AAD
Problemes. Sessió al laboratori: Codis de caràcters i correu SMTP i MIME
(1hora)

Autors: Joan Manuel Marquès i Leandro Navarro.

Introducció

Aquesta sessió de problemes ens ajudarà a entendre les diferents formes de codificació de text que s'utilitzen a gairebé totes les aplicacions com correu electrònic, directori, web, etc. També practicarem i veurem el format d'enviament que utilitzen l'SMTP i la codificació MIME.

Objectius

- Veure la implicació d'utilitzar diferents formes de codificació
- Veure l'enviament de correu utilitzant SMTP
- Veure com es codifica un missatge MIME

Tasques

1.- Veure que el navegador reconeix diferents repertoris de caràcters.

- Connecta't a la pàgina d'unicode *What is unicode?* (http://www.unicode.org/unicode/standard/WhatIsUnicode.html) utilitzant el navegador Netscape o Mozilla.
  - A la franja esquerra de la pàgina apareix un llistat d'idiomes en els que es pot consultar l'explicació d'aquesta pàgina.
  - Depenent de si el navegador reconeix o no els caràcters de cada idioma, es veu la pregunta *What is unicode?* en l'idioma corresponent o un seguit de caràcters erronis.

- Ara connecta-t'hi amb el MS-Explorer i veuràs que els idiomes reconeguts són diferents.

- Si voleu reconèixer un repertori de caràcters que el navegador que utilitzeu no us permet llegir cal instal·lar-se els codis de caràcters corresponents a aquell repertori de caràcters.

- A títol de curiositat...
  ... Windows 2000, Office 2000 porten Arial Unicode MS, amb 51180 caràcters, basat en Unicode 2.0. El fitxer ocupa 23 Mb. Si el tens instal·lat pots anar al menú **inserir – símbol**, escollir el font **Arial Unicode MS** i pots inserir algun dels 51180 símbols (= molts alfabets, i caràcters usats arreu del món. Incloent ideogrames japonesos, xinesos, koreans, ...)

2.- Veure les diferències entre utilitzar un codi de caràcters o un altre

- Obre el **WordPad**
- Escriu el text: Què tal estàs?
- Desa el text (**`Archivo -- guardar como`**) en tres fitxers diferents usant els formats:
  - `Unicode`
  - `Text - format MS-DOS`
  - `Text`
- [Si es fa en Windows'98] Obre una finestra MS-DOS i mira el contingut de cadascun dels tres fitxers que acabes de crear. Per a veure el contingut de cada fitxer fes:
  - `Type <nom_fitxer.extensió>`
  - Veuràs que el contingut de cada fitxer no correspon al text introduït.

    - Donat que els codis de caràcters en que està configurat l'MS-DOS és diferent de l'Unicode o de la configuració de windows, només surt bé el text en format MS-DOS
- Mira què ocupa cadascun dels fitxers. T'adonaràs que, segons la codificació escollida, els fitxers ocupen una mida diferent:
    - En el cas de l'Unicode que utilitza el WordPad, cada caràcter es representa en 2 bytes, A més, l'Unicode afegeix uns codis a l'inici del fitxer per a identificar quina codificació s'usa.
    - En els altres dos casos, s'utilitza un byte per a representar cada caràcter

3.- Efecte que té l'ús de diferents codificacions de caràcters en el correu electrònic.

L'objectiu d'aquesta activitat és que vegeu l'efecte que té usar diferents codificacions en l'enviament de missatges per correu electrònic.

Farem aquesta activitat usant el correu del Netscape.

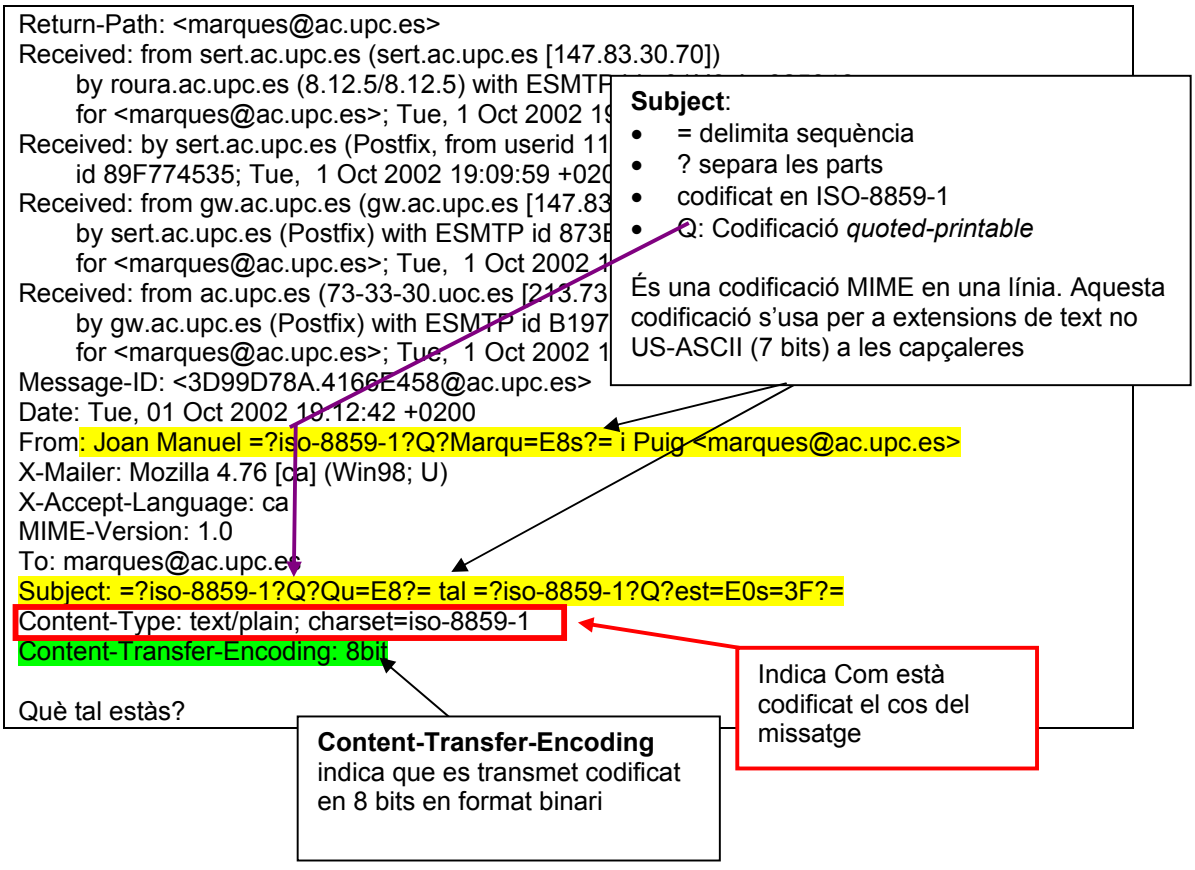
3.1 Enviament d'un missatge i veure què passa si el receptor canvia la codificació a l'hora de presentar el missatge

- Crea un missatge nou
- Posa a l'assumpte i al cos del missatge el text: **Què tal estàs?**
    - (no formategis el text, ja que el missatge s'enviaria en format HTML)

- Envia't el missatge i llegeix-lo des del gestor de correu de Netscape.
- Un cop l'hagis llegit, visualitza'l usant diferents codificacions. P.ex. posa la codificació: **Centreeuropeu (ISO-8859-2)**

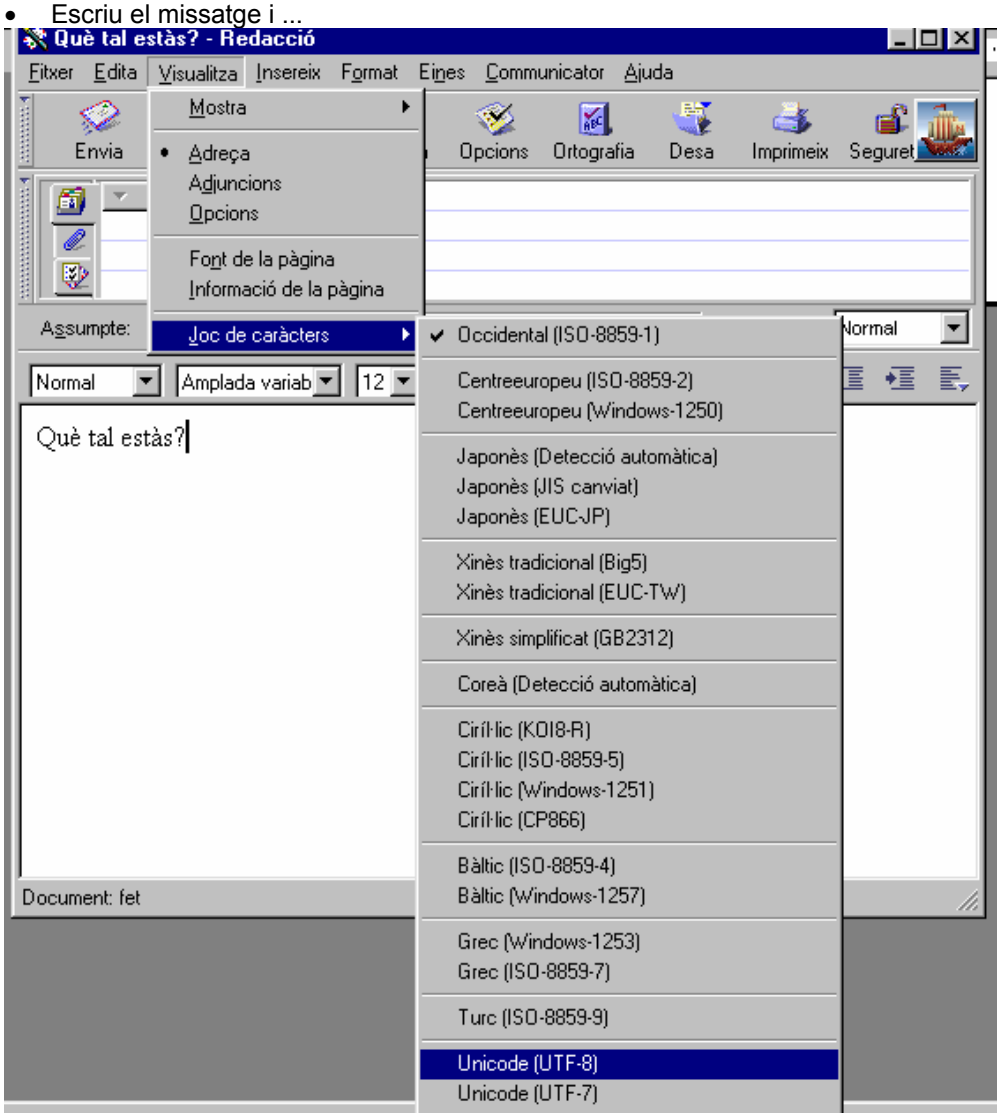
Veuràs que el text no surt correctament.

- Si mires el codi font de la pàgina (menú **Visualitza – font de la pàgina**) apareix el següent text:

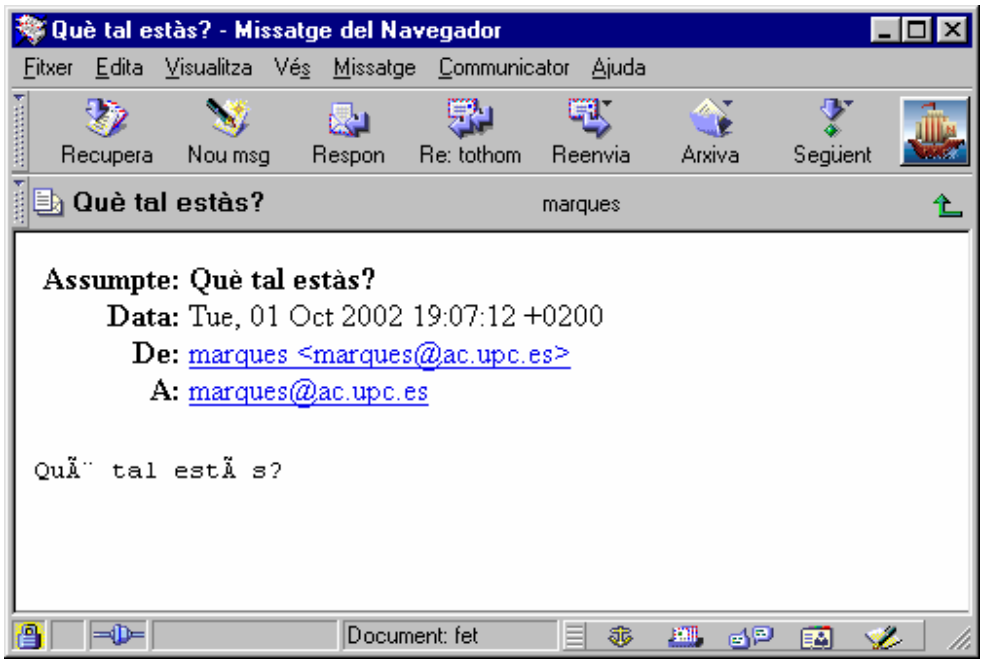


3.2 Canvia la codificació usada abans de crear i enviar el missatge.

- Crea un missatge nou
- Abans d'escriure res, escull la codificació Unicode UTF-8
- Escriu el missatge i ...



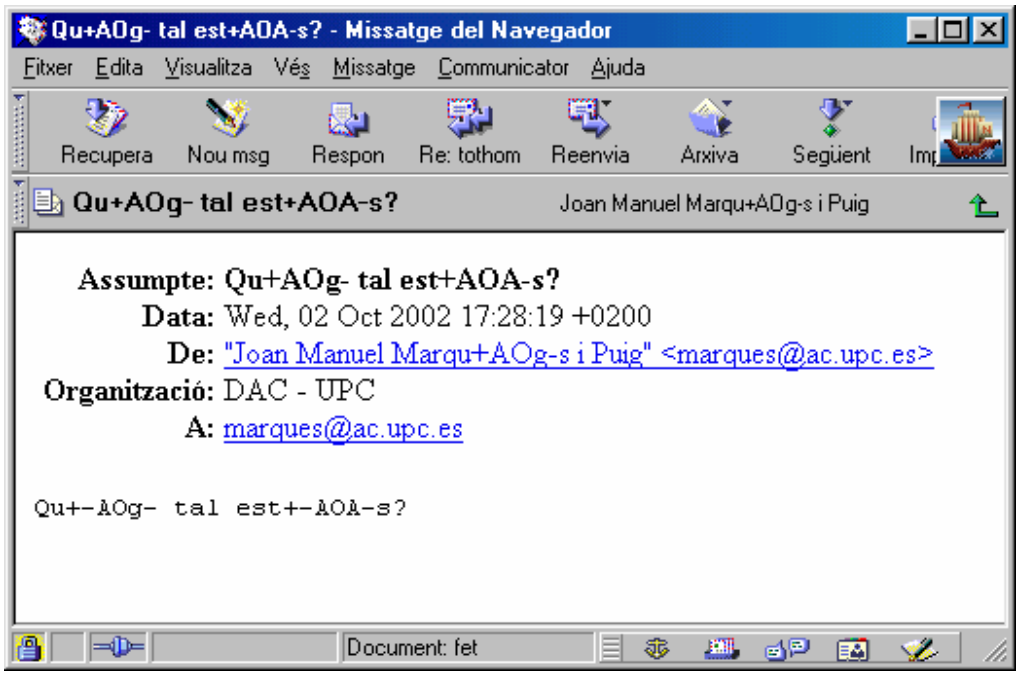
- Des del gestor de correu del Netscape, llegeix el missatge.



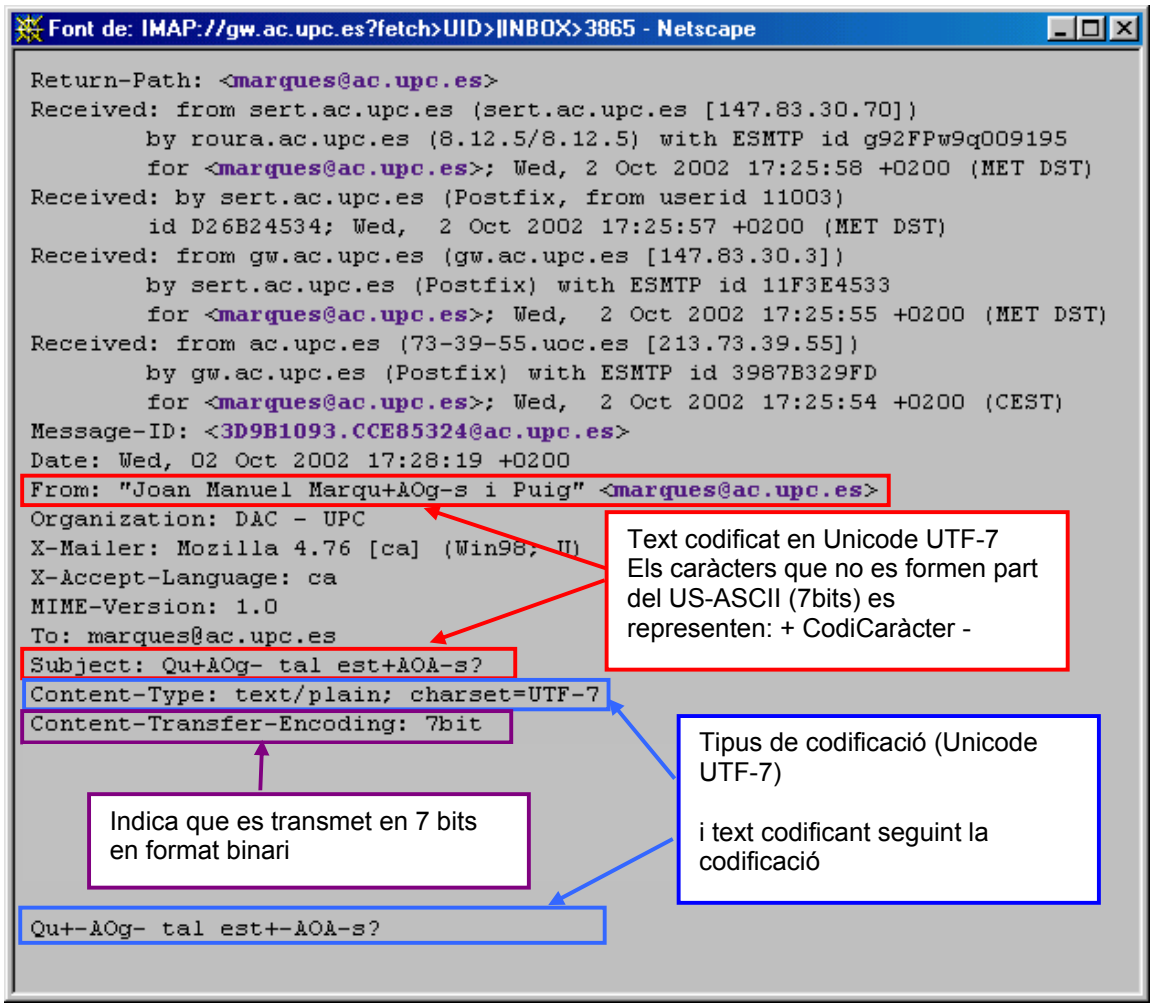
- Perquè surt així? Per a entendre què ha passat, mira el codi font del missatge i compara el que hi posa amb el joc de caràcters que tens seleccionat.

3.3 Enviament d'un missatge en UTF-7

- Repeteix el mateix del cas anterior però enviant el missatge en Unicode UTF-7



- Mira el codi font del que has rebut.



- Canvia la visualització

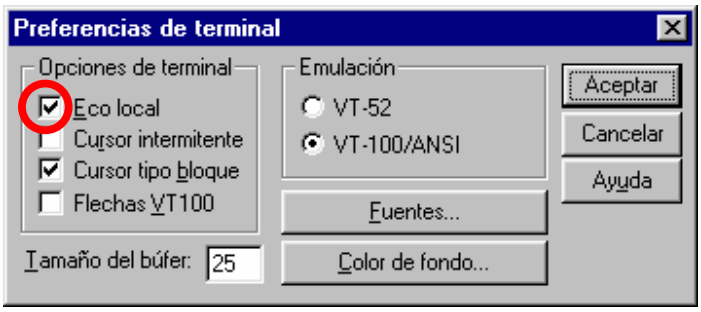
4.- Enviament d'un missatge connectant-se directament amb el servidor SMTP via telnet

L'objectiu d'aquesta activitat és que envieu un missatge de correu electrònic connectant-vos directament al servidor SMTP via telnet.

Farem aquesta activitat obrint una finestra MS-DOS i executant la comanda:

```
c:\> telnet dimoni.upc.es 25
```

(Un cop al telnet, si no veieu el que teclegeu, cal que activeu l'echo local:
(**Terminal -- preferències**)



Sessió d'enviament de correu:

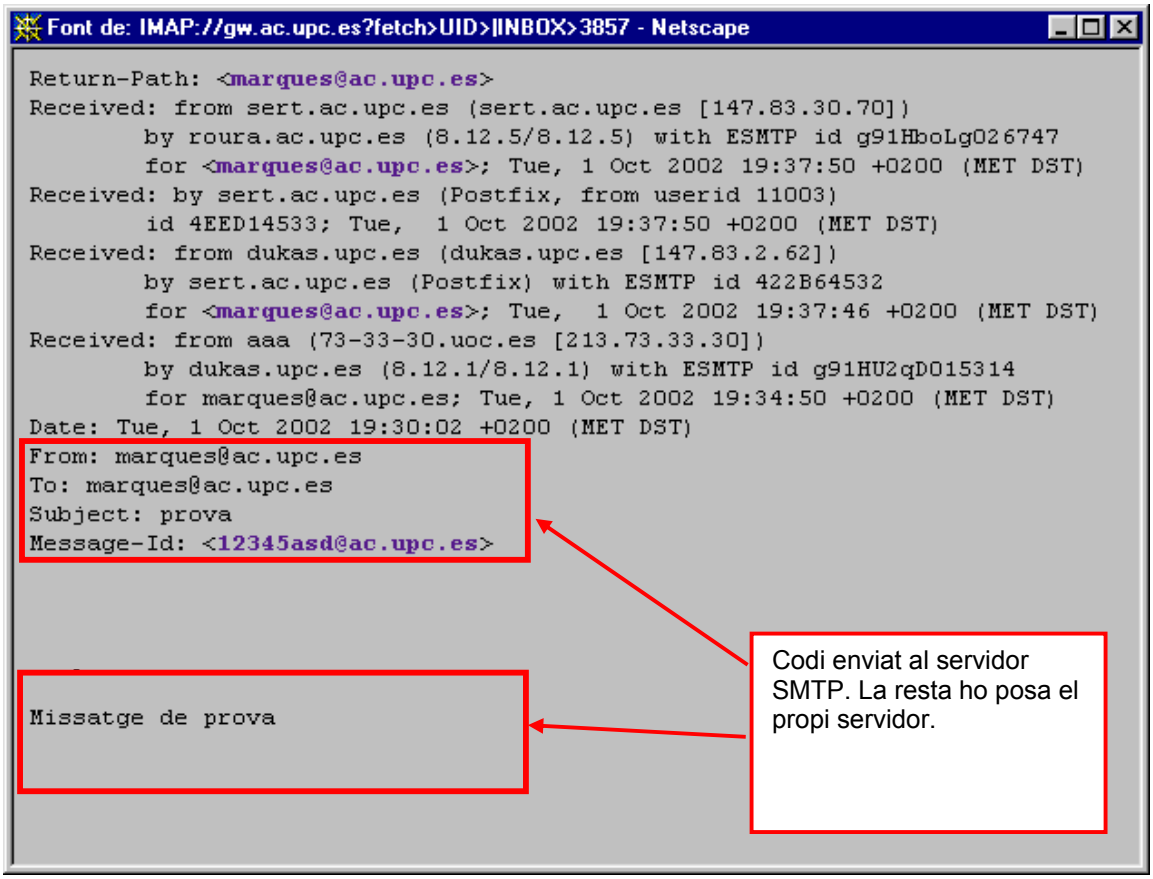
```
220 dukas.upc.es ESMTP Sendmail; Tue, 1 Oct 2002 19:23:45 +0200
(MET DST)
EHLO aaa
250-dukas.upc.es Hello usuari.upc.es [195.77.33.30], pleased to
meet you
250-ENHANCEDSTATUSCODES
250-PIPELINING
250-8BITMIME
250-SIZE 10000000
250-ETRN
250-DELIVERBY
250 HELP
mail from:marques@ac.upc.es (substitueix-ho per la teva adreça)
250 2.1.0 marques@ac.upc.es... Sender ok
rcpt to:marques@ac.upc.es (substitueix-ho per la teva adreça)
250 2.1.5 marques@ac.upc.es... Recipient ok
data
354 Enter mail, end with "." on a line by itself
From:marques@ac.upc.es (substitueix-ho per la teva adreça)
To:marques@ac.upc.es (substitueix-ho per la teva adreça)
Subject: prova
Message-Id:<12345asd@ac.upc.es>

Missatge de prova
.
250 2.0.0 g91HU2qD015314 Message accepted for delivery
quit
221 2.0.0 dukas.upc.es closing connection
```

Domini. Es pot posar qualsevol cosa ja que el servidor s'adona del nostre domini

Entre la capçalera i el cos del missatge cal posar una línia en blanc

Un cop enviat el missatge, el llegeixes des del correu electrònic i esculls l'opció de veure el codi font del missatge. Ha de sortir:



- Pots fer altres proves posant altres camps a la capçalera.

- Fes una prova sense posar cap camp a la capçalera (= sense posar **From:**, **To:**, **Subject:** ni **Message-Id:**. Recorda,però, de deixar una línia en blanc abans de començar el cos del missatge)

- Si vols fer més proves, pots consultar l'RFC 822 o provar les comandes:
  - **HELP**: per obtenir ajuda.
  - **NOOP**: no operació. S'utilitza per a informar que la connexió encara està oberta

5.- Veure quina codificació s'aplica a cadascuna de les parts d'un missatge MIME

L'objectiu d'aquesta activitat és veure que en l'enviament d'un missatge que conté dades en llocs diferents (subject, text, fitxer adjunt,...) s'apliquen diferents codificacions a cadascuna de les parts.

Envia el mateix missatge de l'apartat 3.1 però aquest cop amb un fitxer adjunt. Utilitza com a fitxer adjunt el fitxer en format Unicode creat amb el WordPad a l'apartat 1. Cal que el fitxer no tingui extensió .txt (per exemple .doc) per a que el correu no el reconegui i l'enviï com a text. (Si vols, un cop hagis fet aquest apartat, pots enviar-lo amb extensió .txt i veure què passa).

```
Return-Path: <marques@ac.upc.es>
Received: from sert.ac.upc.es (sert.ac.upc.es [147.83.30.70])
        by roura.ac.upc.es (8.12.5/8.12.5) with ESMTP id g92H9f9q023105
        for <marques@ac.upc.es>; Wed, 2 Oct 2002 19:09:41 +0200 (MET DST)
Received: by sert.ac.upc.es (Postfix, from userid 11003)
        id 098D14535; Wed,  2 Oct 2002 19:09:39 +0200 (MET DST)
Received: from gw.ac.upc.es (gw.ac.upc.es [147.83.30.3])
        by sert.ac.upc.es (Postfix) with ESMTP id 52EA04533
        for <marques@ac.upc.es>; Wed,  2 Oct 2002 19:09:36 +0200 (MET DST)
Received: from ac.upc.es (73-39-55.uoc.es [213.73.39.55])
        by gw.ac.upc.es (Postfix) with ESMTP id 81347329FD
        for <marques@ac.upc.es>; Wed,  2 Oct 2002 19:09:35 +0200 (CEST)
Message-ID: <3D9B28F5.A20E5821@ac.upc.es>
Date: Wed, 02 Oct 2002 19:12:21 +0200
From: Joan Manuel =?iso-8859-1?Q?Marqu=E8s?= i Puig <marques@ac.upc.es>
Organization: DAC - UPC
X-Mailer: Mozilla 4.76 [ca] (Win98; U)
X-Accept-Language: ca
MIME-Version: 1.0
To: marques@ac.upc.es
Subject: =?iso-8859-1?Q?Qu=E8?= tal =?iso-8859-1?Q?est=E0s=3F?=
Content-Type: multipart/mixed;
 boundary="----------82D03CE771FD237DDAAF3B81"

Aquest missatge esta en format MIME i consta de varies parts.
----------82D03CE771FD237DDAAF3B81
Content-Type: text/plain; charset=iso-8859-1
Content-Transfer-Encoding: 8bit

Què tal estàs?


----------82D03CE771FD237DDAAF3B81
Content-Type: application/msword;
 name="Unicode.doc"
Content-Transfer-Encoding: base64
Content-Disposition: inline;
 filename="Unicode.doc"

//5RAHUA6AAgAHQAYQBsACAAZQBzAHQA4ABzAD8A
----------82D03CE771FD237DDAAF3B81--
```

Indica que és un missatge MIME multipart

Separador utilitzat per a indicar on comença cada part del missatge

Contingut del fitxer que s'envia codificat en *base64*

6.- Qüestió per a contestar

En aquesta darrera activitat cal que, a partir de tot el que hem vist en la sessió, empleneu les caselles de la taula següent posant la codificació de Què tal estàs? en cadascun dels formats i en cadascuna de les codificacions de transmissió. (Si alguna codificació l'heu de calcular a ma, no cal que traduïu tota la frase.)

Què tal estàs?	ISO-8859-1	UNICODE UTF-16	UNICODE UTF-8	UNICODE UTF-7
Quoted-Printable				
Base64				
8bit binary				

A partir de la taula, indica els avantatges de codificar text utilitzant la codificació UTF-8 de Unicode (ISO 10646), enlloc d'usar ISO Latin-1 (ISO-8859-1) i UTF-16.

Quina diferència hi ha entre enviar per correu electrònic texte Unicode amb UTF-8 + "quoted printable" o directament amb UTF-7. Pot dependre de l'idioma en que estigui escrit el missatge?

Envieu la taula i els vostres comentaris a: marques@ac.upc.es (en un missatge de text per grup indicant a més els noms dels membres del grup. En el subject ha de posar: **AAD:problemes sessió codis de caràcters i correu SMTP i MIME**)

Bibliografia

- A tutorial on character code issues: http://www.cs.tut.fi/~jkorpela/chars.html
  - This document tries to clarify the concepts of character repertoire, character code, and character encoding especially in the Internet context. It specifically avoids the term character set, which is confusingly used to denote repertoire or code or encoding. ASCII, ISO 646, ISO 8859 (ISO Latin, especially ISO Latin 1), Windows character set, ISO 10646, UCS, and Unicode, UTF-8, UTF-7, MIME, and QP are used as examples.
- Web d'Unicode: http://www.unicode.org/
- RFC 822: http://www.ietf.org/rfc/rfc0822.txt (especificació de l'estàndard SMTP)