€uropäische Standardisierung oder der ganz normale Wahnsinn

Wie es begann

Seit nunmehr über einem Jahr verbringt ein Mitarbeiter des ZDV einen nicht unbeachtlichen Teil seiner Zeit damit, sich durch Berge von E-Mails − momentaner Rekord: 110 im Laufe eines einzigen Tages, von denen etwa 40 persönlich zu beantworten waren − durchzuwühlen, seine Meinung zu verschiedensten Dingen abzugeben und auf Konferenzen präsent zu sein. Viele der Fragestellungen, mit denen er sich in diesem Zusammenhang zu beschäftigen hat, wirken nicht nur auf den ersten Blick abstrus und scheinen kaum mit seinen übrigen Dienstpflichten, zu denen die Gestaltung einer graphischen Oberfläche für TUSTEP gehört, vereinbar zu sein.

Ich will nicht leugnen, daß ich auch jetzt noch bei vielen der Probleme und der prozeduralen Selbstverliebtheit, mit denen ich in diesem Zusammenhang zu tun habe, nur verständnislos den Kopf schütteln kann. Trotzdem erscheint sowohl dem ZDV als auch mir persönlich das Engagement sinnvoll zu sein.

Alles fing Anfang letzten Jahres an, als wir planten, einen neuen Filter / Viewer für TUSTEP zu implementieren, eine Aufgabe, die mir zufiel. Um dem Bedarf nach den verschiedenen, von TUSTEP unterstützten Sprachen − neben allen Sprachen lateinischer Schrift auch Griechisch, Kyrillisch (modern und altkirchenslawisch), Hebräisch, Arabisch, Koptisch, Altsyrisch / Estrangelo − gerecht zu werden, habe ich von Anfang an konsequent auf eine Unicode-konforme Implementierung geachtet. Unicode, dessen Pendant in der ISO-Welt als ISO 10646 firmiert, repräsentiert jedes dargestellte Zeichen mit 16 Bit, was über 60.000 Zeichen abdeckt. Durch sog. *Surrogate Pairs* (Kombinationen aus zwei 16-Bit-Zeichen) können sogar noch über eine Million weitere Zeichen kodiert werden. Dies reicht für alle Schriftzeichen, die es auf dieser Welt gibt und je gegeben hat. Normale Zeichensätze, wie sie unter DOS und UNIX üblich sind, benötigen hingegen nur 8 Bit pro Zeichen, haben damit aber auch nur 256 Codepositionen. Das genügt für viele Schriften, insbesondere Silben- und Bilderschriften, nicht annähernd.

Aber auch Unicode, ein typisches Produkt der amerikanischen Großindustrie, hat seine Mängel, die uns besonders bei einer unbefriedigenden Realisierung des Syrischen auffielen. Entsprechende Kritik führte graduell zu einer verstärkten Einbindung in relevante Gremien, die im Oktober 1997 in der Entsendung als offizieller deutscher Vertreter auf eine Konferenz von CEN/TC304 gipfelte.

Das Normierungswesen

Nicht nur für einen Einsteiger sind die verschiedenen Gremien im Normierungswesen nicht immer leicht durchschaubar. Es gibt Organisationen auf nationaler Ebene, so in Deutschland das altbekannte *Deutsche Institut für Normung (DIN)*, das wohl jeder vom DIN A4-Papier her kennt. Für ganz Europa ist das *Comité Européen de Normalisation (CEN)* zuständig, und international hat die *International Standards Organization (ISO)* das Sagen. Standards, die von höheren Gremien verabschiedet worden sind, sind auch für die darunterliegenden Instanzen bindend.

Die Standardisierungsorganisationen selbst sind wieder in Kommissionen unterteilt, die für bestimmte Gebiete zuständig sind. So heißt das Gremium, das sich in Deutschland mit Zeichensätzen u. ä. beschäftigt, z. B. NI-02, sein europäisches Gegenstück nennt sich TC304, wobei TC für *Technical Committee* steht. Auf internationaler Ebene nennt sich das Pendant ISO IEC JTC1/SC2, das *Joint Technical Committee No 1, Subcommittee 2*.

Da Standards beachtliche Auswirkungen auch finanzieller Art für alle Mitgliedsstaaten haben können, ist der Verabschiedungsweg sehr formell und oft langwierig. Die nationalen Mitgliedsorganisationen haben viel mitzureden, und die Verabschiedung eines Standards kann sich leicht über mehrere Jahre hinziehen. Das ist besonders für sich derart rasch wandelnde Felder wie die Informationstechnologie oftmals

zu langsam. Deshalb ist man in Europa dabei, das traditionelle Prozedere um sogenannte *Workshops* zu ergänzen, deren Mitglieder sich hauptsächlich aus Industrie, Verwaltungen und Standardisierungsorganisationen zusammensetzen. Diese *Workshops* können nur Empfehlungen, sogenannte *CEN Workshop Agreements (CWAs)*, aussprechen, die aber in offiziellen Dokumenten referenziert werden können.

Zur Vorbereitung dieser Veranstaltungen, aber auch zur forcierten Ausarbeitung von »echten« Standards werden verstärkt von der EU finanzierte Projektgruppen mit i. d. R. drei bis vier Mitarbeitern eingesetzt, die sich aus einem Projektmanager, einem sog. *Editor* und einem oder mehreren *Reviewer* zusammensetzen. Der Editor ist für die tatsächliche Ausarbeitung der Dokumente zuständig, während dem Manager in erster Linie Verwaltungsfunktionen zukommen, er damit aber durchaus bedeutenden Einfluß auf das endgültige Produkt nehmen kann.

CEN/TC304

Mehr zufällig entschieden wir uns zur Mitarbeit in CEN/TC304. Dieses Gremium sei, so wurde mir mitgeteilt, der ideale Ort, eigene Ansichten und Ideen möglichst vielen einflußreichen Mitarbeitern zu Gehör zu bringen. Das erwies sich als wahr. Im Rahmen der Globalisierung werden paradoxerweise Probleme der lokalen Adaption von Software mit jedem Tag dringlicher. Gerade weil Softwarepakete weltweit vertrieben werden, müssen örtliche Bedürfnisse verstärkt berücksichtigt werden. Auch ein amerikanisches Softwareprodukt muß wissen, wie in Deutschland üblicherweise Datumsangaben ausgegeben werden oder wie in unserer Kultur Namen mit Umlaut behandelt werden − daß beispielsweise in einer Teilnehmerliste Küster nach Kuester, nicht nach Kuster einsortiert werden muß.

Die Bedürfnisse vieler Nationen wie z. B. Griechenland sind selbst heute noch viel elementarer: Übliche Softwarepakete unterstützen nicht einmal das volle Alphabet, ein Problem, das wir bis vor kurzem mit Umlauten auch noch hatten. Daraus wird auch der Name von TC304 verständlich: *European Localization Requirements*.

Dublin

Unser Debut hatten wir letzten Oktober in Dublin. TC304 befand sich gerade in dieser Zeit in einer Phase massiven Umbruchs. Die alten Arbeitsformen erwiesen sich in den Augen der meisten Teilnehmer als völlig unbrauchbar.

Folglich wurde das Dubliner Treffen von heftigen Streitereien fast zerrissen; an ernsthaftes Arbeiten war kaum zu denken. Zum ganz großen Disput kam es, als der britische Delegierte vorschlug, alle *Working groups* aufzuheben, um gänzlich neue Arbeitsweisen zu ermöglichen. Daß dieser Vorschlag, auch mit starker deutscher Unterstützung, nach fast schon handgreiflichen Diskussionen Erfolg hatte, war der Hauptgewinn dieser Sitzungswoche. So konnte man wenigstens für die Zukunft mit Resultaten rechnen.

Der Vorschlag hätte aber kaum Erfolg gehabt, wenn die Europäische Kommission nicht bereit gewesen wäre, die neuen Gruppen finanziell zu fördern. Der EU ging und geht es dabei nicht zuletzt um die Standardisierung des Eurozeichens (s. u.); die positiven Nebeneffekte sind aber durchaus begrüßenswert.

Für das ZDV bedeutete das Ergebnis den Schnellstart in Sachen Normierungsarbeit in Europa. Nicht nur wurde die Expertise der Abteilung für Literarische und Dokumentarische Datenverarbeitung durch die Übertragung verantwortungsvoller Posten anerkannt, sondern ich als deutscher Vertreter wurde auch in viele Entscheidungsprozesse eingebunden.

Der große Euro-Streit

Der zweite große Streitpunkt, in dem Deutschland zusammen mit einem renommierten amerikanischen Experten gegen den Rest der Anwesenden stand, war die Auseinandersetzung um das (damals noch brandneue) Eurozeichen, das nach dem Willen der Europäischen Kommission auch in den 8-Bit-Standard ISO 8859 eingefügt werden soll. Der damals hitzig diskutierte Vorschlag zu Latin-0, das Latin-1 ersetzen sollte, wurde zwar von den drei

führenden Industrienationen USA, Japan und Deutschland abgelehnt, hatte aber aus politischen Gründen die Unterstützung nicht nur der Europäischen Kommission, sondern auch verschiedener kleinerer Staaten. Dabei wurde klar, daß sich viele dieser Vertreter der entstehenden Probleme kaum bewußt waren. Latin-1 ist nämlich voll belegt; das Eurozeichen und andere Buchstaben, die gleich mit eingebaut werden sollten (z. B. Œ und ž), hätten daher existierende Zeichen ersetzen müssen, das Eurozeichen beispielsweise das Zeichen ±.

Praktisch alle momentan existierenden Datenbanken in Westeuropa und den USA basieren entweder auf reinem ASCII (ISO 646) (oder sogar nur Untermengen davon) oder aber Latin-1. Ein Großteil der E-Mail-Kommunikation ist auf die Interpretation der ankommenden Daten als Elemente von Latin-1 angewiesen. Die Kosten, die Deutschland allein durch eine solche de-facto-Ersetzung entstehen würden, sind horrend. Selbst konservative Schätzungen gehen weit in dreistellige Millionenbeträge.

Industrievertreter unterschiedlichster Provenienz − darunter Repräsentanten von Konzernen wie IBM und HP − betonen gleichfalls, daß sie wenig gewillt sind, derartige Ausgaben für eine Technologie auf sich zu nehmen, deren Tage gezählt sind. Sie möchten ihre Energien darauf konzentrieren, konsequent Unicode − oft über den Umweg des *UCS Transformation Format 8-bit form (UTF8)* − zu implementieren. Nicht nur moderne Programmiersprachen wie Java und Betriebssysteme wie Windows NT und AIX, sondern auch HTML 4 und viele der großen Datenbankhersteller setzen nachdrücklich auf Unicode.

Latin-0 ist nicht zuletzt durch das deutsche Engagement gescheitert, wenn auch nur knapp. Der jetzt favorisierte Entwurf Latin-9 gibt wesentlich weniger Anlaß zu Bedenken, da er nicht länger von sich behauptet, Latin-1 ersetzen zu wollen. Es ist zu bezweifeln, daß er einen großen Stellenwert gewinnen wird, zumal ein Gigant wie Microsoft das Eurozeichen bereits auf nichtstandardisierte Weise in seine proprietären Codepages (z. B. CP1252) untergebracht und entsprechende Software ausgeliefert hat.

Es ist wichtig, sich ins Gedächtnis zu rufen, daß Deutschland 50 Jahre lang gut ohne Währungssymbol, wie das € eines ist, ausgekommen ist. Für internationalen Zahlungsaustausch wird man sowieso das internationale Währungskennzeichen EUR (genau wie heute schon DEM oder USD) verwenden, da nur dies eindeutig ist. Das Symbol könnte bestenfalls auf inoffiziellen Preisauszeichnungen seinen Platz finden, und dort kann man es bereits jetzt mit handelsüblicher Software erzeugen.

Wirtschaftliche Notwendigkeit für das Eurozeichen in 8-Bit Zeichensätzen läßt sich schwerlich zeigen, die wirtschaftlichen Nachteile sind hingegen offensichtlich. Das neue Währungssymbol ist ein rein politisch gewolltes Logo, dessen forcierte Durchsetzung meiner persönlichen Ansicht nach der Währungsunion eher schadet als nützt.

Die Treffen in Brüssel und Reykjavik

In den folgenden beiden Treffen in Brüssel (Februar 98) und Reykjavik (Juni 98) trugen die Entwicklungen von Dublin Früchte. Die Projektteams nahmen ihre Arbeit auf, u. a. auch das Team, dessen Editor ich bin. Wir beschäftigen uns mit den *European Ordering Rules*, den Regeln für paneuropäisches Sortieren. Andere Projekte, auf die ich gleich eingehen werde, werden folgen.

Das heißt nicht, daß in TC304 jetzt alles optimal wäre. Zwar können interessante Aufgaben jetzt endlich mit der notwendigen Energie angegangen werden, aber es gibt immer noch Projekte, deren Notwendigkeit nur schwer einsichtig ist. Der Papierkrieg vor Ort ist ungeheuer − auf der letzten Reise wurden 13 kg neuer Dokumente verteilt. Aber es geht voran; Fortschritte sind überall sichtbar.

TUSTEP und multilinguales Sortieren in Europa

Benutzerdefiniertes, »kulturell korrektes« Sortieren und Registererstellung sind von jeher Stärken von TUSTEP. Diese Fähigkeit ist auch außerhalb der wissenschaftlichen Textdatenverarbeitung plötzlich topaktuell. Programmiersprachen wie Java und Organisationen wie das *Unicode Consortium* geben diesem Problem momentan höchste Priorität. Es ist nur konsequent, daß wir uns auf einen Posten in der europäischen Projektgruppe für multilinguales

Sortieren beworben haben. Allerdings hätten wir selbst nicht damit gerechnet, daß unsere Expertise mit dem Posten des Editors honoriert würde.

Das Projekt verläuft in zwei, formal voneinander unabhängigen Phasen: Sortieren des *Multilingual European Subsets (MES) No 2* und danach von MES-3. MES-2 − früher bekannt als das *Minimal European Subset* − ist eine für die Bedürfnisse der Bewohner der EU zugeschnittene Untermenge von Unicode, die im wesentlichen lateinische, griechische und kyrillische Buchstaben und Diakritika (einschließlich der für polytonisches Griechisch benötigten Akzente) sowie Sonderzeichen abdeckt. MES-3, bisher *Extended European Subset* genannt, geht einen Schritt weiter: Mit diesem Zeichensatz sollen alle einheimischen Sprachen Europas geschrieben werden können. Neben einer Menge weiterer »lateinischer« Buchstaben und Sonderzeichen kommen hier vor allem noch Georgisch und Armenisch hinzu.

Für dieses Vorhaben hat sich das Projektteam eine Reihe von Zielen gesetzt:

– Die Sortierregeln sollen den verschiedenen Bedürfnissen in Europa gerecht werden, ohne sich zu stark an nationale Gepflogenheiten einer Nation anzulehnen.
– Die Sortierkriterien sollen einsichtig sein und konsequent angewandt werden.
– Auch jemand, der einen speziellen Buchstaben wie etwa das ð (Eth) oder das ʒ (Ezh) nicht kennt, soll es in einem Katalog finden können.

Zusammengefaßt bedeutet dies: Klarheit für den Implementierer und einfacher Zugang für den Benutzer.

Mir als Editor obliegt der Eiertanz der Umsetzung. Einen ersten Entwurf für das Sortieren von MES-2 konnte ich Ende April vorgelegen; er ist für alle Interessierten im Netz erhältlich (*http://www.stri.is/TC304/EOR/*), bildete die Grundlage für die Diskussionen in Reykjavik und ist die Basis für weitere Arbeit.

In Reykjavik gemachte Vorschläge werden in die hoffentlich bald fertige nächste Version aufgenommen. Natürlich bin ich für alle Kritik und Kommentare auch von anderer Seite dankbar.

Sortieren international

Seit vielen Jahren wird auch auf internationaler Ebene an einem Sortierstandard gearbeitet, der unter dem Namen ISO FDIS 14651 firmiert. Dieser Entwurf definiert eher eine Sortiersyntax, tat dies allerdings bisher, indem er auf einen anderen, extrem POSIX-lastigen Standardentwurf (ISO CD 14652) rekurrierte. Ferner bietet er eine anpaßbare Default-Sortiertabelle mit einer ebenfalls sehr POSIX-orientierten Syntax.

Von inhaltlichen Schwierigkeiten abgesehen war der Entwurf bisher wegen seiner extrem aufwändigen Syntax inakzeptabel. Um dem Standard zu genügen, hätte es nicht ausgereicht, äquivalente Resultate zu garantieren, sondern man wäre gezwungen gewesen, die volle Syntax zu implementieren. Da der Sortierstandard die Syntax aber durch Referenz auf Erweiterungen der POSIX-Locale definierte, hätte eine konforme Sortierapplikation notwendig auch diesen ganzen Overhead an Bord nehmen müssen − und das in einer Zeit, in der UNIX und damit POSIX täglich an Boden verlieren.

Natürlich wäre es prinzipiell sehr wünschenswert, die europäischen Sortierregeln formal als eine Anwendung von FDIS 14651 zu definieren und somit potentielle Konflikte zwischen den Standards von vorneherein auszuschließen. Dies wurde auch von der Industrie − IBM, HP etc. − massiv gewünscht. Unter den gegebenen Bedingungen war dies aber in meinen Augen unmöglich. Es erschien daher sinnvoll, den Editor dieses Entwurfs zu treffen, idealerweise auf einer offiziellen Tagung des zuständigen ISO-Gremiums mit dem schönen Namen ISO IEC/JTC1/SC22/WG20 *Internationalization.* Das DIN unterstützte diese Ansicht.

Eine solche Tagung fand im Anschluß an das Reykjavik-Treffen von TC304 in Dublin statt.

Der Erfolg des Treffens übertraf meine kühnsten Erwartungen. Ich war ohne die Hoffnung hingegangen, wesentlich mehr tun zu können, als die deutsche Besorgnis zum Ausdruck zu bringen. Aber die Mängel des Entwurfs waren auch dem *Unicode Consortium* und ANSI aufgefallen, die einen ihrer besten Experten, Mitarbeiter von Sybase und Vizepräsident des *Unicode Consortium,* entsandt hatten. Auch Großbritannien hatte einen skeptischen Delegierten nominiert. In Zusammenarbeit gelang es binnen zweier Tage, einem unbrauchbaren Entwurf einen völlig neuen Ansatz zu geben. Die Syntax wurde völlig von POSIX abgekoppelt und im Standard selbst

formal neu definiert. Im Entwurf enthaltene Programmschnittstellen / APIs wurden restlos eliminiert. Die Defaulttabelle wurde gleichfalls radikal entschlackt und um viele Fehler auch prinzipieller Art bereinigt. Gleichzeitig nahm auch das *Unicode Consortium* europäische Hinweise zu seinem Sortierentwurf ernst und modifizierte ihn entsprechend.

Unter diesen Bedingungen erscheint es mir nun sehr vielversprechend, die europäischen Sortierregeln als Anwendung dieses Entwurfs aufzufassen. Solche Resultate kann man aber auch in der Zeit virtueller Realität nur zusammen um einen Tisch herum erarbeiten.

Browsing and Matching

TUSTEP hat schon seit langem überaus mächtige Suchmechanismen, sowohl was Wildcard-Suchen fast beliebiger Komplexität als auch was Fuzzy-Searching angeht. Dies in Verbindung mit TUSTEPs Freitextdatenbankfunktionen und seiner neuen CGI-Schnittstelle macht das System sehr attraktiv für den Einsatz als Suchmaschine auch im WWW.

Handelsübliche Software auf diesem Gebiet erfüllt selbst heute noch nicht die grundlegendsten Anforderungen kulturell korrekter Suche. Zwar sind die meisten Suchmaschinen zu primitiven Vereinfachungen in der Lage und finden auch Hinweise auf *Gothe*, wenn man *Göthe* sucht, aber kaum ein mir bekanntes Produkt liefert auch *Goethe*. Spätestens wenn man in Bereiche kommt, wo mehrere Schriften eine Rolle spielen − z. B. *Περικλῆς*, Perikles, Pericles oder Périclès −, müssen auch die letzten Systeme ihre Waffen strecken.

In einer Zeit, in der der Zugriff auf multilinguale Informationssammlungen täglich wichtiger wird, hat sich die Europäische Kommission entschlossen, eine Projektgruppe einzurichten, die sich zumindest einmal einen Überblick über die momentane Lage verschaffen soll. Natürlich wird man in dieser Gruppe nicht bei den hier angerissenen relativ primitiven Problemen stehenbleiben, sondern auch Felder wie multilinguale Thesauri, Lemmatisierung u. ä. angehen. In möglichen Folgeprojekten könnte man sich dann an die Erarbeitung von konkreten Lösungsstrategien machen.

Es liegt auf der Hand, daß das ZDV mit seiner Erfahrung auch an diesem Projekt interessiert ist und sich beworben hat. Wieder übertraf das Resultat die Erwartungen. Unter einer großen Bewerberzahl wurden wir als Projektmanager ausgewählt. Es ist hier meine Aufgabe, die verschiedenen Zielsetzungen zu koordinieren und für eine rasche Durchführung des Projekts zu sorgen, damit man diesem sich derart rasant entwickelnden Markt schnell Informationen zur Verfügung stellen kann.

Zusammenfassung

Der Artikel hat Ihnen hoffentlich einen kurzen Überblick über die Aktivitäten des ZDV im Normierungswesen gegeben. Gewinn und Verlust kann man hierbei nicht rein in Mark und Pfennig ausdrücken: Der verlorenen Arbeitszeit und den Reisekosten stehen die Refinanzierung durch die Projektmittel der Europäischen Union entgegen; beides dürfte sich in etwa die Waage halten. Der potentielle Gewinn liegt aber nicht in diesem Gebiet:

Einerseits freut es uns natürlich, daß unsere Expertise auf europäischer Ebene durch die Projekte anerkannt wird. Andererseits konfrontiert es uns auch mit einer der Universität oft sehr fremden Welt, die von anderen, oft von der Industrie geprägten Denkweisen dominiert wird. Und es ist nie schlecht, wenn man sich neuen Fragen stellen muß.

Marc Wilhelm Küster
kuester@zdv.uni-tuebingen.de