyssna (och söka via tal) – ska kunna göra det på sitt eget språk. I ett globalt perspektiv, och givet Wikimediaplattformarnas stora användning, är detta en väldigt stor målgrupp.

Tjänster och produkter för institutionella innehållspartnerskap

Vi utvecklar mjukvara, dokumentation och processer för publicering av institutionella samlingar på Wikimediaplattformarna (ffa Wikidata och Wikimedia Commons). Vi prioriterar under hubbens första tid är att vidareutveckla verktyg till att kunna dra nytta av Wikimedia Commons nya förmåga att hantera strukturerade (eller “länkade”) data. Andra prioriteringar är att förbättra möjligheterna för institutioner att få tillgång till användnings- och datakvalitetstatistik samt att hämta hem data som

John Andersson, 073-39 65 189

john.andersson@wikimedia.se

2020-04-27

Wikimediavolontärerna lagt till de institutionella samlingarna t.ex. nya nyckelord, exakta positioner, och översatta beskrivningar.

De primära målgrupperna är institutioner (arkiv, bibliotek, museer, forskningsinstitutioner, etc.) och Wikimediavolontärer med ett specialintresse för deras samlingar. Sekundära målgrupper är bl.a. tredjepartsutvecklare, lärare och forskare. Slutanvändare är all de som använder Wikipedia som uppslagsverk, Wikimedia Commons som mediebibliotek, Wikidata som källa till länkad data, etc.

Överlappningen mellan talsyntes och annan språkteknologi och innehållspartnerskapen är omfattande. Det kan t.ex. handla om uppladdningar av ljudinspelningar från olika arkiv eller inkludering av specialiserade ordlistor. Men även för att de verktyg vi utvecklar för våra partners behöver vara tillgänglighetsanpassade så att all personal kan använda dem (vilket är ett krav genom webbtillgänglighetsdirektivet (EU) 2016/2102).

Tematiskt område

Artificiell intelligens

För att artificiell intelligens skall fungera krävs stora mängder träningsdata. Detta är sant både vad gäller språkteknologi (NLP), bildigenkänning, automatöversättning m.m. Ett vanligt återkommande problem har varit den “bias” som finns i existerande dataset där exempelvis vita män är kraftigt överrepresenterade i träningsdata vilket ex. försämrar ansiktsigenkänning på historiskt bildmaterial. Bildigenkänning för kulturhistoriskt material är generellt underutvecklat då kommersiella tjänster som t.ex. Google Vision och Microsoft Vision är tränade på samtida bildmaterial. För att komplettera de dataset som finns behöver antingen stora summor investeras eller alternativa lösningar undersökas. Inom ramen för hubben kommer vi att identifiera sätt att nyttja crowdsourcing för att allmänheten skall kunna bidra till att utveckla bättre, öppen och mer etisk AI. Denna alternativa lösning bedömer vi har mycket goda möjligheter.

Ett liknande problem är att engelsktalande män är överlägset mest representerat bland de språkresurser som finns. Detta gör att AI-tillämpningar inte fungerar väl för alla, vilket skadar jämställdheten och kan leda till utanförskap. Med genomtänkta crowdsourcingverktyg och -kampanjer kan problemet minskas genom att den träningsdata som behövs förbättras och breddas.

John Andersson, 073-39 65 189

john.andersson@wikimedia.se

2020-04-27

Wikimediarörelsens volontärer och personal producerar tillsammans redan mycket stora mängder texter, media samt metadata som redan används som träningsdata för maskininlärning/AI.

WMSE har erfarenhet av att skapa språkresurser lämpade som träningsdata för AI och med anpassning av automatisk igenkänning av klotter på svenskspråkiga Wikipedia med verktyget ORES. Vidare har vi de senaste åren arbetat med Wikimedia Foundation för att implementera insamlandet av strukturerad data för mediefiler på Wikimedia Commons, som kan användas för utveckling av bildigenkänning. I takt med att vi etablerar oss som hubb avser vi tillämpa metoder, processer, och träningsdataresurser som tagits fram i Wikimediaprojekt, som samarbetsprojektet med Met Museum och Microsoft, på svenska kulturarvssamlingar som delats på Wikimediaplattformarna.

Hos våra samarbetspartners, som KTH, och i den internationella Wikimediarörelsen finns såväl avlönade utvecklare som volontärutvecklare med stor erfarenhet av AI-tillämpningar.

Avancerade digitala färdigheter

Wikimediarörelsen har en nästintill unik position vad gäller flera avancerade digitala färdigheter. Genom att engagera ett stort antal volontärer i aktiviteter som research (efter datakällor, referenser, citat, mm.), artikelskrivande, mediefångst- och uppladdning (samt metadatasättning av medierna), OCR-korrektion, strukturerad datafångst, m.m. bidrar vi till att medborgare lär sig avancerade digitala färdigheter inom flera områden. Och detta på ett sätt som genererar stora mängder öppen data. Öppen data som i många fall fungerar som träningsdata för maskininlärning/AI.

WMSE har spetskompetens inom medborgardriven kunskapsproduktion och hur den kan samverka med kultur- och naturarvsarbete (ofta kallat “GLAM-wikisamarbete”) och språkteknologisk forskning. Denna erfarenhet kommer att användas när vi engagerar allmänheten i att ta fram material för AI.

Geografiskt område

Wikimedia Sverige har ett nationellt uppdrag och samverkar med såväl nationella institutioner som med regionala och lokala, t.ex. folkbiblioteken.

På europeisk nivå samarbetar vi med flera andra Wikimediaorganisationer och är även engagerade i av EU-finansierade projekt och nätverk (t.ex. Europeana). Wikimedia Deutschland, som har ett stort utvecklarteam, är en sådan partner.

VINNOVA
Sveriges innovationsmyndighet

John Andersson, 073-39 65 189
john.andersson@wikimedia.se

2020-04-27

Globalt agerar vi tillsammans med Wikimedia Foundation, den globala Wikimediagemenskapen och dess samarbetspartners. Vi samarbetar även med Mozilla Foundation, UNESCO, och andra FN-organisationer.