ΜΕΡΟΣ Ι:

Κωδικοποίηση Χαρακτήρων και Ταυτοποιητών Πόρων στο Διαδίκτυο

Πίνακας Περιεχομένων

ΚΕΦΑΛΑΙΟ 1:
Κωδικοποίηση Χαρακτήρων

1.1. Εισαγωγή

Η ύπαρξη των προτύπων έχει βελτιώσει βασικούς τομείς συνεργασίας των βιβλιοθηκών και των αρχείων όπως η περιγραφή τεκμηρίων και εγγράφων, η ανταλλαγή εγγραφών και εγγραφών καθιερωμένων τύπων, διαδανεισμού κ.λπ. Με τον όρο συνεργασία δεν εννοείται μόνο η δυνατότητα επικοινωνίας, αλλά η εύρεση κοινών τρόπων αντιμετώπισης προβλημάτων. Ωστόσο είναι γεγονός, ότι η ύπαρξη και η χρήση πολλών διαφορετικών προτύπων για την ολοκλήρωση παρόμοιων διεργασιών, δημιουργεί νέο πρόβλημα διαλειτουργικότητας των συστημάτων.

Οι ηλεκτρονικοί υπολογιστές, χειρίζονται απλώς αριθμούς. Αποθηκεύουν γράμματα και άλλους χαρακτήρες αντιστοιχώντας στο καθένα τους από έναν αριθμό (ονομάζουμε μία τέτοια αντιστοιχία κωδικοσελίδα). Πριν την εφεύρεση του Unicode, υπήρχαν εκατοντάδες διαφορετικές κωδικοσελίδες. Λόγω περιορισμών μεγέθους όμως, σε καμία κωδικοσελίδα δεν χωρούσαν αρκετοί χαρακτήρες: λόγου χάριν, η Ευρωπαϊκή Ένωση χρειαζόταν πλήθος διαφορετικών κωδικοσελίδων για να καλύψει όλες τις γλώσσες των χωρών-μελών της. Ακόμα και για μία και μόνη γλώσσα, όπως π.χ. τα Αγγλικά, μία κωδικοσελίδα δεν επαρκούσε για να καλύψει όλα τα γράμματα, σημεία στίξης και τεχνικά σύμβολα ευρείας χρήσης.

Εκτός αυτού, οι κωδικοσελίδες αυτές διαφωνούσαν μεταξύ τους. Έτσι, δύο κωδικοσελίδες μπορούσαν κάλλιστα να χρησιμοποιούν τον ίδιο αριθμό για δύο διαφορετικούς χαρακτήρες, ή να χρησιμοποιούν διαφορετικούς αριθμούς για τον ίδιο χαρακτήρα. Κάθε υπολογιστής (και ιδίως εάν ήταν διακομιστής) έπρεπε να υποστηρίζει πλήθος διαφορετικών κωδικοσελίδων. Ταυτόχρονα, κάθε φορά που δεδομένα μεταφέρονταν μεταξύ διαφορετικών κωδικοσελίδων ή λειτουργικών συστημάτων, τα δεδομένα αυτά κινδύνευαν να αλλοιωθούν.

Συμπεραίνοντας, η απαίτηση για υποστήριξη χαρακτήρων διαφορετικών αλφαβήτων σε μια βιβλιογραφική εγγραφή ή ένα τεκμήριο είναι ένα από τα βασικότερα προβλήματα που οφείλουν να αντιμετωπίσουν τα σύγχρονα πληροφοριακά συστήματα. Η κατάσταση δυσχεραίνει με την ύπαρξη ποικίλων και διαφορετικών προτύπων κωδικοποίησης για το ίδιο σύνολο χαρακτήρων.

1.2. Ορισμοί

Κωδικός χαρακτήρα (character code): είναι μια ένα προς ένα αντιστοίχηση ενός συνόλου χαρακτήρων στο σύνολο των θετικών ακεραίων αριθμών. Γίνεται ανάθεση μιας *θέσης κωδικού* σε ένα χαρακτήρα.
Κωδικοποίηση χαρακτήρων (character encoding): Οι σειρές των bits που δέχεται ή αποστέλει ένας υπολογιστής οργανώνονται σε οκτάδες που σχηματίζουν τα bytes. Από ένα byte μπορούν να αναπαρασταθούν 2^8=256 αριθμοί και επομένως σε ένα byte μπορούν να αντιστοιχηθούν 256 κωδικοί χαρακτήρων (0-255). Με τον όρο κωδικοποίηση χαρακτήρων εννοούμε μια μέθοδο αναπαράστασης των κωδικών των χαρακτήρων στον υπολογιστή.

1.3. ASCII: American Standard Code for Information Interchange

Όταν αναπτύχθηκε ο κώδικας ASCII, από ένα byte χρησιμοποιούνταν μόνο τα 7 bits για την κωδικοποίηση των χαρακτήρων, ενώ το άλλο ένα bit αποτελούσε ψηφίο ισοδυναμίας (parity bit) για τον έλεγχο μετάδοσης των bytes. Αποτέλεσμα αυτού του τεχνολογικού περιορισμού ήταν ότι το επιτρεπόμενο πλήθος κωδικών ήταν 2^7=128, με εύρος από 0-127. Στον κώδικα ASCII οι κωδικοί 0-31 και 127 αντιστοιχούν σε *control χαρακτήρες* (βλ. παρακάτω στην ομότιτιλη παράγραφο) και το σύνολο των χαρακτήρων ήταν το ακόλουθο.

```
 ! " # $ % & ' ( ) * + , - . /
0 1 2 3 4 5 6 7 8 9 : ; < = > ?
@ A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z [ \ ] ^ _
` a b c d e f g h i j k l m n o
p q r s t u v w x y z { | } ~
```

1.4. Εθνικές παραλλαγές του ASCII

Η αρχική έκδοση του ASCII ονομάστηκε πρότυπο ANSI X3.4-1986. Ωστόσο υπήρχε αδυναμία κωδικοποίησης πολλών χαρακτήρων ακόμα και βασισμένων στο λατινικό αλφάβητο. Για αυτό το λόγο αναπτύχθηκε το πρότυπο ISO 646 το οποίο παρείχε παρόμοια κωδικοποίηση με ASCII εκτός των χαρακτήρων @[\]{|} που αντιστοιχούν σε κωδικούς εθνικής χρήσης. Επιπλέον παρείχε ελευθερία στην αντιστοίχηση των χαρακτήρων #$^`~.

Παράδειγμα 1.1. Στον κωδικό 35 που αντιστοιχούσε ο χαρακτήρας #, υπήρχε δυνατότητα αντιστοίχησης του χαρακτήρα που συμβολίζει το βρετανικό νόμισμα (£)

Το πρότυπο ορίζει και μια International Reference Version (IRV) η οποία δεν διέφερε από τον ASCII. Στον επόμενο πίνακα παρουσιάζονται μερικοί εναλλακτικοί εθνικοί χαρακτήρες που μπορούν να χρησιμοποιηθούν στη θέση των «ελεύθερων χαρακτήρων» του ISO 646.

κωδικός ASCII	**χαρακτήρας**	**εθνική παραλλαγή**
35	#	£ Ù
36	$	¤
64	@	É § Ä à ³

1.5. Η οικογένεια προτύπων ISO 8859

Όταν η τεχνολογία επέτρεψε η κωδικοποίηση των χαρακτήρων να γίνεται από 8-bit , τότε διπλασιάστηκε το πλήθος των χαρακτήρων από τους 128 στους 256 με κωδικούς 0-255. Έτσι αναπτύχθηκε η οικογένεια των προτύπων 8859 σύμφωνα με την οποία

1. οι κωδικοί από 0-127 αντιστοιχούν στους χαρακτήρες του ASCII,
2. οι κωδικοί 128-159, αντιστοιχούν σε *control χαρακτήρες,*
3. ενώ στους κωδικούς 160-255 υπάρχει δυνατότητα αναπαράστασης χαρακτήρων ενός άλλου αλφάβητου.

Παράδειγμα 1.2. Στις θέσεις των κωδικών 160-255 του ISO 8859-1 (Latin 1) αντιστοιχούν οι χαρακτήρες:

¡ ¢ £ ¤ ¥ ¦ § ¨ © ª « ¬ - ® ¯
° ± ² ³ ´ µ ¶ · ¸ ¹ º » ¼ ½ ¾ ¿
À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï
Ð Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß
à á â ã ä å æ ç è é ê ë ì í î ï
ð ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ÿ

1.6. Τα μέρη του ISO 8859

ISO 8859-1 Latin alphabet No. 1 "Western", "West European»
ISO 8859-2 Latin alphabet No. 2 "Central European", "East European"
ISO 8859-3 Latin alphabet No. 3 “South European"; "Maltese & Esperanto"

ISO 8859-4 Latin alphabet No. 4"North European"
ISO 8859-5 Latin/Cyrillic alphabet (Slavic languages)
ISO 8859-6 Latin/Arabic alphabet (Arabic language)
ISO 8859-7 Latin/Greek alphabet (modern Greek)
ISO 8859-8 Latin/Hebrew alphabet (Hebrew and Yiddish)
ISO 8859-9 Latin alphabet No. 5 "Turkish"
ISO 8859-10 Latin alphabet No. 6 "Nordic" (Sámi, Inuit, Icelandic)
ISO 8859-11 Latin/Thai alphabet (Thai language)

(Το 12ο μέρος δεν έχει ακόμα οριοθεί.)
ISO 8859-13 Latin alphabet No. 7 Baltic Rim
ISO 8859-14 Latin alphabet No. 8 Celtic
ISO 8859-15 Latin alphabet No. 9 "euro"
ISO 8859-16 Latin alphabet No. 10 Albanian, Croatian, English, Finnish, French, German, Hungarian, Irish Gaelic (new orthography), Italian, Latin, Polish, Romanian, and Slovenian.

Το πρότυπο ISO 8859-7 που παρουσιάζεται στο επόμενο σχήμα, υιοθετήθηκε

A0	A1 ‘	A2 ’	A3 £			A6 ¦	A7 §	A8 ¨	A9 ©		AB «	AC ¬	AD -		AF ―
B0 °	B1 ±	B2 ²	B3 ³	B4 ΄	B5 ΅	B6 Ά	B7 ·	B8 Έ	B9 Ή	BA Ί	BB »	BC Ό	BD ½	BE Ύ	BF Ώ
C0 ΐ	C1 Α	C2 Β	C3 Γ	C4 Δ	C5 Ε	C6 Ζ	C7 Η	C8 Θ	C9 Ι	CA Κ	CB Λ	CC Μ	CD Ν	CE Ξ	CF Ο
D0 Π	D1 Ρ		D3 Σ	D4 Τ	D5 Υ	D6 Φ	D7 Χ	D8 Ψ	D9 Ω	DA Ϊ	DB Ϋ	DC ά	DD έ	DE ή	DF ί
E0 ΰ	E1 α	E2 β	E3 γ	E4 δ	E5 ε	E6 ζ	E7 η	E8 θ	E9 ι	EA κ	EB λ	EC μ	ED ν	EE ξ	EF ο
F0 π	F1 ρ	F2 ς	F3 σ	F4 τ	F5 υ	F6 φ	F7 χ	F8 ψ	F9 ω	FA ϊ	FB ϋ	FC ό	FD ύ	FE ώ	

από τον Ελληνικό Οργανισμό τυποποίησης και αποτελεί το εθνικό πρότυπο κωδικοποίησης χαρακτήρων ΕΛΟΤ-928.

1.7. Κωδικοσελίδες DOS, Windows

Το λειτουργικό σύστημα MS DOS χρησιμοποίησε διαφορετικούς κωδικούς χαρακτήρων με 8-bit κωδικοποίηση που ονομάζονται code pages. Στο code page 437 περιλαμβάνονται μαθηματικά σύμβολα και ελληνικοί χαρακτήρες, ενώ στο code page 850 περιλαμβάνονται σχεδόν όλοι οι χαρακτήρες του Latin 1 αλφαβήτου αλλά σε διαφορετικές θέσεις κωδικών από το ISO 8859-1.

Τα Windows χρησιμοποιούν άλλη κωδικοποίηση για τους χαρακτήρες, δημιουργώντας μια πληθώρα από νέες κωδικοσελίδες οι οποίες δεν είναι συμβατές με τα διεθνή πρότυπα. Οι κωδικοσελίδες των Windows συμβολίζονται από τέσσερα ψηφία εκ των οποίων τα τρία πρώτα είναι τα 125 (π.χ. cp-1252 για Latin 1, cp-1253 για ελληνικά). Η διαφορά των κωδικοσελίδων των Windows από το ISO-8859 είναι ότι αναθέτουν στους χαρακτήρες σε διαφορετικούς κωδικούς.

Παράδειγμα 1.3. Στο ISO 8859-1 οι κωδικοί 128 - 159 αντιστοιχούν σε control χαρακτήρες που δεν έχουν γραφική παράσταση, δηλ. δεν μπορούν να τυπωθούν στην οθόνη του υπολογιστή. Στο αντίστοιχο cp-1252 των Windows, οι κωδικοί αυτοί αντιστοιχούν σε χαρακτήρες που μπορούν να τυπωθούν στην οθόνη του υπολογιστή.

1.8. ISO 10646 και Unicode

Το ISO 10646 καθορίζει το Universal Character Set, που είναι ένα μεγάλο σύνολο χαρακτήρων (καλύπτει πολλά αλφάβητα) με ενιαία κωδικοποίηση.
Προέκυψε από:
- την πληθώρα των 8-bit κωδικοσελίδων η οποία παρουσίαζε ασυμφωνίες στους κωδικούς ίδιων χαρακτήρων
- την ανάγκη κωδικοποίησης πολλών χαρακτήρων σε μια κωδικοσελίδα

Το πρότυπο προτείνει 32-bit κωδικοποίηση (UCS-4) αλλά στην πράξη χρησιμοποιείται η 16-bit κωδικοποίηση (UCS-2), η οποία ορίζει το Basic Multilingual Plane (BMP). Σύμφωνα με το BMP, τα πρώτα δύο bytes θεωρούνται 0 0, ενώ τα υπόλοιπα αντιστοιχούν σε χαρακτήρες.

Το Unicode υλοποιεί το ISO 10646, προτείνοντας 16-bit κωδικοποίηση, και αποδίδει ένα μοναδικό αριθμό για κάθε χαρακτήρα, καλύπτοντας περισσότερους από 65.000 χαρακτήρες.

1.9. Unicode

Συγκεκριμένα το Unicode version 3.0 (ήδη ανακοινώθηκε η version 4.0) περιλαμβάνει:
- 49.194 χαρακτήρες αλφαβήτων και γραφών από την Ευρώπη, τη Μέση Ανατολή (συμπεριλαμβανομένων γραφών από δεξιά προς τα αριστερά) και την Ασία (π.χ. το Han subset περιέχει 27.484 ιδεογράμματα από την Κίνα, την Ιαπωνία, την Κορέα, το Βιετνάμ, την Ταϊβάν και τη Σιγκαπούρη).
- Σημεία στίξης, μαθηματικά και τεχνικά σύμβολα, γεωμετρικά σχήματα.
- Άλλα αλφάβητα όπως Ethiopic, Canadian Aboriginal Syllabics, Cherokee, Sinhala, Syriac, Myanmar, Khmer, Mongolian, Braille και άλλα ιδεογράμματα.

Επίσης κρατά 6.400 ιδιωτικής χρήσης κωδικούς ενώ υπάρχουν ακόμα 7.827 αχρησιμοποίητοι κωδικοί για μελλοντική επέκταση.

Το Unicode παρέχει τη σημαντική δυνατότητα συνδυασμού 2.048 (16-bit) κωδικών *σε ζεύγη* (pair codes). Με αυτόν τον τρόπο αποδίδει επιπλέον 1.048.544 χαρακτήρες (για ειδικά σύμβολα με τόνους και αρχαίες γραφές), καλύπτοντας πρακτικά σχεδόν όλα τα αλφάβητα και τους τύπους γραφής.

Οι κωδικοί των χαρακτήρων στο Unicode συμβολίζονται με τη μορφή: U+nnnn, όπου το nnnn είναι δεκαεξαδικός αριθμός (π.χ. ο κωδικός U+0020 αντιστοιχεί στον χαρακτήρα «κενό» (space)).

1.10. Υλοποιήσεις του Unicode

Το Unicode έχει υλοποιηθεί με διάφορους τρόπους, που ονομάζονται Unicode Transformation Format (UTF). Η κάθε υλοποίηση έχει στόχο την αναπαράσταση των κωδικών των χαρακτήρων από ομάδες bytes, έτσι ώστε να επιτυγχάνεται η ελάχιστη χρήση bytes για κάθε χαρακτήρα. Σημαντικότερες υλοποιήσεις αποτελούν τα UTF-16, UTF-7 και UTF-8.

1.10.1. UTF-16

Το UTF-16 υλοποιεί το Unicode χρησιμοποιώντας 2 bytes (16-bits) για κάθε χαρακτήρα. Θεωρείται αντιοικονομική κωδικοποίηση αφού ακόμα και χαρακτήρες που ανήκουν στον ASCII, οποίοι απαιτούν 7-bits, κωδικοποιούνται από 16.

1.10.2. UTF-7

Στο UTF-7 κάθε χαρακτήρας κωδικοποιείται από ένα ή περισσότερα bytes. Οι χαρακτήρες ανάλογα με τον κωδικό τους οργανώνονται σε ισοπληθείς ομάδες από 128 κωδικούς.

Οι πρώτοι 128 κωδικοί συμφωνούν με τον ASCII και κωδικοποιούνται από ένα byte. Για την κωδικοποίηση των υπόλοιπων κωδικών χαρακτήρων (που δεν ανήκουν στην πρώτη ομάδα) προηγούνται bytes που αντιστοιχούν σε κωδικούς control χαρακτήρων οι οποίοι παραπέμπουν (δηλ. λειτουργούν ως δείκτες) στην αντίστοιχη ομάδα που ανήκει ο κάθε κωδικός χαρακτήρα.

Το UTF-7 δεν χρησιμοποιείται ευρέως ούτε συνιστάται από διεθνείς οργανισμούς προτυποποίησης.

1.10.3. UTF-8

Το UTF-8 θεωρείται η βέλτιστη υλοποίηση του Unicode και συνιστάται από τους διεθνείς οργανισμούς προτυποποίησης. Η κωδικοποίηση που προσφέρει είναι μεταβλητού μήκους από 1-6 bytes και έτσι επιτυγχάνεται οικονομίας στη μνήμη. Το UTF-8 πρακτικά καλύπτει όλα σχεδόν τα αλφάβητα και τους συνδυασμούς ζευγών κωδικών για την αναπαράσταση τονούμενων χαρακτήρων.

Συγκεκριμένα, οι χαρακτήρες του ASCII (κωδικοί 0-127 ή οι αντίστοιχοι δεκαεξαδικοί U-00000000 - U-0000007F) έχουν το πρώτο bit 0 και κωδικοποιούνται από ένα μόνο byte.

Οι υπόλοιποι χαρακτήρες (εκτός ASCII) κωδικοποιούνται από 2-6 bytes ως εξής:

Το πλήθος των άσσων του πρώτου byte δηλώνει το πλήθος των bytes που κωδικοποιούν τον χαρακτήρα. Μετά τους άσσους ακολουθεί 0 και κατόπιν ακολουθούν τα bits που κωδικοποιούν τον κωδικό του χαρακτήρα σε δυαδική μορφή. Κάθε byte που ακολουθεί έχει αρχικά bits τα 10.

Στον παρακάτω πίνακα παρουσιάζονται οι ομάδες κωδικών των χαρακτήρων, σε δεκαεξαδική μορφή, σύμφωνα με το πρότυπο Unicode καθώς επίσης και η δομή των bytes που τους κωδικοποιούν. Οι χαρακτήρες *x* αντιστοιχούν στα bits που κωδικοποιούν κάθε χαρακτήρα σύμφωνα με το πρότυπο Unicode.
U-00000000 - U-0000007F: 0*xxxxxxx*
U-00000080 - U-000007FF: 110*xxxxx* 10*xxxxxx*
U-00000800 - U-0000FFFF: 1110*xxxx* 10*xxxxxx* 10*xxxxxx*
U-00010000 - U-001FFFFF: 11110*xxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx*
U-00200000 - U-03FFFFFF: 111110*xx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx*
U-04000000 - U-7FFFFFFF: 1111110*x* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx*

Παράδειγμα 1.4. Ο χαρακτήρας Copyright sign (©) έχει κωδικό U+00A9 και επομένως ανήκει στην 2η γραμμή του παραπάνω πίνακα, δηλ. κωδικοποιείται από 2 bytes ως εξής:
ο δεκαεξαδικός αριθμός 00A9 στο δυαδικό σύστημα έχει τη μορφή 00010101001 δηλ. απαιτεί 11 bits και άρα θα «αναλωθούν» 2 bytes. Σύμφωνα με την παραπάνω ανάλυση των αρχικών bits του κάθε byte έχουμε:
1ο byte: 110 και αρχίζουν τα bits του κωδικού 00010 δηλ. 11000010
2ο byte: 10 και συνεχίζουν τα bits του κωδικού 101001 δηλ. 10101001.
Άρα ο χαρακτήρας στο UTF-8 κωδικοποιείται: 11000010 10101001.

Παράδειγμα 1.5. Το σύμβολο Not equal to (≠) έχει κωδικό U+2260 και επομένως ανήκει στην 3η γραμμή του παραπάνω πίνακα, δηλ. κωδικοποιείται από 3 bytes ως εξής:
Η δυαδική του μορφή του κωδικού U+2260 είναι 10001001100000 και αποτελείται από 14 bits. Ξεκινώντας από την τελευταία θέση του τρίτου byte, παραθέτουμε τα bits του κωδικού
3ο byte: 10100000 (τα δύο πρώτα bits είναι τα 10 και ακολουθούν τα έξι τελευταία bits του κωδικού).
2ο byte: 10001001 (τα δύο πρώτα bits είναι τα 10 και ακολουθούν τα έξι επόμενα bits του κωδικού).
1ο byte: 11100010 (τα τέσσερα πρώτα bits είναι τα 1110, τα δύο τελευταία είναι τα 10 και ολοκληρώθηκε ο κωδικός του χαρακτήρα. Στις θέσεις 5 και 6 του πρώτου byte που περισσεύουν, τοποθετούμε το bit 0, αφού τα μηδενικά μπροστά από ένα αριθμό δεν τον αλλάζουν).
Άρα ο χαρακτήρας στο UTF-8 κωδικοποιείται: 11100010 10001001 10100000

1.11. Control χαρακτήρες και ακολουθίες διαφυγής

Όπως αναφέρθηκε παραπάνω όλες οι κωδικοποιήσεις χαρακτήρων περιλαμβάνουν control χαρακτήρες. Αυτοί είναι μη ορατοί χαρακτήρες που

χρησιμοποιούνται για τον έλεγχο συσκευών (devices, οθόνη, πληκτρολόγιο) και διεργασιών (processes). Για παράδειγμα στον ASCII ο κωδικός (3) σταματά την τρέχουσα διεργασία (σε συστήματα Unix ο χαρακτήρας ενεργοποιείται με το ταυτόχρονο πάτημα των πλήκτρων Ctrl+C), ο κωδικός (13) αντιστοιχεί στο carriage return (πλήκτρο enter, αφήνει μια κενή γραμμή), ενώ τέλος ο κωδικός (9) αντιστοιχεί στο horizontal tab.

Σημαντικό ρόλο παίζουν οι ακολουθίες χαρακτήρων μετά τον χαρακτήρα escape (ESC, ο χαρακτήρας με κωδικό 27 στον κώδικα ASCII). Οι ακολουθίες αυτές ονομάζονται ακολουθίες διαφυγής (escape sequences) και χρησιμοποιούνται για τον έλεγχο των τερματικών σταθμών. Έτσι μια escape sequence μπορεί να αλλάζει το χρώμα των γραμμάτων σε μια οθόνη τύπου VT100, ή να αλλάζει τη θέση του κέρσορα στην οθόνη (π.χ. να πηγαίνει τον κέρσορα στην αρχή μιας γραμμής).

Οι ακολουθίες διαφυγής χρησιμοποιούνται και για την αλλαγή αλφαβήτων (κωδικοσελίδων). Η εμφάνιση μιας συγκεκριμένης ακολουθίας διαφυγής αποτελεί μια οδηγία (εντολή) που δηλώνει ότι οι χαρακτήρες που ακολουθούν μέχρι την εμφάνιση μιας νέας ακολουθίας διαφυγής, ανήκουν στο αλφάβητο που αντιστοιχεί σε αυτή την ακολουθία. Πριν από την εμφάνιση του Unicode και όταν η κωδικοποίηση των χαρακτήρων απαιτούσε 7-bit και 8-bit, η αλλαγή αλφαβήτου δηλώνονταν και επιτυγχανόταν με την πρόθεση μιας ακολουθίας διαφυγής. Το πρότυπο ISO 2022, όριζε συγκεκριμένα αλφάβητα (π.χ. ASCII, ISO 8859-7, Latin-1 κλπ) και τις αντίστοιχες escape sequences για κάθε αλλαγή κάθε αλφάβητου.

Παράδειγμα 1.6. Η ακολουθία διαφυγής π.χ. η “ESC F”, δηλώνει την αλλαγή της κωδικοποίησης χαρακτήρων σε ελληνικά (ISO 8859-7, μόνο τους χαρακτήρες που τυπώνονται, όχι τους control χαρακτήρες). Αυτή η ακολουθία διαφυγής δίνει τον έλεγχο στην ελληνική κωδικοσελίδα και έτσι μπορούμε να γράψουμε και να διαβάσουμε ελληνικά στην οθόνη μας. Η εμφάνιση της ακολουθίας “ESC (B” επαναφέρει το σύνολο των ASCII χαρακτήρων.

Η έννοια των ακολουθιών διαφυγής γενικεύεται από τα λογισμικά επεξεργασίας κειμένου προκειμένου να αναπαραστήσουν συγκεκριμένους χαρακτήρες που δεν καλύπτονται από το δεδομένο αλφάβητο. Στο πρόγραμμα LATEX ο χαρακτήρας διαφυγής (αντίστοιχος του “ESC”) είναι ο “\”.

Παράδειγμα 1.7. Στο πρόγραμμα επεξεργασίας κειμένου LATEX ο χαρακτήρας Copyright sign (©) αναπαρίσταται από την ακολουθία “\copyright”.

Η γλώσσα HTML παρέχει τη δυνατότητα ορισμού της κωδικοποίησης μιας σελίδας, προκειμένου ο φυλλομετρητής (browser) που θα παρουσιάσει τη

σελίδα να χρησιμοποιήσει την αντίστοιχη κωδικοποίηση. Η δήλωση κωδικοποίησης του κειμένου μιας σελίδας, γίνεται στην αρχή της, μέσα σε meta ετικέτες. Αν στη σελίδα υπάρχουν χαρακτήρες που δεν ανήκουν στη δηλωθείσα κωδικοποίηση, τότε η «διαφυγή» από τη δήλωση γίνεται με τη χρήση του & και μιας σειράς χαρακτήρων, ή με τη χρήση των &# και ενός κωδικού αριθμού.

Παράδειγμα 1.8. Η δήλωση <meta content="text/html; charset=1253"/meta> σημαίνει ότι οι χαρακτήρες της σελίδας αντιστοιχούν στο ελληνικό αλφάβητο. Αν στο κείμενο υπάρχει ο χαρακτήρας Ä, τότε θα γίνει «κατανοητός» από τον φυλλομετρητή αν έχει γραφεί στη σελίδα με την ακολουθία διαφυγής &Auml ή `&#196.`

1.12. Γραμματοσειρές (fonts)

Η γραμματοσειρά είναι μια σχηματική αναπαράσταση ενός συνόλου χαρακτήρων. Κάθε λογισμικό που διαχειρίζεται χαρακτήρες, διαθέτει και τις αντίστοιχες γραμματοσειρές για την εμφάνισή τους. Οι γραμματοσειρές δεν πρέπει συγχέονται με τους χαρακτήρες και την κωδικοποίησή τους. Για ένα χαρακτήρα που αντιστοιχεί σε ένα κωδικό υπάρχουν διαφορετικές σχηματικές απεικονίσεις του από διαφορετικές γραμματοσειρές. Τα λογισμικά (λειτουργικά συστήματα, προγράμματα εφαρμογών και αυτοματισμού γραφείου κλπ) επιτυγχάνουν την εναλλαγή των γραμματοσειρών μέσα σε ένα κείμενο ή μια ακολουθία χαρακτήρων, χρησιμοποιώντας ακολουθίες διαφυγής (escape sequences). Αυτές οι ακολουθίες διαφυγής δεν διέπονται από κάποιο πρότυπο αλλά ορίζονται από τον κατασκευαστή του αντίστοιχου λογισμικού.

Παράδειγμα 1.9.
- τα Z, **Z,** *Z*, ***Z*** είναι απεικονίσεις του ίδιου χαρακτήρα U-00E9.
- το λατινικό και το ελληνικό «Α», έχουν την ίδια μορφή αλλά διαφορετικούς κωδικούς
- στο Unicode, χαρακτήρες με την ίδια μορφή, αλλά με διαφορετικό νόημα, έχουν διαφορετικούς κωδικούς (π.χ. το γράμμα «Ν» και το σύνολο των φυσικών αριθμών Ν θεωρούνται άλλοι χαρακτήρες).

1.13. Πληροφορίες για την Κωδικοποίηση Χαρακτήρων

Από τα μέχρι στιγμής αναφερθέντα είναι προφανές ότι μια ακολουθία από bytes που αντιστοιχούν σε χαρακτήρες, μπορεί να αποκωδικοποιηθεί με διαφορετικούς τρόπους. Για την αποφυγή προβλημάτων στη μετάδοση και αποκωδικοποίηση χαρακτήρων οι χρησιμοποιούμενες κωδικοσελίδες αναφέρονται σε κάθε εφαρμογή που διαχειρίζεται χαρακτήρες. Ειδικότερα για τις ιστοσελίδες, όπως προαναφέρθηκε η δήλωση κωδικοποίησης του κειμένου μιας σελίδας, γίνεται στην αρχή της, μέσα σε meta ετικέτες.

Για τα μηνύματα ηλεκτρονικού ταχυδρομείου έχει προβλεφθεί το Multipurpose Internet Mail Extensions (MIME) format που επιτρέπει στο ηλεκτρονικό ταχυδρομείο του Internet να αποστέλλει δεδομένα που δεν ανήκουν στον ASCII κώδικα. Αυτό επιτυγχάνεται με τη δήλωση Content-type (τύπος περιεχομένου) η οποία περιέχεται στην επικεφαλίδα (header) του μηνύματος. Η επικεφαλίδα ενός μηνύματος προηγείται του κυρίως μηνύματος και περιέχει πληροφορία για τον τρόπο τον τρόπο κωδικοποίησης του περιεχομένου του μηνύματος.

Παράδειγμα 1.10. Ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου (e-mail client) βοηθάει ένα χρήστη να αποστείλει μηνύματα σε άλλους χρήστες μέσω του διαδικτύου. Το πρόγραμμα αυτό σχηματίζει για κάθε μήνυμα μια επικεφαλίδα που περιλαμβάνει δηλώσεις MIME της μορφής
Content-Type: text/html; charset=iso-8859-1
Η παραπάνω δήλωση αναφέρει τον τύπο του μηνύματος και την κωδικοποίηση των χαρακτήρων – στην συγκεκριμένη περίπτωση το μήνυμα είναι τύπου κειμένου html με κωδικοποίηση χαρακτήρων iso-8859-1.
Ο παραλήπτης ενός τέτοιου μηνύματος μπορεί να διαθέτει ένα διαφορετικό πρόγραμμα ηλεκτρονικού ταχυδρομείου, το οποίο αν υποστηρίζει μηνύματα με μορφή MIME, θα πληροφορηθεί από την επικεφαλίδα του μηνύματος που θα λάβει για την κωδικοποίηση των χαρακτήρων του μηνύματος και θα το αποκωδικοποιήσει ανάλογα με την πληροφορία αυτή.

Κάτι ανάλογο συμβαίνει και με τις βιβλιογραφικές εγγραφές που ακολουθούν το πρότυπο UNIMARC. Το UNIMARC προβλέπει διαπραγμάτευση προτύπου χαρακτήρων και η δήλωση των χρησιμοποιούμενων κωδικοσελίδων γίνεται στο πεδίο 100 «Γενικά Δεδομένα Επεξεργασίας». Ως επί το πλείστον τα συστήματα που το υλοποιούν UNIMARC, χρησιμοποιούν την 8-bit κωδικοποίηση. Οι 256 κωδικοί χαρακτήρων χωρίζονται σε δύο πίνακες των 128 θέσεων που ονομάζονται χαμηλή σελίδα (κωδικοί 0-127) και υψηλή σελίδα (128-255). Οι πίνακες οργανώνουν τους χαρακτήρες σε 16 γραμμές και 8 στήλες. Σύμφωνα με το πρότυπο ISO 2022, οι δύο πρώτες στήλες (32 χαρακτήρες) της κάθε σελίδας περιέχουν control χαρακτήρες και έχουν κωδικούς C0 και C1, ενώ οι υπόλοιπες στήλες περιέχουν ομάδες από 94 χαρακτήρες που έχουν τους κωδικούς G0 και G1 (graphic characters 0 and 1).

Το πεδίο 100 του UNIMARC δέχεται μέχρι τέσσερις κωδικοσελίδες των 128 χαρακτήρων με μοναδικό κωδικό (01: ISO 646 βασικά λατινικά, 03: ANSEL εκτεταμένα λατινικά, 05: ISO 5428-1984 τονούμενα ελληνικά, 04: κυριλλικά). Αν και σε μια εγγραφή επιτρέπεται η χρήση τεσσάρων αλφάβητων (κωδικοσελίδων), μόνο δύο από αυτά είναι ενεργά και καταλαμβάνουν τις ομάδες χαρακτήρων G0 και G1. Οι σελίδες που δεν είναι ενεργές μεταφέρονται σε ενεργές θέσεις (πρώτη ή δεύτερη θέση, δηλ. σύνολα G0 και G1) με τη χρήση κατάλληλων ακολουθιών διαφυγής (escape sequences).

Παράδειγμα 1.11. Οι τιμές «010305 » στο πεδίο 100 δηλώνουν ότι θα χρησιμοποιηθούν οι σελίδες με κωδικούς 01, 03 και 05 δηλ. βασικά λατινικά, εκτεταμένα λατινικά και ελληνικά αντίστοιχα (τα κενά στο τέλος δηλώνουν ότι δεν υπάρχει τέταρτη σελίδα χαρακτήρων). Για να γράψουμε ελληνικούς χαρακτήρες πρέπει να μετακινηθεί η σελίδα με τους ελληνικούς χαρακτήρες από την τρίτη θέση στην πρώτη. Αυτό επιτυγχάνεται με τη χρήση συγκεκριμένης escape sequence που εκτελεί την εντολή: «μετακίνησε τη σελίδα με κωδικό 05 δύο θέσεις αριστερά».

Με αυτόν τον τρόπο το UNIMARC εξασφαλίζει τη συνύπαρξη και διαχείριση διαφορετικών συνόλων χαρακτήρων, ανεξάρτητα αν ο υπολογιστής μπορεί να τους προβάλει. Για παράδειγμα ο ελληνικός χαρακτήρας «λ» που μπορεί να υπάρχει σε μια εγγραφή, μπορεί να μην εμφανίζεται, επειδή το σύστημα στο οποίο προβάλλεται δεν υποστηρίζει έχει την κωδικοσελίδα του ISO 5428-1984.

1.14. Γιατί δεν εμφανίζονται οι χαρακτήρες

Πολλές φορές εμφανίζονται αντί για αναγνώσιμους χαρακτήρες διάφορα σημάδια ή σύμβολα (π.χ. «?») ή τίποτε. Αυτό σημαίνει ότι το πρόγραμμα που καλείται να απεικονίσει τους χαρακτήρες:

- είτε δεν έχει πληροφορηθεί για τον τρόπο κωδικοποίησής τους
- είτε από κατασκευής δεν υποστηρίζει τη χρησιμοποιούμενη κωδικοποίηση
- είτε δεν διαθέτει κατάλληλες γραμματοσειρές για την απεικόνισή των χαρακτήρων.

Στις δύο πρώτες περιπτώσεις το πρόγραμμα εμφανίζει τους χαρακτήρες που αντιστοιχούν στην κωδικοποίηση που υποστηρίζει εκείνη τη στιγμή.

1.15. Βιβλιογραφία

1. B. Bemer, That powerful ESCAPE Character – Key and sequences, Computer History Vignettes, http://www.bobbemer.com/ESCAPE.HTM (ημ/νία πρόσβασης 3.3.2004).
2. Guide to the Unicode 3.0 standard, http://www.unicode.org (ημ/νία πρόσβασης 24.2.2003).
3. J. Korpela, A tutorial on character code issues, http://www.cs.tut.fi/~jkorpela/chars.html (ημ/νία πρόσβασης 21.2.2003).
4. M. Kuhn, UTF-8 and Unicode FAQ for Unix/Linux, http://www.cl.cam.ac.uk/~mgk25/unicode.html (ημ/νία πρόσβασης 21.2.2003).
5. RFC 1554-ISO-2022-JP-2: Multilingual Extension of ISO-2022-JP, http://www.faqs.org/rfcs/rfc1554.html (ημ/νία πρόσβασης 3.3.2004).

6. Δ. Γαροφαλίδης. 2000. Υποστήριξη και διαχείριση χαρακτήρων από συστήματα βιβλιοθηκών και τα διεθνή βιβλιογραφικά πρότυπα. 9ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών, 203-212.
7. Μ. Σφακάκης. 2000. Προβλήματα διαλειτουργικότητας βιβλιογραφικών δεδομένων και βιβλιοθηκών και το ολοκληρωμένο σύστημα Αυτοματισμού Βιβλιοθηκών Εθνικού Κέντρου Τεκμηρίωσης (ΑΒΕΚΤ). 9ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών, 183-191.

ΚΕΦΑΛΑΙΟ 2:
Uniform Resource Identifier (URI), Uniform Resource Locator (URL), Uniform Resource Name (URN)

2.1. Εισαγωγή

Ο Παγκόσμιος Ιστός περιλαμβάνει ηλεκτρονικά αντικείμενα τα οποία προσπελαύνονται από ολοένα και αυξανόμενα πρωτόκολλα. Η προσπέλαση των αντικειμένων γίνονται με τη χρήση των ονομάτων τους ή των διευθύνσεών τους. Για αυτό το λόγο κάθε πόρος στο διαδίκτυο, πρέπει να ταυτοποιείται είτε μέσω ονόματος είτε μιας διεύθυνσης στην οποία θα βρίσκεται ο πόρος αυτός. Είναι λοιπόν απαραίτητη η ύπαρξη ενός καθολικού και ενιαίου τρόπου σύνταξης και ανάθεσης ονομάτων και διευθύνσεων στους πόρους του διαδικτύου, ο οποίος θα είναι ανεξάρτητος από τα υπάρχοντα πρωτόκολλα και από το αν ο πόρος είναι προσπελάσιμος. Το World Wide Web Consortium (W3C), από τις αρχές της δεκαετίας του 1990, έχει περιγράψει ένα ενιαίο τρόπο σύνταξης για την κωδικοποίηση διευθύνσεων και ονομάτων των πόρων στο διαδίκτυο.

Στόχος του κεφαλαίου αυτού είναι η παρουσίαση των προτύπων κωδικοποίησης των ταυτοτήτων των πόρων του δικτύου Uniform Resource Identifier (URI), Uniform Resource Locator (URL) και Uniform Resource Name (URN). Το πρότυπο URI καθορίζει τη γενική σύνταξη ταυτοποιητών πόρων, οι οποίοι διακρίνονται σε ταυτοποιητές διεύθυνσης, των οποίων η σύνταξη περιγράφεται από το πρότυπο URL, και ταυτοποιητές ονομάτων, των οποίων η σύνταξη περιγράφεται από το πρότυπο URN.

2.2. URI – Ενιαίοι Ταυτοποιητές πόρων

2.2.1. Ορισμοί

Τα URIs είναι σειρές από χαρακτήρες που ακολουθούν ένα συγκεκριμένο συντακτικό και ταυτοποιούν ένα πληροφοριακό πόρο στο Διαδίκτυο. Γενικά οι ταυτοποιητές πόρων μπορεί να αναφέρονται είτε στη διεύθυνση, είτε στο όνομα του πόρου, είτε και στα δύο.
Αναλυτικότερα με τον όρο «πόρος» (resource) εννοείται οτιδήποτε μπορεί να ταυτοποιηθεί στο διαδίκτυο. Μπορεί να είναι ηλεκτρονικό κείμενο, ψηφιακή

εικόνα, μια υπηρεσία (π.χ. ένα δελτίο πρόγνωσης του καιρού), ή μια συλλογή πόρων. Ένας πόρος μπορεί να είναι φυσικός, δηλ. να είναι ένα ψηφιακό αντικείμενο που μπορεί να ανακτηθεί από το διαδίκτυο, ή αφηρημένος δηλ. μη ανακτήσιμος, π.χ. ένα βιβλίο σε μια βιβλιοθήκη ή ένας οργανισμός.
Με τον όρο «ταυτοποιητής» (identifier) εννοείται ένα αντικείμενο με το οποίο αναφερόμαστε σε ένα πόρο. Στην περίπτωση των URI, τα αντικείμενα είναι μια σειρά χαρακτήρων που ακολουθούν ένα καθορισμένο συντακτικό.
Ο «ενιαίος» τρόπος σύνταξης ταυτοποιητών πόρων παρέχει διάφορα πλεονεκτήματα: Επιτρέπει στο ίδιο περιβάλλον και για τους ίδιους πόρους τη χρήση διαφορετικών ταυτοποιητών που αντιστοιχούν σε διαφορετικούς τρόπους πρόσβασης τους. Επιπλέον δίνει τη δυνατότητα επαναχρησιμοποίησης των ταυτοποιητών σε διαφορετικά περιβάλλοντα εφαρμογών ή ανάπτυξης νέων ταυτοποιητών πόρων χωρίς να απαιτείται παρέμβαση στη σύνταξη και τη σημασιολογία των ήδη υπαρχόντων.
Η σύνταξη ενός URI είναι της μορφής:

<URI scheme>:<scheme specific part>

Ο όρος scheme (σχήμα) καθορίζει ένα χώρο των ονομάτων (name space) που αναφέρονται στο μηχανισμό προσπέλασης του πόρου. Το σχήμα καθορίζει τη σύνταξη και τη σημασία των ταυτοποιητών που ανήκουν σε αυτό. Ξεκινά με ένα πεζό χαρακτήρα του ASCII κώδικα και τελειώνει με τον χαρακτήρα ":" (colon). Γνωστά σχήματα είναι τα http:, ftp: mailto: κ.ά. Το σχήμα "urn: " αντιστοιχεί αποκλειστικά στους ταυτοποιητές που ακολουθούν το πρότυπο URN.
Για τη σύνταξη του μέρους των URI μετά τη δήλωση σχήματος, δεν υπάρχουν συγκεκριμένοι κανόνες. Το μέρος αυτό εξαρτάται από τους κανόνες σύνταξης που επιβάλλει το σχήμα.

Παράδειγμα 2.1. Τα ακόλουθα αποτελούν URIs.
ftp://ftp.is.co.za/rfc/rfc1808.txt το σχήμα ftp καθορίζει υπηρεσίες μεταφοράς αρχείων (file transfer protocol)
mailto:mduerst@ifi.unizh.ch το σχήμα mailto καθορίζει διευθύνσεις ηλεκτρονικού ταχυδρομείου
news:comp.infosystems.www.servers.unix το σχήμα news καθορίζει ομάδες ανταλλαγής ειδήσεων στο USENET
telnet://melvyl.ucop.edu/ καθορίζει υπηρεσίες διάδρασης μέσω του πρωτοκόλλου telnet
http://www.pgm3.com/varna/school/index.htm καθορίζει υπηρεσίες μεταφοράς υπερκειμένου μέσω του πρωτοκόλλου HTTP. Συγκεκριμένα αυτό το URI αναφέρεται σε ένα κείμενο με όνομα index.htm, που μπορεί να προσπελασθεί μέσω του πρωτοκόλλου HTTP στη μηχανή www.pgm3.com και στο μονοπάτι /varna/school/

2.2.2. Δεσμευμένοι χαρακτήρες και ακολουθίες διαφυγής

Μια από τις βασικές αρχές που διέπουν τη σύνταξη των URIs είναι ότι πρέπει να μπορούν εκτυπώνονται. Για αυτό το λόγο συνίσταται η χρήση χαρακτήρων που ανήκουν στον ASCII κώδικα, ο οποίος είναι αναγνώσιμος και εκτυπώσιμος σχεδόν από κάθε υπολογιστικό σύστημα.

Ένα URI σχήμα ξεκινά με ένα πεζό γράμμα του ASCII κώδικα και ακολουθούν γράμματα (πεζά ή κεφαλαία), ψηφία και οι χαρακτήρες «+», «-» και «.». Στη δήλωση του σχήματος επιτρέπονται και χαρακτήρες του ISO Latin1 συνόλου αλλά απεικονίζονται ως ακολουθίες διαφυγής (escape sequences). Χαρακτήρας διαφυγής στο πρότυπο URI έχει ορισθεί ο χαρακτήρας «%» και πρέπει να ακολουθείται από δύο δεκαεξαδικούς αριθμούς που αντιστοιχούν σε κωδικούς χαρακτήρων του ISO Latin1 συνόλου.

Παράδειγμα 2.2. Η ακολουθία διαφυγής "%20" αντιστοιχεί στον χαρακτήρα το κενό (space), ενώ η ακολουθία "%7e" αντιστοιχεί στον χαρακτήρα "~". Η χρήση του χαρακτήρα «%» σε ένα URI, χωρίς τη σημασία του χαρακτήρα διαφυγής, γίνεται χρησιμοποιώντας την αντίστοιχη ακολουθία διαφυγής "%25". Στον ASCII κώδικα ο δεκαεξαδικός κωδικός 25 αντιστοιχεί στον χαρακτήρα «%».

Εκτός από τον χαρακτήρα «%», το συντακτικό των URI δεσμεύει και άλλους χαρακτήρες αποδίδοντάς τους ειδική σημασία. Αυτοί οι χαρακτήρες είναι:

"/": χωρίζει σύνολα χαρακτήρων με ιεραρχική σχέση, δηλ. καθορίζει απόλυτα (absolute) ή σχετικά (relative) μονοπάτια (paths). Σχετικό μονοπάτι είναι το URI που δεν περιλαμβάνει το πρόθεμα του σχήματος (scheme:). Σε ένα τέτοιο URI οι χαρακτήρες (".", "..") έχουν ειδική σημασία.
"#": ταυτοποιεί ένα απόσπασμα (fragment) ενός URI.
"?": ταυτοποιεί ένα πόρο που είναι η απάντηση σε ένα ερώτημα (query) που οι παράμετροί του (π.χ. λέξεις – κλειδιά, πεδία κλπ) έπονται από το χαρακτήρα "?".
"*", "!": έχουν διαφορετικές σημασίες ανάλογα με το URI scheme.
Επίσης δεσμευμένοι χαρακτήρες είναι και οι: ";", "=", ":", "@", "&", "+", ",", "$".

Παράδειγμα 2.3. Το URI http://info.cern.ch/albert/marie-claude ορίζει ένα απόλυτο μονοπάτι. Ισοδύναμο URI είναι το URI http://info.cern.ch/albert/marie%2Dclaude, στο οποίο ο χαρακτήρας «-» έχει αντικατασταθεί από την αντίστοιχη ακολουθία διαφυγής (ο δεκαεξαδικός αριθμός «2D» στον κώδικα ASCII αντιστοιχεί στο χαρακτήρα «-»).
Το URI http://info.cern.ch/albert/marie-claude#pos1 αναφέρεται σε ένα σημείο-απόσπασμα ενός αρχείου που ονομάζεται pos1.

Όμως το URI http://info.cern.ch/albert%2Fmarie-claude είναι λάθος γιατί αντί του δεσμευμένου χαρακτήρα «/» που καθορίζει μονοπάτια στα URI χρησιμοποιείται η ακολουθία διαφυγής («%2F»), η οποία δηλώνει ότι ο χαρακτήρας «/» έχει άλλη σημασία από την προκαθορισμένη (δηλ. τον ορισμό μονοπατιών).

2.2.3. Σύνταξη

Όπως προαναφέρθηκε για τη σύνταξη του μέρους των URI μετά τη δήλωση σχήματος δεν υπάρχουν συγκεκριμένοι κανόνες. Ωστόσο η σύνταξη αυτού του τμήματος αναλύεται σε τρία επιμέρους συστατικά:

<authority><path>?<query>

Έτσι η γενική σύνταξη των URI *<scheme>:<scheme specific part>* αναλύεται σε:

<scheme:> //<authority>/<path>?<query>

Πριν από το συστατικό *<authority>* τοποθετούνται οι χαρακτήρες "//" και μετά από αυτό ακολουθεί ο χαρακτήρας "/" αν ακολουθεί κάποιο μονοπάτι, ή/και το "?" αν ακολουθούν οι παράμετροι κάποιου ερωτήματος (π.χ. οι λέξεις κλειδιά του query), ή το τέλος του URI . Ένα authority μπορεί να είναι είτε κάποιος server στο διαδίκτυο, είτε κάποιο μητρώο οργανισμού που διαχειρίζεται σχήματα (registry name).

Στην περίπτωση που το συστατικό <authority> αντιστοιχεί σε κάποιο server, η βασική του σύνταξη πρέπει να είναι της μορφής:

<userinfo>@<host>:<port>

όπου:

- *userinfo* είναι ένα όνομα (user name, ή διακριτικό ή λογαριασμός) κάποιου χρήστη ακολουθούμενος από το "@".
- *host* είναι είτε το όνομα ενός υπολογιστή (server) που φιλοξενεί τον πόρο ακολουθούμενο από το όνομα της περιοχής του διαδικτύου που ανήκει (domain name) είτε η διεύθυνση (IP address) του υπολογιστή στο διαδίκτυο.
- *port* είναι ο αριθμός της δικτυακής πόρτας του server η οποία σχετίζεται με το σχήμα. Αν δεν αναφέρεται εννοείται η προκαθορισμένη πόρτα του server που διαθέτει για την προσπέλασή του από το σχήμα.

Τα ερωτήματα (queries) ξεκινούν με τον χαρακτήρα "?" και αποτελούνται από συμβολοσειρές. Γενικά στη σύνταξη των URI δεν πρέπει να χρησιμοποιούνται οι δεσμευμένοι χαρακτήρες που αναφέρθηκαν στην προηγούμενη παράγραφο.

2.2.4. Μονοπάτια

Το συστατικό <path> των URIs σχετίζεται άμεσα με το <authority> συστατικό και αποτελείται από τμήματα (segments) που διαχωρίζονται με το χαρακτήρα "/" και ορίζουν μια ιεραρχία. Μέσα σε ένα τμήμα μπορεί να υπάρχουν παράμετροι που διαχωρίζονται με το χαρακτήρα ";". Μέσα σε ένα τμήμα μονοπατιού δεν επιτρέπεται η χρήση των δεσμευμένων χαρακτήρων "/", ";", "=" και "?".

Πολύ συχνά για λόγους βέλτιστης οργάνωσης και εύκολης πρόσβασης στους πόρους του διαδικτύου, δημιουργούνται ομάδες ή δενδρικές δομές από συναφείς πόρους. Το βασικό πλεονέκτημα αυτής της δενδρικής οργάνωσης είναι ότι οι ομάδες των πόρων που σχηματίζονται είναι ανεξάρτητες από την τοποθεσία τους ή το σχήμα πρόσβασης σε αυτές. Σε αυτές τις περιπτώσεις υπάρχει η δυνατότητα ταυτοποίησης των πόρων σε σχέση με κάποιους άλλους. Σε αυτή την περίπτωση ορίζονται *σχετικά μονοπάτια* (relative paths) στα οποία δεν υπάρχει το πρόθεμα **<scheme>**: σε αντίθεση με τα *απόλυτα μονοπάτια* (absolute paths) τα οποία αναφέρονται σε πόρους ξεκινώντας με το πρόθεμα του σχήματος.

Στα σχετικά μονοπάτια οι χαρακτήρες "." και ".." αποκτούν ειδική σημασία. Συγκεκριμένα ο χαρακτήρας "." αναφέρεται στο τρέχον επίπεδο της δενδρικής δομής (ιεραρχίας) και η σειρά ".."αναφέρεται στο προηγούμενο επίπεδο (πατέρας) της ιεραρχίας. Για τη χρήση σχετικών μονοπατιών απαιτείται η δήλωση ενός απόλυτου μονοπατιού το οποίο ονομάζεται βασικό μονοπάτι (base path) και χρησιμεύει ως βάση για την αναφορά στους πόρους μιας δενδρικής δομής μέσω σχετικών μονοπατιών. Σε διαφορετικά πρότυπα δημιουργίας, πρόσβασης και διαχείρισης πόρων (π.χ. http, e-mail, MIME κλπ) υπάρχει δυνατότητα ορισμού του βασικού μονοπατιού έτσι ώστε να είναι εφικτή η σχετική αναφορά σε πόρους.

Παράδειγμα 2.4. Η HTML μας δίνει τη δυνατότητα να ορίσουμε το βασικό μονοπάτι μιας σειράς αντικειμένων στο τμήμα *HEAD* μιας σελίδας, με την ετικέτα **<base>**:

```
<html>
 <head>
      <title>An example</title>
      <BASE href="http://www.ionio.gr/Test/a/b/c">
</head>
<body> ...
```

Παράδειγμα 2.5. Έστω το βασικό μονοπάτι **http://a/b/c/d/e/f**

το σχετικό μονοπάτι **#s** είναι ισοδύναμο με το απόλυτο μονοπάτι **http://a/b/c/d/e/f#s**

το σχετικό μονοπάτι **g** ή **.g** είναι ισοδύναμο με το απόλυτο μονοπάτι **http://a/b/c/d/e/g**

το σχετικό μονοπάτι **g#s** είναι ισοδύναμο με το απόλυτο μονοπάτι http://a/b/c/d/e/g#s
το σχετικό μονοπάτι /g είναι ισοδύναμο με το απόλυτο μονοπάτι http://a/g
το σχετικό μονοπάτι //g είναι ισοδύναμο με το απόλυτο μονοπάτι http://g
το σχετικό μονοπάτι .. είναι ισοδύναμο με το απόλυτο μονοπάτι http://a/b/c/d
το σχετικό μονοπάτι ../g είναι ισοδύναμο με το απόλυτο μονοπάτι http://a/b/c/d/g
το σχετικό μονοπάτι ../../g είναι ισοδύναμο με το απόλυτο μονοπάτι http://a/b/c/g

2.3. URL – Ενιαίοι Ταυτοποιητές Τοποθεσίας

2.3.1. Ορισμοί

Η μεγάλη ποικιλία μεθόδων προσπέλασης πόρων στο διαδίκτυο έχει σαν αποτέλεσμα την ύπαρξη διαφορετικών σχημάτων περιγραφής των τοποθεσιών των πόρων. Η ενιαία περιγραφή των τοποθεσιών των πόρων και των μεθόδων προσπέλασής τους εξασφαλίζει ότι τα πληροφοριακά συστήματα θα μπορούν με ένα κοινό και διαφανή τρόπο να έχουν πρόσβαση στους πόρους του διαδικτύου. Το URL είναι ένα αντικείμενο που αποτελείται από μια σειρά χαρακτήρων και περιγράφει την τοποθεσία ενός πόρου στο διαδίκτυο.

Οι βασικές ιδιότητες των URL συνοψίζονται στα ακόλουθα:
- Ένα URL ταυτοποιεί την τοποθεσία ενός και μόνο ενός πόρου. Ωστόσο ο ίδιος πόρος μπορεί να έχει περισσότερα από ένα URL, ανάλογα με τα σχήματα που τον προσπελαύνουν.
- Η πληροφορία ενός URL για ένα πόρο περιορίζεται στο μηχανισμό προσπέλασής του
- Ένα URL δεν εξασφαλίζει την προσπέλαση του πόρου, αλλά αναφέρεται στον τρόπο προσπέλασής του και την τοποθεσία του.

2.3.2. Σύνταξη

Τα URLs αποτελούν υποσύνολο (ή υποκατηγορία) των URIs και προσδιορίζουν πόρους μέσω του μηχανισμού προσπέλασής τους. Για αυτό το λόγο ακολουθούν τους βασικούς κανόνες σύνταξης των URIs. Έτσι, στα URL ισχύουν όλα όσα αναφέρθηκαν για τους δεσμευμένους χαρακτήρες, την απαίτηση αναγνωσιμότητάς τους από ηλεκτρονικά και συμβατικά μέσα, τη σύσταση για χρήση χαρακτήρων του κώδικα ASCII, τις ακολουθίες διαφυγής και τα μονοπάτια στις προηγούμενες παραγράφους για τα URIs.

Η βασική σύνταξη των URL είναι:

<scheme:> //<user>:<password>@<host>:<port>/<path>

όπου:

- *scheme* (σχήμα) δηλώνει τον τρόπο προσπέλασης του πόρου (π.χ. ftp, http κλπ).
- *user* είναι ένα όνομα (user name, ή διακριτικό ή λογαριασμός) κάποιου χρήστη, password είναι προαιρετικό και αντιστοιχεί στον συνθηματικό κωδικό αναγνώρισης του χρήστη.
- *host* είναι είτε το όνομα του υπολογιστή που φιλοξενεί τον πόρο ακολουθούμενο από το όνομα της περιοχής του διαδικτύου που ανήκει (domain name) ο υπολογιστής, είτε η διεύθυνση (IP address) του υπολογιστή στο διαδίκτυο.
- *port* είναι ο αριθμός της δικτυακής πόρτας του host η οποία αντιστοιχεί στο σχήμα. Τα περισσότερα σχήματα καθορίζουν πρωτόκολλα που έχουν μια συγκεκριμένη πόρτα επικοινωνίας (π.χ. η προκαθορισμένη πόρτα του σχήματος ftp είναι η 21, του http η 80 κλπ). Αν δεν αναφέρεται πόρτα στο URL, εννοείται η προκαθορισμένη πόρτα του πρωτοκόλλου. Σε ένα URL επιτρέπεται η δήλωση και δεύτερης εναλλακτικής πόρτας. Σε αυτή την περίπτωση μετά από τη δήλωση της πόρτας ακολουθεί ο χαρακτήρας ";"και μετά η δεύτερη εναλλακτική πόρτα (:<port>;<optional port>).
- *path* είναι το μονοπάτι, που έχει άμεση σχέση με το σχήμα και περιέχει πληροφορίες για το πώς ο συγκεκριμένος πόρος μπορεί να προσπελασθεί. Τονίζεται ότι ο χαρακτήρας "/" πριν το μονοπάτι δεν είναι μέρος του μονοπατιού.

Παράδειγμα 2.6. Η σύνταξη URL για το σχήμα http είναι http://<host>:<port>/<path>?<searchpart>, όπου τα <host> και <port>, έχουν αναλυθεί παραπάνω. Αν δεν ορίζεται συγκεκριμένη πόρτα, τότε η προκαθορισμένη πόρτα του σχήματος είναι η 80. Το σχήμα δεν προβλέπει τη χρήση username και password, ενώ και η δήλωση μονοπατιού είναι προαιρετική. Το <searchpart> ακολουθεί τον χαρακτήρα "?" και δηλώνει τις παραμέτρους ενός ερωτήματος (query).

Παράδειγμα 2.7. Το URL ftp://myname@host.dom/%2Fetc/motd αναφέρεται στον υπολογιστή host.dom στον οποίο θα συνδεθεί ο χρήστης με username myname. Όταν κληθεί το πρωτόκολλο ftp για την ανάκτηση του πόρου motd τότε ο host.dom θα ζητήσει το password του χρήστη myname. Επίσης το URL ορίζει το μονοπάτι /etc προκειμένου να προσπελάσει τον πόρο. Η ακολουθία διαφυγής "/%2F" αντιστοιχεί στο χαρακτήρα "/" και χρησιμοποιείται για να δηλώσει ότι δεύτερος χαρακτήρας "/" ανήκει στο μονοπάτι και δεν σχετίζεται (διαφεύγει) με τον πρώτο στη σειρά χαρακτήρα "/", έτσι ώστε να σχηματίσουν τη δεσμευμένη συμβολοσειρά "//", η οποία έχει άλλη σημασία. Επίσης Το URL ftp://myname@host.dom/%2Fetc/motd είναι διαφορετικό από το URL ftp://myname@host.dom/etc/motd, γιατί το

πρώτο ορίζει ότι ο πόρος ανήκει στο μονοπάτι `/etc/motd`, ενώ το δεύτερο ορίζει ότι ο κατάλογος etc βρίσκεται ένα επίπεδο ιεραρχίας κάτω από τον προκαθορισμένο κατάλογο του χρήστη `myname`.

2.3.3. Σχήματα

Η σύνταξη URL δημιουργήθηκε προβλέποντας ότι θα δημιουργηθούν και θα χρησιμοποιούνται νέα σχήματα και πρωτόκολλα για την προσπέλαση πόρων στο διαδίκτυο. Για τη δημιουργία, αποδοχή και καθιέρωση ενός νέου σχήματος, το W3C έχει συστήσει συγκεκριμένες διαδικασίες. Η καθιέρωση ενός σχήματος προϋποθέτει την έγκρισή του από κάποιο οργανισμό που είναι εξουσιοδοτημένος για την εργασία αυτή και διατηρεί μητρώο σχημάτων που ονομάζονται δέντρα μητρώων (registration trees). Ένας τέτοιος οργανισμός είναι ο Internet Engineering Task Force (IETF), ο οποίος έχει καταγράψει στο μητρώο του τα ακόλουθα σχήματα:

- http: hypertext transfer protocol
- ftp: file transfer protocol
- gopher: Gopher protocol
- mailto:Electronic mail address
- mid: Message identifiers for electronic mail (ταυτοποιητές μηνυμάτων ηλεκτρονικού ταχυδρομείου)
- cid: Content identifiers for MIME body part (ταυτοποιητές περιεχομένου που υπάρχει σε μηνύματα ηλεκτρονικού ταχυδρομείου που έχουν διαφορετική κωδικοποίηση από τον ASCII κώδικα)
- news: Usenet news
- nntp: Usenet news for local NNTP access only (νέα του Usenet που προσπελαύνονται μέσω NNTP πρωτοκόλλου)
- prospero: Το prospero περιλαμβάνει τα πρωτόκολλα telnet, rlogin και tn3270 για τη διάδραση με (απομακρυσμένους) υπολογιστές
- wais: Wide Area Information Servers

2.4. URN – Ενιαίοι Ταυτοποιητές Ονομάτων

2.4.1. Ορισμοί

Στόχος των URN είναι η απόδοση μιας καθολικά μοναδικής και μόνιμης ταυτότητας σε ένα πόρο που θα επιτρέπει την πρόσβαση είτε σε αυτόν είτε στα χαρακτηριστικά του. Ένα URN αποδίδει ένα μόνιμο όνομα σε ένα πόρο ή μια "πληροφοριακή μονάδα", ανεξάρτητα από την τοποθεσία που βρίσκεται. Συνοπτικά τα URN είναι το υποσύνολο των URI που προσδιορίζει πόρους ή πληροφορίες, ακόμα και όταν δεν είναι διαθέσιμοι, μέσω ενιαίων, καθολικών και μόνιμων ονομάτων. Η βασική διαφορά των URN με τα URL είναι ότι για ένας πόρος μπορεί να διαθέτει ένα μόνο URN, αλλά μπορεί να βρίσκεται σε περισσότερες από μία διαφορετικές τοποθεσίες ή να μην είναι διαθέσιμος, δηλ. να διαθέτει κανένα ή περισσότερα από ένα URLs.

Η ιδέα της ύπαρξης καθολικών ονομάτων για τους πόρους του διαδικτύου βρήκε θερμούς υποστηριχτές τις επιστημονικές κοινότητες των βιβλιοθηκών και των δικτύων υπολογιστών, οι οποίες διατύπωσαν τα βασικά χαρακτηριστικά των URN. Τα URN χρησιμεύουν σε εφαρμογές διαχείρισης περιεχομένου στο διαδίκτυο, όπως την ανάπτυξη καταλόγων και αποθηκών κατανεμημένων πόρων στο δίκτυο καθώς επίσης στην ανάπτυξη πολιτικών ασφάλειας με χρήση ψηφιακών υπογραφών (digital signatures) και εφαρμογών ελέγχου αυθεντικότητας (authentication) και ακεραιότητας (integrity) των πόρων.

Ένα URN αποδίδεται σε ένα πόρο από μια ανεξάρτητη αρχή ονοματοδοσίας με βάση κάποια διαδικασία δημιουργίας και ανάθεσης ονομάτων που ονομάζεται σχήμα ονοματοδοσίας (naming scheme). Κάθε σχήμα ονοματοδοσίας διαθέτει ένα μοναδικό τρόπο σύνταξης των ονομάτων που δημιουργεί και καθορίζει ένα χώρο ονομάτων (name space).

Τα κυριότερα χαρακτηριστικά των URN είναι:
- *καθολικότητα*, ένα URN είναι ένα όνομα που χαρακτηρίζει ένα πόρο οπουδήποτε βρίσκεται.
- *μοναδικότητα*, δεν επιτρέπεται σε δύο πόρους να μοιράζονται το ίδιο URN.
- *μονιμότητα*, δεν υπάρχει χρονικό όριο για τη ζωή ενός URN. Αυτό σημαίνει ότι το URN μπορεί να συνεχίζει να χρησιμοποιείται ως αναφορά σε ένα πόρο πέρα από το χρονικό περιθώριο της ζωής του πόρου.
- *διάρκεια*, τα URNs μπορούν να ανατίθενται σε οποιοδήποτε πόρο, που μπορεί να είναι διαθέσιμος στο δίκτυο για απεριόριστο χρονικό διάστημα.
- *κληρονομιά*, υποστηρίζονται και διατηρούνται τα ονόματα πόρων που παρέχονται από υπάρχοντα συστήματα ονοματοδοσίας, εφόσον αυτά τα συστήματα ικανοποιούν και τις υπόλοιπες ιδιότητες των URN. Τέτοια συστήματα είναι τα οι αριθμοί ISBN, ISO κά.
- *επεκτασιμότητα*, ένα URN βασίζεται σε κάποιο σχήμα (scheme), το οποίο με τη σειρά του ορίζει ένα χώρο ονομάτων, που μπορεί να επεκτείνεται.
- *ανεξαρτησία*, τα URNs αποδίδονται στους πόρους αποκλειστικά από οργανισμούς – αναθέτουσες αρχές, οι οποίες αποφασίζουν αυτόνομα για τον τρόπο απόδοσης του ονόματος.
- *ανάλυση*, για τη χρήση των πόρων πρέπει να υπάρχει κάποια διαδικασία η οποία με δεδομένο το όνομα ενός πόρου να μπορεί να τον εντοπίζει στο διαδίκτυο. Η διαδικασία αυτή ονομάζεται ανάλυση (resolution) και αρχικά εστίαζε στην αντιστοίχηση των ονομάτων των πόρων (URN) με τις τοποθεσίες (URLs) τους.

2.4.2. Ανάλυση και Μητρώα URN

Προβλέπεται η δημιουργία διάφορων συστημάτων ανάλυσης των URNs, τα οποία θα είναι ανεξάρτητα από τα σχήματα, δηλ. θα έχουν τη δυνατότητα να αναλύουν URN από οποιοδήποτε σχήμα ονοματοδοσίας. Επιπλέον προβλέπεται και η ύπαρξη διάφορων αρχών που θα διαχειρίζονται τα σχήματα ονοματοδοσίας και θα διαθέτουν μοναδικά ονόματα σε πόρους.

Λόγω του πλήθους και της ποικιλίας των σχημάτων ονοματοδοσίας και των συστημάτων ανάλυσης, είναι απαραίτητη η ύπαρξη μητρώων URN, (URN registry), τα οποία είναι μηχανισμοί που καταγράφουν τα συστήματα ανάλυσης και τις αρχές διαχείρισης σχημάτων ονοματοδοσίας. Η ύπαρξη των μητρώων διευκολύνει τους χρήστες ενός URN να ανακαλύψουν τα υπάρχοντα συστήματα για την ανάλυση του URN. Επιπλέον μέσω των μητρώων URN, είναι δυνατή η ταυτοποίηση των ανεξάρτητων αρχών που διαχειρίζονται τα σχήματα ονοματοδοσίας και αποδίδουν URN.

2.4.3. Σύνταξη

Τα URN αποτελούν υποκατηγορία των URIs και ακολουθούν τους βασικούς κανόνες σύνταξής τους. Έτσι, στα URN ισχύουν όλα όσα αναφέρθηκαν στις προηγούμενες παραγράφους για τα URIs. Τα URN πρέπει να είναι μεταφέρσιμα μέσω email, ftp κλπ. Για αυτό πρέπει να χρησιμοποιείται μικρό αλφάβητο, χωρίς διάκριση πεζών κεφαλαίων, με λίγους ειδικούς χαρακτήρες.

Το βασικό συντακτικό των URN είναι:

<urn>:<NID>:<NSS>

όπου:

- *urn* είναι το URI σχήμα που δηλώνει ότι ό,τι ακολουθεί είναι URN.
- *NID* (namespace identifier), προσδιορίζει το σχήμα ονοματοδοσίας και καθορίζει το συντακτικό του NSS. Όλα τα URN που ακολουθούν αυτό το σχήμα συντάσσονται με τον ίδιο τρόπο.
- *NSS* (namespace specific string), είναι μια ακολουθία χαρακτήρων που ακολουθεί το συντακτικό που ορίζει το σχήμα ονοματοδοσίας (NID) στο οποίο ανήκει το URN.

Παράδειγμα 2.8. Τα ακόλουθα είναι URN:

urn:path:/A/B/C/doc.html

urn:inet:cnri.dlib/august98

urn:isbn:0-07-112779-8

Στις παραπάνω δηλώσεις μετά το σχήμα urn: ακολουθεί ο ταυτοποιητής του σχήματος ονοματοδοσίας (NID path, inet), ο χαρακτήρας ":" και μια ακολουθία χαρακτήρων που το συντακτικό τους καθορίζεται από το σχήμα ονοματοδοσίας (NSS). Μερικά NSS αποτελούνται από δύο μέρη. Το πρώτο μέρος δηλώνει το δικτυακό μονοπάτι (π.χ. /A/B/C/) ή την αρχή

ονοματοδοσίας (π.χ. cnri.dlib) και το δεύτερο μέρος είναι μια ακολουθία χαρακτήρων (π.χ. doc.html), που δηλώνει το μοναδικό όνομα.

2.4.4. Σχήματα ονοματοδοσίας

Υπάρχουν διάφορα σχήματα ονοματοδοσίας, με αντίστοιχους ταυτοποιητές, (NID) τα οποία καταγράφονται από τον οργανισμό Internet Assigned Numbers Authority (IANA). Όπως στα URL σχήματα απαιτούνται συγκεκριμένες διαδικασίες για την αποδοχή και καταγραφή ενός σχήματος, έτσι και ο IANA ακολουθεί συγκεκριμένες διαδικασίες για την αποδοχή ενός σχήματος ονοματοδοσίας και την παραχώρηση ενός NID. Ο οργανισμός διακρίνει τα URN στις ακόλουθες κατηγορίες:

- Καταγεγραμμένα (registered) τα οποία διακρίνονται σε δύο υποκατηγορίες: Τα formal URN (μέχρι το Σεπτέμβριο του 2001 ήταν οκτώ, μεταξύ αυτών και το issn) που αποτελούν τον επίσημο κατάλογο των σχημάτων ονοματοδοσίας και τα informal (μέχρι το Σεπτέμβριο του 2001 ήταν τρία τα urn-1, urn-2, urn-3).
- Εκκρεμή (pending) π.χ. μέχρι το Σεπτέμβριο του 2001 δεν είχε αποφασισθεί αν το "isbn" θα αποτελούσε καταγεγραμμένο NID.
- Μη καταγεγραμμένα (unregistered), για τα οποία δεν έχει ξεκινήσει η διαδικασία καταγραφής και απόδοσης NID.

2.5. Βιβλιογραφία

1. T. Berners-Lee, R. Fielding, L. Masinter. 1998. Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396, IETF, http://www.ietf.org/rfc/rfc2396.txt (ημ/νία πρόσβασης 27.02.2004).
2. T. Berners-Lee. 1994. Universal Resource Identifiers in WWW. RFC 1630, http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1630.html (ημ/νία πρόσβασης 09.03.2004).
3. T. Berners-Lee, L. Masinter. 1994. Uniform Resource Locators (URL). RFC 1738, IETF.
4. J. Kunze. 1995. Functional Recommendations for Internet Resource Locators. RFC 1736, http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1736.html (ημ/νία πρόσβασης 09.03.2004).
5. L. Masinter, H. Alvestrand, D. Zigmond, R. Petke. 1999. Guidelines for new URL schemes. RFC 2718, IETF.
6. R. Moats. 1997. URN Syntax. RFC 2141, IETF, http://www.ietf.org/rfc/rfc2141.txt (ημ/νία πρόσβασης 16.03.2004).
7. R. Petke, I. King. 1999. Registration Procedures fo URL Scheme Names. RFC 2717, IEFT, http://www.ietf.org/rfc/rfc2717.txt (ημ/νία πρόσβασης 27.02.2004).
8. K. Sollins. 1998. Architectural Principles of Uniform Resource Name Resolution. RFC 2276, IETF, http://www.ietf.org/rfc/rfc2276.txt (ημ/νία πρόσβασης 27.02.2004).

9. K. Sollins, L. Masinter. 1994. Functional Requirements for Uniform Resource Names. RFC 1737, http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1737.html (ημ/νία πρόσβασης 09.03.2004).
10. URN Implementors. 1996. Uniform Resource Names: A Progress Report. D-Lib Magazine, February 1996.

2.6. Ευρετήριο Όρων