Deliverable

Project: IIIF-Beeldinfrastructuur gekoppeld aan de VKC-Datahub. Fase 1: proefopstelling
Werkpakket: WP2 Selectie van IIIF-compatibele componenten.

Subsidie: Projectsubsidie voor een internationaal cultureel-erfgoedproject
Periode: 2 juli 2018 - 31 augustus 2019

D1: Architectuur van de proefopstelling

Auteurs: Matthias Vandermaesen (VKC), Jolan Wuyts (VKC), Pascal Ennaert (VKC), Sarah Eycken (VKC) en Nastasia Vanderperren (PACKED (VIAA+))

Doel van deze deliverable

Deliverable D1 heeft als doel de architectuur van Fase 1 van de IIIF-beeldinfrastructuur op te lijsten en te bespreken. De componenten die in het project gebruikt worden, zijn geselecteerd na uitgebreid overleg en consultatie van standaarden en na vergelijking van verschillende potentiële componenten. De geselecteerde componenten zijn doorheen het project getest en indien nodig herzien. Het volledige proces is beschreven in dit document. Deze deliverable kan als leidraad dienen voor volgende fases van het IIIF-beeldinfrastructuurproject. Andere culturele instellingen of dienstverlenende organisaties kunnen de deliverable gebruiken bij het kiezen van technische componenten voor het opzetten van een IIIF-beeldinfrastructuur.

Scope van het project

Aanleiding

Sinds de doorlichting van het digitaal ecosysteem van de Vlaamse Kunstcollectie (VKC) door PACKED vzw in 2012, werkt de VKC gestaag aan een flexibel en open informatiesysteem. Een dergelijk systeem heeft verschillende doelstellingen. Enerzijds moet het de geautomatiseerde doorstroom mogelijk maken van gedigitaliseerde content (data en gedigitaliseerde media (voornamelijk afbeeldingen)) uit de musea (en hun digitale infrastructuur) naar de hergebruikstoepassingen van die musea zelf (websites, digitale bezoekersgidsen,...). Anderzijds moet de geautomatiseerde doorstroom van diezelfde bronnen ook mogelijk zijn naar tal van andere externe hergebruikers (aggregatoren, online tools, onderzoeksomgevingen…). Zo wordt de meest actuele en beschikbare content steeds op een kosten- en tijd-efficiënte manier ter beschikking gesteld.

De geautomatiseerde doorstroom kan niet in een enkele stap van collectieregistratiesystemen naar hergebruikstoepassingen gebeuren. Daarvoor is er een uitwisselingslaag nodig: een laag met middleware die content van de collectieregistratiesystemen verzamelt, verrijkt, structureert en ontsluit. Die uitwisselingslaag dient als onderbouw voor het laten functioneren van online hergebruikstoepassingen.

De VKC, met de steun van PACKED vzw, hecht veel belang aan de duurzame onderbouw die nodig is om tal van hergebruikstoepassingen te laten functioneren. Het open en flexibel informatiesysteem voorziet twee essentiële delen: een datahubcomponent en een beeldinfrastructuurcomponent. Immers, de registratiegegevens zitten vandaag opgesloten in registratiesystemen. Musea of derden kunnen ze niet consumeren in andere toepassingen zonder stevige investeringen. De datahubcomponent zet in op een automatische, flexibele uitwisseling op basis van open standaarden - zoals LIDO XML - om dit probleem op te lossen. De beeldinfrastructuurcomponent automatiseert de uitwisseling van beeldmateriaal.

Die uitwisseling gebeurt vandaag zeer moeizaam en is arbeidsintensief. De oorzaak van dit probleem zijn de vele alom gebruikte manuele processen (e-mail, WeTransfer, USB-sticks,

harde schijf,...) om beeldmateriaal te delen . Door een beeldinfrastructuurcomponent te compileren aan de hand van open standaarden, willen de projectpartners dat probleem verhelpen. Er wordt hier geïnvesteerd in de bouw van een infrastructuur die beantwoordt aan de IIIF-standaard. IIIF staat voor International Image Interoperability Framework en is uitgegroeid tot de internationale standaard voor gedistribueerd beeldbeheer. (zie: http://iiif.io)

Dit projectvoorstel voor een IIIF-beeldinfrastructuur is dus een logisch vervolg op de projectenreeks Datahub (1 & 2) die de VKC samen met PACKED vzw opzette. In dit project zetten we samen met PACKED vzw, het Rijksmuseum en andere beeldpartners de eerste stappen naar een vlotte, efficiënte en geautomatiseerde ontsluiting van beeldmateriaal in het belang van de huidige en toekomstige (her)gebruikers. We houden daarbij rekening met de (technische) aanbevelingen (de zgn. modelarchitectuur) uit de “Blauwdruk gedistribueerd beeldbeheer” zoals door PACKED vzw opgeleverd.

Doelstellingen van het project

De projectaanvraag omschreef in Fase 1 volgende doelstellingen:

- De verkenning van verschillende IIIF-beeldservercomponenten in de internationale academische en erfgoedwereld en de selectie van de meest bruikbare componenten in functie van de IIIF-beeldinfrastructuur.
- De compilatie van dergelijke IIIF-componenten in een proefopstelling van een beeldinfrastructuur, complementair aan de Datahub.
- De koppeling van dergelijke IIIF-beeldinfrastructuur aan de VKC-datahub in een proefopstelling.
- Een beknopte analyse van de koppeling met de onderbouw.
- De publicatie van de technische informatie van die proefopstelling op Github en de bekendmaking ervan op relevante nationale en internationale fora.
- Tegelijk het voorbereiden van de musea en andere leveranciers van beeldmateriaal op de noodzakelijke IIIF-metadata.

Onderzoeksvragen

Binnen de scope van de doelstellingen van het project willen we een aantal concrete onderzoeksvragen beantwoorden.

Vraag 1: Hoe zorgen we voor een performante en flexibele, geautomatiseerde ontsluiting van gedigitaliseerd beeldmateriaal via beeldservercomponenten die aan de IIIF Image API specificatie beantwoorden?

Momenteel blijven gedigitaliseerde beelden van cultureel erfgoed grotendeels opgesloten binnen de systemen van de musea zelf. Digitale reproducties/afbeeldingen van culturele erfgoedobjecten zijn van grote waarde voor hergebruik door en voor het bredere publiek,

onderzoekers, erfgoedprofessionals en andere hergebruikers.. Vaak hebben zij daar echter geen rechtstreekse en geautomatiseerde toegang toe. Niet alleen wil de VKC bekijken hoe deze beelden online ontsloten kunnen worden voor hergebruik. De VKC wil zich ook conformeren tot een aantal belangrijke criteria waaraan die ontsluiting zou moeten voldoen, met name performantie, flexibiliteit, automatisering en schaalbaarheid.

- **Performantie**

Ontsluiting van digitale beelden voor eindgebruikers heeft enkel waarde als die ontsluiting moeiteloos, snel, en in hoge resolutie kan gebeuren. Digitale reproducties online uitwisselen voor een groot publiek is echter niet eenvoudig, omdat beelden in hoge resolutie doorgaans zware bestanden zijn. De componenten die instaan voor het ontsluiten van beeldmateriaal moeten om die reden in staat zijn om zware beelden in hoge resolutie op een snelle manier geautomatiseerd aan te leveren wanneer ze worden opgevraagd. Tijdens het project willen we onderzoeken wat op het vlak van performantie de gevolgen en vereisten zijn om grote bestanden aan te leveren. Daaruit willen we dan weer de optimale balans vinden die nodig is voor het leveren van hoge kwaliteitsbeelden aan een hoge performantie.

- **Flexibiliteit**

Voor musea is het aangewezenom met software te werken die beelden ontsluit via open informatiesystemen. Dat zie je in steeds meer state-of-the-arttoepassingen in het buitenland (The Met, Cleveland Art Museum). Het informatiesysteem moet voor verschillende doelen en situaties kunnen worden ingezet. Dergelijke flexibiliteit moet ertoe leiden dat musea met een andere technische context diezelfde infrastructuur kunnen opzetten. Als een museum van een bepaald systeem overstapt naar een ander systeem zou dit ook geen probleem mogen vormen wat betreft de beeldinfrastructuur. De beeldinfrastructuur zou dus moeten kunnen inpassen op verschillende (DAM-)systemen. Het project wil onderzoeken welke componenten makkelijk inzetbaar zijn en uitgewisseld kunnen worden met andere componenten.

- **Automatisering**

Met automatisering wordt bedoeld dat de beeldinfrastructuur voor de uitwisseling van digitale beelden geen tussenkomst van museale medewerkers vergt. Musea, of derde organisaties die daartoe gemandateerd zijn, moeten beelden kunnen aanleveren op een gemakkelijke en intuïtieve manier, waarna de beeldinfrastructuur diezelfde beelden herkent en automatisch laat doorstromen naar middleware en hergebruiksapplicaties. Tijdens het selecteren van de componenten voor dit project is het belangrijk om ons af te vragen welke componenten kunnen zorgen voor een automatische doorstroom, alsook na te denken over hoe de architectuur van die doorstroom er moet uitzien.

- **Schaalbaarheid**

Voor dit project gebruikten we ongeveer tweehonderd afbeeldingen om een proefopstelling te testen. De afbeeldingen stelden een representatieve sample voor van gedigitaliseerde werken uit verschillende musea. In volgende fases van het project zou deze opstelling opgeschaald kunnen worden naar een groter volume. Als voor elk object dat momenteel in de Datahub zit minstens één gedigitaliseerde reproductie aanwezig is, dan bevat een productieomgeving van

de beeldinfrastructuur minstens 15.000 afbeeldingen. Bij de keuze van de componenten voor de proefopstelling is het dus belangrijk om met schaalbaarheid van de componenten rekening te houden. Een belangrijke vraag is hoe we componenten en een architectuur kunnen kiezen die performant blijven en beantwoorden aan de noden van een beeldinfrastructuur, zelfs als er duizenden afbeeldingen moeten verwerkt en ontsloten worden.

Vraag 2: Hoe zorgen we voor een performante en flexibele, geautomatiseerde ontsluiting van relevante inhoudelijke, technische en administratieve metadata die aan de IIIF Presentation API specificatie voldoet?

De projecten rond de Datahub hadden als onderwerp de ontsluiting van metadata gekoppeld aan culturele objecten. Maar ook bij het maken van digitale reproducties van culturele objecten worden objectspecifieke en beeldspecifieke metadata aangemaakt de titel van het werk, de vervaardiger, datum van vervaardiging, maar ook de datum van reproductie van het kunstobject, de maker/fotograaf ervan, het gebruikte fototoestel, de resolutie van een afbeelding, de vindplaats van het moederbestand van de foto, de persistente URI, de auteursrechtenstatus, de externe instelling of interne afdeling die het beeldbeheer waarneemt, de eventuele hergebruiklicentie die de voorwaarden duidt waaronder het beeld voor hergebruik ter beschikking wordt gesteld, … Die metadata moeten ook gestructureerd, bewaard en ontsloten worden. Ze voegen immers belangrijke waarde toe aan de digitale reproductie van het werk. IIIF voorziet in een afsprakenkader hoe dergelijke metadata kunnen/dienen ontsloten te worden. Een tweede onderzoeksvraag, die we in dit project willen beantwoorden, is welke weg de metadata die ontstaan bij digitalisering van een cultureel object moeten afleggen voor ontsluiting via IIIF.

Vraag 3: Kunnen we meertalig verrijkte objectgegevens, zoals ontsloten via de Datahub in LIDO XML-formaat, valoriseren binnen de proefopstelling?

In de context van Belgisch/Vlaams cultureel erfgoed komen we onvermijdelijk in aanraking met meertalige metadata.Titels, beschrijvingen, namen van instellingen en meer worden in collectieregistratiesystemen meertalig ingegeven. Het is belangrijk om de toegevoegde waarde die deze meertalige metadata creëren mee te nemen in het ontsluiten van de digitale objecten om zo een breder internationaal publiek aan te spreken. Technische componenten zijn daarentegen niet altijd uitgerust met functionaliteiten die meertaligheid mogelijk maken, bijvoorbeeld omdat ze ontwikkeld werden in een context waar meertaligheid geen prioriteit is. Dit is heel vaak het geval bij applicaties die ontwikkeld zijn in Engelstalige landen. Drie belangrijke onderzoeksvragen die we ons stellen, zijn dus welke componenten meertaligheid wél ondersteunen, hoe ze dit doen, en of ze compatibel zijn met andere componenten in de proefopstelling. Ook kunnen we onderzoeken welke meertalige authorities (AAT, RKDartists, Wikidata, …) de verzamelde metadata op een geautomatiseerde wijze kunnen verrijken.

Vraag 4: Wat is een werkbaar beleid voor de toepassing van auteursrechten gekoppeld aan de ontsluiting van beeldmateriaal via een IIIF compatibele beeldinfrastructuur?

Een laatste onderzoeksvraag in dit project heeft te maken met auteursrechten. Langs de ene kant is er auteursrechtelijke informatie verbonden aan een cultureel object en de metadata van dat object. Hier is zeker nog werk rond te verrichten. In dit project willen we vooral kijken naar auteursrechtelijke informatie die vasthangt aan gedigitaliseerd beeldmateriaal van culturele objecten. Die auteursrechten creëren een extra laag van complexiteit voor digitaal cultureel erfgoed. Tijdens het digitalisatieproces kan een fotograaf, in bepaalde gevallen, auteursrecht claimen op een digitale reproductie van een werk, ook al ligt dat werk zelf in het publieke domein of is het op een andere manier al helemaal vrijgesteld van auteursrecht. In dit project houden we ons bezig met die extra complexiteit en stellen we ons de vraag hoe we auteursrechten op digitaal beeldmateriaal kunnen vaststellen, structureren, borgen en op een duidelijke manier ontsluiten. Het einddoel daarvan is om aan gebruikers op een heldere en eenduidige manier mee te delen of ze een beeld kunnen gebruiken, en op welke manier ze dat dan mogen doen. Bij het maken van deze oefening houden we ons zo veel mogelijk aan de stipulaties van het gezamenlijk standpunt beeldbeleid, zoals dat bij de partnermusea van de VKC tot stand kwam.

Omschrijving van de proefopstelling

Scope van de proefopstelling

De scope van de proefopstelling bestaat uit de bouw van een infrastructuur op basis van diverse, reeds bestaande, softwaretoepassingen en -bibliotheken. De voorkeur gaat uit naar pakketten onder een open source licentie. Door de componenten te ‘compileren’, komen we tot een geheel dat voldoet aan de vereisten van de vooropgestelde business case.

De business case bestaat uit de online ontsluiting van beeldmateriaal en de bijhorende technische, administratieve en inhoudelijke metadata volgens de regels vooropgesteld door de IIIF-specificatie. Concreet gaat het om de ontsluiting van gedigitaliseerde beelden van de kunst- en erfgoedobjecten aanwezig in de musea voor schone kunsten. In de eerste plaats gaat de aandacht uit naar de ontsluiting van beelden op hoge kwaliteit, conform het geldende auteursrechtelijke kader. Om de onderzoeksvragen goed te kunnen beantwoorden werden ook beelden van objecten die nog in copyright zijn aan de proefopstelling toegevoegd.

Aangezien het om een proefopstelling gaat, beperkten we de selectie tot tweehonderd werken waarvan digitaal beeldmateriaal voorhanden is. Deze werken werden geselecteerd o.m. op basis van volgende vereisten:

- Er moet diversiteit zijn in auteursrechtelijke status (in copyright of publiek domein);
- Er moet een voldoende grote en diverse set van werken zijn, die structureel gerelateerd zijn aan elkaar (bv. schetsboeken, veelluiken, etc.)
- De afbeeldingen moet uit diverse instellingen komen;
- De afbeeldingen moeten uit diverse periodes (15de - 20ste eeuw) stammen;
- Er moeten zowel topstukken en sleutelwerken, alsook minder bekende werken in de selectie zitten.

Naast ontsluiting worden ook de eerste stappen gezet wat betreft de aansluiting op de onderbouw. Er worden verschillende technische oplossingen voorzien om beeldmateriaal in de proefopstelling op te laden en te beheren. De scope van het proefopstelling voorzag initieel geen DAM-opstelling.

Dat toch een DAM-opstelling werd voorzien, heeft twee redenen:

- Via de subsidielijn *"Inhaalbeweging digitale collectieregistratie"* (2018) werd binnen het KMSKA een traject opgestart om een duurzame oplossing voor beeldbeheer te operationaliseren. Om de middelen van zowel de Vlaamse Kunstcollectie en haar partners, als het KMSKA optimaal te kunnen benutten, sloegen we de handen in elkaar en breidden we de proefopstelling uit met een beheercomponent (DAM), zodat het KMSKA de opstelling ook kon inzetten.
- Lukas, Art in Flanders beschikte nog niet over een operationele DAM waarop kon worden ingeplugd.

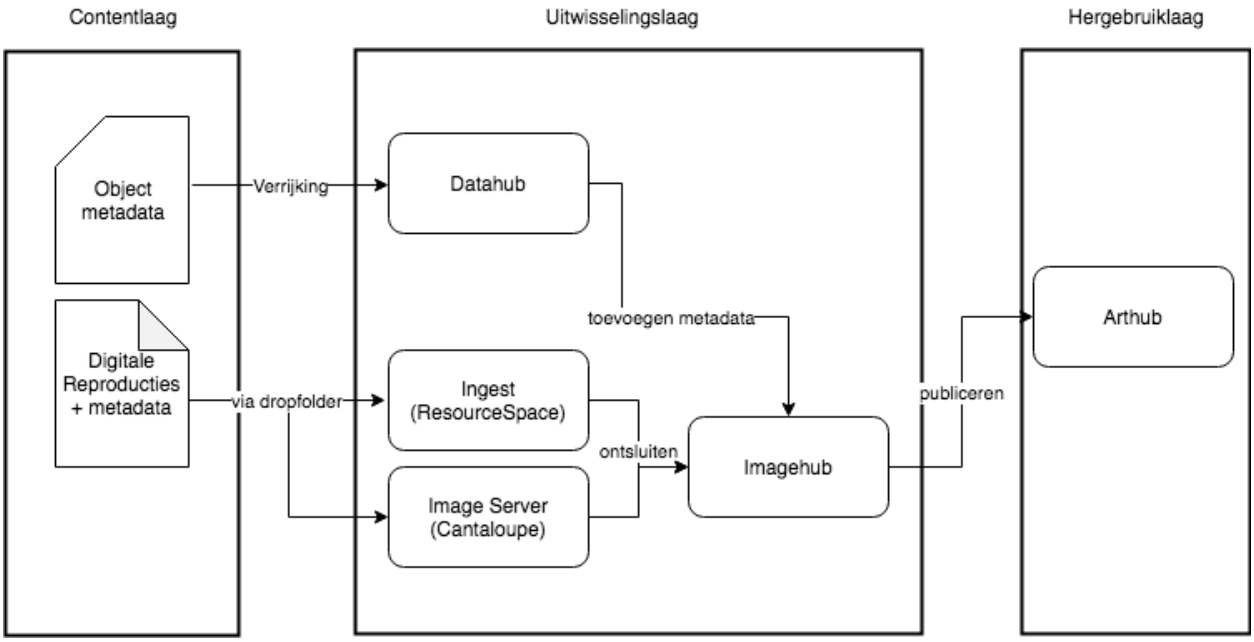
Authenticatie werd binnen dit project niet opgenomen. Alle beelden die via de proefopstelling ontsloten worden, zijn openbaar en worden niet beschermd door een login of een andere vorm van authenticatie.

Basiscomponenten

De architectuur van de proefopstelling bestaat uit diverse componenten die o.a. via ETL-pijplijnen met elkaar zijn gekoppeld. Die pipelines zorgen voor een geautomatiseerde uitwisseling van data en beeldmateriaal tussen de componenten, de onderbouw en de eindgebruikers.

We gaan uit van een architectuur die 'service oriented' is. Elke component verricht een op zichzelf staande dienst, en is verbonden met pipelines. Het belang van een servicegeoriënteerde architectuur ligt erin dat elke component op zichzelf staat en kan ingewisseld worden voor een andere component, zonder dat de overkoepelende structuur moet worden aangepast.

Schematisch ziet de architectuur er als volgt uit:



Elk afgerond vierkant in het schema stelt een onafhankelijke, op zichzelf staande component voor. Elke pijl staat voor een pipeline die de componenten met elkaar verbindt en data doorgeeft van de ene naar de andere component. De componenten zijn opgedeeld in drie lagen. De contentlaag bevat de digitale reproducties en de metadata, die aangemaakt en beheerd worden in de collectieregistratiesystemen en in de DAM's van de musea. De uitwisselingslaag verzorgt de structurering, verrijking en ontsluiting van die data. Eindgebruikers komen enkel in contact met de hergebruikslaag waarin digitale reproducties en metadata ontsloten worden voor het publiek. Het uiteindelijke doel is ervoor zorgen dat de uitwisselingslaag gebruikt kan worden om verschillende hergebruikstoepassingen te ondersteunen/voeden. In dit project zetten wede Arthub in als hergebruikstoepassing . De componenten, hierna beschreven in deze deliverable, zijn:

- de Datahub module;
- de Ingest module, ook wel het DAM-systeem genoemd;
- de Imagehub module;
- de Arthub module.

Daarnaast volgt ook een korte beschrijving van de pipelines.

Datahub module

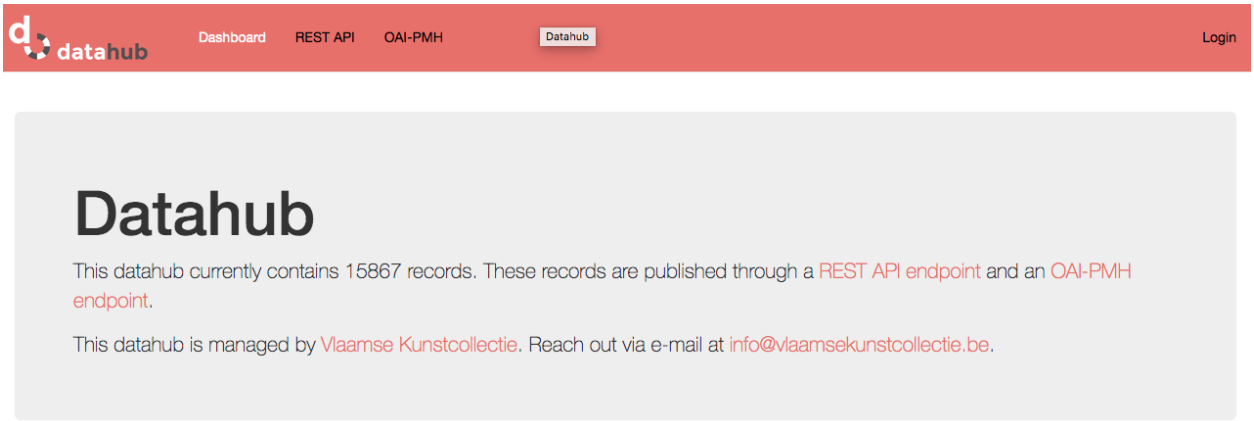
De proefopstelling bevat gedigitaliseerd beeldmateriaal van kunst- en erfgoedobjecten waarvan de inhoudelijke, vaktechnische (bijvoorbeeld gebruikte materialen) en administratieve metadatering gebeurt in de collectieregistratiesystemen (Adlib, TMS). Deze systemen vormen de centrale inventarissen van de musea. De data in die systemen wordt via de Datahub ontsloten in het LIDO XML-formaat via het OAI-PMH harvesting protocol.

De ruwe data bevatten evenwel een aantal lacunes die het gebruik in de beeldinfrastructuur belemmeren. De relaties tussen de fysieke objecten zijn vaak onvoldoende gecapteerd in metadata uit de registratiesystemen. Het gaat om structurele metadata: bv. linkerluik, voorzijde, achterzijde, middenpaneel, de voorkant van folio 10, etc.

Om die reden voegden we aan de proefopstelling een tweede datahub toe die enkel de metadata van de tweehonderd geselecteerde fysieke objecten bevat. De ruwe metadata werd daarbij verrijkt met extra metadata met oog op een 'ideale' ontsluiting. Het normaliseren, standaardiseren en verrijken van de metadata in de registratiesystemen behoorde immers niet tot de scope van dit project. Bovendien zou het verwerken van de aanpassingen in de registratiesystemen voor een onnodige vertraging hebben gezorgd.

Deze tweede datahub fungeert in de eerste fase als de bron waaruit de proefopstelling inhoudelijke, vaktechnische en administratieve metadata put.

In een volgende fase zou deze tijdelijke installatie verwijderd moeten worden , de noodzakelijke verrijkende metadata in de registratiesystemen toegevoegd en via de oorspronkelijke Datahub ontsloten.



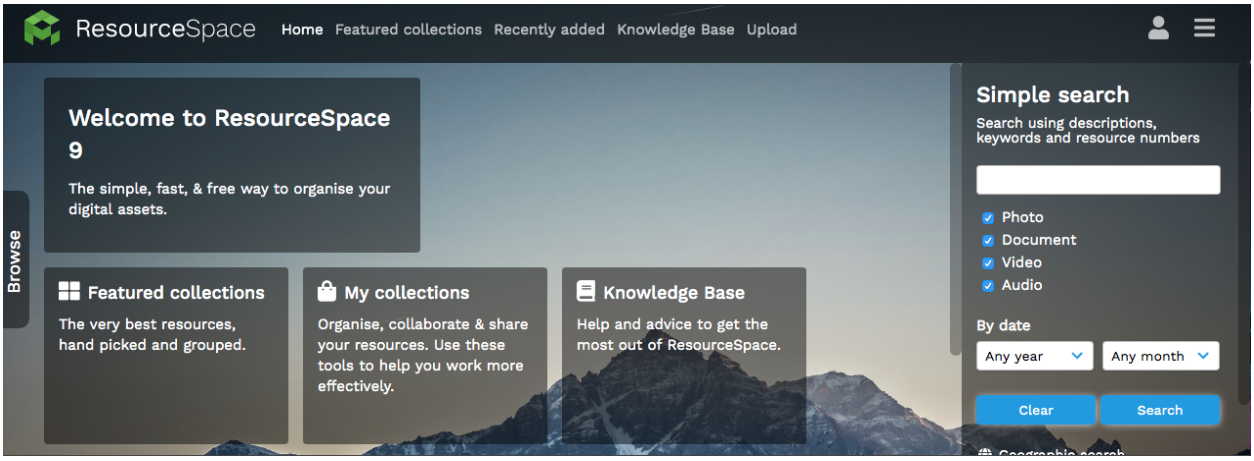
Ingest module

De ingest module voorziet het **opladen en het beheren van beeldmateriaal** in de proefopstelling. Oorspronkelijk was deze module beperkt tot het opladen van beeldmateriaal en het koppelen met een onderbouw. Aangezien het voorlopig ontbreken van een DAM bij VIAA-Lukas (thans VIAA+) een probleem vormde voor de samenwerking met het KMSKA, waren we genoodzaakt om deze module uit te breiden en een bescheiden beheer van het door VIAA+ (Lukas) aangeleverde beeldmateriaal op te zetten.
ResourceSpace werd gekozen als ingest module. Het is een systeem waarin beelden opgeladen en beheerd kunnen worden. In tegenstelling tot een collectieregistratiesysteem worden géén fysieke objecten, maar digitale beelden beschreven. Dit verschil in perspectief

zorgde ervoor dat er een apart metadatamodel, waarbij digitale beelden centraal staan, als vertrekpunt werd genomen. Binnen het traject "*inhaalbeweging digitale collectieregistratie*" ontwikkelde het KMSKA reeds een metadatamodel dat alle belangrijke metadatavelden inbedt. In dit project gebruikten we datzelfde door het KMSKA ontwikkelde metadatamodel, alsook het door het KMSKA ingezette DAM-systeem, zodat de gebouwde beeldinfrastructuur probleemloos toepasbaar is op de componenten die het KMSKA ontwikkelde. In een latere doorontwikkeling moet de beeldinfrastructuur kunnen aansluiten op diverse soorten DAM's zoals ook de datahub aansluit op diverse registratiesystemen.

ResourceSpace komt in elk geval tegemoet aan de volgende gebruikersscenario's:

- Een museummedewerker wil een groot aantal digitale reproducties in één keer inladen in de beeldinfrastructuur. De medewerker doet dit door de beelden in een folder te plaatsen die daarvoor is klaargezet. De beeldinfrastructuur zorgt ervoor dat de beelden van die folder naar ResourceSpace en de rest van de beeldinfrastructuur verplaatst worden en worden verrijkt met metadata uit de Datahub.
- Een museale medewerker wil een klein aantal foto's opladen in ResourceSpace en de rest van de beeldinfrastructuur. De museale medewerker doet dit door de afbeeldingen direct in ResourceSpace in te laden gebruikmakend van het ResourceSpace-uploadformulier. De beeldinfrastructuur verrijkt de data met metadata uit de datahub.
- Een afbeelding dient gepaard te gaan met ingebedde metadata. De beeldinfrastuctuur bewaart die metadata in het metadatamodel van de ingest module. De metadata wordt bewaard en gebruikt in de IIIF Presentation API.



Imagehub module

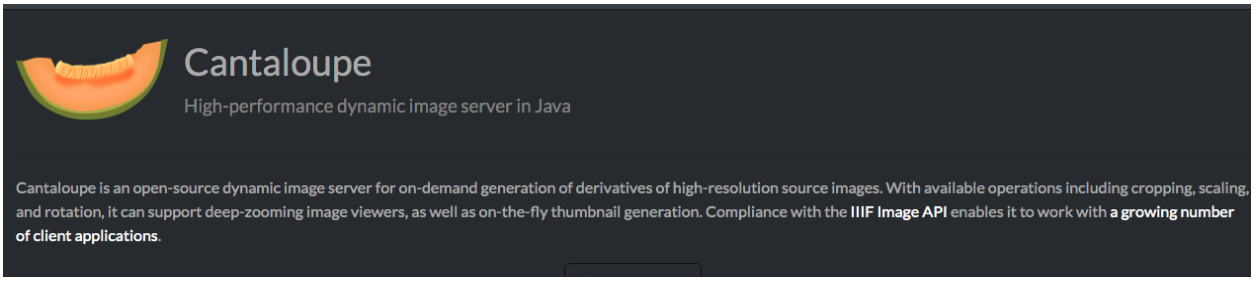
De Imagehub kent twee belangrijke onderdelen: De IIIF Image API en de IIIF Presentation API

IIIF Image API

De IIIF community formuleert een set van voorwaarden waaraan een Image API moet voldoen. Een IIIF Image API laat toe dat beelden worden uitgewisseld via een URI. De URI maakt het mogelijk om beelden op te vragen. Via een URI kunnen de grootte, uitsnijding, kwaliteit en

bestandsformaat van het opgevraagd beeld verder gespecificeerd worden. Er is een groot aanbod van open source beeldenservers die beelden via de IIIF Image API verspreiden. Voor dit project is het belangrijk dat de Image API grote hoeveelheden TIFF-bestanden op een snelle en performante manier kan aanbieden aan hergebruiksapplicaties. Voor dit project kozen we daarom voor de Cantaloupe Image Server. Cantaloupe is een dynamische open source image server die voldoet aan de IIIF Image API-specificatie. Ze garandeert een hoge performantie in het aanbieden van afbeeldingen, waaronder ook in het TIFF-formaat.

De image server heeft als doel om afbeeldingen uit de ingest module te ontsluiten en online aan te bieden via een gestandaardiseerde URL-structuur. Bij het ontsluiten van die afbeeldingen moet ook technische metadata bewaard en ontsloten worden. Inhoudelijke en administratieve metadata worden niet door deze component ontsloten, dat wordt overgelaten aan de IIIF Presentation API module.



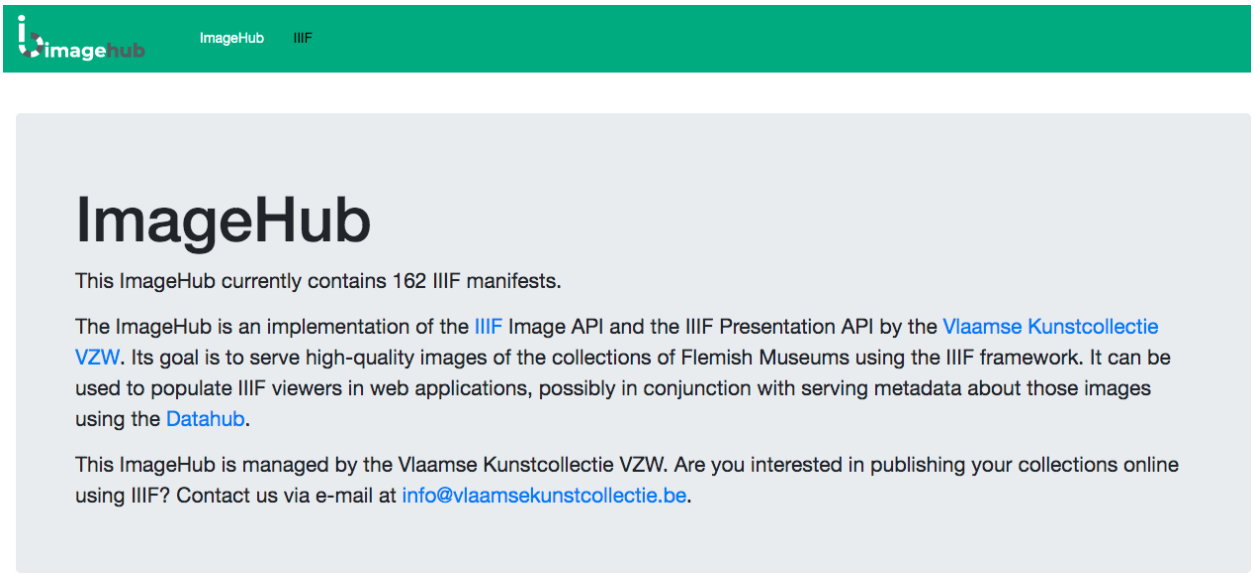
IIIF Presentation API

De IIIF Presentation API-specificatie is een set van standaarden die beschrijft hoe een API moet gestructureerd zijn om afbeeldingen (mogelijks komend uit de IIIF Image API) en hun bijhorende metadata moeten ontsloten worden op een interoperabele manier. Het biedt een high-level metadatastructuur aan die toelaat om zowel technische, inhoudelijke als administratieve metadata mee te geven. Zo kunnen een of meerdere afbeeldingen en hun bijhorende metadata gepresenteerd worden aan gebruikers in hergebruiksapplicaties. De informatie die een IIIF Presentation API-document meegeeft, moet verschillende relaties kunnen aangeven, zoals: of een afbeelding links of rechts naast een andere afbeelding moet getoond worden als de afbeeldingen elk een pagina van hetzelfde boek voorstellen, welke afbeelding het begin van het boek weergeeft, en in welke volgorde de paginering van het boek moet gebeuren. Al die informatie wordt meegegeven via de API in een manifest.json file. Een manifest is een machineleesbaar document dat aan hergebruiksapplicaties alle info meegeeft die nodig wordt geacht om gedigitaliseerde objecten weer te geven. De precieze metadatavelden die gebruikt worden en de manier waarop ze gebruikt worden, kan later nog worden beschreven in een metadatahandboek.

Belangrijk voor dit project is dat de IIIF Presentation API geen beheer van afbeeldingen of metadata inhoudt. Het is enkel een manier om automatisch de informatie te genereren die nodig is voor online ontsluiting van afbeeldingen en hun metadata. De Presentation API is ook geen uitwisselingsstandaard voor metadata.

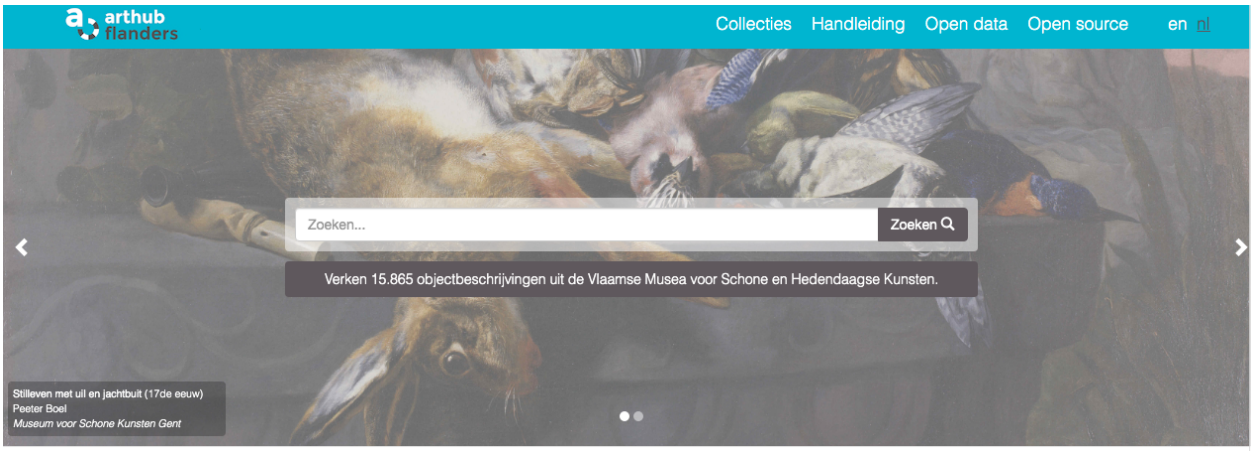
Er dient te worden meegegeven dat, gezien de visie op hergebruiks van de VKC-musea, gedeeltelijke ontsluiting of beveiliging van de beelden niet werd meegenomen binnen de scope van het project. Alle gegenereerde manifests zijn open raadpleegbaar en herbruikbaar. We kozen ervoor ze niet te beschermen met een login of een andere vorm van authenticatie.

De Voor dit project ontwikkelde een freelance developer expliciet een eigenIIIF Presentation API. Deze implementatie komt terecht in een module die we Imagehub noemden, conform aan de naamgeving van Datahub en Arthub. Dit is een PHP-applicatie (Symfony) waarin voor ieder werk een manifest wordt gecreëerd.



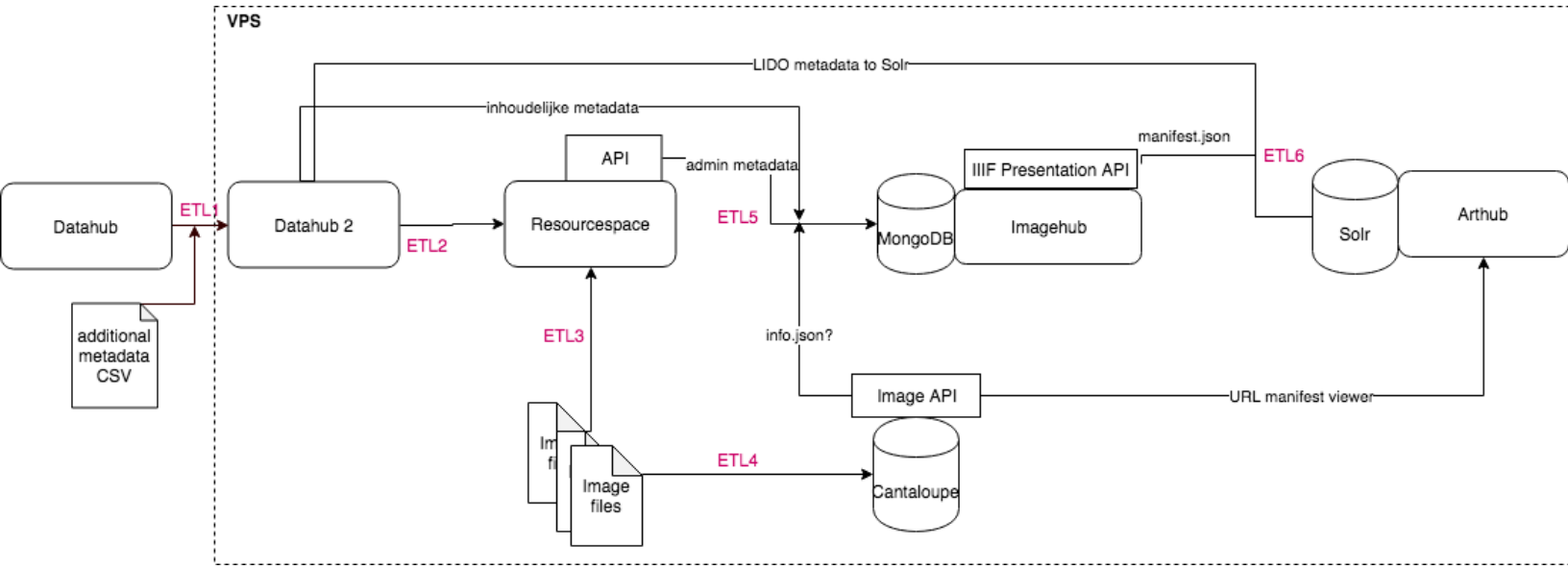
Arthub Flanders

De al eerder ontwikkelde Arthub Flanders module is de hergebruiksapplicatie die toont hoe de uitwisselingslaag kan geïmplementeerd worden. Tijdens dit project zetten we in op de doorontwikkeling van Arthub Flanders door een betere IIIF-conforme viewer te voorzien in een kopie van de originele Arthub Flanders. Tot nu toe werd IIIF gebruikt voor het aanleveren van afbeeldingen aan gebruikers via de Mirador v2 viewer. Mirador, met zijn uitgebreide mogelijkheden (vergelijken van beelden, annotaties, plugins, etc.), wordt echter hoofdzakelijk door onderzoekers gebruikt. De eenvoudigere, maar kwaliteitsvolle, Universal Viewer is gebruiksvriendelijker. We kozen daarom voor Universal Viewer. Universal Viewer ontsluit aan de hand van de Image en Presentation API beeldmateriaal op hoge resolutie. Naast die afbeeldingen geeft Universal Viewer metadata weer in verschillende talen, inclusief de juiste rechtenattributie. Door de Presentation API kunnen gebruikers alle luiken van polyptieken en alle bladzijden prentenboeken gemakkelijk langs elkaar bekijken. *(nvdr. Bij de proefopstelling ondervonden we toch nog een aantal problemen met de Universal Viewer. Zo is die voor de Angelsaksische wereld gebouwd en geeft dat zo zijn negatieve gevolgen wanneer we meertaligheid proberen te realiseren.)*



ETL-pipelines

Pipelines zijn stukken code die data van de ene component naar de andere doorsluizen op vaste tijdstippen. De pipelines worden elke nacht in een vaste volgorde uitgevoerd door één enkel script dat elke nacht loopt. Dit zorgt voor een geautomatiseerde doorstroom van data. Het nadeel van deze aanpak is dat er geen robuuste update en deletefuncties zijn voor data. Als museale medewerkers afbeeldingen of data updaten, moeten ze dus minstens een dag wachten tot alle aanpassingen zichtbaar zijn in de hergebruiksapplicaties. Immers, alle record worden telkens opnieuw opgeladen.
Er zijn zes verschillende pipelines te definiëren. Ze worden beschreven in dit diagram:



ETL1: Van datahub 1 naar Datahub 2
Zoals eerder vermeld in deze deliverable,werden de data van de datahub verrijkt met extra metadata die relaties tussen objecten beschrijven en meertalige metadata inlezen. ETL1 leest LIDO XML-files uit de Datahub en voegt daaraan velden en verrijkte inhoud toe die uit CSV-

bestanden komen. Die verrijkte LIDO XML-files worden daarna ingeschreven in een nieuwe kopie van de datahub, die op de Virtual Private Server (VPS) geïnstalleerd staat.

ETL2: van Datahub 2 naar ResourceSpace
Deze pipeline loopt tegelijkertijd met ETL3, en verzorgt de verrijking van afbeeldingen die opgeladen worden in ResourceSpace. ETL3 laadt de eigenlijke afbeeldingen en hun ingebedde metadata op naar ResourceSpace. Als er een inventarisnummer als ingebedde metadata wordt meegegeven, kan ETL2 de corresponderende LIDO XML-file in de datahub opvragen om de afbeelding te verrijken met relevante inhoudelijke en administratieve metadata. Deze pipeline gebruikt het metadatamodel van het KMSKA.

ETL3: van een drop folder naar ResourceSpace
Als afbeeldingen in een specifieke folder worden geplaatst zodat de beeldinfrastructuur ze kan verwerken, dan is het de taak van deze pipeline om de afbeeldingen in te laden in de ResourceSpace ingest module. Nieuwe afbeeldingen worden nachtelijk ingeladen, hun ingebedde metadata wordt onttrokken aan de afbeeldingen en die metadata wordt ingevoegd volgens het metadatamodel dat binnen het KMSKA wordt gehanteerd. Als afbeeldingen worden geüpdatet of verwijderd, moet deze pipeline de relevante informatie en afbeeldingen ook updaten of verwijderen.

ETL4: van een drop folder naar de IIIF Image API
Misschien is dit niet echt een pipeline. Niettemin zorgt de IIIF-beeldenserver voor de ontsluiting van afbeeldingen die in een specifieke folder worden geplaatst via de Cantaloupe IIIF Image API.

ETL5: Van de ingest module, de IIIF Image API, en de Datahub naar de IIIF Presentation API
Deze pipeline haalt data en metadata uit verschillende bronnen en compileert ze in manifests (.json files) volgens de IIIF Presentation API. De Imagehub module haalt iedere nacht alle afbeeldingen, die zich in ResourceSpace bevinden, op, zodat ze als basis kunnen dienen voor de creatie van manifests. Deze pipeline zorgt ervoor dat elk object in de Arthub een manifest heeft dat verwijst naar één of meer afbeeldingen, geassocieerd met dat object.

ETL6: Van de IIIF Presentation API en de Datahub naar de Arthub
Deze pipeline laadt al de informatie opgeleverd via de uitwisselingslaag op in de database van de hergebruiksapplicatie, in dit geval de Arthub. Het maakt databaseobjecten aan op basis van de LIDO XML-documenten in de Datahub, zoals voorheen ook het geval was. Nu wordt er daarnaast ook de locaties van manifest.json-files ingeladen, die ervoor zullen zorgen dat de correcte afbeeldingen en hun metadata getoond zullen worden op de resultatenpagina's van digitale culturele objecten in de Arthub. Deze pipeline is de laatste stap in de ontsluiting van afbeeldingen voor onlinegebruikers.

Technische implementatie

Dit hoofdstuk overloopt de technische implementatie van de componenten en pipelines die beschreven werden in het vorige hoofdstuk. De implementatie zal besproken worden door van de meest high-level generieke delen van de infrastructuur af te dalen naar de meest gedetailleerde low-level delen van de infrastructuur.

Hosting

De volledige IIIF-beeldinfrastructuur bevindt zich op een Virtual Private Server (VPS). Deze VPS wordt gehuurd bij VIAA, de inhoud wordt beheerd door de VKC.
Aangezien de Vlaamse Kunstcollectie en PACKED voorheen reeds een goede samenwerking hadden met VIAA en er al verschillende stukken infrastructuur op VIAA-servers gehost worden, was het een logische keuze om VIAA opnieuw in te schakelen voor de hosting van de infrastructuur van dit project.

De VPS is een Ubuntu 18.04 installatie met 2 GB RAM en 82 GB opslagcapaciteit.
Technische specificaties:
CPU: Intel(R) Xeon(R) CPU E5-2650L 0 @ 1.80GHz (quad-core)
System Memory: 1GB
RAM: 4GB DDR4
System Storage: 22GB EFI partition + 60GB Virtual Disk 72K RPM lvm2
Network: 10GBps Ethernet VMXNET3 controller

Software-infrastructuur

De software werd lokaal ontwikkeld en periodiek uitgerold op de VPS. De algemene software-infrastructuur wordt uitgerold via Ansible. Ansible is hulpsoftware voor automatisering. Het zorgt ervoor dat alle nodige onderliggende softwarecomponenten, zoals correcte installaties van PHP, Mongo, Cantaloupe… makkelijk geïnstalleerd en geüpdatet kunnen worden.

Technische specificaties:
Ansible version: 2.7.
Ansible plays: te vinden op Github [Imagehub-Box repository]

De keuze van technische componenten is in het voorgaande hoofdstuk al grotendeels geduid. We kozen de componenten volgens bepaalde criteria. Een van de belangrijke criteria is dat componenten open source moeten zijn. Opensourcecomponenten kunnen het makkelijkst via Ansible geïnstalleerd worden, aangezien ze geen licenties nodig hebben. Op die manier kunnen we verzekeren dat de infrastructuur een lage kost heeft en onderhouden zal blijven zonder de nood voor het vernieuwen van licenties.

De configuratie van de technische componenten gebeurt ook via Ansible. Configuratiefiles zorgen ervoor dat de software aangepast kan worden aan de noden van de proefopstelling. Als iets aan de configuratie veranderd moet worden, kan dit lokaal aangepast worden in de Ansible-configuratie. Daar kan Ansible de infrastructuur van de VPS updaten.

Webcomponenten

De belangrijkste componenten, die de freelanceontwikkelaar voor ons bouwde, zijn webgebaseerd. De Imagehub en aangepaste Datahub zijn allebei webcomponenten gebouwd in PHP volgens het Symfony-framework. Dit framework laat toe om pipelines aan te maken en te configureren. Die pipelines worden als commands toegevoegd aan het Symfony-framework en lopen op basis van cron-jobs in een bash script.

Technische specificaties:
PHP version: 7.0.
Symfony version: 2.7.

Frontendcomponenten

Als frontendcomponent wordt er verder gebouwd aan de Arthub, een Ruby on Rails-webapplicatie die een aangepaste versie van Blacklight bevat. Blacklight is een opensourcewebplatform dat functionaliteit voorziet voor het ontsluiten en doorzoekbaar maken van collecties. Het is gebouwd op een Apache Solr-database. Blacklight wordt vaak gebruikt voor de creatie van collectiewebsites voor cultureel erfgoed. De Arthub werd uitgebreid met de installatie van een nieuwe IIIF Viewer en een verbeterde functionaliteit. We integreerden meertaligheid, en een autocompletefunctie om zoektermen te voorspellen.

Technische specificaties:
Ruby version: 2.3.3.
Ruby on rails version: 5.1.
Blacklight version
Apache Solr version: 7.3.1.
Universal Viewer version: 2

Metadata componenten en mappings

In de beeldinfrastructuur moeten data en metadata een aantal formaatwissels ondergaan tussen de contentlaag en de uitwisselingslaag. Ten eerste moet metadata uit de collectieregistratiesystemen gehaald worden om dan, omgezet naar LIDO XML, in de Datahub terecht te komen. Hier zijn pipelines voor gecreëerd. Die pipelines werden uitgebreid om verrijking van data mogelijk te maken. LIDO XML-files worden in deze aangepaste pipeline verrijkt met meertalige data, correcte auteursrechterlijke attributie en een modellering van de relatie tussen objecten. Hoe deze mappings precies gebeurd zijn, kan later nog worden

beschreven in een metadatahandboek. Een PHP-script in de Datahub Symfony-applicatie zorgt voor deze pipeline. De data wordt via de Adlib API uit de collectieregistratiesystemen gehaald, verrijkt met metadata uit een CSV-bestand, en via de Datahub API naar de Datahub gepusht.

Uit de datahub moet metadata worden omgezet naar het KMSKA-model van ResourceSpace. Samen met de afbeeldingen moeten ze uit de dropfolder gehaald worden. Hiervoor zorgen de pipelines ETL2/ETL3, PHP-scripts in de Imagehub Symfony-applicatie. Eerst worden de afbeeldingen opgeladen uit de dropfolder, en wordt ingebedde metadata uit die afbeeldingen ingeladen in het KMSKA-metadatamodel. Administratieve en inhoudelijke metadata komen uit de OAI-PMH API van de Datahub op basis van een lijst van work-PIDs die ingebed zijn in de metadata van de eerder opgeladen afbeeldingen.

Een andere pipeline zorgt ervoor dat metadata uit de Datahub, ResourceSpace en uit de Cantaloupe Image server samenkomen in manifest.json-files in de Imagehub. Dit is weer een PHP-script in de Imagehub. Dit script maakt eerst een array aan op basis van de metadata uit ResourceSpace, en creëert een leeg manifest.json per work-PID. Op basis van deze work-PIDs worden inhoudelijke en administratieve metadata uit de Datahub gehaald. Technische metadata en links naar de afbeeldingen komen uit de info.json-files van de Cantaloupe Image server. De manifest.json-files worden opgeladen naar een MongoDB database in de Imagehub. Deze manifest.json-files worden dan online gepubliceerd op de Imagehub volgens een vaste URL-structuur gedicteerd door de IIIF Presentation API 2.1. -specificatie. Als laatste worden de URLs van de manifest.json-files terug ingeschreven in de LIDO XML-files van de Datahub, zodat de manifestbestanden kunnen opgeroepen worden op basis van de LIDO XML-files. De mappings en stappen die deze pipeline zet, kunnen later nog worden beschreven in een metadatahandboek.

Een laatste pipeline creëert Solr-documenten in de Arthub hergebruiksapplicatie op basis van de data uit de Datahub en de manifestbestanden in de Imagehub. Die Solr-documenten bevatten alle data en metadata van zowel de afbeeldingen als het object zelf. Hier komt alles samen voor openstelling naar gebruikers via de Arthubinterface. Deze pipeline is een PHP-script in de Imagehub Symfony-applicatie.

Oplevering en volgende stappen

Alle geschreven code en opgemaakte architectuur is te vinden in een paar verschillende Github repositories:

- De algemene software-infrastructuur via ansible is te vinden in de Imagehub-box repository.
- De CSV files en voorbeelden van data-mappings zijn te vinden in Imagehub-Fixes en Datahub-Fixes.
- De code voor de imagehub zelf kan gevonden worden in de imagehub repository.
- De geüpdatete code van de Arthub zit in de Arthub-Frontend repository.

Een werkende versie van alle componenten werd eind augustus 2019 online beschikbaar gesteld. De volgende URLs geven toegang tot die componenten:

- Datahub module: http://datahub.iiif.vlaamsekunstcollectie.be/
- Ingest module: http://ingest.vlaamsekunstcollectie.be
- Imagehub module (bevat de IIIF Image en Presentation API modules): http://imagehub.vlaamsekunstcollectie.be
- Arthub module: http://arthub.vlaamsekunstcollectie.be

Samen met de werkende componenten en de code op Github werden deze deliverables opgeleverd voor het einde van de projectperiode.

Na dit project zou een mogelijke volgende fase van de IIIF beeldinfrastructuur het opschalen en doorontwikkelen van de huidige proefopstelling kunnen behelzen. De proefopstelling zou zo kunnen evolueren naar een productieversie die meer dan 200 proefobjecten bevat. Het doel in een volgende fase zou zijn om minstens alle 15.000 objecten uit de Arthub te voorzien van afbeeldingen uit de IIIF-beeldinfrastructuur.