
TOELICHTING
BIJ DE OVERGANG OP UNICODE
IN
CBS 2.1
NOVEMBER 2004

Aan: GEBRUIKERS GGC EN NCC
Van: Daniël van Spanje
cc: Servicedesk OCLC PICA
Datum: 22 november 2004
documentnaam: Toelichting bij de overgang op unicode in CBS 2.1
documentstatus: Extern
versie: 1.0

INHOUDSOPGAVE

1	INLEIDING	2
2	UNICODE	3
2.1	OPSLAG IN DE DATABASE	3
2.2	UNICODE EN LBS3 (OUM)	3
2.3	UNICODE EN OUF	3
2.4	UNICODE EN UITVOER	4
3	INVOER MET WINIBW	4
3.1	INVOER MET WINIBW 2.4.X	4
3.2	INVOER MET WINIBW 3.0	4
4	INDEXEN EN ZOEKEN	5
4.1	ALGEMEEN	5
4.2	DE UMLAUT EN DE TREMA	5
4.3	DE ANGSTROM (Å)	6

1 Inleiding

Deze notitie bevat een korte toelichting bij de overgang op Unicode in GGC en NCC/IBL in CBS-release versie 2.1.

Voor een overzicht van de conversie van de Pica characterset naar de Unicode tekenset en voor de afhandeling van de Unicode tekenset in de index is een aparte tabel beschikbaar (Conversion table Pica characters to Unicode).

Een overzicht van de overige functionaliteit in deze systeemversie is in een apart document te vinden (Overzicht oplevering CBS 2.1 2004).

Het gedeelte over Unicode uit die notitie is ook in deze notitie opgenomen of bijgewerkt.

2 Unicode

2.1 Opslag in de database

Vanaf de oplevering van CBS-versie 2.1 worden alle records in het GGC, dus zowel bibliografische als thesaurus-records in Unicode opgeslagen. OCLC PICA maakt voor de opslag gebruik van UTF-8, de de facto standaard voor Unicode.

Maar aangezien UTF-8 invoer en bewerkingen pas volledig mogelijk zijn met WinIBW 3.0 worden de records bij gebruik in het GGC in een groot aantal situaties teruggeconverteerd naar de Pica-characterset. Deze Pica-characterset kan globaal gelijk gesteld worden aan de ISO-Latin1-characterset. Dit terugconverteren in een bewerkingsscherm gebeurt indien de client-software, zoals bijvoorbeeld WinIBW versie 2.4.x en ouder, niet Unicode-compliant zijn. Er zijn andere situaties, bijvoorbeeld in titelpresentaties waarbij Unicode-characters die niet gepresenteerd kunnen worden als escape-code worden gepresenteerd ("&#x<nr>", waarbij <nr> staat voor de hexadecimale notatie). In weer andere situaties, bijvoorbeeld bij OUM en OUF, zullen Unicode-characters die niet in de Pica-characterset voorkomen, bij uitvoer worden verwijderd.

Zodra een GGC-record na invoer of mutatie in de database wordt opgeslagen, wordt aan het record een systeemcode toegekend met als inhoud "utf8". Zo kan altijd nagegaan worden of een record al een keer geconverteerd en in UTF8 is opgeslagen. De database zal dus niet in één keer worden geconverteerd, maar de komende jaren langzamerhand tijdens en door gebruik geconverteerd worden naar UTF-8.

De nieuwe kmc's zijn:

	Pica3 code (titels)	Pica3 code (ingangen)	Pica+ code
Algemeen nivo	000K	00K	001U \$0
Lokaal nivo	0247	00L	101U \$0
Exemplaar nivo	0248	00M	201U \$0

Deze kenmerkcodes worden in het GGC niet gepresenteerd, maar zijn voor de systeembeheerder wel zichtbaar.

2.2 Unicode en LBS3 (OUM)

De titeldatabase in LBS3 is gebaseerd op de Pica-characterset en niet op Unicode.

De record-uitvoer ten behoeve van het update-mechanisme voor LBS3 (OUM) is gebaseerd op de Pica-characterset. In het OUM zal een Unicode-switch worden aangebracht. Zolang het update—mechanisme niet expliciet vraagt om Unicode-records zullen de records worden uitgevoerd in de Pica-characterset – en zullen de records dus worden teruggeconverteerd. Zodra het OUM niet meer om Pica-characterset vraagt, zullen de records in Unicode worden aangeleverd.

Komen er inmiddels in de records in het GGC Unicode-tekens voor die niet in de Pica-characterset ondersteund worden en wordt door het OUM gevraagd om uitvoer in de Pica-characterset, dan zullen de Unicode-tekens worden verwijderd uit de records, en worden deze dus niet aangeleverd aan het LBS.

2.3 Unicode en OUF

Gebruikers die voor de online export van records uit het GGC gebruik maken van het OUF (Online Update Fetch, de gestandaardiseerde versie van het OUM) bevinden zich in een soortgelijke situatie als de gebruikers die ten behoeve van het LBS gebruik maken van OUM. Ook in OUF wordt de

mogelijkheid ingebouwd om de gewenste characterset van de records nader te bepalen. In de huidige OUF-versie is de uitwisseling van records gebaseerd op de Pica-characterset en verandert de export derhalve niet.

Ook hier geldt dat indien er in de records in het GGC Unicode-tekenen voorkomen die niet in de Pica-characterset ondersteund worden en indien gevraagd wordt om uitvoer in Pica-characterset, dat dan de Unicode-tekenen zullen worden verwijderd uit de records, en dus niet worden aangeleverd.

Over de implementatie van de nieuwe versie van OUF zal met de betreffende gebruikers bilateraal contact worden opgenomen.

2.4 Unicode en uitvoer

De offline uitvoer bestaat in twee varianten: XML-uitvoer en lijstuitvoer. De XML-uitvoer is reeds in Unicode en zal dus niet wijzigen. De lijstuitvoer gebruikt de Pica-characterset en dat zal ook zo blijven. Bij lijstuitvoer zullen Unicode tekens worden teruggeconverteerd naar de Pica-characterset en Unicode tekens die niet in de Pica characterset voorkomen worden weggefilterd.

3 Invoer met WinIBW

3.1 Invoer met WinIBW 2.4.x

Online invoer van tekens in Unicode is pas mogelijk met WinIBW 3.0. Records ingevoerd met WinIBW 2.4.1 (en eerdere versies) zijn gebaseerd op de Pica characterset. Deze records worden nadat ze worden opgestuurd naar de database, geconverteerd naar Unicode. Bij het ophalen van records uit de database met datzelfde WinIBW worden de tekens weer teruggeconverteerd naar de Pica characterset.

Dit mechanisme van heen- en terugconversie is een proces dat wordt ondersteund zolang nog gewerkt wordt met de huidige versie van WinIBW.

Worden tekens ingevoerd die niet voorkomen in de Pica-characterset dan zullen deze binnen WinIBW 2.4.x worden gepresenteerd als escape-sequence (&#lt;nr>).

3.2 Invoer met WinIBW 3.0

Zodra WinIBW 3.0 gereed is kunnen ook andere tekens dan die nu door de Pica characterset ondersteund worden ingevoerd worden. Bijvoorbeeld wiskundige formules of niet-westers schrift zoals chinees, arabisch of hebreuws. Hoe het activeren van een ander schriftsoort gaat, zal in het kader van de introductie van WinIBW 3.0 nog worden toegelicht.

Zodra deze tekens worden ingevoerd kunnen deze gegevens niet goed gepresenteerd worden in oudere WinIBW-versies. Daar worden deze tekens dan als escape-sequence gepresenteerd (&#lt;nr>).

WinIBW 3.0 zal in het najaar van 2004 gereed komen. Dan zal deze versie eerst getest worden voor de invoer van Chinees en Hebreeuws.

WinIBW 3.0 werkt volledig in Unicode en daarin zijn geen voorzieningen ingebouwd voor een conversie naar de Pica characterset, bijvoorbeeld om records te downloaden in de Pica characterset.

M.a.w.:

- WinIBW 2.4.x: altijd Pica characterset + evt. Escape-sequences

- WinIBW 3.0: altijd Unicode

4 Indexen en zoeken

4.1 Algemeen

Met CBS-versie 2.1 kan ook in de indexen en bij het zoeken gebruik gemaakt worden van Unicode. De nieuwe Unicode versie van de zoek- en indexengine (PSI) is daartoe in het CBS ingebouwd. Alle indextabellen zijn opnieuw opgebouwd waarbij uitgegaan is van de bestaande indexdefinities. Er zijn dus geen extra indexen aangemaakt. De oude indexen worden wel op nieuwe Unicode wijze aangemaakt.

Dat de indexen nu in Unicode kunnen worden gemaakt, betekent niet dat dat ook echt gebeurt. Per index (auteursindex, titelindex, titelsleutelindex etc.) kan nl. ingesteld worden of de indexen gebaseerd zijn op Unicode of op de wijze zoals in Pica3 gebruikelijk was. Zo is het bijvoorbeeld nu mogelijk om aparte indexen te maken op "élevé" en "élève" en op "the" en "thé". Het is echter niet per se nodig om dat te doen.

Op dit moment zijn de regels die bij indexeren en zoeken gebruikt kunnen worden voor alle zoek sleutels gelijk.

Bij het maken van de Unicode indexen is zoveel mogelijk aangesloten bij de bestaande praktijk zonder de mogelijke resultaten van Unicode helemaal onzichtbaar te maken.

Wel is besloten om bepaalde bijzondere tekens niet als zodanig te indexeren indien deze gegevens sowieso niet via WinIBW 2.4 in te voeren zijn.

Niet helemaal bevredigend opgelost is de situatie bij de umlaut en de angstrom.

Deze tekens kunnen worden ingevoerd of moeten weggelaten worden. Ze moeten niet uitgeschreven worden tot bijv. ue of aa! Zie hierover hieronder.

"Vreemde" tekens zoals ligaturen, servische d, poolse l, griekse alpha en beta zijn platgeslagen, d.w.z. in hun 'hollandse' vorm in de index opgenomen.

Verder zijn de accenten zoveel mogelijk in de indexen opgenomen zolang zij niet de sortering beïnvloeden (zoals het geval is met de umlaut en de angstrom).

Een volledig overzicht van wat er gebeurt bij het indexeren is te vinden in een apart opgestelde de conversie-tabel.

Bij het zoeken zonder accenten worden in principe alle index-ingangen gevonden. Dus zoeken met "eleve" levert zowel "eleve", als ook "élevé" en ook "élève". Zodra met een accent gezocht wordt, wordt ook letterlijk met die string gezocht. Dus zoeken met "élevé" levert niet "élève".

4.2 De umlaut en de trema

In de officiële versie van Unicode wordt op dit moment geen onderscheid gemaakt tussen de trema en de umlaut.

In de zomer van 2004 is echter op initiatief van Die Deutsche Bibliothek door een werkgroep van een van de Unicode subcommittee's opnieuw gesproken over de umlaut (ISO/IEC/JTC1/SC2/WG2) en daar is afgesproken om het teken dat nu voorgesteld is als umlaut/trema (U + 0308 combining diaeresis) te reserveren voor de umlaut en voor de trema een extra combinatie af te spreken (U + 034F combining grapheme joiner + U + 0308 combining diaeresis). De umlaut is kortom de standaard, voor het trema wordt iets extra's gedaan.

Voor het GGC is besloten vooruitlopend op een te verwachten formele erkenning van het in de werkgroep genomen besluit het hierboven genoemde besluit in het GGC te honoreren. En dus het onderscheid tussen umlaut en trema te handhaven.

Het gevolg daarvan is dat in de database altijd een onderscheid gemaakt kan worden tussen beide tekens.

Dit heeft echter niet als resultaat dat bij het indexeren of het zoeken ook een onderscheid gemaakt kan worden tussen deze tekens!

Bij het zoeken volgen wij de Unicode-standaardimplementatie waarvoor door IBM een aantal tabellen zijn opgesteld, de zg. ICU-library (zie: www-124.ibm.com/icu) . In die tabellen wordt geen onderscheid gemaakt tussen umlaut en trema en is voor OCLC PICA de keuze ofwel de umlaut behandelen als trema of de trema als umlaut. Daarbij is gekozen om de umlaut als trema te behandelen.

Bij het zoeken moet de ü ingevoerd worden als ü of als u, maar niet als ue!

Bij het zoeken moet de ö ingevoerd worden als ö of als o, maar niet als oe!

Bij het zoeken moet de ä ingevoerd worden als ä of als a, maar niet als ae!

En, bij het zoeken moet de ë ingevoerd worden als ë of als e, maar – uiteraard - niet als ee!

Aangezien er echter bij offline invoer ook zgn. platgeslagen umlauten ingelezen worden, dus bijv. records met auteur Mueller, moet bij het zoeken op deze namen beide varianten gebruikt worden om er zeker van te zijn dat alle records gevonden worden!

Voorbeeld: z aut müller of mueller

Dit geldt voor alle zoek sleutels. Zoeken van de titel "Der Wanderer" van Heinrich Böll zal dus gezocht kunnen worden met auteur/titelsleutel: "boll/wand" en niet "boelwand"!

4.3 De angstrom (å)

Voor de angstrom geldt weer een andere oplossing die eigenlijk het gevolg is van de keuze voor de afhandeling van de umlaut.

Algemeen principe is nu dat in het GGC het "accent" of in Unicode-termen, het " combining character", weggelaten kan worden. Je vindt dan alle varianten.

Bij de angstrom werkt die oplossing niet omdat de a-angstrom op een andere plek sorteert dan de letter "a" .

Vanwege het algemene principe heeft OCLC PICA er voor gekozen om de a-angstrom te indexeren als "a". En dus niet als "aa"!

Aangezien er wel woorden in het GGC en NCC voorkomen die als "aa" zijn ingevoerd in plaats van als å moeten beide varianten gebruikt worden bij het zoeken om alle voorkomens te vinden.

Voorbeeld: zoek ttl århus of aarhus