UNIVERSITÉ PAUL VALÉRY MONTPELLIER

# COMPLÉMENTS FORMATS DE FICHIERS, INTERNET, RECHERCHE D'INFORMATION

Stage – Semaine 5

# RAPPEL : AU SEMESTRE 2 !

**Au semestre 2**, vous serez en niveau « Standard ». Ce niveau reprend le contenu du stage pour aller ensuite au-delà.
En fonction de votre résultat, vous aurez 6 ou 10 semaines d'enseignement (en d'autres termes, vous serez dispensé ou non de la première partie du niveau).

- **Par défaut, l'enseignement dure tout le semestre.** Si vous êtes en « Standard 10 semaines », il n'y a pas de vagues d'enseignement.
- Si vous êtes en « Standard 6 semaines », vous restez dans la même vague qu'au premier semestre.
- A suivre sur le site d'enseignement
  - Rubrique « Résultats » pour connaître votre niveau au semestre 2
  - Rubrique « Calendrier » pour connaître les principales dates du niveau dans lequel vous serez.

# ATTENTION !

Pour l'examen (15 points de la note finale), la semaine prochaine, vous devez :

- Avoir **votre carte d'étudiant**
- Connaître **votre mot de passe** pour accéder à **Moodle**

D'ici l'examen :

- Pensez à **terminer les tests de contrôle continu** (5 points de la note finale)

# FORMATS DE FICHIERS

Stage – Semaine 5

# FORMATS DE FICHIERS

- Déjà vu : il existe plusieurs formats d'images ?
  - à cause de choix de codage différents
    - Codage matriciel versus codage vectoriel
    - Codage de couleurs (niveaux de gris, modèle RVB, …)
    - Utilisation ou non d'algorithme de compression (potentiellement destructrice)
- La multitude de formats existe pour d'autres types de documents numériques
  - La raison est toujours des codages différents…
  - … en suite de 0 et de 1,
    le bit étant l'information de base pour un ordinateur

# CODAGE ?

Codage d'une information :
Règle d'écriture à l'aide d'un nombre fini de symboles permettant de désigner de manière unique l'information
(le codage est contextualisé (codage de nombre, de couleurs, de textes, …))

Exemple : codage d'entiers

123 ⟺ Cent vingt trois ⟺ CXXIII

Trois écritures de la même valeur

à l'aide d'une liste de symboles indécomposables :

- des chiffres (0,1,2, …)
- des mots (un, deux, dix, cinquante, …)
- des lettres (I,V,X,L,C, …)

# PRINCIPAUX FORMATS POUR LES TEXTES

Trois niveaux de formats dépendant du type de contenu

- contenu seulement (texte brut)
  - Format dit TEXTE : txt
  - Texte formaté par l'utilisateur : html, xml, …
  - Programmes informatiques : php, c, …
- + les enrichissements (police, corps, couleurs, …)
  - Format TEXTE ENRICHI: rtf
- + fonctionnalités avancées (pagination, tables, index, …)
  - Format de travail de bureautique : doc, docx, odt
  - Format d'échange et de distribution : pdf

# LE FORMAT TEXTE SEULEMENT

- La seule information qu'il contient = une suite de caractères
- Conseillé d'utiliser comme logiciel, un « éditeur de texte » (TextWrangler, BBedit, emacs, NotePad++, …) et non un « logiciel de traitement de texte » (TextEdit, Open Office Writer, Libre Office Writer, Microsoft Word, …)
- Chaque caractère est codé par (i.e. associé à) une suite de 0 et de 1

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| B | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

Exemple :
codage de A et B en code ASCII ou UTF-8

- … mais plusieurs tables de codage possibles
  - ASCII
  - Unicode (utf-8, …)
  - …
    (voir diapositive suivante)

# DE L'ASCII À L'UTF-8

- ASCII d'origine (aperçu à droite)
  - 7 bits par caractère : caractères latins sans accent.
  - 128 caractères
- ASCII étendu
  - nombreuses extensions non compatibles entre elles (une par langue)
  - 8bits par caractère
  - 256 caractères
- « utf-8 » :
  - une table universelle (recommandée W3C)
  - Codage variant de 8 à 32 bits pour un caractère ou idéogramme

盲人摸象 - 各執一端

| Character | Decimal Number | Binary Number | Character | Decimal Number | Binary Number |
|---|---|---|---|---|---|
| blank space | 32 | 0010 0000 | ^ | 94 | 0101 1110 |
| ! | 33 | 0010 0001 | - | 95 | 0101 1111 |
| " | 34 | 0010 0010 | ` | 96 | 0110 0000 |
| # | 35 | 0010 0011 | a | 97 | 0110 0001 |
| $ | 36 | 0010 0100 | b | 98 | 0110 0010 |
| A | 65 | 0100 0001 | c | 99 | 0110 0011 |
| B | 66 | 0100 0010 | d | 100 | 0110 0100 |
| C | 67 | 0100 0011 | e | 101 | 0110 0101 |
| D | 68 | 0100 0100 | f | 102 | 0110 0110 |
| E | 69 | 0100 0101 | g | 103 | 0110 0111 |
| F | 70 | 0100 0110 | h | 104 | 0110 1000 |
| G | 71 | 0100 0111 | i | 105 | 0110 1001 |
| H | 72 | 0100 1000 | j | 106 | 0110 1010 |
| I | 73 | 0100 1001 | k | 107 | 0110 1011 |
| J | 74 | 0100 1010 | l | 108 | 0110 1100 |
| K | 75 | 0100 1011 | m | 109 | 0110 1101 |
| L | 76 | 0100 1100 | n | 110 | 0110 1110 |
| M | 77 | 0100 1101 | o | 111 | 0110 1111 |
| N | 78 | 0100 1110 | p | 112 | 0111 0000 |
| O | 79 | 0100 1111 | q | 113 | 0111 0001 |
| P | 80 | 0101 0000 | r | 114 | 0111 0010 |
| Q | 81 | 0101 0001 | s | 115 | 0111 0011 |
| R | 82 | 0101 0010 | t | 116 | 0111 0100 |
| S | 83 | 0101 0011 | u | 117 | 0111 0101 |
| T | 84 | 0101 0100 | v | 118 | 0111 0110 |
| U | 85 | 0101 0101 | w | 119 | 0111 0111 |
| V | 86 | 0101 0110 | x | 120 | 0111 1000 |
| W | 87 | 0101 0111 | y | 121 | 0111 1001 |
| X | 88 | 0101 1000 | z | 122 | 0111 1010 |
| Y | 89 | 0101 1001 | { | 123 | 0111 1011 |
| Z | 90 | 0101 1010 | \| | 124 | 0111 1100 |
| [ | 91 | 0101 1011 | } | 125 | 0111 1101 |
| / | 92 | 0101 1100 | ~ | 126 | 0111 1110 |
| ] | 93 | 0101 1101 | | | |

Traduction : mángrén mō xiàng... gè zhí yì duān

# RECONNAISSANCE DE FORMATS

- Extension de fichiers : suffixe de 2 à 4 lettres indiquant le format utilisé pour coder l'information contenue dans le fichier
  - Exemples : txt, odt, docx, pdf, …, jpg (pour format jpeg), …
- Le logiciel système associe un logiciel à chaque format
  (le logiciel associé par défaut peut-être modifié).
- Si le suffixe associé à un fichier est erroné (ne correspond pas au format réellement associé), le logiciel par défaut risque de ne pas savoir ouvrir le document ➔fichier inutilisable
- Conséquence : conseil =
  laisser les logiciels ajouter eux-mêmes les extensions
  (Ne saisissez que les noms hors suffixes)
  car ils connaissent l'extension adéquate pour le format qu'ils utilisent

# TRANSFORMATION DE FORMATS

- Changer (ou saisir) un suffixe ne change pas le format réel du contenu d'un fichier.
- Par contre, les logiciels connaissent généralement :
  - Plusieurs formats
  - Les traitements pour passer de l'un à l'autre (via l'enregistrement ou l'export du fichier… en précisant le « type de fichier » ou le « format de fichier »)
- Ces fonctionnalités d'enregistrement ou d'export sont à utiliser pour changer le format d'un fichier.

# INTERNET

## Stage – Semaine 5

## Est-ce qu'internet = web ?

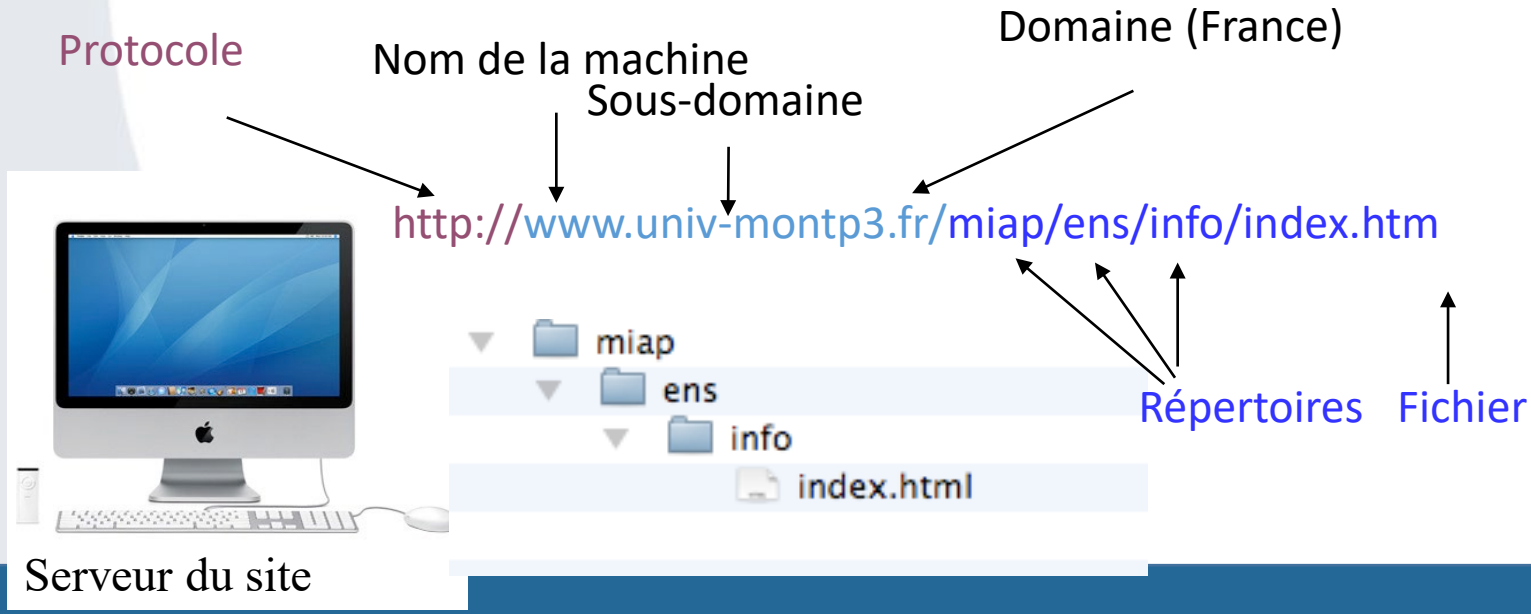
- NON
  - Internet = réseau physique mondial d'ordinateurs (ou plus précisément réseau de réseaux)
    - Internet à plus de 50 ans !
    - Le 29 octobre 1969, premier transfert (du texte « lo ») entre machines connectées à distance dans le cadre du projet ARPANET, pré-version d'internet
    - Voir par exemple le site de l'ICANN
  - Web (World Wide Web) = réseau d'informations constitués des documents mis à disposition sur les serveurs d'internet et reliés par les hyperliens qu'ils contiennent.
    - Créé au début des années 1990

# RAPPELS : CLIENT-SERVEUR

- Principe de base :
  - Un ordinateur (le client) demande un service (exemple : une page web) à un serveur
  - L'échange est effectué en suivant un ou des protocoles, langages gérant le dialogue entre machines
- Quelques protocoles :
  - IP (Internet Protocol) : gestion d'internet
  - http (HyperText Transfert Protocol) : gestion web
  - https : idem sécurisé (crypté)
  - imap (Interactive Message Access Protocol) : accès à des messageries électroniques à partir de logiciels clients
  - smtp (Simple Mail Transfer Protocol) : envoi de mails
  - ftp (File Transfert Protocol) : transfert de fichiers (utile quand vous hébergez un site web chez un fournisseur de services)
- Remarque : pare-feu : permet de gérer les protocoles autorisées

# ADRESSAGE

- Nécessité de localisation : chaque ordinateur est associée à une adresse l'identifiant gérée par le protocole IP
  - Adresse IP = suite de 32 bits (représentés sous forme de 4 blocs de nombres entre 0 et 255 (8 bits)) en IPv4
    ou 128bits en IPv6 en cours de déploiement)
  - Exemples : 193.52.137.213, 10.3.7.12 …
- Pour le protocole http, localisation d'une page web :



# SITE ?

- Un site :
  - Ensemble de pages web dont l'URL débute par un même début, appelé racine du site ou URL du site.
  - Exemples :
    - http://www.univ-montp3.fr/ : site de l'université
    - http://www.univ-montp3.fr/miap/ens/info/index.htm site des enseignements d'informatique

# TRACES

- Lors d'une navigation sur le web, les données échangées avec les serveurs interrogés passent par des serveurs intermédiaires
  - Les données peuvent être lues par d'autres.
  - Vos requêtes peuvent être stockées (obligation légale pendant 1 an des fournisseurs de services – LCEN : Loi pour la confiance dans l'économie numérique)
- Autres traces
  - Cookies (avec RGPD – Règlement général sur la protection des données – vous connaissez !)
  - Historique de navigation
  - Métadonnées dans les fichiers
  - …

## Que se passe-t-il quand on saisit l'adresse IP d'une machine au lieu de l'URL d'un site ?

- Si l'IP correspond à l'adresse d'un site web, l'adresse IP est remplacée par l'URL du site et la page d'accueil du site est affichée (d'autres mécanismes peuvent être mis en place).
- Exemple : 193.52.137.213 … serveur web de l'université (mais des redirections qui mettent en alerte Firefox)

# QUELS OUTILS UTILISEZ-VOUS LORS D'UNE RECHERCHE D'INFORMATION SUR LE WEB ?

1. Un ordinateur (de bureau, portable, smartphone, …) connecté à internet
2. Un navigateur Web :
   - Chrome, Firefox, Edge, Safari, Opéra, …



3. Un moteur de recherche :
   (site contenant un index
   d'une partie du web)

   Google Français  Qwant  Bing  …

   ou un méta-moteur de recherche
   (site construisant sa réponse
   en recoupant les résultats
   de requêtes à différents
   moteurs de recherche)

   DuckDuckGo  lilo  …

- Remarque : moteur de recherche ≠ navigateur web

# EST-CE QU'UN MOTEUR DE RECHERCHE INDEXE TOUT LE WEB ?

- Non
  - Le web est dynamique :
    - Des pages se créent (et disparaissent) régulièrement
    - Les crawlers ou spiders, robots d'indexation, ne peuvent pas tout recenser en temps réel
  - Potentiellement un problème de place dans la base de données d'indexation (et donc des choix sur ce que conserver)
  - Des sites non reliés aux autres (web invisible)
  - Des pages payantes ou privatisées par mot de passe
  - …

# DANS LE TD

- Un zoom sur des requêtes dans un moteur de recherche
  - But = réduire le nombre de réponses et obtenir des réponses plus précises
- Réflexion sur la qualité des sites consultés :

La qualité des ressources est très variable.
C'est l'internaute qui doit la juger.