Informatique et Sciences du Numérique
Représentations structurées des données

Alexandre Termier / Cyril Labbé

2017-2018

Les grands thèmes

- Différent type d'information
  - données / méta-données
  - Information de présentation
- Web
  - HTML
  - CSS
- Modèle de données : semi-structuré (XML)
- Modèle de données : relationnel
  - Algèbre relationnelle
  - Language SQL

De l'information et des caractères "cachés"

- Pour nous : texte en langage écrit (ici le français)
- Pour la machine : une chaîne de caractères
- Jeux de caractères contiennent des caractères de contrôle qui permettent le formatage du texte : `\n` : retour à la ligne, `\t` : tabulation, ...

Example (Ex: texte formaté)

```
Vocabulaire et énoncés

Un triangle rectangle est un triangle admettant un angle droit (c'est-à-dire de mesure 90°).

Les deux côtés adjacents sont appelés cathètes et le côté opposé est l'hypoténuse.

Théorème

La forme la plus connue du théorème de Pythagore est la suivante :

  Théorème de Pythagore - Dans un triangle rectangle, le carré de la longueur de l'hypoténuse est égal à la
  somme des carrés des longueurs des côtés de l'angle droit.

En particulier, la longueur de l'hypoténuse est donc toujours supérieure à celle de chaque autre côté.
```

Fichiers texte

- Format de fichier pour stocker des lignes

Format d'un fichier texte

```
ligne 1EOL
ligne 2EOL
...
dernière ligneEOL
EOF
```

- Caractéristiques
  - `EOL` : fin de ligne (**E**nd **O**f **L**ine), `\r\n` DOS, `\n` Unix, `\r` vieux Macs
  - `EOF` : fin de fichier (**E**nd **O**f **F**ile)
  - Lisible par les humains
  - Standard tant que l'on utilise des caractères ASCII / UTF-8

Code ASCII

ASCII CONTROL CODE CHART

b7 b6 b5 / BITS b4 b3 b2 b1	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
	CONTROL		SYMBOLS NUMBERS		UPPER CASE		LOWER CASE	
0 0 0 0	0 NUL 0 0	16 DLE 10 20	32 SP 20 40	48 0 30 60	64 @ 40 100	80 P 50 120	96 ‘ 60 140	112 p 70 160
0 0 0 1	1 SOH 1 1	17 DC1 11 21	33 ! 21 41	49 1 31 61	65 A 41 101	81 Q 51 121	97 a 61 141	113 q 71 161
0 0 1 0	2 STX 2 2	18 DC2 12 22	34 ” 22 42	50 2 32 62	66 B 42 102	82 R 52 122	98 b 62 142	114 r 72 162
0 0 1 1	3 ETX 3 3	19 DC3 13 23	35 # 23 43	51 3 33 63	67 C 43 103	83 S 53 123	99 c 63 143	115 s 73 163
0 1 0 0	4 EOT 4 4	20 DC4 14 24	36 $ 24 44	52 4 34 64	68 D 44 104	84 T 54 124	100 d 64 144	116 t 74 164
0 1 0 1	5 ENQ 5 5	21 NAK 15 25	37 % 25 45	53 5 35 65	69 E 45 105	85 U 55 125	101 e 65 145	117 u 75 165
0 1 1 0	6 ACK 6 6	22 SYN 16 26	38 & 26 46	54 6 36 66	70 F 46 106	86 V 56 126	102 f 66 146	118 v 76 166
0 1 1 1	7 BEL 7 7	23 ETB 17 27	39 ’ 27 47	55 7 37 67	71 G 47 107	87 W 57 127	103 g 67 147	119 w 77 167
1 0 0 0	8 BS 8 10	24 CAN 18 30	40 (28 50	56 8 38 70	72 H 48 110	88 X 58 130	104 h 68 150	120 x 78 170
1 0 0 1	9 HT 9 11	25 EM 19 31	41) 29 51	57 9 39 71	73 I 49 111	89 Y 59 131	105 i 69 151	121 y 79 171
1 0 1 0	10 LF A 12	26 SUB 1A 32	42 * 2A 52	58 : 3A 72	74 J 4A 112	90 Z 5A 132	106 j 6A 152	122 z 7A 172
1 0 1 1	11 VT B 13	27 ESC 1B 33	43 + 2B 53	59 ; 3B 73	75 K 4B 113	91 [5B 133	107 k 6B 153	123 { 7B 173
1 1 0 0	12 FF C 14	28 FS 1C 34	44 , 2C 54	60 < 3C 74	76 L 4C 114	92 \ 5C 134	108 l 6C 154	124 | 7C 174
1 1 0 1	13 CR D 15	29 GS 1D 35	45 – 2D 55	61 = 3D 75	77 M 4D 115	93] 5D 135	109 m 6D 155	125 } 7D 175
1 1 1 0	14 SO E 16	30 RS 1E 36	46 . 2E 56	62 > 3E 76	78 N 4E 116	94 ^ 5E 136	110 n 6E 156	126 ~ 7E 176
1 1 1 1	15 SI F 17	31 US 1F 37	47 / 2F 57	63 ? 3F 77	79 O 4F 117	95 _ 5F 137	111 o 6F 157	127 DEL 7F 177

LEGEND: dec / CHAR / hex oct

Victor Eijkhout
Dept. of Comp. Sci.
University of Tennessee
Knoxville TN 37996, USA

Code UTF-8

Bits of code point	Bytes in sequence	Byte 1	Byte 2	Byte 3	Byte 4
7	1	0xxxxxxx			
11	2	110xxxxx	10xxxxxx		
16	3	1110xxxx	10xxxxxx	10xxxxxx	
21	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Introduction des métadonnées

Métadonnées d'un document

Données fournissant des informations sur un ou plusieurs aspects du document, sans faire explicitement partie du contenu du document.

- Quelques métadonnées
  - Titre et auteur d'un document
  - Date de création du document
  - Données EXIF d'une photo
  - ...

Utilisation des métadonnées

- Utilisations type
  - Catalogage/indexation de grandes quantités de documents
  - Filtrage et recherche dans ces documents

Utilisation des métadonnées

- Utilisations type
  - Catalogage/indexation de grandes quantités de documents
  - Filtrage et recherche dans ces documents
- Notion ancienne !
  - Un livre a des métadonnées sur sa couverture
  - Catalogue d'une bibliothèque
  - Système Dewey

Utilisation des métadonnées

- Utilisations type
  - Catalogage/indexation de grandes quantités de documents
  - Filtrage et recherche dans ces documents
- Notion ancienne !
  - Un livre a des métadonnées sur sa couverture
  - Catalogue d'une bibliothèque
  - Système Dewey
- Difficiles à repérer dans un fichier texte

Bilan sur les fichiers texte

- Fichiers texte
- Points positifs
  - Faciles à stocker
  - (Assez) standards
  - Quelques informations de formatage
  - Faciles à lire par un humain
- Points négatifs
  - Formatage insuffisant pour des documents complexes (**gras**, *italique*,...)
  - Pas de métadonnées
  - Pas d'images
  - Informations redondantes : perte de place en stockage

Naissance de HTML

- Origine d'HTML
  - CERN : beaucoup de chercheurs, beaucoup de systèmes différents
  - Communications par échange de documents et mails (via Internet)
- HTML : réponse à deux besoins cruciaux
  - Pouvoir faire **facilement** un formatage avancé, "comme sur papier"
    - cela inclut l'intégration d'images
    - ...tout en devant rester portable entre systèmes différents
    - $\rightarrow$ Tim Berners-Lee : faire un sous-dialecte de **SGML**
  - Pouvoir "relier" des documents les uns aux autres
    - hyperliens
    - $\rightarrow$ reprise d'idées anciennes en recherche documentaire !

Caractéristiques de HTML

- **Hypertext Markup Langage**
  - hypertext : documents peuvent être reliés entre eux
  - markup : texte est annoté, en HTML les annotations décrivent la présentation du document
- Forme des annotations : balises

  `Théorème de <bold>Pythagore</bold>`
- Intégration des métadonnées dans le document
  - balises `<title>`, `<author>`, `<meta>`,...
- Informations de structuration
  - balises `<title>`, `<body>`, `<h1>`, `<h2>`, `<ul>`, `<li>`...
- Informations de présentation
  - balises `<bold>`, `<italic>`,...
- Liens
  - balise `<a>`