

LES HUMANITÉS NUMÉRIQUES

Formats

Marine DELABORDE



Format de données

- **Format de donnée** = manière utilisée pour représenter des données sous forme de bits
- Suite de **bits** → nombre binaire
- Exemple : $A \rightarrow 01000001 \rightarrow 65$

Format de données

- Format de donnée =
convention de représentation des données
- Convention qui permet l'interopérabilité
- Données stockées dans un fichier → **format de fichier**

Format de données

- **Format ouvert** : spécification publiquement accessible

(= standard ouvert, format ouvert, spécification ouverte)

- Sans restriction d'accès ni de mise en œuvre (droit d'auteur, brevet, copyright)
- But : améliorer l'interopérabilité → format indépendant du logiciel pour l'exploiter
- Format approuvé par une organisation internationale de standardisation
- Ex : ASCII, TeX, OpenDocument Text, HTML, CSS, PDF, EPUB, CSV, JSON, PNG, GIF, MP3, TAR, ZIP...

VS

- **Format fermé** : spécification secrète

- Seul un logiciel est capable de pleinement l'exploiter
- Utilisation restreinte par le propriétaire
- Migration vers d'autres logiciels limité → emprise sur l'utilisateur
- Enjeu commercial important
- Ex : Microsoft Office, Adobe Photoshop...

Format de données

- Alternatives :

https://fr.wikipedia.org/wiki/Correspondance_entre_formats_ouverts_et_formats_ferm%C3%A9s

Format de données

- **Format normalisé** : normalisation par une institution publique ou internationale (ISO, W3C...)

VS

- **Format propriétaire** : appartient à une entreprise, développé dans un but essentiellement commercial
 - **Format propriétaire ouvert** : spécifications publiées (ex : PDF d'Adobe)
 - **Format propriétaire fermé** : spécifications gardées secrètes (ex : DOC de Microsoft)

Formats de données

- Formats de nombres (entiers, fractionnaires, à virgule)
- Formats de texte
- Formats d'image
- Formats de vidéo
- Formats de scène 3D
- Formats de son
- Compression des données

Formats d'images

- Représentation des images basé sur la **géométrie analytique** (objets représentés par des équations ou inéquations)
- **Image matricielle (bitmap) :**
 - Découpage de l'image en points élémentaires (**pixels**)
 - Format en **carte de points**
 - Chaque pixel → une information de position et de couleur
 - Ex : GIF, JPEG, TIFF
- **Image vectorielle :**
 - Image décrite par des ensembles de **coordonnées mathématiques**
 - Format **économe** : images facilement réduites à des formes géométriques
 - Avantage : Rendu final indépendant de la résolution du périphérique de sortie
 - Inconvénient : Ne permet de représenter que des formes simples
 - Ex : SVG, Adobe PDF

Formats de texte

- **Texte** = formé de **caractères** en nombres finis
 - lettres, diacritiques, ponctuation, idéogrammes...
- **Simplicité** d'attribution d'un nombre à chaque caractère
- **Conversion** caractère → nombre = définie par une **convention** (table, page de code)
- Plus courants : ASCII, Unicode

Formats de texte

- Les textes comprennent :
 - **mise en page**
 - **mise en forme**
- Comment enregistrer cette information ?
 - Définir des **mots de commande** / **instructions** séparées du texte par un caractère spécial
 - **HTML** : instructions = « balises » mises entre < > (chevrons)
 - **LaTeX** : instructions = introduites par \
 - Caractères réservés aux instructions → ne peuvent pas faire partie du texte. Solutions :
 - « codes d'échappement »
 - Instructions spéciales pour les représenter

HTML

- **Hypertext Markup Language** : langage de balisage d'hypertexte
- **Hypertexte** = texte avec des liens vers d'autres pages
- **Code HTML** :

```
<p>Ceci est <i>un</i> paragraphe.</p><p>Ceci est un <i>autre</i> paragraphe.</p>
```
- **Représentation par un navigateur** :
Ceci est *un* paragraphe.
Ceci est un *autre* paragraphe.

Microsoft .doc

- Format propriétaire
- Format de Microsoft Word = **secret** !

Microsoft .doc

- Aujourd'hui d'autres logiciels peuvent ouvrir un .doc
 - ex : OpenOffice → logiciel libre
- MAIS : résultat obtenu en regardant les fichiers .doc. (reverse engineering)
 - Microsoft n'a jamais publié ce format.
- Si on envoie un fichier .doc (ou .ppt ou .xls) à quelqu'un, on suppose que cette personne possède M\$-Office.
 - Acheté : 149€ (ou 69€/an pour Office 365) + éventuellement 145€ (Windows 10)
 - Copié : illégal

=> C'est malpoli !
- De plus : Si le récepteur d'un fichier .doc n'a pas les bonnes polices, tout s'affiche mal.

Microsoft .doc

- Pour un fichier Word qui contient seulement :
Ceci est *un* paragraphe.
Ceci est un *autre* paragraphe.
- Word enregistre en plus beaucoup d'informations...
FPf6].h0. A!n"n#n\$n2P180p3P(20~.+,D~.+,

.rtf et .pdf

- **.rtf** : ancien format de Word
 - seul avantage : lisible par un humain
- **.pdf** : Portable Document Format = format de document portable
 - Polices incluses → l’affichage est le même partout
 - Format ouvert : Adobe a publié les spécifications
 - Lisible sur tous les systèmes d’exploitation (M\$-Windows, Linux, Mac...)
 - Peu de logiciels peuvent le modifier
 - Utilisation : distribution de documents complets, terminés à des tiers

OpenDocument

- Importance d'un **format standardisé** pour la bureautique (Tim Bray) :

« Qui d'entre vous est sûr de posséder des documents auxquels il voudra pouvoir accéder dans dix ans ? »

OpenDocument

- Importance d'un **format standardisé** pour la bureautique (Tim Bray) :

« Qui d'entre vous est sûr de posséder des documents auxquels il voudra pouvoir accéder dans dix ans ? »

« Qui d'entre vous est sûr d'utiliser la même application bureautique dans dix ans ? »

OpenDocument

- **Format ouvert** de données pour les applications bureautiques
 - Traitements de texte, tableurs, présentations, diagrammes, dessins et bases de données bureautique.
- Premier effort de **standardisation** dans le domaine de la bureautique.
- Certification auprès de l'**ISO** (International Organization for Standardization) attribué le 1^{er} mai 2006.
 - Format OpenDocument = norme ISO (ISO 26300) comme HTML (ISO 15445).
- Format de texte = **.odt** (open document texte – texte en document ouvert)

Sémantique des balises

- Mélange :
 - Description de contenu
 - Citation, titre...
 - Description de mise en page
 - Italique, gras...
 - Les deux
 - Paragraphe, tableau, titre du document, hyperlien...
- Description du contenu (au moins partiellement)
- Le navigateur peut choisir comment afficher certaines informations sur le contenu

Côté client : le navigateur

- Interprète le code HTML
- N'affiche pas les balises
 - Ce qui se trouve entre < >
- Doit parfois charger d'autres fichiers du serveur web pour afficher correctement la page :
 - Images, animations, musique...
 - « Feuilles de style »
 - Scripts : Javascript
 - Sert à :
 - contrôler les données saisies dans les formulaires HTML,
 - Interagir avec le document HTML (on parle parfois d'HTML dynamique)
 - Réaliser des services dynamiques de mise en page (menus...)
 - Code JavaScript :
 - Peut être intégré directement dans les pages Web
 - S'exécute sur le poste client (dans le navigateur)

HTML en détails

- Structure basique :

```
<!DOCTYPE html>
```

```
<html>
```

```
  <head>
```

```
    <title>Titre de la page</title>
```

```
  </head>
```

```
  <body>
```

```
    Texte de la page
```

```
  </body>
```

```
</html>
```

HTML en détails

- Commentaires :

```
<!--
```

```
*****
```

```
*   Partie 1   *
```

```
*****
```

```
-->
```

HTML en détails

- Liens :

`Lien vers le site de Paris III`

- Images :

` Une belle image.`

HTML en détails

- Liste non ordonnée :

``

`Premier élément de la liste `

`Deuxième élément `

`Etc.`

``

Donne :

- Premier élément de la liste ;
- Deuxième élément ;
- Etc.

HTML en détails

- Liste numérotée :

Premier élément de la liste

Deuxième élément

Etc.

Donne :

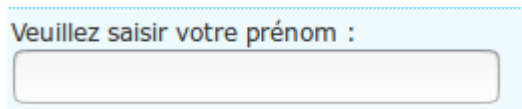
1. Premier élément de la liste ;
2. Deuxième élément ;
3. Etc.

HTML en détails

- Formulaires :

Veillez saisir votre prénom :

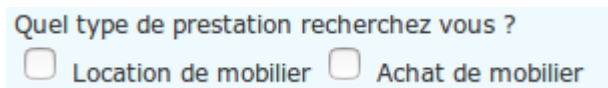

```
<input type="text" name="prenom" value="" />
```

A screenshot of a web form element. It consists of a light blue rectangular container. Inside, at the top, is the text 'Veillez saisir votre prénom :'. Below this text is a white rectangular text input field with a thin grey border.

Quel type de prestation recherchez vous ?


```
<input type="checkbox" name="interet" value="loc" /> Location de mobilier
```

```
<input type="checkbox" name="interet" value="achat" /> Achat de mobilier
```

A screenshot of a web form element. It is a light blue rectangular container. At the top, it contains the text 'Quel type de prestation recherchez vous ?'. Below this text are two radio buttons. The first radio button is followed by the text 'Location de mobilier'. The second radio button is followed by the text 'Achat de mobilier'.

Caractères spéciaux & entités

- Entités = mécanisme de type macro
 - Début → &
 - Milieu → Symbole
 - Fin → ;
- Déclaré dans la DTD
- Mécanisme d'échappement :
 - Si $a < b$ alors...
 - Si $a \< b$ alors...
- Liste :
https://fr.wikibooks.org/wiki/Le_langage_HTML/Entit%C3%A9s

CSS

- **CSS** : feuilles de style en cascade
- Utilisé pour définir les caractéristiques de présentation d'un document : couleurs, polices, rendu (alignement du texte : taille, position...)
- Objectif = séparer :
 - **Structure** : écrite en HTML ou similaire
 - **Présentation** (en CSS) du document
- Bénéfices :
 - Amélioration de l'**accessibilité**
 - **Changement** de structure et de présentation plus facile
 - Meilleure **adaptation** aux caractéristiques du récepteur

CSS

- **HTML** = architecture interne
- **CSS** = tous les aspects de présentation
- **Structure et présentation** : peuvent être gérées dans des fichiers séparés
- Site internet → **présentation uniformisée** :
 - Tous les documents (= pages html) font référence à la même feuille de style
 - Permet un « relookage » plus rapide
- Un même document = possibilité de choix entre **deux feuilles de style** (ex : impression ou lecture à l'écran)
- Complexité du code **HTML réduit** : HTML ne contient plus de balises de présentation

CSS

- Exemple d'une portion de feuille CSS :

```
p { font-size: 110%; font-family: Helvetica, sans-serif; }  
h1 { color: white; background: red; }
```

- Ce code CSS définit
 - l'élément p (paragraphe) avec une taille de 110% et une police Helvetica, ou, si Helvetica est indisponible, une police générique.
 - titres (éléments h1) : blancs, sur fond rouge.

CSS

- Permet aussi de définir **ses propres styles**
- **Firebug**
 - Pluging à Firefox
 - On ne peut normalement pas changer les feuilles de style
- Le W3C déconseille maintenant les éléments et les attributs de présentation en HTML, comme align= » » ou
- Les descriptions CSS peuvent être données à **l'intérieur** d'un document HTML, ou importées **séparément** dans le lecteur.

Encodage

- Combien de caractères différents ?

a A à á â ã ä å ā ǻ ą

a a a a a a **A** a a a ~

Encodage

- **Glyphe** (du grec : γλυφή ; ciselure, gravure) = représentation graphique (parmi une infinité possible) d'un signe typographique (Wikipédia).
- **Caractère** = la plus petite partie de la langue écrite qui porte une valeur sémantique (qui a un sens).
 - Mais aussi : un chiffre, l'espace, un accent seul...
- **Police** : définit des glyphes pour des caractères.

Pour s'entraîner

- Trouver comment afficher le code source d'une page web (firefox : ctrl-u).
- Retrouver le mot qui se trouve sur la page dans le code source.
- Regarder toutes les balises qui entourent ce mot.