



Nationaal Archief
*Ministerie van Onderwijs, Cultuur en
Wetenschap*

Solution Architecture

Zoeken en tonen van transcripties

Datum	16 mei 2019
Versie	1.0
Status	Definitief (vastgesteld in Stuurgroep dd. 16-5-2019)
Auteur	Gijsbert Kruithof

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

Documenthistorie

Versie	Status	Datum	Wijzigingen
0.1	Concept	22-02-2019	Eerste versie
0.2	Concept	April 2019	Eerste concept ter review
0.3	Concept	26 april 2019	Feedback adviesgroep en architectuurteam verwerkt
1.0	Definitief	16 mei 2019	Vastgesteld in stuurgroep project <i>De Ijsberg zichtbaar maken</i>

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

Inhoudsopgave

1	Inleiding	4
1.1	Wat is een transcriptie?	4
1.2	Solution Architectuur	5
1.3	Doel	5
1.4	Scope	5
1.5	Doelgroep	6
1.6	Referenties	6
2	Kaders	7
2.1	MARA	7
2.1.1	Drielagenmodel	7
2.1.2	Diensten	8
2.1.3	Bedrijfsfuncties	8
2.2	DERA	9
2.2.1	Diensten	10
2.2.2	Bedrijfsfuncties	12
2.3	BIR	14
2.4	Open standaarden	15
2.4.1	Samenvatting open standaarden	15
2.5	Uitgangspunten voor zoeken en tonen van transcripties	15
2.5.1	Generiek	15
2.5.2	Gebruik van Linked Open Data	16
3	Functionele eisen	17
4	Niet functionele eisen	18
5	Doelarchitectuur	19
5.1	High-level overview	19
5.1	Bedrijfsarchitectuur	21
5.1.1	Bedrijfsarchitectuur: informatiemodel	21
5.2	Applicatiearchitectuur	21
5.2.1	Transcriptiebeheersysteem	21
5.2.2	Publiekswwebsite	23
5.2.3	Bekijken (IIIF) server applicatie	24
5.2.4	Annotation server	24
5.3	Webservices	24
5.3.1	GetTranscription	25
5.3.2	SearchTranscription	25
5.3.3	ShowTranscription	26
5.3.4	UploadTranscription	26
5.4	Technische architectuur	27
	Bijlage I: Functionele wensen NA	28

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

1 Inleiding

Er is de laatste jaren veel archiefmateriaal gedigitaliseerd. Het streven is zelfs om 10% van de totale collectie te digitaliseren. Wat veel mensen zich niet realiseren is dat de informatie in deze documenten niet doorzoekbaar is. Scans zijn niet machine leesbaar en zijn digitaal dus niet vindbaar voor onderzoekers, studenten, journalisten of grasduiners. Handschriftherkenning gaat hier verandering in brengen.

Het project "De ijsberg zichtbaar maken" gaat aan de slag met handschriftherkenning¹. Grote datasets van het Nationaal Archief en de Regionaal Historische Centra zullen worden ontsloten met behulp van het semi-automatisch transcriptie platform Transkribus. Na een training herkent de software het handschrift en de taal. De scans van handschriften worden dan automatisch getranscribeerd. Vervolgens worden de transcripties doorzoekbaar gemaakt en getoond in combinatie met de scan.

Een deel van het project is het ontwikkelen van functionaliteit voor het zoeken in de transcripties en het tonen van de resultaten. Een eerste projectresultaat is het opstellen van een Solution Architectuur die de contouren en de kaders van een oplossing hiervoor schetst. Op basis van deze Solution Architectuur kan een Proof of concept ontwikkeld worden.

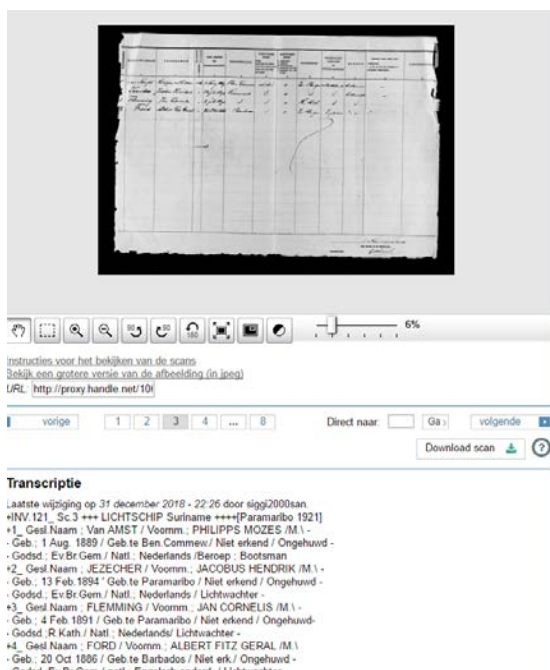
Dit document is de Solution Architectuur voor het zoeken en vinden van transcripties.

1.1 Wat is een transcriptie?

In dit document hanteren we de definitie uit de archiefwiki, die afkomstig is uit de *Archiefterminologie voor Nederland en Vlaanderen (2003)*

*"Een transcriptie is een afschrift in een ander type schrift of symbolen dan in het document waarnaar het is vervaardigd."*²

Doorgaans zijn transcripties "omzettingen" van een moeilijk leesbaar oud handschrift naar moderne machine leesbare tekens. Op dit moment stelt het Nationaal Archief een beperkt aantal transcripties beschikbaar, zie bijvoorbeeld: een scan van de Eerste Algemene Volkstelling van Suriname.



Bron:

https://www.nationaalarchief.nl/onderzoeken/archief/2.10.19.01/inventaris?inventarisnr=121&activeTab=gahetnascan&afbeelding=NL-HaNA_2.10.19.01_121_0002

Figuur 1, Schermafbeelding van transcriptie op website van het Nationaal Archief

¹ Bron: Projectplan De ijsberg zichtbaar maken. Versie 5.1 (15 november 2018)

² Archiefwiki: <https://archiefwiki.org/wiki/Transcriptie>. Zoals geraadpleegd op 22 februari 2019

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

De [MARA] beschrijft transcripties als een manifestatie van een archiefstuk. (Zie MARA par. 8.1. en bijlage C.1) De transcriptie van het archiefstuk is tot op zekere hoogte equivalent aan andere manifestaties zoals een scan of het papieren archiefstuk zelf.

Bij het project moet onderscheid gemaakt worden tussen:

- transcripties die gecontroleerd zijn en waarvan de archiefinstelling garant wil staan dat ze volledig en correct zijn
- transcripties die door de computer zijn gegenereerd of door bezoekers gemaakt en die niet volledig hoeven te zijn en ook niet gecontroleerd zijn door archiefinstelling.

Het eerste type transcriptie kan beschouwd worden als een equivalent van het papieren stuk, vergelijkbaar met een scan of een microfilm. Bij het tweede type transcripties kunnen we er niet voor in staan dat ze 100% correct de inhoud van het stuk weergeven. Het streven van het project is minimaal 80% accuratesse, maar het betekent dat computer gegenereerde transcripties niet zonder meer gelijkwaardige manifestaties zijn die de andere manifestaties kunnen vervangen.

1.2 Solution Architectuur

Deze Solution Architectuur beschrijft de eisen die worden gesteld aan de oplossing voor het realiseren van de functionaliteit voor het zoeken en tonen van transcripties.

Het veld van machine learning, handschriftherkenning en zoektechnologie ontwikkelt zich razendsnel. Deze Solution Architectuur is noodgedwongen gebaseerd op de huidige inzichten en kennis.

1.3 Doel

De beschreven Solution Architectuur vormt het architectuurraster waarbinnen de functionaliteit voor het zoeken en tonen van transcripties wordt gerealiseerd. Het geeft samen met de functionele en de niet-functionele specificaties richting aan de uiteindelijk te realiseren Software Architectuur.

1.4 Scope

De scope van deze Solution Architectuur beperkt zich tot het doorzoekbaar maken en tonen van de transcripties. Dit betekent dat het scannen en maken van de transcripties geen onderdeel is van deze architectuur en dat de architectuur ervanuit gaat dat er transcriptiebestanden zijn inclusief de pointers naar de locatie van tekst op de afbeelding.

Binnen scope is:

- Het opslag- en uitwisselingsformaat van de transcripties;
- Het indexeren van de transcriptie in een zoekmachine;
- Het doorzoeken van de transcripties;
- Het tonen van zoekresultaten in de transcripties;
- Het tonen van de transcripties bij de scans waar ze bij horen;
- Beheer en opslag van transcripties

Buiten scope is:

- Het scannen;
- Het maken van transcripties;
- Het bewerken van transcripties door gebruikers

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

1.5 Doelgroep

Dit document is geschreven voor de projectmedewerkers en de stuurgroep van het Project *De IJsberg zichtbaar maken*, het Ministerie van OCW die de middelen voor het project beschikbaar stelt, ontwikkelaars van functionaliteit (softwareontwikkelaars, testers enz.) en medewerkers van Infrastructuur & Services (Technisch Beheer, Informatiemanagers en architecten).

1.6 Referenties

ID	Titel	Versie
[Projectplan]	Projectplan <i>De IJsberg zichtbaar maken</i>	5.1 (15 november 2018)
[BIR]	https://www.earonline.nl/index.php/BIR-Baseline_Informatiebeveiliging_Rijksdienst	2017 November 2017
[DERA]	Digitaal Erfgoed Referentie Architectuur	2.0 September 2018
[MARA]	Model Architectuur Rijks Archiefinstellingen	1.0 Februari 2016
[NFEs]	Niet functionele eisen generiek en e-Depot https://drive.nationaalarchief.nl/index.php/apps/files/?dir=/Architectuur/6%20-%20solutions/NFE&fileid=167500	1.0 versie September 2018

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

2 Kaders

Dit hoofdstuk bevat de relevante richtinggevende kaders voor de oplossing.

2.1 MARA

De Model Architectuur Rijks Archiefinstellingen [MARA] is een belangrijk kader voor de solution architectuur. Het geeft de gedeelde visie van Convent-RHC's en Nationaal Archief weer.

Het is een consolidatie van inzichten en uitgangspunten die op verschillende plaatsen in de organisaties zijn opgedaan. Behalve als beschrijving van de gedeelde visie wordt deze referentiearchitectuur concreet gebruikt om projecten bij de start te voorzien kaders en richtlijnen alsmede als toetsingskader voor projecten.

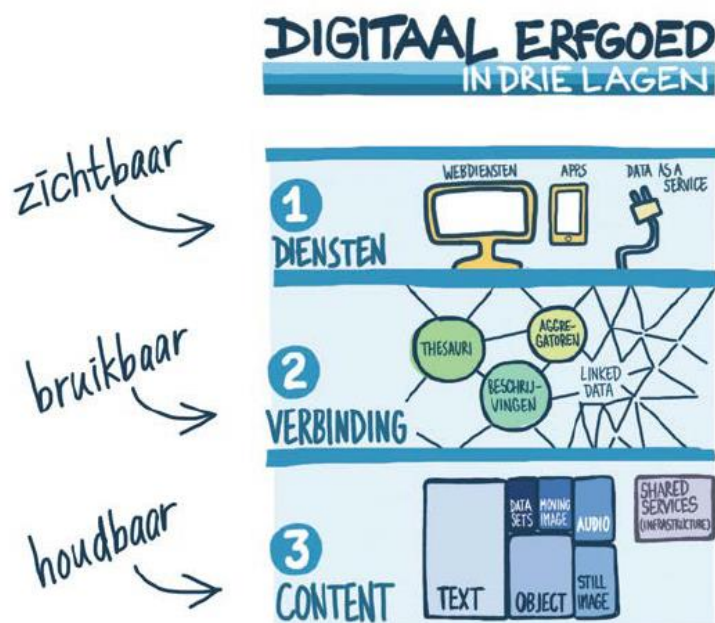
Belangrijk uitgangspunt van de MARA is dat Nationaal Archief en Regionaal Historische Centra gebruik maken van dezelfde e-Depot voorziening.

In de volgende paragrafen worden enkele voor de SA relevante zaken uitgewerkt.

Allereerst wordt het drielagenmodel digitaal erfgoed gepresenteerd. Hierna komen achtereenvolgens de diensten en bedrijfsfuncties aan bod. Tenslotte wordt per laag de benodigde functionaliteit opgesomd. Requirements en constraints die gesteld worden aan de functionaliteit worden eveneens aangegeven.

2.1.1 Drielagenmodel

Belangrijk uitgangspunt is het drielagenmodel, Zie Implementatie drielagenmodel (Zie Thema F en Hoofdstuk 9 in [MARA] en ook [DERA]).



Figuur 1 Drielagenmodel

Binnen het project worden de lagen "houdbaar" en "bruikbaar" voor transcripties gerealiseerd. Voor het zichtbaar maken worden webservices gerealiseerd waarmee NA en RHC's transcripties binnen hun eigen website zichtbaar kunnen maken. De webservices kunnen ook gebruikt worden

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

door derden om andere websites en applicaties te ontwikkelen met de brondata van de transcripties.

2.1.2 Diensten

De RHC's en het NA kunnen de volgende diensten leveren (Zie MARA voor details):

- Houden van toezicht
- Leveren van advies
- Beheer en onderhoud van uitgeplaatst archief
- Inzag faciliteren in uitgeplaatst archief
- *Presenteren van materiaal*
- *Beschikbaar stellen van materiaal*
- *Beschikbaar stellen van data voor hergebruik*

Met name de drie laatste (cursieve) diensten zijn relevant voor het doorzoeken en tonen van transcripties.

De volgende paragraaf geeft een opsomming van de hoofdfuncties van archiefinstellingen. Zij corresponderen met de te leveren diensten.

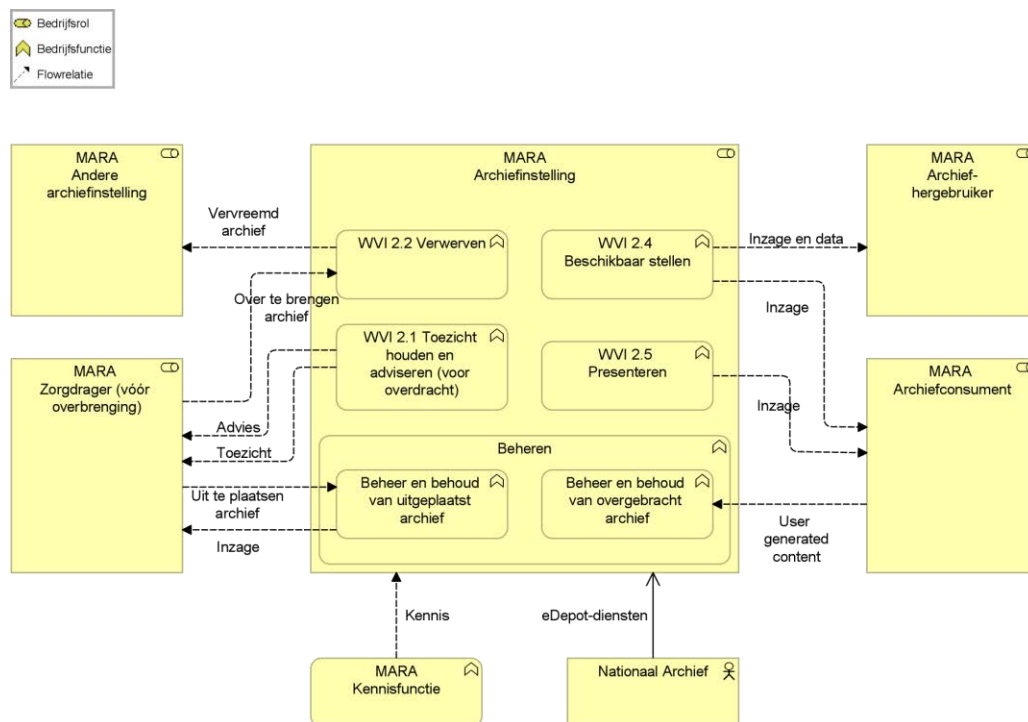
2.1.3 Bedrijfsfuncties

Hoofdfuncties archiefinstelling (NA en RHCs)

- Toezicht houden
- Adviseren
- Verwerven
- *Beheren*
 - Beheer en onderhoud van uitgeplaatst archief
 - *Beheer en onderhoud van overgebracht archief*
- *Beschikbaar stellen*
- *Presenteren*

Onderstaand diagram toont de functies in samenhang.

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019



Figuur 3 Hoofdfuncies archiefinstelling

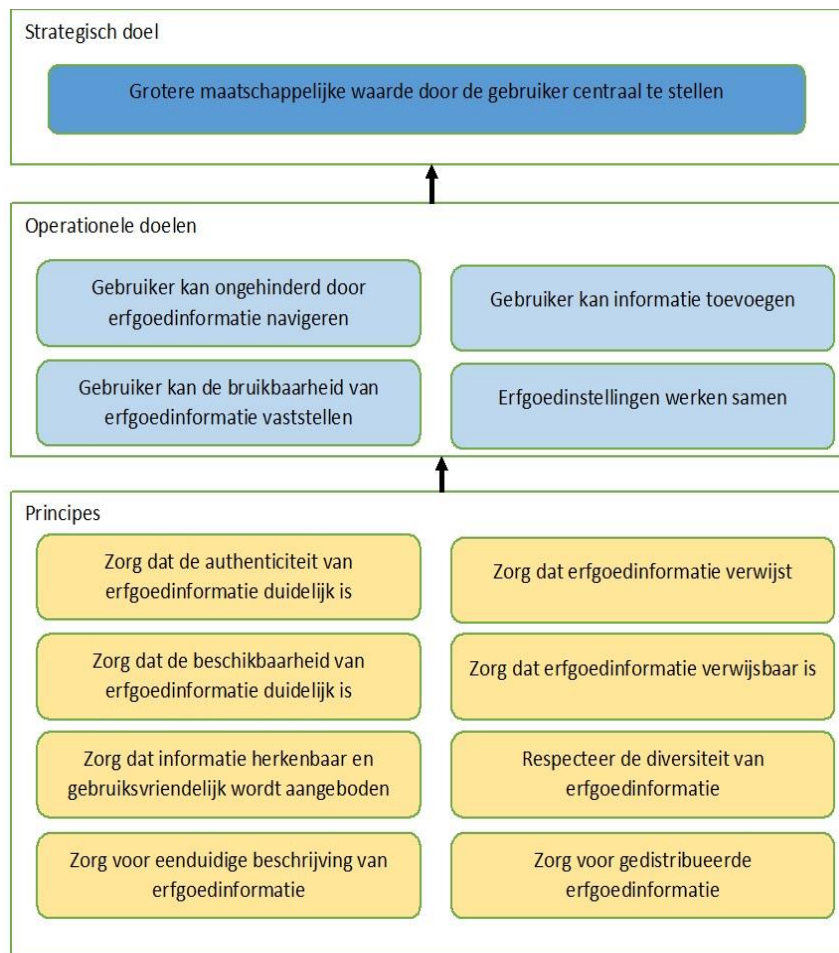
2.2 DERA

De Nationale Strategie beoogt de toegankelijkheid van het Digitale Erfgoed te verbeteren en streeft naar de realisatie van een nationale infrastructuur. Aanbieders en afnemers van erfgoed informatie moeten op deze infrastructuur aansluiten.

De Digitaal Erfgoed Referentie Architectuur (DERA) heeft als doel architectuurkaders op te stellen die bijdragen aan de concretisering van de Nationale Strategie Digitaal Erfgoed. Deze geeft een kader voor de ontwikkeling van een domein overstijgende ICT-infrastructuur binnen het erfgoed. De Nationale Strategie onderscheidt de domeinen Zichtbaar, Bruikbaar en Houdbaar.

DERA 2.0 is een uitwerking van de architectuurvisie (DERA 1.0) voor de Bedrijfslaag en beschrijft de gewenste situatie. De bedrijfsarchitectuur beschrijft wat er moet gebeuren maar niet hoe (applicatiearchitectuur en technische architectuur). In deze Solution architectuur is een koppeling gelegd met implementatievraagstukken.

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019



Figuur 4 Overzicht van de strategische doelen, de operationele doelen en de principes van de DERA.

Voor details zie [DERA]. Achtereenvolgens worden de diensten, bedrijfsfuncties en de paragraaf over linked data uit DERA uitgewerkt voor transcripties.

2.2.1 Diensten

Onderstaande tabel somt de diensten op en geeft aan welke diensten van toepassing zijn voor transcripties.

Dienst	Eisen
--------	-------

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

Gepubliceerde metadata van datasets	Functionele eisen: - De datasets die een bronhouder beschikbaar stelt zijn opvraagbaar. - De datasets worden op een gestandaardiseerde wijze beschreven. <i>Invulling</i> - Een set van transcripties die op basis van een bepaalde Ground truth³ met één Handwritten Text Recognitionmodel zijn gemaakt, worden als een dataset beschouwd en ontsloten. - De Ground truth voor een HCR-traject wordt als afzonderlijke dataset aangemerkt en beschikbaar gesteld. De beschrijving van de datasets gebeurt met DCAT volgens de richtlijnen van de DERA 2.0 Niet-functionele requirements: - De datasets zijn voorzien van duurzame identifiers. - De metadata worden door de bronhouder gepubliceerd als Linked Open Data. <i>Invulling</i> De (data)set van transcripties krijgt een UUID vanuit het Transcriptiebeheersysteem. Deze UUID vormt de basis voor een URI.
Gepubliceerde digitale informatieobjecten	Functionele eisen: - Afnemers kunnen met een duurzame identifier een digitaal erfgoedobject opvragen bij de bronhouder. Via deze duurzame identifier kunnen zij transcripties vinden <i>Invulling</i> Een transcriptie is te beschouwen als een afzonderlijk digitaal informatieobject en is als zodanig ook te vinden. - Elke transcriptie krijgt een eigen URI - Transcripties zijn daarnaast vindbaar via de duurzame identifier van het oorspronkelijke archiefstuk
Gepubliceerde metadata van cultuurhistorische objecten	Niet van toepassing voor transcripties
Gepubliceerde metadata van informatieobjecten	Niet-functionele requirements: - Afnemers kunnen met een duurzame identifier metadata van informatieobjecten opvragen bij de bronhouder. - De metadata worden door de bronhouder gepubliceerd als Linked Open Data. <i>Invulling</i> Transcripties hebben als afzonderlijke informatieobjecten hun eigen metadata, onder meer: datum, gebruikte software, gebruikt HCR-model en ground truth.

³ Ground truth is de term die gebruikt wordt voor de trainingsset waarmee de software het handschrift leert. Het is een door mensen gemaakte, gecontroleerde en geautoriseerde transcriptie van een aantal pagina's van een handschrift.

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

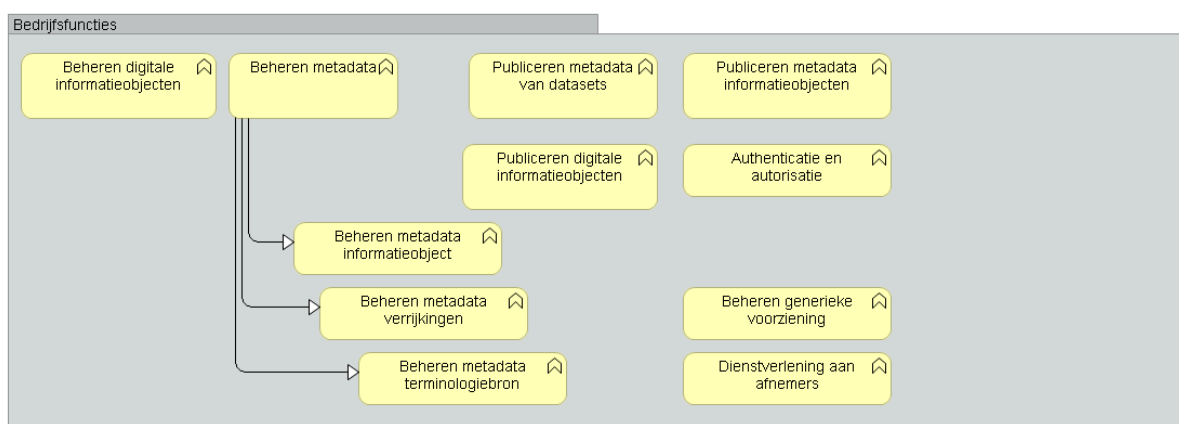
	- Met de URI van de transcripties kan de metadata worden opgevraagd. - Transcripties worden beschikbaar gesteld met de Web Annotation Standard. Dit is een Linked Open Data Vocabulaire. Daarmee worden Transcripties ook beschikbaar als RDF.
Gepubliceerde termen uit terminologiebronnen	Niet van toepassing voor transcripties
Gepubliceerde metadata	Niet van toepassing voor transcripties
Geautoriseerde toegang	Niet van toepassing voor transcripties
Vindbare metadata van termen	Niet van toepassing voor transcripties
Dienstverlening aan afnemers	Functionele requirements: - het afhandelen van informatievragen van afnemers; - het bijhouden van profielen van gebruikers; - het gebruikmaken van gedeelde termen voor het vinden van erfgoedobjecten; - het ophalen van de metadata van erfgoed bij de bron; - het ophalen van digitale erfgoedobjecten bij de bron; - het ondersteunen van de registratie van verrijkingen door afnemers. Niet-functionele requirements: - het bijhouden van indexen (als dit nodig is om gewenste performance te kunnen bieden); - het maken van afspraken met bronhouder over de wijze van authenticatie; - het maken van afspraken met bronhouder over de wijze waarop erfgoedobjecten worden opgevraagd. Invulling - Transcripties worden beschikbaar gesteld met de Web Annotation Standard⁴. Dit is een directe beschikbaarstelling vanuit de bron. De eigen website zal daar gebruik van maken, maar de Annotation Server zal ook opengesteld worden voor anderen, middels API's. Daarmee kunnen zijn de annotations rechtstreeks op te vragen en te gebruiken voor eigen toepassingen - Door gebruik te maken van een afzonderlijke Annotation Server (Zie doelarchitectuur) kan een afdoende performance worden gerealiseerd. Bovendien kan er eventueel gebruik worden gemaakt van authenticatie en procesafspraken.

2.2.2 Bedrijfsfuncties

Het volgende diagram geeft een overzicht van de bedrijfsfuncties zoals gedefinieerd in DERA.

⁴ <https://www.w3.org/TR/annotation-model/>

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019



Figuur 5 DERA bedrijfsfuncties

In onderstaande tabel staat een overzicht van de bedrijfsfuncties en de relevantie voor transcripties.

Bedrijfsfunctie	Requirements vanuit DERA
Beheren digitale informatieobjecten	Geen
Beheren metadata	Niet van toepassing voor transcripties
Beheren metadata cultuurhistorische objecten of informatieobjecten	Niet van toepassing voor transcripties
Beheren metadata verrijking	Niet van toepassing voor transcripties
Beheren metadata terminologiebronnen	Niet van toepassing voor transcripties
Publiceren metadata van datasets	Niet van toepassing voor transcripties
Publiceren metadata cultuurhistorische objecten en informatieobjecten	Niet van toepassing voor transcripties
Publiceren digitale informatieobjecten	- Zorg dat de auteursrechtelijke status van het digitale object wordt vastgelegd. Het gaat hier om de rechten op gebruik van digitale erfgoedobjecten (bijvoorbeeld beperkt toegankelijk i.v.m. wetgeving, auteursrechten of privacy). - Voeg waar mogelijk - zowel technisch als , dan wel auteursrechtelijk - een directe verwijzing naar de digitale representatie toe (bestand). <i>Invoering</i> Eventuele beperkende rechten i.v.m. auteursrecht, of openbaarheid die voor het originele stuk gelden, gelden ook voor transcripties en worden in de metadata van de transcripties overgenomen. Over het algemeen zullen er weinig beperkingen op transcripties berusten

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

	omdat het over het algemeen om oud materiaal gaat. Het Nationaal Archief heeft als beleid om transcripties zoveel mogelijk als open data voor hergebruik beschikbaar te stellen.
Authenticatie en autorisatie	Niet van toepassing voor transcripties
Beheren generieke voorziening	Niet van toepassing voor transcripties
Dienstverlening aan afnemers	• gebruik waar mogelijk erfgoedinformatie die beschikbaar wordt gesteld door het NDE; • gebruik de gedeelde voorzieningen voor de selectie van erfgoedinformatie; • draag bij aan het verbeteren van de toegankelijkheid van het netwerk door verrijking van de metadata. Invulling Het is de intentie om transcripties niet alleen beschikbaar te stellen via de eigen website, maar ook als open data via publiek toegankelijke webservices.

2.3 BIR

Het zoeken en beschikbaar stellen van transcripties dient in het kader van [BIR] te voldoen aan minimaal basisbeveiligingsniveau BBN2. BBN2 richt zich o.a. op departementaal vertrouwelijke informatie waarvan is vastgesteld dat die documenten niet interessant zijn voor statelijke actoren of criminele organisaties. Daarom zijn detectie maatregelen voldoende voor BBN2.

In Hoofdstuk 13 van MARA wordt de informatiebeveiliging uitgewerkt. De volgende tabel geeft de drie gebieden van informatiebeveiliging en de invulling voor transcripties weer.

Gebied	Transcripties
<i>Beschikbaarheid</i> , de mate waarin informatie en informatiesystemen op de juiste momenten beschikbaar zijn voor degenen die bevoegd zijn de informatie te raadplegen of informatiesystemen te gebruiken.	Beschikbaarheid (en performance) van het doorzoeken en tonen van transcripties. Maatregelen: • Monitoren beschikbaarheid • Schaalbaarheid • Hersteltijd • Clustering • Etc. De Non-Functionele eisen voor beschikbaarheid moeten gedurende het project nader worden bepaald en vastgelegd. Zie ook hfst 4.
<i>Integriteit</i> , betreft de juistheid en volledigheid van de informatie in	Géén maatregelen Automatische transcripties zijn in principe openbaar en

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

een informatiesysteem.	per definitie niet 100% correct
<i>Vertrouwelijkheid</i> , betreft de mate waarin de toegang tot informatie beperkt is tot degenen die daartoe bevoegd zijn.	Géén maatregelen Het raadplegen van transcripties is in principe beschikbaar voor iedereen en kent geen beperkingen.

2.4 Open standaarden

Open standaarden zijn publiekelijk beschikbaar. De specificaties van de standaard mogen vrij van licentierechten worden toegepast, gebruikt en gehanteerd.

MARA stelt dat voor het uitwisselen van data zoveel mogelijk gebruik gemaakt wordt van open standaarden. Onderstaande tabellen geven een overzicht van de standaarden en hun scope.

Er zijn een aantal Niet Functionele Eisen (zie [NFEs] par. 1.6) die het gebruik van open standaarden afdwingen.

2.4.1 Samenvatting open standaarden

De volgende open standaarden en richtlijnen zijn van toepassing voor transcripties:

- WCAG 2.1 (webrichtlijnen) niveau A en AA, zie [WCAG]
- Overheidsrichtlijnen DigiToegankelijkheid (<https://www.digitoegankelijk.nl/>)
- OpenAPI specificatie (voormalig Swagger), zie [OAS]
- JPEG2000 voor zichtexemplaren (de afbeeldingen op IIIF), zie [JPG2K]
- IIIF (International Image Interoperability Framework) (<https://iiif.io/>)
- W3C standaarden voor Web Annotation: het Web Annotation Data Model⁵, het Web Annotation Vocabulary⁶ en het Web Annotation Protocol⁷
- BIR2017

2.5 Uitgangspunten voor zoeken en tonen van transcripties

De volgende uitgangspunten zijn van toepassing voor transcripties:

2.5.1 Generiek

- De gerealiseerde oplossing moet zo generiek mogelijk zijn zodat hij niet alleen bruikbaar is voor transcripties die automatisch zijn gegenereerd, maar ook voor handmatig gemaakte transcripties en Optical Character Recognition (OCR).
- De gerealiseerde oplossing is herbruikbaar voor andere erfgoedinstellingen
- Door het project ontwikkelde software komt als Open Source beschikbaar voor anderen
- De functionaliteit die in dit project buiten scope is – zoals het maken van transcripties door publiek – moet later zonder al te veel problemen ingepast kunnen worden in de oplossing die wordt gerealiseerd.

⁵ <https://www.w3.org/TR/annotation-model/>

⁶ <https://www.w3.org/TR/annotation-vocab/>

⁷ <https://www.w3.org/TR/annotation-protocol/>

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

2.5.2 Gebruik van Linked Open Data

- Het gebruik van Linked Open Data (LOD) staat als requirement in MARA en DERA. LOD wordt gebruikt voor inhoudelijke gegevens over afbeeldingen, kaarten, archiefstukken, indexen en andere objecten die niet wordt door ontsloten door de van toepassing zijnde beschrijvingsstandaard, zoals EAD, MARC21.
- Transcripties worden ook ontsloten met Linked Open Data. Door gebruik te maken van de Web Annotation standaard kunnen ze eenvoudig beschikbaar worden gesteld als Linked Open Data. Het format van Web Annotations is namelijk gebaseerd op Linked Open Data⁸.

⁸ <https://www.w3.org/TR/annotation-vocab/>

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

3 Functionele eisen

Onderstaande tabel geeft de functionele eisen weer. Het is gebaseerd op de requirements die zijn opgesteld in verschillende requirementsessies (7-02-2019, 1-03-2019) met vertegenwoordigers van Collectiebeheer, Publiek, I&S en Digitalisering. Sommige van deze functionele wensen zijn echter niet in scope voor dit project. Deze zijn in onderstaand overzicht weggelaten. (Zie voor het volledige overzicht Bijlage I)

De eisen voor de van toepassing zijn voor dit project zijn hieronder weergegeven, aangevuld met requirements met betrekking tot het hergebruik van de transcripties. Deze volgen uit de DERA.

#	Categorie	Project Prioriteit
A - Algemeen		MSCW
A1	De gerealiseerde oplossing moet zoveel als mogelijk generiek zijn voor het zoeken naar, vinden van en tonen van transcripties voortgekomen uit handmatig of automatisch gegenereerde transcripties.	M
A2	De gerealiseerde oplossing moet duurzaam, herbruikbaar en in open source software gerealiseerd worden die aansluit bij de architectuur van het NA en de RHC's.	M
B - Zoeken door eindgebruiker		MSCW
B1	Een transcriptie moet vindbaar zijn via de zoekmogelijkheden op de website.	M
B2	Een transcriptie moet vindbaar zijn via Google.	C
B4	Gevonden resultaten kunnen gefilterd worden op toegang en inventarisnummer	M
B5	Gevonden resultaten kunnen gefilterd worden op "wie, wat, waar wanneer" dus: persoon, onderwerp, plaats, tijd.	S
B6	Ná het uitvoeren van de zoekopdracht worden suggesties gedaan zoals spellingsvarianten.	M
B7	Bij het zoeken wordt je zoekterm na drie karakters automatisch aangevuld.	C
B8	Zoeken met wildcards en boolean is mogelijk.	M
B9	Minimaal is de meest recente versie van een transcriptie vindbaar.	M
B10	Meerdere versie van transcripties worden geïndexeerd en zijn vindbaar	C
B11	Bij het zoeken en indexeren kan gebruik worden gemaakt van soundex, woordenboeken oud Nederlands en woordenlijsten	C
C - Tonen zoekresultaten aan eindgebruiker		MSCW
C1	Zoekresultaten en informatie worden geïntegreerd met resultaten in de archieftoegangen weergegeven. Een transcriptie is in feite niveau dieper dan een inventarisnummer of een item binnen een inventarisnummer.	C
C2	Er wordt informatie aangeboden of een transcriptie gevalideerd is.	M
C3	Wanneer een tekst automatisch is gegenereerd (OCR, HTR, etc.) wordt, indien beschikbaar, de probabiliteit dat een woord goed geluid is meegegeven (keyword-spotting). Een zoekterm kan hierdoor ook gevonden worden als die niet helemaal correct getranscribeerd is of als er sprake is van een spellingsvariant.	C
C4	Wellicht kun je hier zelfs op filteren ('geef mij alle teksten met minimaal 90% correctheid') zodat je grip hebt op fouten in de transcriptie en spellingsvarianties.	C
C6	Alleen de meest recente versie van een transcriptie wordt als zoekresultaat getoond.	M
D - Tonen transcripties aan eindgebruiker		MSCW
D1	Wanneer een gebruiker een scan opent waarvan een transcriptie (zoals transcripties via OCR, HTR of handmatig, etc.) bestaat dan wordt deze tekst getoond.	M
D2	De meest recente versie van de transcriptie bij die scan wordt getoond.	M
D3	Het gebruikte zoekwoord wordt geïnhoudt in de tekst.	S
D4	Een beschikbaar polygon wordt gebruikt om het de locatie van het zoekwoord in de scan zo nauwkeurig als mogelijk wordt weergegeven.	S
D5	Een bezoeker kan zelf instellen of hij/zij bij een scan direct de handmatige transcriptie, OCR of HTR wil zien in een splitscreen of niet.	M
D6	Een gebruiker kan een tekst en de getranscribeerde regel in een scan met elkaar vergelijken. Wanneer de gebruiker óf in een regel in de transcriptie klikt óf op een tekstregel in de scan klikt wordt de corresponderende tekstregel of scanregel in geel gearceerd weergegeven.	W
D7	Kwaliteit van automatisch getranscribeerde teksten moet inzichtelijk worden gemaakt (zie probabiliteit), opnemen van 'disclaimer'.	S
F - Beheer		MSCW
F1	Er is een systeem beschikbaar waarin transcripties kunnen worden beheerd. Vanuit dit systeem kan een backend-component teksten en bijbehorende metadata ophalen voor beschikbaarstelling in de website.	M
F2	Een beheerder kan op een uniforme en geautomatiseerde manier één of meerdere bestanden (in bulk) die binnen een project gevormd zijn in dit systeem opnemen.	M
F4	Een beheerder kan verschillende transcriptie-versies van één bepaalde scan of bestand beheeren.	W
F5	Een beheerder kan nieuwe versies aanmaken van transcripties aanmaken, bekijken, wijzigingen aanbrengen en verwijderen.	W
F6	Bij de transcripties kan metadata worden bewaard en beheerd over diverse zaken rondom de creatie en het gebruik van de transcripties. Vastgelegd moet kunnen worden: Oorsprong (project, website, etc.), Machinaal of handmatig gevormd, Oorspronkelijk type bestandformaat (txt, xml, etc.); het beheer gebeurt wel in één formaat, nog bekijken of, hoe en waar een omzetting naar dit ene formaat plaats moet vinden, Kwaliteit (geautoriseerd na handmatige invoer, kwaliteit software van automatische omzetting, etc.), Mutatie (wie mag wanneer wijzigingen aanbrengen, worden deze direct doorgevoerd of pas na modereren, etc.). Deze de metadata moet door een beheerder bewerkt kunnen worden, zowel per individueel record als in bulk.	S
F7	Er moet een identifier vastgelegd kunnen worden die een koppeling tussen scan en transcript mogelijk maken.	M
F8	Het moet mogelijk zijn om geautomatiseerd uit dit systeem de informatie te halen voor het doorzoekbaar maken, presenteren van en als open data aanbieden van de transcripties.	M
G - Beschikbaarstellen voor hergebruik		MSCW
G1	Een set van transcripties die op basis van een bepaalde ground truth met één HTR-model zijn gemaakt, worden als een dataset beschouwd en ontsloten	S
G2	De Ground truth voor een HTR-traject wordt als afzonderlijke dataset aangemerkt en beschikbaar gesteld	M
G3	Elke transcriptie krijgt een eigen URI	S
G4	Transcripties zijn behalve via een URI ook vindbaar via de duurzame identifier van het oorspronkelijke archiefstuk	S
G5	Eventuele beperkende rechten i.v.m. auteursrecht, of openbaarheid die voor het originele stuk gelden, gelden ook voor transcripties en worden in de metadata van de transcripties overgenomen.	S

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

4 Niet functionele eisen

Voor de uiteindelijk op te leveren oplossing gelden de generieke non-functionele eisen zoals die zijn opgesteld voor e-Depot en GAF [NFES]. Gedurende het project moeten de specifieke niet functionele eisen aan het zoeken en tonen van transcripties nader worden gespecificeerd en toegevoegd aan het NFE document.

Afgezien van de generieke non functionele eisen [NFES] zijn de op dit moment bekende specifieke niet functionele eisen aan dit project:

- de oplossing moet kunnen worden opgeleverd als Linux Redhat Packages (RPM's) die automatisch uitgerold kunnen worden met Ansible
- de oplossing maakt bij voorkeur gebruik van een Open Source Database, zoals MariaDB of PostgreSQL.
- de oplossing is geschreven in de bij het Nationaal Archief gangbare programmeertalen: Java, PHP en Angular
- De documentatie moet aan het eind van het project minimaal kunnen worden ingelezen in het Confluence platform van het NA, bij voorkeur wordt de documentatie tijdens project direct binnen dit platform aangemaakt
- User stories, bugs en releases worden bij voorkeur gedocumenteerd met behulp van JIRA

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

5 Doelarchitectuur

5.1 High-level overview

Onderstaand figuur geeft een high-level overview van de doelarchitectuur met betrekking tot transcripties.

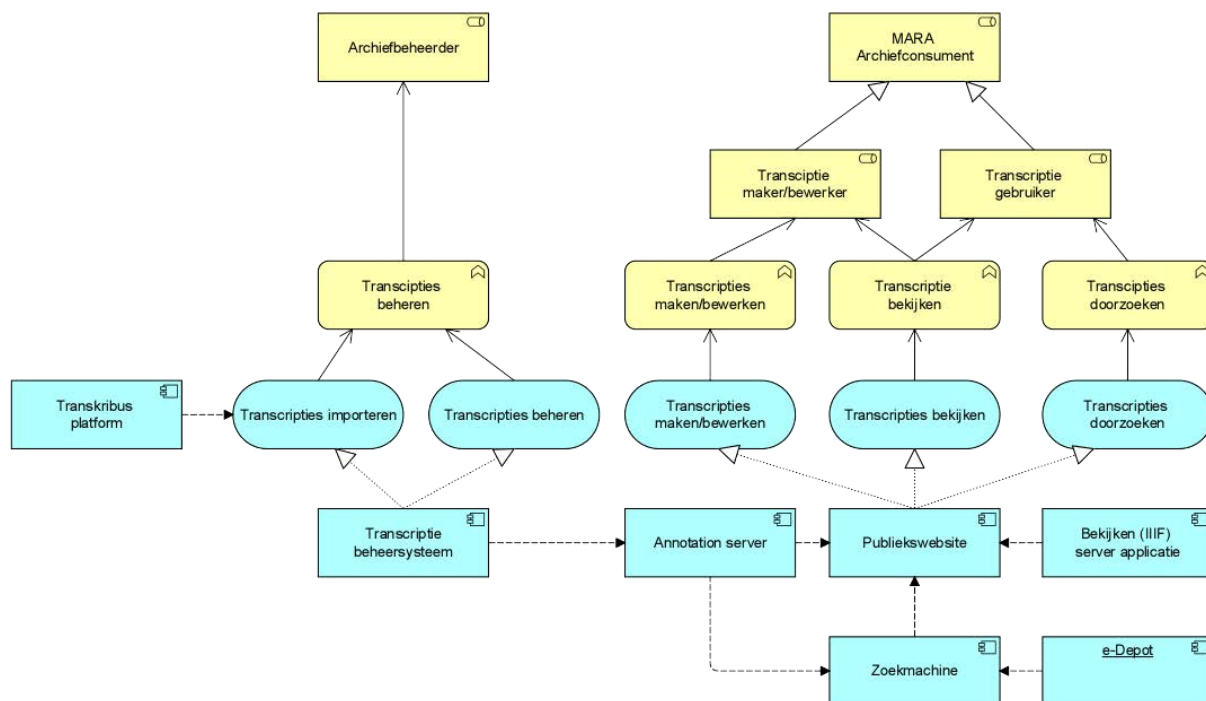
De belangrijkste bedrijfsfuncties - afkomstig uit de functionele wensen - zijn:

- Transcripties tonen
- Transcripties doorzoeken
- Transcripties maken/bewerken
- Transcripties beheren

Ze worden uitgevoerd door twee type rollen:

- Archiefgebruikers (MARA Archiefconsument)
- Archiefbeheerder

Deze staan in de gele blokken bovenaan in de figuur.



Figuur 6 Architectuuroverview Transcripties

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

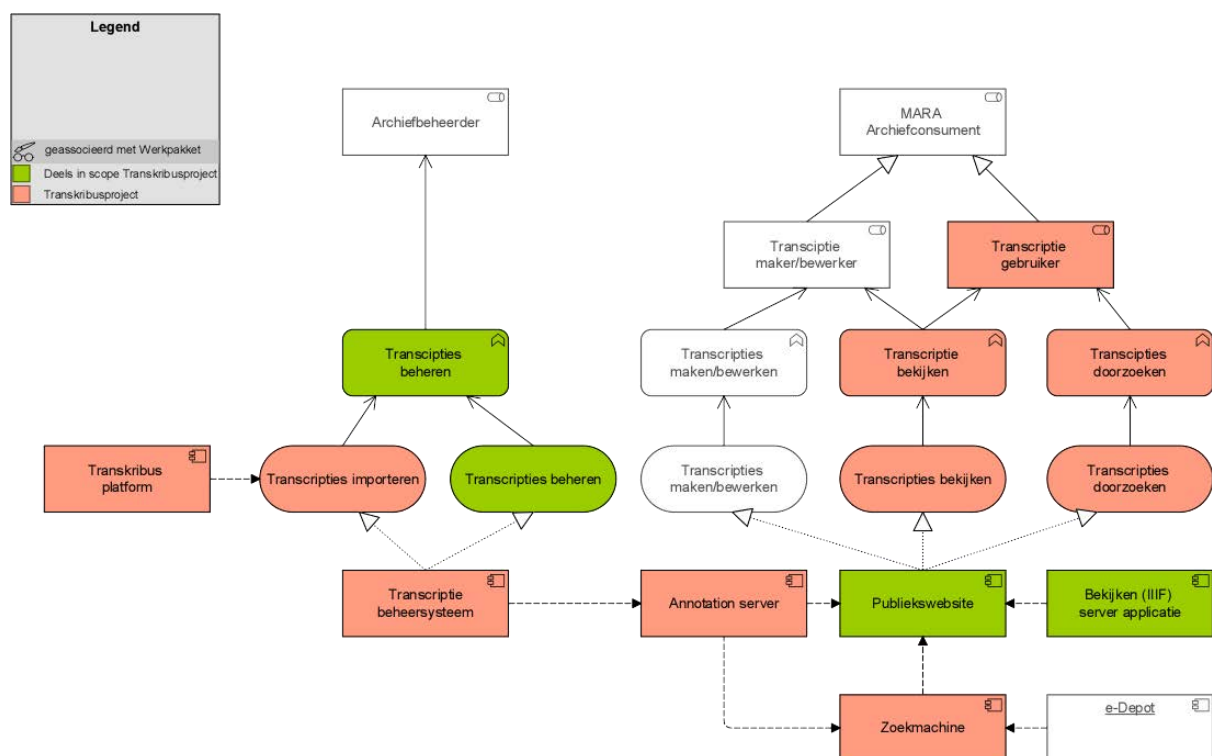
De bedrijfsfuncties maken gebruik van de blauw gekleurde applicatieservices:

- Transcripties importeren
- Transcripties beheren
- Transcripties maken/bewerken
- Transcripties tonen
- Transcripties doorzoeken

De applicaties (de blauwe vierkante symbolen in figuur 6) die een rol spelen bij transcripties zijn:

- Transkribus platform – voor het maken van transcripties. Dit platform speelt natuurlijk een rol in het maken van de transcripties maar niet in de rest van het proces: het bewaren, beschikbaarstellen en doorzoekbaar maken
- Publiekswegsite
- Zoekmachine
- Bekijken (IIIF) server applicatie
- Annotation server – voor het beschikbaarstellen van de transcripties
- E-Depot – als bron van de afbeeldingen op de IIIF server en als bron voor born digital bestanden
- Transcriptiebeheersysteem

N.B. Niet al deze functies en componenten zijn in scope van het project. Hieronder is in oranje/roze weergegeven wat in de scope van het project valt.



Figuur 7 Projectscope architectuuroverview Transcripties

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

De functionaliteit om transcripties te laten bewerken/maken door de archiefconsument is buiten scope en ook de beheerfunctionaliteit is grotendeels buiten scope.

Het indexeren van born digital bestanden uit het e-Depot is buiten scope. De full tekst indexering van transcripties is natuurlijk wel binnen scope.

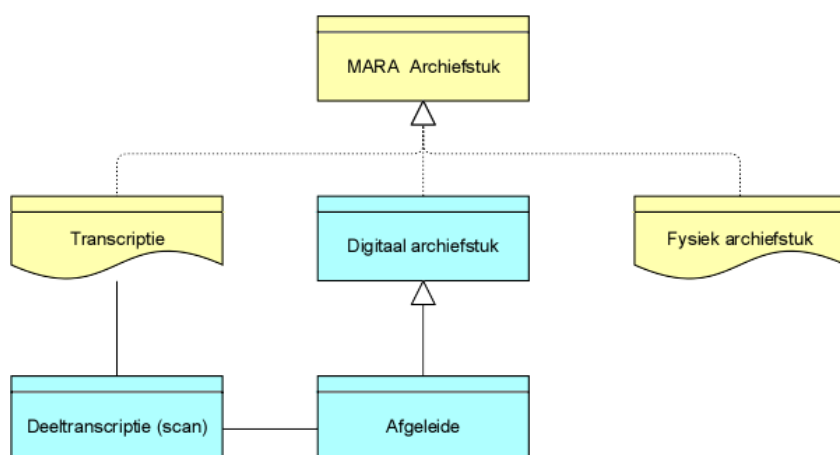
Wel nodig is een systeem om transcripties te kunnen importeren, bewaren en beschikbaar te stellen aan een publiekswaasite en een zoekmachine.

Hieronder wordt per architectuurlaag de doelarchitectuur verder uitgewerkt.

5.1 Bedrijfsarchitectuur

5.1.1 Bedrijfsarchitectuur: informatiemodel

Onderstaand figuur geeft het (bedrijfs)informatiemodel weer. De kleinste opvraagbare eenheid volgens de MARA is een archiefstuk. Daarvan kunnen meerdere manifestaties zijn: het fysieke (papieren) archiefstuk en een digitale afgeleide. Een volledige transcriptie is te beschouwen als een derde manifestatie.



Figuur 8 Informatiemodel Transcripties

Op dit moment zijn er alleen deeltranscripties. Deze deeltranscripties zijn altijd gekoppeld aan één scan, een afgeleide.

5.2 Applicatiearchitectuur

Voor transcripties spelen verschillende systemen een rol. Hieronder volgt per systeem een uitwerking.

5.2.1 Transcriptiebeheersysteem

Het beheren van transcripties zou theoretisch kunnen plaats vinden in het Collectiebeheersysteem of in het e-Depot. Geen van beide systemen is hier echter geschikt voor.

Tegen het beheer van transcripties in het e-Depot pleit:

- Transcripties zijn – op dit moment - geen equivalente manifestaties van archiefstukken. Dit betekent:
 - o In het informatiemodel van het e-Depot (volgens OAIS) zijn het geen manifestaties
 - o Een transcriptie kan niet als vervanging voor een archiefstuk gepresenteerd worden, zoals een scan

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

- Machine gemaakte transcripties zijn een momentopname die afhankelijk is van de kwaliteit van de op dat moment gebruikte software. Als dezelfde transcripties over een aantal jaren weer gemaakt worden, zal de kwaliteit vermoedelijk veel beter zijn.
- Het huidige e-Depot van het NA heeft op dit moment geen functionaliteit om van buitenaf per scan een transcriptie toe te kunnen voegen. Deze functionaliteit wordt ook niet op korte termijn verwacht.

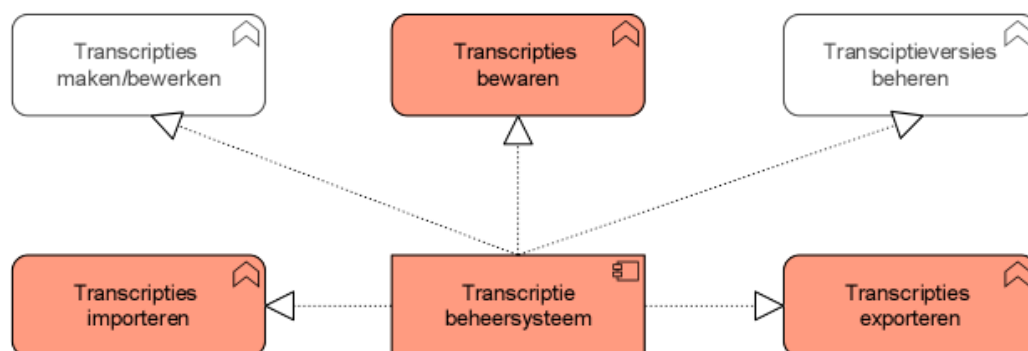
Tegen het beheer van transcripties in het Collectiebeheersysteem pleit:

- Machine gemaakte transcripties zijn geen gecontroleerde en door de archiefinstelling geautoriseerde informatie. Ze kunnen per definitie niet allemaal handmatig door Collectiebeheerders worden gecontroleerd en gecorrigeerd
- Machine gemaakte transcripties zijn een momentopname die later vervangen kunnen worden door betere versies.
- Het huidige Collectiebeheersysteem heeft moeite met grote hoeveelheden (XML)-data. Naar verwachting zal het huidige CBS de hoeveelheid data niet aankunnen

Een afzonderlijk transcriptiebeheersysteem lijkt de beste oplossing. Op basis van de functionele wensen dient dit systeem de volgende functionaliteiten te hebben:

- Transcripties importeren: functionaliteit om transcripties die afkomstig zijn uit een andere willekeurige bron in bulk te kunnen importeren
 - Het Transcriptiebeheersysteem moet in ieder geval transcripties in de formaten Page en Alto-XML kunnen importeren.
- Transcripties bewaren: functionaliteit om transcripties langdurig en veilig te kunnen bewaren
- Transcriptieversies beheren: functionaliteit om verschillende versie van dezelfde transcripties te kunnen bewaren/beheren/exporteren
- Transcripties maken/bewerken: functionaliteit om handmatig nieuwe transcripties te maken en bestaande transcripties aan te passen
- Transcripties exporteren: functionaliteit om in bulk transcripties te kunnen exporteren. Met name ten behoeve van indexering door een zoekmachine of ten behoeve van het beschikbaar stellen als open data.
 - Het Transcriptiebeheersysteem moet transcripties in het Web Annotation format kunnen exporteren

Voor het project *De IJssberg zichtbaar maken* zijn alleen de functies importeren, bewaren en exporteren in scope.



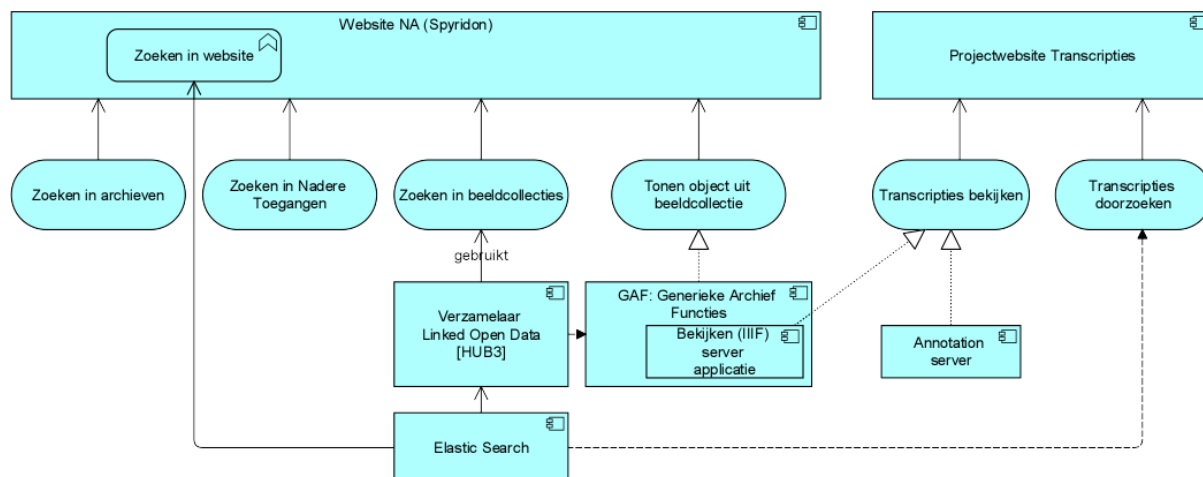
Figuur 9 Projectscope Transcriptiebeheersysteem

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

5.2.2 Publiekswegsite

Transcripties worden uiteindelijk getoond en doorzoekbaar gemaakt op de publiekswegsite van het Nationaal Archief. Het project *De IJssberg zichtbaar maken* werkt daar naar toe. De integratie van de zoekoplossing in de NA-website (Spyridon) behoort niet tot de scope van het project, het aanleveren van een integreerbare oplossing wél.

Onderstaande figuur geeft een beeld van de huidige NA publiekswegsite in combinatie met de te realiseren oplossing voor transcripties.



Figuur 10 Samenhang website NA en projectwebsite Transcripties

De website kent vier “types” materiaal:

- archieven: doorzoekbaar gemaakt via de inventarissen in het zogenaamde archievenoverzicht middels EAD
- nadere toegangen of indexen: doorzoekbaar via de Linked Open Data Verzamelaar (HUB3)
- de beeldcollectie: foto’s en kaarten: doorzoekbaar via de Linked Open Data Verzamelaar (HUB3)
- website content: geïndexeerd met behulp van Elastic Search

Dit correspondeert met vier verschillende manieren van presenteren en vier verschillende manieren van indexeren. Een algemene zoekopdracht op de website betekent onder water een zogenaamde “federated search” in verschillende “bakken”. N.B. De transcripties op de huidige website (nationaalarchief.nl) worden in het archievenoverzicht (zoeken in archieven) beschikbaar gesteld, maar worden niet door een zoekmachine geïndexeerd.

Voor de transcripties in het kader van het project komt er een vijfde type content bij die in eerste instantie afzonderlijk wordt geïndexeerd.

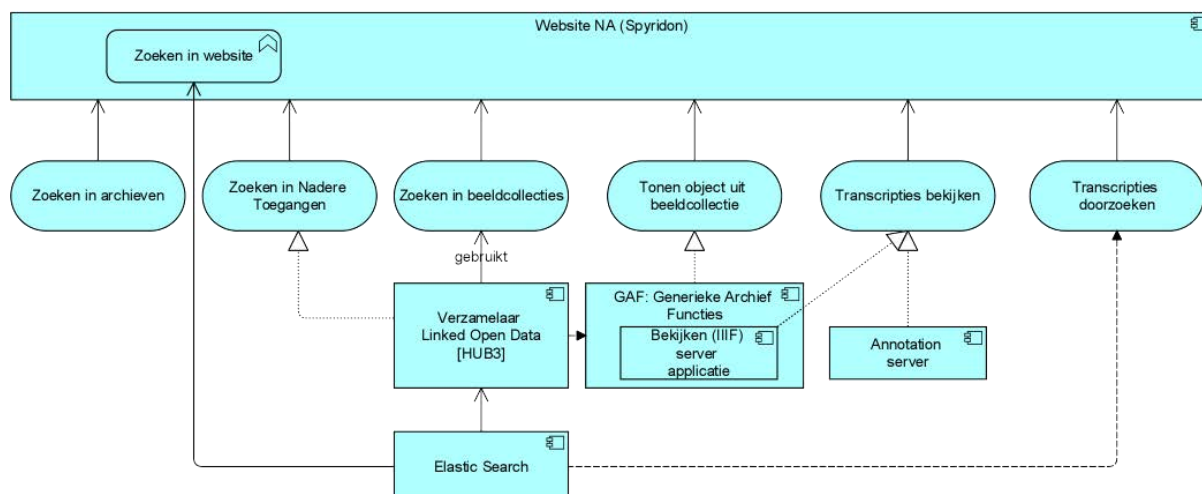
De indexering vindt plaats met behulp van de zoekmachine Elastic Search zodat het op eenvoudige wijze geïntegreerd kan worden met andere zoekapplicaties van Spyridon

De transcripties worden in het kader van dit project getoond en doorzoekbaar gemaakt met een projectwebsite. Deze website kan een niet publiek deel zijn van Spyridon of het kan een afzonderlijke projectsite zijn. De transcripties worden wél getoond op basis van de afbeeldingen op de IIIF Server.

Na afronding van het project kan de oplossing geïntegreerd worden in de publiekswegsite met gebruik making van de services die al in het kader van het project zijn gerealiseerd: de service voor het bekijken van transcripties en de service voor het tonen van transcripties.

Het resulteert in het volgende plaatje:

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019



Figuur 11 Doelmodel van toekomstige integratie van transcripties in website NA

5.2.3 Bekijken (IIIF) server applicatie

Voor digitale objecten met mime/type 'image', die op de publiekswaarsite kunnen worden getoond, wordt een zichtexemplaar gemaakt en opgeslagen volgens IIIF. IIIF is een veel gebruikte standaard om snel plaatjes te kunnen serveren. De zichtexemplaren op IIIF worden opgeslagen in het formaat JPEG2000 en geconverteerd naar 'Tiled Pyramidal'. IIIF wordt door het NA gebruikt om zichtexemplaren van scans en gedigitaliseerde foto's te tonen.

Transcripties worden getoond met de afbeeldingen op de IIIF-server. Dit is een bestaand onderdeel van de huidige infrastructuur. NB: dit deel hoeft dus niet meer in het kader van het project gerealiseerd te worden. Het project maakt hier gebruik van.

Het bekijken van transcripties vergt het bij elkaar komen van twee zaken: de afbeelding en de tekst.

Voor het bekijken van een afbeelding op de huidige website wordt de Universal Viewer gebruikt. Dit is een open source viewer die gebaseerd is op het gebruik van een IIIF oplossing. Idealiter maakt de oplossing in het project ook gebruik van de Universal Viewer, maar in ieder geval wordt in de oplossing ook een viewer gebruikt die gebaseerd is op de IIIF standaard.

5.2.4 Annotation server

Binnen het IIIFramework is rekening gehouden met het toevoegen van (tekstuele) content aan afbeeldingen. Dit werkt volgens de W3C webstandaard voor Annotation⁹. Elke vorm van tekstuele inhoud, zoals commentaar op foto's en ook transcripties kan worden toegevoegd. Elke annotation heeft een specifieke pointer naar een bepaalde plek op de afbeelding. Deze annotaties worden omwille van snelle beschikbaarheid geserveerd vanuit een zogenaamde Annotation Server. Door gebruik te maken van Web Annotations kunnen de transcripties ook beschikbaar gesteld worden als Linked Open Data.

5.3 Webservices

"Een webservice kan omschreven worden als een interface van een applicatiecomponent die toegankelijk is via

⁹ Specifiek het Web Annotation Data Model, het Web Annotation Vocabulary en het Web Annotation Protocol

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

standaard webprotocollen waarbij zonder menselijke tussenkomst wordt gecommuniceerd ..."¹⁰

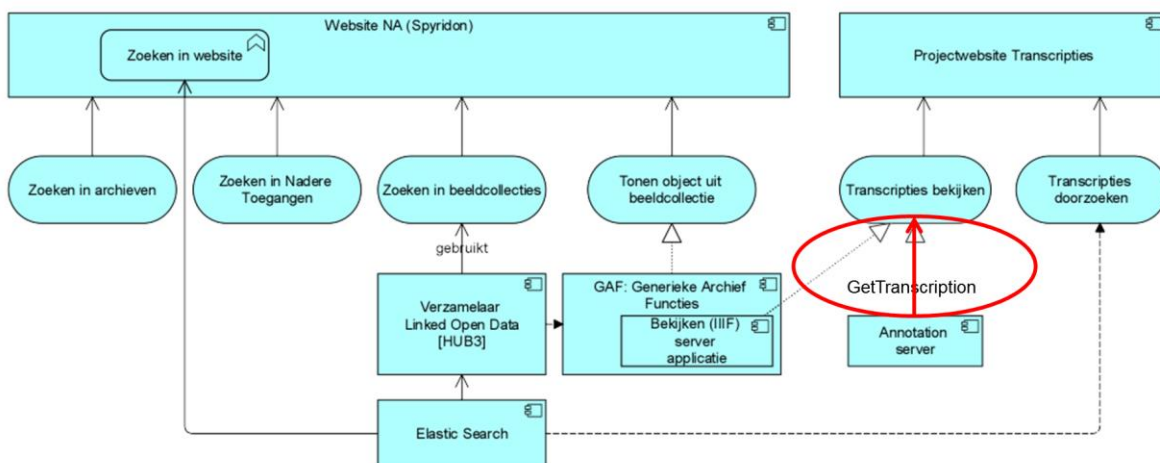
Een webservice maakt het mogelijk om via internet bepaalde functionaliteit aan te roepen. Een webservice kan gebruikt (ingebouwd) worden in verschillende applicaties, zoals een website. Door webservices aan te bieden, biedt je flexibiliteit aan andere partijen om functionaliteit op eigen manier te realiseren.

Een webservice is een verbijzondering van een Application Programming Interface (een API), namelijk een API die via het web publiekelijk benaderbaar is.

De Solution Architectuur voor transcripties voorziet in de realisatie van een aantal webservices. De webservices worden binnen het project gerealiseerd en gebruikt voor de projectwebsite volgens het principe "Eat your own dogfood". De webservices kunnen na afronding van het project gebruikt worden door NA en RHC's om de oplossing te integreren in de eigen website.

5.3.1 GetTranscription

Deze service wordt geleverd door de Annotation Server en maakt het mogelijk om een transcriptie op te vragen.



Figuur 12 Webservice GetTranscription in kader van architectuur

Deze webservice zou gebruikt kunnen worden in combinatie met afbeeldingen die op een andere plek dan de IIIF-server van het NA staan. Met behulp van deze webservice en een eigen viewer kunnen afbeeldingen en transcripties dan getoond worden.

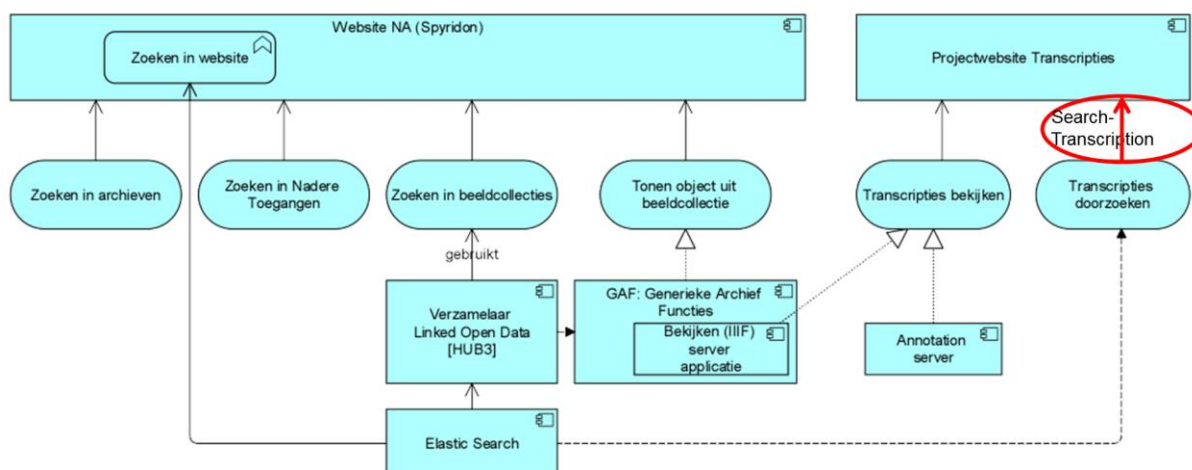
5.3.2 SearchTranscription

Deze service wordt geleverd door de zoekmachine. Het maakt gebruik van de standaard out-of-the-box Search API op Elastic Search¹¹. Met behulp van deze service zoek je binnen Elastic Search specifiek op transcripties.

¹⁰ Bron: <https://nl.wikipedia.org/wiki/Webservice>

¹¹ Zie bijvoorbeeld: <https://www.elastic.co/guide/en/elasticsearch/reference/current/search-search.html>

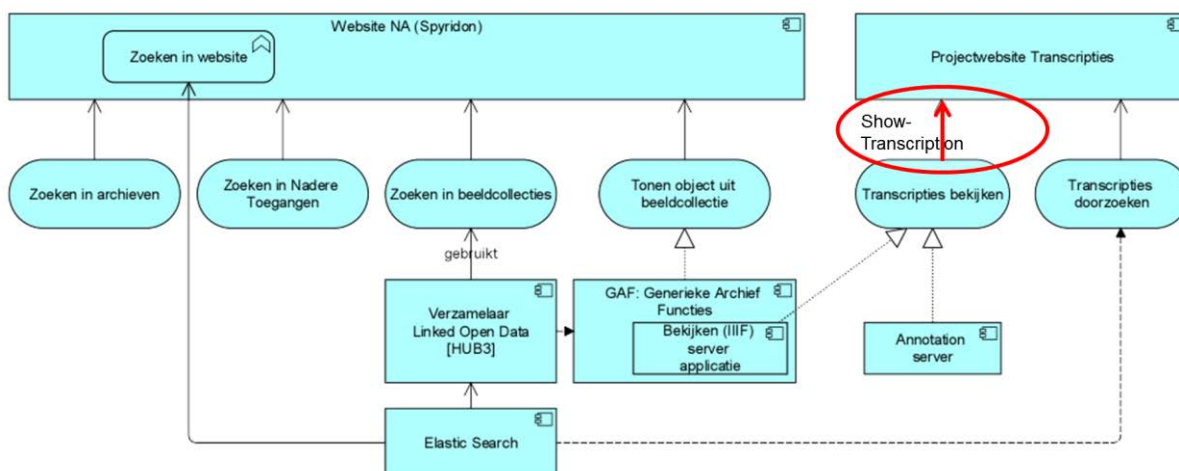
Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019



Figuur 13 Webservice SearchTranscription in kader van architectuur

5.3.3 ShowTranscription

Deze service toont de transcriptie samen met de afbeelding. De service maakt gebruik van de GetTranscription service van de Annotation Server maar ook van de reeds bestaande “Default webservice” van GAF om een afbeelding van de IIIF-server op te halen. De ShowTranscription service maakt gebruik van de IIIF-server van het NA.

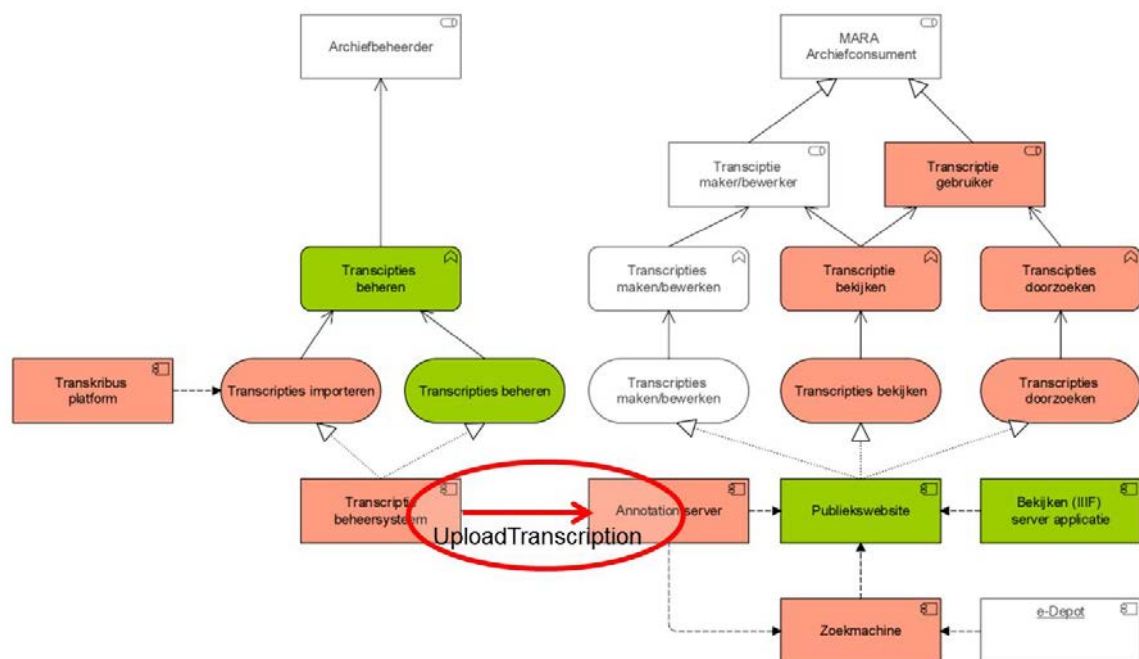


Figuur 14 Webservice ShowTranscription in kader van architectuur

5.3.4 UploadTranscription

Dit is een service van de Annotation Server waarmee een Transcriptie op de Annotation server gezet wordt en geïndexeerd kan worden door de zoekmachine. Deze service wordt gebruikt door het Transcriptiebeheersysteem van het NA om transcripties beschikbaar te stellen, maar deze service kan ook worden gebruikt door een RHC om vanuit een ander systeem transcripties up te loaden en te indexeren.

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019



Figuur 15 Webservice UploadTranscription in kader van architectuur

5.4 Technische architectuur

De Technische architectuur is in deze Solution Architectuur doelbewust nog niet heel precies ingevuld om ruimte te laten aan het project bij de realisatie van de oplossing.

Om de oplossing goed te kunnen integreren in de infrastructuur van het Nationaal Archief gelden wel de volgende voorwaarden:

- de oplossing moet werken op het Linux Redhat platform 7.0
- de oplossing wordt bij voorkeur gerealiseerd op de infrastructuur van het Nationaal Archief
- de oplossing maakt bij voorkeur gebruik van de al operationele Elastic Search instance van het NA

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

Bijlage I: Functionele wensen NA

Onderstaande tabel geeft het complete overzicht van functionele wensen zoals dat is opgesteld door het NA. In de tabel staan ook de prioriteiten aangegeven vanuit het perspectief van het NA. In hoofdstuk 3 is het gedeelte nomen dat in scope is voor het project *De IJsberg zichtbaar maken* overgenomen.

Transcripties	Version: 1.0	Definitief
Solution Architecture Document	Date:	16 mei 2019

#	Categorie	Project Pr	Prio NA
	A - Algemeen	MSCW	MSCW
A1	De gerealiseerde oplossing moet zoveel als mogelijk generiek zijn voor het zoeken naar, vinden van en tonen van transcripties voortgekomen uit handmatig of automatisch gegenereerde transcripties.	M	M
A2	De gerealiseerde oplossing moet duurzaam, herbruikbaar en in open source software gerealiseerd worden en die aansluit bij de architectuur van het NA en de RHC's.	M	M
A3	De oplossing moet aansluiten bij de oplossing voor de doorzoekbaarheid van born digital bestanden en digital born-bestanden.	W	M
	B - Zoeken door eindgebruiker	MSCW	MSCW
B1	Een transcriptie (handmatige transcriptie, OCR, HTR, etc.) moet vindbaar zijn via de zoekmogelijkheden op de website.	M	M
B2	Een transcriptie (handmatige transcriptie, OCR, HTR, etc.) moet vindbaar zijn via Google.	C	S
B3	Op de website kun je in het archievenoverzicht en binnen de toegangen op transcripties zoeken.	C	S
B4	Gevonden resultaten kunnen gefilterd worden op toegang en inventarisnummer	M	M
B5	Gevonden resultaten kunnen gefilterd worden op "wie, wat, waar wanneer" dus: persoon, onderwerp, plaats, tijd.	S	S
B6	Ná het uitvoeren van de zoekopdracht worden suggesties gedaan zoals spellingsvarianten.	M	M
B7	Bij het zoeken wordt je zoekterm na drie karakters automatisch aangevuld.	C	S
B8	Zoeken met wildcards en boolean is mogelijk.	M	M
B9	Minimaal is de meest recente versie van een transcriptie vindbaar.	M	M
B10	Meerdere versie van transcripties worden geïndexeerd en zijn vindbaar	C	C
B11	Bij het zoeken en indexeren kan gebruik worden gemaakt van soundex, woordenboeken oud nederlands en woordenlijsten	C	C
	C - Tonen zoekresultaten aan eindgebruiker	MSCW	MSCW
C1	Zoekresultaten en informatie worden geïntegreerd met resultaten in de archieftoegangen weergegeven. Een transcriptie is in feite niveau dieper dan een inventarisnummer of een item binnen een inventarisnummer.	C	S
C2	Er wordt informatie aangeboden of een transcriptie gevalideerd is.	M	M
C3	Wanneer een tekst automatisch is gegenereerd (OCR, HTR, etc.) wordt, indien beschikbaar, de probabiteit dat een woord goed geduid is meegegeven (keyword-spotting). Een zoekterm kan hierdoor ook gevonden worden als die niet helemaal correct getranscribeerd is of als er sprake is van een spellingsvariant.	C	C
C4	Wellicht kun je hier zelfs op filteren ("geef mij alle teksten met minimaal 90% correctheid") zodat je grip hebt op fouten in de transcriptie en spellingsvarianties.	C	C
C5	In de toegang wordt, net zoals bij de aanwezigheid van scans, aangegeven dat bij een bepaald inventarisnummer transcripties bestaan.	C	M
C6	Alleen de meest recente versie van een transcriptie wordt als zoekresultaat getoond.	M	M
C7	Zoekresultaten uit alle versies van transcripties worden getoond	C	C
C8	Gegevens worden geduid/gesorteerd met named entity recognition.	W	W
	D - Tonen transcripties aan eindgebruiker	MSCW	MSCW
D1	Wanneer een gebruiker een scan opent waarvan een transcriptie (zoals transcripties via OCR, HTR of handmatig, etc.) bestaat dan wordt deze tekst getoond.	M	M
D2	De meest recente versie van de transcriptie bij die scan wordt getoond.	M	M
D3	Indien van toepassing of gewenst worden ook andere versies de meest recente transcriptie getoond	C	C
D4	Het gebruikte zoekwoord wordt gehighlight in de tekst.	S	S
D5	Een beschikbaar polygoon wordt gebruikt om het de locatie van het zoekwoord in de scan zo nauwkeurig als mogelijk wordt weergegeven.	S	S
D6	Een bezoeker kan zelf instellen of hij/zij bij een scan direct de handmatige transcriptie, OCR of HTR wil zien in een splitscreen of niet.	M	M
D7	Een gebruiker kan een tekst en de getranscribeerde regel in een scan met elkaar vergelijken. Wanneer de gebruiker óf in een regel in de transcriptie klikt óf op een tekstregel in de scan klikt wordt de corresponderende tekstregel of scanregel in geel gearceerd weergegeven.	W	W
D8	Kwaliteit van automatisch getranscribeerde teksten moet inzichtelijk worden gemaakt (zie probabiteit), opnemen van 'disclaimer'.	S	S
D9	Wanneer er van een scan géén transcriptie bestaat dan wordt een module aangeboden om een transcriptie te maken. De input hiervan worden in de keten opnemen (eventueel met autorisatiemoment) en daarna op zelfde manier aanbieden als andere transcripties.	W	S
	E - Bewerken transcripties door eindgebruiker	MSCW	MSCW
E1	Een gebruiker kan, wanneer hem dat is toegestaan, een transcriptie bewerken. Wanneer een transcriptie wordt bewerkt en opgeslagen wordt een nieuwe versie van de transcriptie aangemaakt en direct als meest recente versie bij de scan getoond.	W	S
E2	Rechtenstructuur: wie mag wat wanneer, ook beheer en autorisatie aan achterkant meenemen.	W	S
	F - Beheer	MSCW	MSCW
F1	Er is een systeem beschikbaar waarin transcripties kunnen worden beheerd. Vanuit dit systeem kan een backend-component teksten en bijbehorende metadata ophalen voor beschikbaarstelling in de website.	M	M
F2	Een beheerder kan op een uniforme en geautomatiseerde manier één of meerdere bestanden (in bulk) die binnen een project gevormd zijn in dit systeem opnemen.	M	M
F3	Tevens moeten transcripties die via de site handmatig gemaakt worden geautomatiseerd opgenomen kunnen worden.	W	M
F4	Een beheerder kan verschillende transcriptie-versies van één bepaalde scan of bestand beheren.	W	M
F5	Een beheerder kan nieuwe versies aanmaken van transcripties aanmaken, bekijken, wijzigingen aanbrengen en verwijderen.	W	S
F6	Bij de transcripties kan metadata worden bewaard en beheerd over diverse zaken rondom de creatie en het gebruik van de transcripties. Vastgelegd moet kunnen worden: Oorsprong (project, website, etc.), Machinaal of handmatig gevormd, Oorspronkelijk type bestandformaat (txt, xml, etc.); het beheer gebeurt wel in één formaat, nog bekijken of, hoe en waar een omzetting naar dit ene formaat plaats moet vinden, Kwaliteit (geautoriseerd na handmatige invoer, kwaliteit software van automatische omzetting, etc.), Mutatie (wie mag wanneer wijzigingen aanbrengen, worden deze direct doorgevoerd of pas na modereren, etc.). Deze de metadata moet door een beheerder bewerkt kunnen worden, zowel per individueel record als in bulk.	S	S
F7	Er moet een identifier vastgelegd kunnen worden die een koppeling tussen scan en transcript mogelijk maken.	M	M
F8	Het moet mogelijk zijn om geautomatiseerd uit dit systeem de informatie te halen voor het doorzoekbaar maken, presenteren van en als open data aanbieden van de transcripties.	M	M