Décrire et manipuler un document numérique

Introduction

Loïc Grobol <lgrobol@parisnanterre.fr>

2022-01-17

Bonjour

- Loïc Grobol (il/iel) <loic.grobol@parisnanterre.fr>
- PHILLIA / MoDyCo (Bâtiment Rémond, 4ème, bureau 404C)
- *Office hours* le mardi après-midi, n'hésitez pas à passer y compris sans rendez-vous (mais je préfère si vous m'envoyez un mail pour me prévenir)
- De manière générale, n'hésitez pas à m'écrire

Le cours

Infos pratiques

- **Quoi** « Décrire et manipuler un document numérique » 4L4SC02P
- **Où** Salle M114, bâtiment Éphémère 1
- **Quand** 2 séances, les lundis de 13:20 à 15:20, du 17/01 au 11/04
  - Voir le calendrier de l'université pour les dates de vacances.

→ Travail sur machine préférable, amener si possible un PC portable

Liens

- La page du cours (slides, documents, nouvelles, consignes…)
  - → https://loicgrobol.github.io/document-numerique
- Le dépôt GitHub (code source, compléments et historique)
  - → https://github.com/LoicGrobol/document-numerique

Objectifs

Connaissances :

- Savoir ce qu'est un document numérique.
- Savoir comment différents types de documents sont représentés dans des systèmes informatiques.
- Connaissances précises sur les représentations numériques des documents textuels.
- Connaissances de bases sur ce que sont Internet et le Web et leurs fonctionnements.

Compétences :

— Créer et modifier des documents XML simples.
— Créer et modifier des pages web simples avec HTML et CSS.
— Créer et modifier des documents structurés en Markdown.
— Représenter des données linguistiques sous forme tabulaire et semi-tabulaire.
— Utiliser des expressions régulières pour faire des recherches et des modifications dans des fichiers textes.

Évaluation

— Exercices en temps libre pendant le semestre
— Examen final

Documents numérique

— Qu'est-ce qu'un document ?

Précédemment, en *Humanités Numériques*

> Un document est une **trace** permettant d'**interpréter** un **événement passé** à partir d'un **contrat de lecture**. (Jean-Michel Salün (2012), « Vu, lu, su. Les architectes de l'information face au monopole du web »)

. . .

À partir de ces éléments, qu'est-ce qui caractérise un document numérique ?

Documents et supports numériques

— Ce qui différencie un document numérique, c'est d'abord le type de **trace** concerné.
— Ce qui a évidemment des conséquences sur ce qu'on peut représenter.
— Dans ce cours, c'est donc sur cette *trace*, ce *support* qu'on va se concentrer.

La question principale pour nous sera en général : « Étant donné un document, comment le représenter sous forme numérique ? »

Fichiers

Le support d'un document numérique est en général un **fichier** (informatique).

. . .

— Ou plusieurs.
— Ou pas.

Abstractions

La mémoire d'un ordinateur peut se concevoir à plusieurs échelles

- Une suite de *bits*, des éléments d'information élémentaire à deux états, 0 ou 1.
- Une série de *multiplets* (ou *byte*), chacun composés d'un nombre fixe de bits (en général 8 : des *octets*).
  - Possèdent une *adresse*, c'est-à-dire un identifiant permettant de les localiser.
  - Plus petites unités *adressables*.
- Une série de *mots*, chacun composé d'un nombre fixe de bytes (de nos jours, 32 ou 64 bits)
  - Plus petites unités traitables par un processeur.

Un fichier informatique :

- Une série de bytes représentant un information.
- Un *format de fichier* : le code permettant de passer de l'information à sa représenation et vice versa.

En pratique, un fichier a en général d'autres propriétés :

- Une *adresse* (inode pour les systèmes Unix par exemple).
- Un *nom*.
- Un *chemin* d'accès dans un système de fichiers (parfois plusieurs).
- Des *métadonnées* qui peuvent être stockées directement dans le fichier ou par le système hôte.

Selon les *permissions*, un·e utilisateurice peut y effectuer des opérations : création, destruction, lecture, écriture, exécution.

Vous connaissez déjà des formats de fichier, listez-en autant que possible et dites quel(s) type(s) d'informations ils permettent de représenter.

Types et formats

Documents textuels

- Au niveau le plus basique `txt` : une suite de caractères.
  - Comment représenter des caractères avec des *bytes* ?
- Structure et sémantique : HTML, Markdown…
- Structure, format et mise en page :
  - OpenDocument (odt, ods, odp…)
  - Office Open XML (docx, xslx, ppts…)
  - Portable Document Format

Documents visuels

- Le plus basique : *bitmap* (pix-map) représente une image comme un tableau rectangulaire de points de couleurs (pixel)
- Extensions : GIF et PNG pour la compression non-destructive et l'animation.
- Plus sophistiqué : JPEG, WEBP, AVIF, HEIF/HEVC pour la compression destructive.
  - Exploitent le fait que les perceptions humaines sont limités et non-homogènes.

Documents sonores

- Non-compressé : WAV, BMF
- Compression sans pertes : FLAC
- Compression avec pertes : Opus, MP3, Vorbis…
  - Mêmes idées que pour les images.

Vidéo et multimédia

- Combinent en général plusieurs flux images et son (et texte/métadonnées)
- Partagent les formats pour le son et les images (en ajoutant la composante temporelle dans le mix pour la compression)
- Beaucoup de formats modernes pour les images et les sons viennent de travaux pour la vidéo.
- Quelques noms : Theora, MPEG-*, AV1, H.26{3,4,5,6}, VP{8,9}, Dirac…

Conteneurs

Pour la vidéo ou d'autres applications, on a souvent besoin de *conteneurs* : des fichiers qui regroupent et enrichissent un ou plusieurs fichiers, en y ajoutant parfois de la compression.

- Multimédia : Matroska (mkv), MPEG-4 (mp4), AVI, Ogg…
- Générique : tar, zip, 7z…

Données textuelles structurées

On peut en fait représenter énormément de choses avec des suites de caractères.

- Génériques : XML, JSON, YAML…
- Spécifique au TAL : CoNLL-*, PTB…

Exécutables

Représentent non pas des données mais des *instructions* pour une machine :

- Formats textuels : scripts
  - Lisibles pour les humain es *et* les machines.
  - En général interprétés à la volée : peu optimal.
- Formats « binaires » : exécutables en langage machine,
  - Conçus pour être optimisés pour les machines : minimum de place, exécution rapide.
  - Très difficilement lisibles et écrivable directement par des humain es : quasi-systématiquement générés (**compilés**) à partir de **code source** textuel.