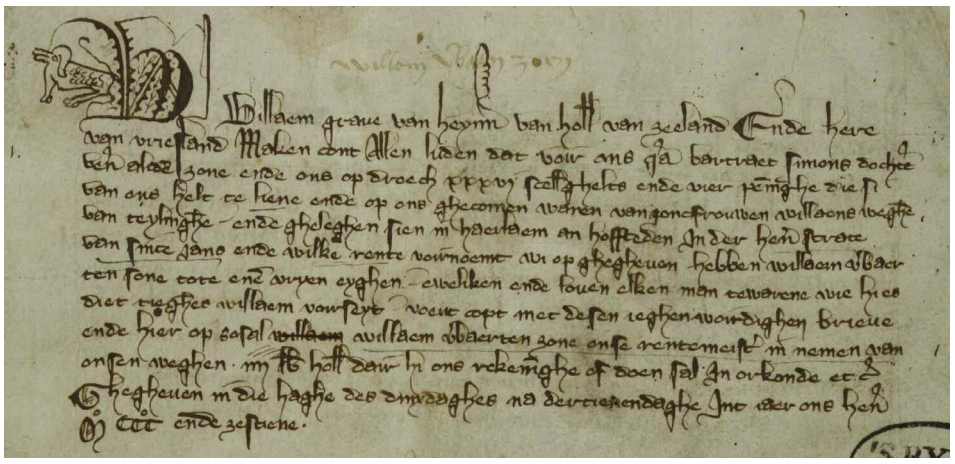
De Hollandse graven hebben hun oorkonden nu op het net

Ze werden in de de veertiende eeuw samengesteld, en ze zijn nu bijna allemaal op het wereldwijde web te bekijken en te lezen: de registers van de Hollandse grafelijkheid. *Jan Burger*



Oorkonde uit 1317, waarin graaf Willem III aan zijn rentmeester een renteleen schenkt van 36 schelling en 4 penning

Bij het Huygens Instituut voor Nederlandse geschiedenis (Huygens ING) wordt sinds najaar 2006 gewerkt aan een digitale editie van die registers, die door de klerken van de graven van Holland, Zeeland en West-Friesland werden bijgehouden. Dit project nadert zijn voltooiing; inmiddels zijn 2953 documenten gepubliceerd. In totaal worden zo'n 3600 documenten uit de periode 1316-1345 ontsloten. Het zijn voornamelijk oorkonden, meestal door de graaf uitgevaardigde stukken met betrekking tot zijn diverse territoria, terwijl er ook registers zijn met teksten die de relaties met het buitenland betreffen.

Oorkonden vormen voor de mediëvist een belangrijke bron. Ze bieden een schat aan feitelijke informatie over allerlei aspecten van maatschappij en bestuur. Daarom worden de oorkonden al sedert de 18e eeuw grote aantallen uitgegeven. Vanaf de 19e eeuw gebeurt dat op een wetenschappelijk verantwoorde wijze, met de Duitse Monumenta Germaniae Historica als lichtend voorbeeld. In Nederland zijn vanaf het begin van de 20e eeuw oorkondenedities verschenen van bijvoorbeeld Holland en Zeeland, van Utrecht, van Gelre en van Noord-Brabant, elk met daarin alle overgeleverde oorkonden over het betreffende territoir. Deze edities eindigen rond 1300, omdat de oorkonden zo talrijk worden dat volledigheid in een wetenschappelijk verantwoorde uitgave onmogelijk wordt.

Geretrodigitaliseerd

Digitalisering opende nieuwe mogelijkheden voor oorkondenedities, en daarbij worden internationaal verschillende strategieën gevolgd. Om te beginnen worden vele van de (honderden!) bestaande papieren edities geretrodigitaliseerd; dergelijke plannen bestaan ook voor de oorkondenboeken die eertijds zijn uitgegeven door het Huygens ING. Een koppeling van dergelijke edities over de landsgrenzen is dan een logische stap, omdat tot in de 13e eeuw oorkonden een grotendeels uniforme Europese cultuur weerspiegelen, en zij merendeels zijn geschreven in dezelfde taal, het Latijn.

Een tweede optie is om door te gaan met het uitgeven van materiaal op de oude, wetenschappelijk verantwoorde en integrale editiewijze, maar nu digitaal: het Oorkondenboek van Noord-Brabant is daarvan een voorbeeld (www.donb.nl).

Blijft het probleem van de grote massa's oorkonden van na 1300. Een mogelijkheid is het om afbeeldingen van zoveel mogelijk oorkonden online te zetten, zonder de pretentie van wetenschappelijk editeren. Voorbeelden hiervan zijn het Gronings-Drentse Cartago (www.cartago.nl), en vooral het Monasterium-project (www.monasterium.net), dat diverse midden-Europese regio's bestrijkt (inmiddels meer dan 220.000 oorkonden).

Voor de digitale editie van de Hollandse grafelijke registers is weer een andere aanpak gekozen, namelijk een wetenschappelijke uitgave, maar dan niet meer gericht op regionale compleetheid. Anders dan bij de papieren voorganger, het Oorkondenboek van Holland en Zeeland tot 1299, worden hier niet alle oorkonden uitgegeven, maar is dit een integrale editie van een afzonderlijk, maar omvangrijk bronnencomplex. Tegenover de driekwart eeuw van het Oorkondenboek, zal het registerproject uiteindelijk zes jaar vergen. En daarbij komen de voordelen van een digitale uitgave: er zijn uitgebreide zoekmogelijkheden in en door de teksten, en de registers zijn compleet afgebeeld. De onderzoeker krijgt op deze wijze een optimale toegang tot het materiaal.

www.historici.nl/Onderzoek/Projecten/ RegistersVanDeHollandseGrafelijkheid1299-1345

Google wil graag iets terugdoen

Jon Orwant is engineering manager bij Google en werkt vooral aan Google Books. Hij was in Nederland op uitnodiging van NWO. E-data sprak met hem. *Peter Boot*

Eén ergernis voor de niet-Amerikaanse gebruikers van Google Books is dat we sommige boeken niet te zien krijgen, terwijl het auteursrecht toch echt verstreken is.

Orwant: "We kennen niet altijd de precieze publicatiedatum van een boek – we verzamelen boekgegevens uit heel veel bronnen, en het is een heel werk om die te matchen. Het volgende probleem is dat we geen betrouwbaar register van overlijdensdata van auteurs hebben. We moeten voorzichtig zijn, want mensen procederen graag tegen Google - zeker omdat in de VS bij schendingen van het auteursrecht schadevergoedingen tot $150 000 worden toegekend."

Bij veel digitaliseringsprojecten laat de kwaliteit van de OCR (Optical Character Recognition) te wensen over. Werken jullie daar aan?

Jon Orwant

"Ja, we hebben een OCR-team, dat trouwens de software als open source ter beschikking stelt. Bij de Gotische letters en klein afgedrukte tekst hebben we de laatste tijd veel vooruitgang geboekt. "

Werkt Google ook aan manieren om tekststructuur in een boek te analyseren?

"Ja. Ik denk dat we onze boeken al kunnen taggen in TEI Lite (TEI: Text Encoding Initiative, standaard voor codering van teksten). Veel verder in de analyse zouden we nu niet durven gaan. We zouden liever een basistekst leveren, en dan de wetenschapper de mogelijkheid geven om de tekst te annoteren. Maar de infrastructuur daarvoor vergt nogal wat. Hoe weten we dat iemand geen onzin typt? Moeten we anoniem commentaar toestaan? Wat moet de eenheid van annotatie zijn: een boek, een bladzijde, een hoofdstuk, een woord? Mag iemand zijn eigen annotaties nog wijzigen?"

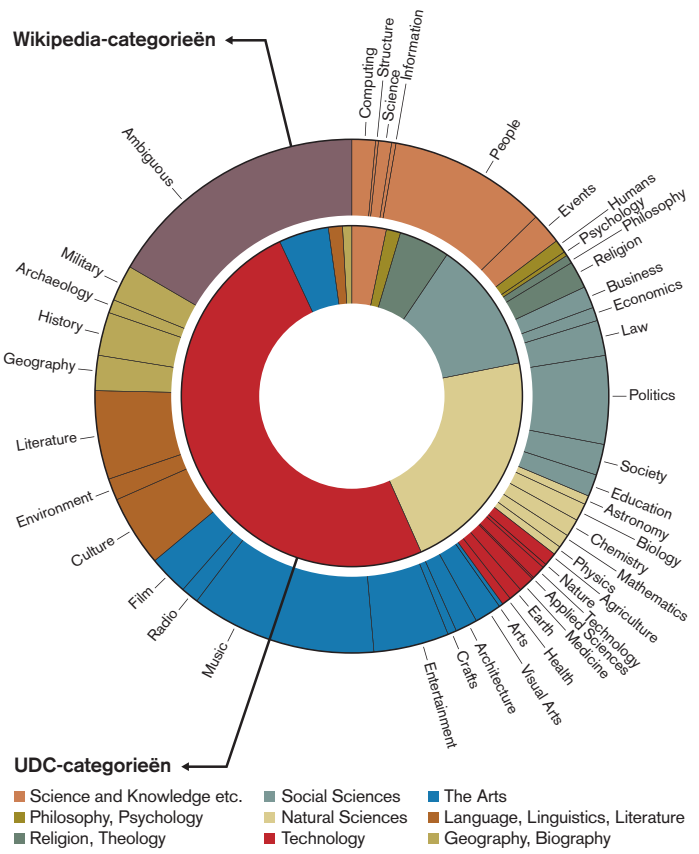
Wordt dat een betaalde dienst?

"Nee, het gaat Google goed, en willen iets teruggeven aan de wereld. We verdienen nu wel iets aan de boeken, door advertenties en het verkopen van digitale versies, maar dat haalt het niet bij wat het scannen gekost heeft. In dezelfde spirit zijn de corpora van onze n-gram viewer (een website waar het gebruik van woorden in de tijd kan worden gevolgd – PB) *open source* beschikbaar. Iedere onderzoeker kan de bestanden downloaden en herpubliceren met een eigen interface."

Wanneer komt de Nederlandse versie?

"Die komt, maar wanneer weet ik nog niet. We willen eerst zeker weten dat de collectie voldoende representatief is. Van de 3,1 miljoen Nederlandstalige boeken hebben we er nu 168.000 gescand, maar overwegend ouder materiaal. Dat is nog niet genoeg. Houd er trouwens rekening mee dat we voor deze bestanden regelmatig met nieuwe versies zullen komen, omdat we nieuwe boeken hebben gescand en onze software weer beter is geworden. Analyses op de bestanden zullen dus altijd moeten aangeven met welke versie ze zijn uitgevoerd. Daarmee wordt onderzoek in de digitale geesteswetenschappen repliceerbaar. En omdat de zichtbaarheid van onderzoek met digitale bronnen enorm toeneemt, willen we er zeker van zijn dat het onderzoek is dat de toets der kritiek kan doorstaan."

http://books.google.com/
http://ngrams.googlelabs.com/
http://googlebooks.byu.edu/



UDC en Wikipedia: ontworpen versus gegroeid

De twee schijven hierboven stellen twee indelingen van informatie voor: de categorie-indeling van de Universal Decimal Classification (UDC, binnenste ring) en die van Wikipedia (buitenste ring). In de twee ringen wordt het contrast zichtbaar tussen het betrekkelijk formele UDC (ontwikkeld aan het begin van de twintigste eeuw en veel gebruikt in bibliotheken) en Wikipedia, dat meer een gegroeid, sociaal systeem is. In UDC is veel ruimte voor technologie (toegepaste wetenschappen), terwijl in Wikipedia kunst, entertainment en sport dominant zijn. Voor een belangrijk deel komt dat doordat Wikipedia alle informatie behandelt, niet alleen de wetenschap. De tekening is een bewerking van een beeldgrafiek van het Knowledge Space Lab project, dat wordt gesubsidieerd door het Strategiefonds van de KNAW. Deelnemers zijn de Virtual Knowledge Studio, the e-humanities group, DANS, de Erasmus Studio Rotterdam en BigGrid NL. De beeldgrafiek is onderdeel van een prachtig gekleurde kaart die onlangs is geselecteerd als een van de tien beste in de jaarlijkse wedstrijd van Places & Spaces.

De originele kaart, met nog veel meer graphics, is te zien op ***http://scimaps.org/maps/map/design_vs_emergence__127/***