Manja Koomen en Renée Verdegaal

Elektronische tekstarchieven

Een pleidooi voor standaardisatie

Er is wildgroei in opzet en inrichting van elektronische tekstarchieven (eta's), menen Manja Koomen en Renée Verdegaal. Om het de gebruiker aangenamer en makkelijker te maken, pleiten ze voor het verhogen van de gebruikersvriendelijkheid en het invoeren van standaardisering bij eta's.

Ook voor menige dichtbundel en literair werk kunnen we tegenwoordig terecht op het world wide web. Alfawetenschappers, die lang de naam hebben gehad terughoudend te zijn in het gebruik van techniek in het algemeen, hebben ingezien hoe ontstellend handig het is om taalonderzoek met de computer te doen en hoe leuk het is om de *Divina Commedia* interactief te lezen.
Op dit moment ontstaan er overal ter wereld elektronische tekstarchieven, vaak van wetenschappers die een eigen onderzoeksobject digitaal willen bestuderen. Deze tekstarchieven krijgen vaak een plaats op het internet omdat men ook andere academici van het eigen werk wil laten profiteren. Het probleem van deze individuele aanpak van projecten is, dat er wildgroei ontstaat rond opzet en inrichting van de tekstarchieven. Een lappendeken van verschillend geïnterpreteerde standaarden is het gevolg en daarvan worden de gebruikers uiteindelijk de dupe. Dat is jammer, want een tekstarchief kan, mits het goed is ingericht, uitstekende mogelijkheden bieden om de gebruiker goed te bedienen en in korte tijd te laten vinden wat hij zoekt.

Waarom een eta?

Elektronische tekstarchieven bieden vele voordelen. Zo zijn digitale bestanden makkelijk op te slaan en nemen zij in tegenstelling tot gedrukte teksten weinig ruimte in. Ook kan in digitale bestanden makkelijk gezocht, geknipt, geplakt en geïndexeerd worden.
Een ander voordeel is dat de gebruikers met velen tegelijk, op elk tijdstip en vanuit elke plaats op de wereld toegang hebben tot het materiaal. Daarbij kan het niet gestolen worden of wegraken. Naast materiaal dat al in digitale vorm beschikbaar is, kunnen digitale archieven ook ingericht worden met een digitale surrogaatvorm van zeldzaam of in slechte staat verkerend materiaal. Wetenschappelijk bronmateriaal dat zich in de kluizen van musea en bibliotheken bevindt wordt zo makkelijker toegankelijk voor een groot publiek en het originele materiaal kan op optimale wijze behouden blijven.

Een elektronisch tekstarchief

Een elektronisch tekstarchief is een bewaarplaats voor digitale informatie. Elektronische tekstarchieven zijn, door het toepassen van diverse conversiemethoden, gezamenlijk verantwoordelijk voor het waarborgen van de integriteit en toegankelijkheid op lange termijn van sociaal, economisch, cultureel of intellectueel erfgoed in digitale vorm.

Ook in het veld van onderwijs en onderzoek biedt een elektronisch tekstarchief tal van mogelijkheden om tekst te analyseren die op een andere manier niet realiseerbaar zouden zijn.

Wat komt er bij kijken?

Een goed tekstarchief op- en inrichten is anno 2000 niet eenvoudig. Er is door veel instellingen echter al uitgebreid gepionierd en geëxperimenteerd en wie slim is kan daar de vruchten van plukken.
Bij het plannen maken voor een elektronisch tekstarchief moet men zich allereerst afvragen wie het tekstarchief gaat gebruiken. De tweede vraag is wat er in het archief opgenomen moet worden, terwijl de derde en meest complexe vraag luidt hoe de teksten gebruikt moeten kunnen worden. Die vraag naar het hoe heeft diverse aspecten.

Kwaliteitseisen

Het voornaamste aspect is de integriteit van de teksten. Gedigitaliseerde documenten kunnen namelijk vele vormen aannemen al naar gelang de software waarmee er naar gekeken wordt. Het is daarom van belang dat er tevoren wordt nagedacht over het behoud van de integriteit van de documenten. Vijf essentiële punten zijn:

Inhoud De inhoud van een document behelst niet alleen de tekst op zich, maar ook het formaat en de structuur daarvan, evenals speciale tekens en afbeeldingen. Deze aspecten van het document moeten ook in een digitale omgeving gewaarborgd kunnen worden.

Onveranderlijkheid In een digitale omgeving is het makkelijk om het originele document te veranderen. Het is belangrijk dat het canonieke object, ofwel de algemeen aangenomen tekst of leestekst behouden blijft en dat alle veranderingen gelinkt worden naar het object in een makkelijk toegankelijke database.

Ontsluiting Een digitaal object is steeds moeilijker te vinden. Dit geldt met name voor documenten die geen papieren equivalent hebben. Het is zinvol om de objecten te ontsluiten en structureren volgens internationale standaarden, bijvoorbeeld TEI SGML of TEI XML.

Herkomst Voor onderzoek is het noodzakelijk dat duidelijk is wie de tekst heeft geschreven en met welke oorspronkelijke papieren uitgave men te maken heeft.

Context Context refereert aan de hard- en softwarevereisten van een digitaal object: het is belangrijk om te

Enkele voorbeelden

Een interessant internationaal tekstarchief is het **Oxford Text Archive** (http://ota.ahds.ac.uk/), een van de bekende tekstarchieven ter wereld. Het heeft door jarenlange ervaring ook een voorbeeldfunctie. Het werd in 1976 opgericht door Lou Burnard, medewerker van de afdeling automatisering aan de universiteit van Oxford. Het archief bevat op dit moment 2500 documenten in 25 talen. Het bestaat uit literaire teksten, dissertaties, tekst corpora, bibliografieën, woordenboeken, essays, kookboeken etcetera.

Een tweede tip is het **Electronic Text Center** van de University of Virginia (http://etext.lib.virginia.edu/). Het Text Center heeft de teksten sinds zijn oprichting in 1992 consistent gecodeerd in SGML, wat betekent dat de data nu nog even bruikbaar zijn als toen. Standaardisering is voor Virginia een kernuitgangspunt. Het Electronic Text Center omvat ongeveer 45.000 on- en offline documenten uit de humaniora met meer dan 50.000 afbeeldingen die op de teksten betrekking hebben (illustraties, covers, krantenpagina's etcetera). Om enigszins structuur aan te brengen in de grote hoeveelheid teksten is er een onderverdeling gemaakt op taal. Eén daarvan is het **Japanese Text Archive** (http://etext.virginia.edu/japanese/index.html). Behalve dat het archief uiterst gebruikersvriendelijk is, is een interessante toepassing dat sommige gedichten worden aangeboden in drie parallelle frames die interactief doorzocht kunnen worden.

Een bijzondere kleine collectie is het **Victorian Women Writers Project** (www.indiana.edu/~letrs/vwwp/index.html). Dit eta bevat teksten van Britse vrouwen uit de negentiende eeuw. De doelstelling van het archief is om 'highly accurate transcriptions' van de teksten te produceren en daarvoor TEI SGML te gebruiken. De teksten blijven zeer dicht bij de gedrukte versies in de zin dat er paginanummers en verwijzingen naar de voetnoten gegeven worden.

Ronduit spectaculair is **The World of Dante** (www.iath.virginia.edu/ dante/). Het archief bevat alle canto's van Dante's *Hel* gecodeerd in TEI SGML. Het bijzondere is dat alle passages die over personen gaan, de geografische plaatsen op aarde en in de hel, mythische figuren, goden en godinnen, architectonische en artistieke structuren helemaal gecodeerd zijn. Een figuur is dus niet alleen gecodeerd wanneer zijn naam wordt genoemd, maar ook wanneer Dante aan hem of haar refereert. Liefhebbers krijgen ook nog de mogelijkheid om door de driedimensionale tekst ervan te navigeren.

Uit dezelfde school komt het **William Blake Archive** (www.iath.virginia.edu/blake/main.html). Het archief bevat een aantal versies van Blake's geïllumineerde boeken en is het ultieme voorbeeld van de mogelijkheden en functionaliteiten die een elektronisch archief op dit moment kan bieden. Verder heeft het Blake-archief een primeur: een zeer geavanceerde zoekmogelijkheid voor afbeeldingen.

The Charrette Project (www.princeton.edu/~lancelot) is een tekstarchief dat geheel gewijd is aan de tekst *Le Chevalier de la Charrette* (Lancelot), geschreven door Chretien de Troyes rond 1180. De moderne versie bevat ongeveer 7100 verzen. Deze edities zijn gebaseerd op acht verschillende manuscripten uit de twaalfde eeuw die alle in dit archief opgenomen zijn. Het archief biedt de mogelijkheid om te wisselen tussen delen uit edities om vergelijkingen te kunnen maken.

De **ALEX** (http://sunsite.berkeley.edu/alex/) database bevat teksten die geclassificeerd kunnen worden als Amerikaanse of Engelse literatuur of teksten uit de westerse filosofie. Alleen teksten die beschouwd kunnen worden als 'Great Literature' mogen opgenomen worden. Een aardige mogelijkheid die ALEX biedt om tekst te kunnen bestuderen en bewerken is het maken van een eigen boekenkastje: een soort persoonlijke collectie met teksten die jezelf kunt bewerken en doorzoeken.

Als laatste nog een archief van eigen bodem: het **Laurens Janszoon Coster** (www.dds.nl/~ljcoster/index.html) project. Dit archief biedt rechtenvrije Nederlandstalige literaire teksten aan in elektronische versie. Het project wordt gerund door vrijwilligers en van standaardisering is geen sprake. Toch is het zeer de moeite waard: het project wordt met veel plezier beheerd en heeft dan ook 1800 bezoekers per dag.

weten hoe het object oorspronkelijk werd aangeboden. Een digitaal object kan namelijk wezenlijk veranderd worden als het met verkeerde software wordt bekeken.

Functionaliteitseisen

Naast het formuleren van kwaliteitsuitgangspunten voor de content van het archief, is het ook belangrijk een goede opbouw van het archief na te streven.
De volgende functionaliteiten zijn noodzakelijk voor een optimaal archief:

Standaardisering Er moet worden gekozen voor internationaal erkende codeertaal. Het meest geschikt hiervoor is TEI SGML of de nieuwe versie TEI XML. Deze vorm van coderen maakt het behoud van de teksten op langere termijn mogelijk, terwijl de teksten goed uitwisselbaar zijn. Daarbij biedt TEI XML tal van interessante verrijkingsmogelijkheden.

Structuur De userinterface is duidelijk en helder gestructureerd. Het elektronisch tekstarchief zit hiërarchisch en logisch in elkaar. Het archief bevat informatie over het elektronisch tekstarchief (welke teksten bevat het en van wie, doelstelling etcetera), het tekstarchief zelf en een zoekfunctie. De site is daarbij goed navigeerbaar, elke pagina biedt buttons om terug te keren naar de homepage en naar de helpfunctie.

Retrieval Er zijn uitgebreide zoekmogelijkheden (booleaans) en er wordt duidelijk uitgelegd waar men precies in zoekt. Daarbij kun je langs de titels en auteursnamen (geautoriseerde lijst met namen) browsen. Een vergevorderd elektronisch archief zou ook de mogelijkheid moeten bieden om te kunnen zoeken in de teksten die een zoekactie hebben opgeleverd als het resultaat te groot is. Ook het maken van een persoonlijke collectie zou wenselijk zijn als het archief veel teksten bevat. Uiteindelijk zou de gebruiker ook heel complexe vragen aan het archief moeten kunnen stellen (bijvoorbeeld: ik wil alle teksten die tussen datum X en Y zijn verschenen waarin het woord A voorkomt en die in land B zijn verschenen).

Editie Er wordt informatie gegeven over de gebruikte edities. TEI XML biedt daarbij de mogelijkheid om verschillende edities op één scherm met elkaar te vergelijken, een toepassing die zeer gewenst is onder wetenschappers. Als men de gebruiker echt goed wil bedienen, kunnen de pagina's gepagineerd en links naar de voetnoten aangebracht worden.

Verrijking Door de teksten te coderen in TEI SGML/XML kan er gemakkelijk informatie worden aangeboden over de gebruikte editie en de structuur van de tekst. De teksten kunnen zowel in SGML/XML als in HTML bekeken en gedownload worden. Voor beginners en gevorderden wordt duidelijke uitleg gegeven over TEI SGML/XML en een link aangeboden naar de plek waar de browsers (tegen betaling) gedownload kunnen worden. TEI XML is uitermate geschikt om teksten en afbeeldingen te ontsluiten met trefwoorden, zelfs op paragraafniveau. Andere vormen van verrijking zijn het aanbieden van facsimile's, secundaire informatie en een zoekfunctie die geheel toegespitst kan worden op de individuele gebruiker.

Interactie Gebruikers kunnen voor veel problemen en vragen komen te staan bij het bezoeken van een elektronisch tekstarchief. Het is van belang dat op deze vragen antwoord wordt gegeven. Denk hierbij aan een FAQ-rubriek of een reply e-mail.

Beheer Het archief wordt up-to-date gehouden, heeft een feedback-mogelijkheid en een goede helpfunctie.

Extra Als alle bovenstaande toepassingen gerealiseerd zijn, beschikt men over een goed tekstarchief. Maar er zijn nog een heleboel leuke en nuttige toepassingen te bedenken, zoals een discussielijst, een tour door het archief, een maandelijkse nieuwsrubriek, het aanbieden van achtergrondinformatie over auteurs, secundaire literatuur en natuurlijk een lijst met relevante websites.

SGML

SGML is een internationale standaard (ISO 8879) die een (computer)taal definieert waarmee tekst beschreven kan worden. Deze standaard biedt handvaten waarmee de informatie-eenheden in een document beschreven kunnen worden.
Doordat SGML-documenten gewone ASCII-bestanden zijn en SGML een internationale standaard is, zijn de bestanden makkelijk overdraagbaar (platform- en software-onafhankelijk) en hoeven er geen ingewikkelde conversies uitgevoerd te worden. Hierdoor gaat er ook geen informatie verloren.
De ontwerper van een SGML-document kan een tekst coderen naar eigen behoefte, mits de standaard wordt aangehouden. Dat kan bijvoorbeeld zijn: titel, hoofdstuk, paragraaf, bladzijde, vers, strofe, lijn, act, scene, naam, citaat, lijst, datum, etcetera. Maar dat kunnen ook analytische kenmerken zijn zoals woordklasse, of vormen van linguïstische, historische of literaire interpretatie. Voor een bepaalde klasse van documenten kan in een Document Type Definition (DTD) worden vastgelegd welke elementen op welke wijze gecodeerd kunnen worden en wat hun onderlinge structuur en relaties zijn. Metadata of beschrijvende informatie over een tekst, zoals de fysieke conditie van een manuscript, de plaats van een annotatie op het manuscript, of de kleur en vorm van een initiaal, kunnen evengoed als SGML-objecten gecodeerd worden.
Doordat SGML platformonafhankelijk is, hebben documenten een lange levensduur.
Documenten hoeven niet meer geconverteerd te worden als een hard- of softwaresysteem verouderd raakt.
Een nadeel van SGML is dat het een grondige kennis vereist van documentstructuren. Daarbij is de conversie naar SGML een kostbare aangelegenheid.

TEI SGML

TEI staat voor Text Encoding Iniative. Het TEI is een voor tekstcorpora, zoals boeken, brieven, gedichten etcetera, ontworpen internationale coderingsstandaard (zie www.uic.edu/orgs/tei/). Het TEI biedt een DTD (of enkele DTD's) om dit soort tekstdocumenten in SGML te documenteren. De richtlijnen zijn gebaseerd op SGML. De TEI SGML is ontwikkeld omdat er vanuit de academische wereld behoefte ontstond aan consensus. Het werk van verschillende werkgroepen, bestaande uit experts uit verschillende vakgebieden, resulteerde in 1990 in de eerste versie van de Guidelines. De richtlijnen van het TEI zijn vrij verkrijgbaar op het web.

XML

De complexiteit van SGML als metataal heeft ertoe geleid dat zij toch nog op relatief kleine schaal wordt toegepast. Dus presenteerde het W3C (World Wide Web Consortium) in 1996 een 'extreem simpel dialect' van SGML, Extensible Markup Language, of XML genaamd (zie www.ucc.ie/xml/). De eenvoud van XML maakt het zeer geschikt voor uitwisseling over het internet. Zij zal dan ook binnen niet al te lange tijd SGML en HTML als webpublishers vervangen. De tools om XML te kunnen lezen worden op dit moment geïntegreerd in standaardbrowsers als Microsoft Internet Explorer. Verwacht wordt dat de software die geschikt is voor XML aanmerkelijk goedkoper zal worden dan die voor SGML. Ook kunnen TEI DTD's, enigszins aangepast, voor XML gebruikt worden.

Gebruiker centraal

Tekstarchieven zijn leuk, er is op het internet al heel wat (gratis) literatuur te vinden die interactief bestudeerd kan worden. Denk hierbij aan gedichten in audiovorm of als animatie, zeer geavanceerde vormen van tekstvergelijking, het creëren van een eigen boekenkastje met zelf verzamelde tekst, discussielijsten over literatuur en zoekscripts voor afbeeldingen. Als er wordt besloten om een tekstarchief te gaan maken, bedenk dan goed dat het verzamelen van zoveel mogelijk teksten geen doel op zich is. De gebruiker moet bij een tekstarchief altijd centraal staan. Het is daarom goed om hen voortdurend te betrekken bij het ontwikkelen en onderhouden van het tekstarchief.

URL's

- Overzicht van tekstarchieven die gebruik maken van SGML/XML standaard: www.oasis-open.org/cover/acadapps.html
- Preserving Digital Information: Report of the Task Force on Archiving of Digital Information, (1 mei 1996): www.rlg.org/ArchTF/tfadi.index.htm

Manja Koomen is documentalist bij het NOS-journaal. Zij deed in opdracht van het ETCL onderzoek naar de functionaliteit van elektronische tekstarchieven die worden aangeboden op internet. Op dit onderwerp is zij inmiddels afgestudeerd aan de IDM te Amsterdam. Zij heeft voor haar scriptie de aanmoedigingsprijs van het Victorine van Schaickfonds gekregen.
Renée Verdegaal is projectleider van het Electronic Text Centre Leiden (ETCL). Het ETCL is een onderdeel van de Universiteit Leiden en houdt zich bezig met de advisering, begeleiding en productie op het gebied van elektronische informatie.

Het Electronic Text Centre Leiden

Het Electronic Text Centre Leiden (ETCL) is aan het eind van 1997 opgericht. Door een samenwerkingsverband te creëren tussen de Universiteitsbibliotheek, de faculteit Letteren en de Informatiseringsgroep werden de voorwaarden geschapen om tot een geïntegreerde aanbieding, verspreiding én archivering van waardevolle wetenschappelijke bronnen en culturele producten van de Leidse universiteit te komen. Vanaf het begin is een van de ambities het op- en inrichten van elektronische tekstarchieven geweest. Leiden heeft daarbij gekozen om zich aan te sluiten bij internationale standaarden. Het ETCL ontwikkelt momenteel samen met andere wetenschappelijk instituten richtlijnen (www.etcl.nl/TEIguide) waaraan proza, poëzie en drama opgenomen in een archief aan zou moeten voldoen.
Het ETCL heeft inmiddels al verschillende projecten op dit gebied gerealiseerd, waaronder het XMLarchief van Constantijn Huygens (www.etcl.nl/goldenage/huygens.stm) dat deel uitmaakt van het Golden Age-project (www.etcl.nl/goldenage). In de zomer gaat ook de Digitale bibliotheek voor de Nederlandse letteren (DBNL) de lucht in (www.dbnl.org). In DBNL zal een breed gevarieerd corpus van Nederlandse literaire teksten worden aangeboden. Naast het digitaal weergeven van oorspronkelijke bronnen voor letterkundig, taalkundig en cultuurhistorisch onderzoek zal van diverse titels tevens een eerder gepubliceerde danwel nieuw te vervaardigen editie raadpleegbaar zijn. Door het gedigitaliseerde materiaal te 'verrijken', niet alleen met de gebruikelijke annotaties, maar ook met illustraties, tekstfragmenten, afbeeldingen van de oorspronkelijke bron en hyperlinks naar andere onderdelen van het digitale archief alsook naar informatiebronnen elders op het internet.
Twee andere projecten van het ETCL zijn *De opstand in de Nederlanden of de Tachtigjarige Oorlog* (http://dutchrevolt.leidenuniv.nl/) en *Bijzonder Collecties* (www.etcl.nl/bc/). De eerste digitale bibliotheek biedt bronnen zoals brieven, staatstukken, ooggetuigeverslagen, geschiedwerken en andere boeken. Bijzondere collecties bevat uitgebreide info en bronnenmateriaal van de bijzondere collecties van de Universiteitsbibliotheek van Leiden.