



## ΨΗΦΙΑΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ

Διδάσκων: κ.Καπιδάκης Σαράντος

Εργασία: «Πολυγλωσσικές και Πολυπολιτισμικές Ψηφιακές  
Βιβλιοθήκες»

**Κουλικούρδη Άννα**

Φεβρουάριος 2004

**Τίτλος: Πολυγλωσσικές και Πολυπολιτισμικές Ψηφιακές Βιβλιοθήκες**

**Περιγραφή:** Εργασία για το μάθημα Ψηφιακές Βιβλιοθήκες στα πλαίσια του *Προγράμματος Μεταπτυχιακών Σπουδών στην Επιστήμη της Πληροφορίας - Διοίκηση & Οργάνωση Βιβλιοθηκών με έμφαση στις Νέες Τεχνολογίες της Πληροφορίας*, για το χειμερινό εξάμηνο του ακαδημαϊκού έτους 2004.

**Θέματα / Λέξεις-κλειδιά:** Ψηφιακές Βιβλιοθήκες, Πολυγλωσσικότητα, Πολυπολιτισμικότητα, Σετ γλωσσών και χαρακτήρων, CLIR

**Δημιουργός:** Άννα Κουλικούρδη, Βιβλιοθήκη Athens Information Technology

**Ημερομηνία δημιουργίας:** 27-01-2004

**Χρόνος έκδοσης:** 2004

**Χώρα έκδοσης:** GR

**Γλώσσα κειμένου:** gre

## ΠΕΡΙΕΧΟΜΕΝΑ:

1. Τι περιλαμβάνει το θέμα;
2. Πώς εντάσσεται το θέμα στην ενότητα των ψηφιακών βιβλιοθηκών, γιατί είναι σημαντικό και γιατί πρέπει να μάθουμε για αυτό;
3. Στατιστικά στοιχεία
4. Πολιτισμός και γλώσσα
5. Από τοπικά συστήματα σε παγκόσμια συστήματα
6. Σχεδιαστικές Προκλήσεις
7. Αντιπροσώπευση σε ψηφιακή μορφή
8. Πολυγλωσσική ανάκτηση πληροφοριών
9. Σετ γλωσσών και χαρακτήρων
10. Μεταγραφή και άλλες απώλειες δεδομένων
11. Μονογλωσσικά, πολυγλωσσικά και παγκόσμια σετ χαρακτήρων
12. Τρέχουσες εφαρμογές και χρήσεις-προσεγγίσεις στο διεθνή χώρο
13. Προσεγγίσεις στον ελληνικό χώρο
14. Περιορισμοί , όρια και εμπόδια
15. Κριτική και σχόλια
16. Σημαντικές πηγές
17. Ερωτήματα και προβληματισμοί
18. Περίληψη και συμπεράσματα
19. Βιβλιογραφία

## Πολυγλωσσικές και Πολυπολιτισμικές Ψηφιακές Βιβλιοθήκες

Η παρούσα εργασία στηρίζεται στο άρθρο «Multi-Media, Multi-Cultural and Multi-Lingual Digital Libraries or How Do we Exchange Data in 400 Languages?/ Borgman C.L.» του περιοδικού D-Lib Magazine, τεύχος Ιουνίου 1997. Το D-Lib Magazine είναι μια αποκλειστικά ηλεκτρονική δημοσίευση με πρωταρχικό ενδιαφέρον στην έρευνα και ανάπτυξη ψηφιακών βιβλιοθηκών και τα περιεχόμενα της είναι διαθέσιμα στο <http://www.dlib.org>.

### Τι περιλαμβάνει το θέμα;

Η παρούσα εργασία έχει ως πρωταρχικό στόχο να εστιάσει την προσοχή στα πολυγλωσσικά ζητήματα που εμπλέκονται στο σχεδιασμό ψηφιακών βιβλιοθηκών προσβάσιμων μέσω του Διαδικτύου. Στοχεύει επίσης στο να αναδείξει μία από τις επείγουσες προκλήσεις για τη δημιουργία ενός παγκόσμιου κατανεμημένου δικτύου, το οποίο να εξυπηρετεί ανθρώπους που να μιλούν άλλες γλώσσες πλην της δεσπόζουσας Αγγλικής.

Πρώτα θα εισάγουμε μερικά γενικά ζητήματα του πολιτισμού και της γλώσσας, μετά θα αναφερθούμε στις σχεδιαστικές προκλήσεις μετάβασης από τοπικά σε παγκόσμια συστήματα, στο θέμα της πολυγλωσσικής ανάκτησης πληροφοριών συνοπτικά (συγκεκριμένα τρεις μεθόδους) και ακολούθως στα τεχνικά ζητήματα. Τα τεχνικά ζητήματα εμπλέκουν την επιλογή σετ χαρακτήρων που αντιπροσωπεύουν γλώσσες, μονογλωσσικά ή πολυγλωσσικά. Τέλος, γίνεται αναφορά σε τρέχουσες χρήσεις και προσεγγίσεις γενικότερα στο χώρο των βιβλιοθηκών διεθνώς, ειδικότερα στον ελληνικό χώρο, σε περιορισμούς, κριτική-σχόλια και σε σημαντικές πηγές, θέματα και προσπάθειες. Επειδή η κλίμακα του γλωσσικού προβλήματος είναι κατά μακράν μεγαλύτερη, γι' αυτό και το ενδιαφέρον της εργασίας αυτής επικεντρώνεται περισσότερο στις πολυγλωσσικές ψηφιακές βιβλιοθήκες.

Προτού προχωρήσουμε στις θεματικές αυτές ενότητες, χρειάζεται να αποσαφηνίσουμε τους κάτωθι όρους οι οποίοι συναντώνται στην παρούσα εργασία αλλά και στο χώρο των ψηφιακών βιβλιοθηκών γενικότερα:

- **Internationalization:** καθιστώντας δυνατή την παγκόσμια επικοινωνία, ανεξαρτήτως γλώσσας
- **Localization:** η προσαρμογή στις τοπικές ανάγκες
- **Multilingual Digital Library - Πολυγλωσσική Ψηφιακή Βιβλιοθήκη:** που περιέχει τεκμήρια σε περισσότερες της μιας γλώσσας, οι λειτουργίες της υλοποιούνται ταυτόχρονα σε τόσες γλώσσες όσες είναι επιθυμητό και οι λειτουργίες αναζήτησης και ανάκτησης είναι ανεξάρτητες από τον παράγοντα γλώσσα.

- Multilingual Document - Πολυγλωσσικό Τεκμήριο: που περιέχει κείμενο σε περισσότερες της μιας γλώσσας
- Cross-Language Information Retrieval (CLIR) - Πολυγλωσσική Ανάκτηση Πληροφοριών: ανάκτηση πληροφοριών σε διαφορετικές γλώσσες από αυτή που εκφράζεται η ερώτηση για αναζήτηση πληροφορίας (query)
- Transliteration (TL): Μεταγραφή ή αλλιώς απόδοση λέξεων μιας γλώσσας με χαρακτήρες άλλης γλώσσας
- Transcription (TS): Η απόδοση των ήχων μιας γλώσσας στους χαρακτήρες ενός αλφάβητου.
- Large Passive Vocabulary: όταν οι χρήστες μπορούν να διαβάσουν μια δεύτερη γλώσσα κατά την πολυγλωσσική ανάκτηση αλλά δεν είναι ικανοί να σχηματίζουν σωστά δομημένα ερωτήματα-queries (small active vocabulary)

**Πώς εντάσσεται το θέμα στην ενότητα των ψηφιακών βιβλιοθηκών, γιατί είναι σημαντικό και γιατί πρέπει να μάθουμε για αυτό;**

Το θέμα αυτό εντάσσεται άμεσα στον τομέα των ψηφιακών βιβλιοθηκών (DL sector) αν θέλουμε να μιλάμε για ψηφιακές βιβλιοθήκες (DLs) προσβάσιμες μέσω του Διαδικτύου, χωρίς εθνικά όρια και σύνορα, με αποτελεσματικότερη ανταλλαγή δεδομένων, επικοινωνία και τέλος για μια διεθνή βιβλιοθηκονομική κοινότητα.

Ως θέμα είναι σημαντικό διότι αν δεν επιλυθεί θα οδηγηθούμε αναπόφευκτα σε ένα πύργο της Βαβέλ στον ψηφιακό χώρο και τότε θα κάνουμε λόγο για αποκλεισμούς ή περιορισμούς πρόσβασης στη γνώση. Επίσης, γίνεται περισσότερο φλέγον για τις «μη-δεσπόζουσες (non-dominant)» γλώσσες και συνεκδοχικά για τις ψηφιακές βιβλιοθήκες, όπου το υλικό τους παρέχεται σε αυτές τις γλώσσες. Η επιβίωση αυτών των γλωσσών, που δεν θα είναι διαθέσιμες για ηλεκτρονική επικοινωνία, θα είναι απίστευτα προβληματική στο μέλλον και αυτό έχει ήδη αναγνωριστεί από προγράμματα επιδοτούμενα από την Ευρωπαϊκή Ένωση (π.χ. για τη συνεργασία μεταξύ Ευρωπαϊκής Ένωσης και Αμερικής με το National Science Foundation).

Το Internet θα πάψει να είναι χρήσιμο αν η επικοινωνία περιορίζεται σε ανταλλαγές κειμένου μεταξύ μόνο εκείνων που ομιλούν την Αγγλική γλώσσα και εκείνων που διαμένουν σε αγγλόφωνες περιοχές (π.χ. Ηνωμένες Πολιτείες). Αντίθετα, η αξία του έγκειται στην ικανότητα του να καθιστά τους ανθρώπους, προερχόμενους από διαφορετικά έθνη, ικανούς να μιλάνε πολλαπλές γλώσσες και να εφαρμόζουν πολλαπλά πολυμέσα στην αλληλεπίδραση τους ο ένας με τον άλλον. Τα δίκτυα υπολογιστών πρέπει να γίνονται ολοένα αποτελεσματικότερα για επικοινωνία με κείμενο (textual communication) όχι μόνο στην Αγγλική αλλά και σε πολλές άλλες γλώσσες, πόσο μάλλον σε αλληλεπιδράσεις μεταξύ πολλών γλωσσών.

Το ζήτημα αυτό χρήζει μείζονος προσοχής διότι οι περισσότερες δραστηριότητες έρευνας και ανάπτυξης ως τώρα έχουν επικεντρωθεί σε μονογλωσσικά περιβάλλοντα και στην πλειοψηφία των περιπτώσεων η κυρίαρχουσα γλώσσα ήταν η Αγγλική. Η υποστήριξη της αναζήτησης και παρουσίασης δεδομένων σε πολλαπλές γλώσσες είναι ένα πολύ σημαντικό και ανερχόμενο θέμα για όλες τις ψηφιακές βιβλιοθήκες. Ακόμη και αν μία ψηφιακή βιβλιοθήκη περιέχει υλικό σε μόνο μία γλώσσα, το περιεχόμενο της πρέπει να είναι αναζητήσιμο και παρουσιάσιμο σε υπολογιστές χωρών που κάνουν χρήση άλλων γλωσσών.

Θέματα διαλειτουργικότητας, φορητότητας και ανταλλαγής δεδομένων σχετικά με πολυγλωσσικά σετ χαρακτήρων είχαν λάβει εντυπωσιακά ελάχιστη προσοχή στο παρελθόν στην κοινότητα των ψηφιακών βιβλιοθηκών ή σε συζητήσεις προτύπων για την πληροφοριακή υποδομή, με εξαίρεση κάποιες χώρες στην Ευρώπη. Γι' αυτό χρειάζεται η ενημέρωση, η πληροφόρηση και το ενδιαφέρον όλων των επιστημόνων του χώρου για τις εξελίξεις αυτές.

Από τη μία πλευρά έχουμε το σκηνικό όπου πρότυπα για την αντιπροσώπευση πολυμέσων υιοθετούνται πολύ γρήγορα, παράλληλα με τη διαθεσιμότητα του πολυμεσικού περιεχομένου σε ηλεκτρονική μορφή και από την άλλη εν αντιθέσει, έχουμε υλικό σε κείμενα αξίας εκατοντάδων ή και χιλιάδων χρόνων σε εκατοντάδες γλώσσες, που δημιουργήθηκαν πολύ πριν τα πρότυπα κωδικοποίησης προϋπάρξουν. Το περιεχόμενο των κειμένων αυτών κωδικοποιείται σε γλώσσα και σε εφαρμογές που είναι δύσκολο να ανταλλαχθεί πλήρως (αν μπορεί να ειπωθεί αυτό) χωρίς να χαθούν δεδομένα.

Όλες οι αναφορές που ακολουθούν είναι δάνειες από την ερευνητική κοινότητα των βιβλιοθηκών, που τυπικά χειρίζεται συλλογές υλικού σε 400 ή παραπλήσιο αριθμό γλωσσών. Αυτά είναι προβλήματα που αντιμετωπίζονται και πρέπει να τα γνωρίζουν όχι μόνο οι σχεδιαστές ψηφιακών βιβλιοθηκών, οι βιβλιοθηκονόμοι και επιστήμονες της πληροφόρησης αλλά και αυτοί που αναπτύσσουν και χειρίζονται κάθε τεχνολογία επικοινωνίας που διασταυρώνει εθνικά ή γλωσσικά σύνορα.

## ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ

Ακολουθούν στατιστικά στοιχεία που στοχεύουν να δώσουν στον αναγνώστη μια εισαγωγική εικόνα για το πολυγλωσσικό περιβάλλον. Σύμφωνα με παγκόσμια έρευνα που διεξήχθη το Μάρτιο του 2002 (<http://www.sis.pitt.edu/~d/wkshop/supplement/8>) οι on-line γλώσσες που χρησιμοποιούνται για τη διάχυση πληροφοριών στο Internet έχουν ως εξής:

1. Αγγλική – 40.2%
2. Κινεζική – 9.8%
3. Γιαπωνέζικη – 9.2%
4. Ισπανική – 7.2%
5. Γερμανική – 6.8%
6. Κορεάτικη – 4.4%
7. Γαλλική – 3.9%
8. Ιταλική – 3.6%
9. Πορτογαλική – 2.6%
10. Ολλανδική – 2.1%

Πληροφοριακά, σύμφωνα με πηγή του CLEF 2000: State-of-the-Art Multilingual Information Access, αν θέλαμε να μεταφράσουμε τις 400.000.000 μη Αγγλικές σελίδες του Διαδικτύου (WWW) θα χρειαζόνταν 100.000 ημέρες (≈300 χρόνια) σε ένα πολύ γρήγορο ηλεκτρονικό υπολογιστή. Ή διαφορετικά 1 μήνα σε 3.600 PCs.

Επίσης, οι χρήστες του Internet που δεν μιλούν Αγγλικά ποσοστιαία έχουν ως εξής:

- § 2001-49%
- § 2003-54%
- § 2005-59%

Ο συνολικός αριθμός των χρηστών του Internet θα αυξηθεί από 171.000.000 σε 345.000.000 μέχρι το 2005.

Τέλος, η ιστορία της πολυγλωσσικής πρόσβασης στην πληροφορία (Multi-lingual Information Access) ανατρέχει στα 1970 όπου ο Salton έκανε τα πρώτα πειράματα ανάκτησης με Αγγλο/Γερμανικό λεξικό ενώ η πρώτη διδακτορική διατριβή πάνω στην πολυγλωσσική ανάκτηση πληροφοριών έλαβε χώρα το 1994 από τον Khaled Radwan (πηγή του CLEF 2000: State-of-the-Art Multilingual Information Access).

## **ΠΟΛΙΤΙΣΜΟΣ ΚΑΙ ΓΛΩΣΣΑ**

Η ομιλία διαφέρει από την γραφή και η προφορική με τη γραπτή επικοινωνία είναι ακόμη πολύ διαφορετικές μεταξύ τους. Η συνομιλία ενός ανθρώπου στην μητρική του γλώσσα με αυτόχθονες του είναι πολύ διαφορετική από την συνομιλία που υλοποιείται μέσω ενός μεταφραστή. Οι γλωσσικές μεταφράσεις, είτε προφορικές ή γραπτές, είτε χειρονακτικές ή αυτόματες, δεν μπορούν να είναι πραγματικά ισάξιες λόγω λεπτών διαφορών μεταξύ των γλωσσών από τις οποίες κατάγονται. Γι' αυτό το περιεχόμενο και η αποτελεσματικότητα των μηνυμάτων είναι αδιαχώριστα από τη μορφή της επικοινωνίας και της γλώσσας στην οποία υλοποιείται η ίδια η επικοινωνία.

Για αυτούς τους λόγους, στόχος όλων είναι να αιχμαλωτίσουμε το περιεχόμενο των Ψηφιακών Βιβλιοθηκών στις πλουσιότερες δυνατές μορφές του με τελική επιδίωξη να διαβεβαιώσουμε το μέγιστο αποτέλεσμα για επικοινωνία. Θέλουμε ακριβείς αναπαραστάσεις και την ελάχιστη αλλοίωση των προθέσεων του δημιουργού (συγγραφέα, καλλιτέχνη, δημιουργού ταινίας, κ.ά.). Την ίδια στιγμή, θέλουμε να παρέχουμε την μεγαλύτερη ποικιλία αναζήτησης, έκθεσης και αιχμαλώτισης των δυνατοτήτων για αυτούς που αναζητούν το περιεχόμενο, για τους ερευνητές ή χρήστες των ψηφιακών βιβλιοθηκών που μπορεί να προέρχονται από διαφορετικούς πολιτισμούς και να μιλούν διαφορετικές γλώσσες από αυτές των δημιουργών.

Εδώ έγκειται το παράδοξο της ανάκτησης της πληροφορίας: η ανάγκη να περιγράψεις την πληροφορία που δεν έχει κάποιος. Έχουν ξοδευτεί δεκαετίες σχεδιάζοντας μηχανισμούς για να συνταιριάξουμε τις εκφράσεις των αναζητητών με εκείνες των δημιουργών των κειμένων (αιώνες, αν ληφθούν υπόψη και τα χειρονακτικά συστήματα ανάκτησης). Αυτό είναι ένα άλυτο πρόβλημα λόγω του πλούτου της ανθρώπινης επικοινωνίας. Οι άνθρωποι εκφράζονται με ιδιαίτερους τρόπους και οι όροι τους συχνά δεν ταιριάζουν με αυτούς των δημιουργών και των ευρετηριαστών της πληροφορίας που ζητείται. Συνεκδοχικά, οι ίδιοι όροι μπορεί να έχουν πολλαπλά νοήματα σε πολλαπλά συμφραζόμενα. Οι ολοένα και καλύτερες αναπτυσσόμενες μέθοδοι, εργαλεία και τεχνικές σίγουρα θα περιορίσουν τις προαναφερθείσες διαφορές μεταξύ των αναζητητών και των δημιουργών περιεχομένου αλλά δεν θα το καλύψουν ποτέ ολοκληρωτικά.

Τα πολιτισμικά θέματα διαπερνούν τις εφαρμογές ψηφιακών βιβλιοθηκών είτε μελετώντας τον πολιτισμό σε τοπικό επίπεδο εφαρμογής (στην τέχνη, μουσεία, επιστημονικές και σχολικές βιβλιοθήκες) είτε σε πολυεθνική κλίμακα (διαφορετικές τακτικές εφαρμόζονται στις Ηνωμένες Πολιτείες και άλλες στην Κεντρική και Ανατολική Ευρώπη). Η προσαρμογή αυτών των τοπικών πολιτισμών και η συνάντησή τους με πρότυπα και μεθόδους για συμβατότητα με άλλα συστήματα και εφαρμογές θέτει το κριτικό πρόβλημα της διαλειτουργικότητας (interoperability). Χαρακτηριστικό παράδειγμα όπου αντιμετωπίζεται μεγάλο πρόβλημα διαλειτουργικότητας παρατηρείται στις βιβλιοθήκες της Ινδίας, όπου χρησιμοποιούν ποικιλία Ινδικών γλωσσών.

## **ΑΠΟ ΤΟΠΙΚΑ ΣΥΣΤΗΜΑΤΑ ΣΕ ΠΑΓΚΟΣΜΙΑ ΣΥΣΤΗΜΑΤΑ**

Είναι ευκόλως κατανοητό πως τα ευκολότερα συστήματα προς σχεδιασμό είναι αυτά που προορίζονται για καλά ορισμένες εφαρμογές και καλά ορισμένους πληθυσμούς χρηστών. Και αυτό διότι οι σχεδιαστές μπορούν να χτίσουν κλειστά συστήματα προσαρμοσμένα σε μια κοινότητα χρηστών. Αυτές όμως είναι συνθήκες που τείνουν να εκλείψουν σήμερα.

Πλέον, η σύγχρονη τάση είναι οι σχεδιαστές να κατασκευάζουν ανοικτά συστήματα (open systems) που να εξυπηρετούν απομακρυσμένους και πιθανότατα αγνώστους πληθυσμούς και όχι ένα τοπικό πληθυσμό ή μια τοπική γλώσσα (minority language). Οι πολυγλωσσικές ψηφιακές βιβλιοθήκες έχουν εξαπλωθεί με αλματώδεις ρυθμούς και ήδη έχουν αρχίσει να γίνονται κοινός τόπος σε χώρες όπου η εθνική γλώσσα και η Αγγλική χρησιμοποιούνται από κοινού ως επιστημονική και τεχνική ορολογία σε πανευρωπαϊκά ιδρύματα όπως διεθνείς εταιρικές συνεργασίες (consortia) έρευνας, σε πολυεθνικές εταιρείες κ.ά

Χαρακτηριστικά παραδείγματα που δηλώνουν τη μετάβαση από τα τοπικά στα παγκόσμια συστήματα είναι οι ψηφιακές βιβλιοθήκες Πανεπιστημίων, οι οποίες χτίζουν τις συλλογές και το υλικό τους αρχικά για την πανεπιστημιακή κοινότητα τους και αργότερα την καθιστούν διαθέσιμη και στο Internet. Η μετάβαση αυτή παρατηρείται και στο χώρο των εταιρειών, οι οποίες αναπτύσσουν βάσεις δεδομένων, αρχικά τις λειτουργούν στο τοπικό τους site και έπειτα παρέχονται σε εταιρικά sites σε όλο τον κόσμο. Επιστημονικές βάσεις δεδομένων σχεδιάζονται για εφαρμογές έρευνας και αργότερα διατίθενται για διάχυση γνώσης. Όλες οι παραπάνω εφαρμογές θα μπορούσαν να έχουν περιεχόμενο σε πολλαπλές γλώσσες, άρα είναι αυτονόητη η σημασία της πολυγλωσσικότητας και της παγκόσμιας πρόσβασης στη γνώση, στην πρόοδο και την ανάπτυξη.

Επιπλέον, το θέμα αυτό θα πρέπει να εξεταστεί όχι μόνο από την πλευρά του χρήστη αλλά και των δημιουργών. Δηλαδή, δεν θα πρέπει μόνο οι χρήστες σε όλο τον κόσμο να έχουν πρόσβαση σε τεράστιους όγκους πληροφορίας όλων των ειδών αλλά και οι παροχείς γνώσης να μπορούν να κάνουν τη δουλειά και τις ιδέες τους διαθέσιμες σε όποια γλώσσα προτίμησής τους, γνωρίζοντας πως αυτό δεν ενέχει αποκλεισμούς ή περιορισμούς πρόσβασης.

Για όλους αυτούς τους λόγους η πολυγλωσσική ανάκτηση πληροφοριών (Cross-Language Information Retrieval – CLIR) έχει γίνει θέμα μείζονος σημασίας στο πεδίο των Ψηφιακών Βιβλιοθηκών.

## **ΣΧΕΔΙΑΣΤΙΚΕΣ ΠΡΟΚΛΗΣΕΙΣ**

Με τη μετάβαση από τα τοπικά στα κατανεμημένα συστήματα (distributed environments), αλλάζει ριζικά και ο τρόπος σχεδιασμού τους ενώ διαφαίνονται αρκετές προκλήσεις για τους σχεδιαστές, κάποιες αρκετά δυσεπίλυτες.

Αναλυτικότερα, σε ένα τοπικό σύστημα οι σχεδιαστές μπορούν να προσαρμόσουν την διεπιφάνεια του χρήστη ( GUI: Graphical User Interface), την παρουσίαση περιεχομένου και τις λειτουργικές ικανότητες στον τοπικό πολιτισμό του και στο διαθέσιμο υλικό και λογισμικό. Μπορούν να καθορίσουν εύκολα τις παραμέτρους εισόδου (input) και εξόδου (output) δεδομένων. Στη συνέχεια μπορούν να ορίσουν τα πληκτρολόγια έτσι ώστε να υποστηρίζουν τις τοπικές γλώσσες των δεδομένων που θα εισαχθούν. Ορίζουν τη τοπική γλώσσα πλοήγησης (language navigation), που είναι η γλώσσα των διεπιφανειών και των μηνυμάτων ενός συστήματος ψηφιακής βιβλιοθήκης.

Οι οθόνες και οι εκτυπωτές ορίζονται επίσης να υποστηρίζουν την κατάλληλη παρουσίαση των τοπικών γλωσσών. Στα συνηθισμένα περιβάλλοντα ηλεκτρονικών υπολογιστών κατά κανόνα είναι προεπιλεγμένη για εγκατάσταση η γραμματοσειρά εκείνη που να υποστηρίζει την αυτόχθονα γλώσσα και επιπροσθέτως την Αγγλική. Σε τέτοιες περιπτώσεις, για την παρουσίαση δεδομένων σε άλλες γλώσσες ο χρήστης πρέπει να εγκαταστήσει επιπρόσθετη γραμματοσειρά για αυτή τη γλώσσα, μια εργασία σχετικά δύσκολη για ένα κοινό χρήστη των υπολογιστών και του Internet.

Η κωδικοποίηση χαρακτήρων σε ηλεκτρονική μορφή περιλαμβάνει υλικό και λογισμικό το οποίο να υποστηρίζει είσοδο, αποθήκευση, επεξεργασία, ταξινόμηση, παρουσίαση και εκτύπωση. Η εσωτερική αντιπροσώπευση κάθε χαρακτήρα προσδιορίζει πώς αυτός μεταχειρίζεται από το υλικό (πληκτρολόγιο, εκτυπωτής) και το λογισμικό. Περιπτώσεις όπου χαρακτήρες δεν μπορούν να παρουσιαστούν λόγω της έλλειψης της κατάλληλης γραμματοσειράς δεν μπορούν να αποφευχθούν. Δύο χαρακτήρες μπορεί να εμφανίζονται το ίδιο σε μια οθόνη αλλά μπορεί να εμφανίζονται διαφορετικά λόγω διαφορετικών τοποθετήσεων τους σε πολλαπλές γλώσσες.

Συμπερασματικά, υπάρχουν αμέτρητες πιθανότητες για λανθασμένα συνταιριάσματα και λάθη πρόσβασης στις ψηφιακές βιβλιοθήκες κατανεμημένων περιβαλλόντων, λαμβάνοντας υπόψη την απίστευτη ποικιλία υλικού και λογισμικού που χρησιμοποιείται από τις ψηφιακές βιβλιοθήκες και τους χρήστες καθώς και την ποικιλία γλωσσών και συστημάτων κωδικοποίησης χαρακτήρων.

Ακόμη και οι σχεδιαστές αντιμετωπίζουν προβλήματα καθότι έχουν πολύ λιγότερο έλεγχο στις ψηφιακές βιβλιοθήκες που προορίζονται για χρήση σε παγκόσμια κατανεμημένα περιβάλλοντα. Οι πλατφόρμες υλικού και λογισμικού των χρηστών είναι τυπικά διάφορες και αλλάζουν γρήγορα. Οι σχεδιαστές συχνά πρέπει να θέτουν προδιαγραφές ή να απαιτούν μια ελάχιστη (minimum) εκδοχή του λογισμικού του

πελάτη, δημιουργώντας trade-offs (εξισορροπήσεις παραγόντων – δηλαδή χαμηλώνοντας τις απαιτήσεις ώστε να απευθύνονται σε ένα μεγαλύτερο πληθυσμό ή αυξάνοντας τις απαιτήσεις για να παρέχουν πιο επιτηδευμένες δυνατότητες). Δηλαδή, από τη μια πλευρά αυξάνεται το ένα σκέλος και από την άλλη μειώνεται το άλλο. Όσο πιο εκλεπτυσμένες είναι οι δυνατότητες της πολυγλωσσικής αναζήτησης, τόσο πιθανότερο είναι οι απαιτήσεις να είναι μεγαλύτερες και οι άνθρωποι στους οποίους απευθύνονται λιγότεροι.

Ένα άλλο δίλημμα τίθεται με το ποια πρότυπα θα εφαρμοστούν σε μια ψηφιακή βιβλιοθήκη και ποιος διευκρινίζει ποια είναι τα ιδανικά. Γιατί ακόμη και τα ιδανικά, είδαμε παραπάνω πως ενέχουν τα δικά τους trade-offs. Ο οργανισμός στον οποίο ανήκει η ψηφιακή βιβλιοθήκη ή ο φορέας που την χρηματοδοτεί έχει σημαντικό μερίδιο στη λήψη μιας τέτοιας απόφασης και γενικότερα στο να νομοθετήσει ποια πρότυπα θα υιοθετηθούν. Για μια τέτοια απόφαση, υπάρχουν αρκετές παράμετροι όπως ποια πρότυπα είναι πιο σταθερά, εφαρμόσιμα και με περισσότερη ικανότητα αλληλεπίδρασης για να ανταλλάσσουν δεδομένα με άλλα συστήματα.

Ως προς το σετ χαρακτήρων για την παρουσίαση του κειμένου, οι σχεδιαστές μερικές φορές κλίνουν προς δύο κατευθύνσεις. Είτε επιλέγουν το πρότυπο που εφαρμόζεται στη χώρα τους και αντιπροσωπεύει την εθνική γλώσσα είτε ένα παγκόσμιο σετ χαρακτήρων, ώστε η εθνική τους γλώσσα να παρουσιάζεται και σε άλλες χώρες. Προς το παρόν, η τάση είναι να υπάρχουν και να δημιουργούνται σχήματα (formats) και εφαρμογές εξειδικευμένες και διαφορετικές για κάθε γλώσσα και χώρα, γεγονός που καθιστά το χάσμα μεταξύ των πολυγλωσσικών ψηφιακών βιβλιοθηκών αγεφύρωτο. Τεράστιες ποσότητες δεδομένων κειμένου παράγονται σε ψηφιακή μορφή και παρουσιάζονται σε ποικίλες διατάξεις κατάλληλες για εφαρμογές, γλώσσες και χώρες.

## **ΑΝΤΙΠΡΟΣΩΠΕΥΣΗ ΣΕ ΨΗΦΙΑΚΗ ΜΟΡΦΗ**

Η πρόσβαση, διατήρηση και ανταλλαγή δεδομένων σε ψηφιακή μορφή επέφερε μαζί της πολλές αλλαγές. Δηλαδή, παρόλο που έχουμε αιχμαλωτίσει κείμενο, εικόνες, ήχο σε μηχαναγνώσιμες μορφές εδώ και αρκετές δεκαετίες, τα θέματα παρουσίασης μόνο τώρα έλαβαν περισσότερη προσοχή.

Στις ψηφιακές βιβλιοθήκες αυτό διαφοροποιείται κατά δύο σημαντικούς τρόπους: 1) από στατική έξοδο δεδομένων (output) σε δυναμική ανταλλαγή δεδομένων και 2) από ένα μηχανισμό μεταφοράς σε μία μόνιμη αρχειακή μορφή. Αναλυτικότερα, σε μια έντυπη εγγραφή, από τη στιγμή που η εγγραφή βγαίνει ή ένα βιβλίο τυπώνεται, δεν έχει πλέον σημασία πώς το περιεχόμενο παρουσιάζόταν σε μηχαναγνώσιμη μορφή. Αντίθετα, σε μια ψηφιακή βιβλιοθήκη, το περιεχόμενο πρέπει να αναζητείται συνεχώς, να επεξεργάζεται και συχνά να ανταλλάσσεται με άλλες εφαρμογές στον ίδιο και σε άλλους υπολογιστές. Γι' αυτό η παρουσίαση καθίσταται σημαντικό ζήτημα.

Παλιότερα τα ηλεκτρονικά μέσα τα χρησιμοποιούσαμε μόνο ως μηχανισμούς μεταφοράς, γι' αυτό και κατεβάλαμε πολύ λίγη προσπάθεια να διατηρήσουμε το περιεχόμενο. Πολλές έντυπες δημοσιεύσεις υπάρχουν μόνο σε χαρτί. Πολλές από τις τηλεοπτικές εκπομπές των τελευταίων ετών χάθηκαν καθώς τα μέσα καταγραφής

χρησιμοποιήθηκαν ξανά ή αφέθησαν στην παρακμή. Τώρα αναγνωρίζουμε ότι τα ψηφιακά δεδομένα πρέπει να εκλαμβάνονται ως μια μόνιμη μορφή παρουσίας που απαιτεί μέσα για να αποθηκεύσει περιεχόμενο σε ολοκληρωτικές μορφές και όχι ως μέσο μεταφοράς.

## ΠΟΛΥΓΛΩΣΣΙΚΗ ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ

Η πολυγλωσσική ανάκτηση πληροφοριών (Cross-Language Information Retrieval-CLIR) είναι ένας νέος κλάδος, για τον οποίο έχει παρουσιαστεί μεγάλο ενδιαφέρον, σχετίζεται άμεσα με τις πολυγλωσσικές ψηφιακές βιβλιοθήκες και αποτελεί αντικείμενο πολλών διεθνών συνεδρίων. Για περισσότερες πληροφορίες για τον κλάδο αυτό μπορεί κανείς να ανατρέξει στο <http://www.clis.umd.edu/dlrg/clir/>.

Οι τρεις κύριες προσεγγίσεις ως προς τη πολυγλωσσική ανάκτηση που αναφέρονται εδώ σχετίζονται με :

1. Μετάφραση κειμένου μέσω μηχανών μετάφρασης (Machine Translation-MT): Η λύση αυτή δεν φαίνεται να είναι ρεαλιστική απάντηση στο πρόβλημα εναρμόνισης δεδομένων και ερωτημάτων σε πολλές γλώσσες. Εκτός του ότι αυτά τα συστήματα απέχουν πολύ ακόμη από το να πετύχουν υψηλά αποτελέσματα, η μετάφραση ολόκληρων συλλογών σε άλλη γλώσσα δεν θα ήταν μόνο πολύ ακριβή αλλά θα ενέπλεκε και έναν αριθμό εργασιών περιττό από την πλευρά της πολυγλωσσικής ανάκτησης, π.χ. χειρισμός της σειράς λέξεων.
2. Τεχνικές βασισμένες στην Γνώση, όπως λεξικά, θησαυροί ή οντολογίες γενικού σκοπού. Η μετάφραση ερώτησης μέσω μηχαναγνώσιμου λεξικού (machine readable dictionary- MRD) έχει αποδειχθεί πως ρίχνει την απόδοση σε ποσοστό 40-60% της ανάκτησης και αυτό για τους εξής λόγους:
  - ο Τα λεξικά γενικού σκοπού (general purpose dictionaries) δεν περιέχουν εξειδικευμένο λεξιλόγιο
  - ο Εντοπίζεται πληθώρα εσφαλμένων μεταφράσεων
  - ο Παρουσία μη ικανοποιητικών δίγλωσσων ηλεκτρονικών λεξικών
  - ο Αποτυχία μετάφρασης συγκεκριμένων όρων με πολλαπλές λέξεις

Μία από τις καλύτερες και δοκιμασμένες προσεγγίσεις στην πολυγλωσσική ανάκτηση είναι αυτή που βασίζεται στους θησαυρούς. Οι θησαυροί είναι ειδικευμένες οντολογίες για την οργάνωση ορολογιών. Ένας πολυγλωσσικός θησαυρός οργανώνει ορολογία για γλώσσες περισσότερες της μιας. Υπάρχει ένας μεγάλος αριθμός συστημάτων βασισμένων σε θησαυρούς, διαθέσιμος στην αγορά. Παρόλο που η χρήση πολυγλωσσικών θησαυρών έχει αποφέρει ικανοποιητικά αποτελέσματα για την πολυγλωσσική ανάκτηση, ο περιορισμός τους είναι πως η δόμηση και η διατήρησή τους είναι ακριβές και χρειάζεται υψηλή εκπαίδευση για την αποτελεσματική χρήση τους.

3. Η χρήση οντολογιών (ontologies): Η μόνη γενικής χρήσης πολυγλωσσική οντολογία για την οποία έχουμε τις περισσότερες πληροφορίες είναι αυτή που έχει αναπτυχθεί στο EuroWordNet project (<http://illc.uva.nl/EuroWordNet>).

Το EuroWordNet είναι μια πολυγλωσσική βάση δεδομένων που αναπαριστά βασικές σημασιακές σχέσεις μεταξύ αρκετών Ευρωπαϊκών γλωσσών. Οι περιορισμοί είναι, όπως και στους θησαυρούς, ότι είναι ακριβές για δόμηση, διατήρηση καθώς και δύσκολες για ενημέρωση. Οι γλωσσικές διαφορές και οι πολιτισμικοί παράγοντες περιπλέκουν ακόμη περισσότερο την κατάσταση και το πρόβλημα διογκώνεται όταν περιπλέκονται πολλές γλώσσες.

Συμπερασματικά, κάθε μέθοδος πολυγλωσσικής ανάκτησης παρουσιάζει περιορισμούς. Αυτό που διαφαίνεται ως λύση είναι ότι πρέπει να πειραματιστούμε με ένα συνδυασμό όλων των προαναφερόμενων μεθόδων, δηλαδή συνδυασμό λεξικών ή θησαυρών. Περισσότερη μελέτη θα πρέπει να εστιαστεί και στον ανθρώπινο παράγοντα (στο χρήστη και τις ανάγκες του) αλλά και στο κατά πόσο μπορούν να βοηθήσουν οι αυτόχθονες ομιλητές μιας χώρας, οι ειδικοί σε γλωσσικά ζητήματα, οι εξειδικευμένοι μεταφραστές και προσωπικό με εμπειρία στην αλληλεπίδραση ανθρώπου-υπολογιστή.

## **ΣΕΤ ΓΛΩΣΣΩΝ ΚΑΙ ΧΑΡΑΚΤΗΡΩΝ**

Η παρουσίαση των σετ χαρακτήρων είναι ένα πρόβλημα παρόμοιο με εκείνο της παρουσίασης πολυμεσικών αντικειμένων στις ψηφιακές βιβλιοθήκες αλλά αυτό είναι ακόμη εντονότερο λόγω του μαζικού όγκου της επικοινωνίας με κείμενο και της ανταλλαγής δεδομένων που λαμβάνει χώρα στα δίκτυα υπολογιστών. Βασική είναι η παράμετρος του πολιτισμού διότι κάθε λαός θέλει να διατηρεί τη γλώσσα του σε μια πλήρη και ακέραια μορφή. Μη ολοκληρωμένες ή λανθασμένες ανταλλαγές δεδομένων καταλήγουν σε αποτυχίες εύρεσης πληροφορίας ή ακόμη και σε μόνιμη απώλεια της πληροφορίας. Ο χειρισμός των σετ χαρακτήρων για πολλές γλώσσες είναι ένα πρόβλημα στην αυτοματοποίηση και μια μεγάλη μέριμνα για τις βιβλιοθήκες, τους σχεδιαστές δικτύων, τις κυβερνητικές υπηρεσίες, τις τράπεζες και τις πολυεθνικές εταιρείες.

Τα πληκτρολόγια των υπολογιστών σχεδιάστηκαν αρχικά για το σετ χαρακτήρων της Αγγλικής γλώσσας, περιλαμβάνοντας μόνο 26 γράμματα, 10 αριθμούς και λίγα ειδικά σύμβολα. Καθώς οι ποικιλίες στο τυπικό πληκτρολόγιο της Αγγλικής γλώσσας χρησιμοποιούνται να δημιουργήσουν λέξεις σε περισσότερες γλώσσες, αυτό οδηγεί είτε σε 1) απώλεια δεδομένων είτε 2) σε κωδικοποίηση χαρακτήρων σε μια συγκεκριμένη γλώσσα ή συγκεκριμένη διάταξη εφαρμογής που δεν είναι μεταφερόμενη σε άλλα συστήματα.

## **ΜΕΤΑΓΡΑΦΗ ΚΑΙ ΑΛΛΕΣ ΜΟΡΦΕΣ ΑΠΩΛΕΙΑΣ ΔΕΔΟΜΕΝΩΝ**

Μια πρώτη βασική κατευθυντήρια γραμμή ως προς τη μεταγραφή είναι ότι οι γλώσσες γραμμένες σε μη Ρωμαϊκά αλφάβητα (Γιαπωνέζικο, Αραβικό, Κινέζικο, Κορεάτικο, Περσικό, Εβραϊκό) μεταγράφονται σε Ρωμαϊκούς χαρακτήρες σε πολλές εφαρμογές.

Η μεταγραφή αντιστοιχεί χαρακτήρες από μια γλώσσα σε μια άλλη. Δεν μεταφράζει νόημα. Το αποτέλεσμα είναι να χάνονται αξιολογώτα δεδομένα κατά την μεταγραφή. Η μέθοδος μπορεί να είναι μη αναστρέψιμη καθώς πολλές παραλλαγές

συμβαίνουν σε πολλαπλά μεταγραφικά συστήματα για μια δεδομένη γλώσσα [παράδειγμα Eking έναντι του Beijing, Mao Tse-tung έναντι του Mao Zedong (Chinese), Tchaikovsky έναντι Chaikovskii (Russian)]. Το χειρίστο θα είναι οι μορφές αυτές να είναι άγνωστες στους ομιλητές των γλωσσών αυτών και έτσι να έχουμε ολοκληρωτικές απώλειες δεδομένων.

Ένα άλλο εμπόδιο παρουσιάζεται στις γλώσσες που γράφονται σε επεκτάσεις του Ρωμαϊκού σετ χαρακτήρων όπως Γαλλικά, Ισπανικά, Γερμανικά, Ουγγαρέζικα, Τσέχικα και Πολωνέζικα. Αυτές δεν διατηρούνται σε πλήρη μορφή σε ορισμένες εφαρμογές και κάποια διακριτικά γνωρίσματα ή σημεία (όπως accents, umlauts= διαλυτικά και άλλα συγκεκριμένα σημεία-διακριτικά γνωρίσματα) που διακρίνουν τους χαρακτήρες παραλείπονται.

Αυτές οι μορφές απώλειας δεδομένων είναι παρόμοιες με την απώλεια που συνεπάγεται η συμπίεση εικόνων, κατά την οποία δεδομένα πετιούνται για να εξοικονομηθούν κόσμη αποθήκευσης και χρόνου μετάδοσης καθώς διατηρούν μια αποδεκτή αναπαραγωγή της εικόνας.

Κάθε είδους απώλεια δεδομένων προκαλεί προβλήματα στις ψηφιακές βιβλιοθήκες. Ποικίλες μορφές λέξεων δεν θα ταιριάζουν και δεν θα αντιστοιχούνται απόλυτα και μη ολοκληρωμένες λέξεις δεν θα ανταλλάσσονται με ψηφιακές βιβλιοθήκες που χρησιμοποιούν ολοκληρωμένες μορφές

Μια άλλη προσέγγιση του Clews John, προέδρου της επιτροπής ISO/TC46/SC2 για την μετατροπή των γραπτών γλωσσών (<http://www.dlib.org/dlib/march97/sesame/03clews.html>) παρουσιάζει το θέμα αυτό απλούστερο και περισσότερο αισιόδοξο. Αναφέρει δηλαδή πως υπάρχουν τρεις τύποι αλφαβήτων, τα ιδεογραφικά αλφάβητα, τα αλφάβητα της Νότιας Ασίας και τα προερχόμενα από το Φοινικικό. Αν μπορούν να ξεπεραστούν τα εμπόδια αυτών των τριών αλφαβήτων (για τα οποία έχει ήδη συντελεστεί αρκετή πρόοδος), τότε θα μπορούμε να έχουμε πρόσβαση στο μεγαλύτερο μέρος του παγκόσμιου πολιτισμού.

## **ΜΟΝΟΓΛΩΣΣΙΚΑ, ΠΟΛΥΓΛΩΣΣΙΚΑ ΚΑΙ ΠΑΓΚΟΣΜΙΑ ΣΕΤ ΧΑΡΑΚΤΗΡΩΝ**

Στην παρούσα πραγματικότητα πολλά πρότυπα και πρακτικές χρησιμοποιούνται για την κωδικοποίηση χαρακτήρων. Κάποια είναι περισσότερο εξειδικευμένα, απευθύνονται δηλαδή αποκλειστικά στο Λατινικό ή Κυριλλικό αλφάβητο και άλλα παγκόσμια, υποστηρίζουν δηλαδή τις περισσότερες από τις παγκόσμια γραπτές γλώσσες. Το πρόβλημα εδώ έγκειται στην ανταλλαγή δεδομένων μεταξύ ψηφιακών βιβλιοθηκών που υποστηρίζουν διαφορετικά σχήματα κωδικοποίησης χαρακτήρων.

Αναλυτικότερα, μονογλωσσικές ψηφιακές βιβλιοθήκες που χρησιμοποιούν όλες το ίδιο σχήμα κωδικοποίησης, όπως το ASCII (American Standard Code for Information Interchange) το οποίο περιλαμβάνει τις περισσότερες Ευρωπαϊκές γλώσσες, θα μπορούν να ανταλλάσσουν άμεσα δεδομένα. Στις μονογλωσσικές ψηφιακές βιβλιοθήκες που χρησιμοποιούν διαφορετικά σχήματα-πρότυπα παρουσιάζεται το πρόβλημα. Για παράδειγμα, ψηφιακές βιβλιοθήκες που κάνουν χρήση του ASCII, πιθανότατα δεν θα μπορούν να ανταλλάξουν δεδομένα με άλλες που κάνουν χρήση

διαφορετικού πολυγλωσσικού σετ χαρακτήρων. Χαρακτήρες παραγόμενοι από ένα συγκεκριμένο πληκτρολόγιο για ένα συγκεκριμένο σύστημα κωδικοποίησης μπορεί να μην ταιριάζουν με χαρακτήρες που χρησιμοποιούνται από άλλο σύστημα κωδικοποίησης, όπως αυτό της Αμερικανικής Ένωσης Βιβλιοθηκών (ALA) που χρησιμοποιείται στις Ηνωμένες Πολιτείες. Επίσης, χαρακτήρες με διακριτικά σημεία-γνωρίσματα μπορεί να εμφανίζονται λανθασμένα ή να μην εμφανίζονται καθόλου.

Προβλέψεις για πολυγλωσσικότητα και κωδικοποίηση χαρακτήρων πρωτοαναφέρθηκαν στο HTTP 1.1. (HyperText Transfer Protocol) (RFC- Request for Comments 2068). Στη συνέχεια, το RFC 2070 προσέθεσε τα απαραίτητα χαρακτηριστικά στο HTML (Hypertext Markup Language) για την περιγραφή πολυγλωσσικών τεκμηρίων, που οδηγούν στην επέκταση του HTML 2.0 (RFC 1886). Στη συνέχεια ακολουθεί το Παγκόσμιο Σετ Χαρακτήρων (Universal Character Set - UCS) του ISO 10646:1993, το οποίο είναι σχεδόν ίδιο με την έκδοση 1.1. ISO 10646/Unicode.

Μετά από χρόνια διεθνούς συζήτησης πάνω στο θέμα, το Unicode εμφανίζεται ως το προτιμώμενο πρότυπο να υποστηρίξει τις περισσότερες γραπτές γλώσσες του κόσμου και ως ο νικητής σε μια μακρά μάχη προτύπων. Το Unicode συγχωνεύτηκε με το ISO 10646 επειδή ήταν ευκολότερο να εφαρμοστεί και συνεκδοχικά να υιοθετηθεί πιο εύκολα. Σημειώτέο πως το Unicode κωδικοποιεί αλφάβητα και όχι γλώσσες, δηλαδή χαρακτήρες και όχι ανάγλυφα. Διατηρείται από το Unicode Consortium (μη κερδοσκοπικός οργανισμός) και η πρώτη έκδοση του ήταν η Unicode 1.00 (Οκτώβριος του 1991) και η τελευταία έκδοση του είναι η 4.0.0 (Απρίλιος του 2003), δηλαδή συνολικά αριθμούνται 15 εκδόσεις. Όταν το Unicode Consortium οραματιζόταν το Unicode, ήθελαν να είναι παγκόσμιο (universal), αποτελεσματικό (efficient), ομοιόμορφο (uniform) και σαφές (unambiguous). Απαιτεί 16 bits για να αποθηκεύσει κάθε χαρακτήρα- τα διπλά από όσα το ASCII- όπου τα 16 bits επιτρέπουν την κωδικοποίηση περισσότερων από 65.000 χαρακτήρων.

Ο βασικός στόχος του Unicode είναι να καλύψει όλα τα αλφάβητα του κόσμου, ιστορικά και μοντέρνα. Το Unicode απαιτεί το μισό χώρο από την προγενέστερη έκδοση του ISO 10646 (32 bits), το ανταγωνιστικό και πιο συνεκτικό παγκόσμιο σετ χαρακτήρων. Προς το παρόν καλύπτει τις κύριες γραπτές γλώσσες της Αμερικής, Ευρώπης, Μέσης Ανατολής, Αφρικής, Ινδίας και Ασίας. Γι' αυτό υπάρχει και το σλόγκαν «Όταν ο κόσμος θέλει να μιλήσει, μιλάει Unicode/ When the world wants to talk, it speaks Unicode».

Αναλυτικότερα, το Unicode προτείνει έναν και μοναδικό αριθμό για κάθε χαρακτήρα, ανεξάρτητα από το λειτουργικό σύστημα, το λογισμικό και τη γλώσσα. Το πρότυπο αυτό αποτελείται από χαρακτήρες, γραπτά δηλαδή συστατικά (αλφάβητα, ιδεογράμματα, σειρές χαρακτήρων συλλαβών, σημεία στίξης, μαθηματικούς τελεστές, κ.ά) που αντιπροσωπεύονται με αριθμητικές αξίες (numerical values). Δηλαδή, για την αντιπροσώπευση ενός χαρακτήρα έχουμε τη μορφή U+yyyy, όπου το U+ είναι η αξία του κωδικού Unicode (Unicode code value) και τα yyyy αποτελούν ένα τετραψήφιο δεκαεξαδικό αριθμό του χαρακτήρα που κωδικοποιούμε. Για παράδειγμα, το Λατινικό κεφαλαίο γράμμα «Α» παίρνει το κωδικό χαρακτήρα U+0041 στο Unicode. Αυτόν τον αριθμό αποθηκεύει ο υπολογιστής και αυτόν χρησιμοποιεί για να αναφερθεί στο «Α». Αποθηκεύει γράμματα και άλλους

χαρακτήρες αντιστοιχώντας στον καθένα τους από έναν αριθμό, όπου μία τέτοια αντιστοιχία την ονομάζουμε κωδικοσελίδα. Το ελληνικό αλφάβητο εντοπίζεται στις θέσεις U+0370 έως U+03FF.

Χάρης στο Unicode ένα και μόνο προϊόν μπορεί να επικοινωνεί με διάφορα λειτουργικά συστήματα, σε διάφορες γλώσσες και χώρες, χωρίς την ανάγκη επαναπρογραμματισμού. Γίνεται έτσι δυνατή η μεταφορά δεδομένων ανάμεσα σε πλήθος διαφορετικών συστημάτων δίχως κίνδυνο αλλοίωσης. Άρα, ένα παγκόσμιο σετ χαρακτήρων σαν το Unicode προσφέρει υποσχέσεις για τη λύση της ανταλλαγής δεδομένων και κατανεμημένες ψηφιακές βιβλιοθήκες. Εάν τα δεδομένα σε όλες τις γραπτές γλώσσες κωδικοποιούνται στο ίδιο σχήμα, τότε μπορούν να ανταλλαχθούν μεταξύ μονογλωσσικών και πολυγλωσσικών ψηφιακών βιβλιοθηκών.

Κάθε λύση που παρουσιάζεται τόσο απλή, πιθανότατα να είναι αλλά πάντοτε υπάρχουν και κάποιοι περιορισμοί ή όρια. Δηλαδή το Unicode θέτει υψηλότερες απαιτήσεις σε θέματα αποθήκευσης και αυτό θα μπορούσε να επηρεάσει το χρόνο μετάδοσης μεγάλων αποστάσεων. Επίσης, επειδή αναπτύσσεται πολύ γρήγορα και συμπληρώνεται συνεχώς, έχει και περισσότερες απαιτήσεις, δηλαδή πρόσφατα λειτουργικά συστήματα (Windows 2000, XP) ή πρόσφατους φυλλομετρητές (Mozilla, Netscape).

Παρόλα αυτά, η επιφυλακτικότητα υιοθέτησης του Unicode φαίνεται να φθίνει καθώς τα πλεονεκτήματα της παγκόσμιας γλωσσικής διαλειτουργικότητας υπερέχουν. Το πρότυπο αυτό το έχουν ασπασθεί κορυφαίοι παράγοντες του χώρου των λογισμικών όπως οι: Apple, HP, IBM, JustSystem, Microsoft, Oracle, SAP, Sun, Sybase, Unisys και πολλοί άλλοι. Η εμφάνιση της κωδικοσελίδας Unicode, και η διαθεσιμότητα εργαλείων που να την υποστηρίζουν είναι από τις σημαντικότερες εξελίξεις της πρόσφατης τεχνολογίας λογισμικών. Γι' αυτό οι περισσότεροι πωλητές υλικού και λογισμικού ξεκινούν να το υποστηρίζουν. Το γεγονός ότι ο Netscape έχει αποφασίσει να σχεδιάσει προϊόντα που να υποστηρίζουν το Unicode θα διαδραματίσει αξιοσημείωτο ρόλο και αρκετά λειτουργικά συστήματα θα αρχίσουν να το υιοθετούν ως εσωτερικό κωδικό χαρακτήρων (internal character code).

## **ΤΡΕΧΟΥΣΕΣ ΕΦΑΡΜΟΓΕΣ ΚΑΙ ΧΡΗΣΕΙΣ-ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΤΟ ΔΙΕΘΝΗ ΧΩΡΟ**

Τα τελευταία χρόνια, έχουμε ένα σημαντικό ανερχόμενο ενδιαφέρον σε όλο τον κόσμο στην ανάπτυξη συστημάτων ψηφιακών βιβλιοθηκών και όχι μόνο σε Αγγλόφωνες περιοχές. Και η Ευρώπη και η Ασία συμμετέχουν ενεργά στο να κτίσουν δικά τους μεγάλα οργανωμένα και κατανεμημένα αποθετήρια γνώσης (repositories of knowledge). Αυτό μαρτυρείται από πρόσφατη αναφορά των Ercim News, του newsletter του European Research Consortium for Informatics and Mathematics ([www.ercim.org](http://www.ercim.org)), το οποίο ήταν αφιερωμένο σε ψηφιακές βιβλιοθήκες και ανερχόμενες πρωτοβουλίες σε Ευρώπη αλλά και Κίνα και Ιαπωνία. Γίνεται επίσης ακόμη περισσότερο εμφανές και από τον αριθμό των διεθνών συνεδρίων που διοργανώνονται σε άλλα μέρη πλην των Ηνωμένων Πολιτειών, για παράδειγμα σημαντικά συνέδρια ψηφιακών βιβλιοθηκών που έχουν λάβει χώρα στην Ιαπωνία.

Επίσης, ο αριθμός των συμμετεχόντων σε συνέδρια όπως αυτά του CLEF (Cross-Language Evaluation Forum) κάθε χρόνο αυξάνεται. Δειγματικά, οι συμμετέχοντες στο CLEF 2003 ανήλθαν στα 43 γκρουπ (10 από Νότια Αμερική, 30 από Ευρώπη και 3 από Ασία).

Η διεθνής βιβλιοθηκονομική κοινότητα άρχισε να αναπτύσσει μεγάλες, πολυγλωσσικές ψηφιακές βιβλιοθήκες στα 1960. Οι βιβλιοθήκες πάντοτε είχαν λάβει υπόψη τη διατήρηση και παροχή πρόσβασης στην πληροφορία. Διαχειρίζονται περιεχόμενο σε πολλές γλώσσες και συνεργάζονται ως μια διεθνής κοινότητα που ανταλλάσσει δεδομένα σε ψηφιακή μορφή. Δεν εντυπωσιάζει που οι βιβλιοθήκες ήταν από τα πρώτα ινστιτούτα που κλήθηκαν να λύσουν το πολυγλωσσικό πρόβλημα. Τα περασμένα τριάντα χρόνια, οι βιβλιοθήκες είχαν δημιουργήσει τεράστιες αποθήκες ψηφιακών δεδομένων. Από εδώ και στο εξής, οι βιβλιοθήκες έχουν τη δύναμη και τη γνώση να επηρεάζουν μελλοντικές αναπτύξεις σε πρότυπα για σета χαρακτήρων και άλλους παράγοντες στην ανταλλαγή δεδομένων.

Ο κόσμος των βιβλιοθηκών αλλάζει καθώς νέες περιοχές του κόσμου έρχονται online. Η Ευρωπαϊκή Ένωση προωθεί το Unicode και χρηματοδοτεί ερευνητικά έργα (projects) τα οποία θα υποστηρίξουν την εφαρμογή του στην αυτοματοποίηση των βιβλιοθηκών. Η αυτοματοποίηση στην Κεντρική & Ανατολική Ευρώπη έχει προχωρήσει γρήγορα από το 1990. Μία έρευνα σε ερευνητικές βιβλιοθήκες σε έξι χώρες της Κεντρικής & Ανατολικής Ευρώπης, κάθε μία με τη δική της εθνική γλώσσα και σета χαρακτήρων, δηλώνει ότι χρησιμοποιείται μια ποικιλία συστημάτων κωδικοποίησης. Στα τέλη του 1994, οι περισσότερες από τις μισές χρησιμοποιούσαν το ASCII, μία χρησιμοποιούσε το Unicode και οι υπόλοιπες ένα εθνικό ή συγκεκριμένο σχήμα. Καμία δεν χρησιμοποιούσε το σета της Αμερικανικής Ένωσης Βιβλιοθηκών. Οι βιβλιοθήκες σε αυτές τις χώρες είναι υπεύθυνες για τη διατήρηση της πολιτιστικής κληρονομιάς που εμφανίζεται σε έντυπη μορφή και αυτό απαιτεί η γλώσσα τους να διατηρείται σε πλήρη μορφή.

Πρότυπα για τη δομή και τα σета χαρακτήρων εγκαταστάθηκαν πολύ πριν το Internet δημιουργηθεί, πόσο μάλλον πριν το Unicode. Ξεκινώντας από φορείς διεθνούς εμβέλειας, αναφέρουμε το OCLC, τη Βιβλιοθήκη του Κογκρέσου και το RLIN και την τριμερή αυτή συνεργασία. Το OCLC ([www.oclc.org](http://www.oclc.org)), εξυπηρετεί πάνω από 17.000 βιβλιοθήκες σε 52 χώρες και περιέχει πάνω από 30.000.000 βιβλιογραφικές εγγραφές σε περισσότερες από 370 γλώσσες. Το OCLC χρησιμοποιεί το πρότυπο σета χαρακτήρων της Αμερικανικής Ένωσης Βιβλιοθηκών, που επεκτείνει το πληκτρολόγιο Αγγλικής γλώσσας.

Ακολούθως, η βιβλιοθήκη του Κογκρέσου (<http://www.loc.gov>), που στέλνει τις εγγραφές της σε ψηφιακή μορφή στο OCLC, στο RLIN και σε άλλους συνεργαζομένους φορείς κάνει πρωτότυπη καταλογογράφηση μη Ρωμαϊκών αλφαβήτων. Το RLIN (Research Libraries Information Network-<http://www.rlig.org/rlin.html>) καινοτόμησε ξεκινώντας κωδικοποίηση των μη Ρωμαϊκών γλωσσών στην πρωτότυπη μορφή τους για βιβλιογραφικές εγγραφές, χρησιμοποιώντας συγκεκριμένα γραφικά πρότυπα. Εγγραφές κωδικοποιημένες σε πλήρη γραφή ανταλλάσσονται μεταξύ της Βιβλιοθήκης του Κογκρέσου, του RLIN, του OCLC που συνεργάζονται και για άλλες βιβλιογραφικές χρησιμότητες στις Η.Π.Α.

Καθώς οι βιβλιοθήκες, τα αρχεία, μουσεία και άλλοι πολιτιστικοί οργανισμοί στον κόσμο ενημερώνονται για τη ανάγκη του να διατηρούν ψηφιακά δεδομένα σε αρχειακές μορφές, η αντιπροσώπευση σει χαρακτήρων γίνεται τόσο πολιτικό όσο και τεχνικό θέμα. Πολλές υπηρεσίες υποστηρίζουν ερευνητικά έργα για να επιβεβαιώσουν τη διατήρηση των βιβλιογραφικών δεδομένων σε ψηφιακές μορφές και ότι μπορούν να ανταλλαχθούν. Μέσα σε αυτές τις υπηρεσίες περιλαμβάνεται η Επιτροπή των Ευρωπαϊκών Κοινοτήτων, η IFLA ([www.ifla.org](http://www.ifla.org)), το Mellon Foundation (το Mellon Foundation παρέχει επιχορηγήσεις στους τομείς της εκπαίδευσης, των μουσείων, της τέχνης, του περιβάλλοντος, των πληθυσμών και των δημοσίων σχέσεων - [www.mellon.org](http://www.mellon.org)) και το Soros Foundation Open Society Institute Regional Library Program ([www.soros.org](http://www.soros.org)). [Το Open Society Institute είναι ιδιωτικός οργανισμός με παραρτήματα σε περισσότερες από 50 χώρες και αναλαμβάνει πρωτοβουλίες σε θεματικά πεδία παρόμοια με αυτά του Mellon Foundation]

## ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΤΟΝ ΕΛΛΗΝΙΚΟ ΧΩΡΟ

Στις πηγές και στα άρθρα τα οποία μελετήθηκαν η πλειονότητα των γλωσσών, που αποτελούν αντικείμενο μελέτης στις πολυγλωσσικές ψηφιακές βιβλιοθήκες, αφορούν ομάδες γλωσσών όπως Αγγλική, Γερμανική, Γαλλική, Ιταλική, Ολλανδική, Σουηδική και Φιλανδική. Για την ελληνική γλώσσα, βρέθηκαν ελάχιστες αναφορές στην έκταση κάποιων γραμμών, από τις οποίες αντλούμε πληροφορίες όπως:

- Για το Ελληνικό αλφάβητο (το οποίο προήλθε από το Φοινικικό και υιοθετήθηκε από Ευρωπαϊκά αλφάβητα όπως το Λατινικό και το Κυριλλικό) η τεχνική επιτροπή ISO/TC46 και η υποεπιτροπή SC2 έχει αναπτύξει το πρότυπο ISO 843:1997 (Information & Documentation-Conversion of Greek characters into Latin characters) για τη μεταγραφή των Ελληνικών χαρακτήρων σε Λατινικούς. Σύμφωνα με πληροφορίες της Βιβλιοθήκης και Κέντρου Πληροφόρησης του ΕΛΟΤ υπάρχει και σε ελληνική μετάφραση. Επίσης, μεταξύ των διαφόρων working groups που περιλαμβάνει η ISO/TC46/SC42 υπάρχει και το working group WG5: Μεταγραφή Ελληνικών.
- Είναι αποδεκτό πως η μεταγραφή μπορεί να λειτουργήσει καλύτερα για φωνητικά αλφάβητα και γλώσσες, όπως το Ελληνικό (<http://www.dlib.org/dlib/march97/sesame/03clews.html>) – πρόβλεψη αισιόδοξη για τη μεταγραφή των Ελληνικών.
- Μια άλλη προσπάθεια στον ελληνικό χώρο ήταν το HELEN (<http://alcyone.cc.uch.gr/~kosmas/Helen>), ένα project 2 χρόνων υπό την αιγίδα του CEC Telematic Systems in Areas of General Interest - Libraries. Το ερευνητικό αυτό έργο άρχισε τον Ιανουάριο του 1993 με ημερομηνία λήξης τον Ιανουάριο του 1995. Συνεργάτες σε αυτό ήταν το Πανεπιστήμιο του Bradford ως συντονιστής (UK), το Kings Κολέγιο του Λονδίνου (UK), το Πανεπιστήμιο Κρήτης, το Κολέγιο Αθηνών και το Κέντρο Νεοελληνικών Ερευνών. Το πρόγραμμα αυτό είχε ως αντικείμενο έρευνας μια ομάδα θεμάτων που είχαν να κάνουν με την μετατροπή μεταξύ Λατινικών

και Ελληνικών αλφαβήτων και την παρουσίαση του Ελληνικού σετ χαρακτήρων. Στόχος ήταν να κάνει το υλικό της Ελληνικής γλώσσας ευρύτερα διαθέσιμο στην Ευρωπαϊκή Ένωση και εφικτή την ανάπτυξη ενός προηγμένου λογισμικού μεταγραφής που θα αντιμετώπιζε τα προβλήματα της χρήσης διαφορετικών σχημάτων μεταγραφής (εξετάστηκαν ποικίλα σετ χαρακτήρων και λύσεις όπως το ISO 10646 και το Unicode). Στα πλαίσια του ερευνητικού έργου εξετάστηκαν ευρωπαϊκές βιβλιοθήκες που έχουν υλικό σε Ελληνική γλώσσα, τα σχήματα μεταγραφής που χρησιμοποιούν και οι γενικότερες τακτικές τους. Δεν εκλείπουν βέβαια και παραδείγματα όπου σε μια βιβλιοθήκη εντοπίστηκαν 6 διαφορετικές μορφές του ονόματος Αριστοτέλη ή ο μεγαλύτερος αριθμός έφθανε τις 19 εκδοχές του ονόματος Καβάφη.

Τίθεται βέβαια το ερώτημα γιατί εφόσον είναι δυνατό οι βιβλιοθήκες να προμηθευτούν υπολογιστές που να υποστηρίζουν Ελληνικούς χαρακτήρες χρησιμοποιώντας βέβαια το σωστό σετ χαρακτήρων, εξακολουθεί να υπάρχει το πρόβλημα της μεταγραφής και της μετάφρασης. Η απάντηση έγκειται στο ότι λίγες βιβλιοθήκες κατέχουν μεγάλες συλλογές υλικού στα Ελληνικά άρα λίγες είναι και θα είναι οι πρωτοβουλίες για την επίλυση αυτού του προβλήματος.

Σημειωτέο, πως η Ελλάδα δεν αναφέρθηκε σε κάποιο άλλο σημείο να είναι μέλος σε κάποια άλλα έργα ή να συμμετέχει σε εργαστήρια (workshops) όπως αυτά του CLEF (Cross-Evaluation Forum).

## **ΠΕΡΙΟΡΙΣΜΟΙ, ΟΡΙΑ ΚΑΙ ΕΜΠΟΔΙΑ**

Οι περιορισμοί και τα εμπόδια για την επίτευξη πολυγλωσσικών δεδομένων απαριθμούνται ως εξής:

- 1) Επικέντρωση ως σήμερα των προσπαθειών σε χώρες όπου η Αγγλική γλώσσα είναι αποδεκτή ως προεπιλεγμένη γλώσσα (default) π.χ. Η.Π.Α.
- 2) Οι πληροφοριακές υποδομές και υπηρεσίες τα τελευταία χρόνια έχουν προσπαθήσει να λύσουν τα προβλήματα πολυγλωσσικότητας, με αποτέλεσμα να μην υπάρχουν ακόμη μακρόχρονες και δοκιμασμένες τεχνικές και εργαλεία
- 3) Η ύπαρξη μαζικού όγκου κειμένων χιλιάδων χρόνων και ο πλούτος της ανθρώπινης επικοινωνίας καθιστά το πρόβλημα ακόμη πιο δυσεπίλυτο
- 4) Η μη αποδοτικότητα των μεταφράσεων, η ύπαρξη πολλαπλών νοημάτων, οι αλλοιώσεις και απώλειες δεδομένων
- 5) Διαφορές αναζητητών και δημιουργών
- 6) Διλήμματα των σχεδιαστών (ποια πρότυπα και σετ χαρακτήρων να επιλέξουν, πού πρέπει να δοθεί περισσότερη έμφαση)

- 7) Απίστευτη ποικιλία υλικού και λογισμικού
- 8) Έλλειψη διαλειτουργικότητας
- 9) Κάθε ψηφιακή βιβλιοθήκη υιοθετεί εφαρμογές και εργαλεία σύμφωνα με τις ανάγκες της, τη γλώσσα της και τον πολιτισμό της, καθιστώντας δύσκολη την επίτευξη μιας παγκόσμιας ψηφιακής βιβλιοθήκης
- 10) Οι ψηφιακές βιβλιοθήκες σε διαφορετικές χώρες στοχεύουν σε διαφορετικούς χρήστες και διαφορετικά σετ γλωσσών. Για παράδειγμα, στη Βραζιλία ένα καλό παράδειγμα σετ γλωσσών είναι τα Πορτογαλέζικα, τα Ισπανικά και τα Αγγλικά
- 11) Στην πλειοψηφία των συνεδρίων και εργαστηρίων οι αξιολογήσεις, μελέτες και έρευνες γίνονται για τα σύνθητα και δημοφιλή ζευγάρια γλωσσών (language pairs) όπως: Ιταλικά-Ισπανικά, κ.ά.
- 12) Καμία μηχανή δεν ευρετηριάζει όλο το περιεχόμενο στο Internet
- 13) Κάθε ψηφιακή βιβλιοθήκη έχει μια δική της αρχιτεκτονική, γεγονός που συνεπάγεται ετερογενή σετ μηχανών αναζήτησης.
- 14) Η διαδικασία ένταξης των αλφαβήτων εκείνων που είναι λιγότερο γνωστά και χρησιμοποιούμενα στο Unicode θα είναι πολύ αργή και ελάχιστα χρηματοδοτούμενη. Ήδη μέλη του Unicode Consortium έχουν δείξει ελάχιστο ενδιαφέρον για την αποπεράτωση κωδικοποίησης τέτοιων αλφαβήτων.
- 15) Τέλος, οι θύρες ένταξης νέων αλφάβητων στο Unicode 5.0 θα είναι πιθανά ανοιχτές μέχρι το 2004. Από εκεί και έπειτα, θα είναι απίστευτα δύσκολο να ενταχθούν αλφάβητα στο Unicode γιατί το ενδιαφέρον για αλφάβητα που αποτελούν την μειονότητα (minority languages) θα είναι ελάχιστο από οργανισμούς και άλλους φορείς. Άρα, αναπόφευκτα αυτό συνεπάγεται και δυσκολία επιβίωσης αυτών των αλφαβήτων.

## **ΚΡΙΤΙΚΗ ΚΑΙ ΣΧΟΛΙΑ**

Πρέπει να αναγνωριστεί ότι τα τελευταία χρόνια έχουν καταβληθεί σημαντικές προσπάθειες πάνω στο αντικείμενο των πολυγλωσσικών ψηφιακών βιβλιοθηκών. Οι σχεδιαστές εφαρμογών δικτύων είναι περισσότερο ενημερωμένοι για διαλειτουργικότητα, φορητότητα και ανταλλαγή δεδομένων από ότι στο παρελθόν. Αλλά το θέμα αυτό πρέπει να κερδίσει και την προσοχή ενός ευρύτερου κοινού και όχι μόνο της τεχνικής ελίτ που ασχολείται με αυτό χρόνια. Πρέπει δημόσιοι και ιδιωτικοί φορείς να χρηματοδοτούν προτάσεις για κωδικοποίηση όλο και περισσότερων αλφαβήτων. Είναι ανάγκη όλοι να δραστηριοποιηθούν διότι η λύση δεν μπορεί να έρθει νυχθημερόν. Χρειάζεται όμως ακόμη περισσότερη εκπαίδευση και εμπειρία πάνω στις εφαρμογές πολυγλωσσικών ψηφιακών βιβλιοθηκών.

Το Unicode εμφανίζεται να είναι η απάντηση για τις νέες τεχνολογίες, παρόλα αυτά οι σχεδιαστές πρέπει να ζυγίσουν παράγοντες όπως το ποσό των δεδομένων που

υπάρχει στην τρέχουσα φάση σε άλλα σχήματα, τα πρότυπα σε χρήση από άλλα συστήματα με τα οποία πρέπει να ανταλλάσσουν τακτικά, και τον ρυθμό με τον οποίο γίνονται όλα αυτά. Το συντομότερο που η κοινότητα των ψηφιακών βιβλιοθηκών αναμειχθεί σε τέτοια θέματα, το συντομότερο θα βρεθεί μια πολυπολιτισμική και πολυγλωσσική λύση στην ανταλλαγή δεδομένων σε όλες τις γραπτές γλώσσες.

Στο μέλλον βέβαια διαφαίνονται και άλλες προκλήσεις, πλην αυτής της πολυγλωσσικής ανάκτησης πληροφοριών. Μια τέτοια είναι αυτή της πολυγλωσσικής ανάκτησης λόγου και ομιλίας και γενικότερα της πρόσβασης σε πολυγλωσσικές πληροφορίες που είναι σε μορφή άλλη πλην κειμένου και τέλος της πολυγλωσσικής ανακάλυψης και πρόσβασης της πληροφορίας (Multi-lingual Information Discovery and Access- MIDAS).

Επίσης, εκφράζεται η προσδοκία η Ελλάδα να διαδραματίσει περισσότερο ενεργό ρόλο στον τομέα των πολυγλωσσικών ψηφιακών βιβλιοθηκών και το ελληνικό αλφάβητο να αποτελέσει τον πυρήνα έρευνας και μελέτης στο μέλλον αλλά και αντικείμενο διεθνών συνεδρίων αξιολόγησης Ευρωπαϊκών γλωσσών.

## ΣΗΜΑΝΤΙΚΕΣ ΠΗΓΕΣ

Σημαντική πηγή πληροφόρησης αποτελεί το ευρωπαϊκό έργο CLEF (Cross-Language Evaluation Forum- <http://www.clef-campaign.org>) του προγράμματος Information Society Technologies (IST). Το πρόγραμμα CLEF αποτελεί παράλληλη δράση του ευρωπαϊκού ερευνητικού ανθρωποδικτύου για ψηφιακές βιβλιοθήκες DELOS (<http://delos-noe.iei.pi.cnr.it/>) και υπάρχει στενή συνεργασία μεταξύ των φορέων που συμμετέχουν στα δύο έργα. Το CLEF ξεκίνησε τη δράση του το 2000 και οργανώνει ετήσιες καμπάνιες και εργαστήρια (workshops) αξιολόγησης:

- § CLEF 2000 (21-22/09/00, Πορτογαλία)
- § CLEF 2001 (3-4/09/01, Γερμανία)
- § CLEF 2002 (19-20/09/02)
- § CLEF 2003 (21-22/08/03, Norway)
- § CLEF 2004 (16-17/09/04, Bath, UK)

Επίσης, προσπαθεί να επεκτείνει τη δράση του και για γλώσσες που δεν χρησιμοποιούνται ευρέως, για παράδειγμα αυτές της Ασίας.

Ακολούθως, αξιόλογες προσπάθειες στο χώρο καταβάλλουν οι κάτωθι φορείς:

- § Η Επιτροπή **ISO/TC46/SC2** η οποία έχει αναπτύξει αρκετά ISO πρότυπα, όπως το ελληνικό το οποίο προαναφέρθηκε, το ISO 9 (Κυριλλικό), ISO 233 (Αραβικό), ISO 3602 (Γιαπωνέζικο) κ.ά.
- § **Winter: Web Internationalization & Multilinguism** (<http://www.w3.org/pub/WWW/International/>): Αποστολή του είναι να εξασφαλίσει ότι τα σχήματα και πρωτόκολλα του WWW χρησιμοποιούνται παγκοσμίως σε όλες τις γλώσσες και όλα τα γραπτά συστήματα.

- § **TREC (Text Retrieval Conference)** (<http://trec.nist.gov>): Στόχο έχει να προάγει την έρευνα στην ανάκτηση πληροφοριών από μεγάλες συλλογές δεδομένων.
- § **NTCIR (NII-NACSIS Test Collection for IR Systems – <http://research.nii.ac.jp/ntcir/>)**: Είναι μια σειρά από εργαστήρια αξιολόγησης που προάγουν την έρευνα στις τεχνολογίες πρόσβασης στην πληροφορία περιλαμβάνοντας πολυγλωσσική ανάκτηση.
- § **ELRA (Evaluations and Languages Resources Distribution Association-<http://www.elra.info>)**: Ασχολείται με την προώθηση, υποστήριξη, προτυποποίηση και βελτίωση γλωσσικών ζητημάτων.
- § **LDC (Linguistic Data Consortium- <http://www ldc.upenn.edu>)**: Υποστηρίζει την εκπαίδευση, έρευνα και τεχνολογία αναφορικά με τις γλώσσες και την ανάπτυξη δεδομένων, εργαλείων και προτύπων για αυτές.

Σχετικά ερευνητικά προγράμματα είναι τα κάτωθι:

- § **HLT Central: Human Language Technologies on the Web (European Commission)** (<http://hltcentral.org>) : Αποτελεί μια on-line πηγή πληροφοριών για τεχνολογίες ανθρώπινης γλώσσας
- § **TIDES: Translignal Information Detection, Extraction and Summarization (DARPA)**

## ΕΡΩΤΗΜΑΤΑ ΚΑΙ ΠΡΟΒΛΗΜΑΤΙΣΜΟΙ

Αυτό που τίθεται ως ερώτημα και προβληματισμός είναι ότι όλες οι προαναφερόμενες μέθοδοι και τεχνικές βρίσκονται σε πιλοτικό και πειραματικό στάδιο. Δεν υπάρχουν ακόμη καθολικά αποδεκτές, παραδεδεγμένες και παγκοσμίως εφαρμόσιμες τεχνικές και συνεκδοχικά δεν μπορεί να προβλεφθεί ο ρυθμός υιοθέτησης τους. Επίσης, τίθεται το ερώτημα πόσο ικανοποιητικά αποτελέσματα αποφέρουν όλες αυτές οι λύσεις.

Περαιτέρω, θα πρέπει να απασχολήσει όλους τους επιστήμονες του χώρου της πληροφόρησης η εύρεση και καθιέρωση νέων γλωσσικών εργαλείων για τον ανερχόμενο κλάδο της πολυγλωσσικής ανάκτησης πληροφοριών.

## ΠΕΡΙΛΗΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Τεράστιοι όγκοι δεδομένων σε πολλές γλώσσες διατίθενται on-line, στις οποίες οι χρήστες από όλο τον κόσμο απαιτούν να έχουν πρόσβαση και να τα ανακτούν.

Στα κατανεμημένα περιβάλλοντα, τα οποία έχουν επικρατήσει, οι ερευνητές από διαφορετικούς πολιτισμούς και με διαφορετικές γλώσσες θέλουν να έχουν πρόσβαση σε υλικό άλλων γλωσσών, ανεξαρτήτως γλώσσας, τοπικού υλικού και λογισμικού. Γι' αυτό κρίνεται επιτακτική η ανάγκη διαλειτουργικότητας και κωδικοποίησης των

χαρακτήρων σε ένα πρότυπο που να υποστηρίζει την πλειοψηφία των γραπτών γλωσσών.

Το Unicode εμφανίζεται να είναι η λύση αλλά υπάρχει πληθώρα παραμέτρων που πρέπει να ληφθεί υπόψιν, όπως ο ρυθμός υλοποίησης αυτής της μετατροπής, αν υπάρχει η απαραίτητη τεχνολογία που να υποστηρίζει τέτοια πρότυπα και δεκάδες άλλες. Παρουσιάζονται επίσης προκλήσεις για τους επιστήμονες της πληροφόρησης και τους σχεδιαστές ψηφιακών βιβλιοθηκών, που απαιτούν λύση στο άμεσο μέλλον. Μόνο με τη συμμετοχή και το ενδιαφέρον όλων, θα μπορέσουν να υπερπηδηθούν όλα αυτά τα εμπόδια και η παγκόσμια ψηφιακή βιβλιοθήκη δεν θα φαντάζει ουτοπία.

## **BIBΛΙΟΓΡΑΦΙΑ**

1. Borgman C.L., Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries or How Do We Exchange Data in 400 Languages?, D-Lib, June 1997 (available at: <http://www.dlib.org/dlib/june97/06borgman.html> )
2. Oard D.W., Ruiz M., Klavans J., Multi-lingual Information Discovery and Access (MIDAS), D-Lib, October 1999 (available at: <http://www.dlib.org/dlib/october99/10oard.html>)
3. Murthy T., Interoperability among Multi-Lingual Digital Libraries through Unicode based metadata: a model for India, Indo-US Workshop on Open Digital Libraries and Interoperability, Virginia Tech, USA, 23-25 June 2003 (available at: <http://fox.cs.vt.edu/IndoUSdl/murthy.pdf>)
4. Oard D.W., Multilingual Information Access: the user's perspective )available at: <http://www.iei.pi.cnr.it/DELOS/CLEF/workshop00.html>)
5. Clews J., Digital Language Access: scripts, transliteration, and computer access, D-Lib, March 1997 (available at: <http://www.dlib.org/dlib/march97/sesame/03clews.html>)
6. Peters C., Picchi E., Across Languages, Across Cultures: issues in multilinguality and digital libraries, D-Lib, May 1997 (available at: <http://www.dlib.org/dlib/may97/peters/05peters.html>)
7. Pavani A., A Model of Multilingual Digital Library, Ci.Inf., Brasilia, v.30, n.3.,p.73-81, Sep./Dec. 2001 (available at: <http://www.ibict.br/cionline/300301/3031001.pdf> )
8. Maeda A., Multi-lingual Information Processing for Digital Libraries (available at : <http://pnclink.org/annual/annual2002/pdf/0921/12/c211206-1.pdf>)
9. Peters C., Cross-Language Evaluation Forum (CLEF): agenda for 2002, D-Lib, February 2002 (available at <http://www.dlib.org/dlib/february02/02inbrief.html>)
10. Peters C., ECDL 2003 Workshop Report: cross-language evaluation forum (CLEF 2003), D-Lib, September 2003 (available at: <http://www.dlib.org/dlib/september03/09inbrief.html>)
11. Peters C., Cross-Language Evaluation Forum, D-Lib, February 2000 (available at: <http://www.dlib.org/dlib/february00/02inbrief.html>)
12. Caidi N., Komlodi A., Cross-cultural Considerations in Digital Library ResearchL report for the JCDL 2003 workshop, D-Lib, July/August 2003 (available at: <http://www.dlib.org/dlib/july03/07inbrief.html>)

13. Dartois M., Maeda A., Sakaguchi T., A Multilingual Electronic Text Collection of Folk Tales for Casual Users Using Off-the-Shelf Browsers, D-Lib, October 1997 (available at: <http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html>)
14. Croft W.B., What Do People Want from Information Retrieval?, D-Lib, November 1995 (available at: <http://www.dlib.org/dlib/november95/11croft.html>)
15. Osawa N., A Multilingual Information Processing Infrastructure for Global Digital Libraries: EPICIST, D-Lib, 1997 (available at: <http://www.dl.ulis.ac.jp/ISDL97/proceedings/osawa/osawa.html>)
16. Java: how to program/ Deitel H.M., Deitel P.J., Prentice Hall PTRM, 5<sup>th</sup> ed., 2002
17. Powell J., Fox E.A., Multilingual Federated Searching Across Heterogeneous Collections, D-Lib, September 1998 (available at: <http://www.dlib.org/dlib/septemeber98/powell/09powell.html>)
18. Anderson D., Unicode and Historic Scripts, Ariadne (available at: <http://www.ariadne.ac.uk/issue37/anderson/>)