Strategische dataproductie: representativiteit van data via crowdsourcing

van der Sijs, N.; Kirstein, Anna; Broeder, D.

2015

document version
Peer reviewed version

Technische Vereisten voor een Crowdsourcing Infrastructuur

Inleiding en scope

De bedoeling van dit document is om een architectuur te schetsen voor een crowdsourcing infrastructuur (CSI) of CS platform die (1) het mogelijk maakt om de CSI te gebruiken met een grote verscheidenheid van omstandigheden en voor veel verschillende taken, (2) modulair van aard is, zodat software voor de verwerking een nieuw type taak makkelijk kan worden geïntegreerd wanneer nodig en (3) de CSI zoveel mogelijk toekomstbestendig maakt.

Aangezien het op dit moment onbekend is welke middelen kunnen worden aangewend om een CSI te realiseren noch wanneer zo'n taak van start zou kunnen gaan, zal de infrastructuur alleen in algemene termen geschetst worden. Zo'n schets die algemene principes neerlegt zal ook meer toekomstbestendig zijn dan een precieze specificatie voor een implementatie. Essentieel voor een CSI is volgens ons dat de infrastructuur schaalbaar is mbt. het aantal vrijwilligers dat er mee kan werken, alsook dat nieuwe typen taken en nieuwe methodes om taken te verrichten, zonder veel moeite in het systeem kunnen worden geïntegreerd. Vanwege de eenvoud voor het software en data management kiezen we voor een CSI model als een centrale web applicatie, die indien nodig ondersteund wordt door lokaal te installeren client software voor bv. ingewikkelde transcriptie verbeteringssoftware en/of audio/video alignment taken.

Om vrijwilligers hun werk op een aansprekende manier aan anderen te kunnen laten presenteren, en ook om nieuwe vrijwilligers aan te trekken, is het van groot belang dat een crowdsourcing platform vergezeld wordt of samengaat met een aantrekkelijke web-site die voldoend informatief is over de verschillende projecten die het crowdsourcing platform uitgevoerd worden of werden. Een dergelijke website kan tevens dienen als instructie en voorbeeld documentatie voor de verschillende project taken.

Kijkend naar de vrijwilligers kunnen crowdsourcing projecten grofweg verdeeld worden in twee klassen: de eerste is een goed gedocumenteerde groep van mensen, die vaak al een langere relatie met het project of een groep van gerelateerde projecten hebben, van vaak al jaren. Deze groep zal op basis van interesse en expertise meedoen aan mogelijk een aantal verschillende projecten. Vaak gaat het hier om transcripties maken (of verbeteren) van manuscripten. Een andere, bredere en grotere groep vrijwilligers is gewild voor taken als het creëren of toevoegen van data (collectie & corpus vorming). Dit kan een grote diverse groep van mensen zijn waar weinig informatie over beschikbaar is en waar de betrouwbaarheid van de bijdragen ook veel minder goed in te schatten is. Vanuit het technisch perspectief zal de eerste groep van projecten een vrijwilligersadministratie vereisen met

daarbij veel mogelijkheden om vrijwilligers te metadateren en ook een sociaalplatform willen voor (onderlinge) communicatie en informatie uitwisseling. Voor vrijwilligers in de laatste groep projecten is er minder noodzaak voor aspecten als een veilige inlogprocedure en langdurige opslag van gebruikersinformatie. Hun relatie met het CS platform is vaak maar tijdelijk. Bij taken als het creëren van nieuwe data e.g. uploaden van video, spraak of images zien we een noodzaak voor goede data management faciliteiten met veel data opslag en een snel netwerk.

Sommige projecten behoeven speciale maatregelen voor het beveiligen privacy of anderzijds gevoelige informatie. Men kan denken aan het transcriberen van geluidsopnamen van patiënt & arts gesprekken, of bv. transcriptie van sommige nog altijd gevoelige historische informatie van het European Holocaust Project Infrastructure (EHRI) project. Dat betekent dat er gezorgd moet worden voor een voldoend veilige inlog mogelijkheid en bv. mogelijkheden voor accreditatie van de vrijwilligers door derden.

De eisen die aan de infrastructuur schets ten grondslag liggen zijn voor een deel afkomstig van en/of getoetst aan de interviews van crowdsourcing platform programmeurs en managers en van bestaande implementaties.

De primaire scope van de CSI zijn toepassingen binnen de Humanities, met nadruk op classificatie, annotatie, transcriptie van tekstuele data en scans en verbeteringen daarvan. Alhoewel onze enquête ook wijst op taken die het uploaden van bestanden noodzakelijk maakt zoals classificatie van gefotografeerde artefacten of van spraak, voor bv. collectie en corpus vorming, ook tot de mogelijkheden behoort.

Beschrijving basis architectuur.

De crowdsourcing infrastructuur kan op grond van functionele en houdbaarheid verwachtingen grofweg verdeeld worden in vijf onderdelen:

- vrijwilligersbeheer
- takenbeheer
- resultaatbeheer
- werkmodulebeheer
- crowdsourcingprocesbeheer

We kiezen waar mogelijk voor een assemblage van los gekoppelde componenten, die indien nodig uitgeruild kunnen worden voor andere, ipv. een monolitisch systeem, zie figuur 1.

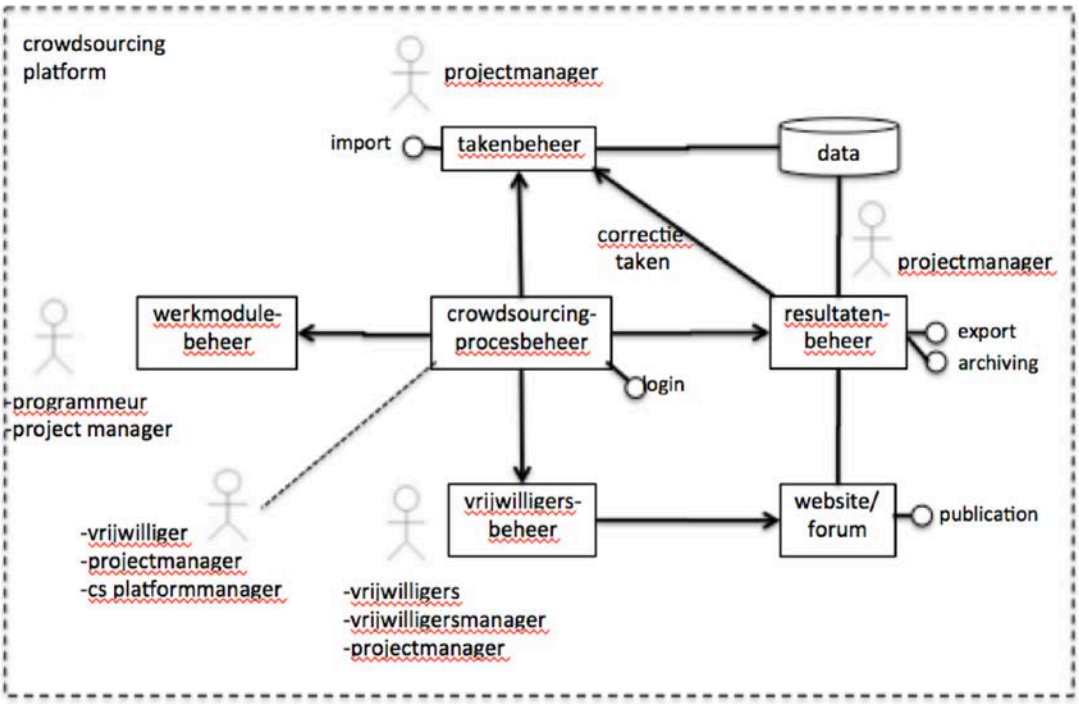

Figure 1 Het Crowdsourcing Platform

A. de vrijwilligersbeheer module bevat gegevens van alle vrijwilligers die een taak kunnen binnen het systeem en houdt rekening met verschillende speciale rollen nodig voor projecten zoals projectleiders, correctors. Alle crowdsourcing projecten onderschrijven het belang van een solide administratie met voldoende kenmerken voor een efficiënte verdeling van taken naar individuele vrijwilligers. De projectleider is in staat om speciale voor het project benodigde kenmerken toe te voegen aan de vrijwilligersinformatie die bv. al door de vrijwilliger zelf is ingevuld (profiel), zoals de gebleken geschiktheid voor een bepaalde taak.

A1. Gekoppeld aan de vrijwilligers administratie, zien we een sociaal-platform (forum) waar de vrijwilligers onderling kunnen communiceren en bv met elkaar voorbeelden van taak gerelateerde problemen en oplossingen kunnen uitwisselen. Ook boodschappen van de projectleider en kunnen via het forum verspreid worden.

B. Werkmodule infrastructuur
Gezien de variëteit van taken die we met het crowdsourcing willen aanpakken en de mogelijkheid van nieuwe, mogelijk ook interactieve hulpmiddelen, om taken te versnellen e.g. gedeeltelijk automatische transcriptie, willen we het zo eenvoudig mogelijk maken om nieuwe werkmodules te maken en in het crowdsourcing platform te integreren. Initieel moeten al een aantal taakmodules voor transcriptie en oplijning en classificatie deel uit maken van het basis-platform. Het integreren van “Transkribus” is bv. een noodzaak.

C. Taakbeheer
Deze stelt de projectleider in staat nieuwe taken te uploaden of te definiëren en te voorzien van speciale kenmerken nodig voor het matchen van taken en vrijwilligers. Om snel veel taken aan te maken moet er een batch import mode beschikbaar zijn. Taken worden toebedeeld aan vrijwilligers via automatische regels of handmatig door de project leider. Eenmaal toebedeelde taken blijven gelinkt aan de vrijwilligers en (gedeeltelijke) resultaten, tenzij de project leider dit veranderd. Een verwerkte taak levert een resultaat, maar kan ook weer dienen als een nieuwe correctietaak.
Taakbeheer zorgt voor het toewijzen van taken en moet in staat zijn om flexibel complexe regels toe te passen bij het toebedelen van taken aan vrijwilligers. Iedere type taak kan gemerkt worden met een aantal eigenschappen e.g. moeilijkheidsgraad. De toebedelingsprocedure is dat de projectleider regels definieert die batches van taken kunnen verdelen door de vrijwilligers kenmerken te matchen met de kenmerken van de taken.

D. Resultaatbeheer
Dit stelt de projectleider in staat om de resultaten per project, per vrijwilliger, per taak te exporteren in een standaard formaat. Ook moet export van resultaten naar een geschikt archief mogelijk zijn. Resultaten moeten geciteerd kunnen worden als voorbeelden voor anderen.

E. crowdsourcingprocessbeheer
Deze module stelt de projectleider in staat tot:
- verzorgen van overzichten van projecten, vrijwilligers, taken en hun onderlinge relaties
- zoeken en aanpassen van alle elementen in het systeem: projecten, vrijwilligers , taken, regels alsook de onderlinge relaties.
Crowdsourcingprocessbeheer zorgt ook voor de authenticatie en autorisatie gegevens van alle componenten van de CSI. Vrijwilligers kunnen en moeten inloggen voor toegang tot het systeem. Voor sommige projecten is een login via een Open Social ID systeem e.g. Google ID voldoende terwijl voor andere een CSI interne login functie gebruikt moet worden.

Noodzakelijke flexibiliteit.
Aangezien we niet kunnen voorspellen welke taak- en vrijwilligerskenmerken we in de toekomst nodig gaan hebben voor een efficiënte matching van vrijwilligers en taken, stellen willen we zowel taken als vrijwilligers kunnen merken met tags uit door de projectleider zelf te creëren vocabulaires. Deze tags kunnen dan gebruikt worden in de vrijwilliger / taak matching regels.

Sociaal aspect
Een van de primaire taken voor het platform is het ondersteunen en motiveren van de vrijwilligers, daartoe is een forum en de mogelijkheid om resultaten van de vrijwilligers als voorbeeld te delen met anderen binnen en buiten de CSI

noodzakelijk. Het zal voor de project manager mogelijk zijn om (groepen van) vrijwilligers te benaderen via het forum of mail, voor terugkoppeling, of om aan te geven dat een nieuwe serie taken beschikbaar is. Een

Ontwikkeling, Onderhoud en beheer.

Bij onderhoud en beheer van het CS platform zien we minimaal de volgende rollen:

- Algemeen projectleider verantwoordelijk voor de hele CSI: 0,2 fte
- Vrijwilligersmanager, verantwoordelijk voor het uploaden/ definiëren van taken en het contact met de vrijwilligers: 0,6 fte
- Een data manager & programmeur voor het onderhoud van de software, en oplossen van problemen met data in/output: 0,4 fte

Het ontwikkelen van nieuwe werkmodules is sterk afhankelijk van de complexiteit en moet op een stukwerk basis beoordeeld worden.
Voor het specificeren en ontwikkelen van de CSI software, mogelijk gebruik makend van al bestaande software (e.g. http://pybossa.com/), kunnen we niet verder tot een redelijke schatting komen dan 1 PY. Er zou ook eerst een uitgebreide inventaris van bestaande CS software moeten komen en daarbij de mogelijkheden voor hergebruik van de code onderzoeken. Tevens zou er gesynchroniseerd en samengewerkt moeten worden met de CLARIAH crowdsourcing taak in WP3.