Εξέταση 102 – Μάθημα 9

107.3 Τοπικοποίηση και Διεθνοποίηση

Τοπικοποίηση και Διεθνοποίηση

- **Τοπικοποίηση (Localization)** είναι η προσαρμογή ενός λογισμικού από μια γλώσσα (συνήθως Αγγλικά) σε μια άλλη. Χρησιμοποιείται και η συντομογραφία **L10n**
- **Διεθνοποίηση(Internationalization)** είναι η προετοιμασία ενός λογισμικού για να μπορεί εύκολα να προσαρμοστεί σε άλλες γλώσσες. Χρησιμοποιείται και η συντομογραφία **i18n**

Ρύθμιση ζώνης ώρας (timezone)

- **/etc/localtime**: το αρχείο αυτό μπορεί να είναι είτε ένα συμβολικός σύνδεσμος που παραπέμπει σε κάποιο αρχείο στον κατάλογο **/usr/share/zoneinfo**, ή ένα αντίγραφο από αυτό τον κατάλογο. Με αυτό τον τρόπο προκαθορίζεται η ζώνη ώρας ολόκληρου του συστήματος
- **/etc/timezone** # είναι ένα αρχείο κειμένου που περιέχει την ζώνη συστήματος. Υπάρχει μόνο σε Debian. Για RedHat υπάρχει το **/etc/sysconfig/clock**:

  $ **cat /etc/timezone**

  Asia/Nicosia

Σημείωση: υπάρχουν άλλα εργαλεία που κάνουν καλύτερη δουλειά από το να αντιγράψουμε ή να δημιουργήσουμε ένα συμβολικό σύνδεσμο από το **/etc/share/zoneinfo**. Αν χρησιμοποιήσουμε αυτή την μέθοδο θα πρέπει να αλλάξουμε και την τιμή στα **/etc/timezone** ή **/etc/sysconfig/clock**

Προβολή ρυθμίσεων τοπικοποίησης με *locale*

- Η εντολή **locale** προβάλλει πληροφορίες σχετικές με την τοπικοποίηση
- **# locale** # προβολή μεταβλητών σχετικών με την τοπικοποίηση
- **# locale -a** # προβολή των εγκατεστημένων τοπικοποιήσεων (locales)

Μεταβλητές τοπικοποιήσεων

LANG=en_US.UTF-8 # Αυτή η μεταβλητή ορίζει την γενική τοπικοποίηση του συστήματος επιτρέποντας παραμετροποιήσεις

LANGUAGE=el:en_US:en # η συγκεκριμένη μεταβλητή είναι μια λίστα με τοπικοποιήσεις και διαβάζεται από εφαρμογές που υποστηρίζουν το σύστημα GNU gettext

LC_CTYPE=el_GR.UTF-8 # καθορίζει ποιοι χαρακτήρες θα είναι αλφαβητικοί ή αριθμητικοί για κάθε τοπικοποίηση

LC_NUMERIC="en_US.UTF-8" # καθορίζει τα διαχωριστικά των δεκαδικών και χιλιάδων. Πχ στα αγγλικά λένε 1,000.00 ενώ στα ελληνικά 1.000,00

LC_TIME="en_US.UTF-8" # καθορίζει την μορφή της ώρας και ημερομηνίας

LC_COLLATE=el_GR.UTF-8 # καθορίζει την αλφαβητική σειρά των χαρακτήρων όταν ταξινομούμε

LC_MONETARY="en_US.UTF-8" # καθορίζει το σύμβολο του νομίσματος ($, €)

LC_MESSAGES=el_GR.UTF-8 # καθορίζει την γλώσσα των μυνημάτων συστήματως και εφαρμογών

LC_PAPER="en_US.UTF-8" # καθορίζει το μέγεθος του χαρτιού, πχ letter ή A4

LC_NAME="en_US.UTF-8" # καθορισμός μορφοποίησης ονομάτων

LC_ADDRESS="en_US.UTF-8" # καθορισμός μορφοποίησης διευθύνσεων

LC_TELEPHONE="en_US.UTF-8" # καθορισμός μορφοποίησης τηλεφώνων

LC_MEASUREMENT="en_US.UTF-8" # καθορισμός μονάδων μέτρησης πχ SI, μετρικό κτλ

LC_IDENTIFICATION="en_US.UTF-8" # μεταδεδομένα και πληροφορίες για την τοπικοποίηση

LC_ALL= # αυτή η μεταβλητή αν καθοριστεί έχει προτεραιότητα σε σχέση με τις υπόλοιπες μεταβλητές οι οποίες θα αγνοηθούν

Ρύθμιση προκαθορισμένης τοπικοποίησης

- Αν καθορίσουμε τις μεταβλητές γλώσσας στο **/etc/profile** ή σε κάποιο αρχείο κάτω από /etc/profile.d πχ **/etc/profile.d/lang.sh** αυτό θα είναι και το προκαθορισμένο για όλους μέχρι κάποιος να καθορίσει διαφορετικές τιμές από το δικό του **.bash_profile** ή **.profile**
- Στα συστήματα βασισμένα σε debian υπάρχει το αρχείο **/etc/default/locale** όπου μπορούν να προκαθοριστούν οι μεταβλητές γλώσσας. Επίσης με την εντολή **'dpkg-reconfigure locales'** μπορούμε να εγκαταστήσουμε επιπρόσθετες τοπικοποιήσεις καθώς και να επιλέξουμε την προκαθορισμένη τοπικοποίηση
- Στα συστήματα βασισμένα σε RedHat υπάρχει το αρχείο **/etc/sysconfig/i18n** όπου μπορούν να δηλωθούν οι μεταβλητές γλώσσας και επίσης μπορούμε να χρησιμοποιήσουμε την εντολή **'system-config-language'**

Προβλήματα με Τοπικοποιήσεις και LANG=C

- Πολλές φορές μπορεί να υπάρχει μια ρυθμισμένη τοπικοποίηση η οποία δεν είναι εγκατεστημένη στο σύστημα. Για παράδειγμα η εντολή "**locale**" να δείχνει **el_GR.UTF-8** αλλά η εντολή "**locale -a**" να μην έχει το **el_GR.UTF-8** στην λίστα των εγκατεστημένων τοπικοποιήσεων. Αυτό θα δημιουργούσε προβλήματα στην χρήση του κελύφους μια και αρκετές εφαρμογές δεν συμπεριφέρονται σωστά όταν δεν υπάρχουν οι σωστές τοπικοποιήσεις. Μπορεί επίσης κάποια σενάρια bash να μην φέρνουν τα αναμενόμενα αποτελέσματα για τον ίδιο λόγο.
- Η λύση είναι να χρησιμοποιήσουμε κάποια τοπικοποίηση που είναι πάντοτε προεγκατεστημένη στο σύστημα. Η τοπικοποίηση αυτή είναι η **C** καθώς επίσης και η **POSIX**. Αν ρυθμίσουμε τις μεταβλητές **LANG** και **LC_*** σε αυτές θα απαλείψουμε τα προβλήματα που αφορούν τοπικοποιήσεις
- Αυτό μπορεί να συμβεί και όταν συνδεόμαστε σε κάποιο άλλο διακομιστή (πχ με ssh) και έχουμε στο τοπικό σύστημα μια τοπικοποίηση η οποία δεν είναι εγκατεστημένη στο διακομιστή

Οι εντολές tzselect και tzconfig

- **$ tzselect** # είναι μια εντολή διεπαφής κονσόλας που σου δείχνει την ώρα σε άλλη ζώνη χωρίς να την αλλάξει. Στο τέλος σου λέει τι πρέπει να κάνεις για να την αλλάξεις μόνιμα.
- **# tzconfig** # η διεπαφή αυτής της εντολής είναι όμοια με της **tzselect** με την διαφορά ότι ή **tzconfig** αλλάζει την ζώνη του συστήματος σε μόνιμη βάση. Επειδή θεωρείται ξεπερασμένη αρκετά καινούργια συστήματα δεν την περιλαμβάνουν. Για παράδειγμα σε Debian χρησιμοποιείται ή '**dpkg-reconfigure tzdata**' και σε RedHat η '**system-config-time**'

Προβολή ώρας με date

- Η εντολή **date** χρησιμοποιείται για τον καθορισμό της ώρας/ημερομηνίας του συστήματος αλλά και για την προβολή της ώρας/ημερομηνίας σε διάφορες μορφές
- **# date** # προβολή της τοπικής ώρας όπως καθορίζεται από το **/etc/localtime**
- **# date -u** # προβολή της διεθνούς ώρας UTC (Universal Time Coordinated) ή Παγκόσμιου Χρόνου
- **# date +%Z** # προβολή του κώδικα της τρέχουσας ζώνης ώρας

Κωδικοποιήσεις χαρακτήρων – ASCII

- Το **ASCII (American Standard Code for Information Interchange)** είναι μια κωδικοποίηση χαρακτήρων (character encoding) για το αγγλικό αλφάβητο που αποτελείται από 7 μπιτ. Αυτό αντιστοιχεί σε 128 χαρακτήρες (2^7) εκ των οποίων οι 33 είναι μη εκτυπωμένοι χαρακτήρες ελέγχου.
- Το πρόβλημα με το ASCII είναι η μη υποστήριξη του για άλλα αλφάβητα όπως Ελληνικό, Κυριλλικό ή ακόμη και λατινικά αλφάβητα όπως τα Γαλλικά που περιέχουν παράξενα σημεία
- Η εντολή '**ascii**' θα μας τυπώσει ένα πίνακα με τις αντιστοιχίες των χαρακτήρων ASCII και το δεκαεξαδικό ισοδύναμο τους

Κωδικοποιήσεις χαρακτήρων – ISO-8859

- Η ακολουθία προτύπων ISO-8859 ήταν μια πρώτη προσπάθεια για διεθνοποίηση κωδικοποιήσεων χαρακτήρων. Κάθε χαρακτήρας αποτελείται από 8 μπιτ και παρέχει περισσότερη ευελιξία από το ASCII με 256 συνδυασμούς (2^8).
- Παρόλα αυτά 256 συνδυασμοί δεν είναι αρκετοί για όλες τις γλώσσες και έτσι δημιουργήθηκε μια διαφορετική παραλλαγή για κάθε σύστημα γραφής. Αυτό σημαίνει ότι πρέπει να επανακαθορίζεται η κωδικοποίηση όταν αλλάζουμε σύστημα:
  - **ISO-8858-1**: Αγγλικά, Γερμανικά, Νορβηγικά, Ισπανικά κλπ
  - **ISO-8859-2**: Σέρβικα, Πολωνέζικα, Τσέχικα, Ουγγρικά κλπ
  - **ISO-8859-5**: Κυριλλικό
  - **ISO-8859-6**: Αραβικά
  - **ISO-8859-7**: Ελληνικά χωρίς πολυτονικά διακριτικά (Σχεδόν ίδιο με την κωδικοποίηση Windows-1253)
  - **ISO-8859-9**: Τουρκικά

Κωδικοποιήσεις χαρακτήρων – Unicode

- Το Unicode είναι ένα, πολλαπλών ψηφιολέξεων (multibyte), πρότυπο για την ενιαία κωδικοποίηση χαρακτήρων κειμένου. Ο σκοπός είναι να προσπεραστεί το εμπόδιο των 7-8 μπιτ που είχαν οι κωδικοποιήσεις ASCII και ISO-8859. Χρησιμοποιώντας από 1 μέχρι και 4 ψηφιολέξεις (bytes) για κάθε χαρακτήρα.
- Τα πρώτα 127 ψηφία έχουν την ίδια κωδικοποίηση όπως και το ASCII και έτσι υπάρχει προς τα πίσω συμβατότητα
- Περιλαμβάνει περιλαμβάνει περισσότερους από 110.000 χαρακτήρες και 100 συστήματα γραφής, από Ελληνικά (συμπεριλαμβανομένων και των πολυτονικών) και Κινέζικα μέχρι εξαφανισμένα συστήματα όπως Γραμμική Β και Κυπροσυλλαβική γραφή.
- Τα Ελληνικά χρειάζονται δύο ψηφιολέξεις για να αναπαρασταθούν και τα Κινέζικα τρεις
- Χρησιμοποιούνται 3 κωδικοποιήσεις χαρακτήρων το **UTF-8**, το UCS-2 και το UTF-16 με το πρώτο να είναι το δημοφιλέστερο.

Κωδικοποιήσεις χαρακτήρων – **UTF-8**

- Το **UTF-8** είναι μια κωδικοποίηση χαρακτήρων μεταβλητού μήκους (1-4 ψηφιολέξεις) που μπορεί να αναπαραστήσει όλη την λίστα χαρακτήρων του Unicode
- Παρέχει το πλεονέκτημα να μπορούμε να αλλάξουμε σύστημα γραφής χωρίς να χρειαστεί να καθορίσουμε κωδικοποίηση χαρακτήρων
- Είναι το δημοφιλέστερο σύστημα κωδικοποίησης σήμερα με ευρεία χρήση στα σύγχρονα λειτουργικά συστήματα (περιλαμβανομένου και του Linux) καθώς και τον Παγκόσμιο Ιστό (World Wide Web)

Μετατροπή χαρακτήρων από μια κωδικοποίηση σε άλλη με *iconv*

- Η εφαρμογή **iconv** μας δίνει την δυνατότητα μετατροπής κειμένου από μια κωδικοποίηση σε άλλη
- **# iconv -f WINDOWS-1253 -t ISO-8859-7 greek-1253.txt** # μετατροπή από Ελληνική κωδικοποίηση Windows-1253 σε ελληνικά ISO-8859-7 και αποτέλεσμα σε τυπική έξοδο (stdout)
- **# iconv -f ISO-8859-7 -t UTF-8 greek-8859.txt -o greek-utf8.txt** # μετατροπή από Ελληνική κωδικοποίηση ISO-8859-7 σε ελληνικά UTF-8 και αποθήκευση στο αρχείο **greek-utf8.txt**
- **# iconv -c -f UTF-8 -t ISO-8859-7 greek-utf8.txt -o greek-8859.txt** # μετατροπή από Ελληνική κωδικοποίηση UTF-8 σε ελληνικά ISO-8859-7 και αποθήκευση στο αρχείο **greek-8859.txt** παραλείποντας μη έγκυρους χαρακτήρες (-c)
- **# iconv -l** # λίστα υποστηριζόμενων κωδικοποιήσεων

Ρύθμιση συστήματος για προβολή ελληνικών σε κονσόλα

- Για **RedHat**:
  - **$ system-config-language** # και επιλέγουμε “Greek”
  - Ρύθμιση των πιο κάτω παραμέτρων σε **/etc/sysconfig/i18n**:
    LANG=el_GR.UTF-8
    SYSFONT=gr737c-8x16
  - Επανεκκίνηση
- Για **Debian**:
  - **# dpkg-reconfigure locales** # εγκατάσταση el_GR.UTF-8 και επιλογή σαν προκαθορισμένη κωδικοποίηση
  - Ρύθμιση των πιο κάτω παραμέτρων στο αρχείο **/etc/default/locales**:
    LANG=el_GR.UTF-8
    LANGUAGE=el:en_US:en
  - Επανεκκίνηση

Εργαστήριο 9

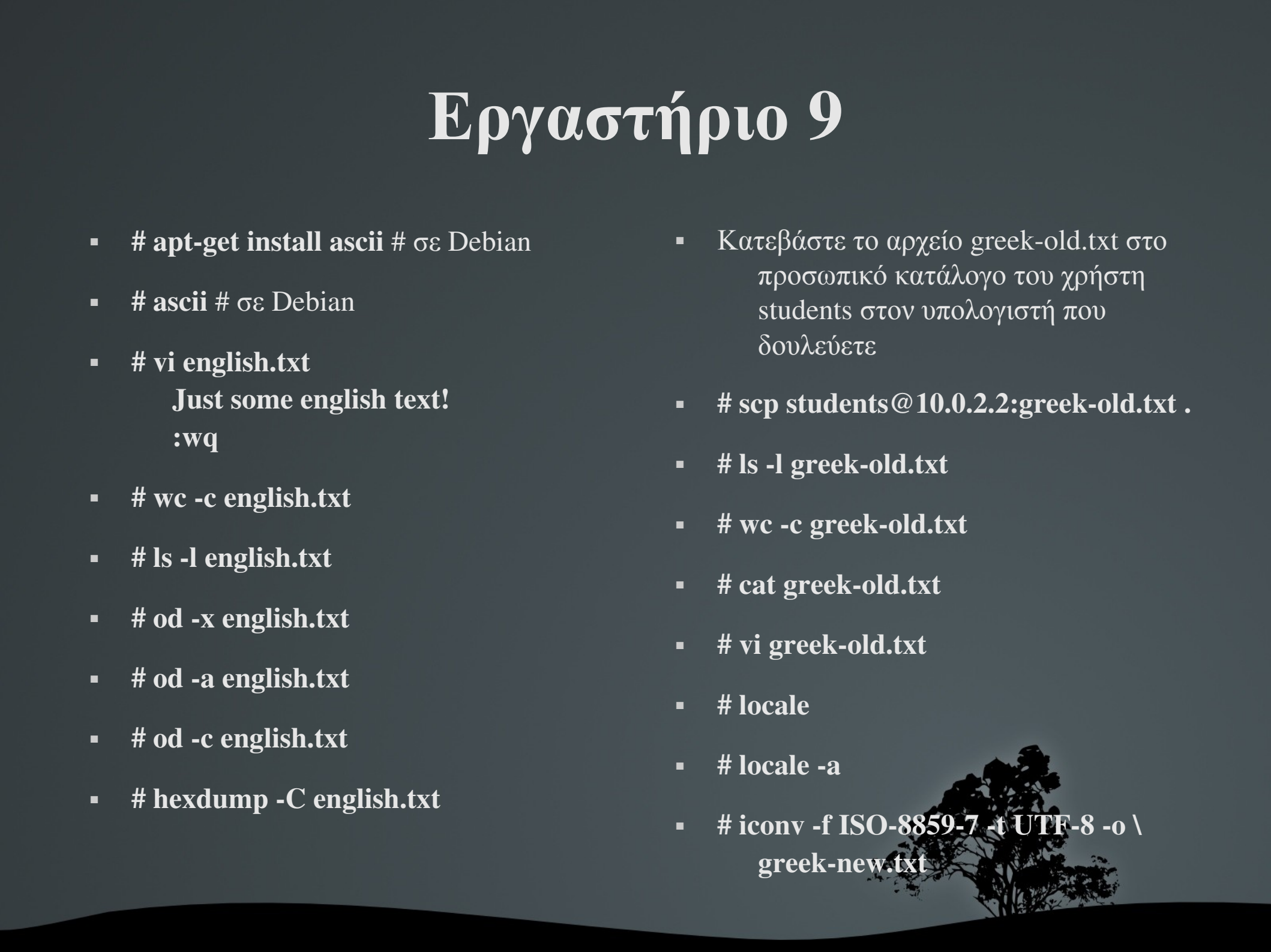
Ξεκινήστε και τις δύο εικονικές μηχανές και συνδεθείτε σαν ”root”

- **# ls -la /etc/localtime**
- **# ls -la /etc/timezone**
- **# file /etc/localtime**
- **# file /etc/timezone**
- **# find /usr/share/zoneinfo -type f**
- **# find /usr/share/zoneinfo -type f | **
  grep Nicosia
- **# locale**
- **# locale -a**
- **# date**
- **# date -u**
- **# date +%Z**
- **# date**
- **# tzselect**
  8 # (Europe)
  16 # (Germany)
- **# date**
- **# tzconfig**
- **# dpkg-reconfigure tzconfig** # σε Debian
  (αλλάξτε την ώρα σε Germany)
- **# system-config-time** # σε RedHat
- **# date**
- **# date -u**
- **# date +%Z**

Εργαστήριο 9

- **# apt-get install ascii** # σε Debian
- **# ascii** # σε Debian
- **# vi english.txt**
  Just some english text!
  :wq
- **# wc -c english.txt**
- **# ls -l english.txt**
- **# od -x english.txt**
- **# od -a english.txt**
- **# od -c english.txt**
- **# hexdump -C english.txt**

- Κατεβάστε το αρχείο greek-old.txt στο προσωπικό κατάλογο του χρήστη students στον υπολογιστή που δουλεύετε
- **# scp students@10.0.2.2:greek-old.txt .**
- **# ls -l greek-old.txt**
- **# wc -c greek-old.txt**
- **# cat greek-old.txt**
- **# vi greek-old.txt**
- **# locale**
- **# locale -a**
- **# iconv -f ISO-8859-7 -t UTF-8 -o **
  greek-new.txt



Εργαστήριο 9

- **# wc -c greek-new.txt**
- **# cat greek-new.txt**
- **# vi greek-new.txt**
- **# file greek***
- **# dpkg-reconfigure locales** # και επιλέξτε el_GR ISO-8859-7 ως προκαθορισμένη (σε Debian
- **# vi /etc/default/locale**
  LANG=el_GR.UTF8
  LANGUAGE=el:en_US:en
  :wq
- **# reboot**
- **# cat greek-old.txt**
- **# vi greek-old.txt**
- **# cat greek-new.txt**
- **# vi greek-new.txt**
- **# dpkg-reconfigure console-setup**
  # επιλέξτε ISO-8859-7
- **# reboot**
- **# cat greek-old.txt**
- **# vi greek-old.txt**
- **# cat greek-new.txt**
- **# vi greek-new.txt**
- **# dpkg-reconfigure console-setup**
  # επιλέξτε UTF-8
- **# reboot**
- **# cat greek-old.txt ; cat greek-new.txt**

Εργαστήριο 9

- **# system-config-language** # και επιλέξτε Greek σε RedHat
- **# vi /etc/sysconfig/i18n**
  LANG=el_GR.UTF-8
  SYSFONT=gr737c-8x16
  :wq
- **# reboot**
- **# locale**
- **# locale -a | grep GR**
- **# cat greek-old.txt**
- **# vi greek-old.txt**
- **# iconv -f ISO-8859-7 -t UTF-8 -o **
  greek-new.txt
- **# cat greek-new.txt**
- **# vi greek-new.txt**
- **# export LANG=NO-SUCH-LOCALE**
- **# locale**
- **# locale -a**
- **# apt-get install nethack-console**
- **# apt-get remove nethack-console**
- **# export LANG=C**
- **# locale**
- **# locale -a**
- **# apt-get install nethack-console**
- **# apt-get remove nethack-console**