

Opzet deelproject 4 - Transcriptiemodule

datum: 17-12-2007

versie 4

auteur: Jules Lauwerier, projectleider Virtuele Onderzoeksruimte Archief4all

Dit document beschrijft de opzet voor deelproject 4 "Transcriptiemodule" van het project "Virtueel Onderzoeksruimte Archief4all".

Doel van het deelproject

In de periode mei tot oktober 2007 is geïnventariseerd op welke wijze het project Virtuele Onderzoeksruimte Archief4all een bijdrage kan leveren aan de ontwikkeling van de samenwerking tussen archiefinstellingen en hun gebruikers. Daarbij is aangegeven dat binnen het op te leveren prototype van een virtuele onderzoeksruimte behoefte is aan "tools" die gebruikers ondersteunen bij hun onderzoek. Er is geconcludeerd dat een tool die gebruikers ondersteunt bij het maken van transcripties van (oude) teksten, daar een belangrijke toevoeging in zal zijn. Met dit deelproject wordt zo'n tool gerealiseerd.

Het primaire doel van dit deelproject is daarom:

Het opleveren van een (software)module die in elke willekeurige web site van een instelling is te integreren of daarvandaan is aan te roepen met als kenmerkende functionaliteit het presenteren van een gescand archiefstuk en het online transcriberen ervan met als eindresultaat een online opgeslagen tekstbestand.

Subdoelen:

1. Aan een transcriptie kunnen (na elkaar) meer (geautoriseerde) personen deelnemen.
2. Het (tussen)resultaat is zichtbaar voor iedereen, maar is beschermd tegen (onbedoelde of bewuste) wijzigingen door anderen dan de geautoriseerde personen.
3. De transcriptie moet ook gedownload kunnen worden (in elk geval voor degenen die er aan gewerkt hebben).

Resultaat: een software-module met documentatie (functioneel en technisch ontwerp, broncode, gebruikershandleiding, beheerdershandleiding, installatie-instructie)

Termijn:

- januari 2008 specificatie maken
- januari-maart 2008: bouw
- maart-april 2008 getest opgeleverd
- april-mei 2008: proof-of-concept door integratie in een andere web site en reacties van geselecteerde eindgebruikers

Afbakening: Het bepalen welke archiefstukken voor transcriptie aangeboden worden is geen onderdeel van dit deelproject. De module kan op de beschreven wijze worden ingezet voor transcripties waarvoor eindgebruikers belangstelling hebben of waarvan een instelling om een andere reden belang hecht aan transcriptie (bijv. gescande originelen van fragiele stukken). Er is binnen de module geen zoekfunctie of overzichtsfunctie van beschikbare scans o.i.d. Scans moeten via een URL benaderbaar zijn.

Aanpak en samenwerking:

Bij dit deelproject wordt samengewerkt met een (nog te benaderen) archiefinstelling. Er wordt in overleg met de partner en de bouwer een functionele specificatie opgesteld en de technische randvoorwaarden worden verder uitgewerkt. Van de partner wordt verwacht dat er één of meer inhoudelijk deskundigen meerwerken aan de specificatie. Er wordt gedacht aan 1 sessie van 4 uur met daarna twee commentaarrondes op het door de geselecteerde bouwers i.s.m. de projectmedewerkers opgestelde concept. Als voorbereiding op de sessie dient deze beschrijving van het deelproject. Verwacht wordt dat de medewerkers de behoeften van de instelling kunnen formuleren en namens de instelling kunnen praten.

We willen bij de specificatie en natuurlijk de testfase ook uitdrukkelijk eindgebruikers betrekken. Het is gewenst dat de contacten die de instelling met eindgebruikers heeft hiervoor worden benut. Aanvullend zal ook direct met eindgebruikers contact gezocht worden door de projectgroep.

De inzet voor de instelling tijdens de bouwfase zal afhankelijk zijn van de wijze waarop de bouw plaats zal vinden. Naar verwachting zal dat betekenen dat er enkele malen tussentijds overleg zal zijn over de tussenresultaten. De medewerkers van de instelling worden ook betrokken bij het testen. Afhankelijk van de behoefte wordt de instelling ook betrokken bij de presentaties aan het eind van het project.

Kenmerken van de functionaliteit:

1. De transcriptie wordt gemaakt door de eindgebruiker. De module bezit (intitueel ?) geen intelligentie om de interpretatie te ondersteunen (zoals OCR, spellingscheck, (oude-) woordenlijsten o.i.d.).
2. De module moet taalonafhankelijk zijn, zodat ook buitenlandse gebruikers er gebruik van kunnen maken of buitenlandse instellingen het kunnen gebruiken in hun omgeving. Bovendien bevinden zich in Nederlandse archieven ook stukken in andere talen (denk aan Franse of Duitse teksten, teksten in (kerk)Latijn en Hebreeuwse teksten). Dit impliceert ook dat de module schrift-onafhankelijk is (Hebreeuws, Cyrillisch, Grieks, Chinees e.d.). Ook in technisch opzicht moet er rekening gehouden worden met "locale instellingen" (toetsenbordindelingen, codepages). De module moet om deze redenen de Unicode-standaard (ISO/IEC 10646 zie <http://www.unicode.org/standard/translations/dutch.html>) ondersteunen in weergave, data-entry en opslag.
3. De module is niet bedoeld als lesmateriaal voor paleografie. Het is echter niet uitgesloten en zelfs waarschijnlijk dat de module zeer goed bruikbaar is in een dergelijke context. Voor goed lesmateriaal paleografie wordt verwezen naar bestaande oplossingen. Zie hiervoor de rapportage van Deelproject 7 "Longlist van bestaande functionele oplossingen", Hoofdstuk Tools, Transcriptie en Paleografie"
4. De module ondersteunt het markeren en labelen van stukken tekst. Zo kan de module ook ingezet worden als hulpmiddel bij het maken van indexen. De markering wordt zowel in de getranscribeerde tekst als (grafisch) bij de scan aangebracht. De scan en de transcriptie worden dan ook als eenheid behandeld.
5. De module ondersteunt het transcriberen van delen van een scan (bijv. doopakte), het transcriberen van een gehele scan (gescande pagina) én het transcriberen van een tekst die (deels) op verschillende - niet noodzakelijk aaneengesloten - pagina's (scans) staat.
6. Gebruikers hebben verschillende voorkeuren. De module probeert gebruikers vrijheid te geven om verschillende voorkeuren te honoreren. Ter illustratie: de gebruiker kan kiezen of de scan naast of boven de tekst komt of eventueel elk in een eigen window of tabblad. Ook het lettertype en de lettergrootte kunnen gebruikersspecifiek ingesteld worden.
7. In de opslag wordt slechts de informatie meegenomen die essentieel is voor de transcriptie. Dit betekent dat lettertype, kleur en lettergrootte niet worden meegenomen.
8. De module is geschikt voor elke soort gescand document. Het ondersteunt zowel gestructureerde tekst als ongestructureerde tekst.
9. Markeringen die kunnen worden aangebracht gaan over (a) de soort akte/bron met citaat-instructie en referentienummers (b) de datering van het opstellen van de tekst zelf en van de beschreven gebeurtenis (c) personen (met onderscheid in onderdelen van persoonsnamen als voornaam, tussenvoegsel, patroniem, geslachtsnaam/achternaam, leeftijd, geboortedatum, overlijdensdatum, geboorteplaats/woonplaats/overlijdensplaats en beroep (d) de locatie (waarop de tekst betrekking heeft cq waar de tekst is gemaakt) (e) opmerkingen over de leesbaarheid als: "doorgehaald", "onleesbaar", "onzeker", "waarschijnlijk", "mogelijk/misschien", "..."
10. De gebruiker kan bepalen welk deel van de scan zichtbaar is: zoomen, pannen, full page / fit window. Beelden in hoge resolutie worden (bij voorkeur ?) tot web-resoluties teruggeschaald om te helpen onrechtmatig gebruik tegen te gaan.

11. Om rechten op het beeldmateriaal te helpen beschermen, kan er een watermerk aan elk beeld toegevoegd worden (aan de serverkant door de eigenaar van de scan). De originele scans mogen hierdoor niet beïnvloed worden.
12. Meta-informatie die in de scan beschikbaar is (bijv. via EXIF-formaat en IPTC-formaat) wordt automatisch herkend.

Technische randvoorwaarden:

1. De module is zoveel mogelijk browser onafhankelijk. Er moet in elk geval ondersteuning zijn voor Internet Explorer, Firefox en Safari. In elk geval moeten de meest recente versie en de voorlaatste versie ervan ondersteund worden. Bij voorkeur moet ook rekening gehouden worden met oudere versies waarvan bekend is dat er nog een significant marktaandeel is.
2. Aan de client-kant moet in elk geval Windows, Linux en Apple ondersteund worden. In elk geval moeten de meest recente versie en de voorlaatste versie ervan ondersteund worden. Bij voorkeur moet ook rekening gehouden worden met oudere versies waarvan bekend is dat er nog een significant marktaandeel is.
3. De module vereist geen specifieke browserinstellingen of plugins. Het gebruik van flash is niet toegestaan (mede i.v.m. Webrichtlijnen voor de overheid). Het gebruik van Java-applets kan belemmerend werken omdat dit specifieke configuratie vraagt aan de client-kant (browser-settings, firewall-settings).
4. De module is zoveel mogelijk platform onafhankelijk. Aan de serverkant moet in elk geval Windows en Linux ondersteund worden. In elk geval moeten de meest recente versie en de voorlaatste versie ervan ondersteund worden. Bij voorkeur moet ook rekening gehouden worden met oudere versies waarvan bekend is dat er nog een significant marktaandeel is.
5. De serverkant verlangt geen dure software van derden. Bij voorkeur wordt gebruik gemaakt van open source componenten. Met name wordt gedacht aan gebruik van PHP/MySQL.
6. De serverkant verlangt geen complex beheer. Met name moet rekening gehouden worden met kleinere instellingen waar vaak geen uitgebreide (web-)technische kennis voorhanden is.
7. De serverkant moet bij voorkeur ook in Saas (software as a service, hosting, asp) aangeboden kunnen worden.
8. In de transfer wordt gebruik gemaakt van XML als middel om de gegevens en tekst gestructureerd te versturen. Dit geldt zowel voor de transfer tijdens de opslag als voor de transfer richting de weergave.
9. De URL kan meegegeven worden tijdens het starten van de transcriptiemodule. Er is dan geen user interface voor bestandselectie.
10. De software komt als open source beschikbaar. Er mogen geen rechten door de bouwer worden voorbehouden. Er mogen geen onderdelen worden gebruikt waarop rechten van derden rusten.

Inspiratie-voorbeelden:

De volgende sites bieden voorbeelden van functionaliteit die hier bedoeld is.

- http://www.cartago.nl/index.php?option=com_content&task=view&id=57&Itemid=83 (oorkonden Drenthe en Groningen)
- <http://www.jacobboerema.nl/Transcript/Freeware.htm> (Transcript desktop software)
- <http://www.scottishhandwriting.com> (Schotse handschriften transcriberen, cursus)
- <http://www.geneaknowhow.net/genea/paleo/kennism-paleo.htm> (cursus paleografie Geneaknowhow.net)