Op en in het web: Hoe de toegankelijkheid van rechterlijke uitspraken kan worden verbeterd

van Opijnen, M.

HOOFDSTUK 6

Uitspraken in context

6.1 Inleiding

'Contextualiteit' hebben we benoemd als een van de deelaspecten van het toegankelijkheidsaspect 'hanteerbaarheid';[1870] het ziet op de mogelijkheid om een uitspraak te beschouwen in relatie tot andere juridische informatieobjecten. Vier soorten relaties zijn voor de jurist in het bijzonder van belang.[1871]

Ten eerste zijn er de formele relaties met andere uitspraken, die ook vallen onder het toegankelijkheidsaspect 'betrouwbaarheid'.[1872] De inzichtelijkheid van deze relaties liet lange tijd te wensen over;[1873] na afronding van het project Nova Porta Iuris worden deze verbanden inzichtelijk gemaakt.[1874]

Het tweede type wordt gevormd door de materiële relaties tussen uitspraken. Deze relaties kunnen betrekking hebben op zaken als een vergelijkbaar feitencomplex of eenzelfde rechtsvraag. Met inzicht in dergelijke relaties kan de jurist bijvoorbeeld kennis vergaren over jurisprudentielijnen. Deze relaties komen aan de orde in § 6.3.

Het derde type relatie betreft de beschouwing van een uitspraak in rechtswetenschappelijke literatuur. Daarbij zijn niet alleen annotaties – die zich meestal exclusief tot de analyse van één uitspraak beperken – van belang, maar alle plaatsen in de doctrine waar een uitspraak aandacht krijgt, zoals compendia, tijdschriftartikelen of handboeken. Deze relaties zijn het onderwerp van § 6.4.

Het vierde type relatie ten slotte is die tussen uitspraken en wet- en regelgeving. Door inzichtelijk te maken welke regelgeving een rechter in zijn oordeelsvorming betrekt, zouden we alle uitspraken kunnen vergaren die betrekking hebben op een specifiek wetsartikel. Deze wetsverwijzingen zijn het onderwerp van § 6.5. In § 6.6 sluiten we het hoofdstuk af met enkele conclusies en een vooruitblik.

De grootste uitdaging bij het inzichtelijk maken van de context van een uitspraak ligt in het expliciteren van de genoemde relaties. Deze verbanden worden door rechters en wetenschappers in door de hen geproduceerde teksten wel gelegd, maar meestal niet op een manier die voor computers makkelijk te begrijpen is. In dit hoofdstuk gaat onze aandacht dan ook vooral uit naar dit technische probleem. Wanneer de relaties eenmaal zijn vastgelegd, kunnen deze op verschillende manieren worden gebruikt, bijvoorbeeld om zoekvragen te formuleren of om netwerkrelaties in beeld te brengen. Dergelijke toepassingen kunnen worden ontwikkeld voor specifieke informatiedomeinen of gebruikersgroepen. Vanwege deze specificiteit zullen we, afgezien van enkele generieke voorbeelden, op dergelijke toepassingen niet verder ingaan. Een meer fundamentele beschouwing van deze relaties volgt in hoofdstuk 7, waar we de gegevens die we met behulp van de in dit hoofdstuk beschreven

1870 *Vide supra*: § 2.4.5.
1871 Daarnaast zijn er natuurlijk andersoortige informatieobjecten waarmee uitspraken relaties kunnen hebben, zoals Kamerstukken of krantenartikelen. Vanwege geringer belang en lagere frequentie laten we deze buiten beschouwing.
1872 *Vide supra*: § 2.4.4.
1873 *Vide supra*: § 3.4.3.3.
1874 *Vide supra*: § 4.5.3.1.

technieken hebben verzameld, mede gebruiken voor het ontwikkelen van een maat voor juridische domeinrelevantie.

Met andere woorden: naast een zelfstandige functie om de context van een uitspraak in kaart te brengen, zijn de wetsverwijzingen en jurisprudentiecitaties die in dit hoofdstuk expliciet worden gemaakt, ook een belangrijke grondstof voor het onderzoek van hoofdstuk 7.

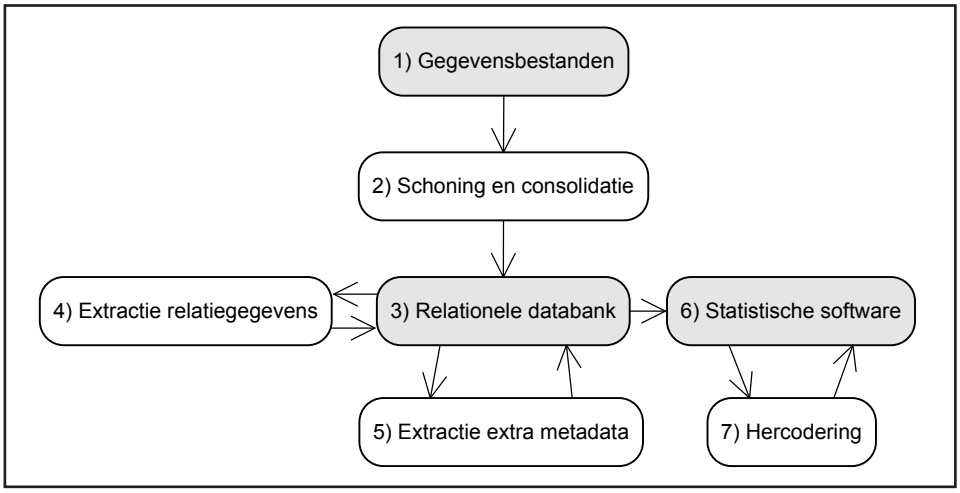
Voordat we beschrijven hoe de relaties worden geëxpliciteerd, beginnen we in § 6.2 met een overzicht van samenstelling en opbouw van de onderzoeksdatabank die we zowel voor hoofdstuk 6 als voor hoofdstuk 7 benodigen.

Ten slotte merken we hier op dat het onderzoek in dit (alsmede in het volgende) hoofdstuk is verricht voordat de ECLI in Nederland werd geïntroduceerd. Het is volledig gebaseerd op het LJN, maar kan zonder problemen worden vertaald naar een informatiearchitectuur die op ECLI is gegrond.

6.2 De onderzoeksdatabank

In deze paragraaf beschrijven we welke gegevensbestanden voor het onderzoek zijn gebruikt en welke generieke IT-processen zijn ontwikkeld om de ruwe bestanden om te zetten in bruikbare data. De afbeelding in Figuur 6-1 kan daarbij als leidraad dienen.

Het proces begint bij de ruwe gegevensbestanden (1), die worden beschreven in § 6.2.1. Deze bestanden konden niet zonder meer in een databank worden ingelezen, maar moesten in veel gevallen eerst worden geschoond en geconsolideerd (2). De hoofdlijnen van dit proces beschrijven we in § 6.2.2, de databank zelf (3) komt aan de orde in § 6.2.3. De volgende



Figuur 6-1. Overzicht van generieke IT-processen bij verzameling onderzoeksgegevens. Grijze vlakken zijn vormen van dataopslag, witte vlakken zijn processen.

stap is het extraheren van de relatiegegevens (4). Omdat dit geen generieke IT-processen zijn, worden deze niet in deze § 6.2 beschreven, maar in § 6.3 tot en met § 6.5. Stap (5), het verzamelen van enkele aanvullende metagegevens, beschrijven we in § 6.2.4. De gegevens die nodig zijn voor het onderzoek van hoofdstuk 7 werden van de databank geëxporteerd naar statische software (6). Daarbij vond soms ook hercodering van data plaats (7) Deze stappen worden besproken in § 6.2.5.

Deze § 6.2 beoogt slechts een beknopt overzicht te geven van de gevolgde stappen. Details van allerlei processen waarbij zich geen bijzonderheden voordeden met betrekking tot de juridische eigenaardigheden van de data, zijn niet beschreven. Een overzicht van alle voor het onderzoek gebruikte software is opgenomen in Bijlage 19.

6.2.1 Gebruikte gegevensbestanden

In § 5.2.2.1.2 is uitgelegd waarom netwerkanalyse zo veel mogelijk data vereist. We hebben er daarom voor gekozen om alle uitspraken te gebruiken die tot onze beschikking stonden:

1. Alle 240.947 uitspraken uit de CJO[1875] voor zover deze op 15 juni 2010 waren gepubliceerd op Rechtspraak.nl, in de databank RO-breed of in een huisdatabank.
2. Alle 523.224 uitspraken uit het E-archief[1876] die daarin op 15 juni 2010 aanwezig waren. Zoals eerder besproken[1877] zijn E-archief en CJO volledig gescheiden, waardoor uitspraken zowel in de ene als in de andere databank voor kunnen komen. Om deze dubbelingen zo goed als mogelijk te elimineren is het algoritme toegepast dat ook wordt gebruikt om te bepalen of uitspraken die met de Bimug worden ingevoerd reeds in de CJO aanwezig zijn.[1878] Uit deze toets kwamen 42.949 dubbelingen tevoorschijn, zodat 480.275 uitspraken uit het E-archief zijn overgenomen.
3. Alle uitspraken die op 15 juni 2010 aanwezig waren in de jurisprudentiedatabanken van Kluwer en Sdu, voor zover tenminste opgenomen in de LJN-index.[1879] Ook voor deze uitspraken heeft een ontdubbeling plaatsgevonden, omdat veel van deze uitspraken bijvoorbeeld reeds waren gepubliceerd op Rechtspraak.nl of in andere tijdschriften. Er zijn 132.640 uitspraakteksten ingelezen die uitsluitend in jurisprudentietijdschriften zijn gepubliceerd.

1875 *Vide supra*: § 3.2.1.1.
1876 *Vide supra*: § 3.2.1.5.
1877 *Vide supra*: § 3.2.1.5.
1878 *Vide supra*: § 4.2.3.1.
1879 Een overzicht is opgenomen in Bijlage 11.

Type	Aantal titels	Aantal bestanden
Annotaties	41	77.498
Tijdschriftartikelen	29	39.704
Handboeken en commentaaredities	114	439.088
Wiki Juridica	1	1.298
Totaal	185	557.588

Figuur 6-2. Ingelezen bestanden met rechtswetenschappelijke literatuur.[1884]

In totaal bevat de onderzoeksdatabank derhalve 853.862 uitspraken.[1880] Naast de uitspraakbestanden zelf hebben we eveneens de volledige (rechtspraak-interne) LJN-index[1881] ingeladen.

Rechtswetenschappelijke literatuur hebben we betrokken van de juridische portalen van de uitgevers Kluwer en Sdu.[1882] Een overzicht van de gebruikte titels staat in Bijlage 12. Ook de lemmata van Wiki Juridica beschouwen we als juridische literatuur.[1883] De omvang van de verschillende bronnen is weergegeven in Figuur 6-2.

De laatste bron die werd ingelezen betrof de actualiteiten die op Rechtspraak.nl worden gepubliceerd. Een groot deel van het historisch archief is bij vernieuwing van Rechtspraak.nl in april 2011 van de website verwijderd, maar in juli 2010 is door ons een afslag gemaakt van alle toen aanwezige 11.337 actualiteiten.

6.2.2 Schoning en consolidatie

Sommige gegevensbestanden waren gemakkelijk te verkrijgen. De rechtspraak-interne uitsprakendatabanken waren bijvoorbeeld als relationele databank aanwezig,[1885] en konden daardoor relatief eenvoudig worden ingelezen.[1886] Voor de vergaring van andere documenten moesten uiteenlopende technische problemen worden opgelost. Zo was één van de uitgevers

1880 Wel beschikbaar in Porta Iuris, maar niet ingelezen zijn de Jura-databank van de AB RvS (*vide supra*: § 3.2.1.1) en EUR-Lex. Naar uitspraken in Jura bestaan geen zelfstandige verwijzingen, naar EUR-Lex-uitspraken wel. Deze verwijzingen zijn natuurlijk wel herkend en in de calculaties meegenomen. Daarnaast zij opgemerkt dat veel HvJ-EU-uitspraken ook zijn gepubliceerd in de commerciële tijdschriften, en daarom wel verder zijn geanalyseerd.

1881 *Vide supra*: § 3.2.1.3.

1882 De overheidsbrede contracten die met deze uitgevers zijn gesloten staan dergelijk hergebruik toe.

1883 *Vide supra*: noot 888.

1884 Bij tijdschriftartikelen en annotaties komt één bestand overeen met één artikel respectievelijk één annotatie, bij de handboeken en commentaaredities bevat één bestand vaak niet meer dan een enkele paragraaf.

1885 *Vide supra*: § 3.2.1.

1886 Het woord 'relatief' wil hier slechts zeggen dat informatie in de bronvelden niet hoefde te worden gesplitst, geconcateneerd of bewerkt. Iedereen die wel eens databankconversies heeft gedaan, weet dat deze vrijwel nooit het predicaat 'eenvoudig' verdienen.

niet in staat om de vakliteratuur in XML te leveren, waardoor een applicatie moest worden ontwikkeld die de HTML-content van het uitgeversportaal schraapte. Een andere uitgever leverde wel XML, maar vanwege de incrementaliteit van de bestanden moest een applicatie worden ontwikkeld die uit soms tienduizenden losse bestandjes geconsolideerde versies van handboeken en commentaaredities opbouwde.

Het inlezen van de bestanden en de essentiële metadata leverde daarna vaak nog aanvullende problemen op. Soms moesten eerst parsers worden geschreven om essentiële gegevens uit slecht gestructureerde HTML of zelfs volledig onopgemaakte tekst te distilleren. Na het onttrekken van deze metadata werden de bestanden van eventueel resterende opmaak geschoond alvorens ze in de databank werden ingelezen. Dergelijke processen zijn hier verder niet beschreven, omdat ze niet worden gedicteerd door het specifiek juridische karakter van de inhoud. Dat neemt overigens niet weg dat deze werkzaamheden technisch gecompliceerd en zeer tijdrovend waren.

6.2.3 Databank

Alle verzamelde gegevens zijn uiteindelijk opgenomen in één relationele databank. Deze bereikte uiteindelijk een omvang van 69 gigabyte. De informatie is opgeslagen in 284 velden, verdeeld over 59 tabellen.

6.2.4 Metadata

Sommige metadata waren – goed gestructureerd – al aanwezig in de bronbestanden, en konden daarom direct in de databank worden ingelezen. Dit betrof bij uitspraken bijvoorbeeld gegevens over rechterlijke instantie, zaaknummer, uitspraakdatum en rechtsgebied, en bij literatuur gegevens over uitgever, publicatiejaar, titel en auteur.

Sommige metadata waren echter niet (altijd) in de bronbestanden aanwezig en moesten daarom met maatwerksoftware alsnog opgespoord worden. Voor het tellen van de lengte van een uitspraak volstond een eenvoudig script, maar bij andere gegevens was dit iets ingewikkelder. We bespreken hier achtereenvolgens: het aantal behandelende rechters (§ 6.2.4.1) en de actualiteiten van Rechtspraak.nl (§ 6.2.4.2).

6.2.4.1 Behandelende rechters

Voor het onderzoek in hoofdstuk 7 willen we graag weten of een uitspraak is gedaan door een enkel- of meervoudige kamer. Voor de uitspraken uit de CJO kon daartoe gebruik worden gemaakt van proceduresoort en sector, waaruit in sommige gevallen ook de enkel- of meervoudigheid van de behandelende kamer blijkt.[1887] In andere gevallen bood alleen

1887 Het aantal rechters kan worden afgeleid bij de proceduresoorten 'eerste aanleg – enkelvoudig', 'eerste aanleg – meervoudig', 'kort geding', 'voorlopige voorziening', en bij de sectoren 'kanton', 'Rechtseenheidskamer', 'president' en 'voorzitter'.

het tellen van de namen van de rechters soelaas. Deze worden in de databanken van de rechtspraak echter niet of slechts zeer sporadisch als afzonderlijk gegeven geregistreerd.[1888] Daarom hebben we een functie ontwikkeld die de namen met behulp van een reguliere expressie uit de uitspraak distilleert. De schrijfwijze van deze namen in de ondertekening van een uitspraak kent een enorme variëteit, inclusief tal van varianten die men redelijkerwijs niet zou verzinnen.[1889]

In de meeste commerciële databanken zijn de namen van rechters wel opgeslagen in een afzonderlijk veld. Alhoewel dit een enkelvoudig veld is, kon voor deze databanken een relatief eenvoudig algoritme volstaan om het aantal rechters te tellen.[1890]

6.2.4.2 Actualiteiten op Rechtspraak.nl

Anders dan de uitspraken hebben de actualiteiten van Rechtspraak.nl zich nooit bevonden in een relationele databank, maar zijn ze als HTML-bestanden in een *content management system* (CMS) opgeslagen. Omdat er in de loop der jaren diverse migraties zijn geweest tussen verschillende (versies van) CMS-en zit er weinig consistentie in URL's en metadata. Daarom moesten we een parser bouwen die de volgende informatie verzamelde: publicerende instantie, datum en de reikwijdte van de publicatie.[1891] Vanwege het louter technische karakter van deze parser zijn de details hier niet beschreven.

Lastiger is dat er vervolgens een onderscheid moet worden gemaakt tussen de uitspra(a)k(en) die de grondslag vorm(en) voor de actualiteit, en de uitspra(a)k(en) die om andere redenen daarin is (zijn) aangehaald. Het onderscheid moet uit tekst en metadata worden afgeleid. In het CMS is weliswaar een speciaal veld gereserveerd voor het 'grondslag-LJN' – zodat er automatisch een hyperlink naar deze uitspraak wordt aangemaakt – maar dit veld is vaak niet of onjuist gevuld. Een meer complex algoritme maakt daarom onderscheid tussen beide soorten. In de tekst van de actualiteit worden eerst alle verwijzingen naar uitspraken gemarkeerd met behulp van de nog te bespreken 'jurpointer-software'.[1892] Het beslissingsschema dat de parser vervolgens gebruikt om onderscheid te maken tussen de twee typen relaties is afgebeeld in het UML *activity diagram* van Figuur 6-3.[1893]

Aangehaalde uitspraken die moeten worden opgevat als citatie, worden als zodanig opgeslagen, aangehaalde uitspraken die worden gekwalificeerd als grondslag voor de actualiteit, worden met die specifieke hoedanigheid geregistreerd.

1888 Alleen in de RO-brede databank en de huisdatabank bestaat er een veld voor.

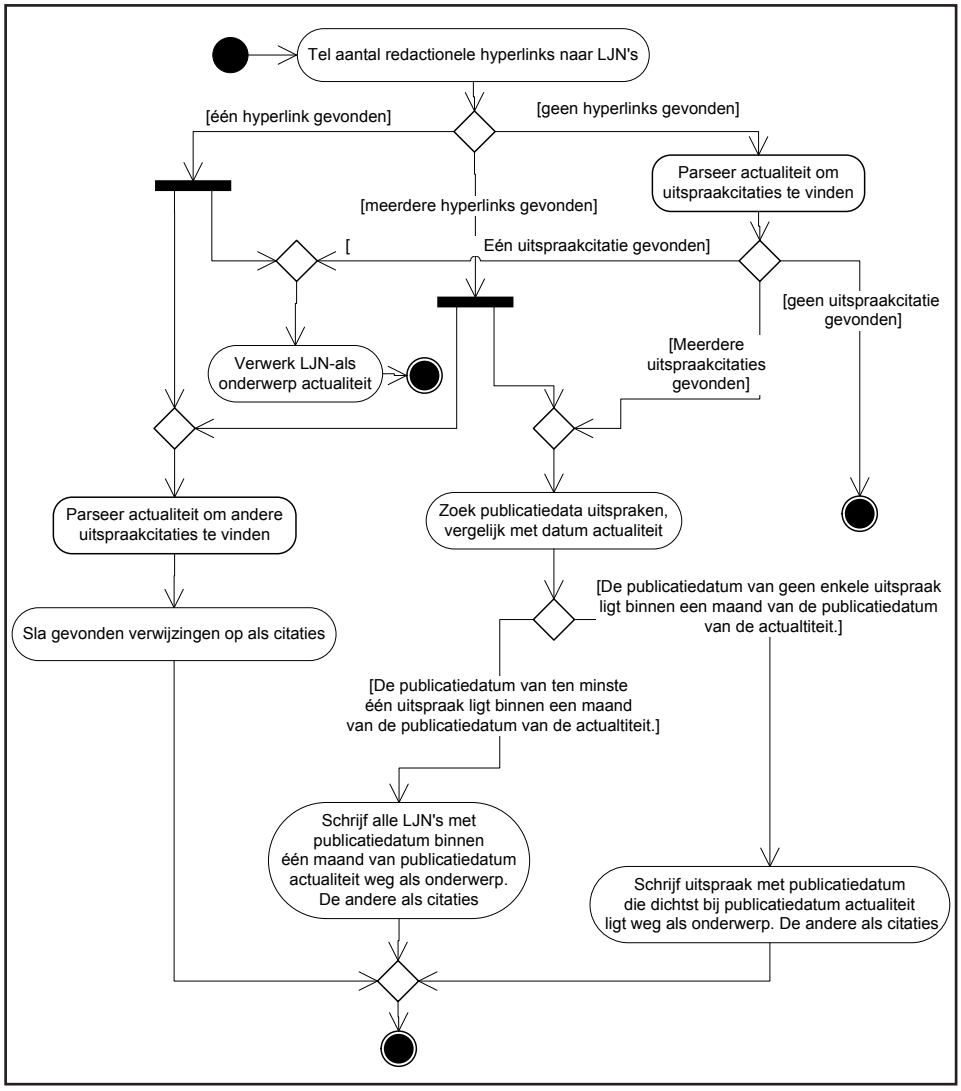
1889 Zoals het gebruik van harde regeleinden midden in een zin en foutieve spelling van academische titels.

1890 In de oorspronkelijke databanken staan de rechters niet altijd in een afzonderlijk metaveld, ze zijn bijv. vaak samengevoegd met AG of annotator. In Ibis (*vide supra*: § 3.2.1.4) worden de rechters evenwel reeds gescheiden van andere gegevens.

1891 Een lokale actualiteit staat alleen op de deelsite van het publicerende gerecht, een landelijke actualiteit staat tevens op de voorpagina van Rechtspraak.nl. In het CMS vervangt een landelijke actualiteit niet de lokale actualiteit maar kopieert deze. Dergelijke actualiteiten moesten daarom worden ontdubbeld.

1892 *Vide infra*: § 6.3.

1893 UML staat voor *Unified Modeling Language*. Het is een ISO-standaard voor het beschrijven van softwarearchitecturen.



Figuur 6-3. UML activity diagram voor het verwerken van uitspraakverwijzingen in actualiteiten op Rechtspraak.nl.[1894]

1894 De in dit diagram gebruikte symbolen:
- gevuld rondje: startpunt van het proces;
- omcirkeld rondje: eindpunt van het proces;
- afgeronde rechthoek: processtap;
- pijl: processtroom;
- ruit: keuzepunt (bij ten minste twee uitgaande pijlen) of samenkomst (bij ten minste twee inkomende pijlen);
- liggend balkje: splitsing van één stroom in twee parallelle stromen.

6.2.5 Omzetting naar statistische software en hercodering

6.2.5.1 Algemeen

Gegevensopslag in een relationele databank is zeer geschikt als basis voor de nog te beschrijven transactionele processen voor het extraheren van relatiegegevens,[1895] maar niet voor statistische analyses. Voor dit laatste moesten de data worden geëxporteerd naar verschillende statistische softwarepakketten.[1896] Dit vergde diverse technische conversies die we hier onbesproken laten. In de statistische software moesten de data vervolgens nog worden nabewerkt, bijvoorbeeld door groepering of hercodering. De enige hercodering waarin inhoudelijke keuzes zijn gemaakt, heeft betrekking op de rechtsgebieden, te bespreken in de volgende paragraaf.

6.2.5.2 Rechtsgebied

Het maken van een indeling in rechtsgebieden is arbitrair; elke keuze is vatbaar voor discussie en commentaar. Een hindernis wordt gevormd door de grote verschillen tussen de verzamelingen waaruit de onderzoeksdatabank is opgebouwd. Sommige hebben een redelijk fijnmazige indeling in rechtsgebieden, andere helemaal niet. En voor zover de verzamelingen wel rechtsgebiedclassificaties hebben, verschillen deze onderling sterk, hetgeen vergelijking bemoeilijkt. We hebben daarom gekozen voor de zeer basale indeling: strafrecht, civiel recht en bestuursrecht. Onderstaand lichten we toe hoe de metadata van de oorspronkelijke verzamelingen hierop zijn gemapt.[1897]

De uitspraken op Rechtspraak.nl hebben een indeling in twaalf rechtsgebieden, maar kennen geen indeling naar sectoren. De uitspraken in RO-breed en de huisdatabanken hebben een indeling van veertien rechtsgebieden. Het E-archief kent geen helemaal geen rechtsgebieden, maar wel een indeling naar zeven sectoren. De mapping tussen al deze metadata en de uiteindelijk gebruikte rechtsgebiedindeling is opgenomen in Bijlage 15.

Voor de databanken van de uitgevers lag de zaak anders. Een rechtsgebied is alleen aanwezig in de NJ, de andere databanken kenden geen indeling naar rechtsgebied.[1898] De meeste tijdschriften bestrijken evenwel een interessegebied dat gelijk is aan, of fijnmaziger dan de door ons gekozen rechtsgebieden. Zo konden we bijvoorbeeld zonder problemen de uitspraken in 'Jurisprudentie Onderneming en Recht' (JOR) indelen onder civiel recht en uitspraken in de 'Nieuwsbrief Strafrecht' onder strafrecht. Echt problematisch waren de tijdschriften die zich niet op een specifiek rechtsgebied richten, zoals het 'Nederlands Juristenblad' (NJB). Voor dergelijke periodieken is de arbitraire keuze gemaakt om de hierin gepubliceerde uitspraken in te delen bij civiel recht. De classificatie van alle tijdschriften is opgenomen in Bijlage 11.

1895 *Vide infra*: § 6.3 t/m § 6.5.

1896 Voor sociaalnetwerkanalyse is andere software gebruikt dan voor traditionele statistische analyse. Zie Bijlage 19.

1897 Voor de betekenis van 'mappen' *vide supra*: noot 907.

1898 De jurisprudentiedatabanken van de uitgevers zijn gebruikt zoals ze beschikbaar waren binnen Porta Iuris. Hierdoor konden we geen gebruik maken van verbeteringen die in de loop der tijd door leveranciers in de (meta)data zijn aangebracht.

6.3 Uitspraakcitaties

6.3.1 Inleiding

Voor een goed begrip van het vervolg beginnen we met een korte terminologische toelichting betreffende jurisprudentiecitatienetwerken.

Materiële relaties tussen uitspraken zijn expliciet of impliciet. Ze zijn expliciet indien de ene uitspraak de andere uitspraak letterlijk citeert. Impliciete relaties komen niet in de tekst tot uitdrukking, maar bestaan alleen in de ogen van een externe beschouwer. Van een impliciete relatie is ook sprake indien de niet-geciteerde uitspraak onderdeel is van een bestendige jurisprudentielijn die bekend wordt verondersteld.[1899] We beperken ons onderzoek tot expliciete materiële relaties.

In Figuur 6-4 is een eenvoudig voorbeeld gegeven van een uitspraak (B) die een andere uitspraak (A) citeert. Uitspraak B noemen we de 'aanhalende -', 'verwijzende -' of 'citerende uitspraak'. Maar we gebruiken ook wel begrippen als 'bronuitspraak' en 'citator'.[1900] Uitspraak A is de 'aangehaalde -', of 'geciteerde uitspraak', waarvoor we ook wel de termen 'doeluitspraak' en 'citandus' hanteren.

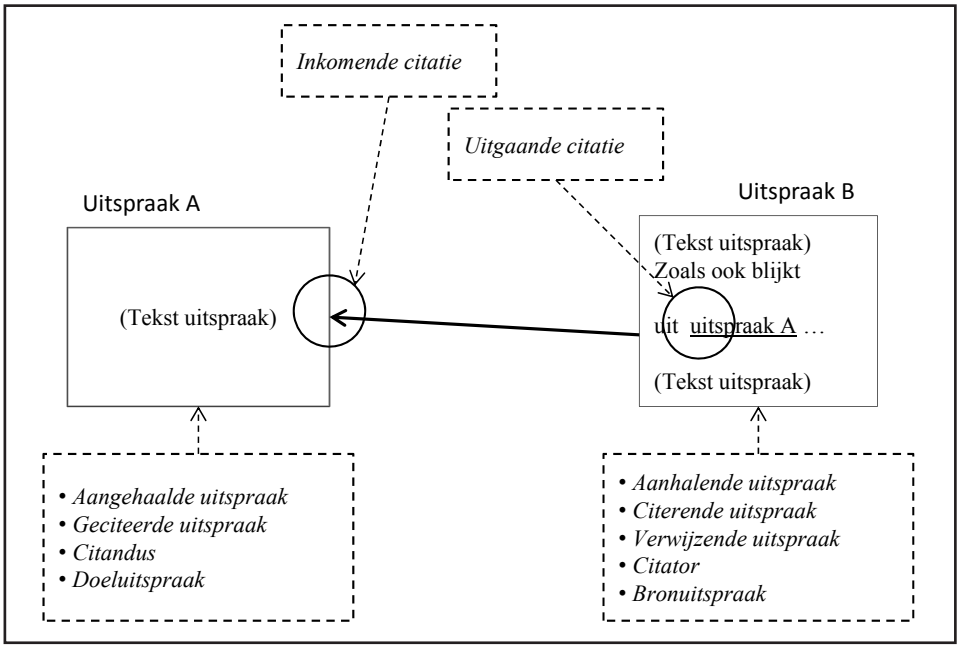
De citatie zelf is een enkelvoudig object, maar kan – afhankelijk van het perspectief van waaruit ernaar wordt gekeken – anders worden betiteld. Bezien vanuit uitspraak B is het een 'uitgaande citatie', bezien vanuit uitspraak A betreft het een 'inkomende citatie'.

Om citatierelaties tussen uitspraken in beeld te brengen en eventueel te analyseren, is een citatie-index nodig: een bestand waarin uitspraakcitaties gestructureerd zijn opgeslagen. Het maken van een dergelijke citatie-index is geen routineklus, want veel uitspraken in onze onderzoeksdatabanken hebben geen computerleesbare verwijzingen; citaties staan in de tekst zoals ze door de rechter zijn opgeschreven. Bij de bespreking van citatievoorschriften[1901] en citatiepraktijk[1902] hebben we gezien dat er vele manieren zijn om een uitspraak te citeren: tripletten, vindplaatsen, LJN of combinaties daarvan.

We illustreerden dit in Figuur 4-4, waarin negen manieren staan voor het citeren van één uitspraak. Verder dient te worden bedacht dat de volgorde van datum, instantie, zaaknummer, vindplaatsen en LJN kan variëren, dat er allemaal verschillende schrijfwijzen zijn voor (onderdelen van) identifiers, en dat deze nog kunnen worden gescheiden door allerlei andere woorden, zoals bijvoorbeeld in de volgende zinsnede:

> *(...) Zoals ook gelezen kan worden in Hoge Raad, LJN AO3858, zaaknr. C02/320HR d.d. 9 april 2004, JOL 2004/199 en NJ 2004/331. Een ander argument voor deze stelling kan worden gehaald uit (...)*

1899 Zie over dit verschijnsel '*obliteration through incorporation*' o.m. [Smith 2007, p. 348].
1900 In *common-law*-landen wordt '*citator*' ook als synoniem voor 'citatie-index' gebruikt, *vide supra*: § 5.2.2.2.
1901 *Vide supra*: § 4.3.1.
1902 *Vide supra*: § 4.3.3.



Figuur 6-4. Terminologie voor subject-predicaat-object-relaties in een jurisprudentiecitatienetwerk.

In dit voorbeeld is bovendien niet duidelijk hoeveel uitspraken er eigenlijk worden aangehaald. Het kunnen er één, twee of drie zijn. Om de bedoeling van de auteur te begrijpen moeten tekstelementen daarom niet alleen als jurisprudentiecitaties worden herkend, maar moeten deze citaties vervolgens ook worden 'gecanonicaliseerd':[1903] geconverteerd naar één canonieke vorm. Het ligt voor de hand om hiervoor het LJN te gebruiken, omdat dit een werkidentifier is. Echter, niet alle uitspraken hebben een LJN. Dat geldt ten eerste voor uitspraken die uitsluitend in het E-archief staan.[1904] Al deze uitspraken hebben een eigen (technisch unieke) sleutel die we in plaats van het LJN kunnen gebruiken. Een tweede groep zonder LJN bestaat uit uitspraken die, ook al staat het publicerende tijdschrift wel in de LJN-index, niet zijn gedigitaliseerd.[1905] Voor deze uitspraken gebruiken we de vindplaats als canonieke identifier. Omdat deze expressie-identifiers de enige bekende identifier zijn, is dit

1903 Het woord 'canonicaliseren' komt niet voor in Groene Boekje of Van Dale. Het is een vertaling van het Engelse IT-begrip *'to canonicalize'* – zie <en.wikipedia.org/wiki/Canonicalization> – dat overigens niet in de *Oxford Dictionary* voorkomt. Zoals uit het vervolg van deze paragraaf zal blijken, zien we, anders dan genoemde Wikipedia-pagina, 'canonicaliseren' niet als synoniem voor 'normaliseren'.

1904 *Vide supra*: § 4.2.3.2.

1905 *Vide supra*: § 4.2.3.2. Het betreft vooral NJ-uitspraken van voor 1965.

niet bezwaarlijk.[1906] Om de registratie van deze drie soorten van identificatie van uitspraken op werkniveau te vereenvoudigen worden ze in het vervolg onder één noemer geschaard: canonieke uitspraakidentifier (CUID).

In de volgende subparagrafen zetten we de stappen van het hele linkproces uiteen.[1907] [1908] Eerst wordt met patroonherkenning gezocht naar alles wat een (deel van een) jurisprudentiecitatie zou kunnen zijn (§ 6.3.3). Deze 'citatiestrings' worden vervolgens genormaliseerd (§ 6.3.4) en daarna gecanonicaliseerd (§ 6.3.5). We beginnen evenwel met enkele complexiteitsreducties (§ 6.3.2).

6.3.2 Beperkingen

De rol van de verwijzing in juridisch taalgebruik kan moeilijk worden onderschat:

> *Legal communication has two principal components: words and citations.*[1909]

Beide componenten zijn vaak niet gemakkelijk te onderscheiden, zeker wanneer een juridisch document geen prescriptief karakter heeft. Verwijzingen in wetteksten naar andere (onderdelen van) regelgeving zijn in hoge mate gestandaardiseerd;[1910] ook al worden vaak relatieve citaties gebruikt (zoals 'laatstgenoemd artikel' en 'het volgende lid'), het aantal constructen dat een parser moet kunnen begrijpen is dientengevolge beperkt. Verwijzingen in rechterlijke uitspraken vormen echter veel meer een integraal onderdeel van de lopende tekst, die bovendien veel vormvrijer is dan regelgevingteksten. Het schier eindeloze aantal verwijzingsconstructen dat daarvan het gevolg is, levert een aantal problemen op.

Een eerste probleem is het veelvuldig gebruik van globale en lokale aliassen. Onder een 'globale alias' verstaan we de roepnaam van een uitspraak, zoals 'Haviltex'[1911] of 'Lijn 10 en de omgevallen weduwe'.[1912] Terwijl een globale alias ook buiten de context van een citatie door een jurist zal worden herkend en begrepen, heeft een lokale alias alleen betekenis binnen een specifieke context. Zo kan een uitspraak bij herhaalde citatie bijvoorbeeld worden aangeroepen met de soortnaam ('het arrest'), met een niet-compleet triplet ('het arrest van de Hoge Raad van maart 2010'), of met hetgeen erin tot uiting wordt gebracht ('de opvatting van de Hoge Raad'). Om het nog ingewikkelder te maken kan laatstgenoemd

1906 Daarnaast zijn er nog citaties die een vindplaats gebruiken uit een tijdschrift dat niet in de LJN-index is opgenomen. Dergelijke citaties gebruiken over het algemeen ook een LJN, een volledig triplet of een andere, wel in de LJN-index opgenomen, vindplaats. Alleen waar deze niet bestaan, kan de citatie niet worden opgelost. De aantallen zijn verwaarloosbaar.

1907 In iets beknoptere redactie is deze paragraaf ook verschenen als [van Opijnen 2010a].

1908 Omdat iedere jurisdictie zijn eigen identificatiesystemen en citatiegewoonten heeft (*vide supra*: hfd. 4) kunnen systemen die in andere landen voor vergelijkbare taken zijn ontwikkeld hooguit ter inspiratie worden gebruikt. Zie voor (summiere) beschrijvingen van andere software voor het extraheren van jurisprudentiecitaties bijvoorbeeld [Leiter 2011a], [Needle 2000] en [Mowbray, Chung en Greenleaf 2009].

1909 [Shapiro 1991, p. 1453].

1910 *Vide supra*: § 4.7.

1911 ECLI:NL:HR:1981:AG4158.

1912 ECLI:NL:HR:1971:AC5093.

voorbeeld niet alleen verwijzen naar een daarvóór geciteerd Hoge-Raadarrest, maar ook naar een verzameling van geciteerde arresten, of op (deels) niet-geciteerde arresten van een bestendige jurisprudentielijn die bij de lezer bekend wordt verondersteld. Een gevolg van deze verwevenheid tussen tekst en citatie leidt tot de vraag wat eigenlijk als (herhaalde) citatie zou moeten tellen.[1913] Het antwoord op deze vraag is subjectief en zal in veel situaties tot discussie leiden.

Globale aliassen worden met de door ons gebouwde parser niet gedetecteerd; de reden daarvoor is niet zozeer gelegen in het niet kunnen herkennen ervan, als wel in het ontbreken van een gestructureerde verzameling met globale aliassen.[1914] Ook lokale aliassen worden niet gedetecteerd. Naast eerdergenoemde problemen is vooral het niet of slordig declareren[1915] van lokale aliassen door uitspraakconcipiënten debet aan die keuze.[1916] De parser herkent dus alleen verwijzingen die met behulp van de verschillende identifiers zijn gemaakt, maar telt daarvan vervolgens wel alle voorkomens.

Een tweede beperking – of zo men wil 'onnauwkeurigheid' – in de hieronder beschreven parser is dat geen onderscheid wordt gemaakt tussen aanhalingen van uitspraak, conclusie en annotatie. Een door de parser herkende verwijzing naar 'NJ 2010, 169' kan een verwijzing zijn naar het onder dit nummer gepubliceerde Hoge-Raadarrest, maar het kan ook een verwijzing inhouden naar de conclusie of een annotatie. Enerzijds omdat er geen specifieke identifiers zijn voor conclusies[1917] en annotaties,[1918] en anderzijds vanwege de benodigde extra complexiteit[1919] van het algoritme, is ervoor gekozen om verwijzingen naar annotaties en conclusies op te vatten als verwijzingen naar de uitspraak zelf.

Naast deze twee beperkingen in het herkennen van verwijzingen, zijn er nog enkele beperkingen ten aanzien van het gebruik van bij citaties behorende attributen. Deze worden besproken in § 7.2.3.1.

6.3.3 Detecteren van citatiestrings

De eerste stap die in het linkproces wordt uitgevoerd, bestaat uit het herkennen van alle strings (tekenreeksen) die mogelijk (onderdeel van) een uitspraakidentifier zijn. Iedere herkende string noemen we een 'citatiestring'. Er zijn vijf soorten citatiestrings: LJN, vind-

1913 Vgl. [Tapper 1982, p. 138].
1914 In het project Nova Porta Iuris (*vide supra*: § 3.2.3) is voorzien in het gestructureerd opslaan van aliassen.
1915 Met de term 'declareren' bedoelen we hier het specificeren van de lokale alias teneinde deze later te kunnen herkennen. Vaak wordt hiervoor een constructie gebruikt als: *"Uitspraak X, hierna te noemen 'de uitspraak'."*
1916 Bij in rechterlijke uitspraken gemaakte wetsverwijzingen is de situatie zowel voor globale als lokale aliassen anders: *vide infra*: § 6.5.
1917 Althans voor de introductie van ECLI, *vide supra*: § 4.5.3.1.
1918 *Vide supra*: § 4.8.2.
1919 Ook bij verwijzingen naar conclusies en annotaties is de schrijfwijze zeer divers, zo wordt bijv. vaak de naam van de AG of annotator gebruikt en niet de rol die door hem wordt bekleed.

plaats en de triplet-onderdelen instantienaam, datum en zaaknummer. De tekstherkenning geschiedt met behulp van reguliere expressies.[1920]

De reguliere expressies voor de verschillende citatiestrings worden in onderstaande paragrafen beschreven. Op deze plaats zij opgemerkt dat één tekenreeks vaak zowel voor zaaknummer als voor een andersoortige citatiestring kan worden aangezien. Om dit probleem op te lossen, zou men de parsers in een bepaalde volgorde door de teksten kunnen laten gaan. Een andere mogelijkheid is om alle parsers tegelijk te laten lopen, en daarna zaaknummer-citatiestrings die andere citatiestrings overlappen te verwijderen. Met het oog op snelheid hebben we voor deze laatste optie gekozen. We bespreken nu kort de parsers die voor de vijf verschillende citatiestrings zijn gemaakt.

6.3.3.1 LJN

Bij het maken van een reguliere expressie voor het LJN moet rekening worden gehouden met enkele bijzonderheden. Zo heette het LJN vroeger 'ELRO-nummer'[1921], wordt het LJN ook wel aangehaald als 'LJ-nummer', 'LJN-nummer', 'LJ-nr', et cetera,[1922] en komen er verschillen voor in interpunctie, spatiëring en hoofdlettergebruik. Zo kan 'LJN:AB1234' bijvoorbeeld ook worden geschreven als:

- ELRO-nummer AB1234
- LJN AB1234
- LJ-nr. AB 1234
- ljn: ab1234

Met de reguliere expressie:

```
((?i)(lj[\s-]?n?|elro)\s?((-?n?)?(umme)?)r?[:\.\s=]{1,2}[a-z]{2})\s?\d{4}
```

worden deze en tal van andere varianten herkend.

6.3.3.2 Vindplaatsen

Vindplaatsen zijn een stuk lastiger te herkennen dan LJN's. Niet alleen omdat er zoveel verschillende bronnen zijn, maar ook omdat er aanzienlijke verschillen bestaan in de schrijfwijze van zowel het naamdeel als het nummerdeel van een vindplaats.

Zo kan het tijdschrift 'Schip en Schade' ook worden aangehaald als:

- Schip & Schade
- SeS
- S&S

1920 Een grammatica voor het herkennen van tekstpatronen. Zie voor de syntaxis onder meer <www.regular-expressions.info/reference.html>. Vrijwel elke programmeertaal heeft zijn eigen dialect van reguliere expressies. Alle voorbeelden in dit boek zijn geschreven in het JGSoft-dialect.

1921 *Vide supra*: noot 898.

1922 *Vide supra*: § 4.3.3.

- SES
- S.E.S.

Bovendien hebben sommige tijdschriften in de loop der tijd het nummerdeel aangepast. Zo werd een uitspraak in VakstudieNieuws van Kluwer tot en met 1997 geïdentificeerd met 'jaartal+paginanummer', maar vanaf 1998 met 'jaartal+aflevering(+groepering)+volgnummer', te vatten in de reguliere expressie:

```
(V(-)?N(,|:)?\s(\d{2,4}(/|,\s)\d{1,2}(\.\d{1,2})*))
```

Een groot aantal tijdschriften heeft een vergelijkbare identifier, bestaande uit jaartal en volgnummer, maar sommige wijken af, zoals de 'Jurisprudentie' (Jur.) van het HvJ EU, waarvoor deze reguliere expressie nodig is:

```
Jur(ispr)?[EG\.\s-]{1,6}\d{4}[,\s-]{1,2}(b(l(ad)?z)?\.|p(ag)?\.)?\s?[IA-]{0,3}\d{1,5}
```

Voor alle in de LJN-index voorkomende tijdschriften, inclusief en JOL[1923] en Jur.[1924] zijn reguliere expressies geschreven.

6.3.3.3 Rechterlijke instanties

Alle rechterlijke instanties waarvan redelijkerwijs kan worden verwacht dat er in uitspraken naar wordt verwezen, worden als citatiestring herkend: Hoge Raad, CRvB, AB RvS (inclusief de rechtsvoorgangers Afdeling Geschillen van Bestuur en Afdeling Rechtspraak), CBb, gerechtshoven, rechtbanken, kantongerechten, EHRM (inclusief Commissie voor de Rechten van de Mens), HvJ EU (en rechtsvoorgangers) en Benelux Gerechtshof. Ook bij de instanties doet zich het probleem voor dat concipiënten een onuitputtelijke fantasie hebben bij het bedenken van afkortingen en het maken van schrijffouten. Zo kunnen in 'College van Beroep voor het bedrijfsleven' alle woorden op verschillende manieren worden afgekort en aan elkaar geplakt. Dat noopt tot de volgende reguliere expressie:

```
(?i:c(oll?(ege)?)?\.?\s?(v(an|\.)?)?\s?b(er(oep)?|\.)?\s?(v(oor|\.)?)?\s?(h(et|\.)?)?\
s?b(edr((ijfs|\.)?leven)?|\.)?)
```

De reguliere expressie voor het HvJ EU is nog een stuk complexer:

```
(?i:(e((u|g)r?(opee?se?)?)\s)?((h(of\s)?v(an)?\s?j(ustitie)?)|(g(er(echt)?)?\.?\
s?(v(an|\.)?|in)?\s?e(erste|\.)?\s?a(anleg|\.)?))(\s?((v(an|\.)?\s?(d(e)|\.)|der)?\
s?)?e(ur(opese)?)?\.?\s?(g(em(eenschap(pen)?)?)?\.?|u(nie|\.)?)))?
```

Vergelijkbare reguliere expressies zijn geschreven voor alle hierboven genoemde instanties.

1923 JOL staat voor 'Jurisprudentie On Line', een op 1 september 1999 gestart initiatief van uitgeverij Kluwer om alle (belangrijke) civiele en strafuitspraken van de Hoge Raad (gratis) online te publiceren ([Hertzberger, van der Wees en Renden 2002, p. 8]). Eind 2008 werd JOL weer opgeheven. Omdat in de eerste jaren van bestaan JOL soms de enige (publieke) bron van een uitspraak was, is er veelvuldig met het JOL-nummer geciteerd.

1924 Alle tijdschriften in de LJN-index zijn ook fysiek als databank bij de Rechtspraak aanwezig (*vide supra*: § 3.2.1.3), Jur. en JOL vormen hierop een uitzondering: alleen de identifiers zijn opgenomen.

6.3.3.4 Datum

Datums in tripletten blijken uitsluitend te worden geschreven met de namen van de maanden in letters, soms voluit geschreven, soms afgekort. De reguliere expressie kon daarom beperkt blijven tot:

```
[1-9][0-9]?\s(jan(uari|\.)?|feb(ruari|\.)?|m(aa)?rt\.?|apr(il|\.)?|mei|juni|juli|aug
(ustus| \.)?|sept?(ember|\.)?|o(k|c)t(ober|\.)?|nov(ember|\.)?|dec(ember|\.)?)
\s((18|19|20)|')[0-9]{2}
```

Hiermee worden natuurlijk ook allerlei datums gevonden die geen onderdeel zijn van een triplet, maar een andere betekenis hebben. Dit probleem zal in een later stadium worden opgelost.

6.3.3.5 Zaaknummer

Reeds in § 4.2.1 zijn we ingegaan op de vele verschillende manieren waarop zaaknummers kunnen worden geschreven. De reguliere expressie is daarom ruim van opzet:

```
([A-Z]*[0-9]+[-/.\s]+[0-9]{2,20}[-/.]?[A-Z]*)|\d{5,20}
```

Net als bij de datums zullen hiermee allerlei andere codes dan zaaknummers worden gevonden. Ook bij het zaaknummer zal dit in een latere fase worden opgelost.

6.3.4 Normaliseren van citatiestrings

Nadat de citatiestrings zijn herkend moeten ze worden genormaliseerd. Dat wil zeggen dat ze zodanig worden herschreven dat ze voldoen aan de schrijfwijze die wordt gehanteerd in de dataverzameling waarin ze in de hiernavolgende stap zullen worden opgezocht.

Voor verschillende citatiestrings gelden verschillende regels.

- Citatiestrings die zijn gevonden met de reguliere expressie voor het LJN, worden herschreven naar de officiële schrijfwijze, waarbij het label 'LJN' wordt weggelaten. 'LJ-nr. AB 1234' wordt derhalve 'AB1234'.
- Bij citatiestrings die zijn gevonden met de reguliere expressies voor vindplaatsen wordt het naamdeel genormaliseerd naar de afkorting die in de LJN-index wordt gebruikt, en het nummerdeel naar de schrijfwijze die daarin wordt gehanteerd, zijnde meestal de citatiemethode die door het tijdschrift zelf wordt voorgeschreven.[1925] 'Ned.Jur. '98/34' wordt dus genormaliseerd naar 'NJ 1998, 34'.

1925 Ibis (*vide supra*: § 3.2.1.4) normaliseert de verschillende schrijfwijzen die door uitgevers in hun bestanden worden gehanteerd reeds naar één formaat, dat in de LJN-index wordt gebruikt.

- Instanties worden genormaliseerd naar de betekenisloze numerieke identifier die ze hebben in zowel E-archief als LJN-index. 'HR', 'Hoge Raad' en 'Hoge Raad der Nederlanden' worden genormaliseerd naar '11'.
- Datums worden genormaliseerd naar ISO 8601: '9 jan. 2009' wordt dan: '2010-01-09'.
- Alle citatiestrings die zijn gevonden met de reguliere expressie voor zaaknummers worden niet genormaliseerd, omdat zaaknummers geen gestandaardiseerde schrijfwijze hebben.[1926]

Nu alle gevonden citatiestrings zijn genormaliseerd, kan worden overgegaan naar de volgende stap: het canonicaliseren.

6.3.5 Canonicaliseren van citaties

De code die in de voorgaande paragraaf is beschreven, leidt tot een XML-document waarin de citatiestrings wel zijn genormaliseerd, maar nog niet gecanonicaliseerd.[1927] In § 6.3.5.1 bekijken we eerst hoe deze XML eruit ziet. Vervolgens formuleren we de eisen waaraan de informatie na canonicalisatie zou moeten voldoen (§ 6.3.5.2), en definiëren we dit in een XML-schema (§ 6.3.5.3). Ten slotte behandelen we in § 6.3.5.4 het canonicalisatieproces zelf.

6.3.5.1 De XML na normalisatie

Het XML-document dat het resultaat is van de in § 6.3.3 en § 6.3.4 beschreven processen voldoet aan de schemastructuur van Figuur 6-5.

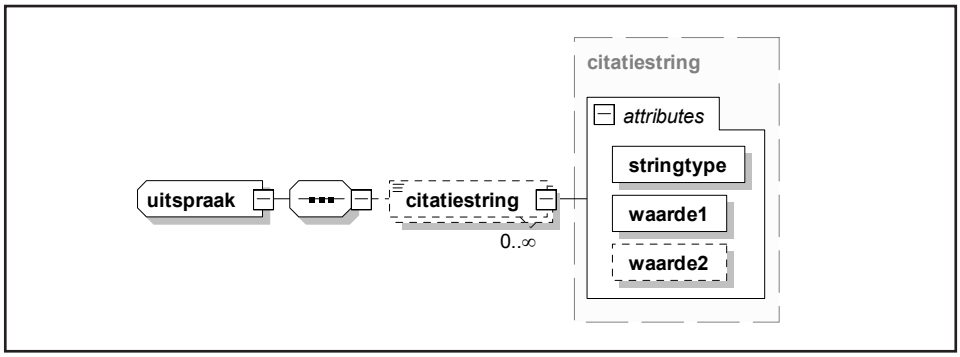
Iedere string in de tekst die (een onderdeel van) een jurisprudentiecitatie zou kunnen zijn, is getagd met het element 'citatiestring'. Dit element heeft twee of drie attributen: 'stringtype', 'waarde1' en 'waarde2'. Het attribuut 'stringtype' geeft weer met welke parser de citatiestring is gevonden; de mogelijke waarden zijn: LJN, vindplaats, instantie, datum of zaaknummer. Het attribuut 'waarde1' bestaat altijd en bevat de waarde van de gevonden citatiestring na normalisatie. Bij het stringtype 'vindplaats' bevat het attribuut 'waarde1' het naamdeel en bevat 'waarde2' het nummerdeel; bij alle andere mogelijke waarden van 'stringtype' is 'waarde2' leeg.

De XML van onze eerder gebruikte voorbeeldtekst:

> *(...) Zoals ook gelezen kan worden in Hoge Raad, LJN AO3858, zaaknr. C02/320HR d.d. 9 april 2004, JOL 2004/199 en NJ 2004/331. Een ander argument voor deze stelling kan worden gehaald uit (...)*

1926 *Vide supra*: § 4.2.1.

1927 Hier begint een zekere discongruentie op te treden tussen de logische beschrijving van de processen en de wijze waarop dit in code is geïmplementeerd. Uit een oogpunt van begrijpelijkheid en reproduceerbaarheid (ook in andere programmeertalen dan het door ons gebruikte C#) geven we de voorkeur aan de logische beschrijving.



Figuur 6-5. XML-schema voor uitspraakdocumenten na het parseren en normaliseren van citatiestrings.

ziet er na parseren en normaliseren dan als volgt uit:

> **<uitspraak>**
> (...) Zoals ook gelezen kan worden in
> **<citatiestring stringtype="instantie" waarde1="11">**Hoge Raad
> **</citatiestring>,**
> **<citatiestring stringtype="LJN" waarde1="AO3858">**LJN AO3858
> **</citatiestring>,** zaaknr.
> **<citatiestring stringtype="zaaknummer" waarde1="C02/320HR">**
> C02/320HR**</citatiestring>** d.d.
> **<citatiestring stringtype="datum" waarde1="2004-04-09">**9 april 2004
> **</citatiestring>,**
> **<citatiestring stringtype="vindplaats" waarde1="JOL" waarde2="2004, 199">**JOL 2004/199**</citatiestring>** en
> **<citatiestring stringtype="vindplaats" waarde1="NJ" waarde2="2004, 331">**NJ 2004/331**</citatiestring>**.
> Een ander argument voor deze stelling kan worden gehaald uit (...)
> **</uitspraak>**

Deze gestructureerde informatie kan als input dienen voor het canonicalisatieproces. Voor dit proces moeten we eerst enkele randvoorwaarden formuleren.

6.3.5.2 Functionele eisen voor canonicaliseren

'Canonicaliseren' hebben we omschreven als het converteren van verschillende identifiers naar één canonieke vorm.[1928] Dit houdt om te beginnen in dat tripletten, LJN's en vindplaatsen moeten worden gekoppeld aan de canonieke uitspraakidentifier (CUID). We kunnen echter

1928 *Vide supra*: § 6.3.1.

niet volstaan met een eenvoudige vervanging; we willen verschillende expressie-identifiers die over hetzelfde werk gaan, ook 'ontdubbelen': twee achter elkaar genoemde vindplaatsen die betrekking hebben op hetzelfde werk, moeten niet beide afzonderlijk, maar beide gezamenlijk worden vervangen door één CUID. Hierdoor ontstaat een zuiverder beeld van de aanwezige citaties, en kunnen – bij gebruik van het document in een gebruikersinterface – hyperlinks op de correcte manier worden getoond. Zonder ontdubbeling zouden we in ons tekstvoorbeeld met vier citaties eindigen, maar met ontdubbeling resteert er slechts één citatie (indien alle citatiestrings tenminste naar hetzelfde werk zouden verwijzen).

Naast de gecanonicaliseerde citaties hebben we echter ook behoefte aan enige aanvullende informatie. Zo willen we de aard van de relatie tussen citator en citandus weten. Deze kan drieërlei zijn: formeel, materieel of zelfreferentieel. Formele relaties hebben we eerder[1929] gedefinieerd als relaties met een wettelijke grondslag. Een citatie is zelfreferentieel als citator en citandus gelijk zijn. Materieel is iedere citatie die niet formeel of zelfreferentieel is.

Ook willen we weten – ten behoeve van onderzoek naar de citatiepraktijk[1930] – wat voor soort identifiers zijn gebruikt, en om technische redenen willen we iedere citatie als zodanig ook uniek identificeren. Ten slotte willen we vastleggen of het canonicalisatieproces succesvol doorlopen is.

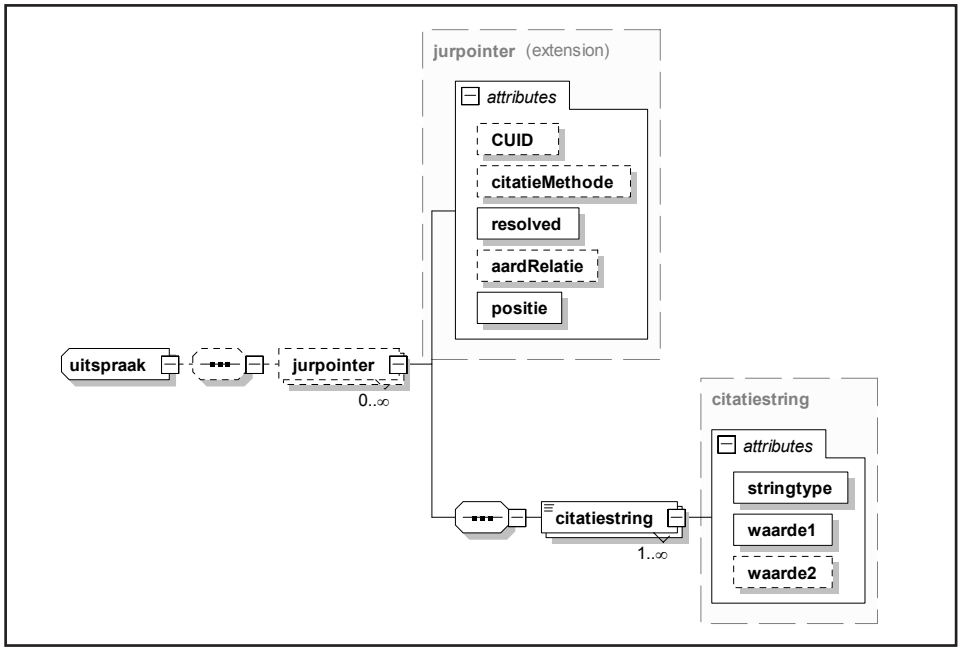
6.3.5.3 Het gewenste eindresultaat

Het XML-schema in Figuur 6-6 geeft weer hoe de XML er, gegeven de in de vorige paragraaf beschreven randvoorwaarden, na het canonicalisatieproces uit moet zien. Het element <jurpointer> verbeeldt hierin de gecanonicaliseerde citatie.

De attributen bij dit jurpointer-element zijn:

- **CUID**: de gecanonicaliseerde identifier, indien deze tenminste gevonden is;
- **citatieMethode**: Indien aanwezig heeft dit attribuut een numerieke waarde, die de volgende betekenis heeft:
  1. LJN, eventueel met (onderdelen van) een triplet;
  2. één of meer vindplaatsen, eventueel met (onderdelen van) een triplet;
  3. uitsluitend een triplet;
  4. LJN en één of meer vindplaatsen, eventueel met (onderdelen van) een triplet.
- **resolved**: met dit attribuut ('*true*' of '*false*') wordt aangegeven of de citatiestrings binnen <jurpointer> herleid konden worden tot een CUID.
- **aardRelatie**: deze is 'formeel', 'materieel' of 'zelfreferentieel'. Als er geen CUID is, bestaat dit attribuut ook niet.
- **positie**: met dit attribuut wordt vastgelegd op welke karakterpositie in de tekst <jurpointer> begint. Hiermee wordt iedere jurpointer uniek geïdentificeerd.

1929 *Vide supra*: § 2.4.4.
1930 *Vide supra*: § 4.3.3.



Figuur 6-6. XML-schema voor uitspraaktekst met gecanonicaliseerde citaties.

Na voltooiing van het canonicalisatieproces zal onze voorbeeldtekst er bijvoorbeeld zo uit kunnen zien:

<uitspraak>(...) Zoals ook gelezen kan worden in
<jurpointer CUID="1" citatieMethode="4" positie="1227"
aardRelatie="materieel">
<citatiestring stringtype="instantie" waarde1="11">Hoge Raad
</citatiestring>,
<citatiestring stringtype="LJN" waarde1="AO3858">LJN AO3858
</citatiestring>, zaaknr.
<citatiestring stringtype="zaaknummer" waarde1="C02/320HR">
C02/320HR**</citatiestring>** d.d.
<citatiestring stringtype="datum" waarde1="2004-04-09">9 april 2004
</citatiestring>,
<citatiestring stringtype="vindplaats" waarde1="JOL" waarde2="2004,
199">JOL 2004/199**</citatiestring>** en
<citatiestring stringtype="vindplaats" waarde1="NJ" waarde2="2004,
331">NJ 2004/331**</citatiestring>**
</jurpointer>
. Een ander argument voor deze stelling kan worden gehaald uit (...)
</uitspraak>

Met opzet staat hierboven 'bijvoorbeeld', want de XML zou er ook anders uit kunnen zien. In bovenstaande XML verwijzen de twee vindplaatsen, het LJN en het triplet naar één werk, maar indien de twee vindplaatsen gezamenlijk naar een ander werk verwijzen dan het werk waarnaar LJN en triplet verwijzen, dan ziet de XML er zo uit:

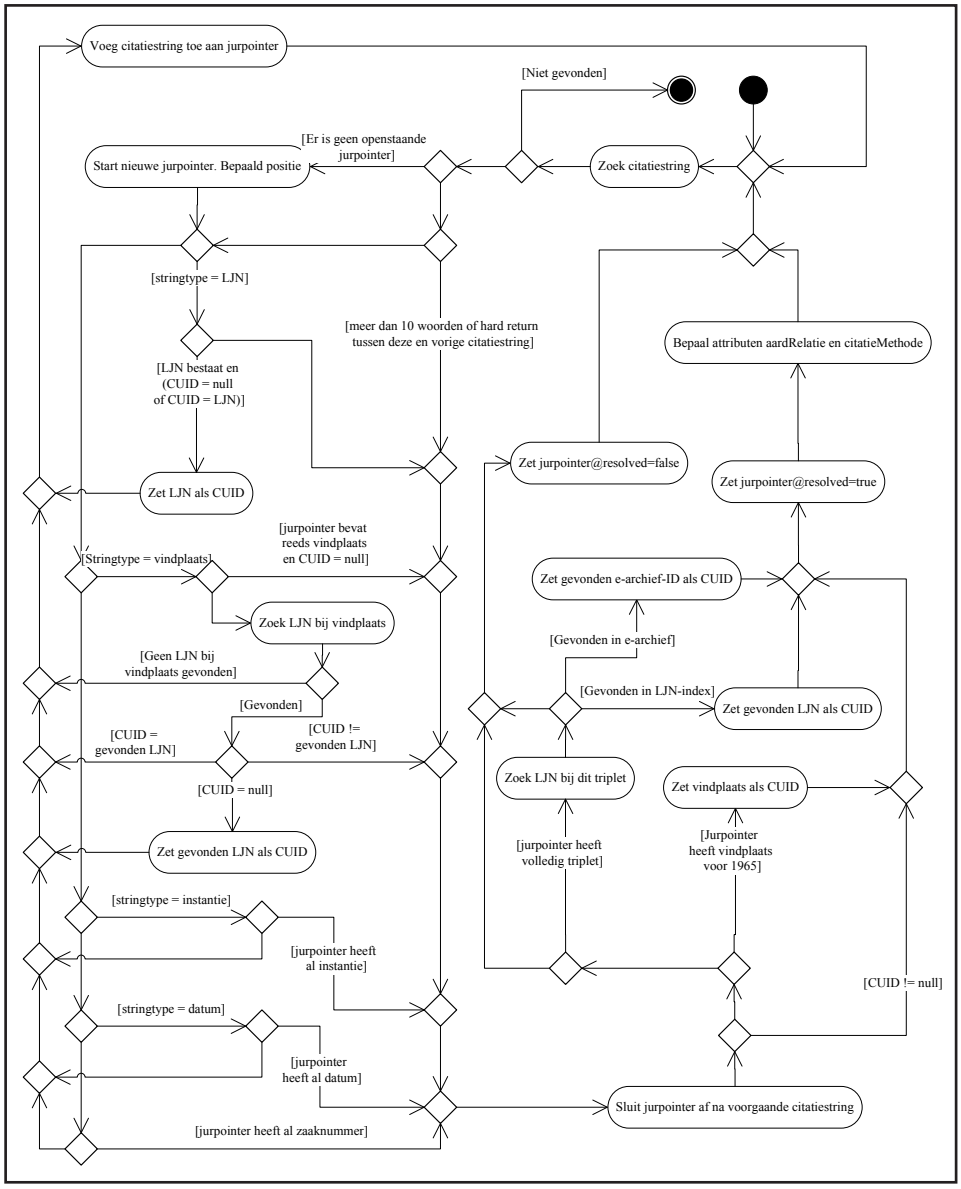
<uitspraak>
(...) Zoals ook gelezen kan worden in
<jurpointer CUID="1" citatieMethode="1" positie="1227"
aardRelatie="materieel">
<citatiestring stringtype="instantie" waarde1="11">Hoge Raad
</citatiestring>,
<citatiestring stringtype="LJN" waarde1="AO3858">LJN AO3858
</citatiestring>, zaaknr.
<citatiestring stringtype="zaaknummer" waarde1="C02/320HR">
C02/320HR**</citatiestring>** d.d.
<citatiestring stringtype="datum" waarde1="2004-04-09">9 april 2004
</citatiestring>,
</jurpointer>
<jurpointer CUID="2" citatieMethode="2" positie="1287"
aardRelatie="materieel">
<citatiestring stringtype="vindplaats" waarde1="JOL" waarde2="2004,
199">JOL 2004/199**</citatiestring>** en
<citatiestring stringtype="vindplaats" waarde1="NJ" waarde2="2004,
331">NJ 2004/331**</citatiestring>**
</jurpointer>
. Een ander argument voor deze stelling kan worden gehaald uit (...)
</uitspraak>

De canonicalisatiefunctie moet onder meer in staat zijn om onderscheid te maken tussen deze twee varianten. In de volgende paragraaf beschrijven we hoe dit in zijn werk gaat.

6.3.5.4 Het canonicalisatieproces

Hoe we van de XML met kale citatiestrings (in § 6.3.5.1) kunnen komen tot de XML met jurpointers (in § 6.3.5.3) is schematisch weergegeven in het UML *activity diagram* van Figuur 6-7.[1931]

1931 Voor een toelichting op het activity diagram: *vide supra*: noot 1894. Bovendien: indien een ruit een keuzepunt is, staan (om het diagram leesbaar te houden) niet alle mogelijkheden genoemd. De ongelabelde pijl is de te volgen processtroom indien niet wordt voldaan aan de voorwaarde(n) van de wel-gelabelde pijl(en).



Figuur 6-7. UML *activity diagram* voor het afleiden van jurpointers uit citatiestrings.

Dit diagram geeft de verwerking van een volledig uitspraakdocument weer. Er wordt gestart met het zoeken naar de eerstvoorkomende citatiestring, en vervolgens wordt op alle knooppunten een keuze gemaakt. Zodra een citatiestring is gevonden, wordt een jurpointer-element aangemaakt, en daaraan worden net zo lang citatiestrings toegevoegd totdat zich

een situatie voordoet die noopt tot het afsluiten van de jurpointer. Zo'n situatie doet zich bijvoorbeeld voor als er tussen citatiestrings meer dan tien woorden staan of bij het begin van een nieuwe alinea. Ook als er een vindplaats wordt gevonden met een ander LJN dan het LJN dat reeds in de openstaande jurpointer aanwezig is, wordt de jurpointer afgesloten.

Na het afsluiten van een jurpointer wordt (in de rechterkant van het schema) de informatie in dit element beoordeeld en eventueel gebruikt om een CUID op te zoeken in de LJN-index of het E-archief. Ten slotte wordt het attribuut '*resolved*' op '*false*' of '*true*' gezet. Indien deze waarde '*true*' is, worden tevens de attributen aardRelatie en citatieMethode gevuld. De citatieMethode wordt bepaald op basis van de in de jurpointer aanwezige citatiestrings. Om de aardRelatie te bepalen wordt eerst onderzocht of de uitspraak zelfreferentieel is. Dit komt regelmatig voor, omdat veel uitspraken beginnen met het vermelden van uitsprekende instantie, zaaknummer en uitspraakdatum. Deze zelfreferenties worden verwijderd.[1932] Is de citatie niet zelfreferentieel, dan wordt gekeken of LJN (indien aanwezig) of triplet voorkomen in de tabel met formele relaties van de CJO. Omdat deze tabel niet volledig is, wordt, indien deze zoekactie geen resultaat oplevert, nog de volgende controle uitgevoerd:

ALS: instantie van de citator = HR, CRvB, gerechtshof, RvS of CBb
EN ALS: jurpointer staat op eerste 20 regels van de tekst
DAN: aardRelatie = formeel
ANDERS: aardRelatie = materieel

Deze regel is gebaseerd op de constatering dat de aangevallen uitspraak vrijwel altijd aan het begin van de uitspraak wordt vermeld, terwijl materiële citaties juist vrijwel nooit in die eerste twintig regels staan.

Mede ten behoeve van de statistische analyses van hoofdstuk 7 wordt de informatie uit de jurpointer-elementen ten slotte weggeschreven naar de databank. Gezien het standaardkarakter van deze activiteit, is deze hier niet beschreven.

6.4 Inkomende citaties uit rechtswetenschappelijke literatuur

Rechterlijke beslissingen worden niet alleen – om formele of materiële redenen – aangehaald in andere uitspraken, maar ook in rechtswetenschappelijke literatuur.

Weliswaar zijn in sommige door uitgevers geleverde bestanden de jurisprudentiecitaties reeds gemarkeerd, maar omdat deze over het algemeen niet zijn gecanonicaliseerd, waren deze voor ons onbruikbaar. Indien de geciteerde uitspraak niet in een tijdschrift van de eigen uitgeverij is gepubliceerd, is een markering bovendien vaak geheel afwezig. Het was daarom het meest efficiënt om eerst alle citaties te verwijderen[1933] en de documenten vervolgens het

1932 Niet alleen omdat het gegeven voor ons niet relevant is, maar ook omdat sommige netwerkalgoritmen niet met zelfreferenties overweg kunnen (vgl. [de Nooy, Mrvrar en Batagelj 2011, p. 8]).
1933 Zoals beschreven in § 6.2.2.

in § 6.3 beschreven proces te laten doorlopen. Een aanpassing was daarbij nodig voor de controle op de aard van de relatie: formele relaties bestaan niet, alle relaties tussen literatuur en uitspraak zijn materieel van aard.[1934] Een bijzonder relatietype komt bovendien voor bij annotaties. Indien een annotatie de uitspraak aanhaalt die de grondslag vormt voor de annotatie, dan werd deze verwijzing verwijderd, omdat dit reeds een eigenstandig gegeven is, dat onder meer in de LJN-index is opgeslagen.

In de lemmata van Wiki Juridica hoefden uitspraakcitaties niet te worden opgespoord; verwijzingen naar uitspraken worden door de wikisoftware reeds gestructureerd opgeslagen. Met een webservice-query konden deze verwijzingen daarom op eenvoudige wijze worden opgehaald.

Uitspraakcitaties in de actualiteiten op Rechtspraak.nl[1935] konden ook met de jurpointersoftware worden gedetecteerd. Eerder[1936] beschreven we reeds hoe onderscheid is gemaakt tussen verwijzingen naar uitspraken die de grondslag vormen van de actualiteit, en uitspraken die om andere redenen in de actualiteit zijn aangehaald.

6.5 Wetsverwijzingen

Net als het herkennen van uitspraakcitaties is ook het detecteren van wetsverwijzingen een ingewikkelde opgave, zij het deels om andere redenen. Een belangrijk verschil vloeit voort uit het feit dat in wetsverwijzingen (bijna) geen numerieke identifiers worden gebruikt.[1937] Een ander verschil zit erin dat naar uitspraken vrijwel altijd in hun geheel wordt verwezen, terwijl van regelingen vaak aan een specifiek onderdeel wordt gerefereerd, zoals een hoofdstuk of een artikel.

Voordat we ingaan op de gevolgde methodiek om wetsverwijzingen computerleesbaar te maken, bespreken we in § 6.5.1 de voor het normalisatieproces benodigde referentiedatabank. Met behulp van deze databank kunnen we de wetsverwijzingen detecteren en normaliseren (§ 6.5.2). Europeesrechtelijke verwijzingen vergen een eigen herkenningsmethodiek, die we beschrijven in § 6.5.2.2. In § 6.5.2.3 bespreken we het afhandelen van lokale aliassen en in § 6.5.2.4 het detecteren van verwijzingen op onderdeelsniveau.

6.5.1 Opbouw referentiedatabank

Eerder[1938] zagen we reeds dat er tal van manieren zijn voor het citeren van regelgeving: opschriften, citeertitels, aliassen, afkortingen en (vooral bij Europese regelgeving) numerieke

1934 Er kan natuurlijk wel een formele relatie bestaan tussen twee of meer uitspraken die in één rechtswetenschappelijke bron worden geciteerd. Hiermee is geen rekening gehouden.
1935 In § 7.3.3 zullen we deze beschouwen als een soort annotatie, en daarmee ook als literatuur.
1936 *Vide supra*: § 6.2.4.2.
1937 *Vide supra*: § 4.7.
1938 *Vide supra*: § 4.7.

identifiers. Teneinde zoveel mogelijk wetsverwijzingen in rechterlijke uitspraken te herkennen, diende daarom de grootst mogelijke verzameling met identifiers te worden opgezet. Voor deze 'Onderzoekswettenverzameling' (OWV) hebben we de volgende bronnen gebruikt:

1. **Basiswettenbestand**
   De benodigde metagegevens uit het BWB zijn via een publieke webservice te verkrijgen. Hiermee werden opschriften, citeertitels, aliassen, afkortingen en BWB-ID's verzameld.
2. **Centrale Voorziening Decentrale Regelgeving**
   Ook voor deze databank[1939] is een webservice beschikbaar. Hiermee zijn citeertitels en numerieke identifiers opgehaald.
3. **EUR-Lex**
   Alle bestanden van EUR-Lex worden door het EU-publicatiebureau in XML geleverd. Deze werden gebruikt voor het inlezen van de Europese regelgeving: sector 1 met de oprichtings- en toetredingsverdragen, sector 2 met de associatieverdragen en door de EU gesloten verdragen, sector 3 met secundaire regelgeving en sector 4 met aanvullende regelgeving.[1940] Van alle bestanden werden CELEX-nummer en titel opgenomen.
4. **Bistro wettentabel (BWT)**[1941]
   Eerder bespraken we de problemen die opgelost moesten worden bij het integraal doorzoekbaar maken van jurisprudentiedatabanken van meerdere leveranciers.[1942] Naast de verschillende identificatiesystemen voor uitspraken moest daarbij ook rekening worden gehouden met verschillen in metadatering, in het bijzonder de 'wetsingang': een redactionele aanduiding van de belangrijkste wetsartikelen die in de uitspraak aan de orde komen. In een geïntegreerde omgeving wil de jurist met één zoekactie alle uitspraken over een bepaald artikel kunnen vinden, ongeacht de schrijfwijze die voor de wetsingang is gehanteerd. Sommige dataleveranciers hanteren een eigen referentielijst met afkortingen en citeertitels, terwijl andere leveranciers de schrijfwijze overlaten aan individuele redacteuren. Tien schrijfwijzen voor één regeling zijn dientengevolge niet ongewoon. Omdat ten tijde van de bouw van Porta Iuris het BWB nog niet bestond, werd door Bistro een eigen referentietabel met wetgevingstitels opgebouwd. Het koppelen van de door leveranciers gehanteerde wetsingangen aan de BWT geschiedde met behulp van Codex.[1943] Naast tal van globale aliassen die in BWB en EUR-Lex ontbreken, bevat de BWT veel regelingen die ouder zijn dan het BWB, alsmede verdragen die niet in BWB of EUR-Lex zijn te vinden.

1939 *Vide supra*: § 4.7.2.
1940 Voor de sectorindeling van EUR-Lex, *vide supra*: noot 1498.
1941 Zie ook [van Opijnen 2010b, p. 31].
1942 *Vide supra*: § 3.2.1.4.
1943 *Vide supra*: § 3.2.1.4.

Voor de integratie van deze vier bronnen in de OWV zijn de volgende stappen gevolgd:

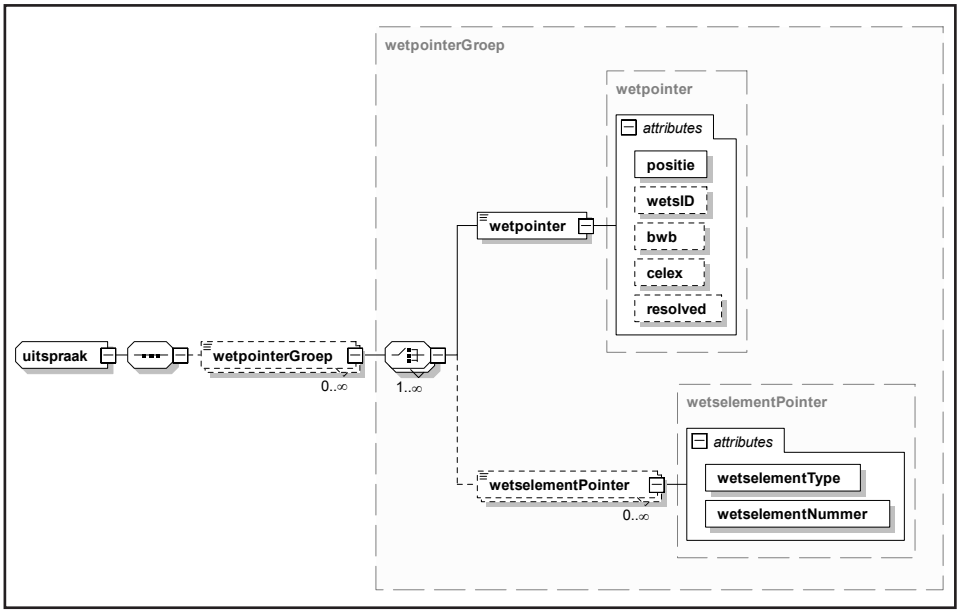
1. Het volledige BWB vormde de basis. Dubbele citeertitels[1944] werden verwijderd.[1945] Ingelezen werden BWB-ID, opschriften, citeertitels, aliassen en afkortingen.
2. Titels en CELEX-nummers van de EUR-Lex-regelingen werden ingelezen. Daarbij werden de verdragen die ook in het BWB voorkomen,[1946] genegeerd.
3. Alle regelingen van de BWT werden vergeleken met de in de eerste twee stappen gevulde OWV. Alle regelingen met een al in OWV aanwezige (citeer)titel werden genegeerd.
4. Alle aliassen en afkortingen uit de BWT die nog niet bekend waren in de OWV, werden daarnaartoe overgezet. Daarbij werd met de hand gecontroleerd op bruikbaarheid; veel waarden in de BWT zijn noodzakelijk voor de productiestraat van Porta Iuris, maar hebben geen relevantie voor de OWV.[1947]
5. Voor zover nog niet in de OWV aanwezig werden alle regelingen van de CVDR ingelezen, met identifier en titel.

De OWV bevat 158.160 regelingen, waarvan er – gelet op de beschreven inleesvolgorde – 114.453 afkomstig zijn uit EUR-Lex, 25.718 uit BWB, 9.684 uit CVDR en 8.305 uit BWT. Naast de (citeer)titels die in de primaire tabel van OWV zijn opgenomen, bevat de OWV aanvullend nog 25.997 alternatieve titels, aliassen en afkortingen, zodat er in totaal 184.157 strings zijn om wetsverwijzingen in uitspraken mee op te sporen. Naast BWB-ID, CELEX-nummer of CVDR-identifier heeft elke regeling in de OWV ook een eigen unieke sleutel.

6.5.2 Het detecteren en normaliseren van wetsverwijzingen

In de uitspraakbestanden in de onderzoeksdatabank zijn wetsverwijzingen niet gemarkeerd.[1948] De te ontwikkelen software diende uiteindelijk XML op te leveren zoals weergegeven in Figuur 6-8.

1944 *Vide supra*: § 4.7.1.
1945 Daardoor ontstaat een (kleine) kans dat verwijzingen aan de verkeerde regeling worden gekoppeld. Alhoewel vervelend, is dat voor ons onderzoek niet cruciaal.
1946 *Vide supra*: noot 1602.
1947 Bijvoorbeeld als er technische codes of andersoortige informatie in het te vergelijken veld waren achtergebleven. Voorbeelden van dergelijke strings: '(BW , , , , , ,)' en 'Fw e.v.'.
1948 In sommige gevallen was er wel markering, maar deze is eerst verwijderd (*vide supra*: § 6.2.2) teneinde alle bestanden volgens één systematiek te kunnen verwerken.



Figuur 6-8. XML-schema voor wetsverwijzingen in uitspraken.

In de volgende paragrafen zullen we stap voor stap bespreken hoe de ontwikkelde software de uitspraken naar dit schema modelleert.[1949] Allereerst dienen wetsaanhalingen als zodanig te worden herkend. Daarbij maken we onderscheid tussen verwijzingen die kunnen worden herkend met de gegevens in de OWV (§ 6.5.2.1) en numerieke verwijzingen naar Europese regelgeving (§ 6.5.2.2). Vervolgens worden lokale aliassen gedetecteerd (§ 6.5.2.3) en de wetselementen verwerkt (§ 6.5.2.4).

6.5.2.1 Het vinden van regelingen met de onderzoekswettenverzameling

Om in de uitspraken van de onderzoeksdatabank te zoeken naar de in de OWV opgenomen termen is gebruik gemaakt van Verity K2 (hierna: 'K2').[1950] Met K2 werd eerst een zoekindex opgebouwd van alle uitspraken. Daarnaast werd een '*topicset*' – een lijst met de te vinden

1949 Ook elders zijn parsers ontwikkeld voor het detecteren van wets(element)verwijzingen (zie bijvoorbeeld [de Maat, Winkels en van Engers 2006]). Veel van dergelijke parsers verschillen echter in twee belangrijke opzichten van de door ons ontwikkelde software. Ten eerste zijn veel van deze parsers ontwikkeld voor het detecteren van verwijzingen in regelgeving, waar de formattering van de verwijzingen – dankzij de Aanwijzingen voor de regelgeving – veel strikter is (*vide supra*: § 6.3.2). Ten tweede zijn dergelijke parsers vaak ontwikkeld voor experimentele omgevingen met kleine datasets. Het zoeken naar circa 200.000 termen in bijna een miljoen bestanden vergde een andere technische aanpak. Overigens kostte ook de ontwikkelde oplossing de nodige computerkracht: drie stevige werkstations hebben vier weken onafgebroken staan stampen om de uitspraken in de onderzoeksdatabank te verwerken.

1950 Commercieel product, bij de Rechtspraak gebruikt voor Porta Iuris en Rechtspraak.nl ([Hoeffnagel 2005]).

zoektermen – opgebouwd uit de OWV. Daarbij werd – om te voorkomen dat korte woorden ten onrechte als wetsverwijzing zouden worden geïnterpreteerd – als bijzondere instructie meegegeven dat alle *topics* met tien of minder karakters hoofdlettergevoelig moesten worden gezocht. Door deze *topicset* aan te houden tegen de zoekindex kwamen alle uitspraken te voorschijn waarin ten minste één regeling uit de OWV wordt aangehaald. Deze uitspraken werden allemaal door een '*highlight-engine*' gehaald, die de gevonden OWV-strings van een markering voorzag.[1951]

Het volgende stuk uitspraaktekst:

> *(...) bestreden besluit in strijd is met artikel 1 juncto artikel 5 van het Algemeen verdrag inzake sociale zekerheid tussen het Koninkrijk der Nederlanden en het Koninkrijk Marokko (Trb. 1972, 34; hierna: het Verdrag).(...)*

komt dan, in XML, als volgt uit de *highlight-engine*:

> **<uitspraak>**
> (...) bestreden besluit in strijd is met artikel 1 juncto artikel 5 van het
> **<wetpointer positie="1854">**
> Algemeen verdrag inzake sociale zekerheid tussen het Koninkrijk der Nederlanden en het Koninkrijk Marokko
> **</wetpointer>**
> (Trb. 1972, 34; hierna: het Verdrag).(...)
> **</uitspraak>**

We hebben nu een tussenproduct waarin alle strings die in de OWV voorkomen als '<wetpointer>' zijn gemarkeerd, met als extra toevoeging een attribuut '@positie', waarmee iedere wetpointer uniek wordt geïdentificeerd. Maar we weten dan nog niet wélke regeling hier is gemarkeerd. De volgende stap is daarom het toevoegen van de unieke OWV-identifier aan de gehighlighte tekst. Daartoe wordt de highlightstring aangehouden tegen de OWV, die vervolgens de gevraagde identifier teruggeeft. Om de herbruikbaarheid van dit tussenproduct te vergroten, voegen we daar nog de externe identifiers aan toe, zoals CELEX-nummer en/of BWB-ID. Bovendien – net als bij de jurisprudentieverwijzingen[1952] – wordt het status-attribuut '*resolved*' opgenomen Ons voorbeeld ziet er nu zo uit:

> **<uitspraak>**
> (...) bestreden besluit in strijd is met artikel 1 juncto artikel 5 van het
> **<wetpointer positie="1854" wetsID="4780" bwb="BWBV0001011"**
> **resolved="true">**

1951 Dit proces is hier vereenvoudigd beschreven. In werkelijkheid werd elk woord afzonderlijk gehighlight en moesten deze door de parser weer worden geconcateneerd. Gezien het algemene IT-karakter van dit proces, en de gebondenheid van deze oplossing aan de specifieke IT-gereedschappen die zijn gebruikt, is het hier verder niet beschreven. Dat geldt ook voor andere (soms zeer tijdrovende) problemen die getackeld moesten worden, zoals het harmoniseren van de *encodings* waarin de gegevens uit de verschillende bronsystemen bleken te staan.

1952 *Vide supra*: § 6.3.5.3.

Algemeen verdrag inzake sociale zekerheid tussen het Koninkrijk der Nederlanden en het Koninkrijk Marokko
</wetpointer>
(Trb. 1972, 34; hierna: het Verdrag).(...)
</uitspraak>

Voordat deze tekst verder verwerkt kan worden, moeten eerst de Europese regelingen in de uitspraak worden gedetecteerd.

6.5.2.2 Detecteren van Europese regelgeving

In § 4.7.3 hebben we geconstateerd dat Europese regelgeving zelden correct wordt geciteerd. Bij gebruik van louter de OWV zouden we veel verwijzingen derhalve over het hoofd zien. Ook het gebruik van de officiële documentnummers zou niet de gewenste resultaten opleveren, omdat deze op talloze – over het algemeen foutieve – manieren worden gespeld.[1953] Met reguliere expressies[1954] kunnen de verschillende spellingsvarianten evenwel worden herkend. De reguliere expressie voor het opsporen van bijvoorbeeld verordeningen ziet er zo uit (de onderstreping van sommige strings maakt geen onderdeel uit van de reguliere expressies; de betekenis ervan wordt later verklaard):[1955]

(V|v)(er)?(o|O)(rd)?(ening(en)?|\.)?\s+((?i)(n(umme)?(r|o)\.?)\s*)?\d{1,4}(/|\.)(19|20)?\d{2}(/|-|\s+)\(?(E(E)?G(\sen\s|,\s|/)Euratom|E(E)?G|Euratom)\)?|(E((E)?G|uropese)(-|\s+))?(V|v)(er)?(o|O)(rd)?(ening(en)?|\.)?\s+((?i)(n(umme)?(r|o)\.?)\s*)?\d{1,4}(/|\.)(19|20)?\d{2}|(V|v)(er)?(o|O)(rd)?(ening(en)?|\.)?(\s+|-)\(?(E(E)?G(\sen\s|,\s|/)Euratom|E(E)?G|Euratom)\)?,?(\s+|/|-)?((?i)(n(umme)?(r|o)\.?)\s*)?\(?\d{1,4}(/|\.)(19|20)?\d{2}\)?

Voor richtlijnen, beschikkingen, aanbevelingen, besluiten, JBZ-kaderbesluiten en gemeenschappelijke acties zijn vergelijkbare reguliere expressies geschreven. Alle strings die hiermee werden gevonden, moesten vervolgens worden genormaliseerd naar een CELEX-nummer, dat voor secundaire regelgeving uit de volgende elementen bestaat: [sectorcode][jaartal][descriptor][volgnummer].

De normalisatie gaat als volgt:

- De sectorcode is altijd '3';
- Het jaartal wordt gevormd door de cijfers die zijn herkend met het enkelvoudig onderstreepte gedeelte van de reguliere expressie. Het wordt aangevuld naar vier karakters;

1953 *Vide supra*: § 4.7.3.
1954 *Vide supra*: noot 1920.
1955 De reguliere expressies zijn gemaakt vóór de inwerkingtreding van het Verdrag van Lissabon; regelingen die 'EU' in de naam dragen (in plaats van 'E(E)G') worden er dus niet door gedetecteerd. Aanpassing van de reguliere expressies aan het Verdrag van Lissabon is echter eenvoudig.

- De descriptor is een 'D' indien de string is gevonden met de reguliere expressie voor beschikkingen, een 'H' indien het een aanbeveling betreft, een 'D' voor besluiten, een 'R' voor Verordeningen, een 'L' voor richtlijnen en een 'F' voor kaderbesluiten en gemeenschappelijke acties.
- Het volgnummer wordt gevormd door de cijfers die zijn herkend met het dubbelonderstreepte gedeelte. Het wordt met voorloopnullen aangevuld tot vier cijfers.

In Figuur 6-9 zijn enkele voorbeelden opgenomen van in uitspraken gebruikte citaties van Europese regelgeving, tezamen met het CELEX-nummer waarnaar ze zijn herleid.

De gevonden strings worden vervolgens – net als de strings die door de *highlight-engine* zijn gemarkeerd[1956] – omsloten door wetpointer-elementen. Daarna wordt – in de OWV – gecontroleerd of het gevonden CELEX-nummer wel bestaat. Als deze verificatie slaagt, worden CELEX-nummer en OWV-identifier als attributen bij de wetpointer opgenomen. Als in de uitspraaktekst dus staat:

(…) zoals kan worden gelezen in EG Verordening nr. 2201/2003 (…)

dan wordt de XML:

<uitspraak>
(…) zoals kan worden gelezen in
<wetpointer positie="6359" wetsID="12730" celex="32003R2201"
resolved="true">
EG Verordening nr. 2201/2003
</wetpointer>(…)
</uitspraak>

Als het CELEX-nummer niet voorkomt in de OWV, dan wordt het attribuut 'resolved' op 'false' gezet.[1957]

6.5.2.3 Verwerken lokale aliassen

Globale aliassen voor regelgeving, zoals de Koppelingswet[1958] en de Bolkestein-richtlijn[1959] zijn in de OWV opgenomen en worden derhalve herkend. Maar meer nog dan bij jurisprudentieverwijzingen[1960] wordt bij wetsverwijzingen gebruikgemaakt van lokale aliassen. We kwamen zo'n alias al tegen in het tekstvoorbeeld van § 6.5.2.1:

1956 *Vide supra*: § 6.5.2.1.
1957 Dergelijke (zeer sporadische) fouten zouden nog handmatig kunnen worden nabewerkt; hiervan is evenwel afgezien.
1958 *Vide supra*: § 4.7.1.
1959 *Vide supra*: § 4.7.3.
1960 *Vide supra*: § 6.3.2.

> *(...) bestreden besluit in strijd is met artikel 1 juncto artikel 5 van het Algemeen verdrag inzake sociale zekerheid tussen het Koninkrijk der Nederlanden en het Koninkrijk Marokko (Trb. 1972, 34; hierna: het Verdrag).(...)*

'het Verdrag' is hierin de lokale alias. Even verderop in de tekst wordt daar dan bijvoorbeeld als volgt aan gerefereerd:

> *In tegenstelling tot het bepaalde in artikel 3 van het Verdrag is in het besluit (…)*

Om dit probleem op te lossen is code geschreven die de declaratie van de lokale alias herkend, en daaropvolgend gebruik ervan van een wetpointer-element voorziet. Voor de laatste voorbeeldzin resulteert dit in:

> **<uitspraak>**
> In tegenstelling tot het bepaalde in artikel 3 van
> **<wetpointer positie="3990" wetsID="4780" bwb="BWBV0001011" resolved="true">**
> het Verdrag
> **</wetpointer>** is in het besluit (...)
> **</uitspraak>**

Daarbij moeten we er wel voor waken dat een globale alias niet ten onrechte als lokale alias wordt getagd. Als voorbeeld diene de volgende zin:

> *Zoals de systematiek van de Algemene wet bestuursrecht (hierna: Awb) duidelijk maakt (...)*

Citatie	CELEX-nummer
EG-verordening 1103/97	31997R1103
EG Verordening nr. 2201/2003	32003R2201
EG-vo. 1408/71	31971R1408
Richtlijn 91/338/EEG	31991L0338
beschikking 95/340/EG	31995D0340
Aanbeveling 2003/311/EG	32003H0311
2002/584/JBZ	32002F0584

Figuur 6-9. Voorbeelden van verwijzingen naar Europese regelgeving en de CELEX-nummers waarnaar deze verwijzingen zijn genormaliseerd.

Omdat 'Awb' reeds voorkomt in de OWV zal deze zin er na parsing als volgt uitzien:

<uitspraak>
Zoals de systematiek van de
<wetpointer positie="3387" wetsID="84" bwb="BWBR0005537"
resolved="true">
Algemene wet bestuursrecht
</wetpointer>
(hierna:
<wetpointer positie="3423" wetsID="84" bwb="BWBR0005537"
resolved="true">
Awb
</wetpointer>
) duidelijk maakt (...).
</uitspraak>

Hierdoor lijkt de Algemene wet bestuursrecht twee keer te worden aangehaald, terwijl dat feitelijk slechts eenmaal geschiedt. Daarom is toegevoegd dat de lokale-aliascode moet controleren of de gevonden lokale alias niet tevens een globale alias is, in welk geval de tweede wetpointer niet wordt aangemaakt. Overal waar verderop in de tekst 'Awb' voorkomt, zal de wetpointer goed worden geplaatst, omdat deze afkorting immers door de *highlight-engine* correct is gemarkeerd.

6.5.2.4 Herkennen wetselementen

We beschikken nu over een uitspraaktekst waarin regelingen zijn herkend. In de volgende stap gaan we op zoek naar de wetselementcitaties. Net als bij het herkennen van wetcitaties is het herkennen van wetselementcitaties gebaseerd op een *best effort*, want de creativiteit van wetgevers om benamingen te verzinnen voor wetselementen,[1961] en die van rechters om elementen in allerlei combinaties en in allerlei stijlen aan te halen, is schier eindeloos. We hebben ervoor gekozen om het artikel als laagste granulariteitsniveau te herkennen. Enerzijds omdat we dit voor ons onderzoek voldoende achten, anderzijds omdat het artikel over de tijd vrij stabiel is. De tekst van een artikel kan weliswaar wijzigen, maar vernummeringen van artikelen en hogere elementen zijn – op grond van de instructie in Aanbeveling 238 eerste lid Aanw. reg. – zeldzaam; vernummering van leden en lagere elementen binnen een artikel is daarentegen gangbare praktijk.

Ook het zoeken naar wetselementcitaties geschiedt met behulp van reguliere expressies. In dit proces zijn er verschillende objecten die – in onderling verband – moeten worden herkend. Bij de hiernavolgende beschrijving van deze objecten kan de volgende (samengestelde) verwijzing ter illustratie dienen:

1961 De Aanwijzingen voor de regelgeving zijn wat dat betreft geen lichtend voorbeeld. Aanwijzing 95 schrijft voor dat een regeling wordt ingedeeld in artikelen – en dus niet in 'aanwijzingen'.

artikelen 11 t/m 13 en artikel 14 eerste lid van de Grondwet

- **Wetselementen**
  Dit zijn de elementen die we herkenbaar willen maken. Ieder wetselement bestaat uit één elementtype (T) en één of meer elementnummers (N). Naast 'artikel' herkennen we als elementtypes: boek, deel, titel(deel), hoofdstuk, afdeling, paragraaf, subparagraaf en bijlage.

  Elementnummers kunnen afzonderlijk opgesomd staan, maar ook als reeks zijn benoemd. In de voorbeeldzin zijn de wetselementen: 'artikelen 11 t/m 13' en 'artikel 14'.
- **Artikelsubelementen (A)**
  Onderdelen van een wetsartikel – subelementen van andere elementtypes worden vrijwel nooit benoemd – die we niet hoeven op te slaan, maar die wel moeten worden herkend omdat ze wetselementen kunnen verbinden.

  In het voorbeeld is 'eerste lid' een artikelsubelement. Omdat de artikelsubelementen niet worden opgeslagen, wordt geen onderscheid gemaakt tussen types en nummers.
- **Koppeltermen (K)**
  Woorden en leestekens die tussen wetselementen en artikelsubelementen kunnen staan. In de voorbeeldzin zijn dit 't/m' en 'en'.
- **Verbindingen (V)**
  Woorden en leestekens die een verzameling van wetselementen en artikelsubelementen verbinden aan de geciteerde regeling. In de voorbeeldzin is de string 'van de' een verbinding.

Als wetselementen vóór de regeling staan, dan kunnen deze voorkomen volgens de reguliere expressie:

```
TN(K(T?N)|(K?A))*V?
```

Als de wetselementcitaties achter de wetcitatie staan dan is de reguliere expressie:

```
V?TN(K(T?N)|(K?A))*
```

Voor de herkenning van de verschillende objecten zijn afzonderlijke reguliere expressies geschreven. Voor de elementnummers zijn bijvoorbeeld alle gangbare nummeringsmethoden tot één reguliere expressie verwerkt:

```
(([1-9]\.)?([1-9][0-9]?)?[A-Z]?(:\s?|\.))?[1-9]\d{0,3}[a-z]{0,9}|
[ABCD]\s?\d{1,2}/([1-9]([0-9])?\.?){1,4}|
H[1-9][0-9]?,[1-9][0-9]?|
[IVXCLDM]{1,10}(-[A-Z])?(,?[1-9]\d{0,3}[a-z]{0,3})?|
[A-Z][1-9][0-9]?[a-z]?|
[1-9]\.[1-9][0-9]?:[1-9][0-9]?|
[1-9]\d{0,2}[a-z]?\.[1-9]
```

Vervolgens worden de elementtypes genormaliseerd (‘art.’ wordt bijvoorbeeld genormaliseerd naar ‘artikel’), en reeksaanduidingen bij elementnummers worden omgezet naar een volledige lijst (‘11 t/m 13’ wordt ‘11’, ‘12’, ‘13’).[1962] De gevonden wetselementen worden nu als XML in de uitspraaktekst opgenomen, volgens het XML-schema van Figuur 6-8. Het element ‘wetpointer’, waarmee de wetcitatie wordt gemarkeerd, wordt daartoe uitgebreid tot een ‘wetpointergroep’, waarin zich naast de wetpointer zelf ook ‘wetpointerelementen’ kunnen bevinden. Een wetpointergroep bevat altijd precies één wetpointer, en geen, één of meerdere wetselementpointers.

Onze voorbeeldzin komt er in XML dan zo uit te zien:

```
<uitspraak>
<wetpointerGroep>
<wetselementPointer wetselementType="artikel"
wetselementNummer="11">
artikelen 11
</wetselementPointer>
<wetselementPointer wetselementType="artikel"
wetselementNummer="12"/>
t/m
<wetselementPointer wetselementType="artikel"
wetselementNummer="13">
13
</wetselementPointer>
en
<wetselementPointer wetselementType="artikel"
wetselementNummer="14">
artikel 14
</wetselementPointer>
eerste lid van de
<wetpointer positie="256" wetsID="1436" bwb="BWBR0001840"
resolved="true">
Grondwet
</wetpointer>
</wetpointerGroep>
</uitspraak>
```

Als laatste stap worden deze in XML gestructureerde gegevens opgeslagen in de relationele databank. Dit is een zuiver technische exercitie die geen verdere beschrijving behoeft.

1962 Hiertoe wordt een reeks van natuurlijke getallen gebruikt, die niet wordt gecontroleerd in de wet in kwestie. Indien in het voorbeeld artikel 12 niet (meer) zou bestaan, dan wordt deze toch opgenomen. Zou er een artikel 11a bestaan, dan wordt deze niet opgenomen.

6.6 Uitspraken in context: conclusies en vooruitblik

In dit hoofdstuk hebben we beschreven hoe jurisprudentiecitaties en wetsverwijzingen uit ongestructureerde teksten kunnen worden geëxtraheerd. Met deze *linked data* kan invulling worden gegeven aan de deelaspecten 'contextualiteit' en 'doorzoekbaarheid' van het toegankelijkheidsaspect 'hanteerbaarheid'.

We hebben in dit hoofdstuk gebruikgemaakt van een bepaalde technologie, maar deze is louter instrumenteel: hetzelfde doel kan ook worden bereikt met andere technologieën. Ook het omzetten van de door ons gegenereerde links naar andere formaten, zoals RDF,[1963] mag geen problemen opleveren.

Eindgebruikerstoepassingen zijn door ons niet ontwikkeld, wel zijn twee eenvoudige applicaties gemaakt waarmee wordt gedemonstreerd:

- hoe een zoekinterface eruit kan zien die alle uitspraken retourneert waarin een door de gebruiker opgegeven wet(selement) voorkomt, ongeacht de schrijfwijze die de rechter in de tekst van zijn uitspraak heeft gehanteerd;
- hoe de citatie-index kan worden gebruikt om te visualiseren waar in jurisprudentie en rechtswetenschappelijke literatuur een uitspraak wordt aangehaald;
- hoe iedere verwijzing naar regelgeving of jurisprudentie met een door de computer aangebrachte hyperlink direct aanklikbaar kan worden gemaakt.

Enkele schermafbeeldingen van deze applicaties zijn opgenomen in Bijlage 13. Na het werk dat in deze paragraaf is beschreven, vergt het ontwikkelen van dergelijke toepassingen weinig aanvullende inspanningen, al vraagt het ontwerpen van goede gebruikersinterfaces natuurlijk de nodige aandacht.

Los van het nut dat eigenstandig gebruik van deze gegevens in toepassingen voor rechtspraktijk en -wetenschap voor kan hebben, zijn deze voor ons vooral een onontbeerlijke grondstof voor het in het volgende hoofdstuk te ontwikkelen model voor het bepalen van de domeinrelevantie van rechterlijke uitspraken.

1963 *Resource Description Framework*, een semantisch-webstandaard waarin verbanden tussen informatieobjecten worden beschreven in 'subject-predicaat-object'-*triples*.