Yves Desrichard

Bibliothèques et écritures, d'ASCII à UNICODE

Paris, Editions du cercle de la librairie, 2009 (Collection Bibliothèques), p. 120, ISBN 978-2-7654-0974-8, € 29,00

La pubblicazione solleva l'aspetto fondamentale della gestione dei caratteri nei sistemi informatici e nei software di gestione delle biblioteche. L'autore non si riferisce solo ai diacritici e alle lettere accentate, ma all'aspetto più complesso legato alla gestione di testi in lingue e conseguentemente scritture diverse, che possono coesistere in un catalogo online. L'esempio scelto dall'autore per illustrare quali problematiche intende affrontare è trasparente: che tasti occorre digitare sulla tastiera – si chiede – per visualizzare a monitor, stampare e poter scambiare tramite internet "l'élection présidentielle française" e non "lâ€™Â©lection prÂ©sidentielle franÂ§aise" (testo codificato con UTF-8, ma visualizzato con ISO 8859-1)?

L'autore parte da un excursus storico sulla nascita della scrittura, da quella cuneiforme a quella fenicia, e sui diversi sistemi di scrittura: quelli basati su ideogrammi e pittogrammi, quelli sillabici e alfabetici. Vengono richiamate poi alcune caratteristiche delle scritture, funzionali alle riflessioni successive: la direzione, la presenza dei caratteri diacritici, le legature o filetti e alcune criticità legate alle trascrizioni fonetiche, ossia gli omografi, che rendono necessaria una elaborazione intellettuale e non la semplice applicazione di regole.

Vengono poi ricordate le principali scritture, alfabeto greco, latino, arabo, ebraico, devanagari e cinese (che ha fortemente influenzato quelle di Corea, Giappone e Vietnam). Questa sezione del libro ha lo scopo di evidenziare alcune criticità essenziali ai fini della codifica dei caratteri, ad esempio in arabo la stessa lettera può essere rappresentata in modo diverso a seconda della posizione che occupa nella parola: inizio parola, nel corpo della parola, alla fine o isolata. Per quanto riguarda le lingue come il cinese, che utilizzano ideogrammi o pittogrammi, una delle criticità legate alla codifica è legata al fatto che altre lingue dello stesso ceppo, come vietnamita, giapponese e coreano, possono condividere ideogrammi o pittogrammi con significato talvolta diverso. Per consentire la rappresentazione di una lingua mediante scrittura diversa da quella originale è possibile utilizzare tre procedimenti diversi:

– *traslitterazione*. Mediante questa operazione si trascrivono i grafemi (o glifi[1]) in modo tale che ad un grafema di un alfabeto o serie di grafemi della lingua di partenza corrisponda lo stesso grafema indipendentemente dalla pronuncia. Le traslitterazioni sono concepite per essere reversibili;

– *romanizzazione*. Si tratta di traslitterazione da scrittura non latina a scrittura latina;

– *trascrizione*. È un processo che mira a rappresentare sia pure in modo approssimativo la pronuncia di una lingua. Questo metodo è strettamente legato alla lingua di partenza e a quella di arrivo ossia al modo in cui viene percepita dalla lingua di arrivo e al suo sistema fonetico. In questo modo ad esempio è possibile leggere il Corano nella propria scrittura e in lingua araba senza sapere l'arabo.

Un esempio di traslitterazione dal cirillico all'italiano è il nome Gorbaciov. Secondo la norma ISO9 deve essere traslitterato (in francese) Gorbačëv, ma potrà essere trascritto Gorbatchof, Gorbachof, Gorbatchev a seconda della lingua di chi effettua la trascrizione.

La norma ISO 15924 si è proposta l'obiettivo di normalizzare il modo in cui rappresentare le scritture, assegnando a ciascuna "variante" di carattere un codice diverso, ossia ad esempio alla stessa lettera in posizioni diverse in lingua araba.

Il primo standard internazionale per la codifica dei caratteri è stato l'ASCII. Il suo principale limite era il fatto di essere stato concepito per codificare tutti i caratteri della lingua inglese, ma quasi esclusivamente quelli. Questo era determinato dal fatto che i primi produttori di elaboratori sono stati in gran parte di lingua inglese e quindi questo per loro non costituiva una limitazione nel progetto iniziale. L'ASCII esteso, adottato ben presto come standard internazionale (ISO 5426) doveva in parte superare questo ostacolo. Presto venne sostituito da ISO 8859 che permetteva non solo di trascrivere lingue neolatine, ma anche lingue che utilizzano altri alfabeti.

UNICODE è uno standard concepito dal consorzio omonimo (costituito da grandi società informatiche quali Adobe, Apple, IBM, Microsoft, Sun, Xerox). UNICODE è un sistema di codifica che assegna un numero univoco ad ogni carattere usato per la scrittura di testi, indipendentemente dalla lingua, dalla piattaforma informatica e dal programma utilizzati.

UNICODE incorpora, nella primissima parte, la codifica ISO/IEC 8859-1, ma va oltre, poiché codifica i caratteri usati in quasi tutte le lingue vive. UNICODE non rappresenta ancora tutti i caratteri in uso nel mondo, poiché è in evoluzione. Tale standard assicura compatibilità e non sovrapposizione con le codifiche dei caratteri già definiti. UNICODE viene supportato dai moderni standard della programmazione e del markup come XML, Java, JavaScript, LDAP, CORBA 3.0, e da vari sistemi operativi.

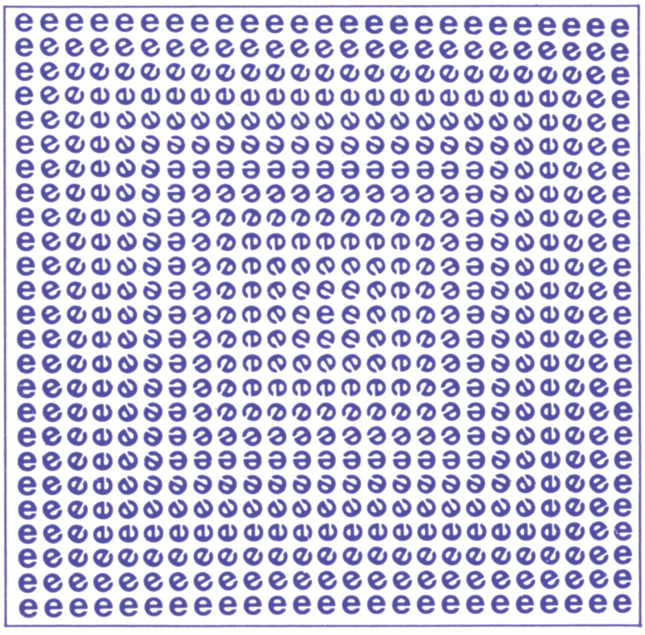
Inizialmente UNICODE era concepito come codifica a 16 bit, con la possibilità di codificare 65.536 caratteri. Ora invece lo standard UNICODE, allineato con la norma ISO 10646, prevede una codifica fino a 21 bit e supporta un repertorio di codici numerici che possono rappresentare circa un milione di caratteri.

È previsto l'uso di codifiche con unità da 8 bit (byte), 16 bit (word) e 32 bit (double word), descritte rispettivamente come UTF-8, UTF-16 e UTF-32.

Alcune entità considerate da altri standard caratteri specifici, come ad esempio parentesi aperta/parentesi chiusa o lettera nelle diverse posizioni nella lingua araba, non lo sono per UNICODE. Ad esempio UNICODE gestisce le parentesi, ma non ha un carattere diverso per parentesi aperta o chiusa. Sarà il sistema che gestisce la parentesi a richiedere una parentesi aperta o chiusa, a seconda della lingua e della direzione della scrittura.

Ugualmente UNICODE ha un codice per la "u" indipendentemente dalla pronuncia che avrà nelle diverse lingue o un codice per un ideogramma, anche se è comune a diverse lingue come giapponese, cinese e coreano. Sarà il contrassegno di lingua a dare l'informa-

zione della lingua del documento. Per le lettere accentate UNICODE offre sia la possibilità di gestire separatamente lettera e accento, sia la possibilità di utilizzare un unico codice per la lettera accentata.
Gli ultimi capitoli evidenziano l'impatto della presenza di lingue diverse, con la conseguente gestione e codifica di caratteri, nei sistemi di gestione automatizzati delle biblioteche e termina con una riflessione su google e su possibili evoluzioni biblioteconomiche.
Come già ricordato, UNICODE può essere usato in modo che uno stesso carattere sia condiviso da diverse lingue, anche con significati diversi, e il contrassegno di lingua dia l'informazione del significato da dare al carattere. In fase di ricerca è sempre più importante che vengano gestiti i caratteri speciali, parte essenziale dei diversi alfabeti. Sarebbe utile che fosse possibile cercare anche con alfabeti diversi da quello latino e con caratteri speciali. Le "stop words" sono anch'esse un aspetto non semplice da gestire nel caso di un catalogo popolato da diverse lingue. Nel caso degli indici per soggetto, sarebbe utile che fossero presenti in diverse lingue, poiché chi cerca documenti in una certa lingua potrebbe trovare utile poter cercare anche con i soggetti in tale lingua e caratteri. Le stringhe di soggetto sono normalmente separate da spazi, ma in alcune lingue (arabo, ebraico) l'articolo iniziale è legato alla prima parola della stringa. Normalmente, purtroppo i sistemi di gestione delle biblioteche utilizzano solo ASCII ridotto per i termini di indice e normalizzano i diacritici, eliminano legature, decodificando ad esempio œ in oe, nonostante non si tratti dello stesso simbolo. Per di più questa scelta potrà non essere trasparente per l'utente finale. Forse potrebbero essere utilizzati indici diversi per scritture diverse.
I motori di ricerca hanno funzioni analoghe ai cataloghi, ma si sa ben poco sul loro funzionamento poiché coloro che li hanno concepiti tengono a mantenere riservate tali informazioni. A maggio 2008 tuttavia Google ha annunciato di aver adottato la versione 5.1 di UNICODE, per la gestione di diverse lingue e scritture, segno che si è adattato ad esigenze internazionali più di quanto abbiano fatto i cataloghi.
Ai tempi dell'uso del solo ASCII le biblioteche che si trovavano a gestire lingue non latine usavano cataloghi manoscritti o dattiloscritti, soluzione che permetterà loro una transizione verso UNICODE senza perdite di informazioni.
Se si utilizza UNICODE in un catalogo che gestisce più scritture, non è possibile usare gli ISBD in senso stretto. Questo standard infatti richiede la trascrizione nella lingua del documento per le aree 1, 2, 4, 6. Nei sistemi di catalogazione usati in Occidente, si trascrive in caratteri latini con una traslitterazione o romanizzazione.
Questo non è prescritto dagli ISBD, ma dai sistemi informatici. Oggi la ISBD, versione francese, distingue la fonte dell'informazione per i documenti in scrittura latina e non latina. Una appendice specifica precisa come comportarsi per i documenti che utilizzano la scrittura nei due sensi. Molti documenti di lingua non latina vengono immessi in alfabeto latino, ma anche nella lingua originale. Il formato MARC infatti supporta anche questo doppio percorso.
Un altro aspetto critico è legato alle aree la cui lingua è del catalogatore. Alcuni paesi hanno due lingue ufficiali, ad esempio il Canada. In questi casi o si sceglie una delle due lingue come lingua "ufficiale" della catalogazione oppure si cataloga in entrambe, come in Canada. In questo caso non ci sono difficoltà, poiché la scrittura è la stessa. UNICODE renderà invece molto più semplice la gestione di diverse lingue e scritture nei paesi nei quali sono tutte lingue ufficiali, come ad esempio l'India.
Per concludere: una delle soluzioni fino ad ora per la gestione di scritture diverse era la traslitterazione, ma per chi cerca i documenti non è invece meglio poterli cercare in una certa lingua con la scrittura originale? Difficilmente vengono cercati da chi non li sa leggere. A dicembre 2007 per la prima volta il numero di pagine disponibili su web in UNICODE era superiore a quello delle pagine codificate con ASCII. Siamo però solo all'alba di una rivoluzione che dovrà gestire non solo lingue in scritture latine.
La pubblicazione termina con una breve bibliografia di testi e di siti utili ad approfondire alcuni dei principali temi affrontati.
La pubblicazione è sicuramente di grande interesse perché evidenzia una serie di criticità legate alla codifica delle lingue, soprattutto non latine, che devono essere tenute in considerazione da chi gestisce un catalogo di documenti in lingue che utilizzino scritture non latine. I cataloghi delle biblioteche avranno sempre più spesso la necessità di gestire tale documentazione. Le avvertenze dell'autore potranno quindi essere preziose per tutti coloro che dovranno progettare l'ingresso di lingue non latine nei cataloghi.



Composizione di Timm Ulrichs

Alessandra Citti

Biblioteca Polo di Rimini
Università di Bologna
alessandra.citti@unibo.it

[1] Per glifo si intende il segno grafico che rappresenta un carattere con specifiche caratteristiche, ossia si distingue il glifo della "a" corsiva dal glifo della "a" in tondo etc.