

Zoeken en interpreteren van jaartallen & plaatsnamen in Nederlandstalige retrieval systemen

Het is vast al eens door iemand gezegd, en zo niet, dan doen we het hier: hoeveel materiaal uit het verleden er ook nog onontdekt in de grond of achter oude gevels mag liggen, het materiaal in de archeologische en historische bibliotheken, door generaties archeologen en andere wetenschappers vlijtig verwerkt tot opgeslagen informatie en kennis, mag evenmin worden veronachtzaamd. En wat in de grond zit vermindert; zelfs de meest waardevolle panden kunnen door brand verloren gaan of worden gesloopt; de informatie in de boeken, artikelen, rapporten groeit daarentegen dagelijks aan en wordt zelfs belangrijker naarmate het originele materiaal onbereikbaar wordt.

SAMENVATTING

De betekenis van een woord of term in onze taal is afhankelijk van de context. Deze bepaalt immers of een getal een jaartal is of een lengtemaat, of dat het woord Utrecht op de stad of op de provincie slaat. De computer is tegenwoordig in staat om die context te interpreteren, maar moet wel 'getraind' worden op teksten uit hetzelfde domein. Dit artikel beschrijft hoe we de computer hebben geleerd om jaartallen en plaatsnamen te herkennen in de tekst van Nederlandstalige archeologische rapporten en op grond daarvan een indexeersysteem hebben gebouwd.



In zekere zin is de informatie in al die gepubliceerde en niet gepubliceerde, 'grijze' rapporten dikwijls al even onontdekt als al het materiaal dat nog in de grond zit. Natuurlijk zijn al die bijdragen in bredere of nauwere kring gelezen en bekend, maar zelfs de meest belezen archeoloog of historicus heeft maar een fractie van al die informatie paraat. De digitale revolutie zorgt er weliswaar voor dat veel van die teksten on-line beschikbaar komen en dank zij moderne 'full text' zoeksystemen kan op ieder voorkomend woord of combinatie van woorden worden gezocht. In veel situaties blijkt het trefwoord alléén echter ontoereikend om de gewenste informatie te vinden.

Het probleem

Als voorbeeld nemen we een schijnbaar triviale zoekopdracht als 'middeleeuwen'. Als we die zoekopdracht in Google opgeven, krijgen we feilloos alle documenten waarin het woord 'middeleeuwen' letterlijk zo voorkomt. En al even feilloos gaat Google voorbij aan elk

jaartal tussen 500 en 1500 en aan uitdrukkingen als de 'elfde eeuw', 'de XI-de eeuw' of '+XI'. Als we alleen geïnteresseerd zouden zijn in een betrekkelijk korte tijd als de jaren tussen 1100 en 1110 zouden we die tien getallen nog wel allemaal in de zoekbalk kunnen intypen, maar dan stuiten we op het volgende probleem: Google weet niet wanneer zo'n getal een jaartal is of een oppervlaktemaat, een prijs in Euro's, een serienummer of het aantal kilometers tussen Amsterdam en Utrecht. Kortom: het automatisch herkennen van tijdsaanduidingen (chronologische expressies) binnen een tekst blijkt allerm minst een triviale aangelegenheid en zelfs een geavanceerde zoekmachine als Google kan er niet mee overweg.

Over Amsterdam en Utrecht gesproken: plaatsnamen dragen een soortgelijke problematiek met zich mee. Het is niet te voorspellen of een plaatsnaam in een tekst ook inderdaad betekent dat het woord een fysieke locatie aanduidt. Het kan evengoed de naam van een persoon zijn, of van een provincie, of de plaats waar een boek is uitgegeven.

In ons project hebben we onder auspiciën van de NRC (Nationale Referentiecollectie¹) en KICH (KennisInfrastructuur CultuurHistorie²) zowel voor de chronologie als voor de locaties bepaalde oplossingen uitgewerkt. Hieronder beschrijven we in detail de ervaringen en oplossingen voor de chronologische expressies; het herkennen van plaatsnamen als echte locaties gaat bijna identiek in z'n werk.

Het project

In 2006 zijn overal in Nederland zogenaamde CATCH projecten gestart. CATCH (Continuous Access To Cultural Heritage) is een NWO initiatief en elk project is een samenwerkingsverband tussen een erfgoed-

instelling en een universiteit. Zo werkt de RACM samen met de Universiteit van Maastricht in het project RICH (Reading Images in Cultural Heritage). Hierin houden we ons bezig met met beeldherkenning en het automatisch classificeren van bijvoorbeeld keramiek.

Helaas blijkt al na enkele maanden dat het voorhanden beeldmateriaal niet geschikt is. Digitaal opgeslagen afbeeldingen zijn er genoeg, maar die blijken bijna allemaal onderdeel van gescande paginas uit, inderdaad, boeken en artikelen. Zulke plaatjes kunnen niet eenvoudig uit hun context worden gehaald, tenminste niet met behoud van de relevante gegevens. Die bijkomende gegevens zijn in de experimenten van groot belang, onder andere om te controleren of de automatische beeldherkenner zijn werk goed heeft gedaan. Daarom richten we ons ook op het probleem van het interpreteren van data die zijn verwoord in de natuurlijke taal in ongestructureerde archeologische teksten. Met onze aanpak willen we genoeg informatie uit de tekst te halen om de afbeeldingen te voorzien van de gegevens uit de oorspronkelijke context.

De interpretatie van natuurlijke taal, zoals het Nederlands, door middel van computers ligt op het gebied van de wetenschap die zich 'Kunstmatige Intelligentie' noemt. En zoals

meestal het geval is binnen wetenschappelijke disciplines zijn ook hier verschillende opvattingen over hoe dat moet worden aangepakt. Binnen CATCH in het algemeen bestaat er een sterke voorkeur voor ontologieën en typologieën, voor thesauri en semantische netwerken. Dit zijn allemaal structuren die berusten op bijna Platonisch geordende werelden van objecten en de relaties daartussen, en regels die daar dan weer van worden afgeleid. Het probleem van deze benadering is dat het heel moeilijk is om bestaande typologieën en thesauri met elkaar in overeenstemming te brengen en in de tussentijd blijven de documenten onontsloten. Voor het onderscheiden van jaartallen en andere getallen binnen lopende Nederlandse tekst hebben we gekozen voor een benadering die het probleem van de ontologieën en typologieën grotendeels omzeilt: de toepassing van het zogenaamde Memory Based Learning (MBL).

De 11-de eeuw, +XI of toch maar 1000-1100?

Laten we even terugkeren naar het probleem van de tijd. Tijdsaanduidingen, chronologische expressies dus, zijn er in twee soorten. De eerste soort bestaat uit de jaartallen zelf en uit de vele varianten waarop ze geschreven worden: 1100, elfhonderd, de elfde eeuw, de 11-de eeuw, +XI, noem maar op. Het is op zich niet zo moeilijk een zogenaamde parser

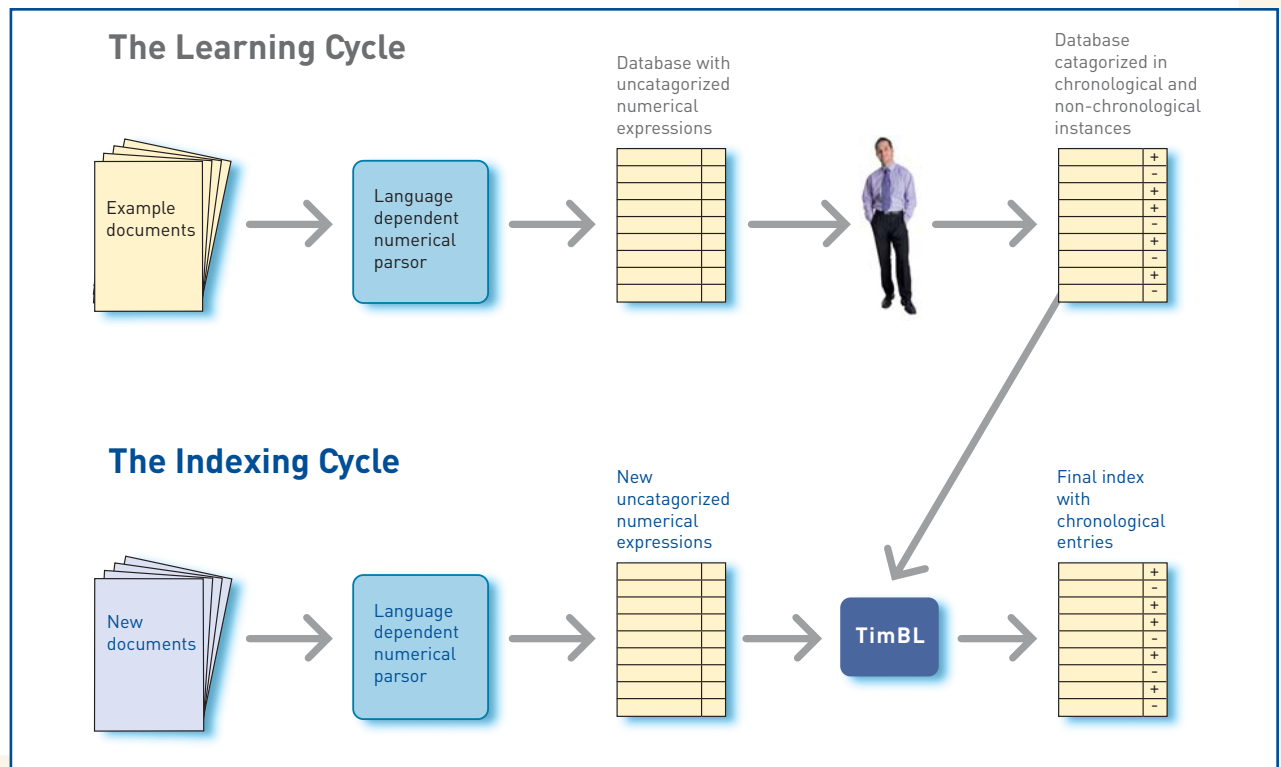
(woordherkenner) te schrijven die voluit geschreven hoofd- en rangtelwoorden herkent en naar getallen omzet, maar het is al een stuk moeilijker om ook rekening te houden met varianten als 'elfde', '11-de', '11^e' of zelfs in superschrift: '11^e' of 'elfde'.

De tweede soort bestaat uit uitdrukkingen en namen zoals 'middeleeuwen', 'de Honderdjarige Oorlog', 'de bronstijd'. Hier is het op zich eenvoudig om een lijst te maken die middeleeuwen vertaalt naar 500-1500 en dergelijke, al doen zich problemen voor bij regionale variaties. Denk hierbij aan de begrenzing van bijvoorbeeld de bronstijd, die alleen binnen Nederland al van plaats tot plaats kan verschillen.

Veel moeilijker is het om getallen in een tekst te interpreteren. Zoals we al zeiden: is 1100 een jaartal, een prijs in Euros of een gewicht in grammen? Het is onverwacht moeilijk een serie regels te bedenken die op grond van de context de jaartallen eruit kunnen pikken, en daar komt ons dan de kunstmatige intelligentie te hulp.

Zoals we al zeiden is Memory Based Learning (MBL) een techniek uit de kunstmatige intelligentie, waarbij het opstellen van de regels eigenlijk aan de computer wordt overgelaten. Om bijvoorbeeld een index te maken van alle tijdsaanduidingen in een tekst, wordt het geheugen van de computer eerst gevuld met een aantal voor-

1
DE MBL-CYCLUS



CODE	CLASS	%	F-SCORE
E42	Object identifier	18.7	0.92
E47	Spatial coordinates	1.2	0.95
E52	Time-span	16.0	0.96
E54	Dimension:		
	- depth/height	2.9	0.86
	- length/width	1.7	0.90
	- diameter	0.8	0.85
	- thickness	0.7	0.79
	- surface	1.7	0.62
	- volume	0.2	0.62
	- weight	0.2	0.85
	- other	1.5	0.88
E60	Number	7.6	0.90
	Reference	6.2	0.99
	Other	40.6	0.95

TABEL 1
GETALLEN EN HUN KLASSEN
IN ARCHEOLOGISCHE PUBLICATIES.

beelden: in dit geval allemaal getallen in hun letterlijke context (zie tabel 2). Vervolgens beoordeelt een mens deze voorbeelden en voorziet ze van het juiste label: jaartal of geen jaartal. Na deze 'learning cycle' heeft de computer dus de beschikking over enkele duizenden gelabelde voorbeelden.

Nu kan het eigenlijke indexeren beginnen. Als er een nieuw geval binnenkomt van een getal in zijn context, dan vergelijkt de computer dat nieuwe geval met alle voorbeelden,

pikt het meest gelijkende voorbeeld eruit en geeft dan dat label ook aan dit nieuwe geval.

Wat betekent dit nu allemaal in de praktijk van ons project. Om te beginnen hebben we de tabel met voorbeelden gemaakt en van labels voorzien. Hiervoor kregen we de medewerking van een student van de Universiteit van Tilburg. Ook het MBL programma, TiMBL (Tilburg Memory Based Learner³), was door Tilburgse wetenschappers geschreven. Binnen enkele weken hebben we dan de beschikking over een grote gelabelde database van vijftienduizend voorbeelden van getallen in een archeologische context, die als basis diende om de beste instellingen van TiMBL vast te stellen.

Aanvankelijk zijn de voorbeelden gelabeld in meerdere categorieën (tabel 1). Binnen deze categorieën is niet alleen plaats voor jaartallen, maar ook voor maten, gewichten en andere getallen. In de derde kolom van tabel 1 is in procenten aangegeven hoe dikwijls zo'n categorie in onze teksten voorkomt.

In de vierde kolom is de mate van succes aangegeven bij het herkennen ervan. Niet onverwacht geldt dat hoe groter de klasse, des te gemakkelijker hij herkend wordt. We zien dus dat 16% van de getallen in archeologische teksten een chronologische expressie behelst, en dat we ongeveer 96% succes hebben bij het herkennen ervan⁴.

In tweede instantie is deze fijnmazige verdeling achterwege gelaten, en hebben we ons beperkt tot slechts twee labels: *jaartal/geen jaartal*. Dit verhoogde de precisie van 96% naar 98% en liet het indexeren veel sneller verlopen.

Het landgoed Ter Meulen is ontstaan uit de watermolen van de Hof van [Almen](#) aan de Berkel bij [Almen](#) (Harenberg, 1992). De hof komt al in 1188 in een goederenlijst voor. De molen zelf wordt voor het eerst genoemd in 1331. Hij wordt later afgesplitst van de Hof van [Almen](#). In 1494 wordt de molen voor het eerst 'Ter Moelen' (Ter Meulen) genoemd (Harenberg, 1975; Anonymus, 1983). Wanneer de watermolen buiten bedrijf is geraakt, is niet bekend. Het moet in ieder geval vóór 1561 zijn geweest. Uit een schriftelijke bron uit 1561 is namelijk sprake van een 'bouman' (boer) op het goed Ter Meulen (Harenberg, 1992). Het zal in die tijd dus geen molen, maar een pachtboerderij zijn geweest. De plek van de mid-

2 TIJDSAANDUIDINGEN IN OPEN
BOEK: MIDDELEEUWEN IS ROOD.

Een praktijktoepassing: open boek

Het principe van Memory Based Learning blijkt dus in theorie te werken. Om de experimenten op het gebied van tijds- en plaatsherkenning ook aan de praktijk te toetsen is een free text information retrieval systeem gebouwd: Open Boek⁵. Dit systeem accepteert pdf-en HTML-documenten, indexeert ze zowel op tijd, plaats als ook 'normaal' op trefwoorden, en stelt de gebruiker vervolgens in staat om te vragen naar 'molens in de mid-

TABEL 2
DE DATABASE MET
VOORBEELDEN.
ELKE REGEL BEVAT
EEN GETAL IN
ZIJN CONTEXT.
DE GELE KOLOM IS
DE FOCUS MET
HET GETAL;
BLAUWE REGELS
ZIJN ALS
TIJDSAAN-
DUIDINGEN
GELABELD.

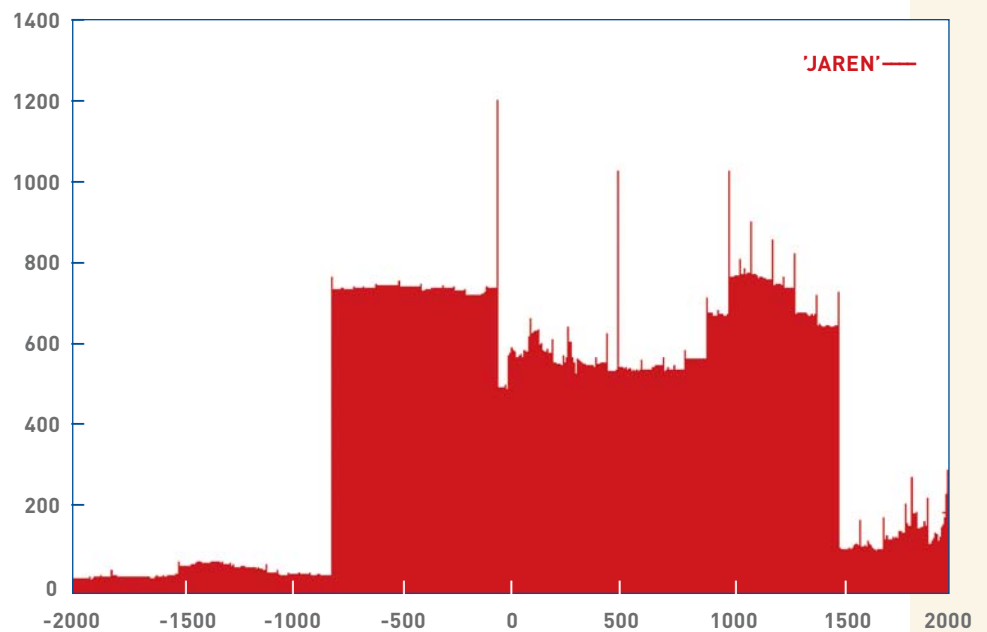
Leiden	2003	Postbus	9515	2300	RA	Leiden	info	[Other]
81,1	Bd	:	beschadigd	59,9	GL	:	beschadigd	[Other]
.	Bloo)	Bijlage	3	Overzicht	van	het	[Reference]
vroegste	complex	(vindplaats	21)	valt	op	[E42_Object_identifier]
op	onge	-	veer	0,2	m	+NAP	,	[E54_Dimension:depth]
slechts	RAAP	-	rapport	969	/	eindversie	12	[E42_Object_identifier]
(Module	3)	3	Vondsten	Het	toekomstige	[Other]
De	Franse	kaart	van	1811	vormt	de	oudste	[E52_Timespan]
te	geven	.	Tabel	6	Aantals	-	en	[Reference]
cultuur	,	uit	de	vierde	eeuw	van	de	[E52_Timespan]
uit	de	Midden	-	Bronstijd	zijn	vijf	sites	[E52_Timespan]
Midden	-	Bronstijd	zijn	vijf	sites	aangetroffen	.	[E60_Number]
IJstijd	:	ca	.	8800	jaar	voor	Chr	[E52_Timespan]
aangetrof	-	fen	.	6	.	conservering	:	[Other]

deleeuwen' en als antwoord dan alle pagina's te krijgen met het woord 'molens' en een referentie naar een jaartal of periode binnen het tijdvak 500-1500 (zie afb. 2). De op het scherm rood onderstreepte woorden en jaartallen beantwoorden aan de zoekvraag; jaartallen en plaatsnamen die niet aan de oorspronkelijke vraag voldoen, maar door het programma wel als jaartal, c.q. plaatsnaam zijn herkend, zijn blauw. In het voorbeeld is overigens niet naar een plaatsnaam gevraagd. In nieuwe teksten wordt 93% van de getallen correct geassocieerd, waarbij de 'missers' vooral liggen bij getallen die ten onrechte als jaartallen worden aangemerkt. Hierbij moet wel worden vermeld dat een correcte classificatie niet noodzakelijk betekent dat het jaartal vervolgens ook juist wordt geïnterpreteerd. Als bijvoorbeeld het jaar '1801' per ongeluk wordt geschreven als '180i', dus met een 'i' in plaats van een '1', zal het vanwege de context wel als jaartal worden herkend, maar vanwege de typefout als '180' worden geïnterpreteerd en als zodanig in de computer worden opgeslagen.

Ons systeem heeft ook nog andere eigenschappen. Zo worden coördinaten eveneens automatisch herkend, en door erop te klikken wordt de gebruiker dan automatisch doorgelinkt naar Googlemaps. Een andere mogelijkheid is om bijvoorbeeld een overzicht te genereren van het aantal keren dat in een selectie van de documenten wordt verwezen naar bepaalde tijdperken. In afbeelding 3 zien we hoe dikwijls naar jaren en tijdperken tussen 2000 voor Chr. en 2000 na Chr. verwezen wordt in een willekeurige verzameling van 100 rapporten uit DANS met bijna 5000 pagina's (zie afb. 3). Opvallend zijn de twee plateaux die worden gevormd door de 'ijzertijd' en de 'middeleeuwen'. Verder zien we elke honderd jaar een piek, die wordt veroorzaakt door de neiging van de auteurs – en ons allemaal – om 'ronde getallen' te gebruiken: het is altijd 'rond het jaar 1000' en nooit 'rond het jaar 1001'. Tenslotte de stijging in de 19-de en 20-e eeuw die wordt veroorzaakt door de literatuurvermeldingen. Zulke 'niet-essencele' jaartallen kunnen overigens worden weggefilterd, zodat alleen de 'echte' jaartallen overblijven.

Het is interessant deze grafiek te bekijken en zich af te vragen hoe deze verdeling tot stand is gekomen. Is er echt een relatief hoogtepunt in de belangstelling tussen 900 en 1300? Wat was er tussen 1500 v. Chr. en 1000 v. Chr. aan de hand? Het zijn vragen die we graag aan de echte archeologen overlaten.

FREQUENTIE VAN JAARTALLEN EN TIJDPERKEN



3 REFERENTIES AAN JAARTALLEN IN 5000 PAGINA'S ARCHEOLOGISCHE PUBLICATIES EN RAPPORTEN.

Zoals we al zeiden indexerend we voor Open Boek alleen pdf (en HTML) documenten; geen Microsoft Word. Daar zijn goede redenen voor. Ten eerste is het zogenaamde format van Word documenten (dat wil zeggen: de interne structuur) niet vrijgegeven en Microsoft verzet zich hevig tegen alle pogingen van derden om dat format toch te lezen. Ten tweede is er binnen de beschermde Windows-omgeving een wildgroei ontstaan van leuke maar niet essentiële toeters en bellen, waar auteurs dan ook graag gebruik van maken. Dit heeft echter tot gevolg dat zo'n document, bijvoorbeeld een CD met de gegevens van een bepaalde opgraving, niet meer buiten een identieke Windowsomgeving bekeken kan worden, en al helemaal niet meer in een geautomatiseerd retrieval systeem kan worden opgenomen.

Alleen op de RACM in Amersfoort liggen zo al honderden CDs met recent materiaal in allerlei Microsoft-gebonden formats, die in snel tempo ontoegankelijk zullen gaan worden. In verband hiermee zijn we dan ook zeer gelukkig met de motie Heemskerk die in december 2007 door de Tweede Kamer is aangenomen, en waarin is vastgelegd dat de Nederlandse overheid vanaf het voorjaar 2008 overgaat op Open document formats en het gebruik van Open Source gaat bevorderen. Alleen zo is duurzame toegang tot de gegevens te garanderen.

Samenvattend: met dit systeem, Open Boek, hebben we een instrument ontworpen om snel en efficiënt chronologische en geografische informatie te vinden in normale, Nederlandstalige boeken en artikelen. Hoewel de toepassing ervan niet beperkt hoeft te blijven tot de archeologische of historische context hopen we toch dat het geadopteerd zal worden door de doelgroep voor wie we het ook hebben ontwikkeld. En dat Open Boek zal uitgroeien tot de 'Google' van de cultuurhistorie. ■

¹ A. Nieuwhof and A.G. Lange: *Op weg naar een nationale referentiecollectie*, Rijksdienst voor Oudheidkundig bodemonderzoek Amersfoort, 2003

² R. Wiemer en P. A. L. M. Janssen, *Cultuurhistorische informatie samen in een nieuw jasje*. In *Geo-Info* 3 (10), oktober 2006.

³ Walter Daelemans, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch: *TiMBL: Tilburg Memory Based Learner version 5.1 Reference Guide*. *ILK Technical Report 04-02*, 2004

⁴ J.J. Paijmans and S. Wubben: *Preparing archeological reports for intelligent retrieval*, in: CAA-2007, Proceedings of CAA-2007 (in press). Berlin, Germany, 2007

⁵ J.J. Paijmans and S. Wubben: *Open Boek: a system for the extraction of numeric data from archeological reports*, in: AHM-2007, 2007

– Open Boek:
www.referentiecollectie.nl/Openboek