

Zur Ordnung und Codierung der Umlautbuchstaben

Bernhard Eversberg

Der folgende Beitrag, von Bernhard Eversberg im Rahmen einer zeitweiligen Arbeitsgruppe der Konferenz für Regelwerksfragen verfaßt, wurde von der Konferenz auf ihrer 5. Sitzung Ende März zustimmend zur Kenntnis genommen. Der Arbeitsgruppe haben außer Bernhard Eversberg Cornelia Katz (BSZ), Dr. Volker Henze (DDB) und Dr. Klaus Haller (BSB) angehört.

Zusammenfassung

Die Umlautbuchstaben stellen eine Eigentümlichkeit der deutschen Schriftsprache dar, zu der es in anderen Sprachen keine Parallele gibt, und zwar ist das Besondere daran die Gleichwertigkeit eines Sonderbuchstabens mit einer Buchstabenkombination. Daraus ergibt sich das Problem der Ordnung oder Indexierung in Katalogen und Verzeichnissen, vor allem Namensverzeichnissen, weil Namen in beiden möglichen Schreibweisen auftreten und auch so erfaßt werden. Hinzu kommt, daß es in den Katalogdatenbanken in großem Umfang Altdaten mit aufgelöst erfaßten Umlauten gibt, da man in früheren Zeiten mit eingeschränkten Zeichensätzen auskommen mußte. Auch die Norm DIN 5007 „Ordnen von Schriftzeichenfolgen“ (1962) trägt diesem Problem Rechnung, indem sie für Namensverzeichnisse die Gleichung ä = ae vorschreibt. Lexika und Wörterbücher dagegen sind nicht primär Namensverzeichnisse, und sie wenden aus guten Gründen die Gleichung ä = a an, die auch überall im Ausland beim Ordnen deutscher Namen gilt. Für den Zugriff zu Online-Bibliothekskatalogen bedeutet dies, daß man in Deutschland nach „mueller“, im Ausland in aller Regel nach „muller“ suchen muß. Zunehmend wird heute grenzüberschreitend gesucht, per WWW und Z39.50, und aus den unvermeidlichen Irritationen ausländischer Nutzer mit deutschen Datenbanken und umgekehrt erwächst der Eindruck, man sollte oder müsse die deutsche Sonderregel aufgeben.

Die vier grundsätzlich möglichen Lösungen des Ordnungsproblems werden diskutiert. Es bestätigt sich, daß die heute praktizierte dem Material und den OPAC-Erwartungen am besten gerecht wird. Zur Verbesserung des grenzüberschreitenden Zugriffs bietet sich nur, als fünfte und neue Lösung, die Doppelindexierung an: Jeder „Müller“ wird sowohl als „muller“ wie auch als „mueller“ indexiert. Die gewohnten Zugriffsqualitäten bleiben dann erhalten. Vollständig wird das Problem auch dadurch nicht gelöst: Zugriffe mit „muller“

finden noch immer die (echten und falschen) „Muellers“ nicht, von denen es aber in manchen Altdatenbeständen sehr viele gibt. Entsprechendes gilt für Titelwörter. Mehr ist jedoch durch kein Verfahren erreichbar, es sei denn Dreifachindexierung (jedes „ue“ auch noch als „u“), aber dadurch würden wohl zuviele unsinnige Einträge entstehen.

Für deutsche OPACs ist dringend zu empfehlen, im Prinzip bei der bisherigen Regelung zu bleiben, u.a. weil eine Änderung wohl nicht flächendeckend durchzusetzen ist, aber auch weil sie die einzige ist, bei der jeder Name zuverlässig mit einem Zugriff gefunden wird, gleich wie er erfaßt wurde. Wo es durchführbar ist, sollte man aber die Doppelindexierung erwägen.

Kein Handlungsbedarf besteht in der Frage der Codierung der Umlaute. Bisher wird, hier wie im Ausland, das Zeichen ü einheitlich erfaßt, unabhängig von der Sprache. In Deutschland gibt es dafür einen eigenen Code, in den MARC-Daten wird die Kombination Trema+u erfaßt. Dies ist eindeutig austauschbar. Würde man im Zuge der UNICODE-Einführung dies ändern, also für finnische, ungarische, türkische und sonstige ü-Zeichen eine andere Codierung einführen als für das deutsche ü, dann würde damit eine Inkonsistenz geschaffen, weil man die Altdaten nicht umarbeiten kann. Wenn man die Doppelindexierung einführt, wird das Suchproblem der ausländischen Namen entschärft, so daß auch von daher gesehen eine differenzierende Codierung keinen Nutzeffekt hätte.

Ein spezifisch deutsches Problem

Die deutschen Umlaute ä, ö und ü sind als Laute jünger als die Grundvokale: sie entstanden in den alt- und mittelhochdeutschen Perioden der Sprachgeschichte. Nachsilben mit dem hellen 'i' wandelten langsam die drei dunklen Laute um - daher der Name, und daher noch heute die Vokalvarianten bei Wörtern mit gleichem Stamm (z.B. Wort - Wörter). Erst spät wurden die umgelauteten Vokale auch in der Schrift sichtbar gemacht, und zwar bildeten sich mehrere unterschiedliche Praktiken für ihre Notierung heraus. Noch in Drucken des 19. Jahrhunderts kann man z.B. Umlaute durch ein kleines, über dem Buchstaben schwebendes 'e' dargestellt sehen. Verbreitet, erhalten und verfestigt haben sich endlich zwei Konventionen: die Pünktchen über dem Grundbuchstaben bzw. das ihm folgende 'e'. Die Gleichwertigkeit dieser zwei verschiedenen Schreibweisen stellt kein Problem für das Lesen und Verstehen dar, jedoch für das Ordnen von Wörtern, insbesondere aber Namen.

Ein analoges Problem gibt es in anderen Sprachen nicht. Diakritische Zeichen treten zwar in vielen Sprachen auf, doch das Spezifikum des Deutschen ist eben die *orthographische Gleichwertigkeit* des modifizierten Buchstabens mit einer *Kombination* aus zwei Buchstaben.

Für die Datenverarbeitung kommt noch das Problem der Codierung hinzu. Die genau gleich aussehenden Sonderbuchstaben der nordischen Sprachen, des Ungarischen, Türkischen, Slowakischen können in diesen Sprachen nicht durch Kombinationen ersetzt werden. In weiteren Sprachen, z.B. Französisch und Spanisch, haben die Pünktchen nur die Bedeutung, die getrennte Aussprache aufeinanderfolgender Vokale zu kennzeichnen. Die Frage ist zu diskutieren, ob diese Unterschiede bei der Datencodierung berücksichtigt werden sollten, und ob für den Datenaustausch daraus Probleme erwachsen, insbesondere im Hinblick auf UNICODE. Sind unsere Daten jetzt und in Zukunft kompatibel, wie sehen die im Ausland produzierten Daten aus, was erwartet man dort von uns, wie sind die Kommunikation per Z39.50 und der Zugriff über WWW davon betroffen?

Zuerst werden die Ordnungsfragen behandelt.

Vier mögliche Ordnungsmethoden

Ohne die besagte Besonderheit, d.h. wenn die Gleichung $\ddot{a} = ae$ nicht existierte, gäbe es nur zwei verschiedene Ordnungsmethoden:

1. *Gleichordnung*: das Ignorieren des Diakritikums bzw.
2. *Separierung*: die Behandlung des modifizierten Buchstabens als eines selbständigen, von allen anderen unabhängigen Zeichens mit eigener Position im Alphabet.

In Wörterbüchern anderer Sprachen sind beide Lösungen zu beobachten, manchmal durchaus für den einen Buchstaben diese, für den anderen jene.

Da wir nun aber die besagte Gleichwertigkeit haben, gibt es für das alphabetische Ordnen deutscher Umlautbuchstaben theoretisch vier verschiedene Methoden:

1. *Konsequente Gleichordnung* mit dem Grundbuchstaben: $\ddot{u} = ue = u$
2. *Ignorieren der Pünktchen* (und damit auch der Gleichwertigkeit): $\ddot{u} = u$
3. *Auflösung des Umlauts*: $\ddot{u} = ue$
4. *Separierung*, d.h. Ordnung als selbständiger Buchstabe: also z. B. $u < \ddot{u} < v$ (wie im Ungarischen) oder gar am Ende des Alphabets wie in den nordischen Sprachen.

Die Unterschiede werden an einem Beispiel sofort klar. Dazu sollen sechs real existierende Namen dienen, und zwar Gulich, A. Güldner, B. Gueldner, Guenther, Gueldenstern und Guerrero. Man erhält dann vier recht unterschiedliche Reihenfolgen, und der einzige Name mit Umlaut landet an vier verschiedenen Stellen:

| 1: GLEICH- ORDNUNG | 2: PUNKTE IGNORIEREN | 3: AUF- LÖSEN | 4: SEPARIE- RUNG |
|--------------------------|-------------------------|------------------|---------------------|
| Guerrero [Ausnahme!!] | Gueldenstern | Gueldenstern | Gueldenstern |
| Gueldenstern | Gueldner, B | Güldner, A | Gueldner, B |
| Güldner, A | Guenther | Gueldner, B | Guenther |
| Gueldner, B | Guerrero | Guenther | Guerrero |
| Gulich | Güldner, A | Guerrero | Gulich |
| Guenther | Gulich | Gulich | Güldner, A |

Ziele des alphabetischen Ordners

Vor einer Diskussion und Bewertung dieser Methoden lohnt es sich, Sinn und Zweck des alphabetischen Ordners kurz zu beleuchten. Nicht nur *ein* Ziel wird damit verfolgt, sondern *zwei verschiedene*:

A. Wiederauffinden ermöglichen (präziser, schneller Zugriff)

B. Zusammengehöriges zusammenführen (Ergebnismengenbildung, browsing)

wobei B nicht immer gleich als Problem bewußt, aber kaum weniger wichtig als A ist, und das zumal in Bibliothekskatalogen wie auch in Nachschlagewerken. Und es ist der Punkt B, der die Entscheidung zwischen den vier Methoden schwierig macht. Jede einzelne ermöglicht ein schnelles und sicheres Auffinden, wenn

1. sie konsequent angewendet wird,
2. das Prinzip dem Suchenden bekannt ist, und
3. dieser exakt weiß, was er sucht.

Diese drei Dinge vorausgesetzt, sind alle vier Methoden für das *Finden* einzelner Datensätze völlig gleichwertig (vom Problem der Ausnahmen bei Methode 1 einmal abgesehen), für das *Zusammenführen* jedoch keineswegs. Das wird klar, wenn man Aufgaben und Erwartungen genauer betrachtet:

Benutzererwartungen

Die (vermuteten oder tatsächlichen) Erwartungen der Benutzer liegen bei Lexika etwas anders als bei Katalogen und Telefonbüchern:

- *Lexika* sollen Begriffe unter ihrem Wortstamm zusammenführen. Weil aber oft der ordnungswichtige oder einzige Vokal des Wortstamms umgelautet wird (Kram - Krämer, Forst - Förster, Kunst - Künstler), bieten sich nur die Methoden 1 und 2 an. Methode 2 funktioniert nur deshalb zufriedenstellend,

weil die Rechtschreibregeln sicherstellen, daß mit ö geschrieben wird, was wie ö ausgesprochen wird: Foerster gibt es nicht. Probleme, und folglich Ausnahmen, bilden aber Fremdwörter und Eigennamen – denn da *gibt* es Foerster.

- *Bücherkataloge* und *Telefonbücher*, die letzteren noch mehr als die ersten, haben vorrangig Namen auffindbar zu machen und zusammenzuführen. Namen unterliegen aber keiner Rechtschreibregelung, und so gibt es Müllers, die sich Mueller schreiben (und darauf Wert legen), Oehlmann existiert neben Öhlmann und Bär neben Baer. (Auch Ölmann, Oelmann, Bähr und Baehr gibt es, aber das ist noch ein anderes Thema.) Methode 2 verteilt die gleichlautenden Namen auf jeweils zwei Stellen. Zum Problem wird dies, weil der Suchende die korrekte Schreibweise oft nicht kennt und es eben keine Regel dafür gibt, anders als für Sachbegriffe. Deshalb sind die Methoden 1 und 3 der zweiten überlegen, denn sie führen all das zusammen, was sich nur in der Umlautschreibweise unterscheidet. Nur Methode 3 hält dagegen bei der *Stichwortsuche* auseinander, was man nicht zusammen finden will: Bär und Bar, Sage und Säge, Saugen und Säugen, Fordern und Fördern, Lösung und Losung, Mull und Müll, Kür und Kur usw.

Folglich ist rein theoretisch Methode 1 zumindest für Namen die einzig Wahre, weil sie allen Anforderungen entgegenkommt. Nur leider ist sie wegen der notwendigen Ausnahmen nicht automatisierbar (siehe unten), scheidet also in dieser unserer Zeit aus. Methode 3 kommt ohne Ausnahmeregeln aus, weshalb sie sich im Bibliothekswesen durchgesetzt hat.

Doch betrachten wir zunächst die Methoden der Reihe nach noch etwas genauer:

Method 1 : Konsequente Gleichordnung

Sie hat die längste Tradition, wird aber heute wohl nirgends mehr benutzt. Die großen Bibliographien des 19. Jahrhunderts, Kayser und Heinsius, ordneten so (außerdem galt $i=j$ und $ij=y$), aber auch die Deutsche Nationalbibliographie bis 1965 (!), d.h. erst die Einführung der EDV beendete diese Praxis zugunsten der Methode 3. (Das aus den alten Bibliographien zusammengeklebte „GV-alt“ übernahm Methode 1.)

Die großen Enzyklopädien, Brockhaus und Meyer, verwendeten ebenfalls Methode 1, bis man in den späten 1970er Jahren, vermutlich ebenfalls bedingt durch die Datenverarbeitung, zu der weniger rigorosen Methode 2 überging.

Hauptproblem der Methode 1 ist natürlich, daß zwar immer ä durch ae ersetzbar ist, umgekehrt jedoch oftmals nicht. Vorwiegend betrifft dies fremd-

sprachige oder mehrteilige Wörter (Aero..., Mensaeßen, Koeffizient, Hydroenergie, Duell, Guerilla), aber vor allem auch Namen (AEG, Israel, Coelho, Roermond, Bluebird, Buenos Aires, Sueton). Manchmal deuten oe und ue in Namen eine Dehnung an (Coesfeld, Suerbier) und keine Umlautung des Grundbuchstabens. Noch andersartige Ausnahmen treten bei der Gleichung ü = ue = u auf: Wörter und Namen wie Bauer, Heuer, sauer, teuer, Treuenbrietzen – praktisch alles mit einem 'e' hinter 'au' oder 'eu'. Diese Ausnahmen kann keine Software alle zuverlässig erkennen, das konnten nur die menschlichen Redakteure der genannten Werke. Für Leser, die das Prinzip nicht kennen, sind in jenen Lexika (aber auch in den o.g. Bibliographien!) Caesar und Goethe die bekanntesten Problemfälle, weil sie keinen Ausnahmestatus erhielten, sondern immer hinter „Casanova“ bzw. „Gotha“ versteckt wurden. Durch den Übergang zur Methode 2 in den neueren Ausgaben sind Dichter und Diktator jetzt da zu finden, wo sie jeder sucht, und das Ausnahmeproblem entfällt. Nicht nur für Ausländer dürften die deutschen Lexika und Kataloge in der Regel heute doch etwas leichter zu benutzen sein.

Methode 2 : Ignorieren der Pünktchen

Dies ist die international übliche Verfahrensweise für Lexika, Telefonbücher und Kataloge, vor allem in der gesamten englisch oder französisch sprechenden Welt, und für Telefonbücher (aber noch nicht Kataloge) übrigens auch in der mehrsprachigen Schweiz. (Österreich: siehe Methode 4)

Rein formal und somit automatisierbar, ist ihr Hauptnachteil die Zerstreuung gleichlautender, aber verschieden geschriebener deutscher Namen. Für Lexika ein quantitativ eher geringes Problem, für Kataloge und Telefonbücher aber ein großes. Deshalb hat DIN 5007 (1962) ausdrücklich für *Namensverzeichnisse* die Regelung nach Methode 3 vorgesehen (Abschnitt 5.1.1.3 der Norm).

Für *Telefonbücher* käme bei Methode 2, würde sie angewendet, erschwerend hinzu: Die Verständigung erfolgt oftmals nur akustisch, wobei zwischen 'ae' und 'ä' kein Unterschied zu bemerken ist. Man müßte dann jedesmal die Schreibweise mit angeben oder aber in vielen Fällen an zwei Stellen nachsehen. Bei der DeTeMedien (zuständig für die Telefonbücher) sieht man einer solchen Änderung, wie durch Anfrage in Erfahrung gebracht wurde, nicht enthusiastisch entgegen, denn jede noch so kleine Änderung löst Proteste aus, und dies wäre wahrlich keine geringfügige.

Für *Bibliothekskataloge* in der heutigen Ausprägung als Datenbanken kommt erschwerend noch etwas ganz anderes hinzu, und zwar das Schlimmste, was es für Datenbanken gibt: Inkonsistenz. Es gibt in großem Umfang Altdaten aus Zeiten, wo die Systeme noch nicht mit Umlauten hantieren konnten. Man

erfaßte folglich jeden Müller als Mueller. (Hätte man das konsequent so weiter gemacht, bräuchte man jetzt nicht über Veränderungen nachzudenken – es wären gar keine möglich.) In nicht so großem Umfang, aber auch nicht vernachlässigbar, gibt es eine quasi „natürliche“ Inkonsistenz: die Schreibweise auf den Titelblättern kann für einen Verfasser mal so und mal anders sein, für Titelwörter ebenfalls. Und Kataloge befolgen (immer rigorosener) das Prinzip der vorlagegetreuen Erfassung, Lexika erfassen ihr Wortgut *norm*getreu nach den Rechtschreibregeln.

Anmerkung zur DIN 5007: Eine Nachfrage bei der zuständigen Abteilung im DIN ergab, daß dort kein Plan besteht, die Norm zu ändern, d.h. die Sonderregel für Namensverzeichnisse abzuschaffen. Dies wird nur dann geprüft und ggf. durchgeführt werden, wenn ein Antrag dazu eingeht. Nach Lage der Dinge kann ein solcher Antrag nur aus dem Bibliothekswesen oder von der DeTeMedien kommen, denn andere größere Anwendungsbereiche gibt es nicht. Da nun die Neigung zu einer Änderung bei der DeTeMedien erkennbar gering ist, wird nichts passieren, solange auch der Bibliotheksbereich das Thema auf sich beruhen läßt.

Methode 3 : Auflösung

Sie ist die heute gängige Praxis der Bibliotheken (begründet durch die lange Zeit maßgeblichen „Preußischen Instruktionen“), aber auch die der Telefonbücher. Die Deutsche Bibliographie (Frankfurt) hat von Anfang an ab 1945 so sortiert. Leipzig zog erst, wie gesagt, 1966 nach. Die großen (zumeist inzwischen eingestellten) Zettelkataloge der Bibliotheken haben gleichfalls so gearbeitet, desgleichen das GV-neu. Der GK gibt sogar die Vorlageform der Namen gar nicht an, sondern nur die aufgelöste!

Speziell für unsere Katalogdaten ist entscheidend: Methode 3 beseitigt zwar nicht die künstliche und natürliche Inkonsistenz der Daten, neutralisiert aber ihren nachteiligen Aspekt: die Auswirkung auf das Auffinden und Zusammenführen. In Online-Katalogen gewährleistet die Zusammenführung der unterschiedlich geschriebenen Namen etwas sehr Wichtiges: bei dem äußerst häufigen Fragetyp Name+Stichwort kann man nur so sicher sein, routinemäßig alles zu „erwischen“. Sonst müßte man oft eine Frage so formulieren: find (mueller or muller) and kernenergie.

Als Schwachpunkt muß man die im Sinne jener Sprachen falsche Einordnung der türkischen, ungarischen und finnischen Wörter und Namen ansehen. „Oezal“ oder „Uecuencue“ wird ein Türke so nicht suchen, vielleicht noch nicht einmal erkennen. (Vermutlich würden wir aber z.B. in einem türkischen Katalog einen Müller gar nicht finden!) Zahlenmäßig viel gewichtiger ist indessen

die Zusammenführung von „Krämer“ mit „Kraemer“, „Förster“ mit „Foerster“ und „Müller“ mit „Mueller“.

Fazit: Unter beiden Aspekten, Wiederauffinden und Zusammenführung, ist die Methode 3 wohl die bestgeeignete für einen Katalog mit hohem Anteil deutschsprachiger Veröffentlichungen für ein mehrheitlich deutsches Publikum. Man muß auch bedenken: sie wurde ja in die PI und dann auch RAK eingeführt auf dem Hintergrund von über hundert Jahren Erfahrung mit Methode 1.

Methode 4 : Separierung

Aus der Entstehungsgeschichte der deutschen Umlaute wird verständlich, warum sich diese Methode, in anderen Sprachen durchaus üblich im deutschen Sprachbereich nirgends herausbildete – sie ist geradezu kontra-intuitiv. In Österreich ist sie dennoch (orientiert an Ungarn?) seit ca. 1986 in den Telefonbüchern realisiert. Weil jedoch keinerlei Vorteile gegenüber den anderen Methoden zu erkennen sind, besonders im grenzüberschreitenden Datenverkehr, erübrigt sich wohl die nähere Erörterung.

Exkurs: Ist „Ordnen“ noch ein zeitgemäßer Begriff?

Weil eine Datenbank anders als ein Zettel- oder Listenkatalog keine lineare Anordnung von Einträgen verkörpert, wird gelegentlich gemeint, man brauche im Zusammenhang mit Online-Katalogen vom „Ordnen“ und von „Ordnungsregeln“ gar nicht mehr zu reden. Das „Finden“ steht im Vordergrund, und niemanden interessiert, wie das unter der Oberfläche abläuft; dort kann durchaus eine ganz andersartige, für Menschen gar nicht nachvollziehbare und in Katalogisierungsregeln nicht beschreibbare Ordnung herrschen. Solange ein Befehl wie z.B. „find müller,fritz“ auch „Fritz Mueller“ liefert, muß man über die Ordnung des 'ü' nichts mehr wissen, ganz anders als beim Zettelkatalog.

Ein voreiliger Schluß. Erstens sind schnelle Zugriffe überhaupt nur möglich, wenn es aufbereitete und geeignet geordnete Indexdateien gibt, und ohne durchdachte Regeln können diese nicht entstehen. Gute OPACs haben zweitens nicht nur einen Find-Befehl oder ein Formular zum Eintragen von Suchwünschen, sie zeigen auch die alphabetisch geordneten Register für das Browsing (äußerst hilfreich bei ungenauer Kenntnis der Schreibweise), und sie stellen auf Wunsch sortierte Trefferlisten her. Bei den angeblich so benutzerfreundlichen WWW-Katalogen dominiert das auszufüllende Suchformular, doch steckt dahinter nichts anderes als ein Find-Befehl. Man erhält meistens Ergebnisse, aber man sieht keinen Kontext, man sieht auch nicht, was man nur knapp verpaßte. Anders gesagt: die Präzision der Suche ist nicht ein-

schätzbar. Ein alphabetisch sortiertes Register bietet mehr: man wird z.B. zwanglos auf abweichende Schreibweisen und auf Erfassungsfehler aufmerksam oder entdeckt überhaupt erst beim Browsing das, was man wirklich suchte. Das Suchformular ist das Paradigma des Datenbankprogrammierers, der von der Vorstellung einer konsistenten Datenbank ausgeht. Die sortierte Liste ist dagegen das Paradigma des Bibliothekars, dem die Komplexität und Uneinheitlichkeit des Materials bewußt ist und auch die unvermeidliche Unvollkommenheit der Erfassung. (Eine Telefonbuch-Datenbank, nebenbei bemerkt, muß unbedingt die Namen als alphabetisch geordnete Liste zeigen können, schon wegen der häufig praktizierten Abkürzung der Vornamen, jedoch auch, weil der Vorname dem Suchenden nicht immer bekannt ist – wenn er dann aber die Liste sieht, fällt ihm dieser oft wieder ein.)

Das Reden über Ordnung ist, kurzum, unverändert wichtig und wird es auch bleiben. Man mag zwar bei Datenbanken präziser von „Indexierungsregeln“ sprechen. Diese umfassen mehr, aber im Kern nichts grundsätzlich anderes.

Herausforderungen der Globalisierung

Handlungsbedarf durch UNICODE?

Der UNICODE-Standard ist für die Codierung von Textdaten gedacht. Das System hat einen, aber *nur* einen Code für das Zeichen ü, differenziert also nicht nach Sprachen – die Zeichen sehen ja auf Bildschirm und Papier alle gleich aus. Es gibt aber einen zusätzlichen Code, 0308, für das Trema („combining diaeresis“ genannt, früher „non-spacing diaeresis“), damit man bei Bedarf auch solche Zeichen damit versehen kann, die als Kombination in UNICODE noch keine Aufnahme gefunden haben. Solche kombinierenden Codes gibt es auf der Codeseite 03xx auch für die anderen Diakritika. UNICODE schreibt vor, die kombinierenden Akzentzeichen bei der Eingabe *hinter* die zu akzentuierenden Zeichen zu setzen. Die gängige Gepflogenheit hier und auch im MARC-Bereich ist bislang die umgekehrte (Akzent *vor* dem Zeichen), doch das ist nur ein Konvertierungsproblem.

Es besteht die Möglichkeit einer Differenzierung: 'ü' nur für deutsche Wörter und Namen, 'u' mit nachgestelltem Trema für alle anderen Sprachen. Deutsche Daten könnten dann weiter nach Methode 3, ausländische aber nach Methode 2 indexiert werden. Davon ist aus sechs Gründen abzuraten:

- Es würde eine internationale Absprache und Einigung erfordern. Die Aussichten dafür dürften nicht besonders hoch sein, denn
- Finnland, Ungarn und die Türkei werden es aus analogen Überlegungen heraus umgekehrt machen wollen (schließlich werden sie ihr eigenes ü für

wichtiger halten als das deutsche), und wenn keine Einigung erzielt wird, werden die Daten doch wieder inkompatibel.

- Man muß davon ausgehen, daß bei der Katalogdatenerfassung in Deutschland in der Regel bisher immer derselbe Code für das Zeichen 'ü' verwendet wurde, trotz RAK 803,3 und obwohl es das Trema in den Zeichensätzen meistens schon gab, mit anderen Zeichen beliebig kombinierbar (DIN 31628/2 und ISO 8859 haben es). Wenn wenigstens z.B. in Namens-Normsätzen die Nationalität der Person festgehalten wäre, könnte man über eine nachträgliche Differenzierung nachdenken; das ist jedoch nicht der Fall. Bis jetzt sind unsere Daten in dieser Hinsicht zwar nicht optimal, aber konsistent. Durch Änderung der Erfassung schaffen wir eine Inkonsistenz. Nur neue Daten werden dann korrekter suchbar als bisher, die alten bleiben falsch. Besser ist, sie sind *alle* falsch, aber man weiß wenigstens, wo man sie findet und findet sie dann verläßlich. Ein Katalog muß ein durchgängiges und ausnahmsfreies Ordnungsprinzip haben.
- Für die USMARC-Daten als umfangreichste Fremddatenquelle gilt: sie sind gleichfalls konsistent und mit unseren in dieser Hinsicht kompatibel, nur umgekehrt: es gibt keine Codes und Symbole für die Umlaute, sondern man setzt in *jedem* Fall ein Trema vor die Grundbuchstaben. Hierzulande wurden und werden solche Kombinationen generell in ä, ö, ü verwandelt und damit stimmen die Daten mit unserer Konvention überein. Weil die MARC-Anwender weltweit nach Methode 2 sortieren bzw. indexieren, also schlicht die Diakritika weglassen, werden sie keinen Handlungsbedarf sehen (und Nachfragen bei Experten bestätigten dies) in Zukunft zwischen deutschen und anderen 'ü' zu unterscheiden – es wäre mehr Arbeit ohne Effekt. Man wird also weiterhin, auch mit UNICODE, undifferenzierte MARC-Daten produzieren.
- Außerdem gibt es Rand- und Zweifelsfälle (z.B. eingebürgerte Türken – nach Staatsbürgerprinzip wie Deutsche zu behandeln!?).
- Führt man die Doppelindexierung ein (siehe unten), entfällt auch das Argument, daß ein türkischer Name wie İnönü besser unter „inonu“ auffindbar sein sollte, statt unter „inoenue“.

Handlungsbedarf durch Z39.50?

Wenn ein ausländisches System per Z39.50 in einer deutschen Datenbank nach einem „Müller“ sucht, wird es in aller Regel nach „muller“ fragen und daher nichts oder sehr wenig finden. Das Konzept der Z39.50-Norm beruht nicht nur (wie das des Abfrageformulars) auf der Vorstellung, die Datenbanken seien inhaltlich konsistent, sondern sogar (noch naiver, ist man versucht zu

sagen), verschiedene Datensysteme seien formal und inhaltlich kompatibel. Abfragen nach Identnummern mögen so gut wie unproblematisch sein, aber alles andere hat seine Haken und Ösen. Diese zu entschärfen oder abzubauen ist deshalb eine verdienstvolle Aufgabe, und die Einführung der Methode 2 in unsere Ordnungsregeln wäre ein nicht ganz kleiner Schritt in diese Richtung. Allerdings könnte ein kritischer Benutzer das Resultat auch so sehen, daß man endlich dieselben Probleme mit den deutschen Namen, die man überall in der Welt hat, dann auch bei uns einführt.

Schlußfolgerungen

Wie oben festgestellt und begründet wurde, ist Methode 3 einem OPAC in Deutschland am besten adäquat. Aus mehreren Gründen bleibt wohl keine andere Wahl, als es bei der bisherigen Sortierregel für Umlaute zu belassen:

- Sie führt zusammen, was zusammengehört, und vermeidet dabei störende Auswirkungen der Inkonsistenz in den Datenbeständen, die durch Einführung der Methode 2 zutage träten.
- Nur diese Methode gewährleistet, daß man bei der sehr häufigen und wichtigen Frage nach Name + Stichwort mit hoher Sicherheit auch bei Namen mit Umlaut zum Erfolg kommt.
- Die Methode ist auf allen Ebenen etabliert und bewährt. Es ist kein einziger Online-Katalog bekannt, der anders ordnen würde – eine beeindruckende Einheitlichkeit!
- Die zentralen Datenbanken müssen so weit wie möglich mit den lokalen OPACs harmonisieren – man zwingt sonst zu zweigleisigem Denken oder programmiert Mißerfolgserlebnisse vor.
- Schon aus reinem bibliothekspolitischen Pragmatismus wird in der Regel die Qualität des Kataloges für die zahlenmäßig größte Benutzergruppe die höchste Priorität erhalten. Die Verantwortlichen der lokalen OPACs werden sich kaum geschlossen hinter eine Maßnahme stellen, die so tief eingreift und eine merkliche Qualitätseinbuße zur Folge hat.
- Ein Problem mit der Codierung wird es auch mit Einführung von UNICODE nicht geben, unabhängig von der Ordnung, für die man sich entscheidet. UNICODE hat mit Ordnungsregeln nichts zu tun und kann davon unabhängig eingeführt werden.
- Für Zettel- und Listenkataloge hätte eine Regeländerung wenig Sinn, denn nur für neu anzulegende könnte man sie einführen. Aber selbst dann wäre zu beachten: die existierenden konventionellen Kataloge folgen alle Metho-

de 3, die alten der Methode 1. Soll noch eine dritte hinzukommen? Das ist auch für DNB-Papierausgaben zu bedenken.

Abhilfe: Methode 5

Es bleibt nur das Dilemma mit der grenzüberschreitenden Suche per Z39.50, in den Web-Katalogen und Diensten wie dem KVK. Wir können den Rest der Welt hinsichtlich Umlautordnung nicht auf unser Gleis überleiten, soviel ist sicher. Es gibt aber immerhin softwaretechnische Möglichkeiten, das Hilfreiche zu tun, ohne das Bewährte und für uns Sinnvolle aufzugeben:

- *Doppelindexierung* zur Verbesserung der Suche vom Ausland aus (von dort kommt die Frage nach „muller“, wenn „Müller“ gesucht wird): Zumindest die Verbunddatenbanken und die DDB können über eine zweigleisige Indexierung nachdenken: jedes Wort mit Umlaut erhält zwei Einträge, einen nach Methode 3 wie bisher, einen weiteren nach Methode 2. Die Muellers mit 'ue' (Altdaten!) bleiben dann aber noch immer unauffindbar, wenn „muller“ gesucht wird! Ob man diese zusätzlichen Einträge aus der normalen (lokalen) Suche ausblendet und nur für Z39.50-Zugriffe zugänglich macht, ist eine noch zu prüfende Frage. (Wer den Katalog und seine Regeln kennt, will vermutlich nicht bei der Suche nach einem „Forster“ auch sämtliche „Förster“ geliefert bekommen.) Im Prinzip dürfte aber eine solche Doppelindexierung in allen Online-Datenbanken durchführbar sein, auch in CD-ROM-Katalogen und -Bibliographien. (Von einer Dreifachindexierung, die aus jedem „ue“ auch noch ein „u“ machen würde, ist abzuraten!)
- *Automatische Doppelsuche*: Umgekehrt, zur Unterstützung unserer Benutzer beim Suchen in ausländischen Katalogen per Z39.50-Gateway, ist eine andere Abhilfe denkbar und zu diskutieren: Wenn ein Benutzer sein Suchwort mit ü eintippt, also z.B. „Müller“, dann könnte unter der Oberfläche eine Suche nach „muller or mueller“ ausgelöst werden. Denkbar ist ferner eine Hilfsfunktion, die jeweils einen Hinweis ausgibt, wenn ein Suchbegriff die Buchstabenkombinationen ae, oe oder ue enthält, z.B.: „Sie könnten auch noch 'muller' versuchen“.

Anmerkungen:

Dieser Aufsatz wird mit zwei Anhängen illustriert, die aus Umfangsgründen hier nicht wiedergegeben werden können, jedoch in der elektronischen Version des BIBLIOTHEKSDIENST unter der URL <<http://www.dbi-berlin.de>>, Gruppe „Publikationen“ enthalten sind. *Anhang 1* besteht aus einer Beispielliste mit Namen, die mit Ü oder Ue begonnen, *Anhang 2* stellt die Behandlung von Diakritika in anderen Sprachen dar.

