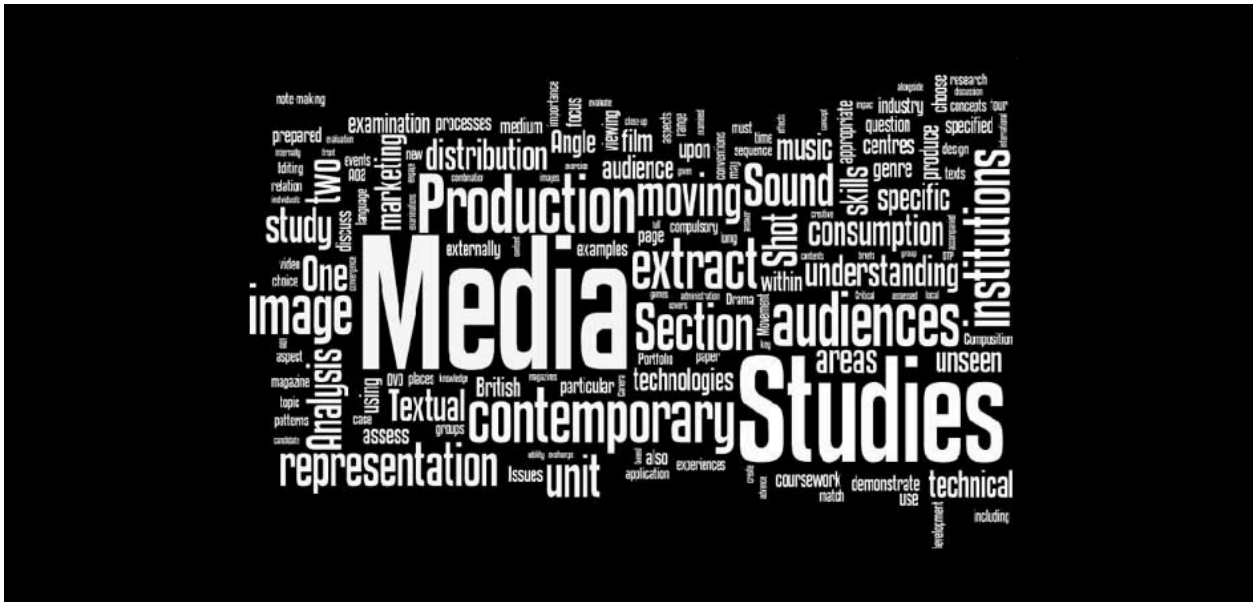
CC 15-091

CLARIAH WP5 Mediastudies/audiovisuele data plan

Julia Noordegraaf, Maarten de Rijke, José van Dijck, Johan Oomen, Jasmijn van Gorp

Versie: 1.05
Datum: 18 maart 2015



Inhoudsopgave

1. Inleiding

Achtergrond

De onderzoeksinfrastructuur CLARIAH beoogt het onderzoek van geesteswetenschappers met digitale data en tools te ondersteunen. De infrastructuur moet faciliteiten bieden voor de workflow van een typisch onderzoeksproject, waarin wordt gewerkt met bestaande en/of zelf gegeneerde data.

Mediawetenschappers werken over het algemeen met bestaande datasets: audiovisuele bronnen (film, radio en televisieprogramma's, online audio en video), gepubliceerde bronnen (kranten, omroepgidsen, sociale media content) en gestructureerde data (zoals kijkcijfers, of een database als Cinema Context met aan kranten en archivalia ontleende data over films, bioscopen, vertoningen en in de filmsector actieve personen en bedrijven[1]).

Vanwege de specifieke aard van audiovisuele bronnen (combinatie van beeld en geluid, tijdgebonden) zijn additionele maatregelen nodig om ze op grotere schaal toegankelijk en doorzoekbaar te maken. Dat gebeurt enerzijds door ze te koppelen aan beschrijvingen die de inhoud van de bron en een aantal contextuele gegevens bevatten (metadata, ondertitelfiles) en anderzijds met behulp van nieuwe technologie, zoals spraakherkenning voor het automatisch genereren van transcripties en beeldherkenning voor het zoeken naar patronen op visuele kenmerken.

Recente gebruikersstudies (Bron et al. 2013; Bron et al. 2015) hebben laten zien dat mediawetenschappers de diverse bronnen in samenhang willen onderzoeken. Een goed begrip van audiovisuele bronnen vereist gedetailleerde kennis over de context waarin ze zijn geproduceerd, verspreid en waargenomen. Tevens zijn onderzoekers die met audiovisuele bronnen werken vaak geïnteresseerd in de rol van media in de constructie, verspreiding en receptie van een bepaald sociaal of cultureel fenomeen, wat vereist dat verschillende perspectieven op hetzelfde fenomeen door de tijd heen kunnen worden opgespoord en in samenhang geanalyseerd. Daarnaast wordt het onderzoeksproces van mediawetenschappers gekarakteriseerd door een *exploratieve fase* van de inhoud van het materiaal (waarin de vraagstelling wordt ontwikkeld en/of aangescherpt) gevolgd door een meer doelgerichte verzameling en analyse van met name contextuele data (*contextualiseringsfase*) – een cyclus die vaak een of meerdere keren wordt herhaald en wordt afgesloten met het presenteren van de gevonden resultaten (*presentatiefase*) (Bron et al. 2015).

De grootste uitdaging voor mediawetenschappers is toegang tot het bronnenmateriaal, dat vaak auteursrechtelijk beschermd is en verspreid over verschillende bewaarplaatsen. Tevens is er behoefte aan instrumenten om het diverse bronnenmateriaal in samenhang te kunnen onderzoeken. De afgelopen jaren zijn er in de context van NWO CATCH, CLARIN-NL en CLARIAH-Seed tools ontwikkeld die in deze behoeften voorzien. Deze tools zijn momenteel alleen beschikbaar als prototype, verkeren in diverse stadia van ontwikkeling en zijn nog niet (allemaal) beschikbaar voor brede groepen onderzoekers. De voorliggende agenda voor mediastudies/audiovisuele data in CLARIAH is gericht op het doorontwikkelen en beschikbaar

[1] www.cinemacontext.nl
[2] CoMeRDa is vanwege copyrightbeperkingen momenteel alleen te raadplegen op locatie bij Beeld en Geluid. Zie

stellen van deze tools, inclusief het trainen van onderzoekers in het werken ermee. De agenda richt zich in eerste instantie op onderzoekers binnen de mediastudies maar daarnaast ook op andere terreinen van geesteswetenschappelijk en sociaalwetenschappelijk onderzoek waar audiovisuele bronnen worden ingezet.

Doelstelling

De agenda voor mediastudies/audiovisuele data heeft als doelstelling vijf bestaande tools voor exploratief en gericht, contextueel mediawetenschappelijk onderzoek te consolideren, verder te ontwikkelen en beschikbaar te stellen, in eerste instantie ten behoeve van onderzoekers in de mediastudies maar met uitbreiding naar alle vakgebieden die audiovisuele bronnen willen betrekken in onderzoek naar specifieke historische thema's of gebeurtenissen. Daarnaast hebben we een overkoepelend project geformuleerd (Media Studies Suite) dat dient om overlap in functionaliteit tussen tools op te lossen via het samenvoegen van tools en nieuwe, state of the art functionaliteiten in die gesynthetiseerde tool te integreren. Het gaat om de volgende tools:

CoMeRDa[2]	Geaggregeerde zoekinterface: exploratief onderzoek in verschillende collecties (beschrijvingen televisieprogramma's, foto's, Beeld en Geluid Wiki, omroepgidsen, kranten); visuele weergave van de bronnen.
AVResearcherXL[3]	Verkennen van beschrijvingen van beschrijvingen radio- en televisieprogramma's, ondertitelfiles en algemene krantenartikelen. Faciliteert comparatieve analyse van televisieprogramma's en krantenartikelen via visualisaties als woordenwolken en tijdslijnen.
TROVe[4]	Multimediale zoekmachine, faciliteert onderzoek naar de verspreiding van nieuws via radio en televisie, Twitter, online kranten en blogs; geeft zicht op ontwikkeling publieke debatten in de loop der tijd en via verschillende media (identificatie belangrijkste topics, spelers en impact via woordwolken, lijn- en puntgrafieken en kijkcijfers, directe toegang tot de items zelf).
DIVE[5]	Presentatie van collectie-items in context (verbinding met relevante contextuele online informatie), faciliteert intuïtief browsen om dieper in de content te "duiken".
Verteld Verleden[6]	Exploratief en gericht zoeken in audiovisuele narratieve interviewcollecties.
Media Studies Suite	Integratie van tools met overlappende functionaliteiten (CLARIAH

[2] CoMeRDa is vanwege copyrightbeperkingen momenteel alleen te raadplegen op locatie bij Beeld en Geluid. Zie voor een beschrijving http://vps46235.public.cloudvps.com/bridge/tools/the-comerda-tool/.

[3] http://avresearcher.clariah.beeldengeluid.nl/; username: j.vangorp@uu.nl, password: AVResearcherXL. Voorheen bekend onder de namen QuaMeRDES/MeRDES, net als CoMeRDa een tool uit het NWO CATCH project BRIDGE, http://ilps.science.uva.nl/node/735. QuaMeRDES is met CLARIN-NL en CLARIAH-Seed-geld doorontwikkeld tot AVResearcherXL.

[4] http://trove.beeldengeluid.nl/; username: trove, paswoord: beg_evort. CLARIAH Seed project.

[5] http://dive.frontwise.com/. Over het project: http://dhcommons.org/projects/dive-dynamically-linking-collections-basis-events. Voorheen bekend onder de naam AGORA, NWO CATCH project.

[6] http://www.verteldverleden.org/. Voorheen bekend onder de naam Oral History Today (OHT), CLARIAH Seed project.

	Core). Uitbreiding met sentimentanalyse en voorspellende analytics van subjectieve aspecten (via extern gefinancierd onderzoeksproject).

De agenda is gericht op de realisatie van de volgende doelen:

- Inrichten van een duurzame beheer-, ontwikkel- en productieomgeving bij B&G waarin de bovengenoemde tools zullen landen
- Consolidatie van de tools (stabiel, goede functionaliteit)
- Doorontwikkeling van de tools (logging en opschalen naar grotere gebruikersaantallen; extra data & functionaliteiten: *data and tool curation*)
- Verankering van de consolidatie, ontwikkeling en doorontwikkeling in gebruikers-scenario's middels user studies en *Research Pilots*
- Overlap in functionaliteit opsporen en oplossen (via overkoepelend project Media Studies Suite en op basis van uitkomsten *Research Pilots*)
- Nieuwe generatie tool ontwikkelen (op basis van bestaande tool(s), *met additionele middelen*)

Omdat bovenstaande tools tot nu toe nog niet breed zijn ingezet, willen we ze in een aantal ‘geoormerkte’ Research Pilots testen met verschillende gebruikersgroepen binnen de geesteswetenschappen. Daarnaast bieden de open Research Pilots ruimte voor de aanpassing van bovenstaande tools op andere data en functionaliteiten.

Scope

Wat leveren we concreet af?

Een **set van tools** voor het doorzoeken, analyseren en visualiseren van audiovisuele collecties en contextuele data die zijn afgestemd op de werkwijze van (mediawetenschappelijke) onderzoekers en de eisen van de specifieke collecties (zoals auteursrecht en privacy). Deze set van tools wordt gemodelleerd naar enerzijds de fases van het (mediawetenschappelijk) onderzoeksproces en anderzijds naar hun functionaliteit in relatie tot specifieke typen collecties. Dit is gevisualiseerd in onderstaand schema:

Functionaliteit en collectie / Onderzoeksfase	Doorzoeken audiovisuele collecties in samenhang met gepubliceerde bronnen	Exploreren van audiovisuele collecties via relevante online informatie	Analyseren van de verspreiding van nieuws over verschillende mediabronnen	Doorzoeken en annoteren van narratieve interview-collecties
Exploratie	CoMeRDa	DIVE		Verteld Verleden
Contextualisering	AVResearcherXL		TROVe	
Presentatie[7]				

Op basis van een analyse van het gebruik van deze vijf tools wordt **een zesde, nieuwe generatie tool** ontwikkeld (Media Studies Suite), waarin mogelijke overlap in functionaliteit tussen twee of meer van de tools wordt opgelost. Daarnaast zullen we die nieuwe tool verrijken met functionaliteiten die de inhoudsanalyse van de bronnen ondersteunen, zoals sentimentanalyse en het integreren van voorspellende analytics met betrekking tot subjectieve aspecten. Voor de integratie van deze *state of the art* technologie wordt in 2015/16 financiering voor fundamenteel onderzoek aangevraagd bij NWO (derde ronde Creatieve Industrie, Open Competitie).

Een productieomgeving voor de implementatie en hosting van de tools alsmede de curatie van audiovisuele en contextuele data bij Beeld en Geluid. Tevens wordt hier een **loket** ingericht waar gebruikers vragen kunnen stellen over gebruik van de tools, dataconversie, licenties etc. Daarnaast wordt een gedistribueerde ontwikkelomgeving ingericht bij de ontwikkelaars van de gebruikte analyseomgevingen.

[7] Momenteel bevatten de vijf tools nog geen mogelijkheden om de gevonden resultaten te exporteren ten behoeve van publicaties. Het inbouwen van deze functionaliteit is onderdeel van de CLARIAH Core Mediastudies-agenda.

De vijf bestaande CLARIAH mediastudies tools bestaan uit twee delen:

1. Een generieke analyseomgeving
2. Toegevoegde data en interfaces

Ad 1: De generieke analyseomgevingen zullen worden beheerd door de onderzoeksgroepen die ze hebben ontwikkeld:

xTAS	UvA Computer Science
ClioPatria semantic search	VU Computer Science
SHoUT	UT Computer Science

Hiertoe zal Beeld en Geluid backend beheerders aanstellen die parttime bij de betreffende onderzoeksgroepen zullen werken aan het onderhoud van de analyseomgevingen.

Ad 2: De bestaande tools zijn gebouwd op de collecties van Beeld en Geluid (omroepmateriaal, foto's, programmabeschrijvingen, omroepgidsen) en contextuele data uit andere collecties (digitale kranten, online informatie), beiden in het Nederlandstalige domein. Daarmee zijn deze tools na implementatie (zie onder 2, Projecten, taken en budget) vooral geschikt voor onderzoek naar aan het nieuws gerelateerde evenementen en naar processen van beeldvorming zoals die via de Nederlandstalige media plaatsvinden. Hiermee zijn ze interessant voor brede groepen onderzoekers in de geesteswetenschappen (mediawetenschappers, historici, kunsthistorici, literatuurwetenschappers, cultural studies onderzoekers) en sociale wetenschappen (communicatiewetenschappers, sociologen, politicologen, antropologen).[8] De opzet van de tools maakt ze echter ook geschikt voor gebruik met andere collecties, zoals internationale collecties radio- en televisiemateriaal, kranten, Wiki's en blogs, evenals collecties films en online video's (YouTube, Vimeo). De productieomgeving bij Beeld en Geluid biedt ondersteuning bij het implementeren van nieuwe data in de bestaande tools.[9]

Tot het project behoort niet:

- Het op het meest basale niveau geschikt maken van nieuwe collecties voor implementatie in de tools (Beeld en Geluid adviseert over standaarden maar kan niet alle conversie zelf uitvoeren);
- Het vormen van nieuwe datasets;
- Het linken van de data over alle onderdelen van CLARIAH (zit in WP2).

[8] Na inbouw van geplande data exportfunctionaliteiten bieden de tools ook toegang tot voor bijvoorbeeld taalkundigen en psychologen relevante collecties gesproken taal, in veel gevallen voorzien van transcripties.
[9] Verteld verleden bevat naast interviews uit de Beeld en Geluid collectie ook al collecties uit diverse andere Nederlandse archieven en is specifiek gericht op een type bron: het narratieve interview, die gebruikt wordt door diverse groepen onderzoekers in de geesteswetenschappen en sociale wetenschappen.

Verwachtingen
We verwachten dat onderzoekers uit de mediastudies aanvankelijk vooral zullen werken met de in CLARIAH aanwezige data en tools. Daarnaast verwachten we dat onderzoekers eigen data in de CLARIAH mediastudies tools zullen willen gebruiken. Ten slotte verwachten we dat onderzoekers *exports* van resultaten uit de mediastudies tools (ten behoeve van de analyse en presentatie) zullen willen combineren met binnen CLARIAH beschikbare tekstuele of gestructureerde data en/of zullen willen invoeren in andere binnen CLARIAH beschikbare tools, bijvoorbeeld voor tekstanalyse of visualisatie.

Randvoorwaarden
Het inrichten van een duurzame beheer-, ontwikkel- en productieomgeving bij Beeld en Geluid is een voorwaarde voor het kunnen realiseren van deze agenda. Omdat Beeld en Geluid in tegenstelling tot de meeste andere datacentra in CLARIAH nog geen gecertificeerd datacentrum is en nog niet volledig beschikt over de benodigde voorzieningen, moet hiervoor een groot deel van de beschikbare middelen worden gereserveerd.

Relaties met andere projecten
1. CLARIAH Core WP2: aansluiting op de centrale infrastructuur (authenticatie en autorisatie), koppeling met andere aanwezige datacollecties via linked open data representaties.
2. CLARIAH Core WP3 en 4: Participatie in twee domeinoverstijgende projecten:
    a. Ontwikkeling tool voor automatische extractie gestructureerde data uit tekst (met als casus de gegevens over bioscopen, films, vertoningsdata in de filmladders van de krantencollectie KB)
    b. Athena project: verzamelen historische data over de relatie tussen mens en natuur
3. CATCH valorisatieproject: toevoegen van programmagidsen aan CoMeRDa
4. Er zal samenwerking worden gezocht met Nederlab voor het toevoegen van de ondertitelfiles aan AVResearcherXL (taak 2.2.10.3, zie onder 2, projecten).

Aansluiting bij DARIAH-EU
Voor het terrein van Mediastudies/audiovisuele bronnen ligt aansluiting bij de DARIAH-EU ERIC het meest voor de hand. Momenteel is de agenda van DARIAH-NL nog in ontwikkeling - In overleg met Peter Doorn wordt een bijeenkomst gepland met alle betrokkenen om te bepalen hoe de aansluiting het beste tot stand kan worden gebracht. We verwachten met name bij te dragen aan activiteiten op het gebied van educatie en training. Binnen CLARIAH Core Mediastudies zal Stef Scagliola optreden als DARIAH-NL liaison (zij is al actief in DARIAH-NL).

Disseminatie
Gezien het feit dat de vijf gekozen tools nog niet beschikbaar zijn voor brede groepen onderzoekers en studenten, zullen we diverse activiteiten organiseren voor de bekendmaking en het gebruik ervan in onderzoek, in de vorm van presentaties, workshops en gastlessen. Ook de geplande gebruikersstudies (zie onder 2, bij project 5) bieden, naast input voor de verdere

ontwikkeling van de tools, ook disseminatiemogelijkheden. Voor deze activiteiten gaan we nauw samenwerken met een adviesraad die de achterban in de Mediastudies vertegenwoordigt (zie onder 3, organisatie), evenals met de relevante onderzoeksscholen (onder meer Research School for Media Studies, Huizinga Instituut voor Cultuurgeschiedenis, Onderzoeksschool Kunstgeschiedenis). De agenda voor deze activiteiten wordt door de disseminatie-liaison (zie onder 3, organisatie) afgestemd met de disseminatiecoördinatoren in WP1.

2. Projecten en taken

De beschreven doelstellingen worden ondergebracht in vijf projecten:

Project		Projectleider
1	Support/productie/ontwikkelomgevingen Beeld en Geluid	Johan Oomen (Beeld en Geluid)
2	2.1 CoMeRDa 2.2 AVResearcherXL 2.3 TROVe	Maarten de Rijke (UvA-CS) Jasmijn van Gorp (UU-GW)
3	DIVE	Lora Aroyo/Victor de Boer (VU-CS) Chiel van den Akker (VU-GW)
4	Verteld Verleden	Roeland Ordelman (UTwente-CS) Stef Scagliola (UvA-GW)
5	Media Studies Suite	Thomas Poell (UvA-GW) t.b.d. (postdoc 0,7; UvA-GW)

Hieronder wordt elk project in meer detail besproken. Daarbij is bij alle taken een prioritering aangegeven: *must haves* (I) en *nice to have* (II).

PROJECT 1. SUPPORT-, PRODUCTIE- EN ONTWIKKELOMGEVINGEN

Om zijn rol als data center in CLARIAH te kunnen vervullen, moet Beeld en Geluid support-, productie- en ontwikkelomgevingen inrichten. Dit omvat de volgende aspecten:

1.10 Backend beheer (beheerder zit parttime bij universitaire ontwikkelaars van de tools) (I)
1.20 Data beheer (technisch, formaten, standaarden, conversies) (I)
1.30 Data beheer (inhoudelijk – selecties) (I)
1.40 Inzet van ontwikkelcapaciteit (I) ten behoeve van

- Inrichten test-, acceptatie en productie omgeving;
- CLARIN compatibel maken van van de infrastructuur (Beeld en Geluid moet nog gecertificeerd worden);
- Installeren van (taal) analysetools in productieomgevingen;
- Installeren en inrichten van zoekmachines en instroom verschillende data (via importers). Optimaliseren van deze omgevingen;
- Ontwikkelen van technische interfaces (tussen databronnen, en met resultaten WP2);
- Koppeling met Academia inrichten en persistent maken;
- Eerstelijnssupport voor de zoekomgevingen en supervisie bij uitbesteding aan derden.

1.50 Hardware en processing (I)
1.60 Bewaken sustainability van tools en beheer/supervisie bij uitbesteding aan derden (I)
1.70 Licenties afkopen (voor zover mogelijk CLARIAH breed) (I)
1.80 Hosted service voor tekstanalyse op basis van de xTAS suite[10] i.s.m. 904Labs[11] (I)

PROJECT 2. TOOLS VOOR SIMULTAAN DOORZOEKEN VAN AV EN GEPUBLICEERDE BRONNEN

2.1 CoMeRDa

Deze tool is vrijwel uitontwikkeld en klaar voor gebruik. Binnen het CATCH valorisatieproject worden momenteel nieuwe jaargangen omroepgidsen toegevoegd en de rechten geregeld. Binnen CLARIAH Core zullen de volgende zaken worden opgepakt:

2.1.10 Noodzakelijke ingrepen (I)
2.1.10.1 Gebruikershandleiding opstellen
2.1.10.2 Live verbinding iMMix en Beeld en Geluid Wiki
2.1.10.3 Rechten regelen voor fotocollecties en kranten
2.1.10.4 Toevoegen krantencollectie KB met live verbinding
2.1.10.5 Koppeling met Europeana maken
2.1.10.6 Toevoegen data export functionaliteit
2.1.10.7 Toevoegen ILPS-logging

2.1.20 Doorzoekbaarheid fotocollectie verbeteren (metadata) (I)

2.2 AVResearcherXL

Deze tool is getest met diverse gebruikers (o.a. mediawetenschappers, journalisten). Dit heeft een lijst met laatste noodzakelijke ingrepen opgeleverd. Daarnaast blijkt er grote behoefte aan een representatie van ontbrekende informatie, om de wel gevonden resultaten goed te kunnen interpreteren en de transparantie te verhogen. Deze tool kan profiteren van een geavanceerde tekst mining tool. Ook zullen nieuwe collecties worden toegevoegd (extra kranten en een radiocollectie).

2.2.10 noodzakelijke ingrepen (I)
2.2.10.1 automatische updates van de data via live verbinding iMMix en KB
2.2.10.2 verbinding met de streaming videos in Academia.nl
2.2.10.3 toevoegen van ondertitelfiles voor doorzoekbaarheid
2.2.10.4 representatie bronnen
2.2.10.5 toevoegen data export functionaliteit
2.2.10.6 toevoegen ILPS-logging

2.2.20 representatie van ontbrekende informatie (I)

[10] http://xtas.net/
[11] http://904labs.com/

2.2.30 geavanceerde tekst mining (iMMix en kranten) (II)

2.2.40 toevoegen nieuwe collecties (II)

2.3 TROVe

TROVe is een veelbelovende tool waar veel vraag naar is. Momenteel is het echter ook de minst uitontwikkelde tool. Eerste urgentie is het stabiliseren van de tool (*sustain liveness*) en het verbeteren van de performance, het aanbrengen van een live verbinding met iMMix, het toevoegen van ILPS-logging en het aanvullen van ontbrekende data en toevoegen van nieuwe collecties, als de kranten van de KB, radio-uitzendingen, etc. (zodat TROVe behalve voor de actualiteit ook bruikbaar is voor de analyse van meer historische publieke debatten). Daarnaast vereist gebruik binnen CLARIAH ook het regelen van rechten voor de in de tool opgenomen tweets, blogs, kranten en kijkcijfers van de SKO[12]. Wenselijk is de implementatie van een zelf-lerend systeem, waarbij zoektermen worden gebruikt voor het crawlen van data, of zoektermen door het systeem worden gesuggereerd, etc. Ten slotte moet een data export functionaliteit worden ingebouwd die te combineren is met diverse bestaande visualisatietools en publicatieformats.

2.3.10 Noodzakelijke ingrepen (I)
2.3.10.1 Monitoren systeem back-end / systeem om *liveness* te ondersteunen
2.3.10.2 Opnieuw ontwerpen normalisatie timeline front-end
2.3.10.3 Opnieuw ontwerpen facets front-end (inclusief apart ondertitelveld)
2.3.10.4 Invoegen functionaliteit voor kwantificeren “bereik” van de content via kijkcijfers, aantal retweets, aantal volgers, etc.
2.3.10.5 Live verbinding met iMMix
2.3.10.6 Rechten klaren tweets
2.3.10.7 Rechten klaren blogs
2.3.10.8 Rechten klaren kranten
2.3.10.9 Rechten klaren kijkcijfers SKO
2.3.10.10 Toevoegen extra jaargangen kijkcijfers (850 euro/jaargang)
2.3.10.11 Ontbrekende data toevoegen (nov. 2013 - jan. 2014 ontbreekt vanwege problemen server)
2.3.10.12 Toevoegen data export functionaliteit (I)
2.3.10.13 Toevoegen ILPS-logging (I)

2.3.20 Zelf-lerend systeem (II)

2.3.30 Toevoegen collecties (I)

2.3.40 Toevoegen kijk- en luisterdata (I)

[12] https://kijkonderzoek.nl/

3. DIVE
Deze tool is in testversie beschikbaar. Om hem te kunnen inzetten in exploratief onderzoek naar de context van specifieke *events* moeten de volgende zaken worden geregeld:

3.10 Noodzakelijke ingrepen (I)
3.10.1 Implementatie van event narratives in de front end
3.10.2 Links naar interne (GTAA) en externe vocabulaires (DBpedia, AAT, ULAN, Geonames)
3.10.3 Links naar collecties (Europeana, Open Cultuur Data)
3.10.4 Integratie van oorspronkelijke beschrijving van elk object
3.10.5 Toevoegen ILPS-logging

Naast deze noodzakelijke ingrepen zal de infrastructuur worden opgeschaald door een API te creëren voor toegang op de verrijkte data, login functionaliteiten in te bouwen en een module te importeren.

3.20 Opschaling (I)
3.20.1 API voor toegang op de verrijkte data
3.20.2 Inbouwen login functionaliteit en importeren module

4. Verteld Verleden
Deze tool is beschikbaar en werkt op de er nu in opgenomen collecties. Binnen CLARIAH Core wordt gewerkt aan een generieke ingest workflow van metadata met specifieke aandacht voor instroom en opslag van tijd-gecodeerde annotaties (handmatig, opgelijnd, automatisch). Daarnaast zal een infrastructuur worden opgezet voor het verwerken van audio via de centrale audio analysedienst (spraakherkenning, emotieherkenning, etc.). Met deze dienst kunnen onderzoekers collecties aanbieden die dan voorzien worden van transcripties.[13]

4.10 Generieke ingest metadata (I)
4.20 Audio-analysedienst (I)
4.30 Semi-automatische transcriptiedienst (II)
4.40 Toevoegen ILPS-logging (I)

Project 5. Mediastudies Suite
Bovenstaande tools zijn (door)ontwikkeld in verschillende, opeenvolgende projecten en maken bijna allemaal gebruik van de collecties en data van Beeld en Geluid en de KB. In dit vijfde, overkoepelende project wordt de overlap in functionaliteit in kaart gebracht. Uitkomst is een typologie van de fases in geesteswetenschappelijk onderzoek (exploratief, analytisch, etc.) waar de tools aan gekoppeld worden, en een op gebruikersonderzoek gebaseerde strategie voor integratie van tools met overlappende functionaliteit (zoals AVResearcherXL en TROVe).

[13] Deze dienst is momenteel opgezet als onderdeel van de spraakherkenningsinfrastructuur van Beeld en Geluid waaraan externe leveranciers zijn gekoppeld. Deze infrastructuur wordt uitgebreid zodat hier ook audio-analyse tools van andere partijen (andere leveranciers, vanuit samenwerking met academische partijen) aan kunnen worden gekoppeld.

Tevens wordt gestreefd naar doorontwikkeling van de geïntegreerde tool via het inbouwen van nieuwe, state of the art search technologie zoals analyse van subjectieve aspecten van informatie (sentiment analyse, framing) – hiervoor zal externe financiering worden gezocht (zie onder 1, Inleiding).

5.10 Gebruikersonderzoek (typologie fases onderzoek; opstellen *user requirements* per fase onderzoek en functionaliteit/type collectie) (I)
5.20 Samenvoegen tools met overlappende functionaliteit (I)
5.30 Integreren nieuwe functionaliteiten (II)
5.30.1 Cross-modaal zoeken
5.30.2 Ondersteuning van multi-sessie zoek- en onderzoekstrajecten
5.30.3 Ondersteuning van collaboratief zoeken

Het gebruikersonderzoek is gebaseerd op twee verschillende typen projecten:

A. Door de postdoc opgezette gebruikersstudies waarbij de bestaande tools worden getest met door gebruikers ingebrachte data (data en tool curatieprojecten zoals voorzien in project 1);
B. Een of meerdere van de Research Pilots die in 2017 zullen worden uitgevoerd.

Planning

	2015-1	2015-2	2015-3	2015-4	2016-1	2016-2	2016-3	2016-4	2017-1	2017-2	2017-3	2017-4	2018
Project 1: NIBG		1.10-1.70: inrichting support-, productie- en ontwikkelomgeving					1.80 xTAS						
Project 2.1: CoMeRDa	(Programmagidsen toevoegen, *CATCH-valorisatie project)		2.1.10: noodzakelijke ingrepen					2.1.20: fotocollectie					
Project 2.2: AVResearcherXL	DTC1: noodzakelijke ingrepen			2.2.10: noodzakelijke ingrepen			DTC2: missing information		2.2.30: geavanceerde tekst mining		2.2.40: toevoegen collecties		
Project 2.3: TROVe		2.3.10: noodzakelijke ingrepen						2.3.40: toevoegen kijk- en luisterdata	Research Pilot: verschillende gebruikersgroepen voor TROVe		2.3.20 zelf-lerend systeem		2.3.30: toevoegen collecties
Project 3: DIVE			3.10: noodzakelijke ingrepen								3.20 opschaling		
Project 4: Verteld Verleden					4.10: Generieke ingest metadata		4.20: Audio-analysedienst		Research Pilot: gebruikersgroepen Verteld Verleden sociale wetenschap		4.30: semi-automatische transcriptiedienst		
Project 5: Media Studies Suite			(aanvraag additionele financiering)	5.10: Use cases CoMeRDa en AVResearcher XL		5.10: Use cases DIVE		5.10: use cases Verteld Verleden GW			5.20: oplossen overlap in funcationliteit		5.30: integratie nieuwe functionaliteit
	2015-1	2015-2	2015-3	2015-4	2016-1	2016-2	2016-3	2016-4	2017-1	2017-2	2017-3	2017-4	2018

3. Organisatie

De hierboven omschreven projecten bevatten taken op de volgende terreinen:

- Infrastructuur
- Data curatie
- Tool curatie

Het plan voor audiovisuele data en mediastudies gaat uit van het stabiliseren, synthetiseren en verder ontwikkelen van bestaande tools. Daarom zijn de data en toolcuratie verdeeld over vijf afzonderlijke projecten, waarbij:
- project 1 zich richt op de curatie van nieuwe data;
- project 2-4 zich richten op de curatie van de bestaande tools en data;
- project 5 een overkoepelende project is bedoeld ervoor te zorgen dat de gebruikers centraal staan in de doorontwikkeling en niets dubbel wordt gedaan.

De bovenstaande data en toolcuratieprojecten worden uitgevoerd door interdisciplinaire teams, bestaande uit:

- GW onderzoeker
- CS onderzoeker
- Programmeur(s)/software engineer(s)
- Data provider

Bij de projecten 2-4 is ervoor gekozen de data en toolcuratie te laten uitvoeren onder leiding van de bij ontwikkeling van de respectievelijke tools betrokken onderzoekers. Bij alle projecten zijn de onderzoeksvragen en –praktijken van geesteswetenschappelijk onderzoekers leidend.

Dit levert de volgende verdeling van taken op:

Trackleider
Julia Noordegraaf (0,2 fte; Universiteit van Amsterdam-GW)

Projectleiders

Project 1:	Johan Oomen (zie onder Technisch coördinator)
Project 2:	Maarten de Rijke (0,1 fte; Universiteit van Amsterdam-CS) en Jasmijn van Gorp (0,2 fte; Universiteit Utrecht-GW)
Project 3:	Lora Aroyo/Victor de Boer (0,1; Vrije Universiteit-CS) en Chiel van de Akker (0,1 fte; Vrije Universiteit-GW)
Project 4:	Stef Scagliola (0,05 fte; Universiteit van Amsterdam-GW), tevens DARIAH-liaison (inhoudelijk)
Project 5:	Thomas Poell (0,1 fte; Universiteit van Amsterdam-GW) en postdoc (n.n.b.) (0,7 fte; Universiteit van Amsterdam-GW), tevens disseminatie-liaison WP1

Technisch coördinator
Johan Oomen (0,2 fte; Beeld en Geluid)

1. Verbinden van de verschillende mediastudies projecten
2. Verbinden van WP5 met andere werkpakketten, in het bijzonder WP2
3. Verbinden van WP5 met DARIAH-EU (technisch)
4. Vertegenwoordigen data centre Beeld en Geluid in technisch overleg o.l.v. CTO Gertjan Filarski

Software engineers en data beheerders

Project 1:	Ontwikkelaar (1,0 fte; Beeld en Geluid)
	Backend beheerder (0,7 fte; Beeld en Geluid)
	Data beheerder (inhoudelijk) (1,0 fte; Beeld en Geluid)
	Data beheerder (technisch) (0,5 fte in WP2; Beeld en Geluid)
Project 2:	n.n.b.
Project 3:	n.n.b.
Project 4:	n.n.b.
Project 5:	n.n.b.

Organisatiestructuur
Om te waarborgen dat bovenstaande agenda wordt gerealiseerd, hanteren we de volgende organisatiestructuur:

CLARIAH Core Mediastudies Kernteam
Rol en taken: voorbereiden strategie en beleid, regie op uitvoering
Overleg: tweewekelijks overleg (1x per maand met projectteam, 1x per maand informeel)
Betrokken personen:

- Julia Noordegraaf (track leader, voorzitter)
- Maarten de Rijke (namens Technisch Advies Panel)
- Johan Oomen (technisch coördinator, liaison WP2)
- Jasmijn van Gorp (liaison gebruikersstudies WP2)
- Thomas Poell (liaison disseminatie WP1)

CLARIAH Core Mediastudies Projectteam
Rol en taken: uitvoering projecten, bewaken voortgang
Overleg: maandelijks

CLARIAH Core Mediastudies Adviesraad
Rol en taken: advies over strategie en beleid vanuit de beoogd gebruikers
Overleg: 2 á 3 keer per jaar
Betrokken personen: nog te benaderen onderzoekers binnen Mediastudies die de achterban representeren (EUR, RUG, RUN, UM, UU, UvA, VU).

4. Budget

Zie bijlage ‘CC WP5 begroting 16 maart 2015’.

De begroting reflecteert de taken zoals genoemd onder 2 en de coördinatie daarvan zoals genoemd onder 3. Er is gewerkt met de meest recent salaristabel van NWO (“Berekening vergoeding met salarispeil 1 juli 2014”[14]). Voor de nog aan te stellen ontwikkelaar (1.40) zal een verzoek worden gedaan bij NWO om de daadwerkelijke salariskosten in rekening te mogen brengen, aangezien hier sprake is van bijzondere expertise noodzakelijk voor de realisering van de faciliteit die niet binnen de normbedragen te verkrijgen is.

[14] http://www.nwo.nl/financiering/hoe-werkt-dat/Salaristabellen

5. Referenties

Bron, M., Van Gorp, J., Nack, F. F., de Rijke, M., & Baltussen, L. B. (2013) Aggregated search interfaces in multi-session tasks. *SIGIR 2013: 36th international ACM SIGIR conference on research and development in information retrieval.* Dublin: ACM.

Bron, M., Van Gorp, J., de Rijke, M. (2015) Media studies research in the data-driven age: How research questions evolve. *Journal of the American Society for Information Science and Technology*. Vol 66, issue 12 (forthcoming).