Unicode

Praktische Auswirkungen im CBS

Karen Hachmann
Verbundzentrale des GBV (VZG)

Göttingen, 28. Februar 2007

Themen

1. Was ist Unicode?
2. Auswirkungen auf die Recherche im CBS
3. Die Indexübersicht im CBS
4. Erfassung nicht-lateinischer Zeichensätze in der WinIBW 3.1 und WinIBW 2.4.1
5. Automatische Transliteration

1. Was ist Unicode?

Was ist Unicode?

*"Unicode ist ein **internationaler Standard**, in dem langfristig für **jedes sinntragende Zeichen** bzw. Textelement **aller bekannten Schriftkulturen** und Zeichensysteme **ein digitaler Code** festgelegt wird. Er will das Problem der verschiedenen inkompatiblen Kodierungen in den unterschiedlichen Ländern beseitigen."*

Λ	U+039B	GREEK CAPITAL LETTER LAMDA
Ж	U+0416	CYRILLIC CAPITAL LETTER ZHE
Ü	U+00DC	LATIN CAPITAL LETTER U WITH DIAERESIS

Was ist Unicode?

"Das gemeinnützige Unicode Consortium wurde 1991 gegründet und ist für den Industriestandard Unicode verantwortlich."

"Bislang, in Unicode 5.0, sind 99.089 Codes individuellen Zeichen zugeordnet."

Zitate aus: Wikipedia Deutschland

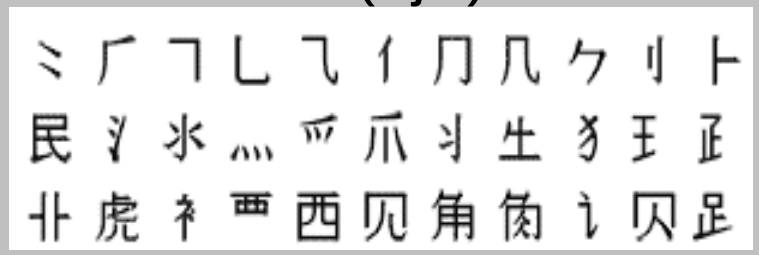
Unicode-Tabellen: http://www.decodeunicode.org/

Beispiele für Unicode-Zeichensätze

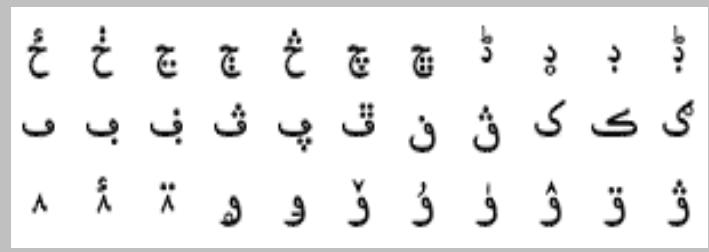
lateinisch

kyrillisch

chinesisch, japanisch koreanisch (cjk)



arabisch



Deutsche Umlaute

Umlaute und Vokale mit Trema werden in Unicode gleich behandelt.

Vergleich ä / ë

a	U+0061	latin small letter a
á	U+00E1	latin small letter a with acute
ä	U+00E4	latin small letter a with diaeresis
e	U+0065	latin small letter e
é	U+00E9	latin small letter e with acute
ë	U+00EB	latin small letter e with diaeresis

2. Auswirkungen auf die Recherche im CBS

Wie soll im GBV mit den Umlauten umgegangen werden?

Vergleich mit anderen Pica-Anwendern:

Die **DNB** indexiert die Umlaute nicht wie in Unicode vorgesehen, sondern löst sie mit "e" auf.

Pica Holland, BSZ:
Die Umlaute werden wie in Unicode vorgesehen behandelt, d.h. eine Recherche mit diakritischem Zeichen oder als Grundbuchstabe ist möglich. Die Umlaute werden nicht mit "e" aufgelöst.

GBV: Indexierung des CBS – 1. Schritt

1. Schritt (Ende November)

Indexierung wie bei der DNB:

Umlaute (d.h. Vokale mit Diäresis) können als Umlaut und aufgelöst mit "e" gesucht werden.

f per schlüter

f per schlueter

Nebenwirkung:

Alle Vokale mit Trema müssen mit "e" aufgelöst werden, auch in den Fällen, bei denen dies sprachlich falsch ist.

Revista de lingüística y lenguas aplicadas

f tit lingueistica

Émile Noël

f per noeel,emile

Das ist falsch!

GBV: Indexierung des CBS – 2. Schritt

2. Schritt (Mitte Dezember)

Umlaute werden nicht mehr mit "e" aufgelöst.

Buchstaben mit Diakritikum können sowohl als Grundbuchstabe als auch mit dem Diakritikum gesucht werden.

Das CBS berücksichtigt, ob in den Suchbegriffen Diakritika verwendet wurden.

f per schlüter	findet Schlüter
f per schluter	findet Schluter und Schlüter
f per schlueter	findet Schlueter

GBV: Indexierung des CBS – 2. Schritt

Nebenwirkung:

Wenn mehrere diakritische Zeichen innerhalb eines **Stichwortes**, einer **Phrase** oder eines **Personennamens** vorkommen, müssen entweder alle Zeichen auf den Grundbuchstaben reduziert oder alle Zeichen mit dem dazugehörigen Diakritikum gesucht werden!

Personenname: Desirée Wüschner

f per wuschner,desiree Treffer

f per wüschner,désirée Treffer

f per wüschner,desiree 00

Das ist zu schwierig!

GBV: Indexierung des CBS – 3. Schritt

3. Schritt (kurz vor Weihnachten)

Die Indexierung bleibt wie sie ist.

Das CBS ignoriert ab sofort die in den Suchbegriffen verwendeten Diakritika.

f per schlüter findet Schluter und Schlüter
f per schluter findet Schluter und Schlüter

f per wuschner,desiree
f per wüschner,désirée
f per wüschner,desiree

} Treffer

Das ist Unicode.

3. Indexübersicht

Die Indexübersicht im CBS

WinIBW 2.000 - Indexübersicht]

2	PER	müller,albert basilius
2	PER	müller,albert burckhard
2	PER	müller,albert c
2	PER	müller,albert franz
4	PER	müller,albert gerhard
4	PER	müller,albert k
1	PER	müller,albert karl
1	PER	mller,jens
1	PER	mller holst,e
1	PER	mller pedersen,poul

WinIBW 2.4.1: Sprünge in der Indexübersicht ab der 20. Zeile

WinIBW 3.1 - Indexübersicht]

2	PER	müller,albert basilius
2	PER	müller,albert burckhard
2	PER	müller,albert c
2	PER	müller,albert franz
4	PER	müller,albert gerhard
4	PER	müller,albert k
1	PER	müller,albert karl
1	PER	müller,albert langen
1	PER	müller,albert lucien
1	PER	müller,albert oskar

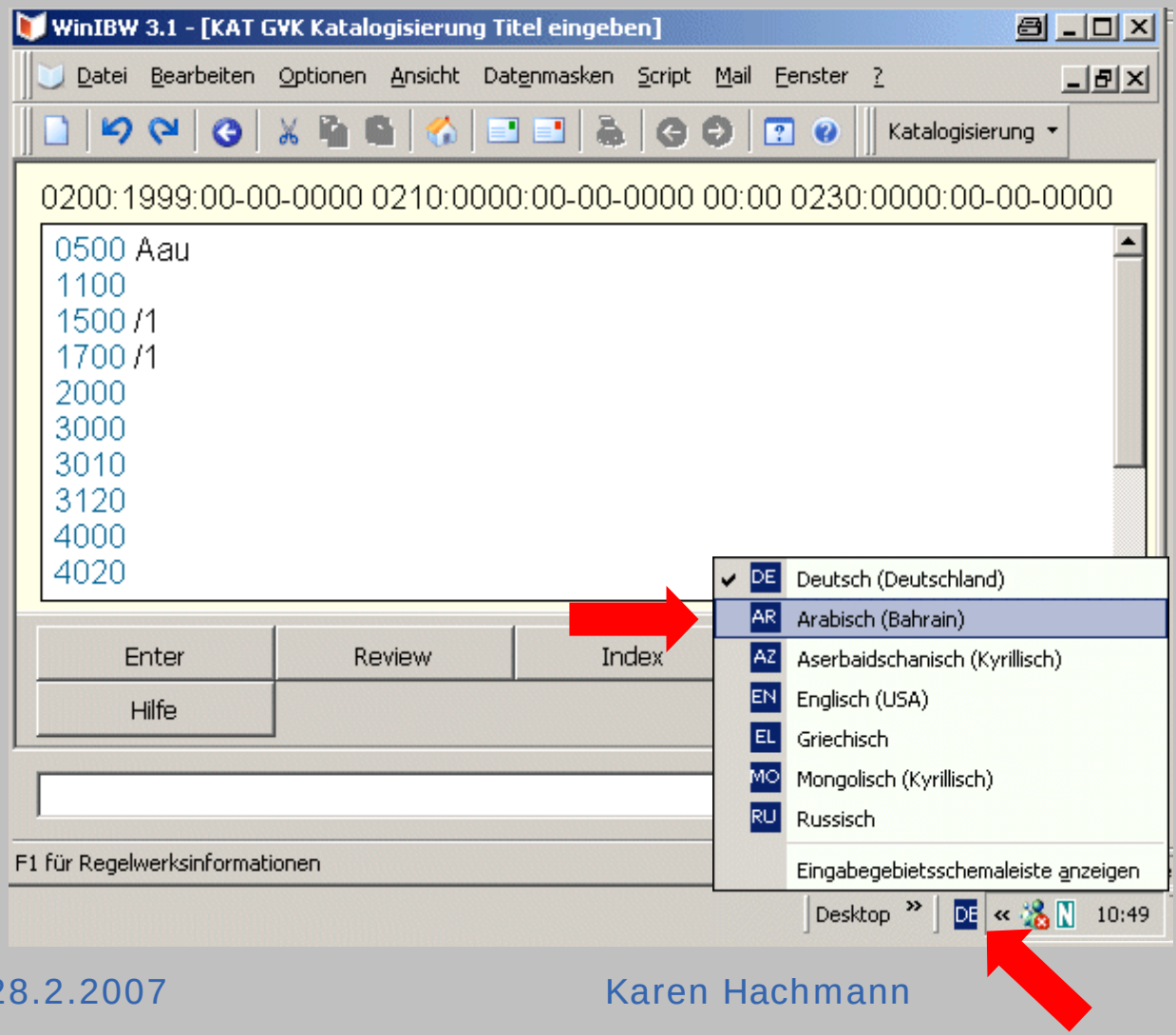
WinIBW 3.1: Korrekt dargestellte Indexübersicht

4. Erfassung nicht-lateinischer Zeichensätze

WinIBW 3.1 und WinIBW 2.4.1

WinIBW 3.1:
Eingabe von nicht-lateinischen Zeichensätzen

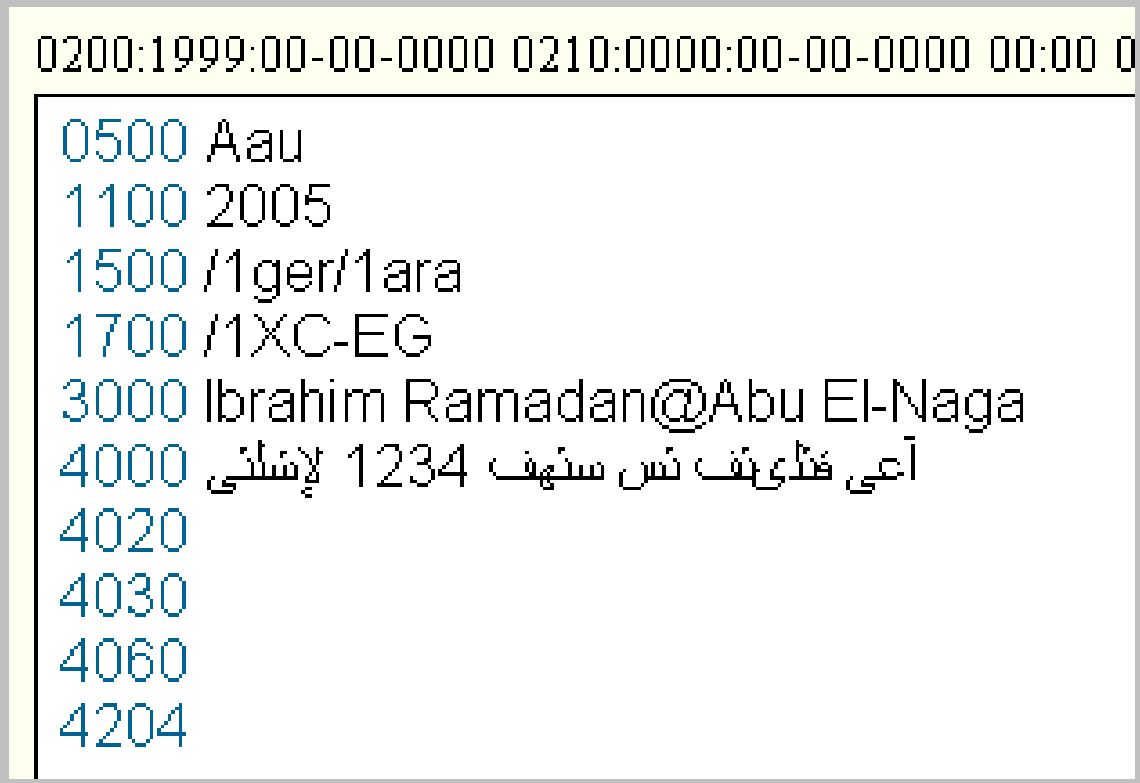
Auswahl des Sprachcodes in der Eingabegebietsschemaleiste



WinIBW 3.1:
Eingabe von nicht-lateinischen Zeichensätzen

Erfassen von arabischem Text.

Beim Schreiben von Buchstaben wird automatisch von rechts nach links geschrieben, bei Zahlen von links nach rechts.



WinIBW 2.4.1: Nicht-lateinische Zeichensätze

Die WinIBW 2.4.1. kann nur den lateinischen Zeichensatz abbilden. Für die nicht darstellbaren Zeichen verwendet sie sogenannte Escape-Sequenzen.

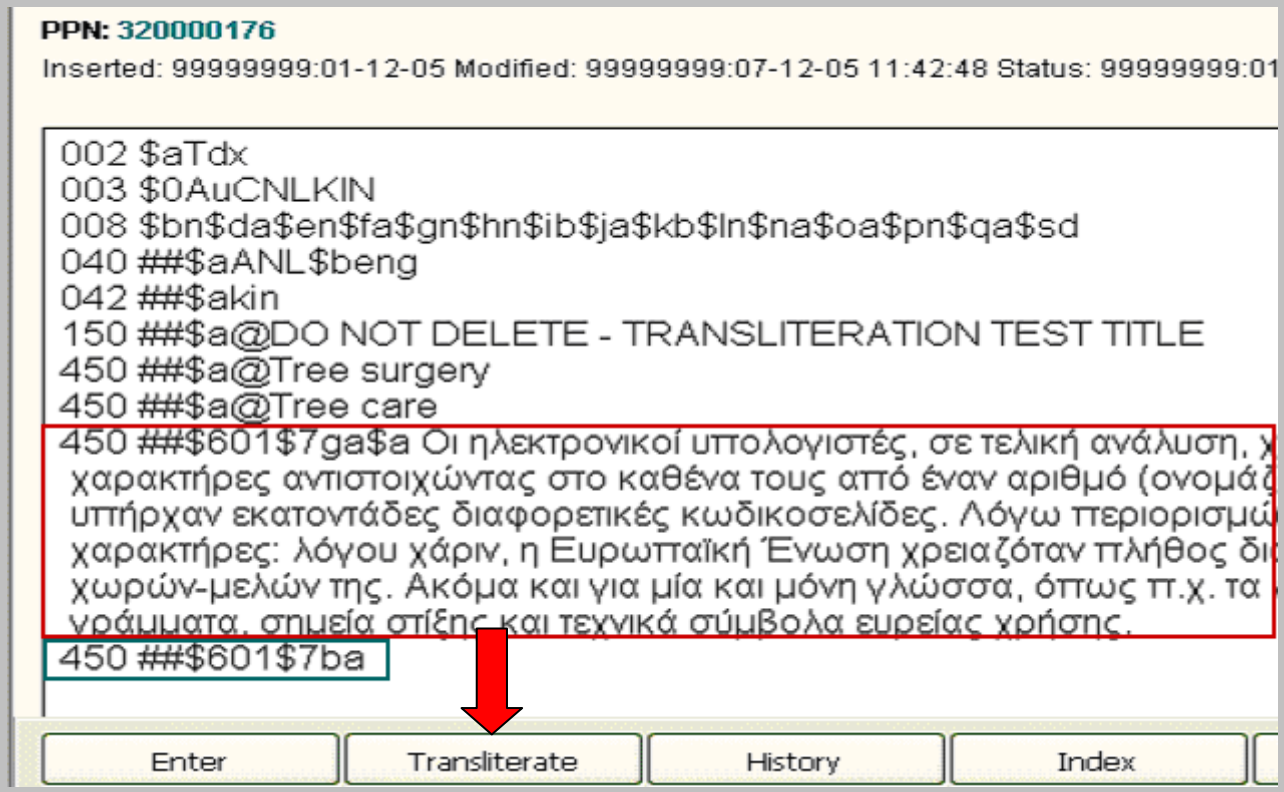
0500 Aau
1100 2005
1500 /1ger/1ara
1700 /1XC-EG
3000 Ibrahim Ramadan@Abu El-Naga
4000 ٓاعى قثل&#x0
سثهف 1234 لٕا&#x0
4020 Als Ms. gedr.
4030 Kairo
4060 II, 116, V Bl.
4204 Kairo, Al-Azhar Univ., Sprachen- und Übersetzungsfak., Magisterarb., 2005
4217 Erscheinungsjahr 1426h

5. Automatische Transliteration

Ein Blick in die Zukunft

Transliteration Schritt 1:

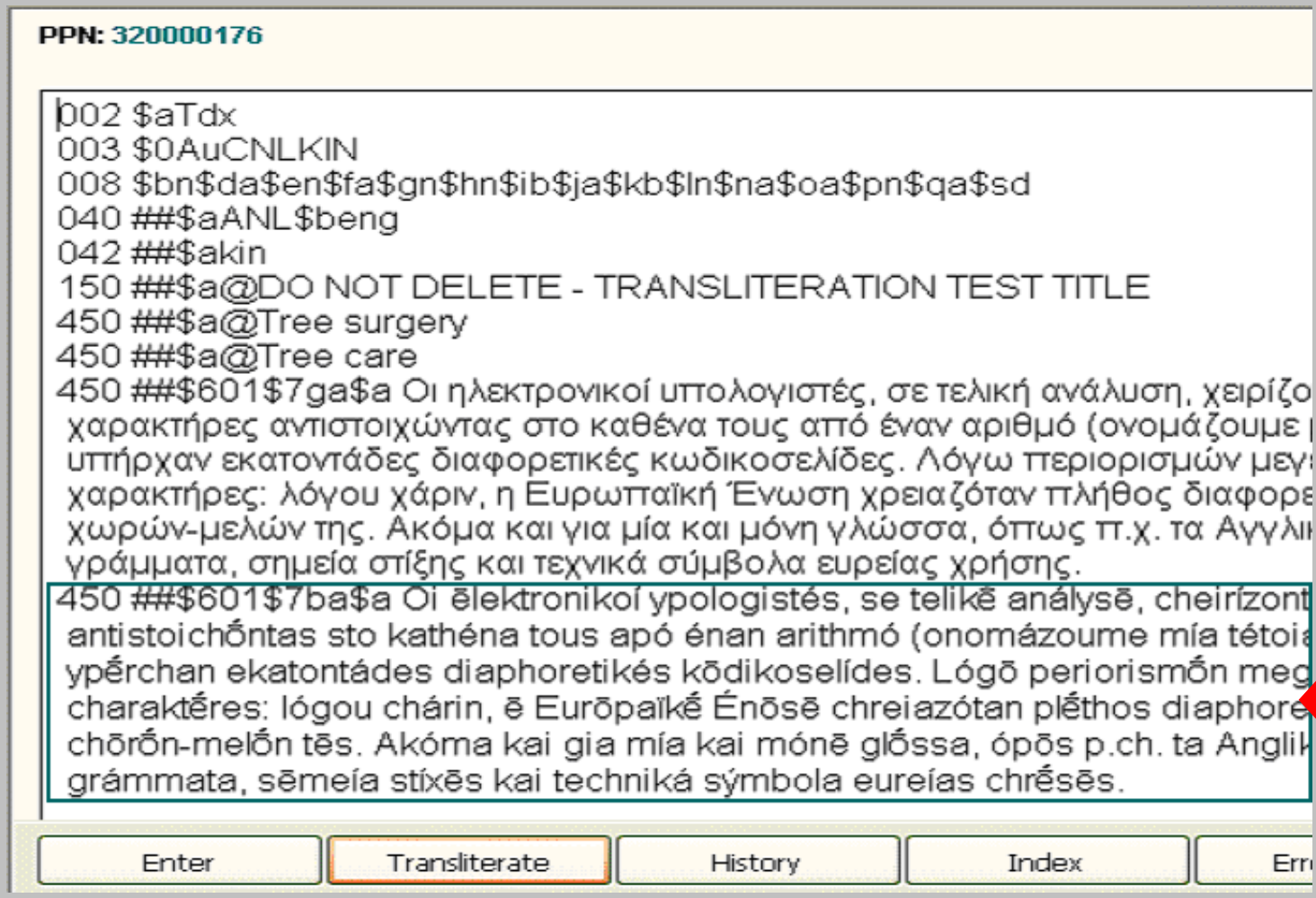
- Kategorie mit originalschriftlichen Text
- Kategorie wird wiederholt
- Sprachcode im Unterfeld (hier in $601$7ba)
- Klick auf Schaltfläche "Transliteration"



(Abbildungen aus der ABES-Datenbank UNM-Format)

Transliteration Schritt 2:

Die Transliterationstabelle im CBS wird gelesen und gibt den Inhalt der Kategorie transliteriert zurück.



Transliteration Schritt 3:

Vollanzeige des Datensatzes nach dem Speichern

cz a22 n 4500
001 000032000017
003 AuCNLKIN
005 20051207115027.0
008 051201 n anannbabn a ana d
040 $aANL$beng
042 $akin
049 $aSH
150 $a@DO NOT DELETE - TRANSLITERATION TEST TITLE
450 $a@Tree surgery
450 $a@Tree care
450 $aOi ēlektronikoí ypologistés, se telikḗ análysē, cheirízontai aplṓs arithmoús. Apothḗ
énan arithmó (onomázoume mía tétoia antistoichía kōdikoselída). Prin tēn epheúresē tou
megéthous ómōs, se kamía kōdikoselída den chōroúsan arketoí charaktḗres: lógou chár
kalýpsei óles tis glṓsses tōn chōrṓn-melṓn tēs. Akóma kai gia mía kai mónē glṓssa, ópō
sēmeía stíxēs kai techniká sýmbola eureías chrḗsēs.
450 $aΟι ηλεκτρονικοί υπολογιστές, σε τελική ανάλυση, χειρίζονται απλώς αριθμούς. Αποθη
αριθμό (ονομάζουμε μία τέτοια αντιστοιχία κωδικοσελίδα). Πριν την εφεύρεση του Unicode,
όμως, σε καμία κωδικοσελίδα δεν χωρούσαν αρκετοί χαρακτήρες: λόγου χάριν, η Ευρωπαϊκ
γλώσσες των χωρών-μελών της. Ακόμα και για μία και μόνη γλώσσα, όπως π.χ. τα Αγγλικά, μί
τεχνικά σύμβολα ευρείας χρήσης.

ISBD-Anzeige kann gedoppelt werden

Notice en écriture originale

Οἱ κῆποι τοῦ διαβόλου [Texte imprimé] : ἀστυνομικά διηγήματα / Αθήνα Κακουρή. - Αθήνα : Βιβλιοπώλειον της Ε
πεζογραφία ; 84) (Sýgkronī ellīnikī pezografia ; 84). - Texte en grec moderne polytonique
ISBN 960-05-0987-5

Collection : Σύγκρονη ελληνική πεζογραφία ; 84

Κακουρή, Αθήνα (1928-....). Auteur

Nouvelles grecques modernes -- 21e siècle

Notice en caractères latins

Ohi kĺpoi toy diabóloy [Texte imprimé] : astynomiká dīgĺmata / Athīna Kakourĺ. - Athĺna : Vivliopōleion tīs Estias,
polytonique
ISBN 960-05-0987-5

Collection : Sýgkronī ellīnikī pezografia ; 84

Kakouri, Athéna (1928-....). Auteur

Nouvelles grecques modernes -- 21e siècle

Auswirkung im LBS

In den LBS-Geschäftsgangmodulen (OWC, ACQ und OUS) sollen die gedoppelten Kategorien, die nicht-lateinische Zeichen enthalten, herausgefiltert werden. Sie können im LBS3 nicht dargestellt werden.

Im OPC4 können die nicht-lateinischen Zeichensätze korrekt abgebildet werden.

Erstes Projekt mit nicht-lateinischen Zeichen

Hebraica der Herzog August Bibliothek Wolfenbüttel

Als Pilotprojekt soll eine eigene Datenbank für den Hebraica-Bestand der HAB Wolfenbüttel aufgebaut werden.

Die Titel liegen bereits sowohl in Transliteration und originalsprachlich in hebräischen Lettern vor.

Die Titelaufnahmen wurden wegen der dort verfügbaren Hebraica-Bestände und der erforderlichen Hilfsmittel in Zusammenarbeit mit der Bodleian Library (Oxford) durchgeführt.