Gebruiksvoorwaarden voor collectiedata van externe providers binnen Nederlab

Versie 9 oktober 2015
Hennie Brugman, Meertens Instituut

1 Inleiding

Nederlab beoogt zoveel mogelijk bestaande digitale collecties met (historisch) Nederlandstalig tekstmateriaal samen te brengen en toegankelijk te maken voor taalkundigen, historici en letterkundigen via een virtuele onderzoeksomgeving.

Veel van deze collecties kennen beperkingen op gebruik vanwege auteursrechten. Deze notitie bevat een gedetailleerde uitwerking van alle soorten datagebruik in Nederlab. Deze inventarisatie kan als checklist worden gebruikt bij het vastleggen van afspraken met tekstleveranciers, waarbij dan binnen zo'n afspraak kan worden gedifferentieerd naar type gebruik.

Nederlab kan worden gezien als een intermediair, die tekstbronnen opwerkt voor wetenschappelijk gebruik door een specifieke wetenschappelijke doelgroep. Het vormt daarmee een extra, ten behoeve van deze doelgroep gespecialiseerde, 'outlet' voor aanbieders van digitale tekstcollecties. Deze rol vergt een apart type licentieovereenkomst, Nederlab is namelijk enerzijds een wetenschappelijk gebruiker van die collecties, maar anderzijds ook een representant van de oorspronkelijke rechthebbende partijen tegenover haar eindgebruikers.

2 Inventarisatie van soorten gebruik

Deze classificatie dient om collectieleveranciers gedetailleerd inzicht te geven in hoe Nederlab met hun data omgaat. Al deze typen gebruik kunnen aanleiding geven tot het maken van precieze afspraken met betrekking tot rechten.

2.1 Dataprocessing en transformatie

Collecties, die in Nederlab worden opgenomen ondergaan transformaties om ze geschikt te maken voor Nederlab doeleinden. Dit heeft mogelijk implicaties voor de herkenbaarheid van de data, en de zichtbaarheid van de oorspronkelijke rechthebbende. Hieronder worden deze transformaties nader toegelicht, onderscheiden naar metadata en tekst.

2.1.1 Metadata

Identificeerbare resources

De relatie tussen identifiers van binnenkomende resources en Nederlab resources is niet altijd 1 op 1. Voorbeelden: we splitsen z.g. onzelfstandige titels af (werken die alleen maar gebundeld in de collectie voorkomen), b.v., we beschouwen krantenartikelen als zelfstandige eenheden van Nederlab, ze krijgen allemaal een eigen Nederlab identifier.
We administreren zo goed als mogelijk de relaties tussen Nederlab identifiers en collectie identifiers, en zorgen dat de herkomst van onze Nederlab resources te allen tijde bekend is, ook voor eindgebruikers.

Verder legt Nederlab relaties tussen resources:

- we brengen links aan tussen titels en auteurs uit verschillende collecties, die de Nederlab-redactie als identiek beschouwt;
- we relateren verschillende incarnaties van titels, zodat we zo goed mogelijk in staat zijn de vroegste versie van een titel te vinden.

Metadataveld structuren
Metadatavelden uit op te nemen collecties worden ofwel *gemapt* op Nederlab 'kernvelden', ofwel overgenomen uit de broncollectie, ofwel genegeerd. Die Nederlab kernvelden bestaan enerzijds om de verschillende collecties binnen Nederlab homogeen te kunnen doorzoeken en analyseren, anderzijds om ze intern eenvoudiger te kunnen verwerken en redigeren.

Opmerking: Nederlab metadata is volledig CMDI compliant.

2.1.2 Wat doet Nederlab met tekstbronnen?

Extraheren van 'basis-FoLiA'
Nederlab en Nederlab analyseprocessen zijn primair gericht op tekstinhoud van de bronnen, voor ieder ander gebruik van de bronnen (inclusief scan-images) linken we door naar de betreffende bron bij de aanleverende organisatie (b.v. Delpher of de DBNL site). We extraheren daarom de tekstinhoud, plus enkele basale kenmerken per tekstblok (zoals koptekst of niet, redactioneel commentaar of basistekst, paragraafindeling). Deze informatie wordt geconverteerd naar het z.g. FoLiA (http://proycon.github.io/folia/index.nl.html) of een equivalent formaat. Dit is nodig voor uniforme verwerking, verrijking, indexering en back-up.

Extraheren van metadata
Soms is in de tekstbronnen zelf nog metadata opgenomen. Voor zover voor ons van nut extraheren we die metadata en behandelen deze als onder 2.1.1. We gaan ervan uit dat metadata rechtenvrij beschikbaar is, dus ook de uit tekstbronnen geëxtraheerde metadata.

Opsplitsen en samenvoegen
Dit verloopt parallel met het opsplitsen en samenvoegen van metadata-beschrijvingen, zie 2.1.1, identificeerbare resources

2.2 Interne opslag

2.2.1 Metadata

Originele, *geharveste* metadata worden alleen bewaard ten behoeve van eventueel snel herconverteren en herindexeren. Van geconverteerde metadata worden alle versies bewaard in de vorm van een interne metadata-database.

2.2.2 Tekstbronnen

Originele tekstbronnen (b.v. TEI documenten) worden momenteel tijdelijk lokaal opgeslagen ten behoeve van conversie- en extractieprocessen. Op termijn slaan we alleen basis-FoLiA's op

- als back-up ten behoeve van eventueel snel herindexeren;
- als onderdeel van de Nederlab index (zie verder 2.4);
- om stapsgewijs te verrijken;
- om als verrijkte teksten te tonen in de Nederlab onderzoeksomgeving, voor zover dat is toegestaan.

2.3 Dataverrijking

Een groot deel van de toegevoegde waarde van Nederlab komt tot stand door dataverrijking. Alle verrijkingen zijn in principe voor iedereen vrij beschikbaar, waarbij we in acht nemen dat uit doorgegeven verrijkingen geen (onderdelen van) integrale teksten mogen kunnen worden gereconstrueerd.

2.3.1 Preprocessing

Extra parallelle tekstlagen worden geautomatiseerd toegevoegd aan de basis-FoLiA's. Voorbeelden: automatisch genormaliseerde tekst (TICCL), woordklassen, lemma's, *named entities*. Daarbij worden verwijzingen naar de oorspronkelijke tekstdocumenten zo goed mogelijk gehandhaafd., waardoor ze mogelijk ook van waarde zijn voor de collectieleveranciers.
Dit is een intern proces, waarbij geen eindgebruikers betrokken zijn.

2.3.2 Verrijking door eindgebruikers

We maken het mogelijk dat eindgebruikers zowel Nederlab metadata als Nederlab teksten handmatig kunnen annoteren. Daarnaast willen we eindgebruikers ook toestaan geselecteerde integrale teksten met in Nederlab ingebouwde automatische tools te annoteren. Omdat deze tools op Nederlab servers draaien hoeven integrale teksten het Nederlab systeem daarvoor niet te verlaten.

2.3.3 Metadata verrijking door de Nederlab redactie

De Nederlab-redactie voert talloze verbeteringen en aanvullingen door, daarbij geholpen door (semiautomatische) redactietools.

2.4 Doorzoekbaarheid en indexering

Grootschalig efficiënt doorzoeken van de combinatie van metadata, fulltext en dataverrijkingen vergt de opbouw van efficiënte en slimme indexen. Bovendien willen we dat deze indexen veel afgeleide en geaggregeerde data opleveren, zoals tellingen van het voorkomen van fenomenen en patronen, en verdelingen (b.v. over tijd en ruimte).
Deze indexen bevatten noodzakelijkerwijs varianten van de aangeleverde metadata en teksten, ook integrale teksten. Dat betekent echter niet dat eindgebruikers teksten die ze hebben gevonden ook als tekst beschikbaar krijgen. We beschouwen de teksten in de indexen als interne kopieën van de data.

2.5 Eindgebruik

2.5.1 Wie zijn de beoogde eindgebruikers?

Beoogd eindgebruikers voor Nederlab zijn onderzoekers (en studenten) in de humaniora, in het bijzonder taalkundigen, letterkundigen en historici. Het soort gebruik zal naar verwachting heel divers zijn, maar altijd voor wetenschappelijke doeleinden. Nederlab heeft niet als doel enig commercieel gebruik te ondersteunen.

2.5.2 Hoe krijgen eindgebruikers toegang tot de (meta)data?

Gebruikers hebben toegang tot de data via een aantal technische kanalen.

Virtuele onderzoeksomgeving
In de verreweg de meeste gevallen zullen mensen Nederlab gebruiken via onze interactieve webapplicatie. Gebruikersauthenticatie (login) in een aantal vormen is al ingebouwd. Er zijn allerhande technische mogelijkheden om te garanderen dat we binnen de met de dataleveranciers overeengekomen licentievoorwaarden blijven. Voor meer details, zie hoofdstuk 3.

Search broker
De belangrijkste bron van informatie voor onze virtuele onderzoeksomgeving is de z.g. search broker, een REST-achtige webgebaseerde zoek-API. In principe is deze API ook direct door eindgebruikers aan te spreken. Door gebruik van API-keys zullen we toegang tot de data controleren. Ook de broker zal overeengekomen licentievoorwaarden in acht nemen. Voor een uitwerking van de daarvoor beschikbare middelen, zie hoofdstuk 3.

CMDI export
Nederlab maakt haar verrijkte metadata vrij beschikbaar in de vorm van CMDI bestanden, onder de aanname dat metadata rechtenvrij zijn. Deze bestanden bevatten ook referenties naar de metadata records waarvan ze zijn afgeleid, liefst in de vorm van *resolvable* links naar de websites van de oorspronkelijke leveranciers.

Downloaden van geselecteerde tekstbestanden
Een beperkte groep onderzoekers heeft behoefte aan het downloaden van geselecteerde sets van tekstbestanden. Het gaat hier vooral om taaltechnologen, die deze collecties batchmatig willen onderwerpen aan hun eigen algoritmen. Er zijn twee mogelijkheden:

- Men wil de volledige rijkdom van de broncollectie gebruiken. In dat geval fungeert Nederlab slechts als een op de doelgroep toegespitste methode om teksten te selecteren. De onderzoeker in kwestie kan dan op basis van deze selectie zelf een gebruiksovereenkomst aangaan met de collectieleverancier(s).
- Men wil profiteren van de door Nederlab toegevoegde meerwaarde. In dit geval dienen extra voorzieningen getroffen te worden. Zie hiervoor 3.4.

3 Middelen om aan eventuele licentievoorwaarden te voldoen

3.1 Gebruikersmanagement

We reguleren wie toegang heeft tot Nederlab, en welke rechten die gebruiker heeft. In principe zijn er drie niveaus van toegang tot de Nederlab virtuele onderzoeksomgeving:

- Vrije toegang, zonder authenticatie.
- Inloggen met een *federated identity* (Surfcontext, CLARIN federatie). Dit garandeert dat de betreffende gebruiker een account heeft bij een

wetenschappelijke instantie. We weten dan wie het betreft, welke instantie en mogelijk wat diens functie is.
- Inloggen met een lokaal Nederlab account, voor beheer en speciale gevallen waar maatwerk is vereist.

Gegeven bovenstaande inlogmethoden registreren we wie wat doet, zodat eventueel misbruik door eindgebruikers kan worden getraceerd.

In onze rol als 'doorgeefluik' van de licentievoorwaarden van collectieleveranciers kunnen we nieuwe eindgebruikers eenmalig vragen in te stemmen met licentievoorwaarden. Eventueel kan deze instemming via email aan gebruiker en/of collectieleverancier worden doorgestuurd.

Het is daarvoor wenselijk deze licentievoorwaarden zodanig op te stellen, dat ze de instemming van alle collectieleveranciers hebben.

3.2 Tonen van teksten

Het meest gevoelige punt is het beschikbaar maken van fragmenten van de oorspronkelijke teksten aan eindgebruikers. Er zijn een aantal middelen denkbaar om dat op een voor alle partijen acceptabele manier te regelen. Wat wel en niet acceptabel is, moet nader worden besproken.

Filteren van resources op basis van metadata
Als de rechtenstatus per resource of (sub-)collectie bekend is, kunnen we deze in de Nederlab metadata opnemen. Dat maakt het mogelijk het vrij tonen van teksten te beperken tot rechtenvrije teksten.

Tonen van tekstfragmenten
Voor een aantal toepassingen is het essentieel dat Nederlab eindgebruikers tekstfragmenten kunnen bekijken in de context van de Nederlab virtuele onderzoeksomgeving. Voorbeelden zijn:

- als onderdeel van resultaatlijsten, om gevonden zoektermen in context te kunnen zien, en daarmee resultaten te kunnen evalueren;
- als onderdeel van concordanties;
- als basis voor het tonen van lagen met verrijkingsdata in context: bijvoorbeeld, een sequentie van woordsoorten is alleen te interpreteren samen met de bijbehorende woordvolgorde.

Als we auteur en herkomst bij deze tekstfragmenten tonen, kunnen we dergelijk gebruik van tekstfragmenten als citaten beschouwen.

Om lezen van grote hoeveelheden tekst te ondersteunen kunnen de volgende mogelijkheden worden overwogen:

- een maximale grootte voor overdracht van tekstfragmenten afspreken: gebruikers kunnen dan slechts één fragment tegelijk opvragen;
- een webpagina van de collectie eigenaar integreren in de Nederlab omgeving;

- volgorde van tekstsegmenten randomiseren waar dat kan (b.v. bij het tonen van concordanties);
- in gevallen waar integratie in de Nederlab omgeving niet van belang is voor de gebruikerservaring kan worden volstaan met het openen van een document uit de bron-collectie in een apart browservenster.

3.3 Web service toegang

Bij toegang via een web-API, zoals de Nederlab search broker, wordt authenticatie door middel van API-keys ingezet om rechtmatig gebruik te garanderen. Daarnaast wordt gezorgd dat reconstructie van bronnen met behulp van *crawling* onmogelijk is, en nemen we de nodige maatregelen om te voorkomen dat search engines data binnen Nederlab kunnen indexeren.

3.4 Downloaden van teksten

In het geval van rechtstreeks downloaden van complete teksten vanuit Nederlab heeft Nederlab geen technische middelen meer voorhanden om legaal gebruik te garanderen. Standaard doen we dat dus niet. Desgewenst kunnen we de gebruiker aangescherpte licentievoorwaarden laten accepteren, die aanvullende beperkingen stellen aan type gebruik, tijdsduur van gebruik en verdere verspreiding.

4 Contracten en contractpartijen

Nederlab streeft na een intermediair te zijn tussen enerzijds meerdere partijen, die collecties beschikbaar stellen en anderzijds een veelvoud aan wetenschappelijke eindgebruikers/wetenschappelijke projecten. Zo beschouwd is Nederlab zelf geen eindgebruiker of onderzoeksproject en is het dus wenselijk licentieovereenkomsten op maat aan te gaan.
In het ideale geval zou er één modelcontract moeten komen voor overeenkomsten tussen collectieleveranciers en de penvoerder van Nederlab (KNAW) en één modelcontract voor gebruik van rechtendragende informatie vanuit Nederlab. Dit laatste contract moet ook alle contractvoorwaarden van collectieleveranciers doorgeven aan eindgebruikers.

5 Tegenprestatie

Het beschikbaar stellen van collecties voor gebruik in Nederlab heeft ook meerwaarde voor de betrokken leveranciers:

- de collecties worden zo door meer mensen en voor een breder scala aan wetenschappelijke toepassingen gebruikt. De collectieleverancier levert zo een zichtbare bijdrage aan vernieuwend geesteswetenschappelijk onderzoek;
- bij ieder dataelement wordt daarom zo goed mogelijk de oorspronkelijke herkomst getoond;
- aanleverende organisaties worden duidelijk op de Nederlab website genoemd;
- dataverrijkingen, crosslinks met andere collecties en curatieresultaten van de Nederlab-redactie zijn vrij beschikbaar en kunnen teruggeleverd worden aan de collectieleveranciers, zo mogelijk in relatie tot hun oorspronkelijke data.