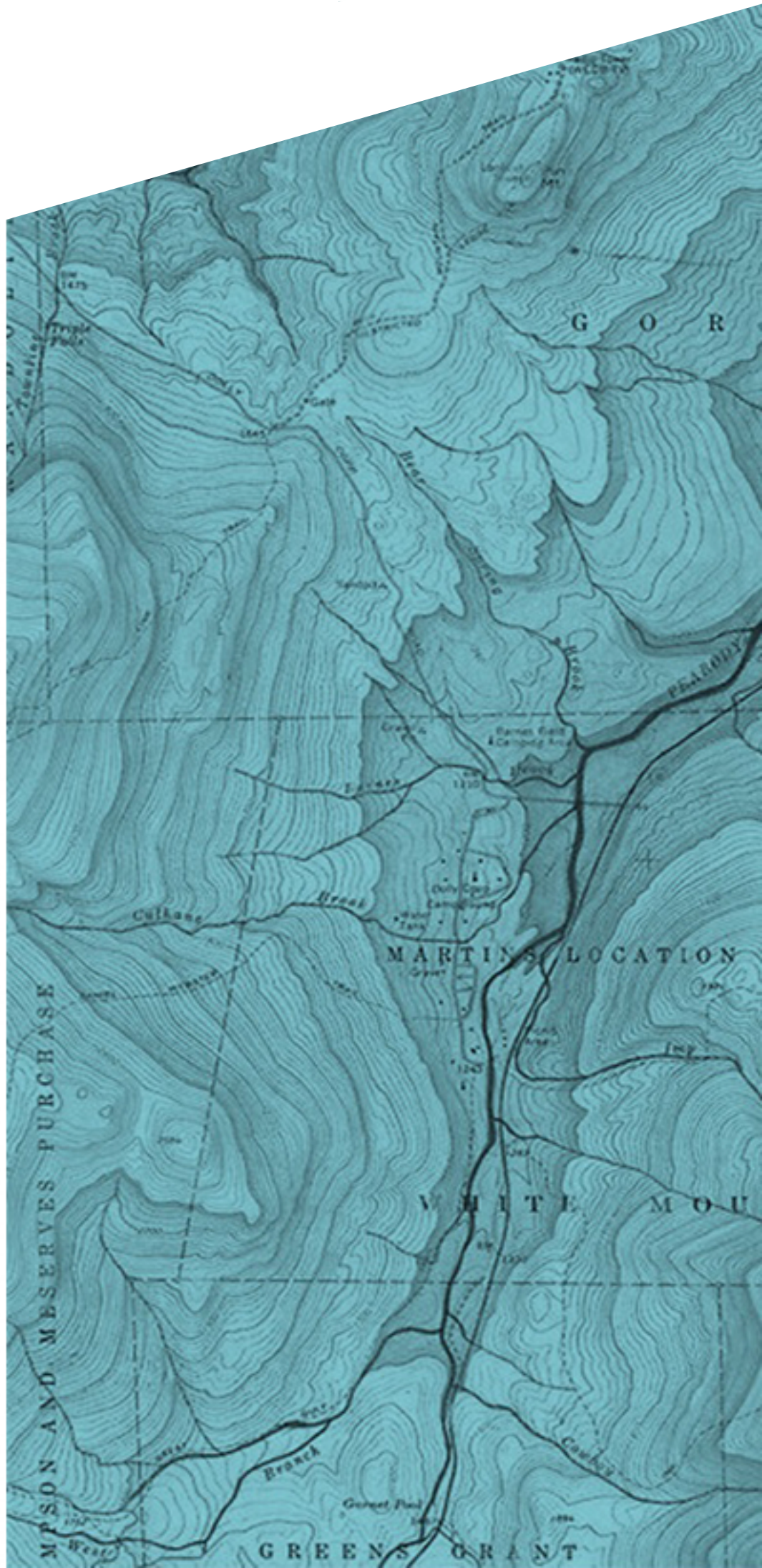


AFSTUDEERERSLAG

Het (semi-) automatisch georeferencen van primaire soortgegevens

JOSINE BLOM, 12082619

Informatiedienstverlening en management, Haagse Hogeschool



Algemene informatie

Versie	1.0
Titel	Het (semi-) automatisch georeferencen van primaire soortgegevens
Afstudeerder	Josine Blom Eiklaan 46 2282 AV te Rijswijk josine.blom@hotmail.com
Studentnummer	12082619
Afstudeerperiode	31 augustus 2015 (2015-2.1) – 8 januari 2016, 2015-2.1 (2015-2.2)
Onderwijsinstelling	Haagse Hogeschool Academie voor ICT & Media Johanna Westerdijkplein 75 2521 EN Den Haag 070 445 88 88
Studierichting	Informatiedienstverlening en –management, voltijd
Afstudeerbedrijf	Naturalis Biodiversity Center Darwinweg 2 2333 CR, te Leiden 071 751 96 00
Bedrijfsmentor	Dr. Marian van der Meij Informatiemanager
Examinatoren school	Klaas Jan Mollema Jochem Mollema

Referaat

Josine Blom, 12082619

Afstudeeropdracht - Het (semi-) automatisch georeferencen van primaire soortgegevens

Naturalis Biodiversity Center, Leiden

September 2015 - januari 2016

Dit document betreft het persoonlijke afstudeerverslag, behorende bij het onderzoek: Het (semi-) automatisch georeferencen van primaire soortgegevens. Dit onderzoek is gedurende 17 weken uitgevoerd, van september 2015 tot januari 2016, bij de organisatie Naturalis Biodiversity Center in Leiden.

Het document fungeert als verantwoording van de uitgevoerde werkzaamheden en opgeleverde producten in het kader van het afstudeertraject bij de Haagse Hogeschool, voor de opleiding Informatiedienstverlening en -management, voltijd. Bijgeleverd bij dit verslag is een document met externe bijlage. Dit zijn de tijdens de stage opgeleverde eindproducten voor de opdrachtgevende organisatie. De externe bijlage hebben de huisstijl van Naturalis en zijn in het Engels geschreven.

Trefwoorden:

- Georeferencing
- Primaire soortgegevens
- Geografische informatie
- Internationale methodes
- Pakketselectie
- Standaarden
- Automatisering

Voorwoord

Voor u ligt het afstudeerverslag 'Het (semi-) automatisch georeferencen van primaire soortgegevens'. Dit verslag is geschreven naar aanleiding van mijn afstuderen voor de opleiding Informatiedienstverlening en -management aan de Haagse Hogeschool. Van september 2015 tot en met januari 2016 ben ik bezig geweest met het onderzoek en het schrijven van dit verslag.

Deze periode viel tussen twee zeer grote projecten binnen Naturalis: net na het digitaliseringproject en vlak voor de verhuizing en verbouwing voor een nieuw museum. Ondanks alle veranderingen en vernieuwingen in de organisatie is er voor mij altijd ruimschoots plaats, tijd en enorme ondersteuning geweest binnen de afdeling Informatiemanagement en de hele organisatie.

Het meest bijzondere was toch wel het enorme enthousiasme dat uit alle hoeken bleek te komen. Georeferencing is een onderwerp dat leeft binnen Naturalis en waar veel draagvlak voor is. Mijn product en onderzoek was zowel bij de opdrachtgevende organisatie, als bij andere biodiversiteitinstellingen een gewenst product en zal daadwerkelijk worden gebruikt en gepubliceerd.

Zonder deze tijd, ruimte, ondersteuning en het enthousiasme was het een veel moeizamer en minder leuk afstudeerproject geweest. Ik wil graag Marian van der Meij, mijn afstudeerbegeleidster bedanken, voor de hulp, de opbouwende kritiek en de vrijheid om zelf mijn proces en producten vorm te geven. Het team informatiemanagement, de afdeling waar mijn afstudeerstage plaatsvond, wil ik bedanken voor het meedenken en de oprechte interesse in de voortgang van het project. Vanuit de Haagse Hogeschool wil ik mijn examinatoren Klaas Jan Mollema en Jochem Mollema bedanken voor de begeleiding vanuit school en de opbouwende feedback tijdens de beoordelingsmomenten. Tot slot gaat mijn dank uit naar iedereen in mijn persoonlijke omgeving, voor de steun, het meedenken en de wijze raad.

Josine Blom
Leiden, 2016

Inhoudsopgave

	Inleiding	6
1	De context van het afstudeerproject	7
	1.1 De organisatie	7
	1.2 De aanleiding	8
	1.3 De opdracht	10
	1.4 Oriëntatiefase en probleemanalyse	11
	1.5 Verantwoording plan van aanpak	13
2	De collectiedata en collectiesystemen	15
	<i>Deelvraag 1</i>	
	2.1 Onderzoeksmethoden	15
	2.2 Resultaten en conclusie	18
	2.3 Evaluatie	21
3	De gebruikers van georeference data	22
	<i>Deelvraag 2</i>	
	3.1 Onderzoeksmethoden	22
	3.2 Resultaten en conclusie	25
	3.3 Evaluatie	27
4	Georeference projecten	28
	<i>Deelvraag 3</i>	
	4.1 Onderzoeksmethoden	28
	4.2 Resultaten en conclusie	30
	4.3 Evaluatie	31
5	Georeference tools, methodes en databronnen	33
	<i>Deelvraag 4 en deelvraag 5</i>	
	5.1 Onderzoeksmethoden	33
	5.2 Resultaten en conclusie	37
	5.3 Evaluatie	38
6	De pakketselectie	39
	<i>Deelvraag 6 en deelvraag 7</i>	
	6.1 Onderzoeksmethoden	39
	6.2 Resultaten en conclusie	41
	6.3 Evaluatie	43
7	De best practice	44
	<i>Deelvraag 8</i>	
	7.1 Onderzoeksmethoden	44
	7.2 Resultaten en conclusie	46
	7.3 Evaluatie	47

8	Eindconclusie en eindadvies	48
9	Discussie	51
10	Evaluatie van het project	53
	10.1 Wat ging er wel goed	53
	10.2 Wat ging er niet goed	54
11	Verantwoording beroepstaken en persoonlijke reflectie	56
	11.1 Reflectie per beroepstaak	56
	11.2 Eindreflectie	61
	Literatuurlijst	62
	Overzicht van tools, & middelen	66
	Verklarend woordenboek	67
	Bijlage:	
1	Overzicht metadata elementen datadictionary	69
2	Codering interviews gebruikers	73
3	Codering interviews CETAF instellingen	77
4	Resultaten deskresearch deelvraag 4	85
5	Tabellen pakketselectie	88
6	Plan van Aanpak	91

Inleiding

Al eeuwen lang worden biodiverse objecten over de hele wereld verzameld door onderzoekers en beheerd in musea zoals Naturalis Biodiversity Center (Naturalis). Met dergelijke collecties wordt de geschiedenis van de biodiversiteit en natuur beschreven en bewaard.

Eén van de belangrijkste toepassingen van dit soort natuurhistorische collecties, is het gebruik van deze data voor wetenschappelijk onderzoek. Naast de taxonomische data, zoals familie, geslacht en soortnaam, zijn de geografische gegevens van deze objecten hiervoor erg belangrijk. Met de geografische verzameldata is het namelijk mogelijk om bijvoorbeeld vroegere en huidige distributie van soorten in kaart te brengen, te analyseren en te combineren met ruimtelijke data van andere soorten en disciplines zoals meteorologie, geologie en medische wetenschappen. Naast fundamenteel onderzoek kunnen hiermee ook vraagstukken over wereldwijde kwesties als natuurbehoud en klimaatverandering worden onderzocht (NBC, 2015b).

Voor het in kaart brengen van objecten zijn geografische referentiepunten zoals coördinaten nodig. Bij grote delen van natuurhistorische collecties zijn deze geografische verzameldata echter alleen in tekstuele vorm, bijvoorbeeld [Leiden, Nederland], geregistreerd.

Naturalis heeft besloten dat de tekstuele geografische objectdata van de collectie op een gestandaardiseerde manier moet worden verrijkt met overeenkomstige geografische coördinaten. Dit kan namelijk de kwaliteit en de toepasbaarheid van de collectiedata vergroten (Arkel et al, 2015). Dit proces, waarbij een tekstuele locatie van een object wordt gekoppeld aan een geografisch referentiesysteem, heet **georeferencing** (Murphey et al, 2004).

Naturalis voerde in 2013 een pilotproject uit waarbij met een testmethode twee deelcollecties werden verrijkt met coördinaten. Ze zijn hiermee niet de enige instelling in Europa die een dergelijke testmethode voor georeferencing onderzochten en uitprobeerde. Uit een meeting tussen verschillende Europese biodiversiteitsinstellingen in 2014 is gebleken dat veel van deze methodes elkaar overlappen en zouden kunnen aanvullen (NBC, 2014a). Mijn afstudeeropdracht volgt uit de behoefte, die tijdens deze meeting naar voren kwam, voor een onderzoek naar de mogelijkheden om deze geteste methodes samen te voegen en te combineren tot 1 internationaal geaccepteerde best practice. Naturalis heeft aangeboden het initiatief te nemen voor het uitvoeren van het benodigde onderzoek. Dit is samen met mij uitgewerkt tot een afstudeeropdracht voor de opleiding Informatiedienstverlening en informatiemanagement aan de Haagse Hogeschool.

Note: Het eindrapport voor de opdrachtgevende organisatie is geschreven in het Engels, aangezien het een internationaal bruikbare best practice diende te worden.

Leeswijzer

In het eerste hoofdstuk beschrijf ik de context, de aanleiding en de hoofd- en deelvragen van het project. Ik beschrijf de oriëntatiefase en verantwoord het plan van aanpak. In hoofdstuk 2 tot en met 7 beschrijf en verantwoord ik per deelvraag de gebruikte onderzoeksmethode, de gegevensanalyse, de resultaten en de conclusies die hieruit volgden. Per deelvraag evalueer ik mijn werkwijze en proces. In hoofdstuk 8 volgen de eindconclusie van het onderzoek en het hierop volgende advies. Hoofdstuk 9 beschrijft kort enkele discussie punten die tijdens het onderzoek aan het licht kwamen. In hoofdstuk 10 evalueer ik het project en bespreek wat er wel en niet goed ging. In hoofdstuk 11 geef ik tot slot een verantwoording van mijn opleiding specifieke beroepstaken en beschrijf ik mijn persoonlijke reflectie op het afstudeerproces.

1 De context van het afstudeerproject

Naturalis is een onderzoeksinstituut en natuur historisch museum dat een lange geschiedenis kent. In 1820 werd het Rijksmuseum van Natuurlijke Historie door Koning Willem I gesticht. Dit was het beginpunt van de collectie die de organisatie die nu Naturalis heet, beheert (NBC, 2015d).

De huidige organisatie, het Nederlands Centrum voor Biodiversiteit Naturalis, ontstond in 2010 uit een samenwerking tussen het Zoölogisch Museum Amsterdam (onderdeel van de Universiteit van Amsterdam), het Nationaal Herbarium Nederland (onderdeel van de Universiteit Leiden, de Universiteit Utrecht en de Wageningen University) en het Nationaal Natuurhistorisch Museum in Leiden. In ruim 250 jaar hebben deze drie instellingen individueel grote collecties natuurhistorische objecten weten te verzamelen (NBC, 2015b). Het hoofddoel van deze samenwerking was het integreren van de collecties van de voormalig losstaande instellingen tot één nationale collectie van ruim 37 miljoen planten, dieren, fossielen, gesteenten en mineralen (NBC, 2015b).

1.1 De organisatie

Missie: “Biodiversiteit beschrijven, begrijpen en verklaren. Voor het welzijn van de mensen en het voortbestaan van de aardse natuur” (NBC, 2012).

Naturalis richt zich op het verzamelen, bewaren en bestuderen van levensvormen. Een deel van de collectie wordt tentoon gesteld in het museum, waar het verhaal van de natuur in zijn natuurhistorische en cultuurhistorische context wordt verteld (NBC, 2015b). Het grootste deel van de collectie wordt echter bewaard voor wetenschappelijk onderzoek, waarbij ze trachten bij te dragen aan de mondiale inspanning om de biodiversiteit te reconstrueren en interactie tussen soorten te begrijpen (NBC, 2015a). Sinds de fusie in 2010 beheert Naturalis één van de grootste natuurhistorische collecties ter wereld, bestaande uit 37 miljoen objecten:

- | | |
|---------------------------------------|---|
| • 18.100.000 insecten | • 250.000 algen |
| • 5.800.000 ongewervelde | • 12.000 gallen |
| • 1.900.000 gewervelden | • 6.000 slijmzwammen |
| • 3.200.000 fossielen | • 121.000 houtmonsters |
| • 800.000 gesteenten en mineralen | • 140.000 banden |
| • 4.600.000 hogere planten | • 14.000 tijdschrifttitels |
| • 605.000 mossen, 282.000 korstmossen | • 57.000 prenten, 13.000 kaarten |
| • 135.000 varens | • 91.500 microfiches |
| • 350.000 schimmels | • 310.000 foto's, dia's en glasnegatieven |

Als **internationaal kenniscentrum en academische werkplaats** voor de universiteiten van Amsterdam, Leiden en Wageningen dragen ze bij aan het opleiden van een nieuwe generatie wetenschappers en academische professionals in de systematiek en biodiversiteit. Ze bieden onder andere moderne onderzoeksfaciliteiten, stagemogelijkheden en verschillende onderwijsmodules en vakken die de bestaande universitaire opleidingen versterken (NBC, 2015a).

Naturalis levert daarnaast ook dagelijks een **bijdrage aan bedrijven, overheden en maatschappelijke organisaties** die in aanraking komen met biodiversiteits materiaal en de determinatie van soorten (NBC, 2015a), zoals:

- Het assisteren van de douane wanneer er lastig te identificeren dieren of planten aan de landsgrenzen worden aangetroffen.
- Het certificeren van hout waarmee ze helpen te bepalen of een partij uit gekweekt of wild hardhout bestaat.
- Het ondersteunen van natuurhobby- verenigingen bij het ontwikkelen en beheren van websites met natuurinformatie.
- Het adviseren van de juweliersbranche bij het bepalen van de echtheid van edelstenen.

De huidige organisatie is ingedeeld in vier sectoren: Onderzoek en Onderwijs, Collectie, Publiek en Markt en ICT (NBC, 2015c). Bij deze afstudeerstage zijn de volgende sectoren betrokken:

- **Sector ICT:** richt zich onder andere op het digitaliseren, beschikbaar en toepasbaar maken van collectiedata en het koppelen ervan aan de digitale bronnen. Hier wordt de afstudeerstage primair uitgevoerd. Specifiek binnen het team datamanagement.
- **Sector Collectie:** verantwoordelijk voor het collectiebeheer en de registratie van de digitale collectie. Deze sector zal worden benaderd voor het project.
- **Sector Onderzoek en Onderwijs:** ruim 120 interne onderzoekers en 200 gastonderzoekers zijn dagelijks bezig met het beantwoorden van uiteenlopende fundamentele en wetenschappelijke onderzoeksvragen. Deze sector zal worden benaderd voor het project.

1.2 De aanleiding

Naturalis voerde in 2013 een pilotproject automatisch georeferencing uit, als onderdeel van een grootschalig digitaliseringsproject.

➤ FCD- collectiedigitalisering project (2010-2015)

Tijdens dit project zijn ruim 7 miljoen van de collectieobjecten van Naturalis gedigitaliseerd. Hierbij is de informatie die bij de objecten hoort, zoals taxonomie, verzameldatum en verzamelplaats, in twee centrale databases geplaatst; één voor zoölogische en geologische data, het Collectie Registratie Systeem (CRS) en één voor botanische data, de Botanical Research And Herbarium Management System (Brahms) (NBC, 2015b). Aanvullend op het grootschalige digitaliseringsproces zijn enkel pilotprojecten voor de verrijking van de digitale informatie uitgevoerd.

➤ FCD- Pilot Georeferencing (2013-2014)

Tijdens dit project is in 2013/2014 een pilot project FCD- Pilot Georeferencing uitgevoerd, door Naturalis, waarin is gekeken wat de mogelijkheden zijn om de vindplaatsgegevens van objecten op een grotendeels geautomatiseerde manier te verrijken met coördinaten en daarmee de datakwaliteit te verhogen (NBC, 2015b). Hiervoor is er onderzoek gedaan naar beschikbare online tools die georeferencing in grote batches tegelijk mogelijk maakt. Hierop volgend is een proces ontwikkeld en getest met de meest bruikbare tool; Google Maps Geocoding API. Hierbij wordt deze API aangestuurd

om grote batches locatiegegevens aan Google Maps aan te bieden en automatisch uitgebreide georeference gegevens terug te geven (Arkel et al, 2015). Deze pilot toonde aan dat productiematig georeferenties met Google Maps een veelbelovende methode is.

Behalve Naturalis hebben veel andere natuurhistorische/ botanische instituten de behoefte om georeference data aan hun objectrecords toe te voegen, omdat ze een duidelijke meerwaarde geven aan de collecties.

➤ **CETAF**

Naturalis Biodiversity Center werkt daarom samen met internationale partners uit de CETAF (Consortium of European Taxonomic Facilities), om de kennis over georeferencing te delen (CETAF, z.d.-a). Binnen deze organisatie zijn er meerdere instellingen die een pilotproject of onderzoek over georeferencing hebben uitgevoerd. Deze vormen samen een interest group, de GIS-experts subgroep, die bereid is samen te werken aan een internationale best practice (CETAF, z.d.-b).

➤ **De GIS-experts subgroep**

Deze onderzoeksgroep heeft besloten dat een internationale best practice kan bijdragen aan een internationale standaardmethode voor (semi) automatisch georeferenties en een infrastructuur voor alle natuurhistorische collecties binnen de EU (NBC, 2014a). Toepassing van de best practice zou moeten leiden tot een verrijking van natuurhistorische databases met betrouwbare en vergelijkbare georeference data. Daarnaast zou een dergelijke standaardmethode en infrastructuur de usability (gebruiksvriendelijkheid) en de kwaliteit van digitale natuurhistorische collecties kunnen vergroten (Arkel et al, 2015).

Naar aanleiding van de resultaten van het pilotproject van Naturalis zijn alle CETAF –GIS- expert instituten benaderd door Naturalis of zij enige ervaring hebben met productiematig georeferenties en of zij interesse hebben om deze informatie te delen (Arkel et al, 2015). Het Natural History Museum London (NHM) en Royal Museum Central Africa Tervuren (RMCA) lieten hun interesse blijken en Naturalis heeft daarop volgend op 2 en 3 december 2014 een meeting over productiematig georeferenties georganiseerd; “The Automated Georeferencing Meeting” (NBC, 2014b).

➤ **The Automated Georeferencing Meeting (2014)**

Tijdens deze meeting is gedetailleerde informatie uitgewisseld over de (technische) methoden die het NHM, RMCA en Naturalis hebben ontwikkeld om grote aantallen collectie-objecten (semi-) productiematig te georeferenties. Uit de meeting kwam naar voren dat NHM, RMCA en Naturalis vergelijkbare en complementaire competenties hebben en gestandaardiseerde GIS tools gebruiken die goed op elkaar aansluiten (NBC, 2014b).

➤ **Afstudeeronderzoek semi- automatisch georeferencing (2015 -2016)**

Verder is de wens uitgesproken om een follow-up onderzoek uit te voeren om de georeference methoden van andere CETAF leden (naast die van het NHM, RMCA en Naturalis) te onderzoeken en al deze projecten met elkaar te vergelijken en te combineren (NBC, 2014a). Hierbij is specifiek behoefte naar een antwoord op de vraag: *“Tot waar kan je op (semi) automatische wijze objecten georeferenties en met welke tools?”*. Naturalis heeft aangeboden het initiatief te nemen voor het uitvoeren van het benodigde onderzoek, hetgeen is uitgewerkt tot mijn afstudeeropdracht.

1.3 De opdracht

Op basis van de eerste gesprekken met de opdrachtgevende organisatie (voorafgaande aan het opstellen van het afstudeerplan) en inzage in een aantal interne documenten over het Pilotproject Georeferencing van Naturalis en de Automated Georeferencing Meeting heb ik de volgende hoofdvraag en deelvragen geformuleerd:

Centrale vraag

“Wat moet je als instelling doen om een collectie geografische verzameldata met tekstuele locatieaanduidingen op een semi automatische manier te verrijken met de overeenkomstige geografische coördinaten op verschillende niveaus/ in verschillende stappen van nauwkeurigheid?”

Deelvragen

1. Wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?
2. Op welke verschillende niveaus willen de gebruikers van Naturalis de geografische data beschikbaar hebben?
3. Welke andere instellingen hebben projecten uitgevoerd om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?
 - a. Wat is de betrouwbaarheid en accuraatheid van de gebruikte methode/tools?
 - b. In welke mate was handmatige correctie nodig om de betrouwbaarheid van de georeferencing data te verbeteren?
 - c. Welke middelen zijn er gebruikt om deze verbeteringen door te voeren?
4. Welke (semi) automatische methodes en tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten?
5. Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren?
6. Welke semi automatische programma's zijn het meest geschikt voor de beschreven collecties?
7. Hoe sluiten deze methodes en programma's het beste op elkaar aan.
8. Hoe kunnen de methodes en programma's het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?

Projectdoel

Het doel van dit project is het maken van een handleiding met methoden en aanbevelingen (best practice) om grote hoeveelheden objectrecords zo geautomatiseerd mogelijk (productiematig) te koppelen aan geografische referentiepunten (coördinaten).

In dit advies moet een use case worden opgenomen, gericht op hoe Naturalis haar collectie met geografische verzameldata kan verrijken met overeenkomstige coördinaten (NBC, 2014). De best practice is herkenbaar en toepasbaar voor de andere geïnteresseerde instellingen van de CETAF GIS-expert sub- groep, waardoor het uiteindelijk kan bijdragen aan een internationale standaardmethode voor (semi) automatische georeferencing en een infrastructuur voor alle natuurhistorische collecties binnen de EU voor de toekomst.

1.4 Oriëntatiefase en probleemanalyse

De oriëntatiefase van het onderzoek heeft plaatsgevonden door middel van een aantal interne bronnen die zijn verkregen via de opdrachtgever bij aanvang van de afstudeerstage en opstartgesprekken met betrokken medewerkers van de afdeling Informatiemanagement. Het doel van deze fase was de breedte van het onderwerp te bekijken, voordat ik dieper op deze situatie inging met de deelvragen. Tijdens de eerstestap van het onderzoek heb ik mij gericht op twee samenhangende vragen die leiden tot de probleemdefinitie en de informatiebehoefte van de opdrachtgever (probleemanalyse). Samen vormden het antwoord op deze vragen de probleemanalyse, die als eerste hoofdstuk van het eindproduct voor de opdrachtgever werd opgeleverd.

Vraag 1: Wat is precies de vraag/probleem van de opdrachtgever?

Het antwoord op de eerste vraag is voornamelijk gebaseerd op bronnen van het pilotproject van Naturalis en verslagen van de bijeenkomsten tussen Naturalis en de CETAF.

Het *'Eindrapport FCD-pilot Georeferencing'* (Arkel et al, 2015), een intern document, beschrijft het volledige onderzoek, alle mogelijk tools/methodes voor georeferencing, het testproces, het datacleaning proces en de uiteindelijke conclusies van het pilotproject van Naturalis. Deze bron gaf zeer uitgebreide achtergrondinformatie over de vraag/het probleem van de opdrachtgever (vraag 1) en daarmee ook een goed inzicht over de reeds intern beschikbare kennis en ervaring in het bedrijf.

Dit eindrapport is alleen gericht op de situatie van Naturalis. Naturalis heeft al eerder een onderzoek gedaan en heeft dus al enige ervaring met georeferencing. De vragen die vanuit deze doelgroep worden aangedragen zijn zeer instelling specifiek en sterk gerelateerd aan hun eigen collectie en collectieregistratiesystemen (Arkel et al, 2015).

Onderzoeksonderwerpen specifiek voor Naturalis:

- Op welke manieren staan de geografische data van de collectie nu geregistreerd in de CRS en het BRAHMS en welke veranderingen zouden nodig zijn om georeference data te ontsluiten?
- Hoe nauwkeurig willen de gebruikers van de collectie dat de objecten worden gegeoreferenced en hoe willen ze dit in het CRS en het BRAHMS geregistreerd hebben?
- Op welke manier(en) kan het aantal dat nodig is voor de zware arbeidsintensieve datacleaning methode, worden gereduceerd.

Dit onderzoek kent echter een breder onderzoeksdomein; namelijk de hele CETAF, waardoor ik het nodig achtte, om in de oriëntatiefase ook informatie te zoeken over dit netwerk.

De CETAF, Consortium of European Taxonomic Facilities, is een Europees netwerk van 33 natuurhistorische museums, botanische tuinen en biodiversiteit- en onderzoeksinstellingen. Naturalis is een van deze aangesloten instellingen. Binnen CETAF is er een subcommissie ISTC, Information Science & Technology Commission, die zich richt op samenwerking tussen de ICT afdelingen/sectoren van de verschillende instituten die aangesloten zijn bij CETAF (CETAF, z.d.-a). Gezamenlijk delen ze expertises en voeren ze onderzoeken uit naar een aantal uiteenlopende onderwerpen: Automated georeferencing is er hier één van (GIS-experts subgroup) (CETAF, z.d.-b).

In 2014 is *The Automated Georeferencing Meeting* georganiseerd, tussen enkele 'expert' instellingen op het gebied van georeferencing, aangesloten bij de CETAF. Naar aanleiding van deze meeting zijn twee documenten (notulen) opgesteld, die zijn gedeeld met alle aanwezige instellingen. Deze heb ik gebruikt voor de probleemanalyse om voor deze doelgroep een specifiekere probleemdefinitie te kunnen maken:

- Follow-up Actions Meeting Automated Georeferencing (NBC, 2014a)
- Technical Summery (NBC, 2014b).

De GIS-expert subgroep van de CETAF vraagt een globaler onderzoek, waarin de focus meer ligt op het hele proces van georeferencing en de precieze rol van belangrijke beheersaspecten van biodiversiteitscollecties in dit proces. Deze doelgroep wil weten welke tools er beschikbaar zijn voor semiautomatisch georeferenzen, hoe deze op elkaar aansluiten, waar belangrijke beheersaspecten van biodiversiteitscollecties, zoals historische collectiedata, nauwkeurigheid, ed, in dit proces passen.

Onderzoeksonderwerpen specifiek voor de CETAF:

- Welke methodes en tools zijn er beschikbaar voor het georeferenzen van grote hoeveelheden data met betrouwbare resultaten en hoe kunnen deze elkaar aanvullen.
- Hoe kun je het beste omgaan met slechte datakwaliteit, zoals missende geografische informatie, spelfouten, historische namen, ed.
- Waar passen de termen nauwkeurigheid (hoe groot is de afstand tussen het georeference punt en de daadwerkelijke vindplaats van het object) en betrouwbaarheid (op welke geografische metadata is dit georeference punt gebaseerd) in het hele proces.

Om tot een eindresultaat te komen, dat voor beide doelgroepen bruikbaar zou zijn en antwoord zou geven op beide soorten vragen, heb ik besloten de situatie van Naturalis te gebruiken als use case, en deze aan te vullen met ervaringen van andere instellingen (via interviews) en externe literatuur.

Vraag 2: Waarom wil de opdrachtgever dit onderzoeken?

Ter beantwoording van de tweede vraag van de probleemanalyse ben ik in gesprek gegaan met medewerkers van de afdeling informatiemanagement die hebben bijgedragen aan het pilotproject georeferencing; Jeroen Creuwels en Marian van der Meij.

Marian van der Meij is in dit onderzoek opdrachtgever en afstudeerbegeleider en was aanwezig bij de Automated Georeferencing meeting. Zij kon daardoor duidelijk beschrijven hoe het vervolgonderzoek precies tot stand is gekomen en welke instellingen hierbij betrokken waren. Daarbij heb ik met haar besproken wat de richtlijnen zijn voor het uitvoeren van het onderzoek en het op te leveren eindproduct. Jeroen Creuwels is één van projectleden van het pilotproject georeferencing geweest en werkt daarnaast aan de datakwaliteit voor het publiceren van de collectiedata op internationale dataportalen. Ik heb met hem gesproken over het belang van georeferenzen en de waardeverhoging van de collectie die deze data kan opleveren.

1.5 Verantwoording plan van aanpak

Het plan van aanpak is op basis van het afstudeerplan uitgewerkt tot een document dat tijdens de doorlooptijd dient als onderzoekplan. Hierin wordt stapsgewijs beschreven hoe de opdracht in de bedrijfscontext wordt uitgevoerd. Het PvA is opgesteld naar aanleiding van de oriëntatiefase naar de opdrachtgevende organisatie, de beschikbare literatuur en oriëntatiegesprekken (zie paragraaf 2.4). In het document staan de volgende projectaspecten omschreven:

- Achtergrond van het project
- Projectactiviteiten
- Projectresultaten en tussenresultaten
- Interne en externe risico's
- Grenzen en randvoorwaarden
- Kwaliteitsbewaking
- Planning
- Projectorganisatie
- Kosten en baten

Voor de opzet van het plan van aanpak heb ik gebruik gemaakt van de methodiek en indeling die Roel Gritt beschrijft in het boek *Projectmanagement* (2011). Deze methodiek en indeling heb ik gebruikt tijdens het merendeel van de blokprojecten voor de opleiding IDM en leek me daarom het meest geschikt om ook voor deze IDM gerichte afstudeeropdracht toe te passen.

Het plan van aanpak is tijdens een presentatie voor de opdrachtgever/ begeleider van de organisatie en de eerste examinerator van de Haagse Hogeschool goedgekeurd op 2 oktober 2015.

De nadruk in het plan van aanpak ligt op hoofdstuk 3, de projectactiviteiten. Het plan van aanpak dient namelijk tijdens dit project ook als onderzoeksplan en beschrijft dus zo volledig mogelijk alle onderzoeksstappen en bijbehorende methodes en op te leveren producten. Hierbij ben ik licht afgeweken van de beschreven methodiek van Roel Gritt. Deze stelt namelijk voor een *work breakdown structure* te maken waarbij de activiteiten op hiërarchisch niveau (deelproduct > week- activiteiten > dagactiviteiten, etc.) beschreven staan (Gritt, 2011). Dit project betreft echter een nieuw soort project zowel voor mij als voor Naturalis. Dit betekent dat de precieze stappen, bijvoorbeeld dagactiviteiten, nog niet te definiëren zijn. Daarnaast is een groot deel van het onderzoek (deelvraag 2 en 3) afhankelijk van de beschikbaarheid en medewerking van externe partijen (gebruikers binnen Naturalis en de CETAF-leden), wat het lastig maakt om dit precies te plannen.

Om toch een mate van hiërarchie, en daarmee overzicht, te creëren in het onderzoeksplan heb ik de activiteiten verdeeld in verschillende stappen die meerdere weken besloegen. De afstudeeropdracht bestaat uit 4 delen:

1. **Oriëntatie:** op het onderwerp, de collectie en de instelling.
2. **Dataverzameling:** van de betreffende collectiedata, de gebruikers en de mogelijke systemen
3. **Gegevensanalyse:** alle verzamelde informatie combineren voor de selectie van de beste tools
4. **Uitwerking:** de conclusies uitwerken tot een best practice die toepasbaar is voor alle CETAF leden

In het plan van aanpak staan bij elk deel de bijbehorende activiteiten als een stappenplan beschreven, inclusief een toelichting op de bijbehorende onderzoeksmethode, beroepstaken en tussenproducten. In tabel 1, hieronder, is een schematisch overzicht te vinden van dit stappenplan, inclusief een globaal tijdschema, de bijbehorende producten, deelvragen en beroepstaken.

Activiteit	Product	Deelvraag	Beroepstaak
Oriëntatie week 1-4			
Plan van aanpak schrijven	PvA		
Aanvullende oriëntatie op probleemdomein uitvoeren	Probleem analyse		Probleem analyse
Dataverzameling week 5-10			
Intern literatuur onderzoek naar betreffende collecties uit voeren	Datadictionary	Wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?	Data-dictionary
Interviews met de gebruikers van georeference data van Naturalis	Overzicht van informatie behoefte van gebruikers	Op welke verschillende niveaus willen de gebruikers van Naturalis de geografische data beschikbaar hebben?	Informatie behoefte
Interviews met CEAT-ISTC instellingen die eigen projecten hebben uitgevoerd.	Interview verslagen en een overzicht van mogelijke methodes/tools.	Welke andere instellingen hebben projecten uitgevoerd om geografische data te verrijken met de overeenkomstige geografische coördinaten? a. Wat is de betrouwbaarheid/ accuraatheid van de gebruikte methode en tools? b. In welke mate was handmatige correctie nodig om de betrouwbaarheid van de georeferencing data te verbeteren? c. Welke middelen zijn er gebruikt om deze verbeteringen door te voeren?	
Literatuuronderzoek naar mogelijke methodes/tools voor (semi) automatisch georeferencing, en middelen om de accuraatheid te verbeteren.	Literatuur review	4. Welke semi automatische methodes en tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten? 5. Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren.	Desk-research
Gegevensanalyse week 11-12			
Vergelijking van de eisen die de gebruikers van collectie stellen en de bevindingen uit de literatuur en interviews	Pakketselectie	6. Bijbehorende deelvraag: Welke semi-automatische methodes en programma's zijn het meest geschikt voor de beschreven collecties? 7. Hoe sluiten deze methodes en programma's het beste op elkaar aan?	Pakket-selectie
Uitwerking week 13-17			
Opstellen van een best practice	Internationale best practice	8. Hoe kunnen de tools en methodes het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige coördinaten?	Para-metriseren

Tabel 1: een schematisch overzicht van de projectactiviteiten

Voor het selecteren en beschrijven van de onderzoeksmethoden heb ik de theorie gebruikt uit *Wat is onderzoek* van Nel Verhoeven (2010). Een boek dat vanuit school is gebruikt tijdens de periode van voorbereiding op het afstuderen (Blok IDM-i/7). Het onderzoek dat bij dit project hoort, was voornamelijk kwalitatief onderzoek (praktijkgericht) en besloeg zowel intern (gebruikers, documenten, systemen) als extern (literatuur, vergelijkbare situaties elders) onderzoek (Verhoeven, 2010).

Tijdens het opstellen van het onderzoeksplan heb ik gekozen voor een combinatie van verschillende onderzoeksmethoden en gegevensanalysemethoden. Zo voer ik tijdens dit project zowel intern literatuuronderzoek uit, neem ik interviews af, doe ik extern literatuuronderzoek om de interviews mee te ondersteunen en voer ik een pakketselectie uit. Deze combinatie aan onderzoeksmethodes draagt bij aan de geldigheid en betrouwbaarheid van het onderzoek (Verhoeven, 2014). Om herhalingen te voorkomen beschrijf en verantwoord ik in de rest van dit verslag de gekozen onderzoeksmethode voor dataverzameling en gegevensanalyse per onderzoeksvraag (hoofdstuk 3-8).

2 De collectiedata en collectiesystemen

Deelvraag 1: wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?

Uit de probleemanalyse bleek één van de vragen van Naturalis uit het Pilotproject Georeferenzen betrekking te hebben op de geografische objectinformatie die is geregistreerd in de collectiemanagement systemen (CMS). De vraag was vooral gericht op benodigde wijzigingen in het datamodel om georeference data op te slaan (Arkel et al, 2015). Hiermee wordt in dit geval bedoeld op zowel de daadwerkelijke geografische metadata als de velden waarin dit wordt ontsloten.

Om dit breder toepasbaar te maken, voor Naturalis en voor de CETAF, heb ik als eerste stap van het onderzoek een inventarisatie gemaakt van de elementen die een rol spelen bij de inrichting van een CMS en de duurzame ontsluiting van georeference data. Hiervoor is een datadictionary gemaakt waarin de gegevens die daarover geregistreerd moeten worden geanalyseerd, gedefinieerd en gespecificeerd ten behoeve van een CMS (Curriculumcommissie IDM, 2010). Deze deelvraag onderzoekt daarbij ook mogelijke beschikbare standaarden voor het ontsluiten van georeference data en de hiervoor benodigde elementen.

Om een compleet beeld te geven van geografische data van een biodiversiteitscollectie is er aan de hand van een use case ook gekeken naar de aanwezige soorten geografische metadata in een dergelijke collectie en hoe dit geïnterpreteerd moet worden voor georeferencing. Naturalis is als use case gebruikt om te analyseren wat de verschillen zijn tussen de huidige en de benodigde metadata elementen.

2.1 Onderzoeksmethoden



Deze deelvraag bestond eigenlijk uit 2 stappen: als eerste de use case, van de collectiedata van Naturalis, waarbij zowel de functionele metadata elementen in de collectiesystemen als de daadwerkelijke geodata werd geanalyseerd. Ten tweede is een literatuuronderzoek gedaan naar internationale standaarden, die gebruikt kunnen worden voor het ontsluiten en beheren van georeference data.

Use case

Naturalis kent sinds het digitaliseringsproject, dat in juli 2015 eindigde, twee centrale databases voor het beheren van de collectiedata (NBC, 2015 b):

- **Het Collectie Registratie Systeem (CRS):** voor zoölogische en geologische data
- **Het Botanical Research And Herbarium Management System (Brahms):** voor botanische data

In tegenstelling tot het pilotproject uit 2013/2014, waar alleen het CRS werd onderzocht, word in dit onderzoek ook het BRAHMS betrokken.

Voor de use case is voor beide collectiesystemen een datadictionary gemaakt. Dit is een overzicht van informatie over de collectiedata die in een collectiemanagement systeem word beheerd, zoals de functionele metadatavelden en beschrijvingen hiervan. Een datadictionary is een zeer bewerkelijk overzicht, dat afhankelijk van het systeem en de hierin beheerde records een bepaalde vorm en grootte kan aannemen (Marsh, 2014). Ik begon daarom met een korte literatuurstudie naar voorbeelden en beschrijvingen van verschillende type datadictionary's. Via de secundaire bron Google (<https://www.google.nl/>), middels de Quick and Dirty methode van (Veen & Westerkamp, 2010) met de zoektermen: 'datadictionary', 'datamodel' en 'metadata model' vond ik enkele voorbeelden:

- Marsh, J. (2014). *Data Dictionary Example: Defining Your Database Tables*. Geraadpleegd op 11 september 2015, via <https://blog.udemy.com/data-dictionary-example/>
- New South Wales schools HSC (z.d.) *Data modelling: Data dictionaries*. Geraadpleegd op 11 september 2015, via http://www.hsc.csu.edu.au/ipt/info_systems/4032/data_dictionaries.htm

Hierbij heb ik mij voornamelijk gericht op afbeeldingen van datadictionary tabellen, in plaats van 'how to' beschrijvingen. Deze beschrijvingen vond ik namelijk erg technisch en te specifiek ingaan op het vastleggen van de specificaties van de elementen, zoals lengte, tekentype, ed. Dit leek mij niet geschikt voor de beantwoording van deze deelvraag, omdat het doel was een beeld te schetsen van mogelijke metadata elementen die belangrijk zijn en gebruikt kunnen worden bij het ontsluiten van georeference data. Specificaties van deze metadata elementen zijn te instellingspecifiek en ik wilde dat het overzicht van benodigde elementen makkelijk toepasbaar zou zijn op situaties in andere instellingen. Hiervoor vond ik de veldnamen en beschrijvingen voldoende. De aanvullende specificaties die aan deze elementen gekoppeld worden, kunnen instellingen zelf bepalen. Op Afbeelding 1 is als eerste een voorbeeld te zien van een datadictionary die ik te technisch vond voor mijn onderzoek en op afbeelding twee is een voorbeeld van een datadictionary te zien met alleen veldnamen en beschrijving.

De enige beschrijving die wel is gebruikt, komt van de beroepstaak omschrijvingen van de opleiding (Curriculumcommissie IDM, 2010). Deze heb ik gebruikt om in het bijbehorende hoofdstuk in het eindrapport en dit afstudeerverslag een datadictionary te kunnen toelichten.

Field name	Data type	Data format	Field size	Description
UserID	Text	NNNN NNNN	8	Unique 8 digit number as text
First name	Text		25	First name of employee
Sur name	Text		25	Surname of employee
DOB	Date	DD/M M/YY YY	10	Date of birth as a short date format
Hourly payrate	Currency	#####. ##	8	Rate of pay in dollars per hour
Height	Real	#.##	3	Height in meters, with two decimal places
Fees Paid	Boolean		1	Y or N for Yes or No

Afbeelding 1: voorbeeld van een datadictionary met veel technische elementen (New South Wales schools, z.d.)

Tabel: Feed_Mapping_Value

Doel: Feed_Mapping_Value: de veldwaarden zoals 'lichtgroen', 'donkergroen', 'kaki' allemaal laten mappen als 'groen'. Of: 'audiomeubels' en 'audiomeubelen' laten mappen naar 'audiomeubilair'.

Id	Unieke identifier (PK)
Feed_id	De unieke identifier van de feed in tabel 'Feed'.
Feed_field_value	De waarde van het bronveld.
Mapped_field_name	Deze kolom bevat de naam van de mapped_field_name zoals deze in tabel Feed_Mapping_Key gemapped is. Het opslaan van die naam in de tabel Feed_Mapping_Value zorgt er louter voor dat er een relatie ontstaat tussen de Feed_Mapping_Key en Feed_Mapping_Value tabel. Wanneer de Key-tabel wordt bijgewerkt moet de Value-tabel dus ook worden bijgewerkt. Tevens moet bij het selecteren (zoeken) naar waarden in de Value-tabel aangegeven worden op basis van welke Key de Value gezocht wordt. Ook daar wordt dit veld weer gebruikt voor het maken van onderscheid.
Mapped_field_value	De waarde van het doelveld.

Tabel: Feed_Data

Id	Unieke identifier (PK)
Feed_id	De unieke identifier van de feed in tabel 'Feed'.
Product_id	De unieke identifier van het product in tabel 'Product'.
Feed_field_name	De naam van het bronveld in de XML-feed binnen de '<Product>' tag.
Feed_field_value	De waarde van het bronveld in de XML-feed binnen de '<Product>' tag.

Tabel: Product

Id	Unieke identifier (PK)
Active	Mogelijke waarden 'Y' / 'N'. Of het product moet worden weergegeven op de publieke website.
Active_from	Kolom 'active_from' kan met een timestamp gevuld worden om producten op moment X live te laten gaan. Standaard = '0'.
Unique_product_identifier	Het unieke kenmerk waaraan een product te herkennen is. Dit kan een identifier zijn in alle vormen en maten. Bijvoorbeeld: 'AUDIOM30291-29938ZW' waarmee de feedleverancier een zwart audiomeubel kenmerkt.

Afbeelding 2: voorbeeld van een datadictionary met weinig technische elementen (Direct websolutions, 2015)

De use case is uitgevoerd door middel van deskresearch, waarbij ik de collecties uit het CRS en het BRAHMS heb geanalyseerd en beschreven. Hierbij heb ik ook gebruik gemaakt van ondersteunende interne literatuur:

- Collectieplan Digitale Collectie (NBC, 2012)
- Data model Brahms (NBC, z.d.a),
- Open Content Beleid (NBC, 2013b)
- Richtlijnen Digitaliseringsproject (NBC, 2013 a)

Deze Documenten zijn gevonden door in de interne Google Drivemappen, van de afdelingen ICT en Collectiebeheer te zoeken met de zoektermen 'collectie', 'collectie- plan', 'BRAHMS' en 'CRS'.

Deze use case is vormgegeven in een hiërarchisch schema, en niet in een tabel, om de hiërarchie en samenhang van de verschillende metadata elementen zo duidelijk mogelijk weer te geven. Daarbij is een beschrijving gemaakt van de huidige stand van zaken van de aanwezige geografische data in de collecties (NBC, z.d.b). Hiervoor heb ik uit zowel het Brahms als uit het CRS uit alle collecties een aantal verschillende soorten locatiebeschrijvingen geselecteerd om een beeld te schetsen van de onderlinge verschillen tussen de aanwezige geodata in de collecties. Ik heb gekozen om dit op te nemen in de datadictionary, om de gebruikers van de best practice een beeld geven van de soorten geografische informatie die voor kunnen.

Standaarden

Deze use case werd aangevuld met een onderzoek naar mogelijke standaarden specifiek voor georeference data van biodiversiteitscollecties. Uit de probleemanalyse bleek namelijk dat de uitwisseling en mogelijkheid om data uit verschillende collecties te combineren, één van de belangrijkste doelstellingen is van de CETAF en van Naturalis (NBC, 2014b). Internationale standaarden bevorderen deze doelstellingen.

Dergelijke standaarden werden al voor dit onderzoek toegepast in de collectie van Naturalis (NBC, 2015b). De nieuwe georeference data moeten hier dus bij aansluiten. In de eindrapportage van het Pilotproject Georeferencing werden 2 mogelijke standaarden genoemd; *ABCD*, *DarwinCore (DwC)*. Door middel van deskresearch heb ik aanvullende informatie gezocht over deze standaarden en de geografische metadata elementen die zij adviseren. Via de secundaire bron Google (<https://www.google.nl/>) middels de Quick and Dirty methode met de zoektermen 'Darwin Core' en 'ABCD Standard' heb ik naam aanvullende informatie gezocht, (Veen & Westerkamp, 2010). Ik wilde hiermee vooral overzichten van de dataschema's van deze standaarden vinden.

2.2 Resultaten en conclusie

Naar aanleiding van de deskresearch kwam ik tot 2 standaarden 1 set guidelines die bruikbaar waren voor het ontsluiten van georeference data in biodiversiteitscollecties. Samen met de resultaten van de use case van Naturalis is een selectie gemaakt van de meest belangrijke (primair) en aanvullende (secundaire) metadata elementen die nodig zijn voor een instelling om de toekomstige georeference data te kunnen beheren.

Use case

De collectiemanagement systemen van Naturalis zijn naar aanleiding van het pilotproject dusdanig ingericht, dat georeference data in de toekomst ontsloten kan worden (NBC, 2013a). In de collecties van Naturalis is een eigen set elementen voor geografische metadata gebruikt, die deels overeenkomt met de metadataseten van ABCD en DwC (NBC, 2012). De set die Naturalis gebruikt bevat echter een aantal aanvullende elementen bovenop de set minimale elementen voorgeschreven door de standaarden.

Alle geografische metadata in het CRS komen onder het element [Gathering site] te staan. Dit veld is opgebouwd uit 33 onderliggende metadata elementen, waarvan er 23 zijn gericht op georeference data (NBC, z.d.b). De samenhang tussen de velden met betrekking op coördinaten is in afbeelding 1 te zien.

Coordinates	
Index	
Method	
Verbatim coordinates	
Accuracy	
Distance in meters	
Lat Long	Lat Long
	Latitude
	Longitude
UTM	System
	UTM
	Zone
	subzone
	NS
	Easting
Grid	Northings
	Datum
	Grid
	Cell system
Grid	Cell code
	Qualifier
	Latitude
	Longitude

Afbeelding 3: Samenhang tussen alle metadata velden met betrekking tot coördinaten in het CRS

In het Brahms zijn er minder velden voor het ontsluiten van geografische objectinformatie; namelijk 13. De volledige overzichten van alle metadata elementen en bijbehorende definities uit het CRS en het BRAHMS zijn te vinden in bijlage 2. Op afbeelding 2 is een screenprint te zien van de geografische metadata elementen uit het BRAHMS. Het grote verschil tussen beide collectiesystemen is de mate van hiërarchie: In het CRS is er wel een hiërarchie waar te nemen en zijn veel velden opgebouwd uit weer onderliggende velden. In het BRAHMS hebben alle metadata elementen hetzelfde niveau.

country	major country area	minor country area	locality	locality notes	lat	ns	long	ew	llunit	lres	lorig	ldatum	q
Unknown				memo	0.000000		0.000000						
Madagascar	Mahajanga (Majunga)			Memo	16.190000	S	46.190000	E	DM	4	sheet		
Madagascar	Mahajanga (Majunga)			Memo	16.190000	S	46.190000	E	DM	4	sheet		
Madagascar	Mahajanga (Majunga)			Memo	15.290000	S	47.350000	E	DM	4	sheet		
Madagascar	Antsiranana (Diego Suarez)			Memo	12.254600	S	49.330500	E	DMS	7	sheet		
Madagascar	Toamasina (Tamatave)			Memo	15.300000	S	49.460000	E	DM	4	sheet		
Madagascar	Toamasina (Tamatave)			Memo	15.300000	S	49.460000	E	DM	4	sheet		
Madagascar	Toamasina (Tamatave)			Memo	15.402400	S	49.575100	E	DMS	5	sheet		
Madagascar	Toamasina (Tamatave)			Memo	15.402400	S	49.575100	E	DMS	5	sheet		
Madagascar	Toamasina (Tamatave)			Memo	15.402400	S	49.575100	E	DMS	5	sheet		
Madagascar	Toamasina (Tamatave)			Memo	15.402400	S	49.575100	E	DMS	5	sheet		
Madagascar	Toamasina (Tamatave)			Memo	15.402400	S	49.575100	E	DMS	5	sheet		
Madagascar	Antsiranana (Diego Suarez)			Memo	14.200000	S	49.400000	E	DM	4	sheet		
Madagascar	Toamasina (Tamatave)			Memo	18.502500	S	48.174500	E	DMS	7	sheet		
Madagascar	Antsiranana (Diego Suarez)			Memo	13.124600	S	49.332100	E	DMS	7	sheet		
Madagascar	Mahajanga (Majunga)			Memo	16.190000	S	46.490000	E	DM	4	sheet		
Guinea	Nzérékoré			Memo	7.424600	N	8.234000	W	DMS	7	sheet		
Madagascar	Mahajanga (Majunga)			Memo	16.371400	S	46.133600	E	DMS	7	sheet		
Madagascar	Antsiranana (Diego Suarez)			Memo	13.135000	S	49.451300	E	DMS	7	sheet		
Madagascar	Antsiranana (Diego Suarez)			Memo	12.233400	S	49.202800	E	DMS	7	sheet		
Madagascar	Mahajanga (Majunga)			Memo	14.542900	S	49.241800	E	DMS	7	sheet		
Madagascar	Mahajanga (Majunga)			Memo	14.550400	S	49.264700	E	DMS	7	sheet		
Madagascar	Antsiranana (Diego Suarez)			Memo	12.374600	S	49.112800	E	DMS	7	sheet		
Madagascar	Antsiranana (Diego Suarez)			Memo	12.361800	S	49.263500	E	DMS	7	sheet		

Afbeelding 4: Samenhang tussen alle metadata velden met betrekking tot coördinaten in het BRAHMS

Met betrekking tot de daadwerkelijke metadata zitten er grote verschillen in de mate waarin deze zijn ingevuld, en verschilt de ingevulde data onderling in vorm, lengte, taal, etc. Zo kunnen de velden data in tekstvorm bevatten, maar ook coördinaten en andere formaten zoals de Rijksdriehoekcoördinaten, of beschrijvingen van plaatsten. Enkele voorbeelden van tekstuele locatiebeschrijvingen in de collectie van Naturalis zijn hieronder te zien:

- "MUNNEKEZIJL; LAUWERSMEER; KOLLUMERWAARD; BREDE SLOOT 05.27. AC 211.735-543.209"
- "HORST-AMERICA; SAAR BIJ SPOORLIJN"
- "ES 4491 Nederland (Z), (MV) Inlaag Hoofdplaat West Langeweg"
- "Noord- Thailand"
- "NEDERLAND (N.H.) IJsselmeeroever bij Broekerhaven 2 km Z.W.v.Enkhuizen"
- "Z.-Spanje, prov. Malaga. Langs pad"
- "wegrand op half-open grond, Kr. wit Pfafleralp (Löhuental)"
- "Prov. : Antalya. Lar (± 10 S.E. of Antalya)"
- "Fron an evident by bind swarm in areble land on the lower of River Copse, above Inkpen, Berks"

Voor het merendeel hiervan geldt dat ze zowel door middel van een thesaurus als verbatim zijn ingevuld (Arkel et al, 2015). Het grootste deel van de bovenstaande voorbeelden komt uit de velden 'Locality' en 'Full locality'. Hierin worden vaak toelichting gegeven op de locatiegegevens die met gecontroleerde lijsten en elementen vast gelegd worden, in vrije tekstvorm (NBC, 2013)

Standaarden

Door middel van deskresearch kwam ik via de websites van GBIF (Global Biodiversity Information Facility) en TDWG (Biodiversity Information Standards) uit op 2 mogelijke standaarden; ABCD en Darwin Core. Naast

de twee standaarden vond ik ook de 'BioGeomancer Workbench Guidelines'. Deze Guidelines combineren de ABCD en DwC tot een meer handbare set elementen. De Workbench is in 2012 stopgezet, na afronding van het project maar de best practice word nog veel toegepast.

- A. **Access to Biological Collections Data (ABCD)** is een zeer geavanceerde data-standaard die het mogelijk maakt om veel informatie gestandaardiseerd te registreren. De standaard bevat meer dan 1000 termen en kan worden geschaald naar specifieke informatie behoeften (TDWG, 2010a). Het is niet mogelijk of zelfs noodzakelijk alle termen te gebruiken. ABCD kent ook extensies om specifieke gegevens van biodiverse collecties verder te beschrijven, zoals de DNA extensie. Het gebruik van de standaard ABCD vereist vaak speciale software.
- B. De **Darwin Core** standaard is een relatief eenvoudige standaard met een totaal van circa 180 termen, onderverdeeld in verschillende categorieën (TDWG Wiki, 2010b). Darwin Core is bijzonder bruikbaar voor de kerngegevens van het verzamelmoment van een soort. Meer informatie kan worden opgeslagen in de zogenaamde Darwin Core extensies, waarbij aanvullende gegevens worden gekoppeld aan "kerngegevens", zoals de Geospatial Element definities (TDWG Wiki, 2010b).

ABCD en Darwin Core kunnen ook worden gecombineerd (GBIF, 2010). Er zijn daarvoor mappings beschikbaar, waarin de overappende velden uit beide standaarden worden gelinkt. Naast de algemene geografische metadata elementen (Higher Geography, Continent, Waterbody, etc.) kent de mapping tussen ABCD en Darwin Core 15 georeference specifieke metadata elementen (TDWG, 2007)

- C. De '**BioGeomancer workbench Guidelines**' betreft een georeferencing tool, die is ontwikkeld door de University of California, Berkeley. De tool wordt aangevuld met een best practice; '*Guide to best practices for georeferencing*', die handleidingen en richtlijnen geeft voor het georeferencen van een biodiversiteitscollecties (Chapman, 2006). Deze bron beschrijft een set van 15 metadata elementen deels gebaseerd op de ABCD en DwC standaarden.

Conclusie

Geografische informatie die de distributie van soorten betreft, kan worden verdeeld in 2 sets elementen die nodig zijn voor ontsluiting. Als eerste de set elementen die de originele locatiebeschrijving, zoals beschreven op het label of in het veldnotitieboek, kunnen vastleggen. Deze set bevat meestal velden zoals [Country], [State], [provincies], [Island], [locality], [Station number], [Full locality], etc. Het is belangrijk ook deze data te behouden tijdens het georeferencen van een collectie door de nieuwe georeference data toevoegen aan een [gathering site] of [locality]. Vaak is dit namelijk van historische waarde. De tweede set elementen beschrijft de daadwerkelijke georeference data en het bijbehorende proces of methode (Chapman, 2006).

Als je de standaarden vergelijkt wordt het duidelijk dat er veel velden zijn die terugkomen in alle drie de overzichten en in de dataset van Naturalis. Deze kunnen daardoor worden beschouwd als primair benodigde data elementen voor het ontsluiten van georeference data in een biodiversiteitscollectie.

Aanvullend daarop zijn er ook enkele velden die in meer dan 1 van de overzichten terugkomen, maar wellicht onderling licht van definitie/beschrijving verschillen. Deze velden kunnen waarde toevoegen aan de collectiedata en zijn onderdeel van de geadviseerde metadata velden, maar de precieze definitie ervan kan per instelling verschillen. Deze secundaire set elementen bevat ook velden die betrekking hebben op het georeference proces en de methode die is gebruikt om de gegevens te genereren. Deze kunnen worden vastgelegd in een collectiemanagement systeem om de voortgang en ontwikkelingen van een collectie bij te houden en eventueel om later de georeference data te aan te passen. In de bijlage 1 is een tabel te vinden met de standaarden, metadata elementen en definities die als primair en secundair worden beschouwd.

2.3 Evaluatie

Voor de datadictionary had ik ruim tijd ingepland, omdat biodiversiteitscollecties vaak taxonomisch zijn ingedeeld en ik geen achtergrondkennis heb met betrekking tot soorten en taxonomie. Een dergelijke collectie heeft een hele eigen manier van beheer en vraagt metadata velden, die je in weinig andere collectie tegenkomt, zoals DNA- codes, determinatiegegevens, mogelijke rang op de IUCN rode lijst. De meeste medewerkers op de afdeling informatiemanagement en het merendeel van mijn interviewpartners hebben wel een achtergrond in biologie, archeologie, e.d. Ik vond het belangrijk om uitgebreid kennis te maken met de collectie, zodat de communicatie tussen mij en de betrokken medewerkers niet in de weg werd gestaan door mijn gebrek aan kennis van de collectie.

Als medewerker op de afdeling informatiemanagement had ik toegang tot beide collectiemanagement systemen van Naturalis, en had ik de rechten om alle velden en functionaliteiten te bekijken, zonder bewerkrechten. Deze systemen speelden een grote rol in mijn onderzoek, en ik heb mij daarom voorafgaand voor deelvraag 1 uitgebreid georiënteerd op de systemen en het leren zoeken in de databases.

Vanaf de aanvang van de stage had ik enkele uren besteed aan het rondklikken en rondkijken in de collectie. Beide systemen zijn in eerste instantie zeer intuïtief en makkelijk in gebruik, maar er is, als je beter bekend ben met de systemen, veel meer mogelijk dan in een eerste oogopslag duidelijk wordt. Om optimaal gebruik te kunnen maken van alle functies heb ik hulp/uitleg gevraagd van de hiervoor verantwoordelijke medewerkers van de afdeling informatiemanagement. Na aanleiding hiervan kon ik gerichter en makkelijker zoeken in de beide systemen om de juist metadata velden en interessante metadata te vinden.

Deze uitleg heb ik gevraagd na een korte literatuurstudie naar datadictionary 's. Hierdoor kon ik goed bepalen welke onderdelen van het CRS en het Brahm's voor mij nuttige informatie zouden bevatten en kon ik ook vragen stellen over bepaalde onderdelen. Ik had door mijn deskresearch al een beeld in mijn hoofd, van hoe ik de datadictionary wilde vormgeven. Na aanleiding van de uitleg moest ik dus omschakelen.

In eerste instantie had ik in gedachte een tabel te maken van alle metadatavelden en definities, maar naar aanleiding van de uitleg ben ik hiervan afgestapt. Ik ontdekte op dit moment pas dat veel van de velden, die direct zichtbaar zijn, bestaan uit enkele onderliggende velden, zoals het voorbeeld hieronder:

- **Coordinates:** all following fields combined:
 - **Verbatim coordinates:** Original coordinate data as appearing on a label or in an original entry
 - **Accuracy:** Textual statement of degree of accuracy
 - **Lat long**
 - **Lat:** Latitude in decimal degrees
 - **Long:** Longitude in decimal degrees

In een tabelvorm kon ik de hiërarchische structuur, zoals deze in het CRS voorkomt, niet duidelijk in beeld brengen, en dus heb ik gekozen voor de huidige vorm. Deze keuze heeft ertoe geleid dat het een simpele en overzichtelijke datadictionary is geworden, die voor andere instellingen bruikbaar is als voorbeeld/ use case en niet onnodig veel tijd heeft gekost om te maken. Door alleen de veldnamen en beschrijvingen te noemen in het onderzoek, kan iedere instelling die hier gebruik van maakt, zelf bepalen in welke taal, lengte tekstvorm, ed. ze deze velden willen invullen. De wijze van ontsluiten van de velden, is namelijk te instelling specifiek en zeer afhankelijk van het collectiemanagement systeem dat wordt gebruikt.

3 De gebruikers van georeference data

Deelvraag 2: op welke verschillende niveaus willen de doelgroepen van Naturalis de geografische data beschikbaar hebben?

‘Fitness for Use’, is een begrip waartegen de kwaliteit van een collectie, van welke aard dan ook, kan worden gemeten (Chrisman 1991). Fitness for use stelt dat de datakwaliteit sterk gerelateerd is aan het gebruik van de data en daarmee dus ook gerelateerd is aan de gebruikers. Dit kan dus niet worden vastgesteld zonder de gebruikers en hun doelen vast te stellen en te analyseren. Voor biodiversiteitscollecties betekent dit dat de waarde van geografische data is gecentreerd rond het verbeteren en verrijken van een verzameling, zodat deze bruikbaar is voor onderzoek, collectiebeheer en de toegankelijkheid van de collectie (Chapman, 2006).

Om te kunnen vaststellen wat de behoefte en doelen van gebruikers van geografische gegevens van biodiversiteitsdata zijn, ben ik in gesprek gegaan met een selectie vaste gebruikers van de collectie van Naturalis. De interviews zijn afgenomen om te bepalen hoe de huidige geografische gegevens moeten worden verrijkt, om de werkzaamheden van deze gebruikers te vergemakkelijken. De verzamelde informatie uit deze interviews is omgezet naar een overzicht van de informatiebehoefte: wensen en eisen, die worden gesteld aan toekomstige georeference data. Op basis van dit overzicht, kunnen bij de pakketselectie tools worden geselecteerd die hieraan voldoen en de gewenste gegevens genereren.

3.1 Onderzoeksmethoden

Voor deze deelvraag is Naturalis, in dit geval de gebruikers van hun collectie, opnieuw benaderd als een use case. De resultaten zijn echter wel ingedeeld in drie algemene toepassingsgebieden die profiteren van het verrijken van de geografische informatie en in de meeste biodiversiteitinstellingen belangrijk zijn; collectiemanagement, onderzoek en toegankelijkheid. Deze drie toepassingsgebieden zijn gebaseerd op informatie uit het *Collectieplan Digitale Collectie*, het *Beleidsplan 2013-2016* en de resultaten van het FCD Digitalisering Project van Naturalis (*Uit het depot, op het web*, 2015). Hierin worden onder andere de doeleinden, plannen, mogelijkheden en resultaten van de digitale collectie beschreven.

Binnen Naturalis zijn er per toepassingsgebied verschillende gebruikersgroepen die gebruik maken van de geografische informatie in de collectie en profiteren van de verrijking hiervan met georeference data:

- **Collectiebeheer (collectiemanagement):** zij beantwoorden en ondersteunen de vragen van zowel interne als externe onderzoekers met behulp van de collectie en hebben informatie over de soorten vragen naar geografische informatie er zijn.
- **Ontwikkeling Bioportaal (toegankelijkheid):** dit is een intern project van Naturalis om de toegankelijkheid van de collectie te vergroten waarbij de basisgegevens van de collectie toegankelijk wordt gemaakt via een web portal (<http://bioportal.naturalis.nl/>). Dit project gebruikt ook de geografische collectiedata om de basisgegevens op een kaart te tonen en geografisch doorzoekbaar te maken.
- **Wikipediabeheer (toegankelijkheid):** dit is ook een intern project van Naturalis om de toegankelijkheid van de collectie te vergroten, waarbij delen van de collectiedata van Naturalis wordt gekoppeld aan Wikipedia pagina's. Dit geldt ook voor spreidingsinformatie van soorten.
- **Onderzoek (onderzoek):** Naturalis heeft ruim 120 onderzoekers en 200 gastonderzoekers die de collectie en collectiedata gebruiken voor wetenschappelijk en fundamenteel onderzoek. Geografische data biedt deze gebruikers de mogelijkheid soorten in kaart te brengen te koppelen aan gegevens uit andere disciplines.

De meeste gebruikers maken momenteel gebruik van de huidige beschikbare geografische data (tekstuele beschrijvingen) in de collectie, maar dit is niet altijd toereikend voor hun werkzaamheden en doelen. Geografische referentiepunten en aanvullende metadata kunnen deze groepen helpen om hun werk beter en efficiënter uit te voeren en creëren zelfs nieuwe mogelijkheden. Om de informatiebehoefte per toepassingsgebied vast te stellen, zijn er interviews afgenomen met één vertegenwoordiger van elke gebruikersgroep. Deze interviews hebben intern plaatsgevonden.

- Collectiebeheer: Hoofd collectiebeheer entomologie
- Ontwikkeling BioPortal: Software ontwikkelaar het Bioportaal
- Wikipedia beheer: Wikipedian in Residence
- Onderzoek: Interne onderzoeker botanie

De interviews zijn in ongestructureerde vorm afgenomen. Hierbij is alleen een topiclijst opgesteld van onderwerpen die aan bod moesten komen in het gesprek (Verhoeven, 2014). Deze topiclijst bestond uit enkele begrippen die een rol spelen bij het gebruik van georeference data. Het was een vrij korte lijst, maar het doel was om meer diepgaande informatie, zoals de motivatie en achterliggende redenen van de informatiebehoefte, te verzamelen over een relatief kleine hoeveelheid deelonderwerpen van het onderzoek.

Deze vorm van interviewen was voor deze situatie het meest geschikt omdat de respondent hierbij zelf veel ruimte heeft om zo veel mogelijk input te leveren en waarbij zijn behoefte aan informatie en ervaringen centraal stonden (Verhoeven, 2014). Door het gebruik van uitsluitend een topiclijst wordt de geïnterviewde niet gestuurd door vragen. Daarbij was het lastig om van te voren in te schatten welke kant de resultaten op zouden gaan. In het vorige onderzoek van Naturalis was namelijk geen aandacht besteed aan de mogelijke gebruikers, dus het was een nieuw onderwerp. Ik wilde me in de interviews niet laten belemmeren door vastgelegde vragen en de ruimte hebben met doorvragen in te spelen op de eerste reacties en antwoorden van de interviewpartners (Baarda, 2014)

De topics op de topiclijst waren als volgt:

- Knelpunten/problemen bij gebruik tekstuele geografische data
- Gebruik toekomstig georeference data
- Belang georeference data
- Gewenste nauwkeurigheid georeference data
- Effect op werkzaamheden

Voor de analyse van de informatie uit de interviews, heb ik samenvattingen (transcripties) gemaakt van de gesprekken. Het maken van letterlijke transcripties was helaas niet mogelijk, omdat ik niet bij elk gesprek een geluidsopname mocht maken. Het leek mij beter om in dit geval alle gesprekken op dezelfde manier uit te werken, zodat ze ook op dezelfde manier konden worden geanalyseerd.

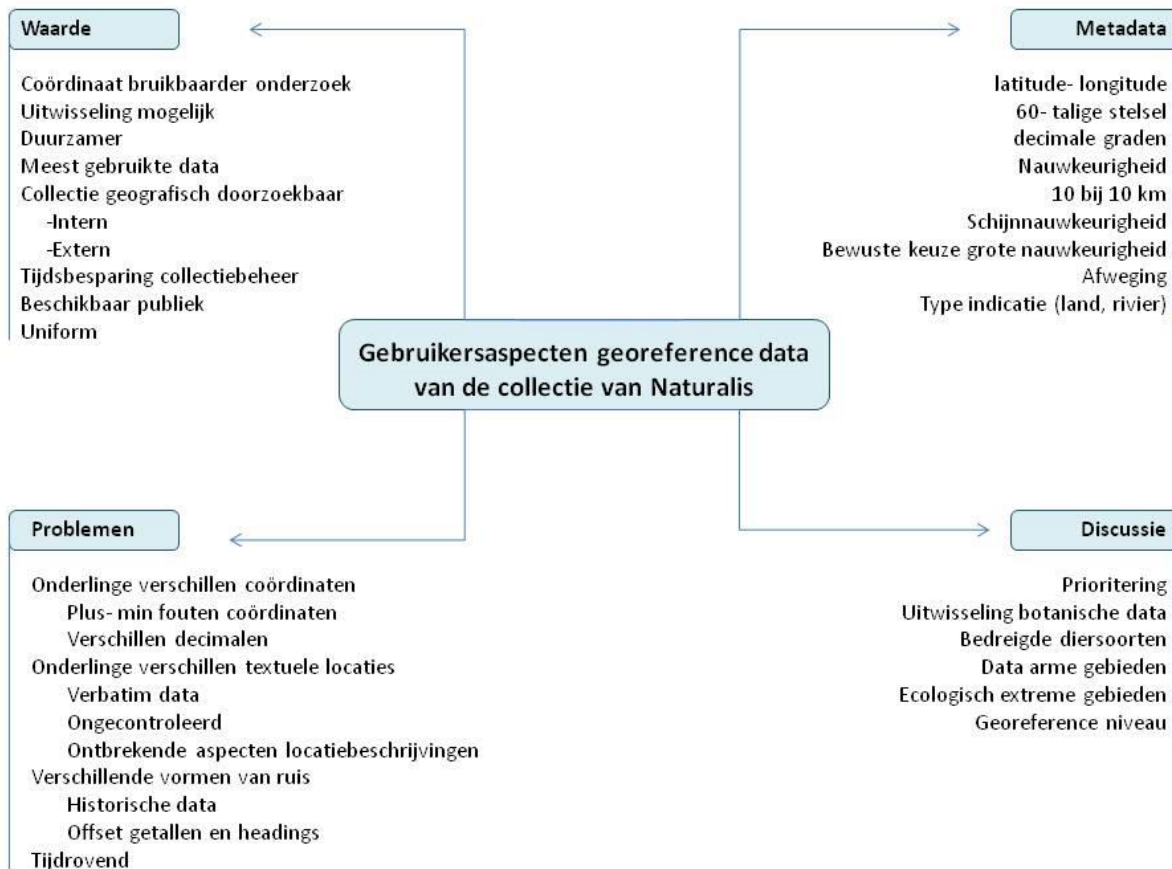
De interviews waren diepte-interviews en betreffen dus kwalitatieve data. Dit is vaak ongestructureerd en niet 'telbaar', waardoor coderen nodig was om patronen te ontdekken in de transcripties (Baarda, 2014). De interviews heb ik niet volledig gecodeerd volgens de methode beschreven in 'Leren Communiceren' van Steehouder (2012). Ik vond dit te tijdrovend voor het resultaat dat het oplevert. Bij een volledige codering van een transcriptie krijgt elke sectie (een stuk tekst, zin of alinea) een label toegewezen waarop in één woord staat welke variabelen in deze sectie staan genoemd (Steehouder et al, 2012). Voor mijn analyse en codering heb ik sommige stukken niet gelabeld omdat ze niet relevant of bruikbaar waren voor mijn onderzoek. Ik heb mij dus alleen gericht op belangrijke begrippen, op basis waarvan ik onderling tussen de interviewverslagen een vergelijking kon maken. De gemarkeerde begrippen betroffen indicatoren van mogelijke:

- Wensen/eisen die aan de metadata elementen werden gesteld
- Problemen die zich nu in de geografische data voordeden
- Waarde van de georeference data als deze wordt toegevoegd (en effect op werkzaamheden)
- Discussiestuk

Het doel van deze interviews was om concrete wensen en eisen te detecteren hoe de collectie verrijkt moet worden met geografische referentiepunten volgens de gebruikers. Alleen begrippen die dergelijke punten omschreven hoefden voor de analyse te worden gebruikt. Daarom heb ik het laatste label: 'discussiestuk', toegevoegd aan het codeboek. Dit betrof problemen, knelpunten vragen ed. die niet onder mijn deelvragen onder te brengen waren maar wel van belang waren voor een instelling tijdens het georeferencen. Deze onderwerpen worden in het discussiehoofdstuk van het eindrapport besproken.

De enige uitzondering op deze werkwijze geldt voor het interview met interne onderzoeker Niels Raes. Hij is namelijk geïnterviewd zowel voor deze deelvraag als voor de volgende deelvraag. Hij heeft namelijk een eigen tool gecreëerd die hem ondersteunt bij het georeferencen. Hij is voor deze deelvraag alleen geïnterviewd voor de topics: 'belang georeferencen' en 'gewenste nauwkeurigheid van de georeference data', omdat de andere topics uitgaan van weinig tot geen ervaring met dergelijke data. In de bijlage 3 is alleen dat deel van de interviewtranscriptie te lezen welke relevant is voor deze deelvraag.

Deze gemarkeerde begrippen uit alle interviews heb ik in het schema hieronder bij elkaar gegroepeerd om de overeenkomsten en verschillen te kunnen analyseren. In bijlage 3 zijn alle transcripties en coderingen van de interviews te vinden.



Abeelding 5: schematische weergave van de codering van de interviews met gebruikers van georeference data.

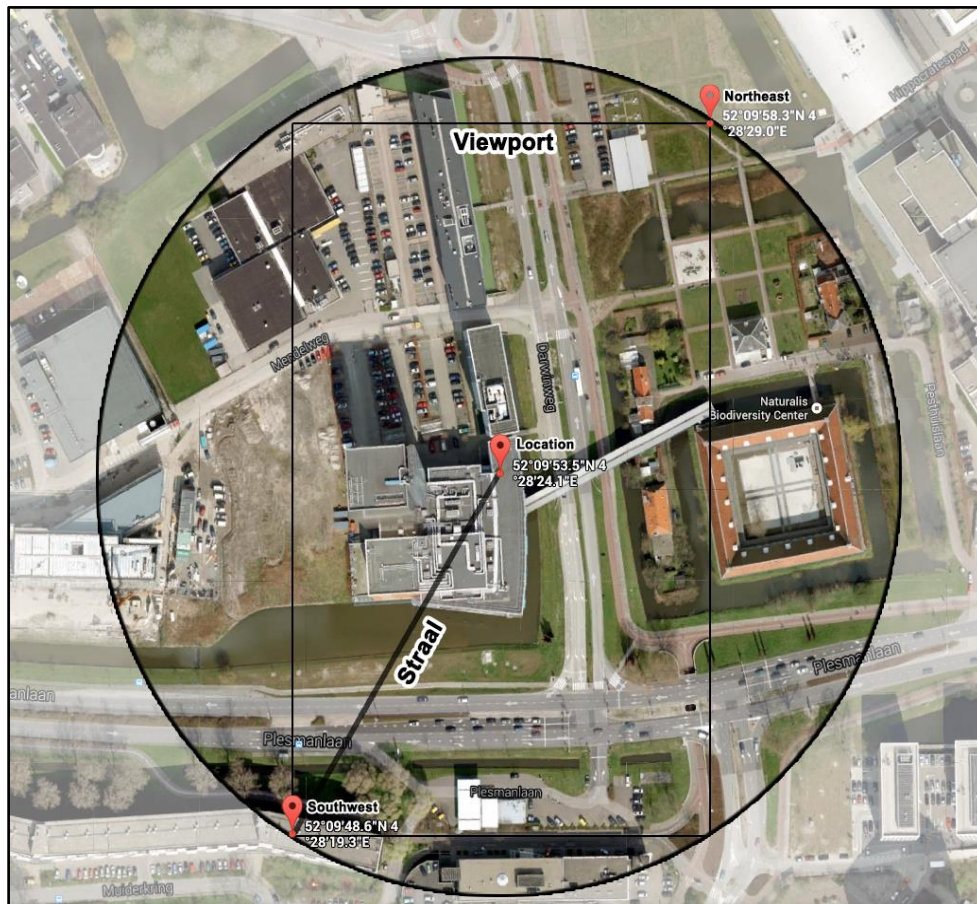
3.2 Resultaten en conclusie

Juiste geografische gegevens zijn essentieel voor het beheren van collecties voor onderzoek naar biodiversiteit. De taxonomische (bijvoorbeeld een soortnaam) en geografische data van objecten zijn de belangrijkste soorten data die nodig zijn voor een onderzoeker om de distributie van soorten te onderzoeken (Persoonlijke mededeling Luc Willemse 2015). De gegevens die verbatim (woordelijk) zijn geregistreerd in de collectie geven enkel richting, maar coördinaten zijn veel bruikbaar voor onderzoek. Het verrijken van een collectie brengt nieuwe mogelijkheden met zich mee voor alle toepassingsgebieden:

- **Collectiemanagement:** in de collectie kan op basis van geografische kenmerken worden gezocht waardoor collectiebeheer makkelijker en sneller onderzoeksvragen met een geografisch component kunnen beantwoorden door data te kunnen selecteren op basis van geografische kenmerken (Persoonlijke communicatie Luc Willemse, 2015).
- **Toegankelijkheid:** in de collectie kan op basis van geografische kenmerken worden gezocht waardoor deze ook toegankelijk wordt voor het publiek zonder taxonomische achtergrondkennis (Persoonlijke communicatie Ruud Altenburg, 2015).
- **Onderzoek:** het in kaart brengen van de verspreiding van soorten, het koppelen aan gegevens uit andere domeinen, zelf de collectie doorzoeken op basis van geografische data (Persoonlijke communicatie Niels Raes, 2015).

Om georeference data “fit for use” te maken voor de drie genoemde toepassingsgebieden moet er meer metadata worden vastgelegd dan alleen de coördinaten van een object. De nauwkeurigheid van de georeference is de meest waardevolle metadata die naast de coördinaten moet worden vastgelegd. Nauwkeurigheden en onnauwkeurigheden markeren de beperkingen van de toepassing van de gegevens. Deze beperkingen bepalen uiteindelijk de kwaliteit van de data en daarmee ook de kwaliteit van de resultaten en hoe een onderzoeker (of collectiebeheerder) deze zou moeten interpreteren (Chrisman, 1991). Door het vastleggen van de nauwkeurigheid van een coördinaat kan dus duidelijk worden gemaakt voor welke soorten gebruik de data “fit for use” is.

De nauwkeurigheid van een coördinaat is sterk gerelateerd aan de onzekerheid. Bij de onzekerheid wordt namelijk een polygoon of cirkel rond de locatie getrokken waarin het object zich moet bevinden. De afstand van het centrum van deze polygoon tot de meest uiterlijke hoeken ervan, is de nauwkeurigheid. Hoe kleiner de polygoon hoe beter de nauwkeurigheid. In afbeelding 3 is een visuele weergave getoond van de nauwkeurigheid van een object. Hierbij is een cirkel getrokken om de gevonden locatie. De straal van de cirkel (afstand van het centrum van de tot rand van de cirkel) kan worden gebruikt om de nauwkeurigheid aan te geven in een meetbare eenheid zoals meters.



Afbeelding 6: Visualisatie van de nauwkeurigheid van een georeference.

Door de nauwkeurigheid in meetbare eenheden uit te drukken voorkom je mogelijke schijnnaauwkeurigheid. De meest objectieve manier is door deze nauwkeurigheid aan te geven met metadata elementen zoals [uncertainty in meters] of [the maximum error distance], en waar mogelijk dit aan te vullen met de gegevens of methode waarop deze nauwkeurigheid is gebaseerd (Persoonlijke mededeling Niels Reas, 2015). Op basis van deze elementen kan een gebruiker zelf bepalen of de coördinaten bruikbaar genoeg zijn voor zijn of haar onderzoek.

Op basis van de interviews, transcripties en coderingen kan het volgende overzicht van wensen en eisen van de gebruikers van georeference data worden opgemaakt:

- Zoveel mogelijk objecten in batches georeferencen
- Tools die ook slechte datakwaliteit, zoals historische data en spelfouten aankunnen
- Objecten zo nauwkeurig mogelijk georeferencen, bij voorkeur minimaal 25x25 km
- nauwkeurigheid in een meetbare standaard aangeven

3.3 Evaluatie



De interviews die voor het beantwoorden van deze deelvraag zijn afgenomen hebben tot zeer bruikbare en heldere antwoorden geleid. Dit zeer concrete antwoord had ik niet van te voren verwacht. Aangezien dit onderwerp, met deze interviews, voor het eerste besproken werd in de organisatie, had ik verwacht meer een discussie te openen waarbij hooguit enkele knelpunten en/of vragen aan het licht werden gebracht. Daarbij werd dit antwoord door enkele interviewpartners ook nog aangevuld met voorstellen voor hoe dit wellicht zou kunnen worden uitgevoerd.

Mede dankzij de vorm van de interviews, zoals ik deze had opgezet, kregen de interviewpartners de ruimte hun voorkeuren en deze voorstellen voor oplossingen aan te kunnen dragen. Het was vooraf lastig vast te stellen of en hoe dit in de interviews naar voren zou komen. Als ik mij aan een (relatief) vaste vragenlijst had gehouden, was hier ook misschien niet de ruimte voor geweest. In een dergelijk geval was dit dus niet aan het licht gekomen. De keuze voor een topic list, in plaats van een half gestructureerd interview, is dus een goed besluit geweest.

Met behulp van de codering van de samengevatte interview transcripties kon een overzicht worden gemaakt van de eisen en wensen, problemen en de werkzaamheden die een rol spelen bij het georeferencen van een collectie als die van Naturalis. In het schema van deze codering werd direct duidelijk dat er over veel wensen en problemen een uniforme mening werd gedeeld. Dit maakt dat het antwoord op deze deelvraag heel concreet en snel geformuleerd kon worden, waardoor dit zeer bruikbaar was voor de rest van het onderzoek, in het bijzonder voor de pakketselectie.

De 'steekproef' die ik voor deze deelvraag heb genomen, heeft zowel voor- als nadelen. De gekozen interviewpartners waren wat mij betreft de meest geschikte kandidaten. Van elk toepassingsgebied zijn het namelijk medewerkers die dagelijks met het onderwerp te maken krijgen en het betreft steeds afdelingshoofden en of projectleiders, waardoor ze niet alleen voor zichzelf kunnen spreken maar ook een algemeen beeld voor hun hele afdeling konden geven. Met een langere doorlooperperiode was het misschien mogelijk geweest om meer gebruikers te interviewen. Dit was nu niet haalbaar, omdat de gebruikers zelf weinig tijd hadden voor interviews: veel werken niet volledig op kantoor of zijn zelfs in het buitenland voor onderzoek. Als ik de steekproef zou uitbreiden, zou ik dit wel doen met medewerkers uit dezelfde toepassingsgebieden en met vergelijkbare werkzaamheden. Dit zou wel de steekproef vergroten en daarmee ook de geldigheid van het onderzoek (Baarda, 2014).

Wat ook een zeer positieve uitkomst was van deze interviews was de zeer enthousiaste reactie van de gebruikers op het onderzoek en de kansen die hieruit kunnen voortvloeien. Het bleek dat veel geïnterviewde al gebruik maken van geografische informatie in de collectie van Naturalis, maar gezamenlijk aangeven dat coördinaten veel kunnen bijdragen aan de werkzaamheden en zelfs nieuwe kansen kunnen opleveren. Alles bij elkaar ben ik erg tevreden over mijn aanpak en het resultaat van dit deelonderzoek.

4 Georeference projecten

Deelvraag 3: welke andere instellingen hebben projecten uitgevoerd om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?

Tijdens de Semi-Automated Georeferencing Meeting op 2 en 3 december, 2014, is er gedetailleerde informatie uitgewisseld over de (technische) methodes die het NHM, het RMCA en Naturalis hebben ontwikkeld en getest voor het automatisch georeferencen van hun collectie. Deze vergadering onthulde dat de drie instellingen vergelijkbare en complementaire competenties hebben en gestandaardiseerde tools gebruiken die goed op elkaar aansluiten (NBC, 2014b). De verwachting ontstond hier dat er meer instellingen zouden kunnen zijn binnen de CETAF die georeferencing projecten hebben uitgevoerd of gestandaardiseerde tools en of methodes gebruiken die de besproken methode van de NHM, RMCA en Naturalis kunnen aanvullen.

Er ontstond een gemeenschappelijke behoefte onder de instellingen die aanwezig waren bij de meeting, om een onderzoek uit te voeren naar het georeference proces en de mogelijkheden om gestandaardiseerde tools in te zetten voor zoveel mogelijk stappen van het proces (NBC, 2014a). Om dit te onderzoeken heb ik internationale biodiversiteitsinstellingen geïnterviewd over de ervaring die ze hebben met georeferencing en de toegepaste tools/ methodes. Deze deelvraag kent 3 sub vragen:

- a) Wat is de betrouwbaarheid/ accuraatheid van de gebruikte methode/tools?
- b) In welke mate was handmatige correctie nodig om de betrouwbaarheid van de georeferencing data te verbeteren?
- c) Welke middelen zijn er gebruikt om deze verbeteringen door te voeren?

4.1 Onderzoeksmethoden

Het doel van de interviews die voor deze deelvraag zijn afgenomen, was een beschrijving te kunnen maken van de tool of de methode die de betreffende instellingen gebruiken en hoe deze is ingezet. Als basis hiervoor heb ik het beschreven proces uit het eindrapport van de FCD Pilot Georeferencing van Naturalis gebruikt. Op basis van deze bron heb ik een interviewscript gemaakt (zie bijlage 4). Ik wilde namelijk van elke instelling vergelijkbare informatie verzamelen, dat ook vergelijkbaar kon worden geanalyseerd en beschreven in het onderzoeksrapport. Het Naturalis project was al beschreven, en dus de maatstaaf voor de andere projecten. Ik ging er hierbij vanuit dat niet alle instellingen een vergelijkbaar eindrapport beschikbaar zouden hebben.

Het verzamelen van informatie van internationale instellingen heeft 3 stappen gekend om zoveel mogelijk verschillende projecten te kunnen onderzoeken:

Stap 1: Directe interviews

Er is in eerste instantie gericht op de 5 instellingen die reeds hadden aangegeven zelf georeference projecten te hebben uitgevoerd en te willen meewerken aan het project. Naturalis wordt hierbij ook meegenomen in het onderzoek, alleen niet geïnterviewd.

- Natural History Museum (NHM), London, Groot Brittannië
- Koninklijk Museum voor Midden-Afrika, Tervuren, België
- Botanische Tuin en Botanisch Museum, Berlijn, Dahlen, Duitsland
- Naturalis Biodiversity Center, Leiden, Nederland
- The Royal Botanic Garden, Edinburgh, Schotland

Deze interviews hadden de vorm van half-gestructureerde interviews (Verhoeven, 2014). Het was namelijk belangrijk dat de vorm van de interviews en de resultaten vergelijkbaar waren voor de gegevensanalyse, maar dat de respondent wel voldoende ruimte had om informatie aan te dragen. De interviews hebben allen plaatsgevonden via Skype en Google Hang outs, omdat alle instellingen zich in het buitenland bevonden, en persoonlijke gesprekken op locatie niet haalbaar waren in de doorlooperperiode van dit project. De precieze vragen verschillen licht per interview. Dit heeft te maken met de hoeveelheid informatie die ik had verzameld uit de verslagen van de *Automated Georeferencing Meeting*, mailcontact met de instelling en de hoeveelheid vooronderzoek dat ik had uitgevoerd.

Stap 2: Algemene oproep CETAF mailinglijst

Naast deze directe interviews heb ik een algemene mail verstuurd naar de hele CETAF-ISTC mailinglijst met de oproep of er nog meer instellingen waren die dergelijke projecten hadden uitgevoerd en hier informatie over wilde delen. Hiervoor had ik een gestandaardiseerde vragenlijst, gestructureerd, opgesteld die via de mail kon worden verzonden en ingevuld. Dit moesten gestructureerde interviews worden, omdat dit zou zorgen voor geloofwaardige en betrouwbare resultaten (Verhoeven, 2014). In de vragenlijsten werden grotendeels open vragen gesteld om de instellingen de ruimte te geven hun ervaringen te delen.

Stap 3: Afsluiting Hacking Heritage Talks, Erfgoed en Locatie

Aanvullend op deze twee stappen ben ik tijdens de doorlooperperiode van dit onderzoek op een afsluitende bijeenkomst van de 'Hacking Heritage Talks', op 17 september 2015 geweest voor het project Erfgoed en Locatie in Amsterdam (Waag Society, 2015). Hier werden verschillende tools gepresenteerd die tijdens dit project zijn ontwikkeld en alle gericht zijn op erfgoed collecties met locatie gegevens. Naturalis heeft eerder meegewerkt aan dit project. Op basis hiervan vond ik 2 middelen voor datacleaning die voor dit onderzoek zeer interessant waren:

- **Histogramh:** <http://histograph.io/viewer/>
- **Plaatsnamen Standaardiseren:** <http://standaardiseren.erfgeo.nl/>

Beide zijn specifiek gericht op Nederland, maar de achterliggende techniek is open source beschikbaar en kan zeer interessant zijn voor andere internationale doeleinden. Van deze bijeenkomst is geen verslag of transcriptie gemaakt. Ik ben hier zonder enige verwachting naar toe gegaan als oriëntatie en kennismaking met Nederlandse projecten, waardoor ik geen aantekeningen of iets dergelijks heb gemaakt. Wel zijn beide tools gebruikt voor de volgende deelvraag.

Gegevensanalyse

Alle afgenomen interviews zijn uitgewerkt tot een samenvattende transcriptie. Ondanks dat het bij deze deelvraag wel mogelijk was geluidsopnames te maken van de gesprekken, zou een letterlijke transcriptie geen bruikbaar verslagen opleveren. Bijna alle gesprekspartners lieten tijdens het interview via opties zoals 'Screenshare', van Google Hang Outs beelden zien van de ontwikkelde of gebruikte tools en de resultaten. Deze kon ik later niet bijvoegen in de transcripties, waardoor eventuele letterlijke verslagen niet voldoende context boden zonder deze screenshots (Baarda, 2014). In de samenvattende transcripties heb ik korte beschrijvingen toegevoegd van de beelden die met mij werden gedeeld tijdens de interviews.

De resultaten heb ik geanalyseerd door de transcripties met een vergelijkbare methode als bij de vorige deelvraag te coderen. Ik heb wederom niet de volledige tekst gecodeerd maar alleen waardevolle begrippen en zinnen gemarkeerd. Voor deze deelvraag heb ik de volgende labels gebruikt:

- **Tool:** om locatiebeschrijvingen om te zetten in coördinaten
- **Methode:** beschrijvingen van hoe om te gaan met verschillende tekstuele locatiebeschrijvingen
- **Databron:** datasets met wereldwijde geografische gegevens
- **Datacleaning:** middelen om het percentage georeference resultaten te verhogen
- **Nauwkeurigheid:** van georeference resultaten
- **Metadata:** aanvullende metadata elementen
- **Beschikbaarheid:** van de gebruikte bron, methode, tool, etc.

Ik heb na de codering verder geen analyse gedaan van de resultaten van de interviews, dit was namelijk onderdeel van de volgende deelvraag. De eerste vier labels: tool, methode, databron en datacleaning, waren het meeste belangrijk voor de beantwoording van de deelvraag. De nauwkeurigheid en de metadata betroffen specifieke kenmerken van de individuele tools, methodes, bronnen en datacleaning middelen. De teksten en begrippen die dit label toegekend hebben gekregen zijn meegenomen voor de volgende deelvraag, waarin de specificaties van de tools, methodes, bronnen en datacleaning middelen centraal staan.

4.2 Resultaten en conclusie



Uit de afgenomen interviews, met 6 internationale instellingen, konden 8 projecten worden vastgesteld, die gericht zijn op het georeferencen van biodiversiteits materiaal:

1. FCD-Pilot Georeferencing, Productiematig Georeferencen: Naturalis
2. SPECimap Georeferencing Software: Royal Botanical Garden of Edinburgh
3. StanDAP-Herb: Botanische Tuin en Botanisch Museum, Berlijn
4. SYNTHESYS NA-D 3.7 "Itinerary" project: RMCA, Tervuren
5. Georeferencing with Google Geocoding API and R: Niels Raes (Naturalis)
6. Data validatie check GBIF: GBIF
7. iCollections, the British and Irish Lepidoptera Project: NHM, London
8. MITCH: Universiteit Tilburg, Naturalis en Universiteit van Amsterdam

Georeferencing is een onderwerp dat veel is onderzocht en getest in de CETAF, omdat er geen kant-en-klare methode beschikbaar is. Elk van deze instellingen gebruikt een andere methodiek voor het georeferencen van collecties. Alleen Naturalis en het NHM gebruiken dezelfde tool. Al werd deze wel op een andere manier gebruikt. Ondanks de grote hoeveelheid tools en onderzoek binnen deze groep, zijn er weinig tot

geen tools en methodes beschikbaar die direct toegepast kunnen worden zonder bugs, problemen of veel technische achtergrondkennis. Veel instellingen hebben namelijk eigen tools of methodes ontwikkeld die alleen binnen de instelling en het project worden gebruikt.

Een voorbeeld hiervan is de Royal Botanical Gardens of Edinburg, die een tool maakte precies passend bij hun collectie, datakwaliteit en wensen. Het resultaat was een tool die objecten kan georeferencen met een nauwkeurigheidsafstand van 5 km, en ook bruikbaar is voor historische collectiedata.

Er zijn slechts enkele instellingen zoals Naturalis, het Natural History Museum in Londen en het Berlin-Dahlem Botanical Garden and Botanical Museum, die online beschikbare, en in sommige gevallen open source, tools gebruiken, zoals de *Google Geocoding API* en *GeoLocate*. Deze tools waren online beschikbaar en grotendeels kosteloos te gebruiken. Ook hierbij geldt echter dat niet alle toegepaste methodes momenteel nog bruikbaar zijn:

Een voorbeeld hiervan is de BioGeomancer tool. Deze werd regelmatig genoemd, als een zeer geschikte en veel gebruikte tool specifiek voor biodiversiteitscollecties. Helaas is deze tool niet meer beschikbaar, sinds het project in 2006 is gestopt en de tool niet meer wordt ondersteund.

De algemene conclusie van deze deelvraag is dat het wel zeker gewenst, maar tegelijk heel ingewikkeld is om objecten met verschillende soorten locatiebeschrijvingen op een automatische wijze te georeferencen. Waarschijnlijk zal zelfs met behulp van één of meerdere tools, een deel van de objecten handmatig moeten worden georeferenced.

4.3 Evaluatie

De resultaten van dit deelonderzoek gaven aan dat er veel gebeurt op het gebied van georeferencing en dat er veel interesse is in verdere ontwikkeling. Het is ook duidelijk dat er sterk een behoefte is naar meer informatie en samenwerking.

Dit laatste kwam dan ook duidelijk naar voren tijdens de interviews. Ondanks dat alle instellingen zelf projecten en onderzoeken hebben uitgevoerd, was hun initiële insteek meer gericht op vragen stellen, knelpunten delen en informatie over mijn onderzoeksresultaten verzamelen. Mijn insteek was juist andersom. Ik wilde achterhalen wat hun ervaringen waren en welke ontdekkingen zij hadden gedaan tijdens hun projecten.

Hierdoor liepen de gesprekken anders dan gepland. In de meeste gevallen ging er ongeveer een kwartier tot een half uur verloren aan het proberen duidelijk te maken wat mijn doel was met deze interviews. Hierdoor bleef er minder tijd over voor mijn vragen. Ik denk dat de reden hiervoor was dat ik in mijn eerste mail noemde dat dit voortkwam uit de Automated Georeferencing Meeting uit 2013. De insteek van deze meeting was informatie uitwisselen en ook plannen maken om op dit gebied internationale samenwerkingen en standaarden te ontwikkelen.

Daarnaast speelde de taalbarrière en het medium waarmee de interviews zijn afgenomen een rol. De meeste interviewpartners zijn van origine niet Engelstalig (ik ook niet), waardoor zeker de vakinhoudelijke

termen soms moeilijk waren uit te leggen. Tegelijkertijd verliepen de interviews vaak zonder beeld van de interviewpartner, omdat deze werden vervangen door beelden van de tools. Een interview zonder de interviewpartner te kunnen zien, blijkt moeilijker dan ik had gedacht. Ik merkte hierbij hoe belangrijk lichaamstaal, houding en gezichtsuitdrukkingen zijn tijdens een gesprek.

Nadat ik mijn insteek duidelijk had gemaakt aan mijn gesprekspartners, schakelde de meeste direct om en waren zeer bereid informatie te delen. De meeste lieten mij via het gebruikte medium beelden van de tool en resultaten zien, en deelden na afloop relevante literatuur met mij. Deze bronnen liepen uiteen van wetenschappelijk artikelen, PowerPoint presentaties, handleidingen en interne documentatie van de functionaliteiten van de tools en methodes.

Uit de interviews bleek dat veel instellingen zelf een tool hadden ontwikkeld voor het georefereren van collecties. Ik was hierover zeer verast en tegelijk teleurgesteld. Veel van deze zelf ontwikkelde tools zijn niet publiekelijk beschikbaar. Ik kon deze dus wel beschrijven in het onderzoeksrapport, maar niet adviseren in de best practice, omdat ze niet beschikbaar zijn voor alle gebruikers. De tools en methodes vanuit de interviews die niet beschikbaar zijn, heb ik toch meegenomen naar de volgende vraag, omdat de technieken hierachter wel bruikbaar en soms zelf open source beschikbaar zijn.

Deze deelvraag resulteerde echter in nog een teleurstelling. Er kwam namelijk geen respons op de algemene oproep die ik had verstuurd naar de CETAF-ISTC mailinglijst. Dit komt omdat ik, als afzender buiten deze lijst, eerst goedkeurig moet krijgen van een 'moderator, voordat de oproep word verspreid. Enkele weken na het versturen van de oproep, kwam ik erachter dat dit niet was gebeurd. In overleg met mijn stagebegeleider heb ik besloten geen verdere acties te ondernemen om deze instellingen te bereiken. Ik had het immers al druk met de interviews van de instellingen die individueel waren benaderd en de hele mailinglijst individueel benaderen, zou teveel tijd kosten. Daarnaast zou de kans dat hier ook binnen de geplande tijd respons op zou komen was te klein.

5 Georeference tools, methodes en databronnen

Deelvraag 4: welke semi automatische methodes/ tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten?

Deelvraag 5: Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren?

Deelvragen 4 en 5 volgden direct op de transcriptie en codering van de interviews die voor de vorige deelvraag (3) zijn gemaakt. Deze laatste stap van het deel 'dataverzameling' bestond uit deskresearch waarmee de verzamelde informatie van de interviews werd aangevuld met verwijzingen naar, en een analyse van relevante informatie over de besproken methodes en tools. Hierbij heb ik mij gericht op informatie over de functionaliteiten, bruikbaarheid, handleidingen en beschikbaarheid.

De interviews uit de vorige stap leverde 6 tools en 7 datacleaning middelen op. Als aanvulling hierop heb ik ook methodes/ best practice en databronnen meegenomen in dit deskresearch onderdeel. Zo kon er een zo compleet mogelijk beeld gegeven worden van het mogelijke aanbod voor productiematig georeferencen. In deze stap zijn alle mogelijke tools/ methodes meegenomen, ongeacht beschikbaarheid en toegankelijkheid. De selectie van best bruikbare tools vind pas in de volgende stap plaats (hoofdstuk 7).

5.1 Onderzoeksmethoden

In eerste instantie zou ik mij tijdens de deskresearch voor deze twee deelvragen richten op de onderzoeksrapporten, pilotverslagen en eindevaluaties die tools en datacleaning middelen beschreven. Ik was er hierbij vanuit gegaan dat in ieder geval een deel van mijn gesprekspartners dergelijke verslagen intern beschikbaar had en kon delen. In zekere mate was dit ook zo. Hieronder is een overzicht te zien van de documenten die ik naar aanleiding van de interviews ontvangen heb:

- Arkel, B. van & Creuwels, J. & Leusen, J. van & Schnörr, S. (2014). *Eindrapport FCD- Pilot Georeferencing: productiematig georeferencen*. Intern document Naturalis, Leiden
- Hine, A. & Penn, M. (2014). *Georeferencing at the NHM*. Presentation. Naturalis Biodiversity center, Leiden
- Schnörr, S. (2014). *Georeferencing Methode Niels Raes*. Summery. Intern document, Geraadpleegd op 7 november 2015
- Chapman, A.D. & Wieczorek, J. (2006). *Guide to best practices for georeferencing*. Copenhagen: Global Biodiversity Information Facility

Voor de projecten waarvan de bovenstaande bronnen niet voldoende informatie boden (of niet een dergelijke bron voor beschikbaar was), is er doormiddel van deskresearch gezocht naar aanvullende informatie. In het algemeen heb ik mij gericht op vakliteratuur, in plaats van wetenschappelijke bronnen, omdat ik op zoek was naar praktische informatie zoals handleidingen (Verhoeven, 2014). Daarom heb ik ook voornamelijk gebruik gemaakt van Google als zoekmachine en de websites van de geïnterviewde instellingen en overkoepelende organisaties als secundaire bronnen.

Ik heb voor deze deelvragen literatuuronderzoek gedaan naar alle georeference middelen in het figuur hieronder. Deze zijn allen afgeleid uit de interviews met CETAF instellingen van de vorige deelvraag.

Tools	Datacleaning middelen
GeoLocate Georeferencing Software Google Geocoding API SPECimap BioGeomancer Georeferencing Calculator MITCH	Manually visual check with Google Maps Plaatsnamen Standaardiseren Open Refine GBIF Specieslink SYNTHESYS NA-D 3.7 'Itinerary' project Crowdsourcing
Databronnen	Methodes
Geonames The Getty Thesaurus of Geographic Names Google Maps The Fuzzy Gazetteer Histogram	MaNIS/HerpNET/ORNIS Georeferencing Guidelines Chapman's <i>Guide to best practices for georeferencing</i>

Tabel 2: overzicht van gevonden georeference middelen uit de interviews met CETAF instellingen.

Gezien de grote hoeveelheid tools, methodes, bronnen en datacleaning middelen heb ik in bijlage 5 een overzicht bijgevoegd van de gevonden bronnen en de gebruikte zoektermen die hiertoe hebben geleid. In dit hoofdstuk leg ik aan de hand van 3 voorbeelden mijn deskresearch methodes uit.

SPECimap

SPECimap, van de Royal Botanical Garden of Edinburg, is een project dat voortkomt uit een onderzoek en samenwerking met de School of Informatics van de University of Edinburgh. Het project leverde uiteindelijk een georeferencing tool op, die binnenkort ook open source beschikbaar word gemaakt.

Van mijn interview partner (Elspeth Haston) kreeg ik enkele interne naamloze documenten toegestuurd waarin delen van de tool beschreven staan. Deze waren weinig bruikbaar voor mijn onderzoek, omdat zonder context de codes niet te begrijpen waren. Wat wel bruikbaar was, was de naam van de student die namens de School of Informatics aan het project heeft gewerkt: Claire Llewellyn.

Met slechts de naam van de tool en de naam van de student kwam ik niet direct tot bruikbare informatie. De tool is namelijk nog niet publiekelijk beschikbaar. De zoektermen die ik hierna gebruikte zijn gebaseerd op de transcriptie van het interview met de betreffende instelling:

- SPECimap
- RBGE
- Llewellyn
- Botanical data
- Botanical specimen
- text analysis
- Geotagger
- Geoparser
- Georeferencing
- textmining

De individuele zoektermen betroffen voornamelijk verschillen technieken (synoniemen) die in de tool worden toegepast om te georeferencen. Dit zijn algemenere technieken, die in meerdere projecten worden

toegepast. Daarnaast zijn termen gebruikt die het onderzoeksdomein specificeren; zoals de naam van de instelling, de tool en projectleden. De tweede zoekronde bestond uit combinaties van termen uit de eerste kolom en de tweede kolom. Dit heet de blokkendoos methode (Veen & Westerkamp, 2010). Dit leverde de volgende resultaten op:

- Llewellyn, C. & Grover, C. & Oberlander, J. & Haston, E. (2012). *Enhancing the Curation of Botanical Data Using Text Analysis Tools*. TPDL 2012, LNCS 7489, pp. 480–485
- Grover, C., Givon, S., Tobin, R., Ball, J. (2008). *Named Entity Recognition for Digitised Historical Texts*. Proceedings of the Sixth International Conference on Language Resources and Evaluation
- Llewellyn, C.A. & Hasten, e. & Grover, C. (2011). *Georeferencing Botanical Data using Text Analysis Tools*. TDWG 2011 Conference, Geraadpleegd op 18 oktober 2015, via <http://www.researchgate.net/publication/282701713Georeferencingbotanicaldatausingtextanalysistools>

De eerste twee bronnen zijn wetenschappelijke artikelen die voortkomen uit het project van Claire Llewellyn, en beschrijven de tool in verschillende fase van het project. De laatste bron beschrijft de twee technieken: Geoparser en Geotagger, die in de tool worden gebruikt voor de tekst mining algoritmes. De derde bron is een PowerPoint presentatie van The Royal Botanical Garden of Edinburg, waarin de werking van de tool word uitgelegd aan de hand van beelden.

Hierin was duidelijk te zien hoe in de tool verschillende lagen kaarten, met verschillende historische achtergronden worden gebruikt, om objecten zeer nauwkeurig te georeferencen. Dergelijke beelden geven een heel concrete toelichting op de werking van de tool en waren dus heel bruikbaar.

MITCH

MITCH is een project dat mijn stagebegeleidster tijdens een van mijn voortgangsgesprekken noemde. Het betreft een onderzoek dat gedeeltelijk intern bij Naturalis is uitgevoerd in het kader van het overkoepelende MITCH project. MITCH, *mining for information in texts from the cultural heritage*, is onderdeel van het CATCH project. Binnen Naturalis is in samenwerking met Marieke van Erp (Universiteit Tilburg) enkele tekstanalyse methodes ontwikkeld voor het verrijken van de collectiedata van Naturalis. Een deel hiervan was voor georeference. Mijn stage begeleidster gaf aan zelf te hebben meegewerkt aan een artikelen dat is gepubliceerd over het onderwerp, maar wist niet of dit al digitaal beschikbaar was.

Mijn deskresearch begon bij de secundaire bron: www.den.nl, de website van Digitaal Erfgoed Nederland. Hier heb ik in mijn 3^e jaar stage gelopen en ik kende hier het project *CATCH* al van. Op deze website wordt een projectbank bijgehouden over projecten in het digitale cultureel erfgoed veld. Via de beschrijving in deze projectenbank kwam ik op de website van het *MITCH* project.

Op deze website vond ik meer achtergrond informatie over het project en de samenhang tussen alle betrokken partners en vond ik een link naar alle publicaties die over het *MITCH* project zijn verschenen. Hier vond ik echter niet het beoogde artikel. Hieronder zie je alle stappen door middels van links weergegeven:

- <http://www.den.nl/>
- <http://www.den.nl/project/166/Mining-for-information-in-texts-from-the-cultural-heritage>
- <http://ilk.uvt.nl/mitch/>
- <http://ilk.uvt.nl/mitch/publications.html>

Uiteindelijk heb ik het beoogde artikel gevonden door te zoeken met de zoekterm combinatie: Georeferencing, Naturalis en van Erp (in het auteur-veld van de geavanceerd zoeken functie van Google Scholar). Google Scholar past automatisch de Booleaanse Operator AND toe bij deze zoekactie, met meerdere termen worden gebruikt (Veen & Westerkamp, 2010). Dit leidde tot de volgende bron:

- Erp, M. van & Hensel, R. & Ceolin. Davide. Meij, M. van der(2014). Georeferencing Animal Specimen Datasets. *Transactions in GIS*, 12/2014: 19(4)

Dit wetenschappelijke artikel beschrijft het onderzoek naar tekstanalyse methodes, en een test met een dataset van Naturalis. Het is eind vorig jaar gepubliceerd in het vak tijdschrift '*Transactions in GIS*'.

MaNIS/HerpNET/ORNIS Georeferencing Guidelines

HerpNET is tegelijk met FishNET2, MaNIS en ORNIS opgericht om een internationaal netwerk voor biodiverse collectie materiaal te creëren. De vier individuele databases bestonden tot 2015, toen ze werden gefuseerd tot VertNET.org. Op het portal promoten de organisaties verschillende best practices voor het georeferencen van biodiversiteitscollecties en delen ze bruikbare databronnen, gazetteers en tools. Dit is in samenwerking met GeoLocate en iDigBio's georeferencing working group.

De *MaNIS/HerpNET/ORNIS Georeferencing Guidelines* kwamen aan het licht in de literatuurlijst van de eindrapportage van het FCD pilot project van Naturalis (sneeuwbalmethode), dat als basis bron is gebruikt in de oriëntatiefase van dit project (van Veen & Westerkamp, 2010). Ook werden de guidelines in verschillende interviews genoemd.

Mijn eerste zoekpoging betrof de termen '**HerpNET**' en '**Georeferencing**'. Deze zoekactie leidde naar de website van het *HerpNET portal*. Hier werd duidelijk dat het portal begin 2015 is gefuseerd met 3 andere portalen tot *VertNET*. Ik ging ervan uit dat de *HerpNET* website niet bestond en niet direct zou leiden tot de guidelines, dit bleek onterecht.

Zowel de *VertNET* website als de oude *HerpNET* website heeft een aparte pagina genaamd 'georeferencing' met overzichten van beschikbare tools, databronnen, methodes, etc. voor georeferencing:

- HerpNET: <http://www.HerpNET.org/Gazetteer/GeorefResources.htm>
- VertNET: <http://georeferencing.org/>

Op het oude portal van HerpNET stond de link naar de MaNIS/HerpNET/ORNIS Georeferencing Guidelines. Daarbij stond ook een link naar een samengevatte variant van de guidelines:

- Wieczorek, J. (2001) "*MaNIS/HerpNET/ORNIS Georeferencing Guidelines*". Geraadpleegd p[2 november 2015, via <http://manisnet.org/GeorefGuide.html>
- Spencer, C. & Yamamoto, K. & Fang, J. & Constable, H. & Koo, M. & Wieczorek, J. (2005) *Georeferencing for Dummies: An Elaboration of the MaNIS/HerpNET/ORNIS Guidelines*. Berkeley: University of California

Beide websites met overzichten van georeferencing tools, methodes, databronnen, etc. hebben in de rest van mijn deskresearch gefungeerd als secundaire bron om informatie te vinden over enkele andere tools. Deze hebben uiteindelijk geleid tot de volgende bronnen:

- Tulane University (2015). *GEOLocate - Software for Georeferencing Natural History Data*. Geraadpleegd op 2 november 2015, via <http://www.museum.tulane.edu/geolocate/>
- Wieczorek, J., D. Bloom. 2011. *Georeferencing Calculator Manual v2*. Geraadpleegd op 18 november 2015, via <http://manisnet.org/GeoreferencingCalculatorManualv2.html>
- Geonames (z.d.). *Geonames, data resources*. Geraadpleegd op 19 november 2015, via <http://www.geonames.org/data-sources.html>
- Christian Kohlschütter (2003). *The JRC Fuzzy Gazetteer*. Geraadpleegd op 18 november 2015, via <http://isodp.hof-university.de/fuzzyg/query/>

5.2 Resultaten en conclusie



Bij de onderzoeksmethode en gegevensanalyse staan van 3 middelen (2 tools en 1 methodiek) de toegepaste deskresearch methodes uitgelegd. Per tool, methode of databron heb ik een vergelijkbaar literatuuronderzoek uitgevoerd. In bijlage 3 is een overzicht te vinden van alle tools en methodes met de gebruikte zoektermen en de gevonden resultaten. Van elk georeference middel heb ik een beschrijving gemaakt waarin de werking en de mogelijke waarde voor het georeference proces word beschreven.

Van de meeste tools heb ik vooral de primaire websites gebruikt, evenals bijgevoegde handleidingen en enkele secundaire bronnen over de functionaliteiten. Van enkele uitzonderingen, zoals te lezen in de vorige paragraaf, zijn ook wetenschappelijke artikelen of vakpublicaties gevonden en gebruikt. Deze laatste waren voornamelijk te vinden voor de georeferencing tools en datacleaning middelen. Voor de databronnen was over het algemeen weinig te vinden. Deze heb ik vooral beoordeeld op basis van hoe vaak ze worden genoemd en gebruikt in de andere methodes en tools. Zo worden Geonames en de TGN gebruikt in Histogram en word Google Maps in veel tools als kaart basis gebruikt.

De onderzoeksmethode en gegevensanalyse heb ik vanuit 3 voorbeelden beschreven. De conclusie zal ik echter wel voor de gehele deelvraag toelichten. Hierin maak ik wel hetzelfde onderscheid als die ik heb gemaakt in het betreffende hoofdstuk (6) van het eindrapport voor de opdrachtgever: tools, methodes, databronnen en datacleaning middelen.

Conclusie

De **tools** verschillen onderling vooral in de hoeveelheid objecten die tegelijk kunnen worden verwerkt. Sommige zijn geschikt om te gebruiken voor het georeferencen van batches, maar er zijn er ook die alleen gericht zijn op individuele objecten. Alle tools genereren meer dan alleen coördinaten, maar per tool verschillen de aanvullende metadata. De meeste genereren wel genoeg data om de geadviseerde velden (zie deelvraag 2) te kunnen vullen.

De **methodes** zijn beschrijvingen van hoe je om moet gaan met verschillende soorten tekstuele locatiebeschrijvingen, en welke aspecten daarvan wel of niet gebruikt kunnen worden voor georeferencen. Er waren slechts twee methodes die uit de interviews voortkwamen. Dit zijn twee losstaande methodieken, van andere auteurs en herkomst, maar ze kunnen elkaar wel aanvullen. Het gebruik van deze methodes leidt tot consistente resultaten die herhaalbaar zijn. De methodes zijn bruikbaar en toepasbaar op georeferencing projecten met verschillende tools en verschillende soorten collecties.

De **databronnen** die zijn beschreven, zijn online beschikbare databases met geografische informatie die nuttig is voor georeferencing, zoals coördinaten en plaatsnamen. Op één van de databronnen na bevatten ze allen geografische informatie van over de hele wereld. Histogram is de enige die gericht is op een specifieke regio, in dit geval Nederland. De andere afwijking is de *Getty Thesaurus of Geographic Names*. Dit is een thesaurus en bevat dus hiërarchische, gelijkwaardige, en associatieve relaties.

De **datacleaning middelen** betreffen methodes en tools die kunnen worden gebruikt om het percentage georeference resultaten van een bepaald proces te verhogen. De beschreven methodes hebben zeer uiteenlopende doeleinden en kunnen op verschillende stappen worden ingezet in het georeferencing proces. Niet alle middelen gaan uit van het gebruik van tools en de meeste vergen nog veel manuele input.

- GBIF, Specieslink en SYNTHESYS zijn methodes om fouten en afwijkingen te detecteren.
- Visual check with Google Maps, Plaatsnamen Standaardiseren en Open Refine zijn softwarematige datacleaning middelen.
- Crowdsourcing is de afwijkende datacleaning middel, omdat deze gebruikt kan worden om resultaten te verbeteren en te valideren, maar op vele manieren kan worden uitgevoerd

Om een collectie objecten volledig te georeferencen en te ontsluiten moet een biodiversiteitsinstelling meer ondernemen dan alleen de tekstuele locatiebeschrijving in coördinaten omzetten. En ook deze stap is lang niet zo simpel als het klinkt. Zo moet de nauwkeurigheid worden berekend, de resultaten worden gecontroleerd, datacleaning worden uitgevoerd, of expert input worden gevraagd. Ook aanvullende metadata, voor het gebruik voor onderzoek en duurzame ontsluiting moet worden vastgelegd. Om dit allemaal te doen heeft een instelling meer dan één hulpmiddel nodig om een collectie te georeferencen.

5.3 Evaluatie

De deskresearch begon al gedeeltelijk tijdens de interviews, waarbij ik van enkele gesprekspartners namen (van personen en tools), interne artikelen en links naar websites kreeg toegestuurd. De informatie die ik hieruit heb gehaald, heeft mij tijdens de deskresearch veel sneller tot bruikbare resultaten kunnen helpen. Ik had namelijk al een startpunt, met zoektermen, namen van auteurs, een bron voor de sneeuwbalmethode en secundaire bronnen.

Aangezien ik vooral de focus wilde leggen op vakpublicaties en handleidingen, in plaats van wetenschappelijk literatuur, had ik verwacht dat er veel minder zoekmethoden en technieken mogelijk waren. Voornamelijk zoektechnieken, zoals Trunkeren, Booleaanse Operatoren ed, gebruik ik meer in wetenschappelijk databronnen, zoals Google Scholar, en Simultaan Zoeken in de HHS Bibliotheek.

Ik heb echter gedurende de deskresearch toch veel verschillende technieken kunnen toepassen. Doordat ik deskresearch uitvoerde naar in totaal 20 georeference middelen (tools, methodes, bronnen, etc.), kon ik voor elke deel een eigen passende zoekmethode toepassen. Zo heb ik uiteindelijk de sneeuwbalmethode, de Blokkendoosmethode, Booleaanse Operatoren, aanhalingstekens en aanvullende zoekleutels (auteur) kunnen gebruiken (Veen en Westerkamp, 2010).

De combinatie van verschillende zoekmethodes draagt over het algemeen bij aan de betrouwbaarheid van een onderzoek en de resultaten die hieruit voortkomen. Ik ben na afloop ook zeer tevreden over mijn deskresearch deel van dit project. Tegen mijn gebruikelijke werkwijze in, heb ik ook een nieuwe onderzoeksmethode toegepast (blokkendoosmethode) en heb ik meer en vaker gebruik gemaakt van extra zoekleutels, zoals auteurs, titels, jaartallen en operatoren. Zo past Google bijvoorbeeld de AND operator automatisch toe, en was dit dus de enige die ik gebruikte. De operatoren OR en NOT (bij Google een -) moet je zelf toevoegen en dit heb ik meerdere keren gedaan (Veen en Westerkamp, 2010). Ook aanhalingstekens voor exacte woordcombinaties heb ik vaker gebruik, dan bij vorige deskresearch.

De hoeveelheid resultaten die de deskresearch opleverde, is groter dan verwacht. In eerste instantie richtte ik mij op de primaire websites van de tools, bronnen en methodes, maar ik had niet verwacht dat ik ook gepubliceerde vakliteratuur en zelfs wetenschappelijke artikelen zou vinden. Over het algemeen heb ik van alle tools, methodes en bronnen uitgebreid de functionaliteiten, bruikbaarheid en beschikbaarheid kunnen beschrijven. Enkel voor de *Fuzzy Gazetteer*, *Specieslink* en *GBIF*, heb ik niet alle mogelijkheden volledig kunnen beschrijven.

Het doel van de deskresearch was uiteindelijk ook om voldoende informatie te vinden die vergelijkbaar genoeg was om de pakketselectie voor de volgende deelvraag op een systematische wijze uit te voeren. Dit doel vond ik na afloop van de deskresearch voldoende bereikt.

6 De pakketselectie

Deelvraag 6: welke (semi-) automatische programma's zijn het meest geschikt voor de beschreven collecties?

Deelvraag 7: hoe sluiten deze methodes en programma's het beste op elkaar aan?

In de pakketselectie worden alle resultaten die voor deelvraag 1 tot en met 5 zijn verzameld gecombineerd. Op basis hiervan wordt een selectie gemaakt van de meest bruikbare middelen voor het georeferencen van primaire biodiversiteitsdata. Om te kunnen beoordelen welke middelen het meest geschikt zijn en welke elkaar kunnen aanvullen, wordt er op systematische wijze een vergelijking gemaakt tussen alle beschreven tools, methodes, bronnen en datacleaning middelen. Deze worden beoordeeld op basis van de wensen en eisen van de gebruiker van de data, de benodigde metadata elementen en kwaliteitsnormen specifiek voor het vakgebied.

Het doel van deze pakketselectie is niet om 1 tool te selecteren die het beste werkt. De verwachting vanuit de opdrachtgever en de CETAF is dat er meerdere tools bruikbaar kunnen zijn en deze gecombineerd kunnen worden toegepast. Hiermee kunnen grote delen van het proces zo automatisch mogelijk worden uitgevoerd met verschillende tools.

6.1 Onderzoeksmethoden

Om te beginnen met de pakketselectie heb ik de opties die niet beschikbaar zijn niet meegenomen in de volgende stappen. Dit betreft de tools *BioGeomancer*, *MITCH* en het *SYNTHESES NA-D 3.7 Itinerary* en het datacleaning middel *crowdsourcing*. De *BioGeomancer* en *MITCH* zijn niet meer beschikbaar, omdat het project waar ze uit voortkwamen is afgerond en de tools niet meer worden ondersteund. Crowdsourcing kan verschillende vormen aannemen en gaat niet uit van 1 beschikbaar platform. Hierdoor is het niet mogelijk om in een tabel de functionaliteiten/ specificaties, te beschrijven. Dit laat de volgende middelen voor georeferencing over:

- | | |
|---------------------------------------|------------------------------------|
| - Geolocate Georeferencing Software | - GBIF data-portal |
| - The Google Geocoding API | - SPieceslink |
| - The Georeference Calculator | - the Visual Check met Google Maps |
| - Getty Thesaurus of Geographic Names | - Open Refine |
| - SPECimap. | - Plaatsnamen-Standaardiseren |

De *SPECimap* tool is op dit moment nog niet beschikbaar, maar de betreffende instelling die de tool heeft gemaakt, gaf aan dat zodra enkele bugs zijn opgelost de tool open source beschikbaar komt. Wat betreft de databronnen en de georeference methodieken zijn er geen mogelijkheden die afvallen vanwege beschikbaarheid. Deze worden dus allemaal meegenomen in de pakketselectie.

Deze overgebleven tools, methodes, databronnen en datacleaning middelen zijn beoordeeld en vergeleken op basis van de wensen en eisen van de gebruiker van de data, de benodigde metadata elementen en kwaliteitsnormen specifiek voor het vakgebied.

Eisen gebaseerd op de Gebruikersbehoefte:

- De mogelijkheid zoveel mogelijk objecten in batches te georeferencen
- Tools gebruiken die ook slechte datakwaliteit, zoals historische data en spelfouten aankunnen
- Objecten zo nauwkeurig mogelijk georeferencen, bij voorkeur minimaal 25x25 km

Eisen gebaseerd op de datadictionary:

- Aanvullende metadata over het georeference proces vastleggen
- De nauwkeurigheid van de coördinaten vastleggen in een objectief format
- De originele geografische locatiebeschrijvingen niet overschrijven.
- Gestandaardiseerde metadata velden gebruiken die aansluiten bij internationale standaarden

Kwaliteitsnormen specifiek voor het vakgebied:

In het discussiehoofdstuk van het eindrapport werd verwezen naar de *“Principles for best practice for georeferencing biological species data”* (Chapman, 2006). Dit zijn richtlijnen met betrekking tot de kwaliteit van data(sets), die een groot deel van de CETAF instellingen toepast bij de digitalisering van collecties. De principes zijn ook specifiek gericht op biodiversiteitscollecties. Hierdoor sluiten deze direct aan op de collecties waarop de internationale best practice zich uiteindelijk richt. In overleg met de opdrachtgever heb ik besloten deze principes ook toe te passen in dit onderzoek. In tabel 2 zijn de acht principes van Chapman en Wieczorek, inclusief een korte beschrijving, te lezen. De effectiviteit van de tools is niet meegenomen in de beoordeling, omdat ik geen testen heb uitgevoerd, of recente testresultaten beschikbaar heb die dit kunnen indiceren.

Principe	Toelichting principe
Nauwkeurigheid (Accuracy)	Hoe goed representeert de gegenereerde data de originele waarden. Aangeduid met nauwkeurigheid
Effectiviteit (Effectiveness)	Percentage records dat aan betrouwbare geogereference data kan worden gelinkt
Efficiëntie (Efficiency)	Inzet die nodig is om acceptabele resultaten te genereren.
Betrouwbaarheid (Reliability)	De reproduceerbaarheid waarmee de georeference actie kan worden gedaan voor dezelfde locatie.
Toegankelijkheid (Accessibility)	Vindbaarheid van de gegenereerde coördinaten en aanvullende metadata
Transparantie (Transparency)	Kwaliteit van de metadata en methodologie waarmee georeference resultaten zijn verkregen
Tijdelijkheid (Timeliness)	Frequentie waarmee gebruikte bronnen worden geüpdate, en/of snelheid waarmee de coördinaten beschikbaar zijn gemaakt.
Relevantie (Relevance)	Is de output ‘fit for use’.

Tabel 2: Principles of best practice for georeferencing biological species data (Chapman & Wieczorek, 2006).

Alle eisen, wensen en de principes van Chapman ed. heb ik in de meest linker kolommen geplaatst van een tabel. Vervolgens heb ik ernaast in elke kolom één van de tools geplaatst en alle velden hiervoor ingevuld. Ik heb voor de tools, methodes, databronnen en datacleaning middelen losse tabellen gebruikt. Hierdoor bleven de tabellen redelijk behapbaar en kon ik per tabel de passende wensen en eisen verwerken. Door deze selectie zo systematisch aan te pakken is makkelijk te zien wat de onderlinge verschillen zijn, welke tools het meest geschikt zijn en waar de eventuele onderlinge overlap zit.

6.2 Resultaten en conclusie

Voor dit hoofdstuk zal ik van de toolselectie een deel van de pakketselectietabel gebruiken als voorbeeld. Er staan 2 tools in beschreven en er is slechts een deel van de wensen en eisen getoond. Dit zijn er namelijk te veel om allemaal in dit hoofdstuk te noemen. In bijlage 6 zijn alle tabellen van de pakketselectie te vinden. Hierin staan alle tools, methodes, databronnen en datacleaning middelen en alle wensen en eisen van de gebruiker van de data, de benodigde metadata elementen en kwaliteitsnormen specifiek voor het vakgebied.

	Google Geocoding API	SPECimap
Accessibility		
Online	Yes	Yes, once released
Offline	No	Unknown
Free	2500 free requests a day	Yes
File formats		
Accepted import formats	HTTPS request via Open Refine.	Individual objects go directly from collection database
Output format	JSON or XML	Unknown
Standard compliant	No	No
Batches	Yes	No
Individual	Yes	Yes
Results		
Accuracy	Based on the outer southwest and northeast corners of the smallest found geographic unit (given in coordinates)	unknown
Datum	WGS 84	unknown
Precision	6 decimals	exact to the nearest second
Data elements	Address components, Formatted address (textual and coordinates), Viewport	latitude and longitude, country, ALTM, National Grid reference, region, habitat
Efficiency	Only requires input to import data in Open refine. Tool takes 1 hour per 1000 objects	Requires allot of manual input, objects are replaced individually and manually
Reliability	Yes	No, because objects are placed manually
Transparency	https://developers.google.com/maps/documentation/geocoding/intro	Will be released open source.
Data quality		
Accepted languages	54 languages	English, the rest is unknown
Historic names	Unknown	Yes, manually
Miss spellings	Built in ranking mechanism that finds well known and populated places	Yes, manually
Other elements	No	Yes, manually

Tabel 3: Een deel van de pakketselectietabel voor de tools.

Op basis van de vier pakketselectie tabellen kon ik voor de tools, methodes, databronnen en datacleaning middelen een conclusie formuleren. Deze beschrijft welke van alle opties het meest bruikbaar zijn voor het georefereren van primaire biodiversiteitsdata.

Tools

De *Google Geocoding API* is de enige beschikbare georeferencing tool die objectrecords in batches kan georeferencen, zonder dat dit teveel manuren vraagt. Als je terugblijkt op de behoeften van de gebruikers, blijkt deze Batch mogelijkheid zeer belangrijk. Met de hulp van *OpenRefine* hoeven gebruikers ook niet veel technische kennis te hebben om de tool te gebruiken. De nauwkeurigheid en mogelijkheid handmatig aanpassingen te doen is echter ook zeer belangrijk voor de datakwaliteit.

Wanneer je dus nauwkeurig wil werken, kunnen de *Georeferencing Calculator* en *SPECimap* worden ingezet als vervolg stap. Deze tools kunnen worden gebruikt voor objecten die niet de gewenste resultaten teruggaven met de *Google Geocoding API*, of een meer handmatige aanpak vereisen. Dit is onder de voorwaarde dat deze tools worden bijgewerkt en/of worden vrijgegeven. *SPECimap* krijgt in dit geval de voorkeur, omdat deze alle data van één object gecombineerd op één scherm laat zien. Hiermee wordt dus de volledige context van het object gebruikt bij georeferencing. De *Georeferencing calculator* kan worden gebruikt voor records met aanvullende aspecten in de originele beschrijving, zoals headings en offsets.

Methodes

Als het aankomt op de besproken georeferencing methoden, wordt het direct duidelijk dat er niet één beter is dan de andere. Beide methodes behandelen belangrijke onderdelen van het georeference proces, en geven vergelijkbare best practices. De *Guide to Best Practice van Chapman* richt zich hierbij op het hele proces en alle mogelijke bijkomstigheden zoals het registreren van GPS data in het veld. Instellingen kunnen dit als raamwerk gebruiken om een projectplan te starten. De *MannisNET/HerpNET Guidelines* zijn daarentegen gericht op specifieke onderdelen van een locatiebeschrijving en hoe hiermee het beste omgegaan kan worden. Deze kan goed worden gebruikt tijdens het georeference proces als ‘spiekbriefje’.

Databronnen

Naast *Histogram* en de *Fuzzy G* zijn alle bronnen bruikbaar voor georeferencing. Het is volledig afhankelijk van de kwaliteit van de oorspronkelijke data welke het beste past. De *TGN* en *Geonames* zijn alleen nuttig voor datasets met gestandaardiseerde plaatsnamen omdat ze termen met een afwijkende spelling niet herkennen. *Google Maps*, is nuttig voor collecties zonder gestandaardiseerde plaatsnamen. Vanwege de rangschikking algoritmes van Google herkent deze bron kleine afwijkingen.

Datacleaning middelen

De data datacleaning middelen zouden bij voorkeur het aantal benodigde manuren voor een georeference project moeten verminderen. Geen van deze mogelijkheden zal echter de datacleaning stappen volledig automatiseren. Er zal dus altijd een bepaalde hoeveelheid manuele input nodig zijn. Door verschillende opties te combineren en in te zetten op verschillende momenten in het proces zal toch de arbeidsintensiviteit afnemen en het aantal resultaten toenemen.

Open Refine en *Plaatsnamen- Standaardiseren* kunnen vooraf helpen het aantal zoekacties te verminderen en afwijkende tekens te verwijderen (zoals getallen en leestekens). De *manuele/ visuele check met Google* is het enige echte datacleaning middel en kan worden ingezet om resultaten van een eerste poging nog preciezer te georeferencen of te valideren. De *GBIF data validatie check* kan tot slot worden ingezet om eventuele fouten te detecteren. De laatste methode, crowdsourcing, kan voor al deze stappen worden ingezet. Hierbij verandert er niks aan de datacleaning stap maar kost het de instelling wel minder tijd.

6.3 Evaluatie

De pakketselectie is nog niet de afsluitende conclusie in het onderzoeksrapport voor de opdrachtgever, maar laat wel al duidelijk zien welke tools en methodes de meest geschikte zijn en hoe deze elkaar kunnen aanvullen. Dat dit antwoord zo duidelijk naar voren komt, komt mede door de tabellen die ik heb gemaakt voor de pakketselectie.

Door alle functionele specificaties over de beschikbaarheid, resultaten, formats en data kwaliteit van alle tools naast elkaar te plaatsen werden de onderlinge verschillen direct duidelijk. Dit maakt het inzichtelijk genoeg om er direct een 'conclusie' uit te kunnen halen.

Bij de twee vorige keren dat ik een pakketselectie heb gemaakt heb ik een vergelijkbare tabel gemaakt met alle specificaties tegenover de wensen en eisen van de gebruikers, volgens de MOSCOW methode (Mollema, 2013). Bij die pakketselecties heb ik middels plusjes en minnetjes en rode en groene kruisjes aangegeven of de betreffende tools goed of slecht scoorden op de wensen en eisen. Bij deze pakketselectie heb ik echter meer informatie in de tabellen toegevoegd. Ik begon zoals bij de vorige keren met plusjes en minnetjes, maar ontdekte al snel dat dit niet voldoende inzicht bood.

Voorbeeld:

Bij The Google Geocoding API, vulde ik bij het vakje 'free' een plusje in, want deze is gratis te gebruiken. Dit geldt echter maar tot 2500 objecten per dag. Voor instellingen zoals Naturalis is dit vaak niet voldoende aangezien het om veel grotere aantallen objecten gaat.

Het gebruik van plusjes en minnetjes had hier wellicht tot een ander resultaat geleid. Sommige van de eisen en specificaties die in deze tabellen een minnetje zouden hebben gekregen, maken de tool niet direct onbruikbaar. Door duidelijke plusjes, minnetjes of de kleuren rood en groen te gebruiken suggereer je dit echter wel.

Voorbeeld:

Bij het veld 'standard compliance' heb ik aangegeven of de tools, ed. metadatavelden generen die overeenkomen met één van de standaarden uit de datadictionary. Voor de meeste was dit een minnetje geweest, omdat de velden niet precies matchen. De meeste tools genereren echter wel voldoende metadata om toch de benodigde elementen uit de standaarden te vullen. De benamingen van deze velden komen echter niet letterlijk overeen.

Ik denk dus dat mijn keuze om extra informatie in de tabellen toe te voegen, heeft bijgedragen aan de volledigheid van de pakketselectie. Zoals verwacht was de conclusie dat er niet maar 1 tool geschikt is voor het georeferencen van biodiverse collecties, maar dat er meerdere tools elkaar kunnen aanvullen. Ook dit laat de tabel goed zien. Door de extra informatie toe te voegen, komen de kwaliteiten van alle tools duidelijk naar voren.

7 De best practice

Deelvraag 8: hoe kunnen de methodes en programma's het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?

In principe vormde het antwoord op de vorige deelvraag van de pakketselectie al een conclusie. Het is duidelijk welke tools het meeste geschikt, wat de kwaliteiten van iedere tool zijn en hoe deze elkaar kunnen aanvullen. De vraag is juist: *“Hoe kunnen deze methodes het beste ingezet worden om primaire soortgegevens te georeferencen?”*. In deze laatste deelvraag worden alle antwoorden en conclusies op de deelvragen samengebracht om dit in één proces te plaatsen.

Het antwoord op deze vraag komt in de vorm van een best practice met een advies voor het toepassen van de methodes en tools die in de pakketselectie zijn uitgekozen. Deze best practice kan als stappenplan worden gebruikt door alle instellingen met een natuurhistorische of taxonomische collectie die een georeference project willen starten. Het bevat richtlijnen en bruikbare tools voor het hele proces, van begin tot eind. Daarnaast worden belangrijke collectie-specifieke, data- specifieke, gebruiker specifieke en instelling- specifieke aandachtspunten en benodigdheden benoemd.

7.1 Onderzoeksmethoden

Als stappenplan kent de best practice een instrumentele bruikbaarheid, waarmee het ingezet kan worden als raamwerk bij het plannen van georeferencing projecten en als basis voor een internationale standaard methode voor het (semi) automatisch georeferencen van natuurhistorische collecties. Er moest echter ook een conceptuele bruikbaarheid in naar voren komen (Verhoeven, 2014).

In overleg met de opdrachtgever, bleek dat de best practice zal worden gepresenteerd tijdens de volgende CETAF meeting in maart 2016. De opdrachtgever wil hiermee een discussie starten over het echte probleem dat in dit onderzoek naar voren is gekomen, namelijk: dat er geen gebrek aan tooling en kennis is, maar een gebrek aan samenwerking om deze tooling en kennis te delen, te onderhouden en te blijven gebruiken.

Uit de pakketselectie bleek namelijk dat er slechts 1 tool echt bruikbaar is voor georeferencing, maar dat er wel meer tools bestaan die zeer nuttig kunnen zijn. Deze zijn echter niet beschikbaar voor het hele publiek, en daardoor niet bruikbaar in de best practice. Andere onbruikbare tools, zijn na afloop van de projecttijd, waarvoor deze is ontworpen, stopgezet en niet langer onderhouden.

Voorbeeld: The Georeferencing Calculator does not function correctly, because it is based on a severely outdated java version (1.1). The most current java version is 8. It is the only tool that can handle additional aspects of geographic location descriptions like headings and offsets. Secondly, the tool complies with the Darwin Core standard.

Dit had natuurlijk ook effect op de indeling, detaillering ed. van de best practice. Om te zorgen dat deze precies aansluit bij de opdrachtgever, heb ik hierover enkele afspraken gemaakt in een gesprek:

- Het stukje 'visie' over het belang van samenwerking, moet duidelijk naar voren komen in de conclusie van het onderzoeksrapport.
- De mogelijkheden van de tools, die wel bruikbaar zijn, maar niet beschikbaar zou in een apart hoofdstuk van de best practice worden genoemd.
- De best practice zou heel praktisch beschrijven welke tools voor welke stappen nodig/ nuttig zijn
- In de bijlage konden gedetailleerdere stappen, zoals formules en bepaalde functionaliteiten van de tools worden beschreven
- Het stappenplan zou ook als schema of workflow worden gepresenteerd in de best practice.

Voor de vorm en indeling van de best practice heb ik verder gekeken naar enkele oudere best practices die ik tijdens de deskresearch was tegenkomen. Deze documenten betroffen vaak een net ander deelonderwerp van georeferencing, maar worden wel op een vergelijkbare manier en door vergelijkbare gebruikers toegepast.

- MaNIS/HerpNet/ORNIS Georeferencing Guidelines. (Wieczorek, 2001)
- Guide to best practices for georeferencing. (Chapman & Wieczorek, 2006)
- Guide to Best Practice for Generalizing Primary Species-Occurrence Data. (Chapman & Grofton, 2008)
- The point-radius method for georeferencing locality descriptions and calculating associated uncertainty (Wieczorek, Guo, and Hijmans, 2004)

Wat mij opviel in deze documenten, is dat de guidelines echt om praktische informatie gaan, die direct toegepast zou kunnen worden. Aanvullende informatie, zoals aandachtspunten, benodigdheden of voorbereidende stappen, worden vooraf of achteraf in een los hoofdstuk beschreven. Daarnaast bleek een zeer duidelijke structuur erg belangrijk te zijn. Er werd veel gewerkt met opsommingen, kaders, richtinggevend leestekens zoals $\rightarrow$, [] > en verschillende soorten tekst. Zo werden voorbeelden bijvoorbeeld in een andere kleur vormgegeven, namen dikgedrukt gemaakt, etc.

Om een hele overzichtelijke structuur aan te brengen heb ik de best practice in slechts 4 hoofdstukken verdeeld:

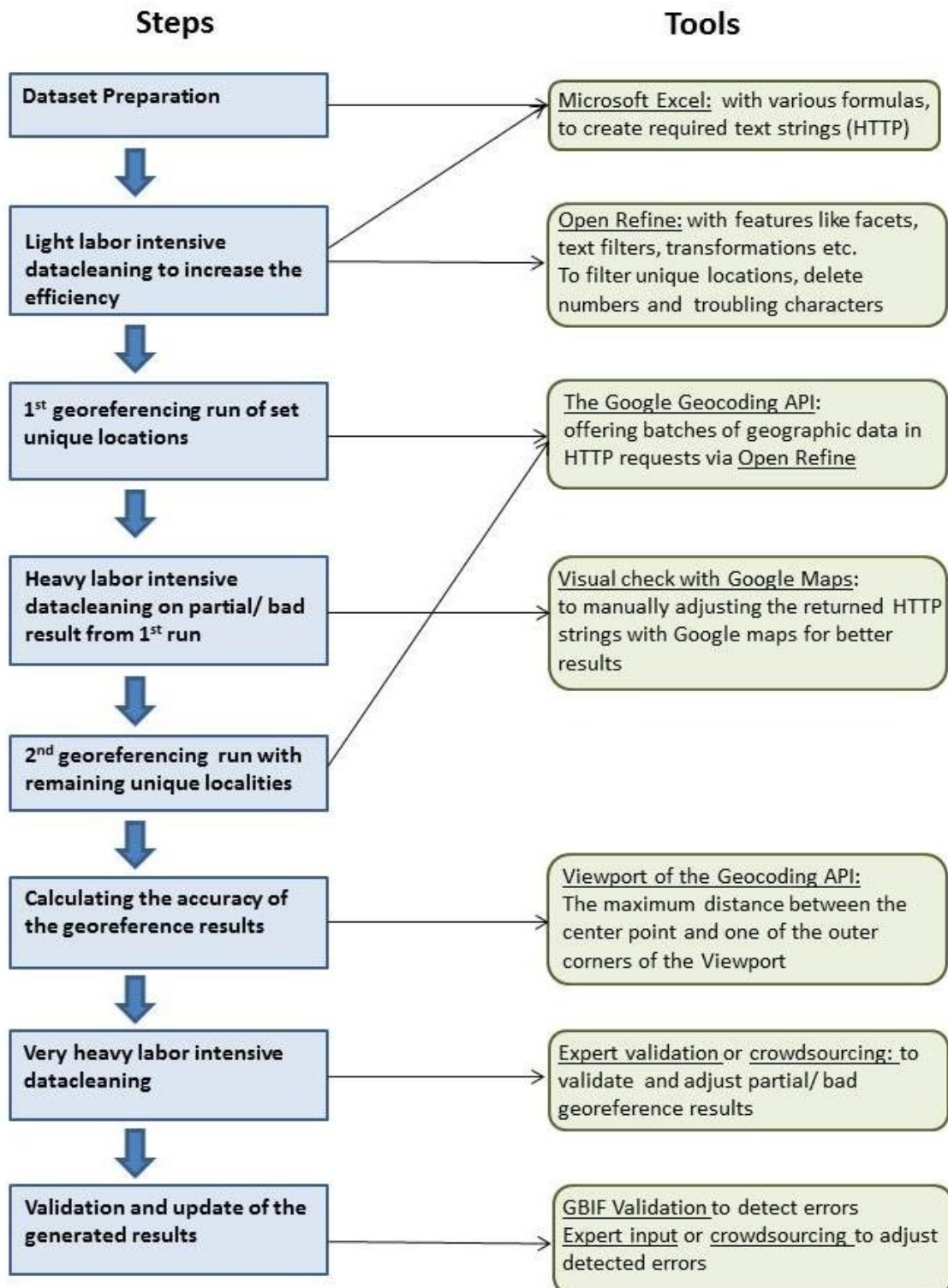
- **Pre georeferencing:** de stappen die nodig zijn om voordat georeferencing kan worden uitgevoerd. Hier komen aandachtspunten over de collectie, databases en gebruikers aan bod.
- **Georeferencing:** de daadwerkelijk stappen en tools hiervoor geschikt zijn
- **Post georeferencing:** stappen die te maken hebben met het valideren, controleren en documenteren van de verzamelde georeference data
- **Discussie:** een beschrijving van de tools, die wel bruikbaar zijn, maar niet beschikbaar zijn en hoe die kunnen worden ingezet, als ze beschikbaar worden gemaakt.

Voor de invulling van de hoofdstukken heb ik voornamelijk de conclusies en de pakketselectie uit het onderzoeksrapport gebruikt. Voor de zeer gedetailleerde stappen die in de bijlage beschreven moesten worden, heb ik gebruik gemaakt van de bijlage van het *Eindrapport FCD- Pilot Georeferencing* (Arkel et al, 2014), voor de precieze formules en stappen die nodig zijn voor het gebruiken van de tools.

7.2 Resultaten en conclusie

Naturalis gebruikte in hun Pilotproject Georeferencing dezelfde tool, die in de pakketselectie het meest geschikt bleek: de *Google Geocoding API*. Het proces dat in het advies wordt beschreven als de best practice voor georeferencing, zal daarom gedeeltelijk vergelijkbaar zijn met het proces dat tijdens dit pilotproject is ontwikkeld. De best practice gaat echter niet uit van het gebruik van 1 tool, maar zal worden aangevuld met andere middelen die de *Google Geocoding API* kunnen complementeren en het proces kunnen verrijken. Dit resulteert in het volgende proces voor het georeferencen van primaire soortgegevens:

Afbeelding 4: Schematische weergave (workflow) van het geadviseerde georeference proces beschreven in de Best Practice.



7.3 Evaluatie



De uiteindelijk meest geschikte tool om grote delen van een collectie te georeferencen in batches is de *Google Geocoding API*. Dit zeer duidelijke antwoord op deelvraag 6 is zowel positief als teleurstellend. Een zeer concreet antwoord maakt het uiteindelijk advies makkelijker en het zorgt ervoor dat de geadviseerde methode in de best practice voor iedereen bruikbaar is. De *Google Geocoding API* is namelijk relatief makkelijk in gebruik (met behulp van *Open Refine*) en online, gratis te gebruiken. Het sluit ook naadloos aan op de enige semi-automatische data cleaning methode: de *Visual Check met Google Maps*.

Tegelijkertijd is dit ook teleurstellend, aangezien dit precies dezelfde tool en methode is die Naturalis heeft beschreven en gebruikt in het FCD-pilot project. Dit betekent dat mijn advies niet geheel vernieuwend is voor de oorspronkelijke opdrachtgever van dit project. Ook de conclusie die voortkomt uit de pakketselectie voor de meest geschikte datacleaning middelen is lichtelijk teleurstellend. Het doel was namelijk om hier softwarematige oplossingen voor te vinden. Elke onderzochte mogelijkheid vergt echter nog een zekere mate van manuele input.

Dat er sinds het Pilot Project Georeferencing van Naturalis niet vele nieuwe mogelijkheden voor automatisch georeferencen zijn opgekomen is natuurlijk niet heel verassend. Het project is slechts een jaar geleden afgerond en de kans dat er in die tijd nieuwe tools worden gelanceerd is zeer klein.

Het mooie aan mijn best practice vind ik dat hij op verschillende manieren gebruikt kan worden. Als stappenplan om direct mee te werken, om eigen beleidsplannen op te baseren en om samenwerkingen aan te gaan. Daarnaast kan hij gebruikt worden voor de huidige situatie, maar hij is ook makkelijk aan te passen of uit te breiden met nieuwe tools, die wellicht in de toekomst beschikbaar komen. Tot slot is hij ook volledig aanpasbaar op basis van de kwaliteit van de originele locatiedata van een instelling.

Mijn onderzoek biedt een redelijk compleet overzicht van georeferencing projecten, tools, methodes, best practices, datacleaning middelen, standaarden, etc. Hier was vanuit de CETAF meeting veel vraag naar en een dergelijk overzicht is nieuw. De enige vergelijkbare best practice zoals "*Guide to best practices for georeferencing*" en de "*Point-radius method for georeferencing*" gaan juist niet uit van een bepaalde tool en spelen meer in op de methodiek van georeferencen. Mijn best practice sluit juist op deze internationaal geaccepteerde methode aan en is dus samen hiermee te gebruiken.

8 Eindconclusie en eindadvies

Meerdere instellingen van de CETAF hebben gewerkt aan projecten en onderzoeken om richtlijnen, instrumenten en procedures te vinden en te creëren voor het verbeteren van het georeferencing proces. Dit onderzoek probeert om een aantal van deze projecten samen te brengen en een overzicht van de beschikbare kennis en ervaring over dit onderwerp te maken om een best practice te creëren die internationaal toepasbaar en herkenbaar is.

In de onderstaande conclusie staan mijn primaire bevindingen beschreven die ik op basis van het onderzoek hebben kunnen formuleren. Aanvullend op deze conclusie heb ik mijn visie gegeven op de mogelijke oplossingen op de gedetecteerde problemen. Het eindadvies van dit onderzoek is de best practice. Mijn werkwijze is te vinden in het vorige hoofdstuk.

Conclusie



Van de relatief lange lijst van tools, methodieken, databronnen en datacleaning middelen, blijft slechts een kleine hoeveelheid nuttige mogelijkheden voor het georeferencen van primaire soortgegevens over, na de pakketselectie. Dit heeft vooral te maken met de beschikbaarheid van de onderzochte hulpmiddelen voor dit specifieke onderwerp. Georeferencing is een onderwerp dat relatief veel is onderzocht en getest, door verschillende instellingen en projecten. Echter veel van deze instelling en projecten beginnen bij het begin, zonder voort te bouwen op de aanwezige kennis, methodes en tools. Zoals in het onderzoek naar de projecten en tools bleek, ontwerpen en creëren veel instellingen een eigen georeference instrument, maar maken deze niet beschikbaar voor andere instellingen. In andere gevallen, zoals bij de *BioGeomancer*, is de tool wel beschikbaar voor andere geïnteresseerden, maar wijzigt dit nadat het initiële project word beëindigd. Op deze manier is er weinig beschikbaar open source, gratis en online voor alle geïnteresseerde partijen. Terwijl hier juist vraag naar is.

Op basis van deze interviews met instellingen die ervaring hebben met het onderwerp en een literatuurstudie naar aanvullende informatie, kunnen de volgende tools en aanvullende hulpmiddelen worden gezien als 'het meest bruikbaar/geschikt':

Stap	Tool
Voorbereiden dataset	➤ <u>Microsoft Excel</u> : om alle losse geografische metadata velden samen te voegen tot 1 veld.
Licht arbeidsintensieve datacleaning om de efficiëntie te verhogen	➤ <u>Open Refine</u> : met functionaliteiten zoals, tekstfilters, transformaties, etc. wordt het aantal unieke locaties gereduceerd en karakters die mogelijke problemen veroorzaken verwijderd of aangepast. ➤ <u>Microsoft Excel</u> : met formules wordt het aantal unieke locaties gereduceerd

	en karakters die mogelijke problemen veroorzaken verwijderd of aangepast. Voor Nederlandse collecties ➤ <u>Plaatsnamen- Standaardiseren</u>: om specifieke kenmerken van locatie beschrijvingen in een dataset te standaardiseren ➤ <u>Histogram</u>: speciaal voor het standaardiseren van historische Nederlandse locaties
1e georeferencing run met unieke locaties	➤ <u>The Google Geocoding API</u>: voor het georeferencen van HTTP strings via Open Refine
Zware arbeidsintensieve datacleaning	➤ <u>Visual check met Google Maps</u>: voor het handmatig aanpassen van de teruggeven HTTP strings met behulp van Google Maps. Voor deze stap kan crowdsourcing worden ingezet om de benodigde manuren terug te brengen.
2e georeferencing run overige met locaties	➤ <u>The Google Geocoding API</u>: voor het georeferencen van HTTP strings via Open Refine van de overgebleven records na de 1e run
Berekenen van nauwkeurigheid	➤ <u>Viewport of the Geocoding API</u>: door de afstand te berekenen van het centrale punt tot één van de meest uiterlijke hoeken van de Viewport in de resultaten van de Google Geocoding API
Zeer zware arbeidsintensieve datacleaning en resultaten controleren	➤ <u>Expert validatie</u>: experts binnen een instelling kunnen de resultaten valideren en records die bad/partial resultaten teruggaven handmatig georeferencen. ➤ <u>Crowdsourcing</u>: om voor hetzelfde proces de benodigde manuren terug te brengen
Update of validatie	➤ <u>GBIF Validatie</u>: als een dataset op het GBIF portal wordt gepubliceerd, voert de web service enkele checks en validatie over de georeference data uit. Mogelijke fouten worden gemarkeerd. ➤ <u>Crowdsourcing</u>: om gemarkeerde fouten aan te passen.

Tabel 4: samenvatting van de meest geschikte tools, methodes, databronnen en datacleaning middelen en het hoe deze in verhouding tot elkaar kunnen worden toegepast

Gebaseerd op de pakketselectie kan worden gezegd dat de *Google Geocoding API* de meest geschikte tool is voor het georeferencen van primaire soortgegevens. Het is momenteel de enige tool die objectrecords in batches kan georefereren met behulp van *Open Refine*. Het sluit daarbij ook naadloos aan bij de enige semi-automatische data cleaning methode: de *Visual Check met Google Maps* (Google Developers, 2015). Ondanks dat de uitkomst van de tool niet volledig aansluit bij internationale standaarden zoals Darwin Core of ABCD genereert de API wel voldoende data om de geadviseerde standaardelementen te vullen (Arkel et al, 2014). Daarmee sluit deze tool goed aan bij de wensen en eisen van de gebruikers.

In het proces dat met het schema in hoofdstuk 8 word afgebeeld staat enkel de *Google Geocoding API* genoemd. De *SPECimap* tool en de *Georeferencing Calculator* zijn echter ook noemenswaardig. Deze twee tools zijn veelbelovend voor het georeferencen van soortgegevens, maar zijn op dit moment niet meer of nog niet openlijk beschikbaar.

De instelling die de *SPECimap* tool heeft ontwikkeld, heeft aangegeven binnenkort een open versie te lanceren. Als deze tool als vervolgstap op de *Google Geocoding API* gebruikt kan worden, zou dit de perfecte combinatie voor georeferencen kunnen vormen. De combinatie zou kunnen leiden tot een proces waarbij de grote bulk objecten automatisch wordt georeferenced (Google) en vervolgens handmatig kan worden aangepast tot een nauwkeuriger resultaat (SPECimap).

De tools zouden echter ook kunnen worden gecombineerd in 1 tool, namelijk door een bulk optie toe te voegen aan SPECimap. Op deze manier zou slechts 1 tool nodig zijn voor het hele georeferencing proces, inclusief de tweede datacleaning stap (arbeidsintensieve datacleaning), en expert input (zeer zware datacleaning stap). Een instelling zou alleen nog *Microsoft Excel* of *Open Refine* nodig hebben voor de preparatie van datasets en collectie specifieke databronnen, zoals historische kaarten als aanvulling/input voor de kaartfunctie van de tool (Llewellyn et al, 2012).

De *Georeferencing Calculator* zou ook veel kunnen toevoegen aan het georeferencing proces. Dit is namelijk de enige tool die bij het 'plotten' van een locatie, ook de headings en offsets uit de originele beschrijving mee kan nemen in de berekening (Wieczorek and Bloom, 2011). De tool kan ook worden aangepast aan alle wensen van de gebruiker: de soort berekening, de coördinaat precisie, het type locatiebeschrijving, het coördinaat system, geodetische datum, etc. Daarnaast levert de tool resultaten die precies aansluiten bij de Darwin Core standaard (Wieczorek et al, 2015). Als de *Georeferencing Calculator* zou worden bijgewerkt naar de meest recente versie van Java (8), zou dit een veelbelovende optie zijn. Het zou wellicht minder tijd kosten om een reeds vrijgegeven en algemeen aanvaardde tool te actualiseren, dan om een nieuwe te creëren.

De meest belangrijke conclusie die ik op basis van alle onderzoeksresultaten kan trekken is dat de meest bruikbare tools voor georeferencing al bestaan. Er hoeven geen nieuwe tools te worden gemaakt, ze hoeven alleen maar te worden onderhouden en beheerd voor een lange periode van tijd en voor het gehele publiek.

Er is al genoeg kennis en tooling verzameld voor het georeferencen van primaire biodiversiteitsdata. Het grote probleem is echter dat de meeste instellingen alleen deze tools creëren en kennis produceren voor zichzelf en hun eigen situatie. Wanneer een project is afgerond, verdwijnt deze kennis en tooling. Een internationale inspanning kan bijdragen aan het behouden van deze kennis en tooling, zodat niet elke instelling opnieuw het wiel moet uitvinden voor zichzelf.

9 Discussie

In dit discussie deel van het onderzoeksproject worden de laatste paar onderwerpen besproken, die aan het licht zijn gekomen tijdens sommige delen van het onderzoek, zoals prioritering, beperkingen bij het publiceren van gegevens met geografische componenten, de waarheid niveaus van de resultaten die georeferencing tools kunnen genereren en 2 datacleaning middelen, die niet zijn meegerekend in het pakketselectieproces. Voor de besproken onderwerpen is geen uitvoerig gedaan. Ze zijn wel belangrijk om kort te noemen, voor de volledigheid van het onderzoek, en omdat ze invloed kunnen hebben op georeferencing.

Hieronder ga ik kort in op de discussieonderwerpen die bij dit onderzoek aan het licht zijn gekomen en waarom deze belangrijk zijn om te noemen.



Prioritering

Een aspect dat heel belangrijk is voor georeferenzen van een collectie voor onderzoeksdoeleinden is prioritering. Er zijn vele manieren waarop een georeferencing project kan worden geïnitieerd in een instelling en daarmee word vaak een selectie proces gestart. Een selectieproces voor welke collectie als eerste aan bod komt (Persoonlijke communicatie, Niels Reas, 2015). Dit kan worden gekozen op basis van:

- De kwaliteit van de originele geografische informatie
- Welke collecties de meeste voordelen halen uit georeference data
- Welke collecties het meest worden gebruikt voor geografisch onderzoek
- De mate van nauwkeurigheid die voldoende is voor de verschillende collecties
- De tijd die nodig is voor het georeferenzen van de collecties

Welke keuzes een instelling hierin maakt is sterk afhankelijk van de soort onderzoeksvragen die ze willen ondersteunen met de collectie en eventuele andere doeleinden waarvoor de georeference data wordt verzameld, zoals publicatie (Persoonlijke communicatie Luc Willemse, 2015).

Publiceren van georeference data

In dit onderzoek zijn voornamelijk de richtlijnen en problemen die komen kijken bij het documenteren en ontsluiten van georeference data besproken. Bij het publiceren van dergelijke informatie komen echter ook dergelijke richtlijnen kijken.

Dit geldt in het bijzonder voor het publiceren van gevoelige gegevens van bepaalde soorten, zoals gegevens van soorten op de rode lijst van IUCN, die leven in beschermde gebieden of met een hoge commerciële waarde (Chapman, Grafton, 2008). De Global Biodiversity Information Facility (GBIF) heeft een best practice opgesteld waarin de gevoeligheid van de exacte locaties van zeldzame, bedreigde soorten wordt besproken en richtlijnen worden gegeven voor hoe om te gaan met deze gegevens: *The Guide to Best Practice for Generalizing Sensitive Species Occurrence Data* (Chapman, Grafton, 2008). Dit kan echter betekenen dat instellingen georeference data moeten publiceren met een grotere nauwkeurigheid, dan dat ze in hun eigen interne collectie beheren, ter bescherming van de soorten (Persoonlijke communicatie Hans Muller, 2015).

Waarheidsgehaltes georeference resultaten

De *Google Geocoding API* geeft georeference data terug met een betrouwbaarheidsscore. Deze is gebaseerd op: of een unieke locatie is gevonden en of deze unieke locatie een volledige match is met de volledige HTTP aanvraag. Echter, de veronderstelling dat alle resultaten met een goede betrouwbaarheid score ook correct zijn, is twijfelachtig (Arkel et al, 2014). Dit gaat vooral op als er wordt gewerkt met collecties waarbij historische plaatsnamen voorkomen of landen ontbreken in de metadata.

Bijvoorbeeld; Een locatiebeschrijving bevat de historische naam van een Indonesisch eiland, zoals Alkmaar, maar Indonesië is niet als land opgenomen in de collectie. Aangezien [Alkmaar] de enige genoemde plaats in de originele data is, zullen de meeste tools coördinaten voor [Alkmaar, Nederland] teruggeven. Dit is namelijk de enige bekende [Alkmaar] in de meeste gazetteers. Het is echter niet de juiste.

Dit is te wijten aan het ingebouwde rangschikking algoritme dat de *Google Geocoding API* kent. Deze techniek richt zich op meer bekende en bevolkte plaatsen. Als [Alkmaar, Indonesië] wordt gebruikt in de HTTP-verzoek, zal Google wel het gewenste resultaat vinden. Dit 'probleem' vermindert het waarheidsniveau van resultaten met een hoge betrouwbaarheid score (Persoonlijke communicatie Luc Willemse, 2014). Wat betekent dat overwogen moet worden of het noodzakelijk is om ook deze resultaten te controleren op dit soort verkeerde interpretaties en te laten valideren door een deskundige.

Crowdsourcing

Veel stappen van het georeference proces kunnen worden uitgevoerd in een crowdsourcing omgeving. De crowd kan bijvoorbeeld unieke locaties filteren, datacleaning stappen uitvoeren of resultaten valideren. Dit zou een instelling veel tijd kunnen besparen. Deze optie gaat er natuurlijk wel vanuit dat de betreffende instelling beschikking heeft tot een crowdsourcing platform en een groep vrijwilligers (van Erp et al, 2014).

Het maken of onderhouden van een platform kost een instelling geld en tijd, maar lever vaak meer op dan alleen de gewenste data. Het leidt tot een grotere betrokkenheid van het publiek, kan nieuwe bezoekers brengen en zelfs publiciteit genereren met het project (Blaauboer, 2014).

Visual Possibilities

Naast de tools die worden beschreven in de conclusie, bestaat de mogelijkheid om een interface te ontwikkelen als een extra laag over deze middelen. De *Google Geocoding API* is namelijk best technisch in het gebruik, en daardoor niet voor iedereen toegankelijk (Arkel et al, 2015). Met de hulp van Open Refine vermindert dit al, maar met een Interface over de *Google Geocoding API*, kun je dit meer gebruikersvriendelijk maken. Voor het gebruik van een interface hebben gebruikers vaak minder technische ervaring of kennis nodig en maak je het toegankelijk voor meer gebruikers. Hiermee kan de groep gebruikers van de tool worden vergroot, en daarmee ook de werklast worden verdeeld.

Histograph en Plaatsnamen- Standaardiseren

Histograph en *Plaatsnamen- Standaardiseren* zijn tools die tijdens de pakketselectie afvielen omdat ze alleen bruikbaar zijn voor een zeer klein geografisch gebied, namelijk Nederland. De technieken achter deze tools zijn echter wel zeer interessant voor andere landen.

Histograph is een tool die kan worden gebruikt om historische geografische locaties te moderniseren en *Plaatsnamen- Standaardiseren* is een tool die sets met locatiegegevens kan standaardiseren (Erfgoed & Locatie, 2015). Door deze bronnen te gebruiken kan de efficiëntie van georeference tools worden vergroot.

Achter beide datacleaning middelen hangen dezelfde databronnen: *Geonames* en *The Getty Thesaurus of Geographic Names*. Deze bevatten geografische informatie van heel de wereld. Er zit alleen in beide tools een focus op de Nederlandse data (Erfgoed & Locatie, 2015). De achterliggende technieken zijn open source en kunnen worden gebruikt voor dezelfde doeleinden voor andere gebieden. Met extra (land) specifieke geografische gegevensbronnen zou het interessant zijn voor instellingen in andere landen.

10 Evaluatie van het project

In de hoofdstukken 3 tot en met 8 heb ik per deelvraag mijn werkwijze, onderzoeksmethode, resultaten en deelevaluatie gegeven. Als afsluiting hierop geef ik ook een eindevaluatie. Deze betreft het volledige onderzoek en het resultaat hiervan in verhouding tot de beginsituatie en het voorgenomen werkplan. Hierbij kijk ik kritisch terug op de dingen die wel goed zijn gegaan en de dingen die minder, of niet goed zijn gegaan.

Ik wil beginnen met een opsomming wat er wel goed ging en wat niet goed ging als een soort samenvatting en terugblik op de evaluatie van de deelvragen. Aanvullend hierop zal ik meer toelichting geven op enkele wat algemene/ overkoepelde aspecten van het project.

10.1 Wat ging er wel goed



- Datadictionary vormgegeven in een hiërarchisch overzicht dat aanpasbaar is voor iedere instelling
- Internationale standaarden gevonden die georeference data kunnen ontsluiten en overlappen
- Heldere antwoorden van gebruikers gekregen. Ook over het nauwkeurigheid probleem
- Enthousiasme van de gebruikers over het onderzoek en aandacht voor hun wensen/eisen
- Codering van de interviews creëerde zeer helder overzicht met wensen/eisen en problemen
- Van de interviews met instellingen veel aanvullende informatie gekregen, zoals links en namen
- Hierdoor relatief snel en makkelijk veel vakliteratuur, functionele informatie en wetenschappelijk bronnen gevonden bij deskresearch
- Veel verschillende onderzoeksmethodes toe kunnen passen
- Duidelijke en volledige pakketselectie tabellen kunnen maken, die alle deelvragen samenvatten
- Duidelijke conclusie gehaald uit de pakketselectie
- Datacleaning middelen en primaire georeferencing tool sluiten perfect op elkaar aan.
- Best practice kan op vele manieren worden ingezet: klaar voor gebruik maar ook klaar voor nader onderzoek of ontwikkeling
- Best practice proces in een helder workflow schema kunnen samenvatten

Gedurende de doorlooptijd van mijn stage heb ik geregeld gesprekken gehad met mijn stagebegeleider/opdrachtgever. Elke twee weken had ik een “bila” (bilateraal) overleg over de voortgang van mijn project. Ik kreeg in deze gesprekken volledig de vrijheid om zelf te bespreken wat ik nodig achtte: mijn voortgang, vragen, feedback, planning enzovoort. Deze vrijheid heb ik gedurende het hele project gemerkt.

Ik werd vanaf dag 1 goed opgenomen in de afdeling en meegenomen in de dagelijkse gang van zaken en, zoals de ‘stand-up’s’ en de afdeling- overleggen. Ik had echter wel volledig de vrijheid zelf mijn onderzoek structuur te geven, dagindelingen te maken, deadlines te plannen en de mijn producten in te delen.

Planning

De mogelijkheid zelf alles in te delen, te plannen en vorm te geven heb ik als zeer prettig en tegelijk uitdagend ervaren. Deze vrijheid gaf mij namelijk de kans zelf van begin tot eind te laten zien, waartoe ik in staat ben en initiatief te tonen. De uitdaging zat hem in het bewaken van de planning en oorspronkelijk werkplan uit het afstudeerplan en het Plan van Aanpak.

Het schema dat ik hiervoor maakte, zie pagina 9, heeft mij enorm geholpen bij het behouden van de structuur en het bewaken van de tijdsindeling. Van week tot week kon ik indelen welke activiteiten ik moest uitvoeren en wat er op een vrijdagmiddag afgerond moest zijn. Dit werkte voor mij niet alleen prettig maar paste ook goed bij de werkwijze op de afdeling, waar ze volgens de Scrum- methode werken, met korte sprints (Wikipedia, 2015). Het heldere schema hielp me ook bij het nemen van beslissingen, wanneer zich een probleem voordeed dat de planning in gevaar bracht. Door regelmatig terug te kijken op de planning kon ik bij problemen makkelijk inschatten of het effect zou hebben op andere activiteiten als ik door zou gaan met onderwerp en of dit de uitloop waard was.

Use Case

De beslissing om Naturalis als use case te bekijken heb ik als zeer waardevol beschouwd. Veel aspecten van mijn onderzoek, zoals de datadictionary, gebruikers wensen en eisen, datakwaliteit, e.d. zijn zeer instellingspecifiek. Naturalis gaf mij een echt kijkje onder de motorkap. Zo kon ik voor aspecten die zeer instellingspecifiek waren toch een beeld schetsen van wat de mogelijkheden waren en hoe instellingen hier in de praktijk mee omgaan.

In eerste instantie zou ik alleen de situatie van Naturalis analyseren voor enkele deelvragen. Ik vond dit zelf niet voldoende, omdat ook deze deelvragen uiteindelijk betrekking zouden hebben op het hele onderzoeksdomein, de CETAF. De resultaten heb ik daarom geformuleerd op een manier, waardoor ze toepasbaar werden op andere situaties door dit aan te vullen met informatie afkomstig uit de literatuur. Met deze combinatie kon ik het uiteindelijke advies ondersteunen met praktische en theoretische kennis.

Technische kant

Dit project betrof meer deelvragen die gericht waren op relatief technische informatie, zoals de datadictionary, functionaliteiten van de tools en het schrijven van de best practice. Dit is technischer dan ik in voorgaande projecten gewend ben. Ik ben zeer tevreden over hoe ik dit heb aangepakt en uitgevoerd. Ondanks dat ik minder technisch ben aangelegd heb ik een duidelijk en volledig overzicht kunnen geven van de eisen die georeference aan een databasestructuur stelt en wat de functionaliteiten van de tools zijn. Door mijn minder technische achtergrond heb ik de best practice op een manier kunnen beschrijven die zowel voor mensen met weinig technische kennis als mensen met veel technische kennis bruikbaar is.

10.2 Wat ging er niet goed

- Het plannen van afspraken voor de interviews met gebruikers en internationale instellingen
- Interviewen via digitale media, zoals Skype, zonder beeld van de gesprekspartner
- Minder grote steekproeven bij de interviews dan gewild
- Geen geluidsopnames bij de interviews met gebruikers
- Algemene oproep aan de ETAF mailinglijst niet doorgekomen
- Weinig tools gevonden die voor iedereen beschikbaar zijn
- Weinig nieuwe datacleaning middelen gevonden die met behulp van software werken

Interviews

Bij de interviews, zowel met gebruikers als met de instellingen, kwam de meeste tegenwerking op mijn pad. Dit had voornamelijk te maken met communicatie en tijd. Het kostte als eerste veel tijd om contact te krijgen met personen die ik wilde interviewen, terwijl mijn stagebegeleider al voorwerk had gedaan. Vervolgens bleken ze allemaal weinig tijd te hebben voor een interview, door drukke agenda's en omdat ze weinig op kantoor werken. In principe was dit geen probleem, en ik had hier al op geanticipeerd. Er waren echter enkele gesprekspartners, die na het plannen van een afspraak, toch afzegden, of niet kwamen opdagen. Dit had een negatief effect op mijn planning en ik voelde me hierdoor niet serieus genomen.

Ten tweede bleek dat de algemene oproep die ik had verstuurd naar de CETAF-ISTC mailinglijst niet te zijn aangekomen bij de personen in deze mailing lijst. Ik zat als afzender niet in de mailinglist (extern) waardoor er goedkeuring nodig was van een 'moderator'. Deze is echter nooit gekomen en ik kwam hier te laat achter om nog actie te ondernemen.

Door deze problemen zijn de steekproeven van beide interviewrondes niet zo groot als ik had gewild. De gekozen interviewpartners waren wat mij betreft de meest geschikte kandidaten. Met meer tijd had ik meer gebruikers met dezelfde soort functies en met dezelfde soort werkzaamheden willen interviewen, en dat zou wel de steekproef groter gemaakt hebben en daarmee ook de geldigheid van het onderzoek.

Van de geïnterviewde internationale instellingen hadden alle gesprekspartners ervaring met georeferencing en zij konden dus inhoudelijk bijdragen aan het onderzoek met informatie over tools, methodieken, datakwaliteit, ed. Van de instellingen die ik wilde bereiken met de algemene oproep aan de CETAF lijst, was het niet zeker of deze ook dergelijke ervaringen hadden met het onderwerp. Als deze oproep dus wel was doorgekomen, zou het niet per se ook bruikbare informatie hebben opgeleverd. Het had wel wederom de steekproef vergroot en daarmee ook de geldigheid van onderzoek. Daarnaast had ik dan ook de vragen en problemen van instellingen zonder ervaring kunnen waarnemen. Hiermee had ik de best practice wellicht anders ingedeeld en meer op deze groep gericht.

Tegenstrijdig aan al deze tegenslagen was de enthousiasme en behulpzaamheid van de gesprekspartners. Uiteindelijk heb ik toch 6 instellingen kunnen interviewen. Alle gesprekspartners waren zeer blij met het onderzoek dat ik uitvoerde en bereid te helpen met namens, documenten, links, screenshares, etc.

Pakketselectie en conclusie

In mijn deel evaluatie van de pakketselectie heb ik al genoemd, dat ik de uiteindelijke conclusie lichtelijk teleurstellend vond. Uit de pakketselectie kwam een zeer eenduidig antwoord. De best bruikbare tool, was de *Google Geocoding API*. Dit antwoord zou onder de gebruikers tot weinig discussie moeten leiden, omdat het de enige beschikbare tool is die aan de twee meest belangrijke eisen voldoet: beschikbaar en de mogelijkheid batches te georeferencen. Dit eenduidige antwoord is belangrijk voor een document dat als internationale standaard gebruikt gaat worden.

De *Google Geocoding API* is dezelfde tool die Naturalis in haar testpilot heeft gebruikt. In principe is dit geen probleem, en zelfs makkelijk aangezien dit betekent dat er al een werkproces klaar lag voor het gebruik van de tool. Het teleurstellende vond ik, de reden dat deze tool de meest bruikbare was. Uit de interviews bleek dat bijna alle instellingen (behalve Naturalis) zelf een tool of een deel van een tool hebben ontwikkeld voor het georeference van collectiedata. Deze zelf ontwikkelde tools zijn alleen niet publiekelijk beschikbaar. De *Google Geocoding API* was dus de enige tool die beschikbaar en toegankelijk is voor alle geïnteresseerde partijen en batches kan georeferencen. Terwijl er dus wel meer tools bestaan. Dit was wat mij betreft het grote probleem en de meest belangrijke conclusie. "Er bestaat zo veel, maar niets word onderhouden en gedeeld". Daarom heb ik in mijn conclusie ook mijn visie gedeeld en in de Best Practice een discussie hoofdstuk toegevoegd. Hierin wilde ik duidelijk maken dat er veel meer kan, als hier aandacht aan word besteed.

11 Verantwoording beroepstaken en persoonlijke reflectie

In het schema uit het Plan van Aanpak op bladzijde 9 staan in de laatste kolom voor elke projectactiviteit de bijbehorende opleiding specifieke beroepstaak genoemd. Voor dit afstudeerproject heb ik deze beroepstaken op minimaal niveau 4 uitgevoerd, zoals deze zijn vastgelegd in de *“Uitwerking van de beroepstaken IDM versie 1.1”* van de Curriculumcommissie IDM (2010). Bij de reflectie in dit hoofdstuk gaat het vooral om het methodische handelen voor elke beroepstaak. Hieronder staan alle 6 de beroepstaken opgesomd.

- Probleemanalyse (analyseren)
- Datadictionary (ontwerpen)
- Informatiebehoefte (analyseren)
- Deskresearch (beheren en exploiteren)
- Pakketselectie (ontwerpen)
- Parametriseren (realiseren)

11.1 Reflectie per beroepstaak



Per stuk zal ik beschrijven aan welke competenties ik tijdens het project heb gewerkt en reflecteer ik op basis hiervan op mijn handelen als aankomend beroepsbeoefenaar.

Probleemanalyse

De eerste fase van het project bestond uit het inventariseren van de achtergrond van het onderwerp binnen Naturalis en binnen de CETAF en de bijbehorende problemen. Voor deze inventarisatie en probleemanalyse heb ik gebruik gemaakt van enkele interne bronnen van het eerder uitgevoerde Pilotproject en de bijeenkomsten van de CETAF en gesprekken met betrokkenen. De informatie uit deze bronnen vormden het startpunt van mijn onderzoek.

Specifiek voor de situatie van Naturalis heb ik gebruik gemaakt van het onderzoeksrapport/eindverslag van het FCD-pilot Georeferencing (Arkel et al, 2015). Deze bron gaf zeer uitgebreide achtergrondinformatie over de vraag/het probleem van de opdrachtgever en daarmee ook een goed zicht op de reeds intern beschikbare kennis en informatie op de afdeling. Er bleek vooral veel behoefte te zijn aan een verband tussen de geteste methode en de benodigdheden van de collectie en collectiesystemen om het proces in te zetten.

Specifiek voor de situatie van de CETAF, heb ik gebruik gemaakt van de verslagen van een bijeenkomst die over het onderwerp in 2013 heeft plaatsgevonden: de *“Follow-up Actions Meeting Automated Georeferencing”* (NBC, 2014a) en de *“Technical Summery”* (NBC, 2014b). Op basis van deze twee bronnen kon ik een beschrijving maken van de algemene problemen en vragen die tijdens de meetings naar voren

zijn gekomen. Hier is vooral behoefte aan een inventarisatie van mogelijkheden en best practices voor semiautomatisch georeferencen en de belangrijkste beheersaspecten die hierbij komen kijken.

Om deze twee losse vragen samen te brengen heb ik besloten de situatie van Naturalis als use case te gebruiken bij enkele deelvragen. Om deze dan toepasbaar te maken voor het hele onderzoeksdomein besloot ik deze resultaten aan te vullen met ervaringen van andere instellingen en externe literatuur. Hiermee zou het advies inzetbaar worden voor vergelijkbare problemen en in relatie worden gebracht met de bedrijfscontext van een biodiversiteitsinstelling.

In gesprek met twee medewerkers van de opdrachtgevende organisatie die betrokken waren bij het Pilotproject heb ik de gedetecteerde problemen teruggekoppeld. Hiermee heb ik bevestigd dat het onderliggende probleem een gebrek aan gedeelde kennis en een standaard methode is.

De gesprekken heb ik zelf geïnitieerd, nadat ik opmerkte dat twee van de medewerkers, op de afdeling waar ik werkte, als auteurs stonden genoteerd op het eindrapport. De documenten, heb ik zelf in de archiefmap van het pilotproject gevonden en als meest bruikbaar ingeschat.

Door deze combinatie van persoonlijke ervaringen en een korte literatuurstudie heb ik een gedegen vooronderzoek uitgevoerd, op basis waarvan ik richting kon geven aan het onderzoek en de eindproducten (Verhoeven, 2014). Hiermee kon ik met een zeker gevoel en inzicht in de volledige aanleiding/ achtergrond, de eerste stappen van het project voortzetten. Mijn bevindingen heb ik beschreven en verantwoord in het eerste hoofdstuk van het onderzoeksrapport.

Datadictionary

Eén van de vragen die specifiek van Naturalis kwam tijdens de probleemanalyse, betrof de metadata die nodig is voor georeferencing. Hierbij heb ik mij gericht op de metadata elementen die worden gebruikt om georeference data te ontsluiten en op inhoudelijk geografische metadata.

Om een volledige datadictionary te maken heb ik de collectiemanagement systemen van Naturalis als use case gebruikt voor het analyseren van de benodigde entiteiten die een rol spelen bij het ontsluiten van georeference data. Vanuit de use case zijn ook de aanwezige soorten geografische metadata en de onderlinge verschillen hierin geanalyseerd. Vanuit deskresearch heb ik deze praktijkgerichte analyse aangevuld met theoretische informatie over beschikbare internationale standaarden geschikt voor het ontsluiten van geografische biodiversiteitsdata.

De geadviseerde metadata elementen zijn gebaseerd op een vergelijking tussen beide overzichten (de use case en de standaarden) en door de overeenkomende elementen te definiëren en te specificeren als primair en secundair.

Alle delen van de datadictionary heb ik beschreven in de vorm van een simpel hiërarchisch overzicht van velden en een beschrijving hiervan. Ik heb hier bewust gekozen om geen entiteitentabel met attribuutvelden te maken, omdat hiermee de hiërarchische structuur en onderliggende relaties van de elementen niet volledig tot zijn recht kwamen (Curriculumcommissie IDM, 2010). Door alleen de veldnamen en beschrijvingen te noemen in het onderzoeksplan, heb ik getracht rekening te houden met de individuele aanpasbaarheid van de geadviseerde elementen. Iedere instelling die hier gebruik van maakt, kan zelf bepalen in welke taal, lengte, tekstvorm, ed. ze deze velden willen invullen.

Deze activiteiten waren een belangrijke basis voor het vervolg van mijn onderzoek, omdat ik hiermee ervaring opdeed en kennis vergaarde met de collectie die een centrale rol speelde in het hele project. Dit

achtte ik ook belangrijk voor de communicatie met de gebruikers van de collectie voor de interviews. Mede daarom heb dit onderwerp zeer uitvoerig behandeld.

Informatiebehoefte

Fitness for use was een van de belangrijkste begrippen in mijn onderzoek (Chrisman, 1991). Het hele georeference proces draait erom de collectie dusdanig te verrijken dat deze nog bruikbaar wordt voor de gebruikers, zowel intern als extern. Het is hierbij essentieel ervoor te zorgen dat de verrijking ook aansluit bij de verwachting en behoefte van gebruikers (Chapman, 2006).

Om dit te garanderen, heb ik de gebruikers van de toekomstige georeference data binnen Naturalis geïnterviewd. Voor deze deelvraag is Naturalis opnieuw benaderd als een use case. De resultaten zijn ingedeeld in drie algemene toepassingsgebieden die profiteren van georeference data en in de meeste biodiversiteitsinstellingen belangrijk zijn; collectiebeheer, onderzoek en toegankelijkheid. Deze toepassingsgebieden zijn gekozen na een analyse van 3 interne bronnen die de doeleinden, plannen en mogelijkheden van de digitale collectie beschrijven.

Het doel van de interviews was om te bepalen hoe en op welke niveaus de huidige geografische gegevens moeten worden verrijkt, om de werkzaamheden van deze gebruikers te vergemakkelijken. Dit bestond dus uit het analyseren van de werkzaamheden van de gebruikers, de toepassing van geografische data hierin, en de invloed van toekomstige georeference data hierop.

De interviews zijn in ongestructureerde vorm afgenomen om de gebruikers volledig de ruimte te geven om te vertellen over welke kennis, informatiebronnen en instrumenten zij nodig achtte (Verhoeven, 2014). Later is deze ongestructureerde data gecodeerd om patronen te kunnen ontdekken. Hiermee kon een overzicht worden gemaakt van de eisen en wensen, problemen werkzaamheden die een rol spelen bij het georeferencen vanuit de drie toepassingsgebieden. Op basis van dit overzicht konden bij de pakketselectie tools/methodes worden geselecteerd die hieraan voldoen en de gewenste (soort) gegevens genereren.

Ondanks de uitgebreide voorbereiding voor dit onderdeel van het onderzoek, is het niet geheel gelopen zoals gepland. Dit kwam voornamelijk door externe factoren, zoals drukke schema's. In enkele gevallen was het voor mij mogelijk de betreffende personen persoonlijk aan te spreken (in de wandelgangen), maar omdat niet iedereen op kantoor werkt, kon dit niet in alle gevallen. Ik heb met verschillende pogingen aangetoond dat ik initiatief toon bij het plannen van afspraken, een professionele houding aanneem ondanks tegenslagen en gerichte beslissingen kan nemen over de kwaliteit en kwantiteit van mijn steekproef.

Deskresearch

Gedurende het onderzoek heb ik meerdere verschillende zoekmethoden toegepast om theoretische kennis te verzamelen, zoals bij de datadictionary, en de probleemanalyse. Voor het onderzoek naar de beschikbare tools, methodes, datacleaning middelen en databronnen betrof het een uitgebreide literatuurstudie in de zin van een analyse van gevonden informatie gericht op input voor de pakketselectie ((Curriculumcommissie IDM, 2010).

Hierbij was het doel het aanvullen van de verzamelde informatie uit de interviews met CETAF instellingen, met informatie over de functionaliteiten, bruikbaarheid en beschikbaarheid van genoemde tools. In het algemeen heb ik mij gericht op vakliteratuur, in plaats van wetenschappelijke bronnen, omdat ik op zoek

was naar praktische informatie, zoals handleidingen (Verhoeven, 2014). Deze informatie diende voor alle tools vergelijkbaar te zijn, zodat deze in een pakketselectietabel kon worden samengevat.

De interviews uit de vorige stap leverde 6 tools en 7 datacleaning middelen op. In aanvulling hierop heb ik ook methodes/ best practices en databronnen meegenomen in dit deskresearch onderdeel. Zo kon er een zo compleet mogelijk beeld gegeven worden van het mogelijke aanbod voor productiematig georeferenzen.

Voor elk deel van de deskresearch (per instrument) heb ik een passende zoekstrategie en zoekmethodes toegepast, zoals de sneeuwbal methode, de blokkendoos methode, booleaanse operatoren, aanhalingstekens, ed. (Veen en Westerkamp, 2010) De combinatie van verschillende zoekmethodes, draagt over het algemeen bij aan de betrouwbaarheid van het onderzoek en de resultaten die hieruit voortkomen (Verhoeven, 2014).

De competenties met betrekking op het uitvoeren van een gedegen literatuuronderzoek zijn voor mij regelmatig aan bod geweest in mijn gehele opleiding en in dit project. De beste combinatie van informatiebronnen, zoekstrategieën en zoekmethoden kan ik goed inschatten en ik pas verschillende methodes automatisch toe. Ik heb meerdere malen de opdrachtgever en afdeling kunnen verrassen met de hoeveelheid informatie die ik vond over de beschikbaarheid en betrouwbaarheid van de georeference instrumenten.

Pakketselectie

Aanvullend op de inventarisatie van tools, ervaringen en kennis binnen de CETAF instellingen, was het einddoel een advies te formuleren over de meest passende tool. De verwachting van de eerdere CETAF bijeenkomsten was dat er wellicht meerdere tools passend en geschikt zouden zijn (NBC, 2014a). Om de lijst met ruim 20 georeference hulpmiddelen te reduceren tot een selectie van de meest bruikbare combinatie van tools heb ik een pakketselectie uitgevoerd. Hierbij heb ik alle tools getoetst op basis van alle belangrijke functionaliteiten en technische aspecten die van een georeference tool worden verwacht (Curriculumcommissie IDM, 2010). Deze heb ik gebaseerd op de wensen en eisen van de gebruiker van de data, de benodigde metadata elementen en kwaliteitsnormen specifiek voor het vakgebied.

De georeferencing hulpmiddelen waren voor de beroepstaak deskresearch uitvoering onderzocht. Deze informatie, aangevuld met de wensen en eisen, heb ik samengevoegd tot een pakketselectie tabel waarin ik elke tool “scoorde” op al deze punten. In deze selectie heb ik de tools, methodes, bronnen en datacleaning middelen die niet beschikbaar waren niet meegenomen.

In plaats van voor elke wens/eis een plusje en een minnetje te gebruiken, heb ik iets meer informatie ingevuld voor alle tools. Bij veel tools golden er namelijk aanvullende voorwaarde voor bijvoorbeeld, de bruikbaarheid, de kosten, het aantal objecten, etc. Met het gebruik van enkel plusjes en minnetjes had ik wellicht een ander resultaat gekregen. Sommige van de eisen die in de selectie een minnetje zouden hebben gekregen, maken de tool niet direct onbruikbaar, terwijl dit minnetje dit wel zou suggereren.

De keuze voor deze systematische aanpak heeft geleid tot een zeer objectief resultaat. Het maakt het selectie proces direct inzichtelijk en het is makkelijk te zien wat de onderlinge verschillen zijn, welke tools het meest geschikt zijn en waar de eventuele onderlinge overlap zit. Door de toevoeging van vakspecifieke kwaliteitsnormen (Principles of best practice for georeferencing) voldoet de selectie en de uiteindelijke conclusie aan internationale standaarden (Chapman, 2006).

Het uitvoeren van een systematisch selectieproces is een competentie die ik in dit onderzoek voor vier soorten pakketten heb kunnen aantonen: tools, databronnen, methodieken en datacleaning middelen. De pakketselectie is nog niet de afsluitende conclusie in het onderzoeksrapport voor de opdrachtgever, maar liet wel duidelijk zien welke tools en methodes de meest geschikte zijn en hoe deze elkaar kunnen aanvullen.

Mijn keuze voor het toevoegen van aanvullende informatie, naast de plusjes en minnetjes, laat zien dat ik voldoende kennis heb van de mogelijkheden van de pakketten en kan inschatten wat hiervan relevant is voor de bedrijfscontext.

Parametriseren

Mijn uitwerking van de beroepstaak Parametriseren is licht afwijkend van de beschreven toelichting in de *“Uitwerking van de beroepstaken IDM versie 1.1”* van de Curriculumcommissie IDM (2010). De bijbehorende competenties zijn daarentegen wel allemaal aan bod gekomen tijdens het uitgevoerde proces.

Het grote verschil zit hem in het de term “applicatie”, die in de uitwerking van school wordt beschreven (Curriculumcommissie IDM, 2010). Mijn project betreft geen system of applicatie maar een Best Practice. Dit is een internationale standaard, uitgewerkt en beschreven in een document in de vorm van een stappenplan/handleiding met richtlijnen voor het gebruik.

Deze best practice volgde uit het advies dat ik kon formuleren naar aanleiding van de pakketselectie. Hierbij is een beschrijving gemaakt van de toepassing van de geselecteerde tools en hoe deze elkaar in het georeferencen proces kunnen opvolgen. De collectie- specifieke eisen, met betrekking tot metadata elementen, eigenschappen van een collectiesysteem, datakwaliteit en de gebruikersaspecten zijn hierin ook beschreven.

Het document heb ik ingedeeld in 3 onderdelen: pre- georeferencing, georeferencing en post-georeferencing. Per deel is als een stappenplan beschreven hoe de tools moeten worden gebruikt en welke soorten data en kennis daarvoor nodig zijn. In de bijlage zijn aanvullende formules en toepassingen beschreven, die kunnen gebruikt voor het aansturen van de tools. Om het nog makkelijker leesbaar te maken, is er veel gewerkt met tekstuele handigheden, leestekens, ed. deze simpele structuur maken het een behapbaar document.

In het onderzoeksverslag, waarin de werkwijze en resultaten van de deelvragen staan beschreven, is documentatie over mijn gemaakte keuzes te vinden. In het conclusie hoofdstuk uit het onderzoeksverslag en de discussie in de best practice is documentatie te vinden over aspecten waarmee voorgebouwd kan worden op het gerealiseerde product, zoals ontwikkeling van tools, of het aangaan van samenwerkingen.

De best practice kan op verschillende manieren gebruikt worden: hij kan worden gebruikt voor de huidige situatie, maar hij is ook makkelijk aan te passen of uit te breiden met nieuwe tools, die wellicht in de toekomst beschikbaar komen. Daarnaast is hij ook volledig aanpasbaar op basis van de kwaliteit van de originele locatiedata van een instelling.

Voor de structuur, inhoud en het uiterlijk van de Best Practice was ik eindverantwoordelijk. Via twee presentaties, één voor de afdeling informatiemanagement en één voor de ICT sector van Naturalis, heb ik binnen de opdrachtgevende instelling aangetoond waar en hoe de best practice optimaal gebruikt kan worden en wat voor resultaten hieruit kunnen volgen.

11.2 Eindreflectie



Het onderwerp georeferencing is er één met veel draagvlak binnen Naturalis en waarvan de waarde hoog word ingeschat. Dit werd mij meerdere keren duidelijk tijdens mijn onderzoek. Bij aanvang van mijn stage bleek het beoogde onderzoek de gesprekken over het onderwerp te hebben aangewakkerd. Tijdens de interviews die ik afnam met interne gebruikers van toekomstige georeference data werd dit enthousiasme en de behoefte aan een dergelijk onderzoek bevestigd.

Dit zelfde enthousiasme bleek nog meer tijdens de interviews met enkele andere CETAF organisaties, die voor deelvraag 3 zijn afgenomen. Georeferencing bleek een zeer veel onderzocht onderwerp, en mijn best practice was een document waar veel mensen naar uitkeken.

Naast dat dit bij mij veel enthousiasme en motivatie opwekte, zorgde dit uiteraard ook voor de nodige spanning. Als zoveel mensen het resultaat tegemoet zien en willen gebruiken, moet het ook van uitstekende kwaliteit zijn, en bruikbaar zijn voor alle geïnteresseerden.

Uiteraard had dit alles veel invloed op mijn werkwijze. Ik moest het zowel toepasbaar maken op de situatie van Naturalis als op de situaties van andere biodiversiteitsinstelling. Het onderzoek moest daarbij ook algemeen genoeg zijn, aan te kunnen passen aan elke instellingspecifiek situatie, en tegelijk specifiek genoeg zijn om direct te kunnen worden ingezet.

De oplossing die ik hiervoor zag was het verzamelen van veel informatie met verschillende invalshoeken: ervaringen, theoretische kennis en technische aspecten. Door deze verschillende invalshoeken heb ik ook veel verschillende onderzoeksmethode kunnen uitvoeren en mijn vaardigheid hierin laten zien: interviews, deskresearch, pakketselectie, ed.

De competenties, die technischer van aard zijn dan ik gewend ben, zag ik vooraf als een drempel, die ik wellicht met veel moeite zou moeten overwinnen. Door rustig en stapsgewijs deze onderdelen aan te pakken heb ik het toch op professioneel niveau kunnen uitvoeren. Hiermee heb ik de competities waar ik mij de rest van de opleiding op heb gericht, voornamelijk “analyseren”, “adviseren” en “beheren/exploiteren”, kunnen uitbreiden met nieuwe competenties gericht op “realiseren” en “ontwerpen” (Curriculumcommissie IDM, 2010).

De spanning heb ik opgevat als iets positiefs en iets wat vanzelfsprekend is bij een afstudeertraject. Ik ontdekte dat een uitdaging als deze bij mij een motivatie en “drive” naar boven haalt om het beste van mezelf te laten zien en duidelijk te maken dat ik in staat ben zoiets professioneel af te ronden. Dit werd versterkt door mijn enthousiasme over het bedrijf. Naturalis is een organisatie die ik altijd zeer interessant heb gevonden en waar ik erg graag mijn afstudeerproject wilde uitvoeren.

Het werken aan een project en product dat daadwerkelijk zal worden gebruikt en gepubliceerd door deze instelling (en andere vergelijkbare instellingen) was voor mijn afstudeeropdracht daarom de perfecte situatie. Ik heb het idee dat ik veel heb kunnen laten zien van wat ik tijdens de opleiding heb geleerd en tegelijk me heb kunnen ontwikkelen als werkende Informatieprofessional.

Na afloop ben ik van menig dat ik met deze insteek een zeer volledig onderzoek heb kunnen afleveren en de inventarisatie van ervaringen, kennis en projecten waar vraag naar was, heb kunnen opleveren.

Literatuurlijst

Gepubliceerde bronnen

Baarda, B. (2014). *Dit is Onderzoek, Handleiding van kwantitatief en kwalitatief onderzoek*. Groningen: Noordhof Uitgevers

Chapman, A.D. and Grofton, O. (2008). *Guide to Best Practice for Generalizing Primary Species-Occurrence Data*. Copenhagen: Global Biodiversity Information Facility

Chapman, A.D. & Wieczorek, J. (2006). *Guide to best practices for georeferencing*. Copenhagen: Global Biodiversity Information Facility

Chrisman, N.R. (1983). *The role of quality Information in the long-term functioning of a GIS. Proceedings of AUTOCART06*, 2: 303-321. Falls Church, VA:ASPRS

Curriculumcommissie IDM (2010). *“Uitwerking van de beroepstaken IDM versie 1.1*. Academie ICT&Media, Haagse Hogeschool

Erp, M. van & Hensel, R. & Ceolin. Davide. Meij, M. van der(2014). Georeferencing Animal Specimen Datasets. *Transactions in GIS*, 12/2014: 19(4)

GBIF (2010). *Darwin Core Archives – How-to Guide, version 1*. Copenhagen: Global Biodiversity Information Facility, 21 pp

Gritt, R. (2011). *Projectmanagement*. Groningen/Houten: Noordhoff Uitgevers

Grover, C., Givon, S., Tobin, R., Ball, J.(2008). Named Entity Recognition for Digitised Historical Texts. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*

Llewellyn, C. & Grover, C. & Oberlander, J. & Haston, E. (2012). Enhancing the Curation of Botanical Data Using Text Analysis Tools. *Theory and Practice of Digital Libraries, volume 7489 of the series Lecture Notes in Computer Science*, pp 480-485

Llewellyn, C.A. & Haston, e. & Grover, C. (2011). Georeferencing Botanical Data using Text Analysis Tools. *TDWG 2011 Conference*, Geraadpleegd op 15 oktober 2015, via

Murphey, P. & Guralnick, R. & Glaubitz, R. & Neufeld, D. & Allen Ryan, J. (2004). Georeferencing of museum collections: a review of problems and automated tools, and the methodology developed by the mountain and Plains Spatio-Temporal Database-Informatics Initiative. *Phyloinformatics*, volume 3, 1-29

NBC (2015b). *Uit het Depot, op het web: Twee eeuwen nationaal natuurhistorisch erfgoed in het digitale domein*. Leiden: Naturalis Biodiversity center

NBC (2012b). *Beleidsplan 2013-2016*. Leiden: Naturalis Biodiversity Center

Spencer, C. & Yamamoto, K. & Fang, J. & Constable, H. & Koo, M. & Wieczorek, J. (2005) *Georeferencing for Dummies: An Elaboration of the MaNIS/HerpNET/ORNIS Guidelines*. Berkeley: University of California.

Steehouder, Jansen, Mulder, van der Pool, de Zeijl (2012). *Leren communiceren*. Groningen/ Houten: Noordhoff Uitgevers.

Veen, M. van & Westerkamp, K. (2010). *Deskresearch, Informatie selecteren, beoordelen en verwerken*. Amsterdam: Pearson Benelux

Verhoeven, N. (2014). *Wat is onderzoek? praktijkboek voor methoden en technieken*. Amsterdam: Boom Uitgevers

Internetbronnen

Blaauboer, R. 1 (2014) *Digitaal erfgoed: zo past crowdsourcing in je bedrijfsmodel*. Geraadpleegd op donderdag 20 november 2015, via <http://www.frankwatching.com/archive/2014/11/03/digitaal-erfgoed-zo-past-crowdsourcing-in-je-bedrijfsmodel/>

CETAF (z.d.-a). *Consortium of European Taxonomic Facilities, What is CETAF*. Geraadpleegd op 2 september 2015, via: <http://cetaf.org/about-us/what-cetaf>

CETAF (z.d.-b). *Consortium of European Taxonomic Facilities, CETAF Bodies*. Geraadpleegd op 2 september 2015, via: <http://cetaf.org/taxonomy/cetaf-bodies>

DEN (2009 a). *De projectenbank: innovatieve projecten van Nederlandse erfgoedinstellingen*. Geraadpleegd op 9 november 2015, via <http://www.den.nl/projecten>

DEN (2009 b). *Mining for information in texts from the cultural heritage, projectgegevens*. Geraadpleegd op 9 november 2015, via <http://www.den.nl/project/166/Mining-for-information-in-texts-from-the-cultural-heritage>

Direct Web Solution (2015). *Webapplicaties, complexe webapplicaties*. Geraadpleegd op 11 september 2015, via <http://www.directwebsolutions.nl/webapplicatie-ontwikkeling/complexe-webapplicaties.html>

Erfgoed & Locatie (2015). *Demo-versie Historische Geocoder online*. Geraadpleegd op 18 november 2015, via <http://erfgoedenlocatie.nl/2015/04/demo-versie-historische-geocoder-online/>

Google Developers (2015). *The Google Maps Geocoding API*. Geraadpleegd op 17 november 2015, via <https://developers.google.com/maps/documentation/geocoding/intro#GeocodingRequests>

HerpNET (2010). *Georeference Resources*. Geraadpleegd op 2 November 2015, via <http://www.HerpNET.org/Gazetteer/GeorefResources.htm>

Marsh, J. (2014). *Data Dictionary Example: Defining Your Database Tables*. Geraadpleegd op 11 september 2015, via <https://blog.udemy.com/data-dictionary-example/>

NBC (2015a). *Wat wij doen: onderzoek*. Geraadpleegd op 4 september 2015, via: <http://www.naturalis.nl/nl/over-ons/wat-doen-wij/onderzoek/>

NBC (2015d). *Wat wij doen: Collecties*. Geraadpleegd op 4 september 2015, via: <http://www.naturalis.nl/nl/over-ons/wat-doen-wij/collecties/>

NBC (2015c). *Organisatie*. Geraadpleegd op 7 september 2015, via <http://www.naturalis.nl/nl/over-ons/organisatie/>

New South Wales schools HSC (z.d.) *Data modelling: Data dictionaries*. Geraadpleegd op 11 september 2015, via http://www.hsc.csu.edu.au/ipt/info_systems/4032/data_dictionaries.htm

Wikipedia (2015). *Scrum (software development)*. Geraadpleegd op 8 september 2015, via [https://en.wikipedia.org/wiki/Scrum_\(software_development\)](https://en.wikipedia.org/wiki/Scrum_(software_development))

TDWG (2010 a). *ABCD - Access to Biological Collection Data*. Geraadpleegd op 14 september 2015, <http://wiki.tdwg.org/twiki/bin/view/ABCD/WebHome>

TDWG Wiki (2010 b). *Geospatial Extension Concept List*. Geraadpleegd op 14 september 2015, via <http://wiki.tdwg.org/twiki/bin/view/DarwinCore/GeospatialExtension>

TDWG (2007). *Mapping between Darwin Core 1.4 concepts (DwC) and ABCD 2.06b*. Geraadpleegd op 14 september 2015, via <http://www.bgbm.org/TDWG/CODATA/Schema/Mappings/DwCAndExtensions.htm>

Tilburg University. (2005). *MITCH, Mining for Information in Texts from Cultural Heritage*. Geraadpleegd op 9 november 2015, via <http://ilk.uvt.nl/mitch/>

Tilburg University. (2005). *MITCH, Mining for Information in Texts from Cultural Heritage Publications*. Geraadpleegd op 9 november 2015, via <http://ilk.uvt.nl/mitch/publications.html>

VertNET (2013). *Georeferencing, Mapping resource hub*. Geraadpleegd op 2 november 2015, via <http://georeferencing.org/>

Waag Society (2015). *Hacking Heritage Talk 17 September*. Geraadpleegd op 10 september 2015, via <https://www.waag.org/nl/event/hacking-heritage-talks>

Wieczorek, J. (2001) "*MaNIS/HerpNet/ORNIS Georeferencing Guidelines*". Geraadpleegd op 2 november 2015, via <http://manisnet.org/GeorefGuide.html>

Wieczorek, J., D. Bloom. 2011. *Georeferencing Calculator Manual v2*. Geraadpleegd op 18 november 2015, via <http://manisnet.org/GeoreferencingCalculatorManualv2.html>

Niet-gepubliceerde bronnen (intern)

Arkel, B. van & Creuwels, J. & Leusen, J. van & Schnörr, S. (2014). *Eindrapport FCD- Pilot Georeferencing: productiematig georeferencen*. Intern document Naturalis, Leiden (NBC, 2012)

Hine, A. & Penn, M. (2014). *Georeferencing at the NHM*. Presentation. Naturalis Biodiversity center, Leiden

Mollema, K.J. (2013). *Documentmanagement en pakketselectie, Blok C college Tools*. Academie ICT&Media, Haagse Hogeschool. Geraadpleegd op 30 november 2015 2.

NBC (2014a). *Follow-up Actions Meeting Automated Georeferencing*. Intern document, geraadpleegd op 2 september 2015

NBC (2014b). *Technical Summery , Automated Georeference meeting December*. Intern document, geraadpleegd op 2 september 2015

NBC (2013a). *Richtlijnen Digitaliseringproject*. Intern document, geraadpleegd op 16 september 2015

NBC (2013b). *Beleidslijn Open Content*. Intern document, geraadpleegd op 18 september 2015

NBC (2012a). *Collectieplan Naturalis Biodiversity center 2013-2016*. Intern document, geraadpleegd op 15 september 2015

NBC (z.d.a). *Datamodel BRAHMS, Botanische gedachten BRAHMS vs BOLD*. Intern document, geraadpleegd op 17 september 2015

NBC (z.d.b). *Help datamodel CRS*. Intern document, geraadpleegd op 15 september 2015

Schnörr, S. (2014). *Georeferencing Methode Niels Raes*. Summery. Intern document, geraadpleegd op 7 november 2015

Interviews

Persoonlijke communicatie (23 September 2015). Luc Willemse, Collectiebeheerder entomologie

Persoonlijke communicatie (25 September 2015). Ruud Altenburg, Ontwikkelaar Het BioPortal

Persoonlijke communicatie (29 September 2015). Agnes Kirchhoff, The Botanical Garden and Botanical Museum, Berlin-Dahlem

Persoonlijke communicatie (30 September 2015). Cees Hoff, Nederlandse Node Global Biodiversity Information facility

Persoonlijke communicatie (6 Oktober 2015). Elspeth Haston, The Royal Botanical Garden of Edinburg

Persoonlijke communicatie (6 Oktober 2015). Hans Muller, Wikipedian in Residence

Persoonlijke communicatie (8 Oktober 2015). Niels Raes, Interne onderzoeker botanie

Persoonlijke communicatie (15 October 2015). Patricia Mergen, The Royal Museum of Central Africa Tervuren

Persoonlijke communicatie (29 October 2015). Malcolm Penn, The Natural History Museum of London

Overzicht van tools, & middelen



Georeference tools

- GeoLocate Georeferencing Software: <http://www.museum.tulane.edu/geolocate/>
- Google Geocoding API: Google (2015).
<https://developers.google.com/maps/documentation/javascript/examples/geocoding-simple>
- SPECimap: *Nog niet beschikbaar*
- BioGeomancer: *Niet meer beschikbaar*
- Georeferencing Calculator: *Niet meer beschikbaar*
- MITCH: *Tool niet beschikbaar, documentatie wel*: <http://ilk.uvt.nl/mitch/>

Georeference methodes

- MaNIS/HerpNET/ORNIS Georeferencing Guidelines: <http://manisnet.org/GeorefGuide.html>
- Chapman's *Guide to best practices for georeferencing*

Georeference databronnen

- Geonames: <http://www.geonames.org/data-sources.html>
- The Getty Thesaurus of Geographic Names
- Google Maps: <https://www.google.nl/maps>
- The Fuzzy Gazetteer: <http://isodp.hof-university.de/fuzzyg/query/>
- Histogram: <http://histograph.io/>

Georeference datacleaning middelen

- Manually visual check with Google Maps:
<https://developers.google.com/maps/documentation/javascript/examples/geocoding-simple>
- Plaatsnamen Standaardiseren: <http://standaardiseren.erfgeo.nl/>
- Open Refine: <http://openrefine.org/>
- GBIF: <http://www.gbif.org/>
- Specieslink: <http://splink.cria.org.br/index?criaLANG=en>
- SYNTHESYS NA-D 3.7 'Itinerary' project: *Tool niet beschikbaar, documentatie wel*:
http://www.biocase.org/products/geo_services/itineraries/

Verklarend woordenboek

In de onderstaande verklarende woordenlijst worden onderzoek specifieke en vakspecifieke termen uitgelicht, zoals deze in dit onderzoek worden gebruikt of bedoeld.



Accuracy (Nauwkeurigheid)	is sterk gerelateerd aan de onzekerheid. Bij de onzekerheid wordt namelijk een polygoon of cirkel rond de locatie getrokken waarin het object zich moet bevinden. De afstand van het centrum van deze polygoon tot de meest uiterlijke hoeken ervan, is de nauwkeurigheid. Hoe kleiner de polygoon hoe beter de nauwkeurigheid.
Datacleaning (correctie en validatie)	een proces dat wordt gebruikt om onjuiste, onvolledige of onredelijke gegevens te bepalen en deze ontdekte fouten en onvolledigheden vervolgens aan te passen om de kwaliteit van de data te verbeteren. Voorbeelden: formaat controles, volledigheid controles, analyse van uitschieters (geografische, statistische, ed.), en evaluatie van gegevens door vakgebied experts.
Distributie	de verspreiding van soorten over een bepaald gebied of over de hele wereld en de veranderingen die hierin optreden.
Extent	het geografische bereik, de omvang, of de afstand die een locatie eigenlijk vertegenwoordigt. In het geval van een stad is de extent een polygoon waarbinnen het gebied tussen de stadsgrenzen valt.
Gazetteers	een geografisch woordenboek of een index met de namen van locaties van een bepaald gebied, inclusief een indicatie van de positie op het aardoppervlak met behulp van één van de verschillende geografische coördinatenstelsels. Veel gazetteers hebben een bepaald onderwerp, zoals: waterlichamen op aarde, bewoonde gebieden, historische plaatsen, bergketens, etc.
Georeference	de gegevens die worden gegenereerd tijdens het georeference proces, zoals: lengtegraad, breedtegraad, nauwkeurigheid, ed. Zie voor een volledig overzicht van mogelijke georeference data bijlage 2
Headings	de richting vanaf een startlocatie, gegeven in de vorm van punten op het kompas, zoals O, NO, W, ZW, etc. Meestal gebruikt in combinatie met offset om een afstand en richting van een aangeduide plaats te geven.
Locality	een nederzetting. Dit is in de geografie, statistiek en archeologie een gebied of regio die bewoond is (of is geweest) door mensen.
Named place	een locatie die aangeduid kan worden, met of zonder eigen naam. Bijvoorbeeld: kruispunten, aftakkingen van rivieren, snelweg meterpalen, cellen van een grid systeem.

Offsets	de afstand vanaf een genoemde plaats waarbij de locatie van de genoemde plaats als startpunt wordt gebruikt. Meestal gebruikt in combinatie met <i>headings</i> om een afstand en richting van een aangeduide plaats te geven.
Plotten	het visualiseren of het vastleggen van een hoeveelheid gegevens, waarbij gebruik gemaakt wordt van coördinaten.
Schijn Nauwkeurigheid	treedt op als de onzekerheid en nauwkeurigheid van een georeference locatie een betere waarde indiceren, dan dat deze daadwerkelijk voorstelt. Dit treedt bijvoorbeeld op als coördinaten met meer decimalen worden geregistreerd dan ooit zijn vastgelegd.
Stand- up	Een dagelijkse korte meeting van maximaal 15 minuten, waar de werkzaamheden van die dag en de dag ervoor worden besproken. Ook wel Daily Scrum of Daily Stand up genoemd.
Taxonomie	(systematiek) is de wetenschap van het indelen van individuen of objecten in groepen. In de biologie is dit een vakgebied dat zich bezighoudt met de verscheidenheid van organismen op aarde en het indelen (classificeren/ determineren) van soorten.
Uncertainty (Onzekerheid)	tijdens het georeferencen wordt er een punt geplaatst op een kaart die een textuele locatie beschrijving representeert. De mate waarin dit punt ook daadwerkelijk de gezochte locatie representeert, wordt onzekerheid genoemd. Dit wordt vaak uitgedrukt in de vorm van een polygoon, of een cirkel waarin de gezochte locatie zich moet bevinden.
Verbatim	Woordelijk of letterlijk. Dit heeft betrekking op de huidige geografische data in de collectie van Naturalis. Deze informatie is namelijk tijdens het digitaliseren letterlijk overgetypt van de etiketten of labels van de objecten. Naar deze data wordt verwezen met de term verbatim data.

Bijlage 1: Overzicht metadata elementen datadictionary



Mapping ABCD en Darwin Core

De georeference specifieke metadata elementen volgens de mapping tussen de DwC- en ABCD standaarden (TDWG, 2007):

- **Decimal latitude:** The latitude of the geographic center of a location where an event occurred (organism collected, observation made), expressed in decimal degrees.
- **Decimal longitude:** The longitude of the geographic center of a location where an event occurred (organism collected, observation made), expressed in decimal degrees.
- **Geodetic datum:** The geodetic datum to which the latitude and longitude refer.
- **Coordinate uncertainty in meters:** The upper limit of the distance (in meters) from the given DecimalLatitude and DecimalLongitude describing a circle within which the whole of the described locality lies
- **Point radius spatial fit:** A measure of how well the circle defined by the coordinates and uncertainty match the original spatial representation, as a ratio of the area of the circle to the area of the original spatial representation.
- **Verbatim coordinates:** A text representation of the coordinate data
- **Verbatim latitude:** A text representation of the Latitude part of the coordinate data from its original source
- **Verbatim longitude:** A text representation of the Longitude part of the coordinate data from its original source.
- **Verbatim coordinate system:** The name of the system in which the verbatim geographic coordinates were recorded
- **Georeference protocol:** A reference to the methods used for determining the coordinates and uncertainties
- **Georeference sources:** A list of maps, gazetteers or other resources used to georeference the locality.
- **Georeference verification status:** A categorical description of the extent to which the georeference has been verified to represent the location where the specimen or observation was collected.
- **Georeference remarks:** Comments about the spatial description determination, explaining assumptions made in addition or opposition to the those formalized in the method referred to in GeoreferenceProtocol.
- **Footprint WKT:** A Well-Known Text representation of the shape (footprint, geometry) that defines the location of the occurrence.
- **Footprint spatial fit:** A measure of how well the geometry expressed in the footprint match the original spatial representation, as a ratio of the area of the footprint given to the area of the original spatial representation.

BioGeomancer Workbench Guidelines

De georeference specifieke metadata elementen volgens de Biogeomancer Workbench Guidelines, op basis van de Guide to Best Practices for Georeferencing (BioGeomancer Consortium, 2006):

- **Decimal Latitude:** the latitude coordinate (in decimal degrees) at the center of a circle encompassing the whole of a specific locality. Decimal latitudes north of the equator are positive numbers less than or equal to 90, and those south are negative numbers greater or equal to -90
- **Decimal Longitude:** the longitude coordinate (in decimal degrees) at the center of a circle encompassing the whole of a specific locality. Decimal longitudes east of the Greenwich Meridian are considered positive and less than or equal to 180, while western longitudes are negative and greater than or equal to -180.
- **Geodetic Datum:** a model of the earth used for geodetic calculations. A geodetic datum describes the size, shape, origin, and orientation of a coordinate system for mapping the surface of the earth
- **Maximum Uncertainty Estimate:** The upper limit of the distance from the given latitude and longitude describing a circle within which the whole of the described locality must lie.
- **Maximum Uncertainty Unit:** The unit of length in which the maximum uncertainty is recorded
- **Verbatim Coordinates:** The original (verbatim) coordinates of the raw data before any transformations were carried out.
- **Verbatim Coordinate System:** The coordinate system in which the raw data were recorded.
- **Georeference Verification Status :** A categorical description of the extent to which the georeference and uncertainty have been verified to represent the location and uncertainty for where the specimen or observation was collected
- **Georeference Validation:** Shows what validation procedures have been conducted on the georeferences – for example various outlier detection procedures, revisits to the location, etc
- **Georeference Protocol:** A reference to the method(s) used for determining the coordinates and uncertainty estimates
- **Georeference Sources:** The reference source (e.g., the specific map, gazetteer, or software) used to determine the coordinates and uncertainties
- **Spatial Fit:** A measure of how well the geometric representation matches the original spatial representation and is reported as the ratio of the area of the presented geometry to the area of the original spatial representation
- **Georeference Determined By:** The person or organization making the coordinate and uncertainty determination
- **Georeference Determined Date:** The date on which the determination was made.
- **Georeference Remark:** Comments on methods and assumptions used in determining coordinates or uncertainties

Naturalis- CRS

De geografische metadata elementen die onder het element [Gathering site] vallen, in het CRS van Naturalis:

- **Geographic place names:** *place name and place type combined*
 - **Place name:** *Name of the gathering site (a geographic*
 - **Place type:** *Classification categories for the class of the gathering named site (local or national subdivision levels*
- **Bioregion:** *Describe the bioregion in which the specimen was collected or observed.*
- **Country:** *Name of country or major region in which the specimen was collected or observed*
- **State provinces:** *Field to indicate the state or province in which the specimen was collected or observed.*
- **Island:** *Island name in which the specimen was collected or observed*

- **Locality:** Field to indicate the locality full name where the specimen was collected or observed. Additional notes e.g. [10 Km South of..]. must be placed in Full locality text.
- **Station number:** Field to indicate the station number of the locality where the specimen was collected or observed.
- **Full locality:** Original locality data as appearing on a label or in an original entry
- **Coordinates:** all following fields combined:
 - **Index:** In case an area is described an index number is added to the data.
 - **Methods:** Coordinates measuring system e.g. GPS
 - **Verbatim coordinates:** Original coordinate data as appearing on a label or in an original entry
 - **Accuracy:** Textual statement of degree of accuracy
 - **Distance in meters:** An estimate of how tightly the collecting locality was specified; expressed as a distance in meters corresponding to a radius around the coordinates
 - **Lat long**
 - **Lat:** Latitude in decimal degrees
 - **Long:** Longitude in decimal degrees
 - **System:** Mathematical surface on which the mapping and coordinate system used for the geocodes of the record are based
 - **UTM**
 - **Verbatim UTM:** Verbatim concatenated text representation of UTM coordinates
 - **Zone:** The numerical zone corresponding to the central meridian and origin upon which the UTM Easting is based
 - **Subzone:** The subzone letter corresponding to one of the 20 North-South divisions of the UTM grid system
 - **NS:** The hemisphere to which the UTM Northing refers (North or South)
 - **Easting:** The distance in meters east of the origin of the UTMZone
 - **Northing:** The distance in meters north of the origin for the UTMSubzone. For the northern hemisphere (UTMSubzones M through X) the origin is at the equator. For the southern hemisphere (Subzones C through L)
 - **Datum:** Mathematical surface on which the mapping and coordinate system used for the geocodes of the record are based.
 - **Grid**
 - **Cell system:** The name of the grid system used for the gathering site coordinates
 - **Cell code:** The code of the grid system used to record the gathering site coordinates
 - **Qualifier:** A grid reference precision qualifier for the gathering site coordinates
 - **Latitude:** Latitude
 - **Longitude:** Longitude

Naturalis- BRAHMS

De geografische metadata elementen in het Brahms van Naturalis:

- **Country:** chose from the look up list: 'ISO Countries or Oceans'
- **Major country area:** highest-order subdivision of the country' political or natural unit
- **Minor country area:** second highest-order subdivision of the country', like 'State/province (by Getty Geographical Thesaurus)
- **Locality :** Third order subdivision of the country. This includes city's and smaller, like a park, lake, bedding or river
- **Locality notes:** additions to the locality to further describe the geographic location of the object
- **LLunit:** type of notation used for the geographic coordinates, use DD-notation.
- **Latt:** Latitude: this depends on the notation used (LLUNIT), hereby N is positive and S is negative

- **NS:** North or South of the equator
- **Long:** Longitude: this depends on the notation used (LLUNIT), hereby E is positive and W is negative
- **Ew:** East or West of the Greenwich meridian
- **LLRES:** The accuracy of the coordinates on a scale of 1° tot 0.01'
- **LLorig:** Required if coordinates are on the label. In that case, fill in "sheet" or blank if there no coordinates
- **LLdatum:** Mathematical surface on which the mapping and coordinate system used for the geocodes of the record are based

Geadviseerde georeference metadata elementen

Alle standaarden en de use case van Naturalis samengevoegd tot één universeel overzicht van primaire en secundaire georeference specifieke metadata elementen voor een biodiversiteitscollectie:

Primary metadata elements	
Fieldname per standard	Description/definition
LatitudeDecimal	The latitude of the geographic center of the locality expressed in decimal degrees.
LongitudeDecimal	The longitude of the geographic center of the locality expressed in decimal degrees.
SpatialDatum	The geodetic datum to which the latitude and longitude refer.
CoordinatesText or Coordinates UTM/UTMText	A text representation of the coordinates. This can be one element for latitude and longitude, but can also be separate elements
CoordinatesGrid/GridCellSystem	The name of the system in which the verbatim geographic coordinates were recorded.
CoordinateMethod	A reference to the methods used for determining the coordinates and uncertainties.
coordinateErrorDistanceInMeters or AccuracyStatement	A measure of the area in which the described locality must lie. Usually expressed by the distance from the coordinates to the upper limit/outer corners of the polygon/radius, in meters. A free text statement of the degree of accuracy of the latitude and longitude coordinates.
Secondary metadata elements	
Fieldname	Description/definition
FootprintSpatialFit	The overlap between the actual locality and the georeference polygon/point
GeoreferenceSources	A list of maps, gazetteers or other resources used to georeference the locality.
GeoreferenceVerificationStatus	The status of the georeference data; "What validation steps have been conducted?"
GeoreferenceRemarks	Comments about the recorded georeference data
Georeferenced By	The person or organization making the coordinate and uncertainty determination.
Georeferenced Date	The date on which the determination was made.

Bijlage 2: Codering interviews gebruikers van geografische collectiedata van Naturalis

Naturalis, collectiebeheer 23-09-2015	
Transcriptie interview	Code label
Gebruik huidige geografische data Geografische gegevens in onze collectie worden door collectiebeheer voornamelijk gebruikt voor het juist determineren van objecten in de collectie omdat het een belangrijk is voor het compleet maken van onze collectiedata. Daarnaast speelt het een rol bij het verzamelen van collectiedata op basis van een onderzoeksvraag. Bij digitalisering is er altijd een strijd tussen kwaliteit en kwantiteit en voor het collectiebeheer ligt vooral de nadruk op kwaliteit. De grote druk achter digitalisering is het toegankelijk maken van de collectie, maar voor ons hoort daar het juist plaatsten en omschrijven (determineren) van onze objecten bij. Digitalisering en georeferenzen zijn iets wat daarbij komt, maar in beide onderdelen is de betrouwbaarheid van gegevens heel belangrijk. Waarde van huidige geografische data De verbatim geo- informatie geeft enige richting en het is wel belangrijk dat dit bewaard blijft, maar een coördinaat is veel bruikbaar voor onderzoek. Als een object alleen tekstuele locatie beschrijving heeft, zoals het grootste deel van onze collectie, zal de onderzoeker in de meeste gevallen zelf de coördinaten hierbij zoeken, of de objecten zelf op kaarten plaatsten. Met tekstuele beschrijvingen, alleen kan je namelijk niet zoveel. Gebruik toekomstig georeference data Voor de afdeling collectiebeheer zitten er ook vast wel enkele mogelijkheden in het gebruik van georeference data. Zo kan de collectie bijvoorbeeld op bepaalde geografische gebieden worden gesorteerd en word het makkelijker in sommige gevallen de gevraagde collectiedata te verzamelen. Stel wederom dat er wordt gevraagd naar collectiedata van vogels uit een bepaald reservaat, dan kost het ons veel tijd omdat we eerste moeten uitzoeken in welk land, provincie, stad, etc. dit ligt en dan de in deze omgeving verzamelde objecten zoeken. Je echter in het CRS niet op geografische metadata zoeken. Je moet dan dus vervolgens uitzoeken, welke soorten hier voorkomen, of welke verzamelaars hier zijn geweest. Met coördinaten kan je als het ware alles tussen bepaalde coördinaatwaarden objectgegevens aanvragen, mits het CRS deze opties ondersteund. Dat scheelt veel tijd. Coördinaten kunnen zeker veel toevoegen aan de collectie, mits deze juist zijn. Er zijn namelijk zo veel fouten die kunnen voorkomen waardoor onbetrouwbare gegevens gegenereerd kunnen worden. Grote fouten, zoals plus min fouten kun je vaak makkelijk zien als collecties op een kaart plaatsen, maar kleinere fouten niet. Stel een vogel heeft een bepaald verspreidingsgebied en een object is georeferenced aan de grens hiervan, het valt dan wel in het juiste gebied, maar misschien hoort het wel helemaal aan de andere kant van het gebied. In dit geval zie je de fout dan niet (of minder snel), maar het is wel een fout. Bij vragen zoals naar vogels in een reservaat, kan dit het onderzoek verstoren.	Waarde Waarde Probleem Problemen

Belang georeference data De taxonomische data (zoals soortnaam) en de coördinaten van objecten zijn de belangrijkste soorten van informatie die een onderzoeker nodig heeft voor onderzoek naar verspreiding van soorten. Dit zijn dus ook de meeste gebruikte gegevens van de collectie. Hiermee kunnen kaarten worden gemaakt, en kan de data worden gecombineerd met andere soorten gegevens, zoals ecologische of klimatologische gegevens. Een coördinaat heeft daarnaast ook als voordeel dat deze vast staat als deze eenmaal is geplaatst. De coördinatensystemen veranderen namelijk niet zo snel en veranderingen in het landschap hebben er geen effect op. Tekstuele beschrijvingen worden hierdoor echter wel beïnvloed. Een voorbeeld: de locatiebeschrijving van een object is 10 km ZW van het centrum van Leiden, en dit is vastgelegd in 1850. Nu is Leiden, in de afgelopen 150 jaar, veel groter geworden, dus 10 km ZW van het centrum ligt ineens veel verder weg, dan dat het in 1850 lag. Met een coördinaat heb je dit probleem minder snel. Coördinaten zijn dus ook duurzamer. Voor planten is het anders, want daar worden duplicaten verzameld. Dus instellingen in Parijs, Amsterdam en Londen hebben dezelfde objecten, hierbij is het ook belangrijk dat de data kan worden uitgewisseld, zodat er geen dubbel werk wordt gedaan. Gewenste nauwkeurigheid georeference data Wat betreft de nauwkeurigheid geldt natuurlijk dat hoe preciezer de coördinaat hoe bruikbaar het is, maar met een vlak van 10 bij 10 km, kom je voor de meeste soorten onderzoek al heel ver. Als je vragen krijgt over de verspreiding van vogelsoorten in een specifiek reservaat is 10 bij 10 km niet precies genoeg, maar daarbij moet je natuurlijk ook een afweging maken tussen het aantal zeer specifieke vragen die je krijgt en de energie/tijd die kost om de hele collectie te georeferencen.	Waarde Waarde Waarde Discussie Metadata Discussie
---	---

Het Bioportaal 25-09-2015	
Transcriptie interview	Code label
Gebruik huidige geografische data Op dit moment is het mogelijk om in het Bioportaal op basis van geografische gegevens de ontsloten collecties te doorzoeken, maar dit geldt alleen voor de collecties met bruikbare geografische gegevens; coördinaten. We gebruiken hiervoor alleen de lat-long coördinaten. Dit is voor het Brahms ongeveer één derde, en voor het CRS, gok ik ongeveer 6 procent. In het geografische zoekstelsel kan je op land, stad, of gebied zoeken naar specifieke soorten, maar je kunt ook zelf een gebied selecteren in een kaart. Hier wordt op de achtergrond van het Bioportaal elasticsearch voor gebruikt. Hoe kleiner de gebieden, hoe moeilijker dit is, maar dit ligt niet aan de gegevens van de collecties maar aan een bug in de huidige versie van Elastic Search. Waarde van huidige geografische data In principe controleren wij de geografische gegevens die via het Bioportaal beschikbaar komen niet, en gebruiken we de data die direct uit het CRS of het Brahms komt. Hierin zitten er veel onderlinge verschillen tussen de collecties. Tijdens het digitaliseren is de data uit de vorige systemen en van de labels/ etiketten gewoon overgenomen, zonder grootschalige controle of interpretatie van de gegevens. Ook tussen de aanwezige lat long coördinaten zitten grote verschillen. Sommige hebben 2 decimalen en andere wel 4, waardoor die veel nauwkeuriger zijn. Daarnaast word de data van sommige collecties door onderzoekers, of collectiebeheer zelf uitgebreid of zelfs 'georeferenced'. Zo weet ik dat er bij een collectie vogels, onlangs, na het digitaliseren extra geografische informatie is toegevoegd. Deze is echter gebaseerd op	Metadata Probleem Probleem Probleem

schattingen en secundaire bronnen. Van deze collectie is bijvoorbeeld bekend, dat ze verzameld zijn tijdens een rondreis, waarvan de route is bekend. Van sommige objecten in de collectie is bekend waar deze precies zijn verzameld en op basis daarvan kan een schatting worden gemaakt welke vogels waar op de route zijn gevonden. Hoe accuraat en betrouwbaar deze gegevens zijn is onbekend, maar ondanks dat, word alles wat coördinaten beschikbaar heeft, gebruikt bij de geografische zoekopties in het Bioportaal. Deze data is echter ongecontroleerd. Gebruik toekomstig georeference data Het Bioportaal is zeer geïnteresseerd in georeference data. Hoe meer collecties er zijn verrijkt met coördinaten, hoe meer er via ons Bioportaal op kaarten toegankelijk gemaakt kan worden, en er op geografische wijze gezocht kan worden, in de collectie van Naturalis. Belang georeference data Dit is niet alleen voor de gebruikers buiten onze instellingen om belangrijk, want het Bioportaal is toegankelijk voor iedereen, maar ook voor gebruikers vanuit onze organisatie. Het is bijvoorbeeld het CRS of het Brahms, geen mogelijkheid om objecten of collecties op basis van geografische gegevens te selecteren. In het Bioportaal kan dit wel. Als een interne onderzoeker, of collectiebeheerder dus specifiek alleen objecten wil van bijvoorbeeld Scandinavische zeevogels, dan moet dit handmatig worden uitgezocht; een proces dat veel tijd kost. Gewenste nauwkeurigheid georeference data Het is bij het georeferencen van de collectie wel belangrijk dat er een bepaalde nauwkeurigheid word meegegeven aan de collecties, of dat er in ieder geval geen schijn-nauwkeurigheid optreedt. Interne onderzoekers hebben inzicht in meer data en meer achtergrond kennis, ook over georeference data. Gebruikers van buiten de instelling, voornamelijk ons publieke bezoekers, hebben dit niet, waardoor er snel schijnnaauwkeurigheid kan optreden. Op dit moment, zitten er grote verschillen tussen de collecties, en dat is op zich niet erg, zo lang dit in het CRS en het Brahms maar wel aangegeven word. In het Bioportaal kan dan uiteindelijk ook worden aangegeven hoe nauwkeurig de georeference data is, of een onderverdeling worden gemaakt op basis van de nauwkeurigheid van coördinaten. Zo kan ook het Bioportaal worden gebruikt voor onderzoek en collectiebeheer.	Probleem Waarde Waarde Metadata Probleem Metadata
--	---

Wikipedia in Residence, Naturalis 06-10-2015	
Transcriptie interview	Code label
Gebruik huidige geografische data Naturalis is begin 2015 een project gestart om een deel van de collectie afbeeldingen en andere multimedia, zoals filmpjes, 3d modellen of geluidsopnames, via Wikimedia toegankelijk te maken voor het bredere publiek. Naast de multimedia bestanden, doneren we ook de bijbehorende metadata van de soorten die hierin worden afgebeeld, zoals de taxonomische gegevens en de geografische beschrijvingen van de verspreiding van de soort. Wanneer er coördinaten aanwezig zijn bij een object, wordt dit ook meegegeven bij de metadata die geleverd worden aan Wikipedia. Dit geldt voor alle pagina's op Wikipedia met een geografische plaatsaanduiding. Voor multimedia worden de coördinaten meegegeven van de locatie waar deze is gemaakt/opgenomen. Wanneer deze data ontbreekt bij de eigenaar van de pagina kunnen gebruikers dit wel voor doen. Om de coördinaten te generen zijn er twee verschillende tools beschikbaar: Tool 1: de Geografische Locatie Bepaler is geschikt voor het aanpassen van reeds aangegeven locaties. Wanneer je via een Wikipedia pagina op deze tool komt, wordt de locatie op een	Waarde

Google Earth kaart getoond. De gebruiker kan deze gepinde locatie verslepen naar het juiste punt. Naast de kaart worden de bijbehorende coördinaten getoond en de code gegeven, die in het sjabloon van Wikipedia kan worden opgenomen.	
<u>Tool 2:</u> de Coördinates- Tool is ook een Google Maps tool, waarmee je simpelweg een punt kan prikken op een kaart die overeenkomt met de locatie. Hieronder wordt dat een regel gegeven met hierin de coördinaten, die in het sjabloon van Wikipedia geplakt kan worden.	
Waarde van huidige geografische data Voor het Wikipedia project plaatsen we alleen geografische metadata van de objecten als deze lengte en breedte- graden bevatten. Tekstuele locatiebeschrijvingen worden hiervoor niet gebruikt. Ik maak de coördinaten dan wel compatible met de standaard die op het Wikipedia sjabloon word gehanteerd.	Waarde
Gebruik toekomstig georeference data Als er meer coördinaten beschikbaar zijn, zouden deze ook meegegeven kunnen worden aan de objecten die op Wikimedia worden geplaatst. Het project loopt echter maar tot het einde van dit jaar, dus het is voor mij onbekend, wanneer en hoe dat wordt voortgezet.	Waarde
Belang georeference data Het mooie aan het project van Wikipedia is dat onze collectie data voor het publiek, in de breedste zin van het woord, beschikbaar komt. Gebruikers hoeven niet maar naar onze eigen websites om toch gebruik te kunnen maken van de kennis van Naturalis. Geografische data is een belangrijk component van Wikipedia. Alle encyclopedie pagina's die een geografisch component kunnen bevatten, dus dat is meer dan alleen steden, zouden ook coördinaten moeten krijgen. Het voordeel van coördinaten is namelijk dat dit uniform is en voor iedereen begrijpelijk.	Waarde Waarde
Gewenste nauwkeurigheid georeference data Voor sommige soorten laat ik de geografische metadata tot op een bepaald niveau achterwege, zo laat ik bijvoorbeeld bij vogels die op de lijst voor bedreigde diersoorten staan, de precieze locaties weg. Ik geef hierbij alleen het land en stad aan waar de vogels bijvoorbeeld in buurt leeft. Zo kies ik dus bewust voor een grote onnauwkeurigheid van de geografische metadata, om te voorkomen dat de dieren nog verder met uitsterven worden bedreigd.	Discussie Metadata
Voor alle onderdelen van een Wikipedia pagina zijn zogenaamde sjablonen beschikbaar. In het sjabloon voor coördinaten wordt er gevraagd om deze in graden (60-tallig stelsel) aan te geven, of om decimale graden te gebruiken. Hiervoor vragen ze 8 parameters verplicht in te vullen. Deze verwijzen naar de graden, minuten en seconden van zowel de lengte- als de breedtegraad. Daarnaast kan je een type meegeven, waar het coördinaat naar verwijst, zoals 'waterbody', 'island', 'river', 'forrest', 'landmark', etc. Dit geeft enige indicatie van de nauwkeurigheid. Wanneer het type 'land' is gekozen zal de nauwkeurigheid namelijk groter zijn, dan wanneer er voor 'city' is gekozen. Dit is alleen niet direct af te lezen en vraagt dus wel om eigen interpretatie.	Metadata Metadata

Naturalis, onderzoek 08-10-2015	
Transcriptie interview	Code label
Belang georeference data Het toevoegen van coördinaten aan de collectie maakt het veel bruikbaar voor onderzoek. Je kunt het combineren met ecologische en klimatologische gegevens. Dit zijn onderwerpen waar nu veel onderzoek naar wordt gedaan. Georeferenzen kost alleen heel veel tijd en is een zeer arbeidsintensieve klus. Dit wordt nog eens extra moeilijk door allerlei vormen van ruis in de collectie, zoals historische locatiebeschrijvingen, ontbrekende geografische informatie, offset getallen, etc. Hierdoor	Waarde Probleem

moet je vaak afwegingen maken tussen verschillende zaken, zoals prioritering en nauwkeurigheid, bij het georeferencen van een collectie. Een afweging die je moet maken, is bijvoorbeeld of je eerst gebieden georeferenced waar weinig geografische data van is, data arme gebieden, zoals tropische gebieden of gebieden die juist heel data rijk zijn. Van de data arme gebieden zijn vaak slechte locatiebeschrijvingen bekend, die historische of lokale namen bevatten, die niet in de standaard gazetteers te vinden zijn. Van datarijke gebieden, zoals Nederland, is juist heel veel geografische data van bekend, tot op adresniveau. Voor mij ligt de voorkeur voor het georeferencen van data arme gebieden. Hierbij levert het georeference data namelijk veel sneller een veel bruikbaarere collectie op. Het is dan misschien moeilijker om te georeferencen, door het gebrek aan geografische data, maar het maakt de collectie, zelf met minder nauwkeurige coördinaten, veel bruikbaarder voor onderzoek op. Daarnaast moet je ook de afweging maken of je op soort niveau of op gazetteer niveau gaat georeferencen. Idealiter wordt er hier gekozen voor soort niveau, maar het meest tijd/kosten efficiënt is om te georeferencen op gazetteer niveau. Hiermee kun je sneller grote stappen vooruit nemen. Dit heb ik ook voor mijn onderzoek gedaan. Gewenste nauwkeurigheid georeference data In mijn onderzoek heb ik iets van 50 km als nauwkeurigheid gekozen, maar ik deed een onderzoek naar een groter gebied, voor veel soorten onderzoek zal 10 km het meest ideaal zijn. Toch moet hierbij wel worden genoemd, dat dit natuurlijk zeer afhankelijk is van de geografische kenmerken van de gebieden. Zo kan in een berggebied 10 km nauwkeurigheid veel te weinig zijn. Hier kan 10 km namelijk het verschil tussen een dal of een bergtop zijn en dat kan voor ecologische omstandigheden zeer veel uitmaken, en dus ook op de aanwezigheid van soorten. Van sommige gebieden is het gewoon heel lastig om een goede nauwkeurigheid te krijgen, zoals van tropische gebieden. Het is dan slimmer om eerst het gros te georeferencen op een nauwkeurigheid van 10 km dan 20 jaar te besteden aan het prefect krijgen van bepaalde gebieden, terwijl hier ondertussen al zoveel weer is veranderd in biodiversiteit en op ecologisch gebied. In hoeverre is de data in het CRS en BRAHMS bruikbaar voor het georeferencing? In het BRAHSM is er al een redelijk percentage van de soortgegevens voorzien van georeference data. Hoe dit in het CRS staat, weet ik niet goed. Het is in ieder geval wel belangrijk dat er gewoon zo snel mogelijk voor het grote deel coördinaten bijkomen. Open Refine is een middel om veel van de fouten of moeilijkheden in locatiebeschrijvingen, zoals spelfouten ed, te filteren. Daarmee verhoog je het percentage dat georeferenced kan worden. Daarnaast moet je altijd uitgaan van een foutmarge van 10-20%. Ik vind het een belangrijkere vraag hoe lang je moet doorgaan en hoeveel tijd je moet besteden aan het georeferencen van alles in de collectie. Soms moet je accepteren dat sommige locatiebeschrijvingen gewoon niet te georeferencen is en een keuze maken waar de prioriteit ligt. Wat mij betreft ligt dat bij collecties die voor onderzoeksdoeleinden de meeste baat hebben bij coördinaten.	Discussie Discussie Wens/eis Discussie Discussie
--	---

Bijlage 3: Codering

interviews CETA5 instellingen

CETAF, Royal Botanical Garden and Botanical Museum, Berlin 29-09-2015	
Transcriptie interview	Code label
In your email you mentioned the StanDAPherp project for georeferencing botanical data. Can you tell me more on that project? On herbarium sheets data, like plant name, species, collection site, collector, etc. can be found, mostly on labels glued on the sheet. The data is thus visible on images taken of the herbarium sheets during the digitization process, but is not available as machine readable metadata. The StanDAPherb project is developing a way to extract the information on labels automatically and to replace the very labor- intensive manual process that is used in most institutions at this moment. Which methods or tools did you use for georeferencing? Thus far we have been looking up some information about georeferencing methods and tools, but have not found exactly what we want. BioGeomancer, seemed like the perfect tool, but is unfortunately no longer available. So now we are aiming towards Geolocate, because this one would be easily incorporated in the workflow we are creating. Can you describe the process, or the application of the tool?? The StanDAP-Herb Project is developing a standard process for (semi-) automatic detection of meta-data on Herbarium specimens. The herbarium sheets we use have already been digitalized, so pictures or scans of the sheets are already available and the data is thus visible on images taken of the specimen. Object detection software detects objects such as labels or barcodes on the scans of the sheets and classifies them in different categories. Text objects are transformed into structured information with the use of text mining algorithms. Some of the herbarium sheets labels are very old and only contain handwritten data, for this we attempt to use author identification and handwriting recognition to classify data on the labels. The aim of the project is to create a workflow to extract this type of data and link different tools/ methods to support this process in our web services, so that other institutions can use it, when it is completed. Have you tested the workflow yet, and what where the results? At this moment we don't have any results that we can share, as the project will take 3 years, and we are only in the first year. We also don't have any georeference data yet, but we would ultimately really like to enrich the extracted data with geocodes, because it can add so much value to the data. What type of georeference tools or methods are you looking for? We want to connect different tools or methods to this workflow that can be used to complete the data extraction and enrichment. So, we do not want to include georeferencing methods in our tool, but link it in the web services of our workflow. Off course georeferencing is part of this process, because some herbarium sheets contain geographic descriptions on the labels, and this is also extracted. The goal of this extracting is to eventually enrich the data. For this enrichment step, georeferencing is a vital part, and we are very interested in exploring this. Is the documentation of the tool available as open source? The workflow and web services are hosted at GitHub, and available as open source, but at this moment we are only in the startup of the project, so there are no test results available yet. As soon as we have any, I will be happy to share them with you.	Tool Beschikbaarheid Tool Beschikbaarheid Tool Beschikbaarheid
CETAF, GBIF- The Global Biodiversity Information Facility 30-09-2015	

Transcriptie interview	Code label
Als je op de webportal van GBIF een dataset bekijkt kan je zien welke delen van deze collectie georeference data bevatten en of hier fouten in voor komen. Hoe komen jullie aan deze data? GBIF heeft in een backbone achter de web service een aantal lijsten (databases) hangen ter controle van de gegevens uit e datasets die instellingen publiceren. Voor taxonomische data gebruiken we bijvoorbeeld de <i>Catalogue of Life</i>. Datasets die in ons portaal worden gepubliceerd worden tegen deze gegevens uit de backbone gecontroleerd. Voor geografische gegevens hebben we bijvoorbeeld een lijst met alle landen en bijbehorende landgrenzen (in coördinaten) in de backbone hangen. Door datasets te checken op basis van onze backbone gegevens, kunnen we “mismatches” vinden en markeren. Kun je voorbeelden geven van de mismatches die jullie kunnen detecteren? Als wij dus een set binnenkrijgen waarvan bekend is dat deze objecten uit Nederland bevat word er een “mismatch” gegeven voor objecten die locatiebeschrijvingen of coördinaat gegevens hebben die niet overeenkomen met de landgrenzen van Nederland uit onze backbone. Nog een voorbeeld is: object gegevens van bijvoorbeeld planten die locatiebeschrijvingen hebben die op basis van onze backbone worden gelinkt aan locaties midden in de oceaan zijn bijvoorbeeld ook mismatches. Tot slot controleren we ook of coördinaten en locatiebeschrijvingen die in een dataset staan wel overeen komen met de coördinaten en landen die in onze backbone staat. Zo kan ook worden gecontroleerd, of coördinaten misschien zijn afgerond, of er plus en min is omgedraaid of welke geografische datum is gebruikt. Worden de gevonden “mismatches” ook aangepast? We doen maar gedeeltelijk verrijking van data en voornamelijk controle ervan. Verrijking vindt eigenlijk alleen plaats als een object coördinaten heeft, die overeenkomen met onze backbone, dan word aan de locatiebeschrijving het bijbehorende land of gebied toegevoegd. Echt georeferenzen doen we niet het is meer een controle check voor geografische data in een dataset. Deze controle check wordt wel voor alle datasets die via ons portaal toegankelijk zijn uitgevoerd. Je gaf in de mail al aan dat er soms foutmeldingen tussenzitten. Kan je hier meer over vertellen? Er is natuurlijk altijd een risico dat er fouten in een dataset zitten, maar de controle hiervan ligt bij de beherende instellingen. Er is eigenlijk bijna altijd 1-20% onjuist in een dataset, dat is gewoon gegeven. Wij linken alleen verschillende databases aan elkaar en geven aan waar de fouten zitten. Een instelling kan dit dus wel gebruiken, bij het verbeteren van de dataset. En onder het kopje ‘verbatim data’, is altijd de originele data van een dataset te zien, zoals die door de beherende instelling is aangeleverd. Wat is er mogelijk met deze resultaten die jullie check oplevert? Via onze web services kan je in principe exports maken van deze controlechecks, waarin uitgebreid word gerapporteerd over de kwaliteit van de datasets en de fouten die hierin zitten. Daarnaast hebben we sinds kort ook een extensie van Google Chrome, die dergelijke informatie van GBIF beter visualiseert en bijvoorbeeld weergeeft hoe groot de set is, hoeveel fouten erin zitten, etc. Op deze manieren kunnen instellingen het dus gebruiken als een extra beveiligingscheck voor de controle van de datakwaliteit. Hoe moet deze data worden opgeslagen om voor jullie bruikbaar te zijn? Je moet goed kijken naar de standaarden die worden gebruikt. Wij gebruiken DwC omdat deze het eenvoudigst is, en maar 8 a 9 velden heeft. Welke je gebruikt ligt aan de instelling, maar er is geen enkele groep die dit allemaal invult, terwijl het heel waardevol is. Er is zo veel meer vast te leggen dan alleen latitude en longitude. Als dit allemaal goed is	Datacleaning Datacleaning Datacleaning Beschikbaarheid Metadata Datacleaning Metadata

vastgelegd maakt het niet zo uit, wat de nauwkeurigheid van een coördinaat is. Als het is aangegeven, kan de gebruiker zelf bepalen of het bruikbaar is voor bepaalde onderzoeken. Als je namelijk een object heb met een bepaalde nauwkeurigheid kan dit meerdere redenen hebben; er kan bijvoorbeeld data missen, het kan komen door de tool die is gebruikt, of het is misschien expres gedaan. Het is veel bruikbaarder voor onderzoek, als dit soort data bekend is. 'Geografic precision in meters' in de belangrijkste om dit vast te leggen. Velden als 'Footprint' ed kunnen dit met polygonen visueel maken, alleen kunnen wij dit niet weergeven op GBIF. Weet u waar ik meer informatie kan vinden over de datasets in de backbone van GBIF? Op het hoofdkantoor in Kopenhagen werken mensen die je meer informatie over de backbone en bijbehorende algoritmes kunnen geven.	Nauwkeurigheid Metadata
---	---------------------------------------

CETAF, The Royal Botanical Garden of Edinburgh 06-10-2015	
Transcriptie interview	Code label
You said in your email that you were interested in the Google Maps method. Why this method specifically? We have been tackling georeferencing from the other end, than you did at Naturalis. In collaboration with the Informatics Department of The University of Edinburgh a student (Clare Llewellyn) developed a georeferencing tool that is more manual than the method you used, but can give really accurate results. Because the tool requires more manual work, it is quite time consuming, and so we wanted to see what other methods there are, that could do bigger sets of objects at once. Our tool is more suitable for small amounts of objects. If you compare it to the Google Geocoding API method, our tool could work great for object that have more difficult locality description, or for checking objects that already have georeference codes, or even to improve the original georeferencing codes that an institutions collected with tools like Google Geocoding API. Can you describe how the tool u created works? The tool is called SPECimap (website), and it is basically open source, but we are still working on some little bugs at the moment. As soon as these are fixed, we will release the tool open source completely. SPECimap takes certain field from the database and uses text strings from these fields for text mining on a known gazetteer. Based on this it plots the localities that were found with text mining, on a map. The tool works with textual locality descriptions as well as with known coordinates. The map is based on Google maps, but is combined with different historical maps (old survey maps) and the UK National Grid reference. In the tool you have the possibility to fade out these different maps or view them layered on top of each other. This way you can see the changes in coastal arrears or city, and replace the pinned locality more accurately. For example: <i>You have an object that was found in 1750 on the outer corner of a city, if you pin this object on the current border of that city it will likely be less accurate or placed even wrong, because most cities have grown in the past 250 years. So be checking the map placed behind Google Maps, from 1750, you can see where the border of that city was in the time the object was collected and place the object very accurately.</i> You can plot the localities from your database that have textual localities. But it also works with known coordinates. The tool will give back a georeference code (coordinates) that matches with pinned location of the historical map with the names and coordinates of the modern map Google Maps).	Tool Datacleaning Tool Beschikbaarheid Tool Databron Metadata

Does the tool deliver the results that you expected and how accurate are they? The tool delivers very accurate results, mostly with an accuracy of about 5 km, but because it requires manual work, it is very time-consuming. So for georeferencing large collections at once, it is to manual. How do you store the results from the tool? For us the biggest question with the storage of georeference data is the how to make accuracy of the georeference code accessible for everyone. Though our tool delivers very accurate georeference data and we do store this in our database, we do not make the accuracy calculation public on our website. We do not know the best ways to do this, but we know it is very important for the users of our data.	Nauwkeurighe id Nauwkeurighe id
---	---

CETAF, The Natural History Museum, London 03-11-2015	
Transcriptie interview	Code label
You mentioned that you use the Google Geocoding API for georeferencing you collection. Why did you use this tool? We did the georeferencing part of the iCollections project with the Google Geocoding API, mainly because of some limitations. We had to come up with a simple and fast way to find really detailed geographic coordinates. The problem for us was the scale, especially considering the limited resources. The Lepidoptera collection contains about 1,5 million objects, and that is only one part of our entire collection. At this moment there is no method or tool that gives us the results we want, with the variables of our collection and the resources we have. This has to do with the quality of our data, not the available tools for georeferencing. None of the existing tools could give us the results we wanted, with the poor data quality of the collection here. We therefore used the Google API as a standard and not per se as a tool. We know that not all results are 100% right but it gives the same type results, for all object records. We can describe the Google API as a method, in our collection database under the name of 'resource' and later update this with the help of a tool that might give better results. Until that time though, we have useful georeference data that is gathered in unified way. I found a presentation online, from Adrian Hine, and saw some screenshots of your own designed interface; can you tell me more about that? The interface itself was built around Microsoft Access forms, but we only created it to make the process go faster and make it more easy to use. For us the concept of how you georeference is more important than the small things like an interface, you build as an extra layer, to make it easier. In the technical report from the Automated Georeferencing Meeting you also wrote you had done some trials with the BioGeomancer Tool. Where there any differences between this tool and the Google API? We also tried the BioGeomancer, but it was not noticeably better for us and our collection. With the BioGeomancer we got allot of partial hits, instead of one full match for the records. Which meant that we still had to do allot of manual work, to select the right location; which still took allot of time. So did the Google API deliver less partial hits?	Tool Tool Metadata Tool Tool

With the Google API it was much easier to select the right location, from the different partial hits that came back in the tool. But this again, depended mainly on our data, because with the Lepidoptera project we looked at the whole label, and not just the locality. So for some objects we immediately knew that there was only one right matches between the partial hits. For example during this project we used the collection of British and Irish moths and butterflies. So if an object was found in London, and one of the hits was [London, US], we knew this could not be the right, since the entire collection comes from the UK. Can you describe how you apply the Google Geocoding API? In our collections we have very many site variants, which differ in accuracy, description way, spelling, etc. So we had to rationalize to get more accurate representations of these sites, without duplicating efforts. We did this by linking all site variants that refer to the same place to a representative site master. This leaves us with only 45 % of the total number of locality descriptions in the collection, and can therefore save a lot of time. After creating the site master, the tool can georeference these localities based on the Google Geocoding API. In the tool, the locality is plotted on a map and can be viewed both visually and in text. Then the right georeference match is selected, by choosing from the partial hits, if there are any. The interface automatically fills in the metadata field we want to store in the databases and selects The Google API as the georeference method. Does the tool deliver the results that you expected? We georeference individual locality descriptions, but because we use site-masters georeferencing one site, could mean georeferencing 500 object records. We are very interested in georeferencing in larger batches, but with our data quality we haven't found a way to do so. We had hoped to georeference objects in an automated way and find coordinates for about 60 – 70 % of the locality descriptions, but we only found about 12%. This is because of the quality of our data, not the Google Geocoding API.	Datascleaning Datascleaning Tool Metadata Results
--	--

CETAF, The Royal Museum of Central Africa, Tervuren 15-10-2015	
Transcriptie interview	Code label
While the NHM and Naturalis mainly presented the tools and algorithms used in the back-end, the RMCA also showed many applications and visualization tools. How much experience do you have with georeferencing? I started working here in 2005, but was involved with the museum since 2002. Since then we have been working on georeferencing projects continually. Both our Biology and Geology department have experience with georeference data from some GIS related activities and projects that we have done. We have a vast collection of maps and aerial photographs that have been digitized and georeferenced to a large extent. Can you tell me more about the projects you did? In the last few years we have been working on EDIT, in collaboration with the BGBM Berlin. Although the project ended a few years ago, in 2011, we still work on the maintenance of the tools on this platform. The workgroups of this project are now incorporated with the CETAF-ISTC. The things that we developed here, we want to compare with projects from other institutions and see where we can help each other, to start new projects. Does the EDIT-Platform refer to a tool that specifically focuses on converting textual localities into coordinates? Edit Cybertaxonomy Platform is a platform that contains many different tools, of which a few are concerned with geographic data and georeferencing like the EDIT map viewer and the	Tool Beschikbaarheid Tool

<i>EDIT MapRest</i>. I am not sure if there are any, and if so, which, tools can be used to convert textual localities to georeference codes. Many of these tools were widely used, by an international community of users, but since the projects ended, the user group slimmed down a lot.	
You mentioned more projects, can you tell me about those?	Tool
Another project that we did was in collaboration with the University of Berkeley and VertNET, using the BioGeomancer tool and the Georeferencing for dummies table, for expanding the database of HerpNET. In collaboration with 8 other institutions we tried a few prototypes for georeferencing herpetology collections.	
For this project the RMCA donated its entire documented amphibian collection to GBIF and HerpNET. This collection contains ca 3000 unique localities from DRC, Burundi, Cameroon, Ivory Coast and Togo, from specimen from with linked locality names, station numbers, and collector coordinates. The collection was georeferenced according to the HerpNET protocol, which includes the checking of coordinates and the addition of radius (maximum error radius). With the use of GIS the localities were georeferenced and the extent and error were calculated. During this project we also gave some training in collaboration with GBIF, on the use of these methods.	Methode
Besides these projects we have also been working in the EU project EU-BON, which is part of the GEO-BON organization (Global Earth observation Biodiversity Observation Network). In this project we sharing our knowledge and experience from the Edit platform project with other participants and are advising on the possibility of delivering data to GBIF.	Tool
Thirdly we did the SYNTHESYS Itinerary project where we aimed to detect patterns in travel itineraries with georeference data collected during the collection event. Many of the georeference data available in our collection has been collected during expeditions and surveys, but this data is usually no longer useable because of historical names or because the data was poorly documented. This project was aimed at georeferencing this “difficult” data. I don’t know if this algorithm is still available.	Beschikbaarheid
To create and use these kinds of applications and tools on the EDIT Platform, you would need coordinates. Do you have parts in the collection that are georeferenced?	
We have done multiple projects for which we needed coordinates, but only had textual localities available. So over the past few years we have collected quite a lot of georeferenced data. A lot of this georeferencing was done during the project with the university of Berkeley and VertNET, so most coordinates are created with the help of BioGeomancer, but this tool is no longer available. The SYNTHESYS project was used more for checking georeferencing results in datasets.	Beschikbaarheid
	Datacleaning

CETAF, Naturalis Biodiversity Center Niels Raes (Onderzoeker Botanie) 08-10-2015	
Transcriptie interview	Code label
<i>NOTE: Vervolg op het interviewverslag in bijlage 3</i> Waarom heb je je eigen tool gebouwd en niet een bestaande tool gebruikt? Er zijn wel enkele tools en pipelines beschikbaar, zoals Biovel, maar deze zijn over het algemeen zeer gegeneraliseerd, waardoor het bruikbaar wordt voor alle gebruikers. Wanneer	

je echter zulke specifieke eisen heb als ik, is dit gewoon niet goed genoeg. Ik heb toen specifiek voor mijn onderzoek een tool gemaakt, geprogrammeerd in R, die op basis van de Google Geocoding API specimen op gazetteer niveau kan georeferencen. Het grootste nadeel van Google Geocoding API is dat je maximaal 2500 object per dag mag georeferencen. Voor collecties zo groot als Naturalis is dat een zeer langdurig traject. Ik heb op een gegeven moment zelf een student gehad die alle koffieshops en internet cafés in Leiden langsging om per IP adres die 2500 objecten te uploaden. Zo hebben we in 1 dag toch heel Thailand kunnen georeferencen. Hoe gaat jouw tool precies in zijn werking? Eerst word de data uit GBIF gehaald, soms ook het CRS en deze data wordt automatisch in grote batches via R scripts aangeboden aan de Google API. Hierin worden de specimen gekoppeld aan coördinaten, die je vervolgens zelf kan accepteren of niet. Je kunt als extra ook eerst de coördinaten nog vergelijken met landsgrenzen. In de tool heb ik ook hoge resolutie satelliet beelden, oude expeditie kaarten en SRTM digital elevation data verwerkt, omdat het soms om oude collecties ging waarvan de vindplaats inmiddels erg is veranderd. Zo waren er specimen gevonden in de bochten van rivieren, terwijl nu, bijna 200 jaar later, deze bochten niet meer bestaan. Met behulp van de satellietbeelden, kon ik de restanten van deze bochten toch nog vinden en dus beter georeferencen. De georeference data die als resultaat uit mijn onderzoek kwamen heb ik in het BRAHMS geplaatst met in de Memo box een opmerking over de bron waarmee de data is verkregen. De nauwkeurigheid heb ik niet aangegeven in het CRS. Deze werd in de scripts wel berekend op basis van de output van de Google Geocoding API, maar wanneer de output bijvoorbeeld 52.03 is, kan dit heel veel dingen betekenen. Zo kan dit aanduiden dat er gegevens missen waardoor het niet nauwkeuriger kon, dat de coördinaat is afgerond door de tool, of dat het daadwerkelijk op de plek 52.03 ligt. Door hier zelf een nauwkeurigheid aan te hangen, creëer je een zeer subjectief waardeoordeel. Die voor de ene onderzoekers wel gelijk is, maar voor andere misschien niet.	Tool Tool Databron Nauwkeurigheid Metadata Nauwkeurigheid
---	---

Bijlage 4: Resultaten

deskresearch deelvraag 4

Tools		
Geolocate	• Rios, N.E. & Bart, H.L. (z.d.). <i>User Manual, GEOLocate Georeferencing Software</i>. University of Tulane, Bell Chase. Geraadpleegd op 17 november 2015, via https://www.museum.tulane.edu/geolocate/standalone/manual_ver2_0.pdf • Tulane University (2015). <i>GEOLocate - Software for Georeferencing Natural History Data</i>. Geraadpleegd op 2 november 2015, via http://www.museum.tulane.edu/geolocate/	Geolocate Tulane georeferencing software
Google geocoding API	• Enipedia TU Delft (2015). <i>OpenRefine Tutorial</i>. Geraadpleegd op 17 november 2015, via http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial • Google Developers (2015). <i>The Google Maps Geocoding API</i>. Geraadpleegd op 17 november 2015, via https://developers.google.com/maps/documentation/geocoding/intro#GeocodingRequests • Hine, A. (2015). <i>iCollections, Mass Digitisation of British & Irish Lepidoptera</i>. Geraadpleegd op 3 november 2015, via http://slideplayer.com/slide/3541251/ • Natural History Museum Londen (2015). <i>iCollections</i>. Geraadpleegd op 3 november 2015, via http://www.nhm.ac.uk/our-science/our-work/digital-museum/i-collections.html • Wikipedia (2015a). <i>R, programming language</i>. Geraadpleegd op 17 November 2015, via https://en.wikipedia.org/wiki/R_(programming_language) • Arkel, B. van & Creuwels, J. & Leusen, J. van & Schnörr, S. (2014). <i>Eindrapport FCD- Pilot Georeferencing: productiematig georeferencen</i>. Intern document Naturalis, Leiden • Hine, A. & Penn, M. (2014). <i>Georeferencing at the NHM. Presentation</i>. Naturalis Biodiversity center, Leiden • Schnörr, S. (2014). <i>Georeferencing Methode Niels Raes</i>. Summery. Intern document, Geraadpleegd op 7 november 2015	Open refine Google Icollections, geocoding API, Google API, georeferencing google, programming language R geocoding service Datum geocoding API
SPECimap	• Llewellyn, C.A. & Hasten, e. & Grover, C. (2011). <i>Georeferencing Botanical Data using Text Analysis Tools</i>. TDWG 2011 Conference, Geraadpleegd op 15 oktober 2015, via http://www.researchgate.net/publication/282701713_Georeferencing_botanical_data_using_text_analysis_tools • Llewellyn, C. & Grover, C. & Oberlander, J. & Haston, E. (2012). Enhancing the Curation of Botanical Data Using Text Analysis Tools. <i>Theory and Practice of Digital Libraries, volume 7489 of the series Lecture Notes in Computer Science</i>, pp 480-485 • Grover, C., Givon, S., Tobin, R., Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. <i>Proceedings of the Sixth International Conference on Language Resources and Evaluation</i>	SPECimap, RBGE, Llewellyn, Botanical data, Botanical specimen, text analysis, Geotagger, Geoparser, Georeferencing, textmining
BioGeomancer	Chapman, A.D. & Wicczorek, J. (2006). <i>Guide to best practices for georeferencing</i> . Copenhagen: Global Biodiversity Information Facility	Biogeomancer, Chapman, georeferencing tool
StanDAP- Herb	GitHub (2011). <i>StanDAP-Herb WebGenesis Web Services</i> . Geraadpleegd op 2 november 2015 via	STandapherb

	https://github.com/gentisaliu/StanDAP_Herb_WebGenesis_WebServices	
Georeferencing Calculator	Wieczorek, J., D. Bloom. 2011. <i>Georeferencing Calculator Manual v2</i> . Geraadpleegd op 18 november 2015, via http://manisnet.org/GeoreferencingCalculatorManualv2.html	Georeferencing Calculator, Wieczorek, HerpNET
Mitch	- Tilburg University. (2005). <i>MITCH, Mining for Information in Texts from Cultural Heritage</i>. Geraadpleegd op 9 november 2015, via http://ilk.uvt.nl/mitch/ - Erp, M. van & Hensel, R. & Ceolin. Davide. Meij, M. van der(2014). <i>Georeferencing Animal Specimen Datasets</i>. Transactions in GIS, 12/2014: 19(4)	MITCH, Georeferencing, Naturalis, van Erp, CATCH
TGN	- The J. Paul Getty Trust (2015). <i>Getty Thesaurus of Geographic Names Online</i>. Geraadpleegd op 18 november 2015, via http://www.getty.edu/research/tools/vocabularies/tgn/about.html - Arkel, B. van & Creuwels, J. & Leusen, J. van & Schnörr, S. (2014). <i>Eindrapport FCD- Pilot Georeferencing: productiematig georeferencen</i>. Internal document Naturalis, Leiden	TGN, Getty Thesaurus, Geographic names, thesaurus geodata
Methodes		
MaNIS/ HerpNET	- HerpNET (2007). <i>Georeferencing – Step by Step & Online Resources</i>. PowerPoint Presentation available at: http://HerpNET.org/HerpNET/documents/StepbyStep.ppt. - Spencer, C. & Yamamoto, K. & Fang, J. & Constable, H. & Koo, M. & Wieczorek, J. (2005) <i>Georeferencing for Dummies: An Elaboration of the MaNIS/HerpNET/ORNIS Guidelines</i>. Berkeley: University of California. Available online at www.HerpNET.org/HerpNET/documents/georeffordummy.xls - Wieczorek, J. (2001)"<i>MaNIS/HerpNET/ORNIS Georeferencing Guidelines</i>. Geraadpleegd op 2 november 2015 via http://manisnet.org/GeorefGuide.html	HerpNET, Wieczorek, Manis, georeferencing for dummies, georeferencing guidelines,
Guide to best practices, Chapman	Chapman, A.D. & Wieczorek, J. (2006). <i>Guide to best practices for georeferencing</i>. Copenhagen: Global Biodiversity Information Facility	Chapman, Biogeomancer, Georeferencing best practice
Databronnen		
Geonames	Geonames (z.d.). <i>GeoNames, data resources</i> . Geraadpleegd op 19 november 2015, via http://www.geonames.org/data-sources.html	Geonames
TGN	The J. Paul Getty Trust (2015). <i>Getty Thesaurus of Geographic Names Online</i> . Geraadpleegd op 18 november 2015, via http://www.getty.edu/research/tools/vocabularies/tgn/about.html	TGN, Getty Thesaurus, geographic names, thesaurus geodata
Google Maps	Wikipedia (2015b). <i>Google Maps</i> . Geraadpleegd op 18 november 2015, via https://en.wikipedia.org/wiki/Google_Maps	Google Maps, Maps API, Google Maps datum,
The Fuzzy G	Christian Kohlschütter (2003). <i>The JRC Fuzzy Gazetteer</i> . Geraadpleegd op 18 november 2015, via http://isodp.hof-university.de/fuzzyg/query/	fuzzy G, data, Fuzzy gazetteer, Isodp, fuzzy geographic
Histogram	- Erfgoed & Locatie (2015). <i>Demo-versie Historische Geocoder online</i>. Geraadpleegd op 18 november 2015, via http://erfgoedenlocatie.nl/2015/04/demo-versie-historische-geocoder-online/ - Histogram (z.d.) <i>Histogram: geocoding places of the past</i>. Geraadpleegd op 18 november 2015, via http://histograph.io/	Histogram, historische geocoder, E&L,
Datacleaning		
Visual check	Arkel, B. van & Creuwels, J. & Leusen, J. van & Schnörr, S. (2014). <i>Eindrapport</i>	Geocoding API,

Google	<i>FCD- Pilot Georeferencing: productiematig georeferencen</i>. Internal document Naturalis, Leiden • Enipedia TU Delft (2015). <i>OpenRefine Tutorial</i>. Geraadpleegd op 17 november 2015, via http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial • Google Developers (2015). <i>The Google Maps Geocoding API</i>. Geraadpleegd op 17 november 2015, via https://developers.google.com/maps/documentation/geocoding/intro#GeocodingRequests	Google API, georeferencing google, geocoding service, Open refine Google
GBIF	<i>NOTE: Alle informatie voor dit datacleaning middel is gehaald uit het interview met Cees Hoff en een gesprek met Wouter Addink (GBIF vertegenwoordiger binnen Naturalis)</i>	
Species link	• Spicieslink (2008). <i>Specieslink, Information about the Project</i>. Geraadpleegd op 19 november 2015, via http://splink.cria.org.br/project?criaLANG=en	Specieslink cria, brazil georeferencing,
itinerary project	• BioCase (2005). <i>SYNTHESYS NA-D 3.7- Providing itinerary related datasets and tools (for integration, visualisation and quality check)</i>. Geraadpleegd op 2 november 2015, via http://www.biocase.org/products/geo_services/itineraries/ • Meganck, B. & Meirte, D. & Mergen, P. & Theeten, F. (2006). <i>Milestone report for SYNTHESYS Network Activity</i>. Geraadpleegd op 2 november 2015, via http://www.biocase.org/products/geo_services/itineraries/files/SYNTHESYS_milestonereportmay31.pdf	Itinerary georeferencing, text analyses RMCA, Theeten, Mergen, itinerary project
Plaatsnamen Standaardiseren	• Engelse, M. den (z.d.). <i>Handleiding Standaardiseertool</i>. Geraadpleegd op 19 november 2015, via http://erfgeo.nl/wat-hoe/standaardiseertool.html • Erfgoed & Locatie (2015). <i>Demo-versie Historische Geocoder online</i>. Geraadpleegd op 18 november 2015, via http://erfgoedenlocatie.nl/2015/04/demo-versie-historische-geocoder-online/	Plaatsnamen standaardiseren, E&L, standaardiseer tool
Crowdsourcing	• Blaauboer, R. 2(2014). <i>Nichesourcing helpt het Rijksmuseum collecties in kaart brengen</i>. Geraadpleegd op donderdag 20 november 2014 http://www.frankwatching.com/archive/2014/03/20/nichesourcing-helpt-het-rijksmuseum-collecties-kaart-brengen/ • Blaauboer, R. 1 (2014) <i>Digitaal erfgoed: zo past crowdsourcing in je bedrijfsmodel</i>. Geraadpleegd op donderdag 20 november 2015, http://www.frankwatching.com/archive/2014/11/03/digitaal-erfgoed-zo-past-crowdsourcing-in-je-bedrijfsmodel/ • Oosterman, J. & Nottamkandath, A. & Dijkshoorn, C. & Bozzon, A. & Houben, G.J. & Aroyo, L. (2014). <i>Crowdsourcing knowledge-intensive tasks in cultural heritage</i>. (online publication) WebSci 2014: 267-268, • Ridge, M. (2007). <i>Sharing authorship and authority: user generated content and the cultural heritage sector</i>. (online publicatie) 2007 Web Adept - UK Museums on the Web, Museums Computer Group conference, • Willemsen, J. (2011). <i>Erfgoed en Crowdsourcing</i>. Geraadpleegd op vrijdag 20 november 2015, http://www.frankwatching.com/archive/2011/02/28/erfgoed-en-crowdsourcing/ • Fleurbaay, E. & Eveleigh, A. (2012) <i>Crowdsourcing: prone to error?</i> (Online publication) International Council on Archives Conference 2012	Crowdsourcing, citizen science, crowdsourcing biodiversity, scientific crowdsourcing, nichesourcing

Bijlage 5: Tabellen

pakketselectie

Tools

	Geolocate	Google Geocoding API	Georeferencing Calculator	SPECimap
Accessibility				
Online	Yes	No	Yes, if updated	Yes, once released
Web service	Yes	Yes	Yes, if updated	Unknown
Offline	Yes	No	Yes, if updated	Unknown
Free	Yes	2500 free requests a day (for educational institutions 100,000 requests a day)	Yes, if updated	Yes, once released
File formats				
Accepted import formats	XML, CSV, or TXT files	HTTPS request via Open Refine.	Not applicable, since it's not possible to upload batches	Individual objects go directly from collection database
Output format	Delimited text files , CSV and specially formatted XML file	JSON or XML	Not applicable	Unknown
Standard compliance	MaNIS/HerpNET/OR NIS Guidelines	No	5 elements align with Darwin Core	No
Batches	Yes, but very difficult.	Yes	No	No
Individual	Yes	Yes	yes	Yes
Results				
Accuracy	Uncertainty in meters and Polygon error	Based on the outer southwest and northeast corners of the smallest found geographic unit (given in coordinates)	Maximum error distance based on the MaNIS/HerpNET/ORNIS Guidelines	Unknown
Datum	WGS84	WGS 84	WGS84	Unknown
Precision	Exact to the nearest second	6 decimals	Multiple possibilities: nearest second, nearest degree, etc.	Exact to the nearest second
Data elements	Latitude, Longitude, Uncertainty in meters and Polygon error	Address components, Formatted address (textual and coordinates), Viewport (accuracy)	Decimal Latitude, Decimal Longitude, Coordinate uncertainty in meters, Geodetic datum, verbatim coordinate system, extent, maximum error distance, distance units, distance precession, coordinate precision	Latitude and longitude, country, ALTM, National Grid reference, region, habitat
Manually adjustable	Yes	No	No	Yes
Geographic focus area	North America, Canada, Mexico	Worldwide	Worldwide	Worldwide, but old maps and grid reference, only for England.

Efficiency	Requires allot of manual input	Only requires input to import/adjust data in Open refine. Tool itself takes 1 hour per 1000 objects	Requires allot of manual input, since it only works for individual objects	Requires allot of manual input, objects are replaced individually and manually
Reliability	Yes	Yes	Yes	Yes
Transparency code/algorithm	Sample source code and documentation: http://www.museum.tulane.edu/geolocate/developers/default.html	Documentation: https://developers.google.com/maps/documentation/geocoding/intro	No	Will be released open source.
Data quality				
Accepted languages	English	54 languages	Spanish, Portuguese, French English, Dutch	English, the rest is unknown
Historic names	No	Unknown	No	Yes, manually
Miss spellings	No	Built in ranking mechanism that finds well known and populated places based on original request	No	Yes, manually
Other elements	No	No	Yes	Yes manually

Methodes

Als het aankomt op de geanalyseerde georeferenzen methodes, was het zonder pakketselectie al duidelijk dat er geen van de twee 'beter' zou zijn dan de andere. Daarom is er geen pakketselectie tabel voor de methodes gemaakt.

Databronnen

	Geonames	TGN	Google Maps	The Fuzzy G	Histogram
Accessibility					
Online website	Yes	Yes	Yes	Yes	Yes
Web service	Yes	Yes	yes	no	Yes
Available as download	Yes	Yes	No	Unknown	No
Free	Yes	Yes	Yes	Yes	Yes
Geographic focus area	Worldwide (better with populated places)	Worldwide	Worldwide	Worldwide	Netherlands
Results					
Data type results	Textual	Text with associative relationships, hierarchical and equivalence terms	Textual and visual	Textual	Textual and visual
Coordinates	Yes	Yes	Yes	Yes	No
Fuzzy data	Doesn't handle misspellings or historic data	Knows terms in different languages and historic names	With the ranking mechanism it can handle minor misspellings	Very good	Knows historic names from different periods of time

Datacleaning middelen

	Visual Check Google API	GBIF	Specieslink	Plaatsnamen standaardiseren	Open refine
Accessibility					
Online	Yes	Yes	Yes	Yes	Yes
Offline	No	No	No	No	
Free	Yes	Yes	Yes	Yes	Yes
File formats					
Accepted import formats	HTTP request with Open Refine or similar	DarwinCore Archive files, offers converter for MSEXcel files	Unknown	CSV.	TSV, CSV , Excel XML, RDF as XML, JSON, Google Spreadsheets, RDF N3 triples
Output formats	JSON or XML	No output possible	Unknown	CSV.	Same as import
Batches	Creating strings in batches, but viewing individually	Yes	Yes	Yes	Yes
Individual	Yes	Yo	Yes	No	Yes
Results					
Type of results	New (more accurate) coordinates	Highlighted mismatches in a dataset	Highlighted error	4 tabs; 1) standardized localities 2) localities with multiple matches. 3) Localities with no results, 4) intended for localities that can't be standardized by expert or tool.	Various, depends on the formulas and features used. examples: structured data, new formats, parsed data, unique localities, georeference http strings
Manual input required	Yes, string has to be adjusted manually for a better results	Yes, GBIF does not adjust, just highlight	Yes, does not adjust, just highlights errors.	Yes, selecting the field that contains the place names and what type of locality it contains	Yes, but only for selecting/creating columns when importing data
Compliance with standards	No	Yes, complies with Darwin Core	Yes, Complies with Darwin Core	No	Depends on Google Geocoding API and column in dataset
Geographic focus area	Worldwide	Worldwide	Brazil	Netherlands	Not applicable. Depends on tool that is used.
Efficiency	Requires manual input to adjust each string	Low, does not correct errors, only highlights them	Low, does not correct errors, only highlights them	High, only requires manual input to create csv file.	High, requires little manual input
Reliability	Low, adjustment depends on manual input of a user	High, backbone of GBIF is consistent	Unknown	High, data sources (Geonames and TGN) are consistent	Only uses imported data, so it depends on the tool that is used and the original data.
Transparency	High, strings are based on Google geocoding API out.	Low, documentation not available. It is unknown how the data check works exactly.	Unknown	High, documentation http://erfgeo.nl/wat-hoe/standaardiseertool.html	High, is open source: https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users

Bijlage 6: Plan van Aanpak



Plan van aanpak

Het (semi-) automatisch georeferencen van primaire
soortgegevens

Afstudeerstage IDM

Josine Blom, 1208619

Organisatie: Naturalis Biodiversity Center

Bedrijfsmentor: Marian van der Meij

Haagse Hogeschool

Informatie dienstverlening en informatie management

Klaas Jan Mollema (Examinator 1)

Jochem Mollema (Examinator 2)

Darwinweg 2
Postbus 9517
2300 RA Leiden

T 071 751 91 02
josine.blom@naturalis.nl
www.naturalis.nl

Naturalis
Biodiversity
Center

Inhoud

Inleiding	1
1. Achtergronden	2
1.1 Organisatie	2
1.2 Geschiedenis	3
1.3 Stakeholders	4
2. Projectresultaat	5
2.1 Doelstelling	5
2.2 Probleemstelling	5
2.2.1 Hoofdvraag	5
2.2.2 Deelvragen	5
2.3 Projectresultaat	6
3. Projectactiviteiten	8
4. Projectgrenzen en randvoorwaarden	13
4.1 Projectgrenzen	13
4.2 Randvoorwaarden	13
5. Tussenresultaten	14
6. Kwaliteitsbewaking	15
6.1 Kwaliteitsnormen	15
6.2 Controle	15
7. Projectorganisatie	17
7.1 Contactgegevens	17
7.2 Communicatie	17
8. Planning	19
9. Kosten & baten	20
9.1 Kosten	20
9.2 Baten	20
10. Risico's	21
10.1 Interne risico's	21
10.2 Externe risico's	21
Literatuurlijst	22
Bijlage 1: Verslag opstartgesprek opdrachtgever	23
Bijlage 2: Afstudeerplan	25

Inleiding

Dit plan van aanpak dient als projectplan/ basisdocument van het onderzoeksproject Automated Georeferencing. Dit project betreft de afstudeerstage van Josine Blom (opdrachtnemer), voor de opleiding Informatiedienstverlening- en management, aan de Haagse Hogeschool. Naturalis Biodiversity Center fungeert tijdens dit project als opdrachtgever.

Het plan van aanpak is opgesteld tijdens de definitiefase van het project om daarin de 5 beheers aspecten; tijd, geld, kwaliteit, informatie en organisatie van dit project duidelijk te beschrijven voor alle betrokkenen (Gritt, 2011). Voor de opzet van het plan van aanpak is gebruik gemaakt van de methode en indeling van Roel Gritt (2011) en de theorie over onderzoeksmethoden van Nel Verhoeven (2010).

Het plan van aanpak dient daarnaast ook als startdocument voor het afstudeerdossier van de opdrachtnemer. Voorafgaande aan de afstudeerstage is er een afstudeerplan (Bijlage 3) gemaakt, waarop dit plan van aanpak is gebaseerd. In het plan van aanpak worden echter alleen de projectonderdelen besproken die direct betrekking hebben op de uitvoering van het project en de projectorganisatie. Bijkomende onderdelen en activiteiten die alleen voor de Haagse Hogeschool bestemd zijn, staan hierin niet vermeld.

In de rest van dit document is als eerste alle benodigde achtergrond informatie over de organisatie, het project en alle belanghebbende te vinden. De beoogde projectresultaten en alle activiteiten die hiertoe moeten leiden worden in de hoofdstukken hierop volgend behandeld. Daarna wordt duidelijk aangegeven welke grenzen dit project heeft en wat de randvoorwaarden zijn. Vervolgens worden de tussenresultaten/mijlpalen en de kwaliteitsbewaking toegelicht. In de hoofdstukken 7 en 8 komen de praktische zaken rondom de projectorganisatie en planning aan bod en tot slot worden de kosten en baten en de mogelijke risico's van dit project besproken.

1. Achtergronden

Het project 'Automated georeferencing' betreft het afstudeerproject van Josine Blom als laatste onderdeel van haar opleiding Informatie-dienstverlening en Informatiemanagement aan de Haagse Hogeschool. De officiële opdrachtomschrijving is als volgt:

"Een best practice opstellen voor het (semi) automatisch georeferencing van de gedigitaliseerde collectie data van Naturalis."

Vanuit het stage bedrijf (Naturalis) zal Marian van der Meij optreden als bedrijfsbegeleider en daarmee ook opdrachtgever van het project zijn. De rol van opdrachtnemer en voornaamste uitvoerende project lid van dit project zal Josine Blom vervullen.

Om toekomstige projectleden, de organisatie waarin het project wordt uitgevoerd en de betrokken school van de opdrachtnemer over het project te informeren wordt een beschrijving van de achtergronden 'de projectomgeving' gegeven. Aan de hand van deze beschrijving dienen niet- ingewijden zich een beeld kunnen vormen van het project en de organisatie waarin het project wordt uitgevoerd (Gritt, 2011).

1.1 Organisatie



Naturalis Biodiversity Center is een stichting die biodiversiteit wil beschrijven, begrijpen en verklaren. Voor het welzijn van de mensen en het voortbestaan van de aardse natuur (Naturalis Biodiversity Center, 2012). Dit doen ze door:

- Collecties: ze verzamelen, bewaren en bestuderen dieren, planten en stenen. De collectie telt inmiddels 37 miljoen objecten en is daarmee van wereldformaat.
- Onderzoek: met wetenschappelijk onderzoek dragen ze bij aan de mondiale inspanning om de biodiversiteit te reconstrueren en de interactie tussen soorten te begrijpen;
- Museum: gebruikmakend van de collectie vertellen ze het verhaal van de natuur in zijn natuurhistorische en cultuurhistorische context, aansprekend voor verschillende leeftijdsgroepen en voeding gevend aan verwondering;
- in een open dialoog kennis te delen over de natuur, taxonomie en biodiversiteit, die bruikbaar is voor de samenleving en een duurzame economie.
- Onderwijs: Als internationaal kenniscentrum en academische werkplaats voor de universiteiten van Amsterdam, Leiden en Wageningen. Zo dragen ze bij aan het opleiden van een nieuwe generatie wetenschappers en academische professionals op het gebied van systematiek en biodiversiteit.
- Ten slotte levert Naturalis een bijdrage aan bedrijven, overheden en maatschappelijke organisaties:
 - o Ze helpen de douane wanneer er lastig te identificeren dieren of planten aan de grens verschijnen.
 - o Bij het certificeren van hout verlenen ze assistentie om te bepalen of een partij uit gekweekt of wild hardhout bestaat.
 - o Ze ondersteunen natuurhobbyverenigingen bij het ontwikkelen en beheren van websites met natuurinformatie.
 - o Ze adviseren de juweliersbranche bij het bepalen van echtheid van edelstenen.

Het Nederlands Centrum voor Biodiversiteit Naturalis werd gevormd op 28 januari 2010 uit een samenwerking tussen het Zoölogisch Museum Amsterdam (onderdeel van de Universiteit van Amsterdam), het Nationaal Herbarium Nederland (onderdeel van de Universiteit Leiden, de Universiteit Utrecht en de Wageningen University) en het Nationaal Natuurhistorisch Museum in Leiden. Het hoofddoel van deze

samenwerking was het integreren van de collecties van de genoemde partners tot één nationale collectie van 37 miljoen planten, dieren, fossielen, gesteenten en mineralen. Het instituut behoort daarmee qua omvang mondiaal tot de top vijf in het natuurhistorische domein (Naturalis Biodiversity Center, 2015b).

De huidige organisatie is ingedeeld in vier sectoren, Onderzoek en Onderwijs, Collectie, Publiek en Markt en ICT. ICT richt zich onder andere op het digitaliseren van collectie- en andere biodiversiteitsinformatie, het beschikbaar en toepasbaar maken ervan en het koppelen ervan aan andere digitale bronnen. De ICT sector van Naturalis bevat een afdeling informatiemanagement bestaande uit twee teams, applicatiebeheer en datamanagement. Deze afstudeerstage zal worden uitgevoerd bij de sector ICT en dan specifiek het team datamanagement (Naturalis Biodiversity Center, 2015a).

1.2 Geschiedenis

Tijdens het FES collectiedigitaliseringsproject zijn ruim 7 miljoen collectieobjecten gedigitaliseerd. Hierbij is de informatie die bij de objecten hoort (zoals taxonomie, verzameldatum en verzamelplaats) in twee centrale databases gezet; één voor zoölogische en geologische data, het *Collectie Registratie Systeem* (CRS) en één voor botanische data, de *Botanical Research And Herbarium Management System* (Brahms). Deze data worden ontsloten via het dataportal van GBIF (Global Biodiversity Information Facility; <http://www.gbif.org>) en de Naturalis BioPortal (bioporal.naturalis.nl).

Eén van de belangrijkste toepassingen van deze data is het gebruik ervan in wetenschappelijk onderzoek. Naast de taxonomische data is de geografische verzameldata van de objecten hiervoor erg belangrijk. Met de geografische data van objecten is het namelijk mogelijk om bijvoorbeeld vroegere en huidige verspreidingsdata van soorten te mappen en te analyseren en ze te combineren met ruimtelijke data van andere soorten en disciplines (bv. meteorologie, geologie, geografie, sociale en medische wetenschappen). Naast fundamenteel onderzoek kunnen hiermee ook vraagstukken over wereldwijde kwesties als natuurbehoud en klimaatverandering onderzocht worden (Stuurgroep FES Collectie Digitalisering, 2015).

Alle geografische data van objecten is conform de internationale standaarden (DarwinCore en ABCD EFG) genoteerd in het CRS en het Brahms. Bij een groot deel staat deze data echter alleen als tekst in de registratiesystemen genoteerd en bevat veelal oude namen (bijvoorbeeld Pruisen, Joegoslavië). Deze data is voor de wetenschap beter bruikbaar als de objectrecords verrijkt worden met de bijbehorende coördinaatgegevens. Om de kwaliteit en de toepasbaarheid van de Naturalis dataset te vergroten is het daarom wenselijk dat de tekstuele geografische objectinformatie van (delen van) de door Naturalis beheerde objectrecords op een gestandaardiseerde manier worden verrijkt met de overeenkomstige geografische coördinaten (Stuurgroep FES Collectie Digitalisering, 2015). Het koppelen van een locatie van een ruimtelijk object, in dit geval biodiverse objecten, aan een geografisch referentiesysteem (zoals coördinaten), zodat deze op een digitale kaart kunnen worden geplaatst, wordt georeferencing genoemd. Het is een belangrijk kenmerk van georeferenzen dat de locatie als een unieke identifier optreedt, er is maar één locatie mogelijk waarvoor de georeference data kan optreden als referentiepunt.

Naturalis heeft nog weinig ervaring met het toepassen van georeferencing op grote hoeveelheden objectrecords. Productiematig georeferenzen is een relatief nieuw concept; er zijn geen standaarden of kant en klare methoden beschikbaar. Daarom is in 2013/2014 een pilot project georeferencing uitgevoerd waarin is gekeken wat de mogelijkheden zijn om de vindplaatsgegevens van objecten op een grotendeels geautomatiseerde manier te voorzien van coördinaten.

Behalve Naturalis hebben veel andere natuurhistorische/ botanische instituten de behoefte om georeferentiedata aan hun objectrecords toe te voegen omdat ze een duidelijke meerwaarde geven aan de collectie. Naturalis Biodiversity Center werkt daarom samen met internationale partners van de CETAF om de kennis over georeferencing te delen. Gezamenlijk heeft deze onderzoeksgroep besloten dat een

internationale best practice kan bijdragen aan een internationale standaardmethode voor (semi) automatisch georefereren en een infrastructuur voor alle natuurhistorische collectie binnen de EU voor de toekomst. Toepassing van de best practice zou moeten leiden tot een verrijking van natuurhistorische databases met betrouwbare en vergelijkbare georeferentiedata. Daarnaast zou een dergelijke standaardmethode en infrastructuur zouden de usability en de kwaliteit van digitale natuurhistorische collecties kunnen vergroten (Stuurgroep FES Collectie Digitalisering, 2015).

Tijdens de Automated georeferencing meeting in 2014, georganiseerd door Naturalis, zijn enkele vergelijkbare projecten van het Natural History Museum Londen (NHM) en het Royal Museum for Central Africa Tervuren (RMCA), besproken. Daarnaast is gekeken of er mogelijkheden zijn om op dit gebied in de toekomst samen te werken. Uit de meeting kwam naar voren dat NHM, RMCA en Naturalis vergelijkbare en complementaire competenties hebben en gestandaardiseerde GIS tools gebruiken die goed op elkaar aansluiten.

Vanuit de resultaten van het FCD Pilot Project Georeferencing en de projecten van NHM Londen en het RMCA Tervuren is gezamenlijk behoefte aan een onderzoek naar hoe de verschillende methodes elkaar kunnen aanvullen en gezamenlijk één compleet proces kunnen vormen. Om dit te realiseren moet er verder onderzocht worden hoe de verschillende methodes elkaar kunnen aanvullen en gezamenlijk één compleet proces kunnen vormen.

1.3 Stakeholders



Bij dit project zullen er naast de opdrachtgever en de opdrachtnemer nog 2 partijen betrokken die zijn die belang hebben bij de resultaten van het project. Daarmee telt het project de volgende vier stakeholders:

- Josine Blom heeft als opdrachtnemer en student veel belang bij het tijdig en kwalitatief goed uitvoeren van het project. Als eerste betekent dit voor haar dat ze het laatste verplichte onderdeel van haar studie kan afronden, en mag afstuderen. Daarnaast doet ze tijdens de uitvoering van project werkervaring en nieuwe kennis op.
Naturalis is het bedrijf waarin het project zich afspeelt en fungeert hierbij als opdrachtgever. Met de resultaten van het project hebben ze een best practice methode om haar de collectie verrijken met nieuwe data, waardoor onderzoekers van Naturalis vroegere en huidige verspreidingsdata van soorten mappen en analyseren en ze combineren met ruimtelijke data van andere soorten en disciplines. Naast fundamenteel onderzoek kunnen hiermee ook vraagstukken over wereldwijde kwesties als natuurbehoud en klimaatverandering onderzocht worden. Daarnaast doen ook zij ervaring en nieuwe kennis op over automatisch georeferenzen door middel van de begeleiding van de opdrachtnemer. Hiermee kunnen zij in de toekomst hun eigen collectie verder uitbreiden of nieuwe projecten/onderzoeken starten naar dit onderwerp.
- De CETAF-ISTC, Consortium of European Taxonomic Facilities, is een Europees netwerk tussen enkele natuurkundige museums, natuurhistorische museums, botanische tuinen en biodiversiteit-onderzoeksinstituten. Naturalis is een van deze aangesloten instellingen. Binnen CETAF is er een subcommissie ISTC, Information Science & Technology Commission, die zich richt op samenwerking tussen de ICT afdelingen/sectoren van de verschillende instituten die aangesloten zijn bij CETAF (CETAF, z.d. a). Gezamenlijk delen ze expertises en voeren ze onderzoeken uit naar een aantal uiteenlopende onderwerpen. Automated georeferencing is hier één van. Tijdens de Automated Georeferencing meeting in december 2014 georganiseerd door Naturalis, zijn enkele pilotproject van aangesloten instellingen besproken. Namens de CETAF heeft Naturalis de lead gekregen in het opstellen van een internationale best practice voor het semi automatisch georeferenzen van collecties biodiversiteits materiaal. Het eindresultaat van dit project (de best practice) zal dus niet alle voordelen opleveren voor Naturalis maar ook voor de andere aangesloten instellingen van de

CETAF. Deze internationale best practice kan bijdragen aan een internationale standaardmethode voor (semi) automatische georeferencing en een infrastructuur voor alle natuurhistorische collecties binnen de EU voor de toekomst. Een dergelijke standaardmethode en infrastructuur zouden de usability en de kwaliteit van digitale natuurhistorische collecties kunnen vergroten.

Gebruikers van de metadata van de collectie

De laatste groep stakeholders zijn de gebruikers van de collectie van Naturalis. De collectie wordt door veel verschillende doelgroepen gebruikt, zoals onderzoekers, universiteiten en amateurwetenschappers. Voor al deze gebruikers leidt het project uiteindelijk tot een verbetering van de kwaliteit van de collectie, die ten goede komt aan de usability en de toepasbaarheid ervan. De gebruikers van de collectie krijgen dus meer data en kwalitatief gezien, meer bruikbare data. Hierdoor kan dit gebruikt worden voor meer en gedetailleerder onderzoek.

2. Projectresultaat

Uit de projectopdracht komt een te beoogd projectresultaat (eindproduct) voort dat bijdraagt aan de doelstelling van de opdrachtgever. Het project Automated georeferencing betreft grotendeels een onderzoek. Naast de te bereiken projectresultaten en de bijbehorende doelstelling zal dus ook een probleemstelling, en bijbehorende hoofd- en deelvragen worden opgesteld en besproken in dit hoofdstuk.

2.1 Doelstelling



Het doel van dit project - het maken van een handleiding met methoden en aanbevelingen (best practice) om grote hoeveelheden objectrecords zo geautomatiseerd mogelijk (productiematig) te koppelen aan geografische referentiepunten (zoals coördinaten), waarbij de locatie als een unieke identifier optreedt. Hierbij is maar één locatie mogelijk waarvoor de georeference data kan optreden als referentiepunt, zodat deze op een digitale kaart kunnen worden geplaatst.

Dit advies moet gericht zijn op hoe Naturalis hun collectie met geografische verzameldata kan verrijken met overeenkomstige coördinaten (NBC, 2014). De gedigitaliseerde collecties van Naturalis kunnen dan worden voorzien van nieuwe geografische verzameldata, waardoor onderzoekers van Naturalis vroegere en huidige verspreidingsdata van soorten analyseren en combineren met ruimtelijke data van andere disciplines.

De best practice is ook herkenbaar en toepasbaar voor de andere geïnteresseerde instellingen van de CETAF, waardoor het uiteindelijk kan bijdragen aan een internationale standaardmethode voor (semi) automatische georeferencing en een infrastructuur voor alle natuurhistorische collecties binnen de EU voor de toekomst. Deze bevat handvaten en richtlijnen voor het volledige proces dat bij productiematig georeferenzen komt kijken en dat zo hoog mogelijk scoort op de *Principles for Best Practice for Georeferencing Biological Species Data* (Chapman, 2006). Zodat de usability en kwaliteit van digitale natuurhistorische collecties wordt vergroot. Hierbij is het voor alle leden van de CETAF belangrijk dat een kwaliteitscontrole wordt uitgevoerd voordat de nieuwe data wordt geaccepteerd. Deze controle moet geleid of gecontroleerd worden door een wetenschappelijk expert en dat de tools die worden gebruikt voor de controle buiten het gebruikte collectieregistratiesysteem moet plaatsvinden, omdat een kwaliteitscontrole een specifieke taak met een eigen workflow is (NBC, 2015-1).

2.2 Probleemstelling



“Wat moet je als instelling doen om een collectie geografische verzameldata met tekstuele locatieaanduidingen op een semi automatische manier te verrijken met de overeenkomstige geografische coördinaten op verschillende niveaus/ in verschillende stappen van nauwkeurigheid?”

Deelvragen

1. Wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?
2. Op welke verschillende niveaus willen de doelgroepen van Naturalis de geografische data beschikbaar hebben?
3. Welke andere instellingen hebben projecten uitgevoerd om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?
 - b. Wat is de betrouwbaarheid/ accuraatheid van de gebruikte methode/tools?

- c. In welke mate was handmatige correctie nodig om de betrouwbaarheid van de georeferencing data te verbeteren?
- d. Welke middelen zijn er gebruikt om deze verbeteringen door te voeren?
- 4. Welke (semi) automatische methodes/ tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten?
- 5. Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren?
- 6. Welke semi automatische programma's zijn het meest geschikt voor de beschreven collecties?
- 7. Hoe sluiten deze methodes/ programma's het beste op elkaar aan.
- 8. Hoe kunnen de methodes/ programma's het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?

2.3 Projectresultaat

Het eindproduct van deze afstudeerstage is een advies voor Naturalis gericht op hoe zij hun collectie met geografische verzameldata kunnen verrijken met overeenkomstige coördinaten. Dit advies kan door Naturalis worden gebruikt bij het uitvoeren van een project om de collectiedata te verrijken en eventuele andere projecten met collecties die in de toekomst nog gedigitaliseerd worden (Blom, 2015). Tijdens dit project zal de opdrachtnemer de volgende (deel) producten opleveren:

- Plan van aanpak
- Onderzoeksverslag met hierin:
 - een probleemanalyse met overzicht van mogelijk te onderzoeken methode en technieken, een zoekwoordenlijst voor het literatuuronderzoek en een lijst met mogelijke te interviewen instellingen.
 - een overzicht/ beschrijving (inventarisatie) op van de collecties die moeten worden verrijkt met coördinaten, op basis waarvan de methoden en tools die later worden onderzocht kunnen worden beoordeeld.
 - een overzicht op van de informatiebehoefte van de gebruikers en medewerkers dat vertaald moet kunnen worden naar een lijst van wensen en eisen op.
 - Een pakketselectie
- De uiteindelijke best practice methode (advies)

Uit dit eindproduct kunnen uiteindelijk nog twee resultaten volgen:

- De gedigitaliseerde collecties van Naturalis kunnen worden voorzien van nieuwe geografische verzameldata, waardoor onderzoekers van Naturalis vroegere en huidige verspreidingsdata van soorten mappen en analyseren en ze combineren met ruimtelijke data van andere soorten en disciplines. Naast fundamenteel onderzoek kunnen hiermee ook vraagstukken over wereldwijde kwesties als natuurbehoud en klimaatverandering onderzocht worden.
- Daarnaast kan dit advies worden gebruikt als internationale best practice voor alle instellingen die aangesloten zijn bij CETAF. Deze internationale best practice kan bijdragen aan een internationale standaardmethode voor (semi) automatische georeferencing en een infrastructuur voor alle natuurhistorische collecties binnen de EU voor de toekomst. Een dergelijke standaardmethode en infrastructuur zouden de usability en de kwaliteit van digitale natuurhistorische collecties kunnen vergroten (Arkel et al, 2015).

3. Projectactiviteiten

Op basis van de probleemanalyse en resultaten van de oriëntatiefase is er voorafgaande aan het onderzoek een onderzoeksplan opgesteld. Deze bevat de te gebruiken onderzoeksmethode, bijbehorende deelproducten en een overzicht van mogelijk te gebruiken trefwoorden, tools en interviewpartners op basis van de probleemanalyse. Onder onderzoeksmethoden worden gekozen methoden verstaan die nodig zijn om alle benodigde gegevens te verzamelen en te analyseren.

Het onderzoek wordt gestart op basis de gegevens die zijn verzameld in de oriëntatiefase van het onderzoek. Deze zijn afkomstig uit verslagen van het FCD pilotproject van Naturalis, verslagen van de Automated Georeferencing Meeting in 2014. Deze bevatten mogelijk te onderzoeken methoden en tools voor productiematig georeferenzen en enkele CETAF partners die pilotprojecten hebben uitgevoerd of willen meewerken aan de interest group.

Om tot de projectresultaten, die zijn beschreven in het vorige hoofdstuk, te komen, zal een aantal activiteiten moeten worden uitgevoerd door de opdrachtnemer. De activiteiten en bijbehorende taken worden hieronder in de vorm van een stappenplan beschreven. De afstudeeropdracht bestaat uit 4 delen; oriëntatie, dataverzameling, gegevensanalyse en uitwerking. Bij elk deel staan hieronder de stappen beschreven met bijbehorende beroepstaken (onderzoeksmethode) en producten (Blom, 2015).

Het onderzoeksproject zal voornamelijk bestaan uit kwalitatief onderzoek (praktijkgericht), waarbij wordt gericht op het analyseren en aanpakken van problemen door het 'waarom' en 'hoe van het onderzoeksdomein 'te onderzoeken (Verhoeven, 2014).

Deel 1 oriëntatie

Stap 1: Plan van aanpak schrijven.

Als eerste stap van het onderzoek wordt er een plan van aanpak opgesteld, waarin nog specifieker dan in dit afstudeerplan de eisen en afspraken samen met de opdrachtgever worden vastgelegd en een planning wordt opgenomen.

Bijbehorend product: een plan van aanpak.

Stap 2: Aanvullende oriëntatie op het probleemdomein uitvoeren.

Het oriëntatiedeel van het onderzoek betreft literatuuronderzoek naar de probleemsituatie. Hierbij wordt specifiek gericht op het eerder afgeronde pilotproject dat Naturalis uitvoerde, verslagen van de Automated georeferencing pilot en informatie over vergelijkbare projecten van de GIS-experts subgroep, die Naturalis al intern beschikbaar heeft.

De probleemanalyse beschrijft de hoofdlijnen van het vraagstuk dat tot deze opdracht heeft geleid en de achtergrond hiervan. Het vormt uiteindelijk een hoofdstuk voor eind onderzoeksrapport en bevat ook een overzicht van mogelijk te onderzoeken methode en technieken en een lijst met mogelijke te interviewen instellingen

Bijbehorend product & beroepstaak: een probleemanalyse.

Deel 2 dataverzameling

Stap 3: Intern literatuur onderzoek doen naar de betreffende collecties uit voeren (inventarisatie).

Als eerste stap van het onderzoek wordt een inventarisatie gemaakt van alle gedigitaliseerde collecties van Naturalis die voorzien dienen te worden van coördinaten en de verscheidenheid van alle vastgelegde gegevens die betrekking hebben op de geografische verzameldata. De collecties kunnen namelijk metadata in tekstvorm bevatten, maar wellicht ook al coördinaten en andere formaten zoals de Rijksdriehoekcoördinaten (Blom, 2015). Het gaat hier om het definiëren en kwalificeren van entiteiten die een bepalende rol spelen bij de inrichting van het CRS zodat de nieuwe geografische data kan worden

opgeslagen. Hiervoor zal een datadictionary worden gemaakt waarin de gegevens die daarover geregistreerd moeten worden geanalyseerd, gedefinieerd en gespecificeerd ten behoeve van het CRS.

Het onderzoek naar de collecties in het webbased CRS zal worden uitgevoerd door middel van deskresearch, waarbij de collecties uit het CRS als het BRAHMS worden geanalyseerd en beschreven. De meeste datadictionary beschrijven de soorten metadata en de kenmerken daarvan (zoals lente, naam, tekst vorm etc.). In dit geval wordt er ook gekeken naar de daadwerkelijke metadata die is opgeslagen over de objecten, omdat deze per collectie kunnen verschillen.

Bijbehorend product en beroepstaak: een datadictionary. Dit is een overzicht/ beschrijving van de benodigde metadata elementen voor het verwerken en beheren van georeference data

Bijbehorende deelvraag:

1. Wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?

Stap 4: Interviews afnemen met de gebruikers van georeference data van Naturalis.

Er zullen interviews worden afgenomen met de gebruikers van de georeference data van de collecties, om vast te stellen op welke niveaus of wijze de huidige geografische verzameldata verrijkt moeten worden, zodat deze voor alle deze doelgroepen bruikbaar zijn. Op basis van de afgenomen interviews zal de behoefte aan kennis en informatie in kaart worden gebracht. Dit onderzoek levert een overzicht op van de informatiebehoefte van de gebruikers en medewerkers dat vertaald moet kunnen worden naar een lijst van wensen en eisen op, wat later word gebruikt voor de pakketselectie.

De interviews voor deze stap vinden intern plaats. Hierbij worden de volgende partijen gesproken:

- Collectie beheer: zij handelen de vragen van zowel interne als externe onderzoekers af en hebben informatie over de soorten vragen naar geografische informatie er zijn.
- Bioportaal: dit is een intern project van Naturalis dat de basisgegevens, inclusief illustratie, van volledige collectie van Naturalis ontsluit en toegankelijk maakt voor het publiek. Dit project werkt ook met de geografische collectiedata om de collecties op een kaart te tonen.
- Wikipediabeheer: dit is ook een intern project van de sector ICT waarbij collectiedata van Naturalis worden gekoppeld aan Wikipedia pagina's. Ook hierbij is er vraag naar geografische informatie.

De interviews zullen in ongestructureerde vorm worden afgenomen. Hierbij word alleen een topiclijst opgesteld van onderwerpen die aan bod moeten komen in het gesprek (Verhoeven, 2014). Deze vorm van interviewen is voor deze situatie het meest geschikt omdat de respondent hierbij een zo groot mogelijk input moet leveren en zijn behoefte aan informatie en ervaringen die zij hebben centraal staan.

Bijbehorend product en beroepstaak: een overzicht op van de informatiebehoefte van de gebruikers.

Bijhorende deelvraag:

2. Op welke verschillende niveaus willen de doelgroepen van Naturalis de geografische data beschikbaar hebben?

Stap 5: Interviews afnemen met CEAT-ISTC instellingen die eigen projecten hebben uitgevoerd.

Tijdens de Automated Georeferencing meeting zijn er kort twee pilotprojecten van (semi) automatische georefencing besproken van het NHM Londen en het RMCA Tervuren (NBC, 2014-1). Naast de overige twee projecten zijn er nog enkele instellingen die een pilotproject hebben uitgevoerd en-/ of onderzoek hebben gedaan. Verwacht wordt dat deze projecten elkaar overlappen en kunnen aanvullen.

In interviews worden met deze instellingen besproken welke methode de voorkeur heeft, of deze nog gebreken of knelpunten hebben en waar de overlappen zitten met eventuele ander methoden of tools. Via de opdrachtgever wordt er in eerste instantie een standaard mail verstuurd met hierin:

- Een korte toelichting op het project
- Een verzoek voor het afnemen van een interview
- Als hiervoor geen tijd is, of de bijgevoegde vragenlijst kan worden ingevuld

- Of er mogelijke onderzoeksrapporten, pilotverslagen en of eindevaluaties van de projecten die de informatie uit de interviews kunnen aanvullen (voornamelijk over technische aspecten) of

Er wordt in eerste instantie gericht op de 5 instellingen die al hebben aangegeven zelf georeference projecten te hebben uitgevoerd en te willen meewerken aan het project. Naturalis wordt hierbij ook meegenomen in het onderzoek.

- Natural History Museum (NHM), London, Groot Brittannië
- Koninklijk Museum voor Midden-Afrika, Tervuren, België
- Botanische Tuin en Botanisch Museum, Berlijn, Dahlen, Duitsland
- Natuur Historisch Museum, Wenen, Oostenrijk
- Naturalis Biodiversity Center, Leiden, Nederland
- The Royal Botanic Garden, Edinburgh, Schotland

Daarnaast wordt er een mail verstuurd naar de hele CETAF-ISTC mailinglijst met de vraag of er nog meer instellingen zijn die dergelijke projecten hebben uitgevoerd en ook willen meewerken aan een interview of vragenlijst. De interviews die worden afgenomen zullen hoogstwaarschijnlijk plaatsvinden via Skype, Facetime of een ander vergelijkbaar medium, omdat alle instellingen zich in het buitenland bevinden, en bezoeken in de doorlooperperiode van dit project niet haalbaar is.

De interviews die “live” worden afgenomen krijgen de vorm van half-gestructureerde interviews (Verhoeven, 2014). Het is belangrijk dat de vorm van de interviews en de resultaten vergelijkbaar voor de gegevensanalyse in deel 3 van het project. De vragenlijst die hiervoor wordt opgesteld, wordt gebaseerd op al verzamelde informatie van het Pilotproject van Naturalis, omdat deze een beeld geeft van de informatie die nodig is voor de gegevensanalyse, zoals missende informatie of vergelijkbare resultaten ed.

De interviews die via de vragenlijst uit de mail verlopen zijn gestructureerde interviews, op basis van enquêtelijsten. Het gebruik van standaardvragenlijsten voor alle instellingen kan de geloofwaardigheid en de betrouwbaarheid van de resultaten vergroten (Verhoeven, 2014). In de vragenlijsten worden grotendeels open vragen gesteld om de instellingen de ruimte te geven hun ervaringen te delen.

Bijbehorend product: Interviewverslagen en een overzicht van mogelijke methodes/tools.

Bijbehorende deelvraag:

3. Welke andere instellingen hebben projecten uitgevoerd om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?
 - a. Wat is de betrouwbaarheid/ accuraatheid van de gebruikte methode/tools?
 - b. In welke mate was handmatige correctie nodig om de betrouwbaarheid van de georeferencing data te verbeteren?
 - c. Welke middelen zijn er gebruikt om deze verbeteringen door te voeren?

Stap 6: Literatuuronderzoek doen naar mogelijke methodes/tools voor (semi) automatisch georeferencing, en middelen om de accuraatheid te verbeteren.

De laatste stap van het deel ‘dataverzameling’ bestaat uit deskresearch waarmee de verzamelde informatie van de interviews word aangevuld met verwijzingen naar, en een analyse van relevante informatie over de besproken methodes en tools. Zo word er een zo compleet mogelijk beeld te geven van het mogelijke aanbod van methodes voor productiematig georefereren.

Hierbij zal in eerste instantie worden gericht op de onderzoeksrapporten, pilotverslagen en of eindevaluaties van de projecten die in de vorige stap zijn verzameld. Voor de projecten waarvan niet een dergelijke bron is beschikbaar is zal er doormiddel van deskresearchmethodes worden gezocht naar aanvullende informatie. Tijdens de deskresearch zal voornamelijk worden gericht op secundaire literatuur (Verhoeven, 2014) met informatie over de beschikbaarheid en functionaliteiten van bepaalde methodes, en tools.

Bijbehorend product: Dit onderdeel levert een literatuur review op die samen met de verslagen van de interviews een onderzoeksrapport vormen

Bijbehorende deelvragen:

4. Welke semi automatische methodes/ tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten?
5. Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren?

Deel 3 gegevensanalyse

Stap 7: Vergelijking maken van de eisen die de gebruikers van collectie stellen en de bevindingen uit de literatuur en interviews.

Op basis van de interviewgegevens en de deskresearchanalyses wordt er een pakketselectie gemaakt. Hiervoor worden alle gevonden methodes en tools getoetst op basis van de wensen en eisen die worden gesteld aan de functionaliteit en de technische aspecten door de geïnterviewde instellingen en interne medewerkers. Hierbij zal ook rekening worden gehouden met de *Principles for Best Practice for Georeferencing Biological Species Data* (Chapman & Wieczorek, 2006), zodat de kwaliteit en betrouwbaarheid van de best practice zo hoog mogelijk is. De selectie zal worden gedaan door op systematische wijze alle gevonden applicaties en methode te scoren op de wensen en eisen die hieraan worden gesteld.

In een tabelvorm worden alle eisen, wensen, principes van Chapman ed. in een lage lijst worden opgesomd. Vervolgens worden hiernaast alle methodes en tools per onderdeel van het gehele georeferencing proces gelegd en op elke punt gescoord. Uiteindelijk kan dan worden berekend welke tools/methoden de meest geschikte zijn voor de best practice en hoe goed deze aansluiten bij de gebruikers.

Door deze selectie zo systematisch aan te pakken is de hele stap overzichtelijker. Daarnaast is de verwachting dat sommige methodes/tools op bepaalde punten elkaar overlappen en dat word op een schematische wijze ook beter inzichtelijk.

Bijbehorend product: een pakketselectie en een adviesrapport, inclusief onderzoeksresultaten

6. Bijbehorende deelvraag: Welke semi automatische methodes/programma's zijn het meest geschikt voor de beschreven collecties?

7. Hoe sluiten deze methodes/ programma's het beste op elkaar aan?

Deel 4: uitwerking

Stap 8: Opstellen van een best practice.

Op basis van de tabel uit de vorige stap kan een advies worden geformuleerd over de beste methoden voor Naturalis en de andere CETAF-ISTC instellingen om collecties op productiematige wijze te georefereren. De verwachting van de eerder CETAF meetings is dat er wellicht meerdere methodes zijn die aansluiten bij de wensen en eisen (NBC, 2014-1). De vraag is dus juist; "hoe kunnen deze methodes het beste ingezet worden en op elkaar aansluiten?". De best practice zal beschrijven welke en hoe de methodes/programma's het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten. Hierbij wordt een technische datadictionary opgesteld die een beschrijving geeft van de benodigde functionaliteiten van een collectieregistratiesysteem. De technische datadictionary beschrijft de wijze waarop de standaard attributen aangepast moeten worden om de georeference data te ondersteunen, en de manier waarop de applicatie ingericht moet worden (parametriseren).

Bijbehorend product: een internationale best practice.

Bijbehorende deelvraag:

8. Hoe kunnen de methodes/ programma's het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?

4. Projectgrenzen en randvoorwaarden

Om onduidelijkheden gedurende het project te voorkomen, worden in dit hoofdstuk de projectgrenzen en randvoorwaarden toegelicht. Duidelijke grenzen en andere afspraken kunnen de kans op eventuele onenigheid en risico's verminderen. Deze projectgrenzen en randvoorwaarden bieden gedurende het project houvast tijdens het uitvoeren van de onderzoeksactiviteiten.

4.1 Projectgrenzen



Het project is, zoals in hoofdstuk 3 te lezen is, opgebouwd uit vier fase, die elk is opgedeeld in een aantal stappen. Deze stappen bevatten op hun beurt de uit te voeren activiteiten en op te leveren producten.

Bij enkele stappen dient wat extra toelichting gegevens te worden waardoor duidelijk word wat niet binnen het betreffende project valt en waarvan de activiteiten dus niet door de opdrachtnemer zullen worden uitgevoerd. Dit project bouwt gedeeltelijk verder op de resultaten het eindrapport van de FCD-pilot georeferencing, maar er zullen wel enkele verschillen zijn met betrekking tot de scope van het project.

- Als eerste is het belangrijk aan te geven dat de opdrachtnemer onderzoek doet bij Naturalis en niet bij de CETAF. In de stappen uit het voorgaande hoofdstuk is te lezen dat de opdrachtnemer wel met enkele instellingen van de CETAF zal praten over hun ervaringen met automatisch georeferencen. Hierbij zal alleen gericht worden op de resultaten van deze projecten en de betrokken medewerkers hiervan. Er zal dus geen gebruikersonderzoek worden gedaan bij de CETAF.
- In het eerder uitgevoerde pilot project (FCD-pilot georeferencing), zijn de tests uitgevoerd met collectiedata uit het CRS. Collecties uit het Brahms zijn hierin niet opgenomen, omdat het niet mogelijk was deze data te importeren. In dit project wordt de gehele collectie van Naturalis gebruikt voor de datadictionary. Het is echter niet onderdeel van deze stage om Naturalis te adviseren over hoe de georeference data kan worden geïmporteerd in het BRAHMS. Dit betreft de functionaliteiten van het systeem, en is aan de betreffende afdeling om dit mogelijk te maken.
- In het FCD- pilot project zijn tests uitgevoerd met de onderzochte collecties en methodes voor automatisch georeferencen. De opdrachtnemer zal deze resultaten en de resultaten van andere pilots gebruiken tijdens het project en niet zelf dergelijke uitgebreide nieuwe tests uitvoeren.
- Dit project betreft een onderzoek en een advies. Een implementatie of uitvoering van het advies is een vervolgstap. Het eindproduct dient dus opgezet te zijn om direct te worden opgevolgd een van deze stappen, maar zullen niet tijdens de doorlooptijd van dit project worden ondernomen.

4.2 Randvoorwaarden



Om de projectactiviteiten te ondersteunen en mogelijk te maken dient er aan enkele randvoorwaarden te worden voldaan:

- De opdrachtnemer moet binnen de afdeling informatiemanagement een werkplek hebben met de benodigde materialen (computer).
- De opdrachtnemer moet beschikken over toegang tot de collecties om een goede beschrijving te kunnen maken van de huidige stand van zaken (probleemanalyse en datadictionary)
- De opdrachtnemer moet toegang hebben tot informatie over het eerder uitgevoerde pilotproject, zodat de opdrachtnemer geen tijd besteed aan al bekende zaken.
- De opdrachtnemer wordt eventueel geholpen bij het contactleggen met samenwerkingspartners uit de CETAF. Hierbij kan worden gedacht aan namen, emailadressen of eventuele mailinglisten.

5. Tussenresultaten

Het eindproduct van dit project is zoals eerder genoemd een onderzoeksverslag, met hierin een internationale best practice (advies) voor automatisch georeferencen van grote collecties. Hierin wordt in kleinere stappen naar toe gewerkt. Deze stappen sluiten elk zelf af met een tussenresultaat/ tussenproduct. In dit hoofdstuk worden deze tussenresultaten/- producten besproken. Elk van deze resultaten is gekoppeld aan een 'deadline', waardoor het meetbare mijlpalen worden gedurende de looptijd van het project.

a. **Probleemanalyse, week 4**

Specifiek gericht op het eerder afgeronde pilotproject dat Naturalis uitvoerde en andere informatie over vergelijkbare projecten die Naturalis al intern beschikbaar heeft. De probleemanalyse bevat daarnaast ook een overzicht van mogelijk te onderzoeken methode en technieken en een lijst met mogelijke te interviewen instellingen.

b. **Datadictionary, week 5**

Een overzicht/ beschrijving (inventarisatie) van de collecties die moeten worden verrijkt met coördinaten, op basis waarvan de methoden en tools die later worden onderzocht kunnen worden beoordeeld.

c. **Informatiebehoefte, week 8**

Een overzicht van de informatiebehoefte van de gebruikers en medewerkers van Naturalis en ervaringen van projecten van aangesloten instellingen van de CETAF, dat vertaald moet kunnen worden naar een lijst van wensen en eisen.

d. **Pakketselectie, week 12**

Een vergelijking van de eisen die de gebruikers van collecties stellen en de bevindingen uit de literatuur naar methodes en tools en interviews met de CETAF instellingen. Dit tussenproduct levert een scoresheet op waaruit de meeste geschikte methodes en of tools kunnen worden opgemaakt.

e. **Best practice methode (advies), week 16**

Een handleiding met methoden en aanbevelingen om grote hoeveelheden objectrecords zo geautomatiseerd mogelijk (productiematig) te georeferencen.

Deze producten worden uiteindelijk gezamenlijk opgeleverd in het eindproduct, als hoofdstuk of als bijlage van het onderzoeksverslag.

6. Kwaliteitsbewaking

Een belangrijk onderdeel van het project is de kwaliteitsbewaking. De opdrachtnemer voert het project uit namens de opdrachtgever en betreffende organisatie, die op haar beurt weer zorg draagt voor het project namens de CETAF. Het is belangrijk dat alle partijen uiteindelijk positief zijn over het uitgevoerde werk en de eindresultaten bruikbaar vinden voor het vooraf gestelde projectdoel.

6.1 Kwaliteitsnormen



Om te zorgen dat kwaliteit van de op te leveren producten goed is en aan de verwachtingen van de opdrachtgever voldoet zijn er een aantal afspraken vastgesteld die de opdrachtnemer kunnen ondersteunen.

- **Beroepstaken:** In hoofdstuk 3 is onder elke stap van dit project de bijbehorende beroepstaak genoemd. Dit zijn taken/competenties waaraan studenten van de opleiding Informatiedienstverlening en Informatiemanagement moeten voldoen voor de afronding van de opleiding. Aan elk van deze beroepstaken is een beschrijving, bijbehorende werkzaamheden en beoordelingscriteria gekoppeld. Gedurende dit project (stage) zal de opdrachtnemer moeten aantonen aan school dat zij aan deze beoordelingscriteria voldoet. Dit is dus een manier om de kwaliteit van de werkzaamheden en de producten te verzekeren. In bijlage 1 vind de uitwerkingen van deze beroepstaken.
- **Verantwoording methoden en technieken:** naast de beroepstaken heeft de opdrachtnemer gedurende de opleiding kennisgemaakt met enkele methoden en technieken. Het is aan de opdrachtnemer om tijdens dit project aan te tonen dat zij de juiste keuzes kan maken van geschikte methoden en deze kan verantwoorden. Wederom bied deze verantwoording en de controle hiervan een mogelijkheid om de kwaliteit van de op te leveren producten te garanderen.
- **Verantwoording van afwijking:** voorafgaande aan het onderzoek zijn de te gebruiken methoden, planning, stappen etc. vastgesteld in dit plan van aanpak. Wanneer hiervan wordt afgeweken, voor welke reden dan ook, dient deze afwijking te worden verantwoord in de discussie van het onderzoeksverslag. Daarnaast dient deze afwijking ook met de opdrachtgever (en eventueel examinatoren) te worden overlegd. Het plan van aanpak wordt goedgekeurd door zowel de opdrachtgever als de begeleider vanuit de Haagse Hogeschool.

6.2 Controle



Gedurende het project zullen er regelmatig voortgangsgesprekken plaats vinden, om te controleren of de kwaliteit voldoende is en of de planning nog wordt nageleefd. Deze controles zullen zowel met de opdrachtgever als met de examinatoren gebeuren.

- **Voortgangsgesprekken:** om de andere week zullen de opdrachtgever en de opdrachtnemer de voortgang en het uitgevoerde werk bespreken. Tijdens deze gesprekken kan de opdrachtgever feedback op de opgeleverde producten en op de werkwijze van opdrachtnemer geven. Tegelijk is dit voor de opdrachtnemer een mogelijkheid eventuele vragen of knelpunten bespreekbaar te maken.

- Terugkoppelen plan van aanpak: in het vorige hoofdstuk zijn de tussenresultaten en de meetbare mijlpalen hiervan genoemd. Na het behalen van een van deze mijlpaal zal de opdrachtnemer terugkoppelen naar het plan van aanpak om te controleren of deze mijlpaal voldoet aan de vooraf gestelde eisen en of het project nog de beschreven koers volgt.
- Terugkoppelen opdrachtgever: alle tussenproducten worden voorgelegd aan de opdrachtgever ter controle. De opdrachtgever kan dit product voorzien van feedback of richtlijnen voor de hierop volgende stappen en -/ of producten. Op basis van deze feedback worden de producten aangepast.

Controles examinatoren school: op 30-40% van het project zal er een voortgangsgesprek worden gepland met de docentbegeleider van de Haagse Hogeschool en zal op 80% al het uitgevoerde werk worden ingeleverd voor een tussentijdse beoordeling.

7. Projectorganisatie

Dit project betreft gedeeltelijk mensen en activiteiten die buiten de 'normale' gang van zaken van de organisatie vallen. Daarom worden in dit hoofdstuk de betrokken personen (inclusief contactgegevens), praktische zaken omtrent manieren van communicatie, rapportage en opslag besproken, die in de voorgaande hoofdstukken nog niet zijn besproken.

7.1 Contactgegevens

Josine Blom

Functie: Opdrachtnemer (stagiaire)
Taken: Uitvoeren projectwerkzaamheden, oplevering (tussen)producten, verantwoording werkwijze
Organisatie: Naturalis & Haagse Hogeschool
E-mailadres: Josine.Blom@naturalis.nl & 12082619@student.hhs.nl
Telefoonnummer: 06-10544551

Marian van der Meij

Functie: Opdrachtgever en bedrijfsbegeleider
Taken: Begeleiding en ondersteuning opdrachtnemer, beoordeling werkwijze en producten namens Naturalis
Organisatie: Naturalis
E-mailadres: marian.vandermeij@naturalis.nl
Telefoonnummer: 071-7519386 & 06 16126148

Klaas Jan Mollema

Functie: Eerste examiner
Taken: Begeleiding opdrachtnemer, beoordeling werkwijze en producten namens de Haagse Hogeschool
Organisatie: Haagse Hogeschool
E-mailadres: K.J.Mollema@hhs.nl

Jochem Mollema

Functie: Tweede examiner
Taken: Beoordeling eindproducten namens de Haagse Hogeschool
Organisatie: Haagse Hogeschool
E-mailadres: Jochem.Mollema@unafact.nl

7.2 Communicatie

De opdrachtnemer rapporteert aan de opdrachtgever op verschillende manieren. Als eerste zal dit gebeuren tijdens de twee- wekelijkse voortgang/controler gesprekken. Ten tweede zal de opdrachtnemer tussen producten voorleggen bij de opdrachtgever voor goedkeuring.

De communicatie hierover zal gedeeltelijk plaatsvinden via de mail en de gedeelde Google Drive omgeving. Met behulp van deze middelen kan de opdrachtnemer de opdrachtgever op de hoogte stellen van nieuwe producten en-/ of nieuwe versies hiervan.

De voortgangsgesprekken en overige gesprekken zullen persoonlijk plaatsvinden gedurende het project. Van deze gesprekken zal een verslag worden gemaakt door de opdrachtnemer. Deze dient ter verantwoording aan school en als verslaglegging van feedback en afspraken.

Alle documenten en definitieve producten zullen worden opgeslagen op Google Drive, in een map die toegankelijk is voor zowel de opdrachtnemer als de opdrachtgever. Dit betreft de volgende documenten:

- Opgeleverde (tussen)producten
- Praktische documenten, zoals de planning, het afstudeerplan en plan van aanpak
- Gevonden literatuur
- Uitwerkingen van interviews
- Verslagen van besprekingen
- Presentaties voor (tussentijdse) voortgangsmomenten
- Documenten van de Automated georeferencing pilot en andere (nodige) CETAF verslagen
- Eventuele templates

Daarnaast zullen alle tussenproducten, verslagen van besprekingen en voortgangsverslagen worden gedeeld met de twee examinatoren van school, via het persoonlijke afstudeerportfolio van de opdrachtnemer.

Oudere versies van (tussen)producten worden, inclusief feedback en opmerking bewaard in een specifiek daarvoor aangemaakte map in Google Drive.

8.Planning

De onderstaande planning geeft een overzicht van alle uit te voeren activiteiten en op te leveren producten (mijlpalen) in de doorlooptijd van het project. In de onderstaande tabel is een strokenplanning gemaakt om alle stappen genoemd in hoofdstuk 3 op een visuele wijze overzichtelijk te maken. Het project heeft een doorlooptijd van 17 werkweken, waarvoor 19 weken beschikbaar zijn tot aan de deadline op 8 januari. De groene blokken zijn de periode waarin de aan deze activiteiten en producten wordt gewerkt en de rode blokken geven de deadlines (mijlpalen) waarin waarop het deelproduct opgeleverd moet worden.

Activiteit	Week																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1: Plan van aanpak schrijven																			
Oriëntatie werkplek																			
Opstellen PVA																			
2: Oriëntatieprobleemdomein																			
Inlezen doc. Auto-geo meeting																			
Probleemanalyse opstellen																			
3: Intern deskresearch collecties																			
Onderzoek bestaande collecties																			
Opstellen datadictionary																			
4: Interviews doelgroep Naturalis																			
Selecteren interview partners																			
Contact opnemen interviewers																			
Opstellen interviewschrift																			
Uitwerken interviews																			
Vertalen naar informatiebehoefte																			
5: Interviews instellingen CETAF																			
Selecteren interview partners																			
Contact opnemen interviewers																			
Opstellen interviewschrift																			
Uitwerken interviews																			
Vertalen naar methode/tools																			
6: deskresearch methode/tools																			
Opstellen onderzoeksplan																			
Deskresearch uitvoeren																			
Lezen en analyseren artikelen																			
Opstellen literatuurreview																			
7: Pakketselectie																			
Overzicht wensen en eisen																			
Uitvoeren pakketselectie																			
Analyseren geschikte methodes																			
Verantwoorden keuzes en scores																			
8: opstellen advies																			
Samenvoegen tussenproducten																			
Opstellen onderzoeksrapport																			
Opstellen best practice methode																			
Voorleggen opdrachtgever																			
Feedback verwerken																			

9. Kosten & baten

Aan het uitvoeren van een project zijn altijd kosten en baten verbonden. In dit geval zijn deze beide aspecten lastig uit te drukken in geld. De kosten van beide partijen, zowel opdrachtnemer- als opdrachtgever, zullen voornamelijk tijd betreffen. De materiaalkosten en overige middelen van beide partijen zijn dusdanig laag dat deze niet in dit hoofdstuk genoemd worden. De opbrengsten van het project gaan voornamelijk uit van vergaring van kennis en verbetering van de kwaliteit van de collectie en betreffen geen winstoogmerk.

9.1 Kosten



Opdrachtgever

De kosten van de opdrachtgever zijn voornamelijk tijd. Naast de initiële opstart van het project zal de opdrachtgever gedurende de looptijd van het project geregeld tijd moeten besteden aan voortgangsgesprekken, ondersteuning en controle van producten die de opdrachtnemer oplevert. Hierbij moet gedacht worden aan ongeveer één halve dag/ één dag per twee weken.

Opdrachtnemer

De kosten van de opdracht nemer betreffen ook alleen tijd, maar liggen hoger dan die van de opdrachtgever. De opdrachtnemer zal zeventien weken lang 36 uur per week besteden aan de uitvoering van het project.

9.2 Baten



Opdrachtgever

Voor de opdrachtgever levert dit project een verbetering van de kwaliteit van het product (de collectie) van de organisatie. Hiervan zal in het bijzonder de usability en toepasbaarheid van de collectie kunnen worden vergroot. Daarnaast kan de organisatie naar aanleiding van het project de afnemers van de collectie beter ondersteunen in hun gebruik van de collectie voor onderzoeken. Naturalis kan tot slot met deze best practice, samen met de andere instellingen aangesloten bij de CETAF een gestandaardiseerde methode en EU infrastructuur kunnen ondersteunen

Opdrachtnemer

Voor de opdrachtnemer betekent dit project, mits het goed wordt afgerond, voornamelijk de mogelijkheid af te studeren en een HBO- diploma te kunnen halen. In meetbare standaarden zou dit 30 studiepunten betreffen.

10. Risico's

In het hoofdstuk projectorganisatie wordt duidelijk dat er bij dit project weinig mensen en partijen betrokken zijn. In principe zijn dit drie partijen, Naturalis, de opdrachtnemer en de Haagse Hogeschool, waarvan 1 partij zich enkel richt op de voortgang van de ontwikkeling van de opdrachtnemer. Het project wordt als succesvol beschouwd, wanneer het juiste projectdoel wordt opgeleverd met de afgesproken kwaliteit, binnen de daarvoor vastgestelde doorlooptijd.

Ondanks de lage kosten, en klein aantal betrokken partijen is er enkele interne en externe risico's die het projectdoel in gevaar kunnen brengen. Door ze hieronder duidelijk in kaart te brengen worden de aandachtspunten van het project duidelijker en kan hier meer aandacht aan worden besteed.

10.1 Interne risico's



- Een van de eisen om het project als succesvol te kunnen beschouwen is het halen van de deadline. Een definitieve deadline biedt echter weinig ruimte, voor onvoorziene zaken, zoals ziekte, vertraagde levering of toegang tot middelen ed. Daar komt bij dat de definitieve deadline voor de opdrachtgever en de opdrachtnemer verschillen. Het project duurt voor de opdrachtnemer namelijk 8 januari 2016. Dit is de definitieve deadline van het afstudeerverslag. Voor de opdrachtgever, geldt echter dat het project verwacht klaar te zijn voor de volgende Automated georeferencing meeting op 16 aart 2016.
- Naast de definitieve deadline is de doorlooptijd van het project aan de korte kant. Het project duurt tot 8 januari, wat slecht 17 weken geeft voor de uitvoering en afronding. Hierbij moet worden genoemd dan de laatste 2-3 weken volledig zullen worden besteed aan de opstellen van het eindrapport en afronding van het project. Er zijn dus in totaal 14-15 weken om het onderzoek uit te voeren.

10.2 Externe risico's



Er is in dit project maar één extern risico's die het project eventueel vertraging zou kunnen opleveren, en dat zijn de betrokken instellingen van CETAF. Deze worden namelijk als onderdeel van het project geïnterviewd over hun eerdere ervaringen, maar zijn alle in het buitenland gevestigd. Er zal worden geprobeerd om in persoon met deze stakeholder te overleggen. De internationale communicatie kan het project vertraging opleveren. Er moet dus voorafgaande aan de interviews (en het initiële contact) worden nagedacht over mogelijke digitale middelen voor het afnemen van de interviews.

Literatuurlijst



Blom, J. (2015) *Afstudeerplan: Opdrachtschrijving afstudeerstage Naturalis*. Intern document

CETAF (z.d.1). *About us: What is CETAF*. Geraadpleegd op 31 augustus 2015, via <http://www.cetaf.org/about-us/what-cetaf>

CETAF (z.d.2) CETAF Bodies. Geraadpleegd op 8 augustus 2015, via <http://www.cetaf.org/taxonomy/cetaf-bodies>

Chapman, A.D. & Wieczorek, J. (2006). *Guide to best practices for georeferencing*. Copenhagen: Global Biodiversity Information Facility

Gritt, R. (2011) *Projectmanagement*. Groningen/Houten: Noordhoff Uitgevers

Naturalis Biodiversity Center (2012). *Beleidsplan 2013-2016*. Geraadpleegd op 31 augustus 2015, via: http://www.naturalis.nl/media/library/2013/09/Beleidsplan_2013-2016.pdf

Naturalis Biodiversity Center (2014). *Follow-up Actions Meeting Automated Georeferencing*. Intern document, geraadpleegd op 2 september 2015

Naturalis Biodiversity Center (2015a). *Sectoren*. Geraadpleegd op 31 augustus 2015, via: <http://www.naturalis.nl/nl/over-ons/organisatie/sectoren/>

Naturalis Biodiversity Center (2015b). *Wat wij doen*. Geraadpleegd op 31 augustus 2015, via: <http://www.naturalis.nl/nl/over-ons/wat-doen-wij/>

Stuurgroep FES Collectie Digitalisering (2015). *Eindrapport FCD- Pilot Georeferencing: productiematig georeferencen*. Intern document Naturalis, Leiden

Verhoeven, N. (2014). *Wat is onderzoek? Praktijkboek voor methoden en technieken*. Amsterdam: Boom Uitgevers

Bijlage 1: Verslag opstartgesprek opdrachtgever

Opstart gesprek afstudeerstage

Dinsdag 1 september 2015

Marian van der Meij, Josine Blom

Naar aanleiding van de rondleidingen en eerste documenten van maandag 31 augustus en het afdelingsoverleg van dinsdag 1 september had ik enkele praktische en inhoudelijke vragen over de stage. Dinsdag 1 september heb ik daarom een kort overleg gehad met Marian over de opstart van het project.

Praktische zaken

- De Google + Community is te vinden via share.naturalis.nl. Hierop kun je vragen stellen en zaken delen met de andere medewerkers van Naturalis. Daarnaast vind je hier de evenementen kalender, die voornamelijk handig is voor de introductie cursussen over de overstap naar Google Drive. Donderdag 3 september is er een cursus en het is handig om hierheen te gaan, ook als je al ervaring heb met Google Drive, omdat de omgeving die specifiek voor Naturalis is ingericht wordt toegelicht.
- Dino is de oude documentenopslag en is gewoon de naam voor de server. Hierop kan je al jou documenten plaatsen en kun je een aantal mappen vinden zoals, sjablonen. Verder zal jij weinig gebruik maken hiervan, omdat we overgaan naar Drive.
- Alle producten die je maakt kan je in een Drive map plaatsen die je deelt met mij. Zo kan ik bij jouw documenten. Misschien is het ook handig om een map aan te maken waar je alle documenten die bedoeld zijn voor de CETAF instellingen kan plaatsen en kan delen met deze instellingen.
- Voor het project zal je de collectie moeten bekijken. Onze collectie staat gedeeltelijk in het CRS (collectie registratie systeem). Hiervoor heb je een inlogaccount nodig om toegang te krijgen. Ik heb deze aangevraagd en zodra die is geregeld geef ik dit door.

Inhoudelijke zaken

- Er is een archiefmap met alle documenten van het pilotproject. Deze kan nu nog niet gedeeld worden, omdat de beheerder deze week afwezig is. De toegang is wel aangevraagd. De meeste belangrijke documenten heb ik gedownload en gemaaild, maar voor de hele map moeten we wachten tot de beheerder terug is.
- Jeroen en Wilfred waren betrokken bij het pilotproject. Hier kan je bij terecht met vragen of als je wil sparren over dit project.
- In de literatuurlijst van het eindrapport van het pilotproject staan een aantal zeer nuttige bronnen. Een aantal liggen hardcopy in de kast. Die kan Jeroen voor je regelen.
- In het onderzoek moet de Naturalis collectie in zijn geheel worden bekeken. In tegenstelling tot het pilot moet dus ook het Brahm's worden bekeken, naast het CRS.
- Het is niet de bedoeling dat je zelf gaat testen, maar de resultaten en ervaringen van bestaande pilotprojecten gebruik. De bedoeling is dat deze vergeleken worden en je uiteindelijk kiest welke het beste op elkaar aansluiten. In de notulen van enkele besprekingen staat al genoemd, dat sommige methode wellicht op elkaar aansluiten, maar dit gaat om verschillende methode van 1 instelling. In dit project worden alle methode van alle CETAF instellingen met elkaar vergeleken.
- Er is al een voorzet gemaakt voor contact met de CETAF instellingen. In de vorige Automated Georeferencing vergadering is al geopperd dat er een aantal graag wilde meewerken. Hiermee kan direct contact worden opgenomen, maar er kan ook contact opgenomen worden met de hele mailinglijst. Mails hiervan stuurt Marian ik door.
- De deadline van jouw stage komt wat ongunstig uit rondom de kerstperiode. Marian is de week rondom kerst vrij. Dit betekent dat het eindrapport voor de kerstvakantie moet zijn afgerond zodat Marian ze kan voorzien van feedback.

Bijlage 2: Afstudeerplan

Informatie afstudeerder en gastbedrijf

Afstudeerblok: 2015-2.1 (start uiterlijk 31 augustus 2015)
Startdatum uitvoering afstudeeropdracht: 31 augustus 2015
Inleverdatum afstudeerdossier volgens jaarrooster: 8 januari 2016
Studentnummer: 12082619
Achternaam: mw Blom
Voorletters: J J
Roepnaam: Josine
Adres: Eiklaan 46
Postcode: 2282AV
Woonplaats: Rijswijk
Telefoonnummer: 0703992425
Mobiel nummer: 0610544551
Privé emailadres: josine.blom@hotmail.com

Opleiding: IDM
Locatie: Den Haag
Variant: voltijd
Naam studieloopbaanbegeleider: René Vogels
Naam begeleidend examiner: Klaas Jan Mollema
Naam tweede examiner: Jochem Mollema

Naam bedrijf: Naturalis Biodiversity center
Afdeling bedrijf: informatiemanagement
Bezoekadres bedrijf: Darwinweg 2
Postcode bezoekadres: 2333 CR
Postbusnummer: 9517
Postcode postbusnummer: 2300RA
Plaats: Leiden
Telefoon bedrijf: 071 751 96 00
Telefax bedrijf: 071 751 96 66
Internetsite bedrijf: <http://www.naturalis.nl>

Achternaam opdrachtgever: mw van der Meij
Voorletters opdrachtgever: M.A.A.
Titulatuur opdrachtgever: Dr.
Functie opdrachtgever: Informatiemanager
Doorkiesnummer opdrachtgever: 071-75179386
Email opdrachtgever: marian.vandermeij@naturalis.nl

Achternaam bedrijfsmentor: mw van der Meij
Voorletters bedrijfsmentor: M.A.A.
Titulatuur bedrijfsmentor: Dr.
Functie bedrijfsmentor: Informatiemanager
Doorkiesnummer bedrijfsmentor: 071-75179386
Email bedrijfsmentor: marian.vandermeij@naturalis.nl
Doorkiesnummer afstudeerder:
Functie afstudeerder (deeltijd/duaal):

Titel afstudeeropdracht

Een best practice opstellen voor het (semi) automatisch georeferencing van de gedigitaliseerde collectie data van Naturalis.

Opdrachtomschrijving

1. Bedrijf

Naturalis Biodiversity is een stichting die biodiversiteit wil beschrijven, begrijpen en verklaren. Voor het welzijn van de mensen en het voortbestaan van de aardse natuur (missie).

Dit doen ze door:

- Collecties: ze verzamelen, bewaren en bestuderen dieren, planten en stenen. De collectie telt inmiddels 37 miljoen objecten en is daarmee van wereldformaat.
- Onderzoek: met wetenschappelijk onderzoek dragen ze bij aan de mondiale inspanning om de biodiversiteit te reconstrueren en de interactie tussen soorten te begrijpen;
- Museum: gebruikmakend van de collectie vertellen ze het verhaal van de natuur in zijn natuurhistorische en cultuurhistorische context, aansprekend voor verschillende leeftijdsgroepen en voeding gevend aan verwondering;
- in een open dialoog kennis te delen over de natuur, taxonomie en biodiversiteit, die bruikbaar is voor de samenleving en een duurzame economie.
- Onderwijs: Als internationaal kenniscentrum en academische werkplaats voor de universiteiten van Amsterdam, Leiden en Wageningen. Zo dragen ze bij aan het opleiden van een nieuwe generatie wetenschappers en academische professionals op het gebied van systematiek en biodiversiteit.
- Ten slotte levert Naturalis een bijdrage aan bedrijven, overheden en maatschappelijke organisaties:
 - Ze helpen de douane wanneer er lastig te identificeren dieren of planten aan de grens verschijnen.
 - Bij het certificeren van hout verlenen ze assistentie om te bepalen of een partij uit gekweekt of wild hardhout bestaat.
 - Ze ondersteunen natuurhobbyverenigingen bij het ontwikkelen en beheren van websites met natuurinformatie.
 - Ze adviseren de juweliersbranche bij het bepalen van echtheid van edelstenen.

Het Nederlands Centrum voor Biodiversiteit Naturalis werd gevormd op 28 januari 2010 uit een samenwerking tussen het Zoölogisch Museum Amsterdam (onderdeel van de Universiteit van Amsterdam), het Nationaal Herbarium Nederland (onderdeel van de Universiteit Leiden, de Universiteit Utrecht en de Wageningen University) en het Nationaal Natuurhistorisch Museum in Leiden. Het hoofddoel van deze samenwerking was het integreren van de collecties van de genoemde partners tot één nationale collectie van 37 miljoen planten, dieren, fossielen, gesteenten en mineralen. Het instituut behoort daarmee qua omvang mondiaal tot de top vijf in het natuurhistorische domein.

De huidige organisatie is ingedeeld in vijf sectoren: Bedrijfsvoering, Onderzoek en Onderwijs, Collectie, Publiek en Markt en ICT. ICT richt zich op het digitaliseren van collectie- en andere biodiversiteitsinformatie, het beschikbaar en toepasbaar maken ervan en het koppelen ervan aan andere digitale bronnen. De ICT sector van Naturalis bevat een afdeling informatiemanagement bestaande uit twee teams, applicatiebeheer en datamanagement. De afdeling informatiemanagement werkt op projectbasis. Dit betekent dat ze 50% collectieondersteuning uitvoeren en de overige 50% besteden aan verschillende projecten die bijdragen aan het informatiemanagement en de toegankelijkheid van informatie. Deze afstudeerstage zal worden uitgevoerd bij de sector ICT en dan specifiek het team datamanagement.

2. Probleemstelling

Tijdens het FES collectiedigitaliseringsproject zijn 7 miljoen collectieobjecten gedigitaliseerd. Hierbij is de informatie die bij de objecten hoort (zoals taxonomie, verzameldatum en verzamelplaats) in centrale databases gezet (Collectie Registratie Systeem (CRS) en Botanical Research And Herbarium Management System (Brahms)). Deze data worden ontsloten via het dataportal van GBIF (Global Biodiversity Information Facility) en de Naturalis BioPortal.

Eén van de belangrijkste toepassingen van deze data is het gebruik ervan in wetenschappelijk onderzoek. Naast de taxonomische data is de geografische verzameldata van de objecten hiervoor erg belangrijk. Met de geografische data van objecten is het namelijk mogelijk om bijvoorbeeld vroegere en huidige verspreidingsdata van soorten te mappen en te analyseren en ze te combineren met ruimtelijke data van andere soorten en disciplines (bv. meteorologie, geologie, geografie, sociale en medische wetenschappen). Naast fundamenteel onderzoek kunnen hiermee ook vraagstukken over wereldwijde kwesties als natuurbehoud en klimaatverandering onderzocht worden.

De collecties van Naturalis worden behalve voor wetenschappelijk onderzoek ook beschikbaar gemaakt voor andere doelgroepen van de instelling, zoals bezoekers, universiteiten, etc. Deze doelgroepen willen ook gebruikmaken van de geografische data van de collecties, maar wellicht op een ander niveau dan nodig is voor het wetenschappelijk onderzoek. Zo zal een natuurhobbyvereniging genoeg hebben aan 'Leiden, Nederland' als verzamelplaats- data, maar zal een wetenschappelijk onderzoeksinstituut de precieze coördinaten willen weten van de straat of zelf het betreffende huisnummer.

Alle geografische data van objecten is conform de internationale standaarden (DarwinCore en ABCD EFG) genoteerd in CRS en Brahms. Bij een groot deel staat deze data echter alleen als tekst in de registratiesystemen genoteerd en bevat veelal oude namen (bv. Pruisen, Joegoslavië). Deze data is voor de wetenschap beter bruikbaar als de objectrecords verrijkt worden met de bijbehorende coördinaatgegevens. Om de kwaliteit en de toepasbaarheid van de Naturalis dataset te vergroten is het daarom wenselijk dat de tekstuele geografische objectinformatie van (delen van) door Naturalis beheerde objectrecords op een gestandaardiseerde manier worden verrijkt met de overeenkomstige geografische coördinaten.

Productiematig georeferenzen is een relatief nieuw concept; er is geen kant en klare methode beschikbaar. Bijkomend probleem is dat er veel oude geografische namen of verwijzingen zijn naar oude kaarten. Naturalis heeft nog weinig ervaring met het toepassen van georeferencing op grote hoeveelheden objectrecords. Daarom is een pilot project georeferencing uitgevoerd waarin is gekeken wat de mogelijkheden zijn om de vindplaatsgegevens van deze objecten op een grotendeels geautomatiseerde manier te voorzien van coördinaten. Deze pilot toonde aan dat productiematig georeferenzen met Google Maps (Google Geocoding API) een veelbelovende methode is die, na realisatie van de benodigde CRS functionaliteiten, toegepast kan worden op objectrecords in het CRS.

Er zijn echter nog veel meer instellingen die dergelijk testprojecten hebben uitgevoerd. Tijdens de Automated Georeferencing meeting in december 2014 georganiseerd door Naturalis, zijn enkele van deze projecten besproken. Resultaten uit deze meeting en de vraag of er meer instituten geïnteresseerd zijn in het ontwikkelen van een best practice voor georeferencing zijn voorgelegd bij de CETAF-ISTC meeting. Meerdere geïnteresseerde hebben zich gemeld en Naturalis heeft de lead gekregen om dit initiatief verder uit te werken.

Hoofdvraag

Wat moet je als instelling doen om een collectie geografische verzameldata met tekstuele locatieaanduidingen op een semi automatische manier te verrijken met de overeenkomstige geografische coördinaten op verschillende niveaus/ in verschillende stappen van nauwkeurigheid?

Deelvragen

1. Wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?
2. Op welke verschillende niveaus willen de doelgroepen van Naturalis de geografische data beschikbaar hebben?
3. Welke andere instellingen hebben projecten uitgevoerd om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?
 - a. Wat is de betrouwbaarheid/ accuraatheid van de gebruikte methode/tools?
 - b. In welke mate was handmatige correctie nodig om de betrouwbaarheid van de georeferencing data te verbeteren?
 - c. Welke middelen zijn er gebruikt om deze verbeteringen door te voeren?
4. Welke (semi) automatische methodes/ tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten?
5. Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren?
6. Welke semi automatische programma's zijn het meest geschikt voor de beschreven collecties?
7. Hoe sluiten deze methodes/ programma's het beste op elkaar aan.
8. Hoe kunnen de methodes/ programma's het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?

3. Doelstelling van de afstudeeropdracht

Het doel van dit project - het maken van een handleiding met methoden en aanbevelingen om grote hoeveelheden objectrecords zo geautomatiseerd mogelijk (productiematig) te georeferencen.

4. Resultaat

Het eindproduct van deze afstudeerstage is een advies voor Naturalis gericht op hoe zij hun collectie met geografische verzameldata kunnen verrijken met overeenkomstige coördinaten. Dit advies kan door Naturalis worden gebruikt bij het uitvoeren van een project om de collectiedata te verrijken en eventuele andere projecten met collecties die in de toekomst nog gedigitaliseerd worden.

Uit dit eindproduct kunnen uiteindelijk nog twee resultaten volgen:

1. De gedigitaliseerde collecties van Naturalis kunnen worden voorzien van nieuwe geografische verzameldata, waardoor onderzoekers van Naturalis vroegere en huidige verspreidingsdata van soorten mappen en analyseren en ze combineren met ruimtelijke data van andere soorten en disciplines. Naast fundamenteel onderzoek kunnen hiermee ook vraagstukken over wereldwijde kwesties als natuurbehoud en klimaatverandering onderzocht worden.
2. Daarnaast kan dit advies worden gebruikt als internationale best practice voor alle instellingen die aangesloten zijn bij CETAF. Deze internationale best practice kan bijdragen aan een internationale standaardmethode voor (semi) automatische georeferencing en een infrastructuur voor alle natuurhistorische collecties binnen de EU voor de toekomst. Een dergelijke standaardmethode en infrastructuur zouden de usability en de kwaliteit van digitale natuurhistorische collecties kunnen vergroten.

Urgentie

De resultaten van het onderzoek worden verwacht op de volgende jaarlijkse meeting van de CETAF-ISTC en deze zal in maart 2016 plaatsvinden. Dit is als de planning goed loopt 2 maanden na de afronding van deze afstudeerstage.

5. Uit te voeren werkzaamheden, inclusief een globale fasering, mijlpalen en bijbehorende activiteiten

De afstudeeropdracht bestaat uit 4 delen; oriëntatie, onderzoek, analyse en uitvoering. Bij elk deel staan hieronder de stappen beschreven met bijbehorende onderzoeksmethode en producten.

Deel 1 oriëntatie:

Stap 1: Plan van aanpak schrijven.

Als eerste stap van het onderzoek wordt er een plan van aanpak opgesteld, waarin nog specifiekere dan in dit afstudeerplan de eisen en afspraken samen met de opdrachtgever worden vastgelegd en een planning wordt opgenomen.

Bijbehorend product: Deze stap levert een plan van aanpak op.

Stap 2: Aanvullende oriëntatie op het probleemdomein uitvoeren.

Hierbij wordt specifiek gericht op het eerder afgeronde pilotproject dat Naturalis uitvoerde en andere informatie over vergelijkbare projecten die Naturalis al intern beschikbaar heeft, zoals de verslagen van de Automated Georeferencing meeting ed.

Bijbehorend product: Het product dat deze stap oplevert is een probleemanalyse met overzicht van mogelijk te onderzoeken methode en technieken, een zoekwoordenlijst voor het literatuuronderzoek en een lijst met mogelijke te interviewen instellingen.

Bijbehorende beroepstaak: Probleemanalyse

Deel 2 onderzoek:

Stap 3: Intern literatuur onderzoek naar de betreffende collecties uit voeren (inventarisatie).

Als eerste onderdeel van het onderzoek wordt er een inventarisatie gemaakt van de collecties die voorzien dienen te worden van coördinaten en de verscheidenheid aan geografische meta data in de collecties. De collecties kunnen namelijk metadata in tekstvorm bevatten, maar wellicht ook al coördinaten en andere formaten zoals de Rijksdriehoek. Het gaat dan voornamelijk om ver verschil tussen nieuw binnengekomen materiaal en ouder materiaal. Collectie materiaal van voor 2003 kan bijvoorbeeld de term 'Joegoslavië' kunnen bevatten. In het nieuwe materiaal kom je die, als het goed is, niet meer tegen.

De beste manier om dit te doen is door de datasets en de collecties zelf te bekijken en op basis hiervan een beschrijving te maken. Zo krijg je echt een beeld van de huidige stand van zaken en het meest realistische beeld van de collectiedata. Ook biedt dit de mogelijkheid om alleen de benodigde informatie voor dit onderzoek eruit te halen en de overbodige informatie weg te laten.

Bijbehorend product: een datadictionary. Dit is een overzicht/ beschrijving (inventarisatie) op van de collecties die moeten worden verrijkt met coördinaten, op basis waarvan de methoden en tools die later worden onderzocht kunnen worden beoordeeld.

Bijbehorende deelvraag:

1. Wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?

Bijbehorende beroepstaak: Datadictionary

Stap 4: Interviews afnemen met de doelgroepen van Naturalis.

Deze stap is nodig om vast te stellen op welke niveaus of wijze de huidige geografische verzameldata verrijkt moeten worden, zodat deze voor alle doelgroepen bruikbaar zijn. Dit wordt gedaan door interviews af te nemen met de gebruikers van de collectie, zowel intern (onderzoekers, medewerkers en de afdeling informatiebeheer) al extern (CETAF- partners).

Bijbehorend product: Deze stap levert een overzicht op van de informatiebehoefte van de gebruikers en medewerkers dat vertaald moet kunnen worden naar een lijst van wensen en eisen op.

Bijbehorende deelvraag:

2. Op welke verschillende niveaus willen de doelgroepen van Naturalis de geografische data beschikbaar hebben?

Bijbehorende beroepstaak: informatiebehoefte

Stap 5: Interviews afnemen met de instellingen die eerder dergelijke projecten hebben uitgevoerd.

Tijdens de Automated Georeferencing meeting zijn er kort drie pilotprojecten van (semi) automatische georeferencing besproken van instellingen die aanwezig waren bij de meeting, één hiervan was Naturalis. Naast de overige twee projecten zijn er nog enkele instellingen die een pilotproject hebben uitgevoerd en/of onderzoek hebben gedaan. Door met deze instellingen in gesprek te gaan, via interviews of een gezamenlijke bijeenkomst (voorstel van de opdrachtgever) kan worden geanalyseerd welke methode de meest gebruikte zijn, of deze nog gebreken of knelpunten hebben en waar de overlappen zitten. Zo kan voorkomen worden dat het wiel opnieuw uitgevonden wordt.

Bijbehorend product: ook deze stap levert interviewverslagen of gecodeerde interviews op.

Bijbehorende deelvraag:

3. Welke andere instellingen hebben projecten uitgevoerd om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?
 - a. Wat is de betrouwbaarheid/ accuraatheid van de gebruikte methode/tools?
 - b. In welke mate was handmatige correctie nodig om de betrouwbaarheid van de georeferencing data te verbeteren?
 - c. Welke middelen zijn er gebruikt om deze verbeteringen door te voeren?

Bijbehorende beroepstaak: Deskresearch

Stap 6: Literatuuronderzoek naar mogelijke methodes/tools voor (semi) automatisch georeferencing, en middelen om de accuraatheid te verbeteren.

Dit onderdeel van het onderzoek zal uit deskresearch bestaan. Het heeft als doel de verzamelde informatie uit de interviews aan te vullen en een compleet beeld te geven van het mogelijke aanbod van methodes voor georeferencing.

Bijbehorend product: Dit onderdeel levert een literatuurreview op die samen met de verslagen van de interviews een onderzoeksrapport vormen. Dit wordt opgenomen in het uiteindelijke adviesrapport.

Bijbehorende deelvragen:

4. Welke semi automatische methodes/ tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten?
5. Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren?

Bijbehorende beroepstaak: pakketselectie

Deel 3: analyse:

Stap 7: Vergelijking van de eisen die de gebruikers van collectie stellen en de bevindingen uit de literatuur en interviews.

In deel 2 van het onderzoek is informatie verzameld via verschillende onderzoeksmethode en in dit deel van het onderzoek wordt al deze informatie vergeleken. Op basis van alle gevonden eisen die de gebruikers van de collectie en de opdrachtgever stelt kan worden vergeleken welke methodes en tools het meest geschikt zijn. Deze stap zal de vorm krijgen van een pakketselectie. De selectie voor de best practice methode zal worden gedaan door het overzicht van wensen en eisen van de toekomstige gebruikers en de resultaten van de eerder uitgevoerde pilotprojecten onder elkaar te plaatsen en de gevonden methoden en tools te beoordelen op al deze punten. De verwachting van de eerder CETAF meetings is dat er wellicht meerdere methodes zijn die aansluiten bij de wensen en eisen. De vraag is dus juist; "hoe kunnen deze methodes het beste ingezet worden en op elkaar aansluiten?".

Bijbehorend product: Uiteindelijk levert dit dus een analyse (pakketselectie), waarschijnlijk in tabelvorm, op met daarin de scores van alle gevonden tools en methode.

Bijbehorende deelvraag:

6. Welke semi automatische methodes/programma's zijn het meest geschikt voor de beschreven collecties?
7. Hoe sluiten deze methodes/ programma's het beste op elkaar aan?

Bijbehorende beroepstaak: pakketselectie

Deel 4: uitvoering

Stap 8: Opstellen van een advies (best practice) voor Naturalis in de vorm van een stappenplan.

De laatste stap van het afstudeerproject is het opstellen van het advies en de daarbij behorende internationale best practice. Op basis van de tabel uit de vorige stap kan een advies worden gedaan over de beste methoden voor Naturalis en de andere CETAF-ISTC instellingen.

Bijbehorend product: deze laatste stap levert een adviesrapport op voor Naturalis met als onderdeel hierin de internationale best practice.

Bijbehorende deelvraag:

8. Hoe kunnen de methodes/ programma's het beste worden ingezet om geografische verzameldata te verrijken met de overeenkomstige geografische coördinaten?

Bijbehorende beroepstaak: Parametriseren

6. Op te leveren (tussen)producten

Tijdens deze afstudeerstage zal de student de volgende (deel) producten opleveren:

- Plan van aanpak
- Onderzoeksverslag met hierin:
 - een probleemanalyse met overzicht van mogelijk te onderzoeken methode en technieken, een zoekwoordenlijst voor het literatuuronderzoek en een lijst met mogelijke te interviewen instellingen.
 - een overzicht/ beschrijving (inventarisatie) op van de collecties die moeten worden verrijkt met coördinaten, op basis waarvan de methoden en tools die later worden onderzocht kunnen worden beoordeeld.
 - een overzicht op van de informatiebehoefte van de gebruikers en medewerkers dat vertaald moet kunnen worden naar een lijst van wensen en eisen op.
 - Pakketselectie
 - De uiteindelijke best practice methode (advies)

7. Te demonstreren competenties en wijze waarop

Beroepstaak 1: probleemanalyse

De probleemanalyse vindt plaats op basis van documenten en gesprekken met betrokkenen op de afdeling informatiemanagement, de CETAF-ISTC en het pilot project. Het doel van de probleemanalyse is het oriënteren op het probleemdomein. Dit zal worden gedaan op basis van de ervaringen en knelpunten van het eerder afgeronde pilot project dat Naturalis heeft uitgevoerd. Het rapport moet voor alle betrokkenen herkenbaar en aanvaardbaar zijn. Het bevat ook een verantwoording van de aanpak. Het resultaat is een hoofdstuk in het uiteindelijk onderzoeksrapport met de probleemanalyse en een overzicht van mogelijk te onderzoeken methode en technieken, een zoekwoordenlijst voor het literatuuronderzoek en een lijst met mogelijke te interviewen instellingen.

Beroepstaak 2: Datadictionary

Als tweede zal na de probleemanalyse een intern literatuuronderzoek worden gestart naar de collecties die op basis van het uiteindelijke advies moeten worden verrijkt met overeenkomende coördinaten.

Het gaat hier om het definiëren en kwalificeren van entiteiten die een bepalende rol spelen bij de inrichting van een documentair informatiesysteem. Deze entiteiten (enerzijds gelieerd aan de beoogde gebruikers,

anderzijds aan de informatie die opgeslagen en toegankelijk gemaakt moet worden) zijn af te leiden uit eerder analyserend onderzoek.

Zoals bij punt 4 in dit onderzoeksplan al genoemd werd, hebben niet alle collecties overeenkomende geografische metadata. Om het advies uiteindelijk toepasbaar te maken voor alle collecties die Naturalis bezit moet zowel een goede beschrijving worden gemaakt van de verscheidenheid aan geografische meta data in de collecties als de verschillende collecties die uiteindelijk moeten worden verrijkt. Bij deze beroepstaak worden de gebruikersaspecten en de gegevens doe daarover geregistreerd moeten worden geanalyseerd, gedefinieerd en gespecificeerd ten behoeve van het collectieregistratiesysteem. Dit deel levert dus een Datadictionary op. Tijdens het onderzoek, dat met behulp van deskresearch zal worden uitgevoerd, zal er worden gericht op de deelvraag: “Wat voor geografische metadata bevatten de collecties die verrijkt dienen te worden met geografische coördinaten?”.

Beroepstaak 3: deskresearch:

Tijdens stap 6 van de afstudeerstage zal deskresearch worden uitgevoerd. Hierbij is het doel het aanvullen van de verzamelde informatie van de interviews, over alle mogelijke methodes van applicaties voor (semi) automatisch georeferenzen van de collecties. Deze stap levert verwijzingen naar, en een analyse van relevante informatie om het onderzoek en het advies voor het eind advies te ondersteunen. Tijdens dit deel van de deskresearch zal er worden gericht op de deelvragen: “Welke (semi) automatische methodes/ tools bestaan er om tekstuele locatie aanduidingen te verrijken met de overeenkomstige coördinaten?” en “Welke middelen zijn er mogelijk om deze verbeteringen op een arbeidsextensieve wijze door te voeren?”.

Beroepstaak 4: Informatiebehoefte

Naturalis kent verschillende doelgroepen, naast interne en externe onderzoekers kunnen bijvoorbeeld ook universiteiten, amateur wetenschappers en het brede publiek gebruik maken van de digitale collecties, via het Bioportaal (<http://bioportal.naturalis.nl/?language=nl>). Door middel van interviews met de gebruikers en medewerkers en analyse van taken van medewerkers en afdelingen zal de behoefte aan kennis en informatie in kaart worden gebracht met het oog op een advies over de beschikbaarstelling van informatiebronnen en passende instrumenten voor het delen van kennis. Hiermee wordt specifiek gericht op de niveaus van geografische data die de gebruikersgroepen het meest bruikbaar vinden. Dit onderzoek levert een overzicht op van de informatiebehoefte van de gebruikers en medewerkers dat vertaald moet kunnen worden naar een lijst van wensen en eisen op. Tijdens dit onderzoek zal er worden gericht op de deelvraag: “Op welke verschillende niveaus willen de doelgroepen van Naturalis de geografische data beschikbaar hebben?”.

Beroepstaak 5: Pakketselectie:

Op basis van de interviews met de gebruikers en medewerkers van Naturalis en de deskresearch naar de mogelijke methodes van applicaties voor (semi) automatisch georeferenzen van de collecties wordt er een pakketselectie gemaakt. Bij deze beroepstaak worden de wensen en eisen die worden gesteld aan de functionaliteit en de technische aspecten en eisen van de betreffende pakketten getoetst. De selectie zal worden gedaan door op systematische wijze alle gevonden applicaties en methode te scoren op de wensen en eisen die hieraan worden gesteld. De verwachting van de eerder CETAF meetings is dat er wellicht meerdere methodes zijn die aansluiten bij de wensen en eisen. De vraag is dus juist; “hoe kunnen deze methodes het beste ingezet worden en op elkaar aansluiten?”.

Beroepstaak 6: Parametriseren

Uiteindelijk wordt er een advies gegeven over de meest geschikte applicaties en methoden voor het (semi) automatisch georeferenzen van de collecties en hoe deze methodes en applicaties op elkaar aan kunnen sluiten. Tijdens dit deel van de deskresearch zal er worden gericht op de deelvragen: “Welke (semi) automatische programma’s zijn het meest geschikt voor de beschreven collecties?” en “Hoe sluiten deze methodes/ programma’s het beste op elkaar aan?”.

Bij deze beroepstaak wordt een technische Datadictionary opgesteld die een beschrijving geeft van een documentair informatie systeem op basis van functionele eisen (zijn gegeven). De technische datadictionary beschrijft de wijze waarop de standaard attributen aangepast moeten worden, wat de wensen en de eisen van de gebruikers zijn en de manier waarop de applicatie ingericht moet worden.

8. Persoonlijke verantwoording

Tijdens mijn stage in het derde jaar van de opleiding kwam ik al in aanraking met Naturalis toen ik hier iemand interviewde over crowdsourcing projecten. Ik werd toen gelijk weer herinnerend aan hoe leuk ik de organisatie vind. In eerste instantie had dit vooral betrekking op de soort informatie waar zij over beschikken. Ondanks dat ik nooit iets met biologie heb gedaan, heeft dit mij wel altijd enorm geïnteresseerd. Naar aanleiding van mijn mailcontact en het eerste gesprek ontdekte ik dat er meer dingen aan Naturalis zijn die mij persoonlijk erg aanspreken. Zo werd mij verteld dat de afdeling informatiemanagement op projectbasis werkt. Dit betekent dat 50% van de tijd die zij besteden wordt gestoken in verschillende projecten. Dit zorgt voor veel afwisseling en de mogelijkheid om steeds in nieuwe projectgroepen werken. Deze dynamiek en afwisseling lijken mij een prettige manier van werken. Verder ben ik maar kort binnen geweest bij Naturalis en de informatieafdeling, maar de manier waarop ik werd ontvangen en waarop ik de medewerkers zag werken, leken mij erg prettig. Het gaf een open gevoel, waarbij ook de afdelingshoofden gewoon tussen de andere werknemers zitten.

De opdracht draait om geografische informatie, een onderwerp waar ik ook al een minor in volgde. Deze minor had ik bewust gekozen, ondanks dat ik de studiepunten niet nodig had. Het onderwerp sprak me aan en het leek mij een goede aanvulling op de andere onderwerpen die tijdens de opleiding worden aangeboden. Daarnaast is het een waardevolle aanvulling als ik me op de erfgoedsector wil blijven richten, aangezien steeds meer erfgoedinstellingen geografische componenten willen linken aan hun eigen collectie. De minor was echter niet erg uitgebreid en ik heb het idee dat ik met behulp van deze stage het onderwerp me nog meer eigen kan maken.