

PROCES-VERBAL

FORMATION EN REALISATION LINEAIRE ET NON LINEAIRE

« DE LA DONNEE A LA CONNAISSANCE »

Le 28 septembre 2023 par vidéoconférence

Destinée aux organismes du CIRTEF :

**ORTB Bénin, RTB Burkina Faso, ORTC Comores, MBC Ile Maurice, RTN Niger, Radio
Lomé Togo – CRTV Cameroun (pas de désignation)**

Présentateur : Ali Oumarou, Directeur Afrique

Formateurs : Roger Roberts, CIRTEF/TITAN
Damien Duriau, RTBF (excusé)

Liste des participants :

ORTB Bénin	Mr Lionel Gbegnonoude (point focal)
	Mme Christelle Sagbohan
	Mr Arnaud Togbenon
	Mr Armel Gbeyihin
	Mr Janvier Tonon
	Mr Thimothée Houton
RTB Burkina Faso	Mme Delphine Yerbenga (point focal)
	Mme Clémence Tuina (excusée)
	Mr Gilbert Tougma
ORTC Comores	Mr Faïçoil Abdallah Mbadjoubme
	Mr Ahamadi Houmadi
	Mr Mohamed Hila
	Mr Irham Mohamed
	Mr Abdillah Mdahoma
	Mr Saindou Abdillah Libasse

MBC Ile Maurice	Mr Rakesh Gooljar (point focal) (excusé) Mr Mahadeven Mootoosawmy
RTN Niger	Mme Hamey Ramatou Doulla (point focal)
TVT Togo	Mr Fidel Mawuena (point focal) Mme Rachel Kpizing (excusée) Mr Egome Amah Mr Mathias Gbegbeni N'Lipibe Mr Dissirama Kossi Sodhar
VOV Vietnam	Mme Dieu Van Mme Thi Thu Ha Mme Thuy Van Nguyen (excusée)
CSTIC/ISTIC (Institut des Sciences et Techniques de l'Information et de la Communication du Burkina Faso)	Mme Mafarma Sanogo
CRPF Cotonou	Mr Anselme Awannou
CRTF Yaoundé	Mr Hubert Atangana
CIRTEF	Mme Lucia Verdone Mr Daniel Brouyère (excusé)

1. CONTENU

Au nom de l'Equipe Formation «CIRTEF» nous aimerions vous remercier de vous être « accroché » au sujet ! C'est une introduction ardue, conceptuelle ... qui touche de près à tout ce qui nous vient de la philosophie et des mathématiques ... et selon des études scientifiques, le monde entier serait mathématique.

Cette introduction est importante car il faut comprendre cet univers informatique qui pilote la publication en non linéaire ... C'est un changement de paradigme (une modification profonde de la façon de penser, d'agir ou de voir le monde qui se produit souvent lorsqu'une nouvelle technologie ou une nouvelle découverte modifie radicalement le processus de production ou de compréhension d'un phénomène).

Dans cette formation, nous avons mis l'accent :

- sur la manière de coder du binaire ... ce n'est pas fondateur, mais cela permet de comprendre le caractère calculateur de la machine informatique et d'interpréter correctement les valeurs de codages des textes (ASCII – UNICODE), images et sons (sur 8 ou 16 bits), des problèmes de transcodages !
- sur la définition de la donnée : savoir comment se fabrique une donnée (à l'aide de connaissances) permet de bien comprendre pourquoi l'échange entre système d'information est aussi « complexe » !
- sur la distinction donnée - métadonnée : le caractère « meta » devrait être réservé aux informations de représentation et de structuration de la donnée ... même « Bing » ne comprend pas la différence !!!

Cette formation explique donc comment des modèles mathématiques peuvent être exprimés par des langages formels sémantiques afin de contribuer à l'échange de données entre systèmes d'information et surtout constituer des outils pour forger des archives pérennes.

Merci donc pour votre résilience ... Les documents ci-joints devraient vous aider à pénétrer davantage dans cet univers virtuel et vous aider dans les productions à venir !

Si vous souhaitez visionner le PPT présenté en séance, veuillez bien à le télécharger et l'enregistrer car le lien expirera **le 31 décembre 2023** : <https://we.tl/t-JEYmbZwgUG>

Le document textuel reprenant une transcription de la présentation PPT figure à la suite de ce PV.

2. ENREGISTREMENT DE LA FORMATION

1^{ère} partie :

https://us02web.zoom.us/rec/share/-Ornk8X7sq6Br1I4DY4-01JI0oIQY1-Jhb_CwOAYAOQdDDfuCMyn_zbzJa7Vuy48.D-3g6SSeXkTSVTZD?startTime=1695891988000

2^{ème} partie :

https://us02web.zoom.us/rec/share/-Ornk8X7sq6Br1I4DY4-01JI0oIQY1-Jhb_CwOAYAOQdDDfuCMyn_zbzJa7Vuy48.D-3g6SSeXkTSVTZD?startTime=1695895458000

3. PROCHAINE FORMATION

DU PIXEL A L'IMAGE : mercredi 04 octobre 2023

HORAIRE EN GMT :

De 8H45 à 9H00 : connexion

De 09H00 à 11H00 : partie théorique

De 11H00 à 13H00 : pause

De 13H00 à 16H30 : partie pratique

Voici déjà le lien qui vous permettra de vous y connecter :

<https://us02web.zoom.us/j/3923235558?pwd=am1aSVBLNDEvTVVHRk90UGFVaXdyUT09>

CIRTEF – TITAN asbl

La formation réalisation linéaire et non linéaire

“De la donnée à la connaissance”

Ali Oumarou et Roger Roberts
27/09/2023

Colophon

Base : document issu base réunion de travail du 26 octobre 2022

Transcription du PPT

Caractéristiques :

Texte :

Sous-titres :

Lien : PPT Formation réalisation pour le linéaire et le non-linéaire V2023_01_11

Nouvelle version :

PPT : CIRTEF formation juin data V2023_06_02 rro aom

Travaux du 2 juin 2023

Correction du 4 juin :

Le texte sur la démonstration dans excel doit porter sur l'importance du formalisme ... à la fois pour les humains et la machine ... sans formalisme pas de communication (idem plus haut) ! J'ai ajouté deux paragraphes dans la préface à ce sujet !

Travaux été automne 2023 :

- Restructuration du codage binaire
- Mise en forme structuration en table vs graphe

Table des matières

Préface :.....	4
Une introduction à la problématique :.....	5
1 Introduction :.....	7
1.1 Vision technologique de l'univers audiovisuel interconnecté:	7
1.2 Les services audiovisuels linéaires et non-linéaires	10
1.3 La représentation et la signification : humaine et machine (exercice excel).....	12
2 De la DONNÉES à la SAGESSE (WISDOM) :	13
3 Numérisation : codage - transcodage– décodage.....	14
4 La représentation et la signification de l'INFORMATION par les DONNÉES	15
De l'importance de la symbolique dans la représentation dans les unités de mesure :	15
4.1 de la représentation arithmétique au codage binaire :	16
4.2 Le code MORSE :.....	17
4.3 Le traitement textuel :.....	18
4.4 Normalisation du traitement de textes : UNICODE.....	21
5 La définition de la donnée :.....	22
6 DATAVISION : la modélisation de l'INFORMATION par les DONNÉES :.....	24
• Niveau Matière.....	24
• Niveau Représentation :.....	25
▪ Niveau Symbolique	25
▪ Niveau Élémentaire.....	26
▪ Niveau Cellulaire	26
▪ Niveau Documentaire	27
○ Niveau Patrimoine Documentaire :	27
▪ Exemple : Patrimoine documentaire : les archives Belgavox - Congovox :	28
7 L'analyse des différents niveaux du modèle DATAVISION :	29
○ M. Niveau « Matière » la capacité d'être porteur de signifiants et son support.....	29
○ Niveau représentation :.....	29
▪ Le niveau Symbolique :	29
▪ Niveau 1 : Élémentaire.....	30
▪ Niveau 2 : Cellulaire	33
▪ Niveau 3 : « documentaires »	33
▪ l'interopérabilité entre les niveaux et intra-niveau :.....	34
○ Niveau 4 : Patrimoine documentaire :.....	35
8 DATAVISION : les couches Meta 1 (Représentation) et Meta 2 (Sémantique) :.....	35
Représentation de l'Information (Formats / Codages / Protocoles / Sémantique / ...) :.....	35
• Représentation de l'Information Meta R et Meta S :	36

• Les « Métadonnées R et S » expriment les contextes et le cadre d'interprétation de « données » :	37
9 Données - Métadonnées : exemples niveaux d'interopérabilité	38
o Exemple : de l'objet donnée à la représentation :	38
10 la structuration des données : modèle table vs graphe	41
10.1 L'indexation : de la « Table » au « Tableau » :	42
10.2 Les bases des données relationnelles :	43
10.3 Les bases de données orientées graphe	44
10.4 La structuration des données en table vs en graphe :	44
10.5 La structuration des données en table vs en graphe :	45
10.6 Les données ancillaires des fichiers :	45
10.7 Classification des données :	46
11 La représentation et la signification :	47
12 Quelques définitions	49
12.1 Symbolique : la relation Concept - Symbole :	49
12.2 Quelques définitions :	49
12.3 Archivage : quelques définitions	50
13 Conclusion :	51

Table des illustrations

Figure 1-1 : l'interaction Homme - Machine.....	7
Figure 1-2 : documents (textes, sonores, vidéo, ...) – structuration et contexte.....	8
Figure 1-3 : organisation des documents non linéaires	9
Figure 1-4 : DNL : Document Non Linéaire : l'expression d'un [Contexte]	10
Figure 1-5 : l'irruption des technologies interactives.....	11
Figure 1-6 : exercice pratique de formalisation dans l'application « Excel »	12
Figure 2-1 : De la DONNÉE à la SAGESSE (WISDOM)	13
Figure 3-1 : le processus de codage numérique.....	14
Figure 3-2 : le processus de décodage numérique	14
Figure 4-1 : différentes représentations du nombre 9018.....	16
Figure 4-2 : une machine à écrire.....	18
Figure 4-3 : ASCII (American Standard Code for Information Interchange)	19
Figure 4-4 : Exemple d'une table de codage de la lettre E – valeur 105	20
Figure 4-5 : barre des outils d'un traitement de textes	21
Figure 6-1 : Le niveau matière : la pierre, le parchemin, les livres, les fichiers,	24
Figure 7-1 : DATAVISION : le niveau « matière »	29
Figure 7-2 : DATAVISION : niveau symbolique	30
Figure 7-3 : DATAVISION : le niveau élémentaire.....	30
Figure 7-4 : l'échantillonnage et la quantification audio	31
Figure 7-5 : l'échantillonnage – la quantification et la compression numérique TV	32
Figure 7-6 : DATAVISION : niveau cellulaire	33
Figure 7-7 : DATAVISION : le niveau documentaire	34

Figure 7-8 : interopérabilité entre niveau 2 et 3 : exemple JPEG	34
Figure 7-9 : DATAVISION : niveau Patrimoine Documentaire	35
Figure 8-1 : La représentation des DONNÉES par des Metadonnées	36
Figure 8-2 : La représentation des DONNÉES par des Metadonnées R et S	36
Figure 8-3 : La représentation des DONNÉES par des Metadonnées	37
Figure 8-4 : DATAVISION : Meta R (Représentation – Meta S (Sémantique)	37
Figure 9-1 : un fichier à l'état brut	38
Figure 9-2 : information de représentation : codé selon la norme ISO 646	38
Figure 9-3 : représentation dans un convertisseur ASCII	39
Figure 9-4 : information de représentation : codé en table	39
Figure 9-5 : le bilan climatologique belge de 1998 à 2013 publié par l'IRM.	40
Figure 9-6 : les normes impliquées dans la représentation de ce fichier ODF	40
Figure 10-1 : représentation de données en mode table	42
Figure 10-3 : description de la définition des tables d'une base de données.....	43
Figure 10-4 : données ancillaires d'une image (fichier) JPEG.....	46
Figure 10-5 : classification de différents types de données !.....	46
Figure 11-1 : TITAN Modèle Conceptuel : un processus intégré ?	47

Préface :

Depuis les Grecs et selon des études scientifiques, le monde entier est mathématique. Il est organisé suivant des règles qui se représentent admirablement par des expressions mathématiques . Il existe une possibilité de fusion entre une représentation formelle du monde à partir des technologies de la mathématique et les commentaires relatifs à ces modèles mathématiques qui peuvent être exprimés par des langages formels sémantiques.

L'approche humaine, humaniste, est de faire un discours de manière sémantique en accrochant à ce discours sémantique des modèles mathématiques qui, en quelque sorte, sont porteurs de toute l'épistémologie scientifique.

Cette approche est cohérente au sens où l'approche mathématique par son essence même déclare qu'elle est incapable de décrire complètement le monde.

Depuis Gödel on sait très bien que quelle que soit l'approche mathématique et axiomatique que l'on prendra, il existera toujours un nombre important (infini dans certains cas) de propositions qui seront indémontrables dans le système mathématique axiomatique concerné !

Guy Maréchal.

Une introduction à la problématique :

L'univers numérique est la source d'un tsunami de données. Une réflexion approfondie passant par la découverte de données, le questionnement sur ces données, leur interprétation et recherche de sens, leur agencement ainsi que la recherche de réponses quant à leurs relations nécessitent des conceptualisations sophistiquées. Il faut donc pouvoir exprimer dans un outil informatique la différence entre une représentation et une signification !

De toutes les formations, c'est la plus conceptuelle ! En effet pour comprendre l'univers dans lequel nous travaillons nous avons besoin de connaissances qui nous aident à comprendre des nouveaux processus comme le passage d'une réalisation linéaire à des contenus non-linéaires ! Les concepts nous aident à comprendre des idées abstraites tels que la communication, le changement et les relations.

Dans le domaine de la communication, ils sont importants car ils nous permettent d'organiser les informations et les idées en catégories, nous aident à comprendre les liens et les interactions entre les propriétés, les objets, les individus et les idées, à communiquer efficacement en utilisant un langage commun, ... En résumé, les concepts sont importants car ils nous aident à développer notre compréhension du savoir.

« De la donnée à la connaissance » ... le titre est à l'image de l'ambition de ce module qui traite des technologies numériques ! Le numérique est omniprésent dans nos sociétés et a un impact majeur sur nos dynamiques sociales, culturelles et cognitives. Cependant, il peut être difficile de saisir la portée et les implications du numérique, d'où l'importance de développer une intelligence du numérique pour le rendre intelligible.

Dans ce contexte distinguer les données des métadonnées est essentiel dans le domaine de l'informatique et de la gestion des informations.

- Une donnée est une représentation formalisée d'informations (faits, concepts ou instructions) exprimée selon un formalisme adapté à la communication, à la mémorisation, à l'interprétation ou au traitement par des êtres humains ou par des moyens automatiques. Par exemple, les données peuvent être une liste de mesures de température ou une description d'un objet.
- D'autre part, les métadonnées sont des données qualifiant des données. Elles décrivent les informations pertinentes en termes de représentation et de structuration, ce qui aide à identifier leur nature et leurs caractéristiques. Les métadonnées fournissent un contexte supplémentaire pour comprendre et interpréter les données.

Par exemple « ASCII » (American Standard Code for Information Interchange) est un standard de codage de caractères utilisé pour représenter du texte dans les ordinateurs sur la base de l'alphabet latin. Il fournit une table de correspondance entre les valeurs numériques et les caractères de l'alphabet latin non accentué. Ainsi, il permet d'interpréter les valeurs attribuées aux caractères et de les représenter correctement. Ce standard remplit donc bien une tâche « meta » à savoir une connaissance de la manière dont les caractères de l'alphabet latin ont été codés !

La modélisation générique des méta-documents sous forme de graphe intègre les relations spatiales (actuelles) et temporelles (dans le temps) des éléments. Les relations spatiales et/ou temporelles expliquent comment les données ou les éléments d'un même document sont reliés. Compte tenu de la représentation du méta-document, les relations spatiales et temporelles sont présentées comme des liens dans un graphe.

Pour exprimer des contenus complexes (avec la phrase comme base sémantique), des structures et des contextes (via des ontologies avec les classes modélisées ainsi que leurs relations et propriétés), il paraît hautement souhaitable d'adopter le concept de « donnée sémantique » qui balise correctement et simultanément les champs de l'opérationnel et de l'archive.

Ce module est essentiel pour faciliter le dialogue dans le futur avec le « data scientist alias le docteur "do" pour données ! Sa mission consiste à transformer les données massives (big data) en informations exploitables, afin d'améliorer les performances d'une organisation de radio-télévision. Pour ce faire, il recueille, trie, traite et analyse les données massives à l'aide de modélisations complexes, accompagnées d'échanges avec les collaborateurs, qui ont une connaissance approfondie des processus de fabrication d'un contenu ! Ce docteur "do" est vital afin de permettre de préconiser des solutions pour régler certaines problématiques de publication et d'archivage.

Bonne lecture ... l'essentiel du travail porte sur le fait d'avoir une compréhension commune des concepts de la science de l'information !

Ali Oumarou
Directeur Afrique

Roger Roberts
Consultant CIRTEF

Septembre 2023

1 Introduction :

1.1 Vision technologique de l'univers audiovisuel interconnecté :

Depuis une vingtaine d'années, le développement des technologies de l'information et de la communication offre de nouvelles situations d'interaction avec la représentation de l'information.

L'**interaction Homme-machine IHM** (ou interaction humain-machine) traite de la conception et du développement de systèmes interactifs. Les humains interagissent avec de multiples périphériques qui les entourent et cette interaction nécessite des interfaces intuitives qui facilitent la communication entre l'humain et la machine. L'IHM repose sur la vision par ordinateur, sur l'humain mais aussi sur la sémantique (la partie de la linguistique qui traite du sens, de la signification des signes, notamment dans le langage) afin d'offrir des relations explicites entre des objets exploitables par l'informatique.

L'interaction homme-machine est un domaine pluridisciplinaire entre ingénierie (informatique, électronique, mécanique...), science de la nature (sciences cognitive, psychologie, sociologie...) et art et design (design de produit, design interactif, ergonomie...).

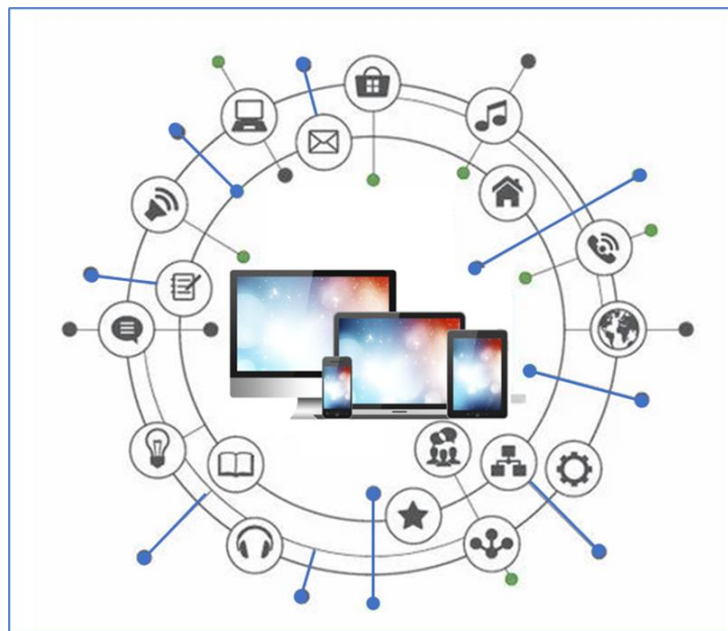


Figure 1-1 : l'interaction Homme - Machine

L'histoire de l'interaction homme-machine est aussi vieille que l'histoire de l'informatique :

- En 1945 : conception d'un système d'information électronique qui invente les concepts de navigation, d'indexation et d'annotation (Vannevar Bush).
- En 1963 : Sketchpad : l'ancêtre des interfaces graphiques modernes (Ivan Sutherland).
- En 1964 : invention de la souris pour désigner des objets sur son écran (Douglas Engelbart).
- Entre 1970 et 80, les laboratoires de Xerox ont révolutionné les systèmes interactifs avec la sortie de Xerox Star (un Windows avant la lettre).
- 12 août 1981 : arrivée du PC d'IBM sur le marché. L'appareil connu aussitôt le succès dans les entreprises et les familles.
- Au début des années 1990, Robert Cailliau et Tim Berners-Lee étudient et développent au CERN un système hypertexte qui entourera la planète, World Wide Web.
- En 1991, Mark Weiser présente sa vision de l'Informatique ubiquitaire qui envisage des écrans et des ordinateurs multiples capables de communiquer entre eux pour permettre

l'utilisateur à accéder à l'information en toute circonstance. Cette vision préfigure clairement l'avènement des assistants personnels, Tablet PC et smartphones d'aujourd'hui.

Ces dispositifs sont de plus en plus présents et importants, vu le nombre croissant d'interfaces numériques utilisés dans la vie quotidienne pour des tâches multiples.

La recherche d'information multimodale constitue un problème très complexe : les données sont ou non structurées, dispersées, hétérogènes, multimédia, etc. C'est dire la multiplicité d'accès, de contenu, de forme qu'elles nécessitent. Elles concernent des experts qui connaissent à la fois les représentations (outils de formalisation des essences : texte – audio - ...) et les significations (sens de ces données).

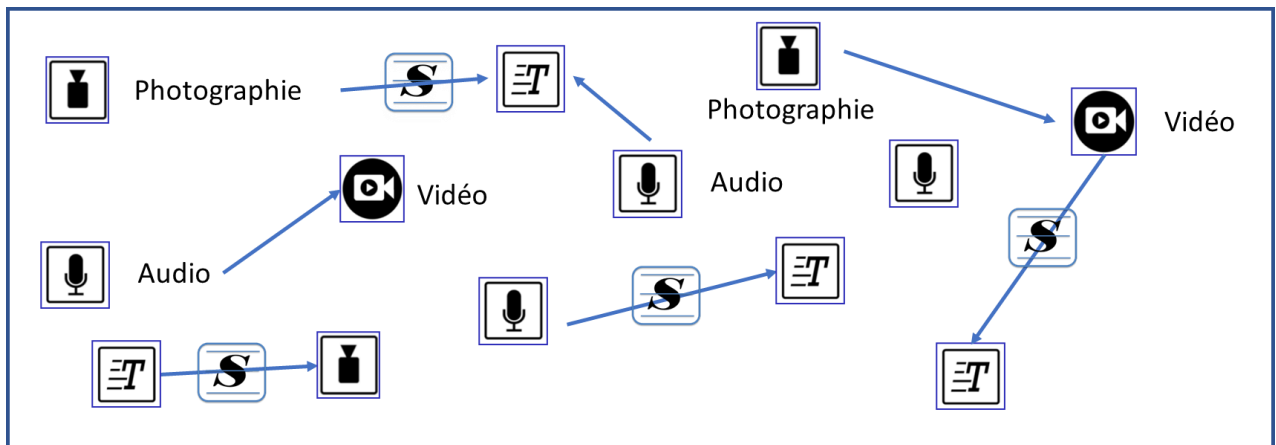


Figure 1-2 : documents (textes, sonores, vidéo, ...) – structuration et contexte

Cela pose des problèmes d'indexation, de traitement de contenu, d'accès individuel ou collaboratif, d'archivage, etc. En plus du texte, l'image, le son, la parole, ... entrent de plus en plus dans le champ des recherches quand il s'agit par exemple de trouver la séquence sonore radiodiffusée par un média (concept de personne morale) : « trouver la déclaration de monsieur X (concept de personne physique) parlant d'un sujet Z (d'événement – date - ... - ...) représentée dans un contexte.

Parmi ces technologies, les documents numériques non-linéaires tiennent une place importante dans les situations de production, de publication ou encore de gestion documentaire. Ces nouveaux types de documents sont aujourd'hui principalement incarnés par les dispositifs hypertextuels.

Dans cette formation, le terme de **DNL (document non-linéaire)** est employé car celui-ci présente l'avantage d'être assez générique pour englober les divers termes employés comme lien, hyperlien, documents hypertextuels ou hypermédias, sémantique.

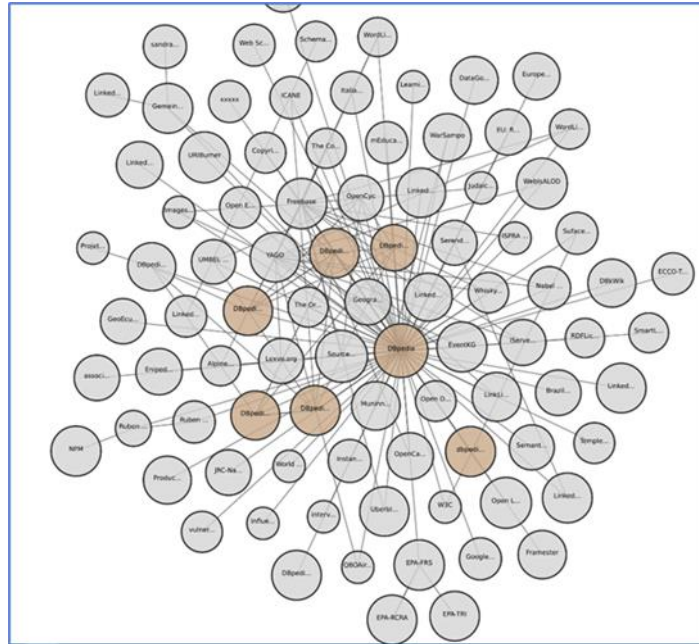


Figure 1-3 : organisation des documents non linéaires

Les DNL sont organisés en réseau dans lequel les nœuds sont :

- des entités nommées (interprétables par l'informatique (exemple : Ouagadougou – ville du BF),
- des segments textuels (ex. pages, paragraphes,),
- des photos,
- des fichiers audio ou vidéo.

Ces nœuds entretiennent, via des structures explicites, des relations généralement d'ordre sémantique (telles que relations, définitions, expansions, exemples, etc.) et expriment des contextes.

L'interaction avec un document non-linéaire consiste donc à passer de l'exploration d'un segment textuel, iconographique à une navigation dans un corpus via des liens proposés par le concepteur ou généré par l'outil informatique (raisonnement : inférences sur les contenus ciblés par un lien) !

L'utilisation de DNL dans le domaine audiovisuel intervient dans deux types de tâches :

- la première concerne **la collecte d'informations** au cours desquelles le réalisateur/journaliste/documentaliste, ... vise prioritairement la localisation d'une information cible à l'intérieur de multiples contenus sur base textuelle (dépêches, sous-titres, indexations,) ou audiovisuelles ;
- la seconde concerne des tâches au cours desquelles le réalisateur/journaliste vise plutôt la **construction d'une représentation** de l'ensemble des contenus de documents divers en fonction d'une ligne éditoriale.

Contrairement à une tâche de recherche d'information précise, les tâches d'expression d'une représentation contextuelle induisent des buts à un haut niveau d'abstraction amenant les acteurs à inférer la nature des informations pertinentes pour le but.

La tâche de compréhension implique alors généralement une activité d'étude des documents pouvant se traduire par une représentation graphique facilitant l'exploration et la navigation dans un corpus de documents dans l'objectif de présenter un réseau (un but ouvert) dans le cadre d'une production collaborative, la base d'une réalisation audiovisuelle.

Les représentations peuvent être structurées en interne et être liées entre elles comme la construction de «réseaux sémantiques» de représentations et de contextes associés [exprimés dans les “Bases de Connaissances”].

Exemple : Trace d'un ÉVÉNEMENT dans la « Marche sur Washington pour l'EMPLOI & LA LIBERTÉ » («March on Washington for JOBS & FREEDOM”) en 1963 :

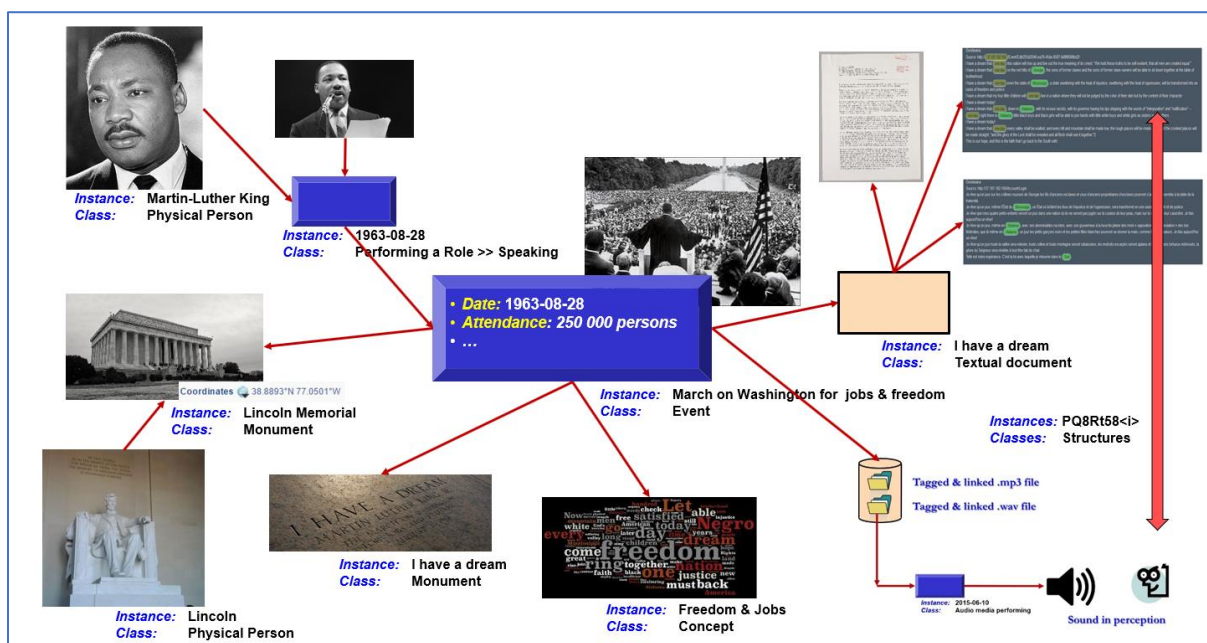


Figure 1-4 : DNL : Document Non Linéaire : l'expression d'un [Contexte]

Dans cette approche, la modélisation décrit des «événements» et des «concepts», comme illustré à la figure « I have a dream ».

« Le 28 août 1963, le jeune pasteur de 34 ans Martin Luther King prononçait un discours emblématique de huit minutes devant le Lincoln Memorial de Washington à l'issue de la marche pour le travail et la liberté qui avait réuni plus de 250 000 personnes.

La traçabilité permet de représenter les liens entre les œuvres, leurs affiliations et leurs contextes. La place dans l'Histoire d'un discours (Martin Luther King sur les marches du Lincoln Center à Washington) ne peut se comprendre sans le relier à son contexte, la présentation de l'événement dans lequel il a été prononcé, les événements qui y sont associés, avec articles dans les journaux, réactions sociopolitiques, écoute de ce discours tel qu'il a été prononcé, possibilité d'accéder à sa transcription (y compris en braille) et à sa traduction dans de nombreuses langues. L'interopérabilité culturelle implique l'intégration de la gestion de [quelque-chose] avec la gestion des contextes.

Le fait de cliquer sur n'importe quel champ provoque une reconstruction d'un nouveau graphe (View) en fonction de la sélection opérée (mécanisme de zoom dans un contexte) !

1.2 Les services audiovisuels linéaires et non-linéaires

Depuis une vingtaine d'années, le développement des TIC (technologies de l'information et de la communication) offre de nouveaux modes d'interaction avec la représentation de l'information. Parmi ces technologies, les DNL (documents numériques non-linéaires) tiennent une place importante dans les situations de production, de publication ou encore de gestion documentaire.

Ces nouveaux types de documents sont aujourd'hui principalement incarnés par les dispositifs hypertextuels. Leur non-linéarité renvoie à une offre de contenus choisis par l'utilisateur parmi un nombre de choix possibles définis et offerts par le fournisseur de services.

Les services audiovisuels linéaires et non-linéaires correspondent à une typologie de services qui prend en compte la diversification des modes de transmission, de réception et de consommation des programmes et services à l'heure du numérique (télévision numérique terrestre, télévision par ADSL, télévision sur téléphone mobile, web, ...), ainsi que la diversification des modes de diffusion (diffusion point – multipoint en mode broadcast ou vidéo à la demande en mode point à point).

- Les **services linéaires** sont les services de radiodiffusion (audio – vidéo) contrôlés par le diffuseur. Le contenu est acheminé en mode diffusé vers un public indéterminé situé dans une zone de réception en mode point vers multi points (push content).
- Les **services non-linéaires** sont les services interactifs où le contenu est contrôlé par l'utilisateur (pull content). Par conséquent, la transmission interactive est en mode « point à point », qu'il s'agisse de VOD reçue sur de multiples périphériques (la télévision, sur un téléphone portable ou tout autre écran informatique), ou de navigation dans des contenus textuels/audiovisuels sur des sites Web, contenant souvent les transcriptions électroniques de journaux d'information de télévisions et de radios.



Figure 1-5 : l'irruption des technologies interactives

Cette dernière transformation d'interactions avec un contenu décliné de manière multiple a notablement modifié les chaînes de fabrication collaboratives des Media traditionnels qui produisaient en silo (une chaîne radio, une chaîne TV, un service Internet).

Par le passé, ces deux tâches s'enchaînaient .. Ce n'est plus le cas aujourd'hui car la première tâche facilite à la fois la construction d'un contenu non linéaire et le formatage d'un contenu linéaire ! Il devient donc possible de concevoir une chaîne de fabrication collaborative pour la création et la préparation qui hérite de toutes les caractéristiques de la construction d'un site web !

1.3 La représentation et la signification : humaine et machine (exercice excel)

Le traitement d'une donnée implique trois étapes :

- la représentation des données dans des formalismes informatiques,
- l'exploitation informatique de ces données formalisées par des applications,
- et enfin l'interprétation des résultats et leur utilisation.

Dans la réalité, ces trois étapes sont fortement liées les unes aux autres et nécessitent des allées et venues nombreuses car l'interprétation consiste à donner du sens aux résultats formels en les comparant à la réalité observée ! Elle tient également compte des choix opérés dans la représentation, lesquels sont fonctions des exploitations envisagées.

La programmation consiste en effet à assembler, à partir d'un formalisme de base, les différents objets calculables dont elle a besoin pour représenter une connaissance donnée. Ces dernières dépendent elles-mêmes des données disponibles et utilisables.

Un des problèmes-clefs de l'informatique est qu'elle ne peut progresser qu'en améliorant ces "schèmes" de base (les formalismes). Ceux-ci montrent souvent leurs limites dès qu'il s'agit de formaliser la complexité des problèmes que posent la représentation des connaissances et leur manipulation par des utilisateurs naïfs.

RUB.	LIBELLES	TOT IMMEUBLE	QUOTITES		
1	ASSURANCES 9880 Q	27.387,03 €	103	/ 9880	285,51 €
2	COMPLEMENT PROVISION SYNDIC 9150 Q	18.300,00 €	103	/ 9150	206,00 €
3	FRAIS RESIDENTIELS 8761Q	45.743,90 €	94	/ 8761	490,80 €
4	COMPLEMENT PROVISION SYNDIC CRECHE 730 0	730,00 €	0	/ 730	0,00 €
5	CHARGES FINANCIERES 9880 Q	-836,64 €	103	/ 9880	-8,72 €
6	FRAIS PRIVATIFS	6.970,10 €	0	/ 0	147,26 €
8	DIVERS - 9880 Q	13.760,08 €	103	/ 9880	143,45 €
9	GESTION APTS - .97	15.132,00 €	1	/ 97	156,00 €
10	ENTRETIEN IMMEUBLE - 9150 0	28.665,70 €	103	/ 9150	322,68 €
11	CONCIERGE 9880 Q	34.954,71 €	103	/ 9880	364,41 €
12	ASCENSEURS - 8761Q	23.502,16 €	94	/ 8761	252,16 €
13	JARDINS - 9880 Q	5.489,37 €	103	/ 9880	57,23 €
14	EAU CHAUDE - 8761Q	33.846,20 €	94	/ 8761	363,15 €
15	CONSOMM."CHAUFFAGE"	108.759,25 €	0	/ 0	
16	EGOUTS - 9491Q	4.303,42 €	94	/ 9491	42,62 €
17	ABONN.REDEVANCE EAU 96 APTS	1.848,41 €	1	/ 96	19,25 €
18	CONSOMM.SUPPL.EAU	21.366,29 €	0	/ 0	260,25 €
19	GARAGES 49 Q	490,00 €	1	/ 49	10,00 €
21	INVESTISSEMENTS - 8761Q	- 713,70 €	94	/ 8761	- 7,66 €
22	INVEST ASC. - 8761Q	- 895,03 €	94	/ 8761	- 9,60 €
TOTAL CHARGES					

Figure 1-6 : exercice pratique de formalisation dans l'application « Excel »

Exercice pratique de formalisation dans l'application « Excel » :

- formalisme des montants représentés dans le tableau (interprétation machine)
- formalisme d'une représentation d'une date !

2 De la DONNÉE à la SAGESSE (WISDOM) :

Les données ne sont pas le pouvoir; la connaissance est le pouvoir, et la connaissance vient des contextes et connexions :

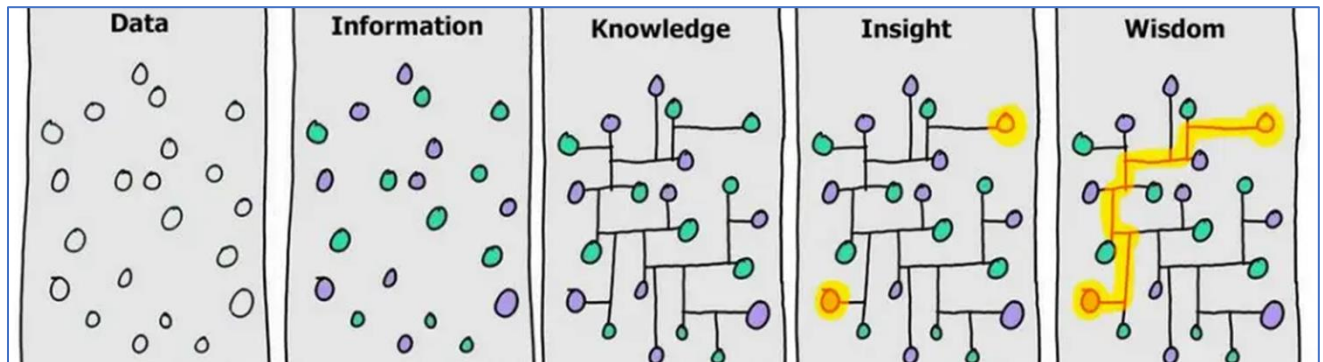


Figure 2-1 : De la DONNÉE à la SAGESSE (WISDOM)

Voici comment fonctionne notre cerveau :

- Reconnaître les choses,
- Connecter les choses,
- Résoudre des problèmes,
- Prédire l'avenir en reliant les concepts d'information et les faits.

Pour être utilisables, les données doivent être :

- classées et identifiées pour représenter l'information,
- liées pour devenir connaissance,
- analysées pour créer des contextes,
- généralisées pour devenir sagesse.

Dans ce schéma, seules les parties donnée – information – connaissance sont dans la machine informatique. Le contexte et la sagesse sont de l'ordre de l'humain.

Le schéma explique bien l'existence d'une différence de taille entre l'information et la connaissance. C'est la "connaissance" qui est le pouvoir; pas des données ou des informations, mais des connaissances. En fait la donnée résulte d'une information qui a été transformée (formalisée) par une connaissance !

Dans un monde où nous disposons d'une quantité incroyable de données (certaines d'entre elles très douteuses) à portée de main, c'est la façon dont nous les rassemblons, les intégrons et les appliquons que ces données procurent des [contextes].

Notre cerveau peut établir des liens entre des données qui étaient déjà présentes dans sa caboche, et créer de nouvelles idées et de nouvelles conclusions à partir de celles-ci et recréer de nouvelles idées et de nouvelles conclusions à partir de celles-ci.

C'est cette capacité à créer ces connexions, et à partir d'elles des idées originales, qui est moteur.

L'informatique a généralisé le terme « data » pour qualifier la description élémentaire d'une réalité, le résultat d'une comparaison, d'une observation ou d'une mesure. Une donnée est une forme contingente (signifiant) associée à un concept (signifié). Dans un système d'information, la donnée est la base de la représentation.

La nécessité du discours sémantique est inéluctable parce qu'elle redonne à l'homme le moyen d'exprimer le sens qu'il donne aux choses et de s'appuyer sur les modèles mathématiques en vue d'être efficace.

La sémantique rend possible de remplir les vides ou de déclarer l'impossibilité de démontrer mathématiquement ou scientifiquement des représentations complexes. Elle est porteuse de sens et donne toute sa dimension au caractère humain. Elle rejoint en cela l'aspect philosophique.

3 Numérisation : codage – transcodage– décodage

La numérisation est un processus qui transforme des données textuelles/graphiques, iconiques, sonores/musicales, audiovisuelles, 3D, ...) ou un signal électrique en données numériques (fichiers/dossiers) que des dispositifs informatiques numériques pourront traiter, rendre accessible et exploiter en ligne !

La **numérisation** traite exclusivement l'aspect « représentation » des documents.

Le **codage** de l'information désigne les moyens de formaliser l'information :

- Un codage des caractères définit une manière de représenter les caractères (lettres, chiffres, symboles) dans un système d'information.
- Un encodage sémantique est un procédé de transformation d'un langage formel (typiquement, un langage de programmation, de description ou un langage mathématique) en un autre langage formel = la capacité de produire une représentation signifiante de quelque chose et à communiquer ce sens.

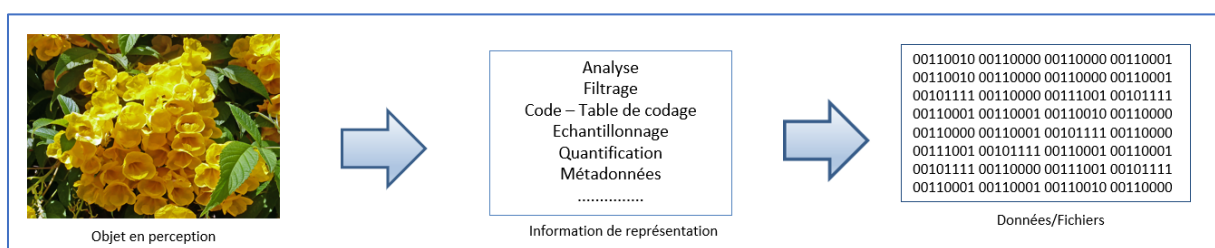


Figure 3-1 : le processus de codage numérique

Le **Transcodage** : la traduction d'une information dans un code différent. Il ne s'agit pas d'un codage au sens strict du terme car le plus souvent le transcodage génère des pertes lors du recodage.

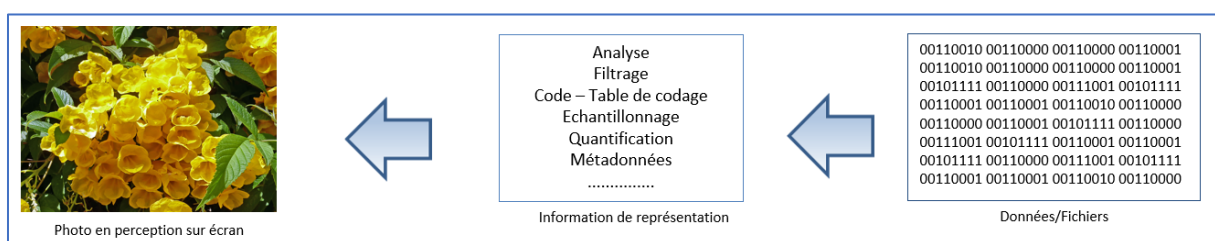


Figure 3-2 : le processus de décodage numérique

Le **Décodage** : le processus inverse s'appelle décodage afin de remettre l'objet codé préalablement dans une forme techniquement comparable à l'original. (Exemple : la carte son qui reconstitue dans un mode synchrone la forme d'onde d'origine pour l'envoyer vers les haut-parleurs en tant que signal analogique).

La numérisation implique l'usage d'un [codec] (COdeur-DEcodeur) qui transforme le signal en fichier et le fichier en signal en vue de représenter l'objet en perception et de codes de correction d'erreur en vue de pouvoir recréer l'objet avec des bits manquants.

La numérisation génère le "code génétique" d'un document (les informations de représentation. La connaissance de cette suite de valeurs, permet de reproduire un document à l'identique, à condition de ne pas coder : transcoder avec pertes ([compression numérique])).

NB : Programmer vs coder : Programmer, c'est créer la logique, coder, c'est traduire cette logique en code

4 La représentation et la signification de l'INFORMATION par les DONNÉES

De l'importance de la symbolique dans la représentation dans les unités de mesure :

Le 22 décembre 1666, Colbert, ministre de Louis XIV, réunissait une quinzaine de savants dans la Bibliothèque du roi, rue Vivienne, à Paris, avec mission « d'avancer et favoriser la science pour l'utilité publique et la gloire de son règne ». Elle va regrouper des scientifiques de tous domaines : astronomie, botanique, chimie, mathématiques, paléontologie, zoologie, ...

C'est cette « Académie royale des sciences » qui publie en 1790 la réforme générale des poids et mesures qui est à l'origine du système international d'unités de mesure encore en vigueur.

Familiarisé depuis des siècles avec les arpents, les pintes, les muids, les setiers, les boisseaux et les livres, le citoyen rencontrait quelques difficultés à admettre qu'un kilo de plumes pèse pratiquement le même poids qu'un kilo de plomb. Le citoyen qui avait coutume d'acheter 1 muid, 4 setiers, 8 minots de farine pour un montant de 2 livres, 15 sous et 6 deniers, devait à présent faire l'achat de 12,5 kg de farine pour 1,2F.

Habitué à compter les distances en lieues et en pas, le citoyen eut quelque résistance à passer au mètre. Mais la Science et la République surent le mettre au pas les citoyens français .. et étrangers qui étaient alors sous domination française.



À l'origine, l'Académie a choisi de définir les unités sur base de grandeurs effectives : c'est pourquoi ils ont mandaté la mesure de la dimension de la terre et de définir arbitrairement qu'elle avait 40 000 km de circonférence.

Il en a été ainsi dérivé un objet de référence, déclaré ayant 1 mètre de long, qui a été déposé au Pavillon de Breteuil à Sèvres en France.

Aujourd'hui, le « Système International des unités de mesure » [MKSA] est basé sur le Mètre, le Kilogramme, la Seconde, l'Ampère, le kelvin, la Mole et le Candela.

Les unités de mesure sont définies arbitrairement sur base de constantes physiques : la vitesse de la lumière « c » ; la constante de Planck « h » ; la charge de l'électron « e » ...

4.1 de la représentation arithmétique au codage binaire :

La notation positionnelle est un procédé d'écriture des nombres, dans lequel chaque position d'un chiffre ou symbole est reliée à la position voisine par un multiplicateur, appelé base du système de numération. Dans le système décimal, la valeur d'une position est celle du symbole multipliée par la base : un multiple de 10.

Dans l'Antiquité, on utilisait exclusivement de nombreux systèmes non-positionnels, dont l'exemple le plus connu est la numération romaine, où le nombre trente-huit, par exemple, s'écrit à l'aide de pas moins de sept chiffres (XXXVIII), tandis que le nombre cinquante, se contente d'un seul (L).

En chiffres romains, vous ne pouvez pas répéter un symbole plus de trois fois consécutives.

En conséquence, la taille d'un nombre que vous pouvez écrire est limitée. Le plus grand chiffre romain est MMMCMXCIX qui est 3,999.

Dans un tel système de notation, une simple opération comme une multiplication se révèle pratiquement impossible à effectuer sans abaque (boulier, tablettes de calcul à jetons, ou autre outil de calcul).

L'application de la notation positionnelle à un système décimal fut initiée par les Chinois au II^{ème} siècle av. J.-C., puis finalisée vers l'an 500 de l'ère chrétienne par les brahmanes de l'Inde. Le système est décrit dans l'Āryabhaṭīya, un ouvrage indien — rédigé par Āryabhata et daté de 499 — considéré comme l'équivalent indien de ce que seront les Éléments d'Euclide.

Le système décimal usuel utilise dix symboles, plus les symboles pour les nombres signés et les nombres à virgule.

Dans un système de numération positionnel, par exemple en base 10, le système décimal utilise 10 chiffres (0 à 9) pour représenter les nombres. Chaque position de droite à gauche est un multiple de 10 !

Représentation du nombre 9018 :

- ❑ 9018 s'écrit 九千十八 avec le système utilisé pour **la numération japonaise** (décimale) :
soit $9 \times 1000 + [1 \times]10 + 8$.
- ❑ 9018 s'écrit 九千 零十八 avec le système utilisé pour **la numération chinoise** (décimale) :
soit $9 \times 1000 + 0 [\times 100] + [1 \times]10 + 8$.
- ❑ 9018 avec le système **arabo-indiens** (décimal) :
 $9[\times 1000] + 0[\times 100] + 1[\times 10] + 8$.
- ❑ 9018 s'écrit en mode binaire : **100011 00111010**

Figure 4-1 : différentes représentations du nombre 9018

Dans le système binaire, chaque position de droite à gauche représente une puissance de la base 2. Il faut donc multiplier chaque chiffre binaire 1 par la puissance de deux correspondante (de droite à gauche) : le chiffre le plus à droite est multiplié par 2^0 , le suivant par 2^1 , puis par 2^2 , etc. .. et additionner l'ensemble des produits obtenus.

La technologie numérique est basée sur la numérisation des symboles (caractères, son, image, vidéo, ...) c'est-à-dire transformés en une succession de 0 et de 1 qu'on appelle des bits ou des octets (c'est-à-dire 8 bits).

Les ordinateurs sont d'abord des calculateurs, comme nos simples calculettes, qui, pour effectuer des calculs, convertissent les chiffres décimaux conformément au système binaire : 0 = (00000000), 1 = (00000001), 2 = (00000010), 3 = (00000011), 4 = (00000100), 5 = (00000101), 6 = (00000110), 7 = (00000111), 8 = (00001000), etc., ... sans perdre la moindre information.

Le numérique est un procédé qui consiste à " décrire/coder " différentes formes d'information : des phonogrammes (caractères), des pixels (images) ... à l'aide de nombres. Si l'utilisation du numérique semble naturelle et légitime pour représenter des éléments déjà " chiffrés " (comme des valeurs mathématiques, des valeurs colorimétriques, des fréquences sonores, des dimensions, etc.), elle l'est moins, a priori, quand on parle de phonogrammes, idéogrammes, icônes

Dans le cas des textes, le processus consiste à attribuer de façon arbitraire un numéro unique (donc une valeur) à chaque symbole (lettre de l'alphabet, signe de ponctuation, fonction traitement de texte, ...).

Le nombre des combinaisons possibles est une puissance de deux : pour 4 bits, on a $2^4 = 16$ combinaisons, pour 8 bits, on a $2^8 = 256$ combinaisons, pour 10 bits, $2^{10} = 1024$ combinaisons, et ainsi de suite. C'est très important de disposer d'un nombre de combinaisons suffisant que pour couvrir l'ensemble des signes / fonctions à représenter.

Sur la base d'une table de correspondance, on peut " coder " n'importe quel caractère, idéogramme, ... dont le nombre résulte d'une formulation mathématique arbitraire.

4.2 Le code MORSE :

Inventé en 1832 pour la télégraphie, il est considéré comme le précurseur des communications numériques.

A	• —	U	• • —
B	— • • •	V	• • • —
C	— • — •	W	• — —
D	— • •	X	— • • —
E	•	Y	— • — —
F	• • — •	Z	— — • •
G	— — •		
H	• • • •		
I	• •		
J	• — — —		
K	— • — —	1	• — — — —
L	• — • •	2	• • — — —
M	— —	3	• • • — —
N	— •	4	• • • • —
O	— — — —	5	• • • • •
P	• — — •	6	— • • • •
Q	— — • —	7	— — • • •
R	• — • •	8	— — — • •
S	• • •	9	— — — — •
T	—	0	— — — — —

Ce codage de caractères assigne à un ensemble défini de lettres, chiffres et signes de ponctuation une combinaison unique d'impulsions intermittentes (point/barre) : quatre impulsions au maximum pour les lettres, cinq pour les chiffres. En vue d'optimiser la transmission :

- les caractères les plus fréquents de la langue anglaise sont codés avec peu de signaux (E = point, T = barre) ;

- les séquences plus longues correspondent à des symboles les plus rares : signes de ponctuation, symboles et caractères spéciaux (codage entropique : une compression réversible et sans pertes d'informations d'un

document (texte, image, ...).

Le code morse dispose donc d'une table de caractères propre à un système de signes (charset) qui délimite strictement les champs des représentations (signifiant) avant la transformation en une valeur (exemple Point/barre du Morse). Il est important de distinguer le code (répertoire fermé de signes) qu'un système supporte, d'un jeu de signes codés ou table de codage des signes qui spécifie comment représenter un signe (idéogramme, phonogramme, pixel, ...) en utilisant une valeur arbitraire.

4.3 Le traitement textuel :

Une machine à écrire est une machine mécanographique qui permet de produire des documents avec des caractères imprimés. Elle se présente sous la forme d'un clavier comportant une cinquantaine de touches représentant les caractères qui seront imprimés sur le papier. Lorsque l'on presse la touche correspondant au caractère, un bras positionné dans une corbeille, sur lequel le caractère est gravé, frappe le ruban d'encre puis le papier, c'est le principe de l'impression.



Figure 4-2 : une machine à écrire

Le passage de la machine à écrire manuelle au traitement de texte numérique a été une évolution majeure dans l'histoire de la production de documents. Les machines à écrire ont été largement utilisées jusqu'à l'avènement des ordinateurs personnels dans les années 1980. Les ordinateurs personnels ont permis aux utilisateurs de créer, modifier et stocker des documents numériques, ce qui a rendu la production de documents plus rapide et plus efficace.

Le passage d'une machine à écrire manuelle à une application informatique a nécessité de prendre en compte :

- des codes de contrôle non imprimables qui sont utilisés pour contrôler de nombreuses fonctions et des périphériques tels que des imprimantes.
- les caractères imprimables et représentent des lettres, des chiffres, des signes de ponctuation et des symboles divers.

La numérisation d'un texte est sa transformation en une suite de caractères prise dans une liste de caractères existant dans la langue écrite, comme une dictée transforme les sons de la parole en une suite de mots existant dans le dictionnaire de la langue écrite.

La transformation de l'écriture par l'industrie informatique a été particulièrement rapide !

Début des années 80, l'IBM PC et ses clones ainsi que le Macintosh d'Apple inondent le marché mondial. Les logiciels de traitement de textes (Lotus Notes, Word Perfect, Word, TextEdit, ..) généralisent la production des fichiers numériques qui vont progressivement éliminer tous les procédés analogiques connus de transcription, de transformation, d'échange, de stockage ou d'archivage d'un message écrit ou imprimé.

L'informatisation du traitement de textes impliquait de pouvoir faire avec un éditeur de texte ce que faisait une machine à écrire : taper un texte en caractères alphanumériques et utiliser des fonctions (effacer, copier, couper, coller, enregistrer...) propre à l'informatique. Ce sont les logiciels de traitement de textes qui ont ouvert la voie à la mise en forme des documents.

Pour convertir, afficher ou imprimer sous la forme de caractères d'imprimerie les signaux émis par les touches du clavier en unités binaires, le logiciel devait disposer d'une table de codage des caractères alphanumériques et des instructions de fonctions.

0	0	0	[NULL]	48	60	110000	0	96	140	1100000	
1	1	1	[START OF HEADING]	49	61	110001	1	97	141	1100001	a
2	2	10	[START OF TEXT]	50	62	110010	2	98	142	1100010	b
3	3	11	[END OF TEXT]	51	63	110011	3	99	143	1100011	c
4	4	100	[END OF TRANSMISSION]	52	64	110100	4	100	144	1100100	d
5	5	101	[ENQUIRY]	53	65	110101	5	101	145	1100101	e
6	6	110	[ACKNOWLEDGE]	54	66	110110	6	102	146	1100110	f
7	7	111	[BELL]	55	67	110111	7	103	147	1100111	g
8	10	1000	[BACKSPACE]	56	70	111000	8	104	150	1101000	h
9	11	1001	[HORIZONTAL TAB]	58	71	111001	9	105	151	1101001	i
10	12	1010	[LINE FEED]	57	72	111010	:	106	152	1101010	J
11	13	1011	[VERTICAL TAB]	59	73	111011	;	107	153	1101011	k
12	14	1100	[FORM FEED]	60	74	111100	<	109	154	1101100	l
13	15	1101	[CARRIAGE RETURN]	61	75	111101	=	109	155	1101101	m
14	16	1110	[SHIFT OUT]	62	76	111110	>	110	156	1101110	n
15	17	1111	[SHIFT IN]	63	77	111111	?	111	157	1101111	o
16	20	10000	[DATA LINK ESCAPE]	64	100	1000000	@	112	160	1110000	p
17	21	10001	[DEVICE CONTROL 1]	65	101	1000001	A	113	161	1110001	q
18	22	10010	[DEVICE CONTROL 2]	66	102	1000010	B	114	162	1110010	r
19	23	10011	[DEVICE CONTROL 3]	67	103	1000011	C	115	163	1110011	s
20	24	10100	[DEVICE CONTROL 4]	68	104	1000100	D	116	164	1110100	t
21	25	10101	[NEGATIVE ACKNOWLEDGE]	69	105	1000101	E	117	165	1110101	u
22	26	10110	[SYNCHRONOUS IDLE]	70	106	1000110	F	118	166	1110110	v
23	27	10111	[END OF TRANS BLOCK]	71	107	1000111	G	119	167	1110111	w
24	30	11000	[CANCEL]	72	110	1001000	H	120	170	1111000	x
25	31	11001	[END OF MEDIUM]	73	111	1001001	I	121	171	1111001	y
26	32	11010	[SUBSTITUTE]	74	112	1001010	J	122	171	1111010	z
27	33	11011	[ESCAPE]	75	113	1001011	K	123	173	1111011	{
28	34	11100	[FILE SEPARATOR]	76	114	1001100	L	124	174	1111100	!
29	35	11101	[GROUP SEPARATOR]	77	115	1001101	M	125	175	1111101	}
30	36	11110	[RECORD SEPARATOR]	78	116	1001110	N	126	176	1111110	~
31	37	11111	[UNIT SEPARATOR]	79	117	1001111	O	127	177	1111111	[DEL]
32	40	100000	[SPACE]	80	120	1010000	P				
33	41	100001	!	81	121	1010001	Q				
34	42	100010	*	82	122	1010010	R				
35	43	100011	#	83	123	1010011	S				
36	44	100100	\$	84	124	1010100	T				
37	45	100101	%	85	125	1010101	U				
38	46	100110	&	86	126	1010110	V				
39	47	100111	'	87	127	1010111	W				
40	50	101000	(	88	130	1011000	X				
41	51	101001	)	89	131	1011001	Y				
42	52	101010	.	90	132	1011010	Z				
43	53	101011	+	91	133	1011011	{				
44	54	101100	,	92	134	1011100	\				
45	55	101101	-	93	135	1011101	}				
46	56	101110	_	94	136	1011110	^				
47	57	101111	/	95	137	1011111	~				

Figure 4-3 : **ASCII** (American Standard Code for Information Interchange)

A partir de 1963, l'univers anglophone a imposé l'ASCII (American Standard Code for Information Interchange) : Code américain normalisé pour l'échange d'information), au grand dam des pays européens ayant des langues avec des caractères accentués.

L'ASCII standardise un codage de caractères sur 7 bits (permettant la représentation de la valeur de 128 signes différents). Le tableau ASCII (American Standard Code for Information Interchange) complet est illustré dans la figure ci-dessus.

Les commandes de contrôle du terminal informatique et les caractères sont listés (de 0 à 127). La table attribue à chaque élément une valeur :

Description ASCII :

- Les caractères de numéro 0 à 31 et le 127 correspondent à des commandes de contrôle du terminal informatique.
 - Le numéro 32 est l'espace, le numéro 127 est la commande pour effacer.
 - Les autres caractères sont les chiffres arabes, les lettres latines majuscules et minuscules sans accent, et des symboles de ponctuation.
 - Il n'existe pas de valeurs se terminant par 8 ou 9 !
- numéro 48, le chiffre 0, valeur 60, soit 110000 en binaire
 - numéro 65, la lettre A majuscule, valeur 101, soit 1000001 en binaire.

Exemple d'un codage de la lettre E majuscule – valeur arbitraire (ASCII) 105 - binaire : 011001 :

2 ⁷	2 ⁶	2 ⁵	2 ⁴	2 ³	2 ²	2 ¹	2 ⁰	
128	64	32	16	8	4	2	(1)	
X	X	X	X	X	X	X	X	
0	1	1	0	1	0	0	1	
=	=	=	=	=	=	=	=	
0 +	64 +	32 +	0 +	8 +	0 +	0 +	1	= 105
Valeur E (ASCII) : 105 (3 positions) - codage binaire : 01101001 (8 positions)								

Figure 4-4 : Exemple d'une table de codage de la lettre E – valeur 105

Utiliser un traitement de texte c'est avoir la possibilité de :

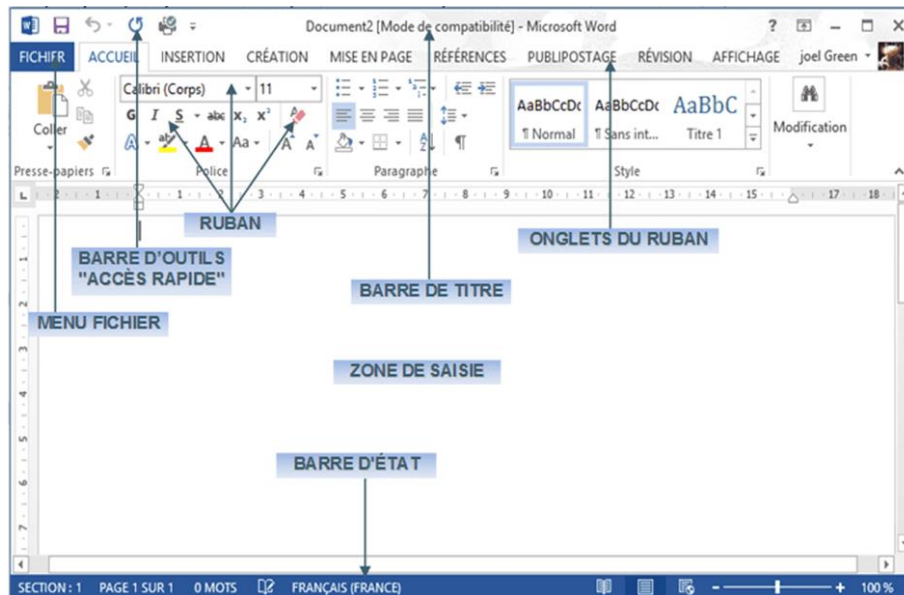


Figure 4-5 : barre des outils d'un traitement de textes

- Modifier, déplacer, supprimer une partie du texte.
- Insérer un bloc de texte, des notes de bas de page, des notes de fin, ...
- Modifier la police, la taille et le style des caractères, l'alignement d'un paragraphe, les marges, le nombre de colonnes. Ajouter des bordures.
- Insérer des sauts de page, de colonne, des sauts de section.
- Créer un tableau.
- Modifier les marques de tabulation.
- Corriger l'orthographe et la grammaire (avec plus ou moins de réussite).
- Intégrer une image, un tableau,ajouter des sous-titres.
- Générer automatiquement une table des matières, un glossaire.
- Modifier l'orientation de la page.
- Utiliser des feuilles de style qui proposent des mises en page types.
- Réaliser de façon automatique un enchaînement de tâches (les "macros").

Si le texte est altéré et contient des points d'interrogation ainsi que des carrés lors de l'ouverture d'un fichier dans un traitement de texte, il est possible que le logiciel n'ait pas correctement détecté la norme de codage du fichier.

4.4 Normalisation du traitement de textes : UNICODE

Comme il n'était pas possible d'introduire les caractères accentués dans un format 7 bits, chaque langue a produit un standard d'encodage sur 8 bits.

En 1998, l'UNICODE a permis d'encoder les plus importantes langues du monde. Cette approche généraliste implique aussi toutes les règles d'interopérabilité qui sont exprimées par la norme UNICODE.

Totalement compatible avec le jeu universel de caractères (JUC) de l'ISO/CEI 10646, la norme UNICODE l'étend en lui ajoutant un modèle complet de représentation et de traitement de textes, en conférant à chaque caractère un jeu de propriétés normalisées ou informatives, en décrivant avec précision les relations sémantiques qui peuvent exister entre plusieurs caractères successifs d'un texte, et en normalisant des algorithmes de traitement qui préservent au maximum la sémantique des textes transformés.

Unicode a pour objet de rendre un même texte utilisable à l'identique sur des systèmes informatiques totalement différents.

La généralisation de l'Internet a conduit à une norme de représentation universelle compatible avec l'ASCII et l'ISO8859 appelée UTF (déclinée en UTF-8 ; UTF-16 et UTF-32).

Non seulement les caractères 'occidentaux' sont représentables mais aussi les caractères arabes, sanscrit, cyrillique, les idéogrammes (chinois, japonais, ...).

5 La définition de la donnée :

Dans le langage courant, une donnée correspond à quelque chose qui est connu, et déterminé, dans l'énoncé d'un problème. Il est également utilisé pour un élément qui sert de base au raisonnement, de point de départ à la recherche.

L'informatique a généralisé ce terme pour qualifier la représentation de toutes choses: d'une réalité, de concepts, du résultat d'une comparaison, d'une observation ou d'une mesure.

Voici la définition que nous utiliserons dans cette présentation/document :

[Data] [Donnée] :

Une représentation formalisée d'informations (faits, concepts ou instructions) exprimée selon un formalisme adapté à la communication, à la mémorisation, à l'interprétation ou au traitement par des êtres humains ou par des moyens automatiques.

Les données sont de natures très différentes suivant leur source et elles sont évidemment étroitement liées à l'utilisateur qui les crée. C'est ce dernier qui, au travers du couple signifiant/signifié, décide en final de la qualification. C'est une étape cruciale dans le développement de tous les outils qui vont traiter et mettre en forme des enregistrements sous de multiples formats.

La donnée brute est dépourvue de toute capacité de raisonnement, de constatation ou de projection statistique. Si elle sert aujourd'hui de base à de nombreuses recherches, c'est dans un premier temps sur sa base syntaxique que la recherche se fait et compte tenu du développement au niveau mondial d'outils capables de l'interpréter sémantiquement dans un univers numérique, elle est au cœur des SEO (Search Engine Optimization), des outils qui permettent d'améliorer le traitement de l'ensemble des pages des sites web en analysant la thématique et le contenu pour alimenter des robots d'indexation.

Les données doivent répondre aux interrogations tant Humaines que Machines aux questions « qui, quoi, où, pourquoi, quand » sur la base d'un modèle de données. Elles sont nodales, elles doivent permettre tant en interne (des documentalistes, des équipes IT, des outils business, ...) qu'en externe (des utilisateurs) de comprendre, de travailler et d'échanger des données pertinentes et de qualité. Elles servent à créer au-dessus des jeux de données des taxonomies qui seront indexées et naviguées par un moteur de recherche ou de navigation.

La gestion des données fait partie intégrante d'une stratégie de gouvernance des données « agiles » (gérer le cycle de vie des données, garantir des règles permettant la bonne utilisation des données et ainsi maximiser la création de valeur des données) d'une entreprise ou d'une institution :

- pour les auteurs/producteurs/éditeurs : elles assurent un bon référencement des documents,
- pour les distributeurs/diffuseurs : elles permettent "l'échange" **de tables de données communes** entre plateformes,
- pour les utilisateurs : elles permettent des recherche/navigation "performante".

Maintenir un répertoire de données à jour assure les multiples acteurs d'exploiter des données fiables et pertinentes pour leurs cas d'usage.

Types de données :

- **Les données descriptives** : fournissent des données à propos d'un document (un livre et son contenu : le titre de l'ouvrage, l'ISBN, le nom de l'auteur, le nom de l'éditeur, la langue utilisée, un extrait du livre, les mentions de licence, etc...) ;
- **Les données techniques** : décrivent le format d'un fichier, la structure d'un jeu de données et les informations liées au stockage ;
- **Les données opérationnelles** : décrivent le cycle de vie d'un document : date de création, analyse statistique de la donnée, date de mise à jour, provenance chaînage, volume, cardinalité, identifiant des traitements ayant créé ou transformé la donnée, statuts des traitements sur la donnée, etc...
- **Les données business** : servent à décrire un contexte métier : des descriptions (contexte et usage), les propriétaires et référents, des tags et propriétés, ...
- **Les métadonnées de contrôle** : niveau de confidentialité des données, type d'accès, ...
- (.....)

Il existe de nombreux modèles de données (format table – flat models) qui peuvent être décrits selon plusieurs standards : syntaxe " meta " HTML et Dublin Core, XML, DTD EAD (Encoding Archival Description), TEI (Text Encoding Initiative), etc.

6 DATAVISION : la modélisation de l'INFORMATION par les DONNÉES :

Cette représentation est organisée en couches. Les explications détaillées de chaque couche sont sommairement présentées ci-après et le seront avec plus de détail dans les schémas suivants :

M. Niveau «**Matière**» : correspond à l'aspect matériel capable d'être porteur de signifiant et son support

R. Niveau «**Représentation**» correspond à la représentation de l'Information par des Données :

- Niveau R0 : « **Symbolique** » : la capacité de produire une re-présentation signifiante de quelque chose et à communiquer ce sens.
- Niveau R1 : « **élémentaire** » : les « Données élémentaires » couvrent le niveau sémantique racine. Un seul type de données est concerné à la fois.
- Niveau R2 « **cellulaire** » : les « Données élémentaires » couvrent le niveau sémantique racine. Plusieurs types de données sont concernées à la fois.
- Niveau R3 : « **documentaire** » les « Données documentaires » sont les données exprimées comme des tous sémantiquement cohérents.

P. Niveau «**Patrimoine documentaire**» comprend des documents uniques – ou groupes de documents – d'une valeur significative et durable pour une communauté, un pays ou pour l'humanité en général, et dont la détérioration ou la perte serait un appauvrissement préjudiciable.

Enfin la nécessité d'implanter une couche d'interopérabilité : Meta R (information de Représentation) et Meta S (Information Sémantique) afin d'assurer l'interopérabilité entre les couches !

- Niveau Matière

Un **document** est le résultat d'une action matérialisé par une trace figée, stable dont il garantit l'authenticité et la lisibilité. Cette trace doit disposer d'une composante suffisamment stable pour pouvoir manifester cette action et être rendue sensible, perceptible par nos sens. Cela implique un support pérenne et tangible pour un contenu défini.



Figure 6-1 : Le niveau matière : la pierre, le parchemin, les livres, les fichiers, ...

Avec le temps, les humains ont façonné des outils complémentaires au texte en vue de figer un espace (la photographie comme évolution de la peinture), d'enregistrer et de reproduire le mouvement (le film puis la vidéo), ainsi que des univers sonores, aujourd'hui rassemblés sous le vocable de médias. Un document est porteur de signes, de symboles... de conventions sémiotiques, linguistiques... exprimant l'intention (les tablettes cunéiformes, les hiéroglyphes...).

La relation entre le contenu et le contenant peut présenter d'importantes qualités, esthétiques, culturelles ou techniques.

Le contenu de ce document est à l'attention d'un public indéterminé. La signification peut découler de l'interprétation particulière de l'humain percevant le document.

- Dans un univers physique, le document est directement perceptible : un livre, une photo ou encore, une peinture faite sur une paroi de roche (pariétale).
- Un univers numérique requiert l'aide d'une application pour, par exemple, rendre perceptible un fichier codé en MP3, en MPEG4, en HEVC, un site web.

- Niveau Représentation :

Le Niveau « Représentation » comprend 4 sous- niveaux :

- Niveau Symbolique



Un idéogramme est un symbole graphique (signe) représentant une idée ou un mot utilisé dans certaines langues vivantes (comme le chinois et le japonais) ou dans des langues anciennes (comme les hiéroglyphes de l'Égypte ancienne).

Les chiffres sont des idéogrammes.

Un phonogramme est un caractère écrit qui est la transcription arbitraire d'un son. Le phonogramme n'a aucun rapport avec le sens : il est utilisé pour transcrire un son quelle que soit la signification à laquelle ce son reviendra.

Un pictogramme est une représentation graphique, un dessin figuratif stylisé ayant une fonction de signe. Utilisé dans l'art rupestre (dessins peints), il est généralement utilisé pour la signalisation, c'est une alternative à la signalisation multilingue, il réduit la quantité d'informations sur un panneau d'affichage.

En mathématiques, de nombreux symboles sont employés avec une signification qui n'est pas toujours précisée dans les documents qui les emploient.

Un chiffre¹ est un signe d'écriture utilisé seul ou en combinaison pour représenter des nombres entiers auxquels on associe une valeur numérique.

Latin / Arabe occidentale	0	1	2	3	4	5	6	7	8	9	10
Hiéroglyphes égyptiens		I	II	III	IIII	IIII	IIII	IIII	IIII	IIII	∩
Latin romain		I	II	III	IV	V	VI	VII	VIII	IX	X
Sinogrammes		〇	一	二	三	四	五	六	七	八	九
Hébreu		א	ב	ג	ד	ה	ו	ז	ח	ט	י
Araméen		Ⲁ	ⲁ	Ⲃ	ⲃ	Ⲅ	ⲅ	Ⲇ	ⲇ	Ⲉ	ⲉ
Grec		Αα	Ββ	Γγ	Δδ	Εε	Ζζ	Ηη	Θθ	Ιι	
Devanagari		०	१	२	३	४	५	६	७	८	९
Brahmi		𑀧	𑀨	𑀩	𑀪	𑀫	𑀬	𑀭	𑀮	𑀯	𑀰
Abjad		א	ב	ג	ד	ה	ו	ז	ח	ט	
Arabe orientale		٠	١	٢	٣	٤	٥	٦	٧	٨	٩

Les chiffres constituent, une primitive de représentation, une spécification à caractère formel et explicite d'une conceptualisation partagée d'un ensemble de connaissances :

Le **système unaire** utilise un seul signe, sous la forme d'un simple bâton, représentant la valeur 1, qui est répétée pour exprimer tous les nombres naturels.

La **numérotation romaine** utilise un code avec 7 symboles numériques (I, V, X, L, C, D, M, de 1 à Mille), avec table de codage : la valeur d'un nombre étant la somme des valeurs des symboles qui le composent sauf pour les symboles précédant un symbole de valeur supérieure qui sont au contraire soustraits.

¹ <https://fr.wikipedia.org/wiki/Chiffre>

Le système décimal, le plus courant des systèmes de numération, utilise dix chiffres (de zéro à neuf) qui permettent de gérer les quatre opérations arithmétiques fondamentales (addition, soustraction, multiplication et division).

L'alphabet latin est un alphabet bicaméral comportant vingt-six lettres de base, principalement utilisé pour écrire les langues d'Europe de l'Ouest, d'Europe du Nord et d'Europe centrale, ainsi que les langues de nombreux pays qui ont été exposés à une forte influence européenne, notamment à travers la colonisation européenne des Amériques, de l'Afrique et de l'Océanie.

	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	{		}	~	

Jeu de caractères du « Latin de base »

La norme ISO 8859-1 (dont le nom complet est ISO/CEI 8859-1, ou Latin-1) définit ce qu'elle appelle l'alphabet latin numéro 1, qui consiste en 191 caractères de l'alphabet latin (avec minuscules, majuscules et accents), chacun d'entre eux étant codé sur un octet (soit 8 bits).

ISO 8859-1 reprend le codage des caractères imprimables d'US-ASCII.

■ Niveau Élémentaire

Les « Données élémentaires » couvrent le niveau sémantique racine. Un seul type de données est concerné à la fois.

Exemples:

- **Nombres entiers** 49 012 Codage hexadécimal de -99 999 999 999 à +99 999 999 999
- **Valeur booléenne** 1 Codage 1 = « vrai » Codage 0 => « faux »
- **Chaine de Caractères** Guy – Noël Codés suivant la norme ISO « UNICODE »
- **Pixel** 9F-14-B3-F0 Codage hexadécimal de R G B L avec 8 bits par critère
- **PCM-S** 45FE187B5A91 Codage hexadécimal à 48 bits par échantillon
- **Année** 2023 Codage de l'année suivant la norme ISO « 8601 »
- **Mois** 07 Codage du mois ('juillet') suivant la norme ISO « 8601 »
- **Hyperlien** <https://www.google.com/> Codage suivant la norme W3C

■ Niveau Cellulaire

Les « Données cellulaires » couvrent le niveau sémantique des Entités et Relations avec leurs Propriétés. Plusieurs types de données sont concernées à la fois.

Exemples :

Date 2023-07-11 Codage suivant la norme ISO « 8601 » [format AAAA-MM-JJ]

Dans l'exemple, le 11 juillet 2023

Emplacement sur terre :

50,8166144 Codage GPS (en décimal de la latitude)

4,3876352 Codage GPS (en décimal de la longitude)

N50°48'59,812" Codage ISO (en sexagésimal de la latitude 'N' => 'Nord')

E 4°23'15,486" Codage ISO (en sexagésimal de la longitude 'E' => 'Est')

Adresse :

Nom de rue	Rue des Hellènes	Codage « UNICODE » [en français]
Numéro de maison	N°14	Codage « UNICODE »
Code Postal	1050	Codage « UNICODE »
Commune	Ixelles	Codage « UNICODE » [en français]

Phrase : Wie geht es dir Roger ? Codage « UNICODE » [en allemand]

...

■ Niveau Documentaire

Les « Données documentaires » sont les données exprimées comme des « TOUS » sémantiquement cohérents.

Il faut d'abord évoquer les « Données documentaires » exprimées au sujet de livres et autres supports physiques. Ce sont elles qui seront scannées, passées en OCR ... pour constituer des « Données documentaires » représentées comme des fichiers au format PDF.

Il faut ensuite évoquer les « Données documentaires » exprimées dans des représentations informatiques. Les plus courantes s'expriment comme des fichiers, des bases de données ou des sites Web.

Dans le cadre des fichiers liés à la préservation et à l'accès, divers formats ont actuellement une place particulière:

- Formats révisables : .odt // .docx ainsi que les suites d'outils associés : Open Office // Office 365
- Formats image : .jpg // .raw
- Formats son : .mp3 // .wav // flac
- Format vidéo : .mp4
- Format de transfert : .pdf/A
- Formats sémantiques : .owl // .rdf

○ Niveau Patrimoine Documentaire :

Le niveau « Patrimoines documentaires » comprend ces documents uniques – ou groupes de documents – d'une valeur significative et durable pour une communauté, un pays ou pour l'humanité en général, et dont la détérioration ou la perte serait un appauvrissement préjudiciable.

Le niveau « Patrimoine documentaire » tente de dépasser la simple représentation de l'information pour couvrir les éléments sociétaux, philosophiques, éthiques ... associés aux documents.

Le niveau « Patrimoine documentaire » couvre les rassemblements de documents et leur contextes documentaires.

Un « Patrimoine documentaire » comprend ces documents uniques - ou groupes de documents - et les artefacts associés d'une valeur significative et durable pour un contexte désigné (une communauté, un pays ou pour l'humanité en général), et dont la détérioration ou la perte serait un appauvrissement nuisible !

Le « Patrimoine documentaire » nécessite :

- des Métadonnées Structurelles et Sémantiques

- des Métadonnées de lignée (dans le temps / l'espace / culturel / historique)

- Exemple : Patrimoine documentaire : les archives Belgavox - Congovox :

La Société belge des actualités Belgavox, est créée en 1937 par Georges Fannoy. Tout au long de ces années la direction de Belgavox a œuvré à la sauvegarde de ces actualités, un patrimoine inestimable (archives films 35 m/m).

En 1955 lors du 1er voyage Royal du roi Baudouin que Pierre Fannoy se prend d'amour pour le pays. En 1957 il s'envole une nouvelle fois pour le Congo pour faire mieux découvrir aux belges tous les trésors du pays. Dès lors, l'actualité congolaise fera partie intégrante du journal d'actualité quotidien.

Belgavox couvre tous les grands événements du Congo puis du Zaïre.

1964 : création de "Congovox" avec ses propres journalistes, cameramen et techniciens. Les sujets étaient traités par les Congolais avec une vision congolaise. Les liens étroits avec Belgavox permettaient des échanges avec des correspondants du monde entier, permettant au Congo de recevoir les informations du monde entier, mais aussi de diffuser l'image du pays partout dans le monde.

Une collection est à la fois un regroupement d'objets en fonction d'une intention (un thème, une valeur documentaire, esthétique, le prix, la rareté, ...) et l'activité qui consiste à acheter, stocker et gérer l'accès à ces objets :

- 1955 – 1964 : Belgavox : toutes les séquences réalisées avant 1964
- 1964 – 1974 : toutes les séquences produites par Congovox entre 1964 et 1974
- 1974 – 1994 : Belgavox : toutes les séquences « Congo » produites par Belgavox

Il existe donc un fonds d'actualités 'Belgavox-Congovox' : une collection de reportages sur la Belgique et le Congo qui a une signification historique pour les deux pays et l'histoire de la décolonisation" dans de multiples journaux d'actualités.

Du fait des échanges avec des correspondants du monde entier, permettant de recevoir les actualités du monde entier, il existe une forte diversité des sujets.

Il n'y a donc pas d'homogénéités dans les différents supports physiques archivés dans la filmothèque (mixité des sujets diffusés dans une actualité) et la collection qui exige un tri).

Une base de données relationnelle peut décrire la structure physique des « Actualités » par semaine. En revanche pour visualiser des collections (entités de données, des attributs, des clés et des relations,) il faut disposer de modèles de données logiques (décrits sémantiquement).

Le développement d'un modèle « DATAVISION » illustre à suffisance l'importance d'une normalisation entre les différents niveaux (matière, représentation, patrimoine documentaire) via des couches Meta Représentation et Meta Sémantique afin de garantir une pérennité de l'exploitation des données.

7 L'analyse des différents niveaux du modèle DATAVISION :

- M. Niveau « Matière » la capacité d'être porteur de signifiants et son support



Figure 7-1 : DATAVISION : le niveau « matière »

L'organisation des surfaces graphiques sur un support est un objet culturel inscrit dans une histoire des savoirs : les formes progressivement décidées et adoptées (idéogramme, phonogramme, icône, ...) témoignent des solutions graphiques apportées à des problèmes de représentations d'un ordre du monde (Christin : 2001).²

- Niveau représentation :
 - Le niveau Symbolique :

² <https://journals.openedition.org/pratiques/1522>

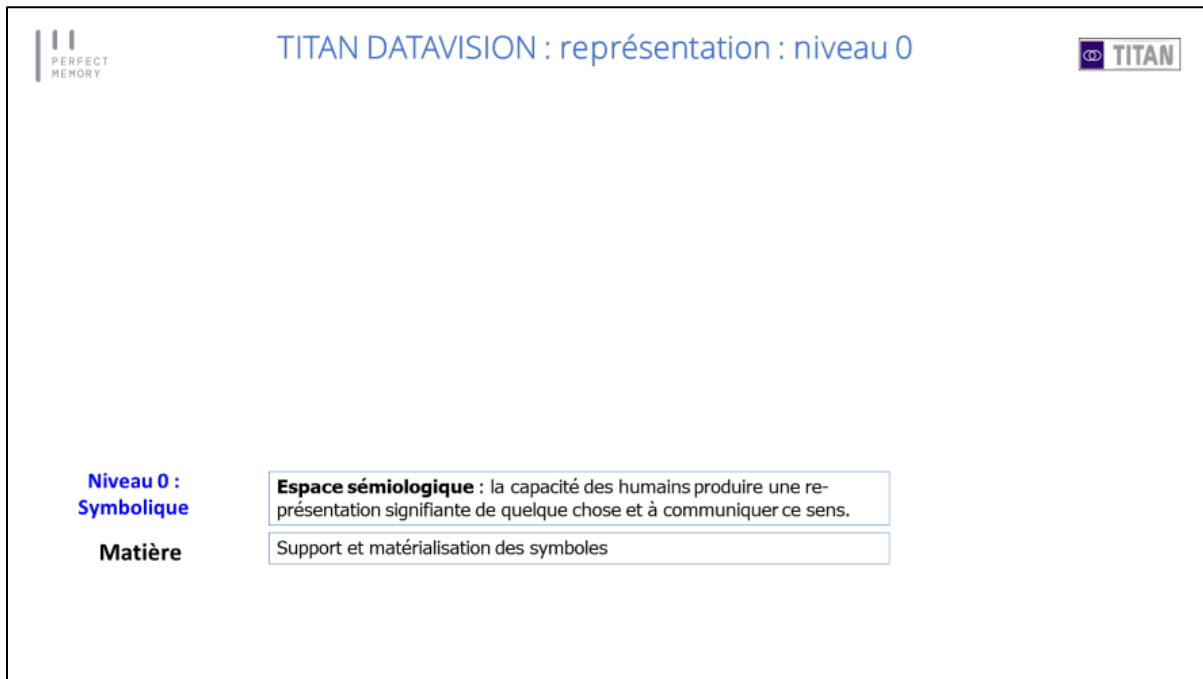


Figure 7-2 : DATAVISION : niveau symbolique

- **[Symbole]** : au niveau d'une langue c'est la syntaxe qui normalise les chaînes de caractères.
- **[Concept]** : au niveau d'une langue c'est la grammaire qui normalise les genres, le singulier/pluriel, l'association mode (indicatif, subjonctif, conditionnel, ..) avec comme temps possible : le présent, le passé-simple, l'imparfait, le passé-composé, Ce sont des **[Connaissances]** : concepts ou instructions qui servent de fabriquer des données textuelles, des énoncés mathématique ou chimique, ..

■ Niveau 1 : Élémentaire

Le numérique est basé sur la numérisation des symboles (caractères, son, image, vidéo, ...) c'est-à-dire transformés en une succession de 0 et de 1 qu'on appelle des bits. On rassemble ces bits sous la forme d'octets, c'est-à-dire 8 bits.

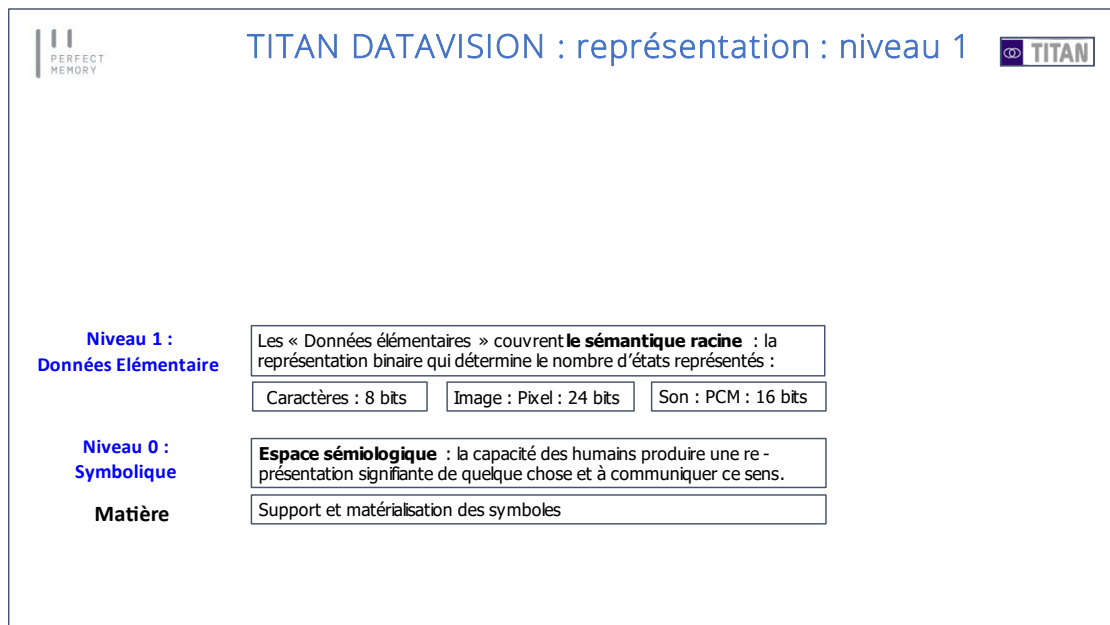


Figure 7-3 : DATAVISION : le niveau élémentaire

Les « Données élémentaires » couvrent le niveau sémantique racine. Un seul type de données est concerné à la fois.

- *Illustration de la quantification d'un signal analogique en PCM :*

Le PCM [Pulse (qui traite des impulsions électriques de transmission) Code Modulation) représente de façon numérique un signal analogique.

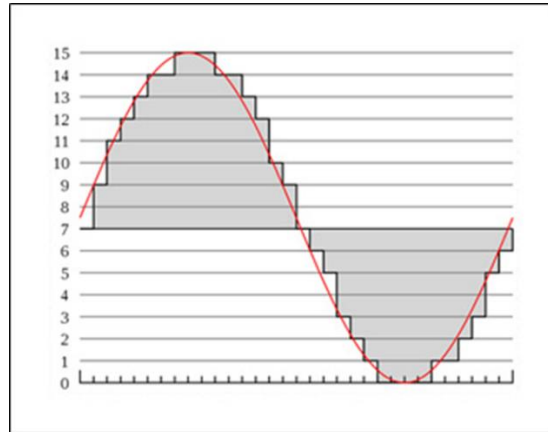


Figure 7-4 : l'échantillonnage et la quantification audio

A intervalles réguliers, la hauteur d'un signal analogique audio est mesurée (échantillonnage et quantification) et transformée d'un signal continu à un signal discret qui se traduit par un ensemble de valeurs fondées sur un point dans le temps.

Ces valeurs sont ensuite codées, généralement en binaire.

Si le nombre affiché par la carte à microcontrôleur du convertisseur analogique-numérique varie entre 0 et 1023 (ce qui correspond à un nombre total de 1024 possibilités) la résolution est donc calculée sur 10 bits car 2^{10} valeurs correspondent à 1024.

Le CD-Audio code le PCM avec 16 bits par échantillon et 44,1 kHz de fréquence. C'est insuffisant pour les besoins professionnels.

L'EBU (Tech 3285) a donc développé le format .bwf . En studio, on utilise des micros et des convertisseurs « analogique – digital » qui ont une dynamique (rapport signal/bruit de fond) de 117 dbA, enregistré avec 24 bits à 48 kHz ou 96 kHz ce qui est suffisant compte tenu des capacités limitées de l'oreille humaine.

L'EBU a également établi une norme de procédure d'établissement de la référence en audio : en effet, l'oreille humaine entend sur une dynamique de 110 dbA (qui est une échelle logarithmique de puissance) mais, comme l'oreille humaine sait régler le niveau minimal dynamiquement, une dynamique de 70 dbA suffit.

En studio son, on travaille en deux fois, de manière similaire à ce qui se fait en image en .RAW vers .jpg ou vers .png ...

- *La transformation "image": l'échantillonnage – la quantification et la compression numérique :*

Les systèmes de télévision en noir et blanc utilisaient uniquement l'information luminance (Y). Les informations de couleur (U et V) ont été ajoutées séparément via une sous-porteuse afin qu'un récepteur noir et blanc puisse toujours recevoir et afficher une transmission d'image en couleur dans le format noir et blanc natif du récepteur.

YUV code une image couleur ou une vidéo (RVB : Rouge Vert Bleu) prenant en compte la perception humaine. Une réduction de la bande passante pour les composants de chrominance, permet ainsi de masquer plus efficacement les erreurs de transmission ou les artefacts de compression par la perception humaine.

D'autres codages de couleur ont des propriétés similaires, et la raison principale pour implémenter ou étudier les propriétés de Y'UV serait pour l'interfaçage avec une télévision analogique ou numérique ou un équipement photographique qui est conforme à certaines normes Y'UV.

La représentation UV de la chrominance a été choisie sur les signaux R et B droits car U et V sont des signaux de différence de couleur. En d'autres termes, les signaux U et V indiquent :

- au téléviseur de décaler la couleur d'un certain pixel sans modifier sa luminosité
- au moniteur de rendre une couleur plus brillante au détriment de l'autre et de combien elle devrait être décalée.

Plus les valeurs U et V sont élevées (ou plus basses quand négatives), plus le pixel est saturé (coloré). Plus les valeurs U et V se rapprochent de zéro, moins la couleur change de couleur, ce qui signifie que les lumières rouge, verte et bleue seront plus brillantes, produisant un pixel plus gris.

L'avantage d'utiliser des signaux de différence de couleur, cela indique combien le pixel est plus rouge que le vert ou le bleu. Cela signifie aussi que lorsque les signaux U et V sont nuls ou absents, cela affiche simplement une image en niveaux de gris.

Dans le cas de l'usage d'un signal RVB, les pixels auraient des valeurs non nulles même dans une scène N & B, nécessitant les trois signaux porteurs de données.

L'échantillonnage permet de découper le signal en échantillons. Le sous-échantillonnage de la chrominance (R-Y et B-Y) permet de réduire le volume d'informations liées à la couleur pour laisser plus de place aux données liées à la luminance (Y) :

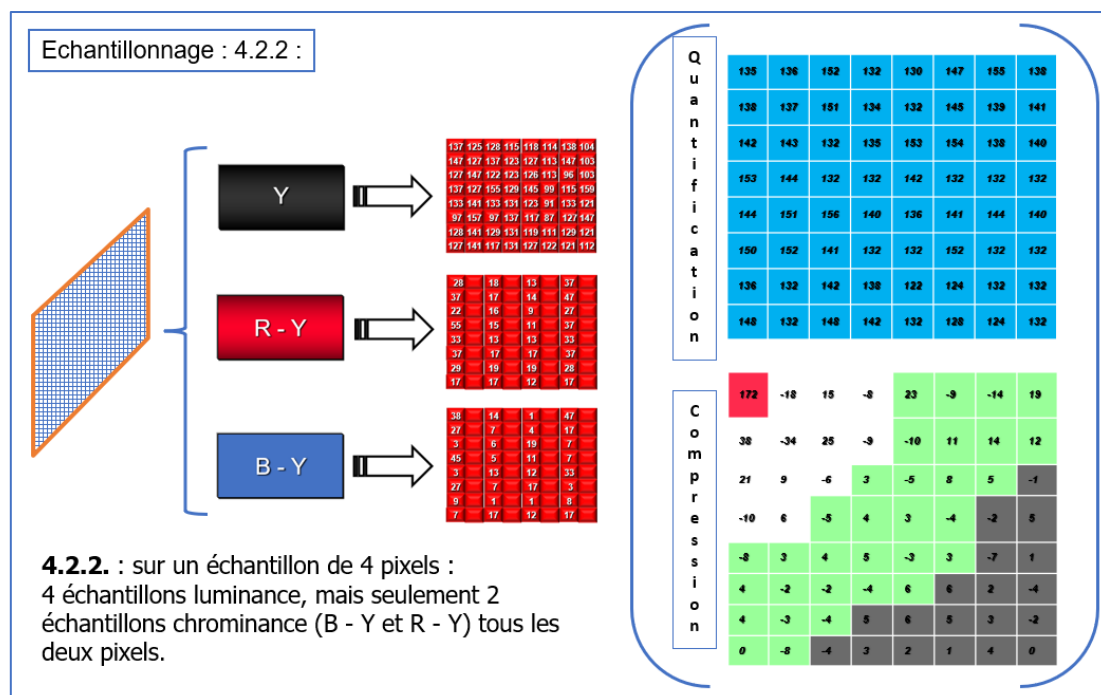


Figure 7-5 : l'échantillonnage – la quantification et la compression numérique TV

Le signal est décomposé en macro-blocs de 8 x 8 pixels pour chaque couleur (RVB - YUV).

L'échantillonnage permet de sélectionner les pixels (tous pour la luminance, 1 sur 2 pour la chrominance) avec une quantification (256 niveaux – soit 28 bits).

DCT (Discrete Cosine Transform) : Transformation d'une représentation spatiale de 64 pixels à une représentation fréquentielle avec une composante continue donnant le niveau moyen du bloc et permettant de traiter les quantifications.

- Niveau 2 : Cellulaire

Les « Données cellulaires » couvrent le niveau sémantique opérationnel racine. Plusieurs types de données peuvent être concernés et enchaînés :

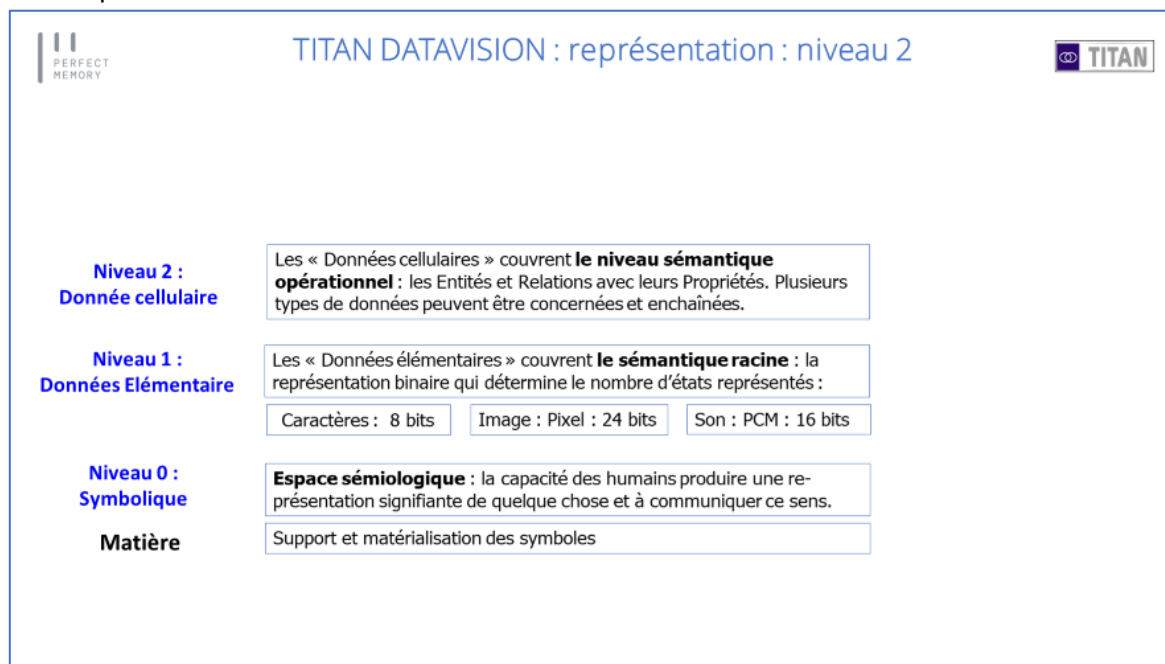


Figure 7-6 : DATAVISION : niveau cellulaire

- Niveau 3 : « documentaires »

Les « Données documentaires » sont les données exprimées comme des tous sémantiquement cohérents.

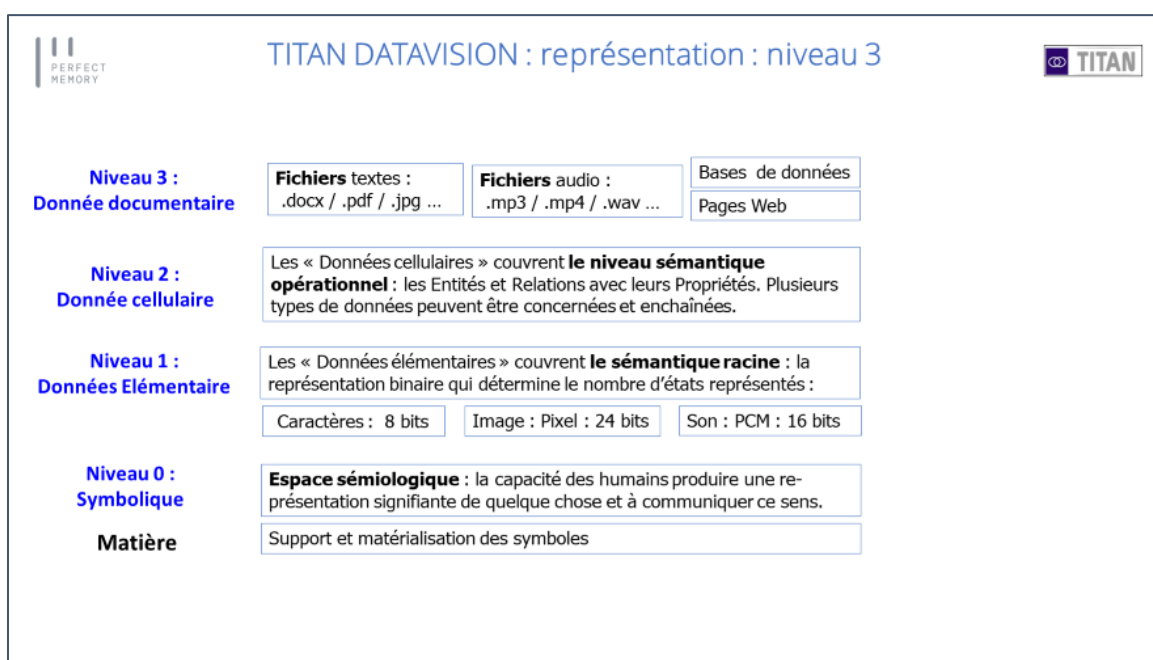


Figure 7-7 : DATAVISION : le niveau documentaire

- l'interopérabilité entre les niveaux et intra-niveau :

Historiquement, les bases de données contenaient des données alphanumériques. Aujourd'hui de la diversité des applications, de nouveaux types de données sont à stocker dans les bases de données, tel que ancillaires, structurées, semi-structurées, non structurées, ...

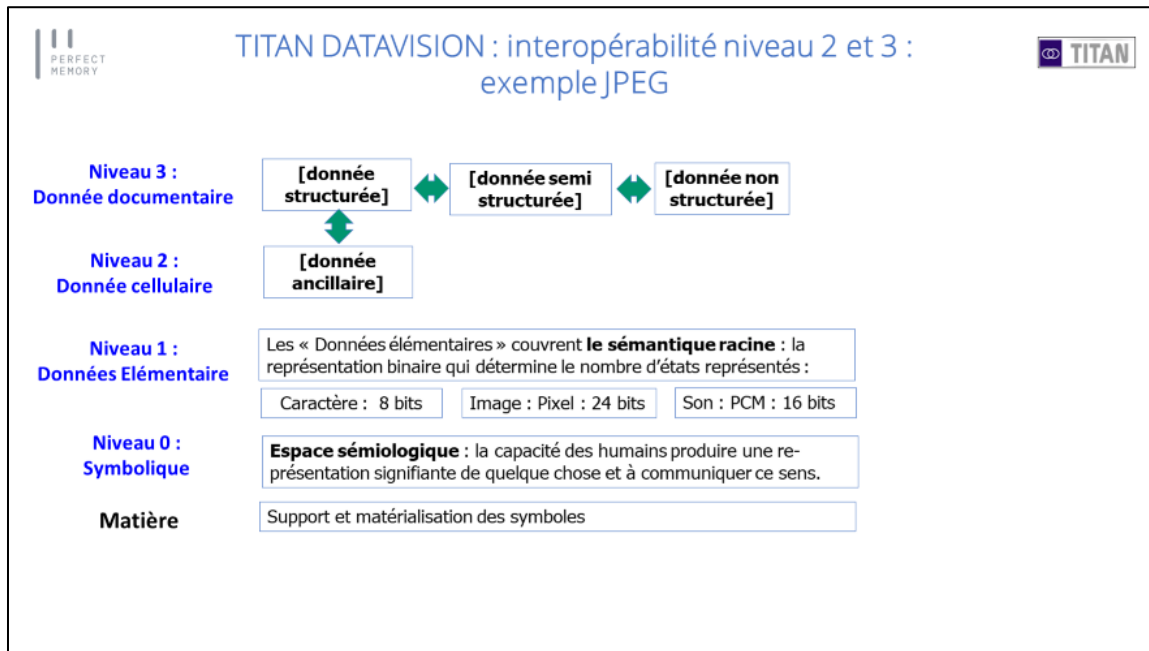


Figure 7-8 : interopérabilité entre niveau 2 et 3 : exemple JPEG

La sémantique joue le rôle d'une cartographie, elle formalise les liens entre les données/informations/connaissances. L'enjeu théorique d'intégration du numérique « concerne la possibilité d'envisager une conception du sens fondée sur la médiation technique », plus précisément de « comprendre comment du sens peut être produit par l'artefact informatique » (Bruno Bachimont, 1996 : 3-4.)

Les défis sont de savoir comment gérer ces données ? Comment les corréler entre elles ? Comment les associer pour les représenter ? Comment les traiter pour pouvoir offrir une navigation ?

Les principaux facteurs de choix d'un logiciel sont la conformité aux besoins, la fiabilité, l'ergonomie (dont la facilité d'emploi), la flexibilité, la maintenabilité, l'intégrité et la disponibilité ... mais pas la conformité à une norme ou un standard, la compatibilité avec d'autres logiciels, ... La majorité des applications sont des silos assurant une fonctionnalité interne (assurant la fonction de représentation dans l'interface utilisateur IHM).

Les défis sont multiples : comment gérer ces données générées de façon propriétaire et sémantiquement « hétérogènes » ? Comment les corréler entre elles ? Comment les associer ? Comment les échanger ?

Le réseau Internet, en tant que « connecteur », constitue un élément fédérateur en termes de normalisation universelle ! L'Internet repose sur de nombreuses technologies normalisées (protocole HTTP, IP ou le DNS). L'Internet connecte des équipements très variés, issus d'équipes différentes, et utilisant des logiciels développés de manière indépendante.

Deux niveaux : l'interopérabilité et l'intelligibilité des données :

- ✓ Pour assurer l'interopérabilité, c'est-à-dire le fait que ces équipements différents puissent communiquer, on utilise la notion de protocole : un protocole est un ensemble de règles que les machines doivent suivre afin de se comprendre. En général, le protocole est décrit par un texte (la spécification), et cette spécification est parfois formalisée dans une norme, lorsqu'elle a été développée et approuvée par un organisme de normalisation (W3C ou ISO).
- ✓ L'intelligibilité est associée à quatre principaux critères : le critère de lisibilité, de cohésion, de cohérence et de pertinence.

[donnée sémantique] : donnée dont le formalisme utilisé dans leur représentation recouvre une définition du sens (jusqu'au niveau de détail choisi) attribué aux données et que ce formalisme est lui-même représenté par des données, via un langage informatique formel utilisant des primitives spécifiques à la logique et aux mathématiques.

L'usage de la sémantique introduit une interopérabilité et une intelligibilité des données ! Dans un univers interconnecté, il est nodal de pouvoir parcourir dans tous les sens le schéma ci-dessus ! L'univers des données est hétéroclite et hétérogène ...

- Niveau 4 : Patrimoine documentaire :

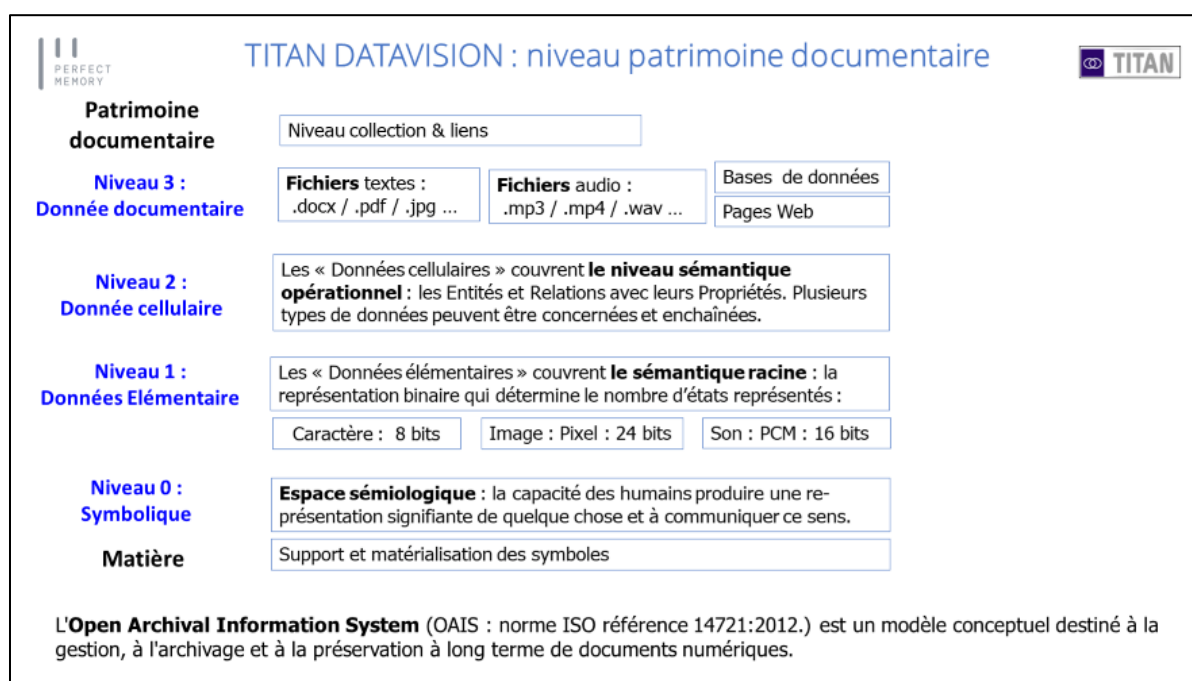


Figure 7-9 : DATAVISION : niveau Patrimoine Documentaire

8 DATAVISION : les couches Meta 1 (Représentation) et Meta 2 (Sémantique) :

Représentation de l'Information (Formats / Codages / Protocoles / Sémantique / ...) :

Les couches Meta R et Meta S « Métadonnées » expriment les contextes et les références d'interprétation de « données »

Toute « Donnée » s'entend en référence aux « Données » nécessaires à définir le formalisme utilisé pour représenter l'Information (Formats / Codages / Protocoles / Sémantique ...).

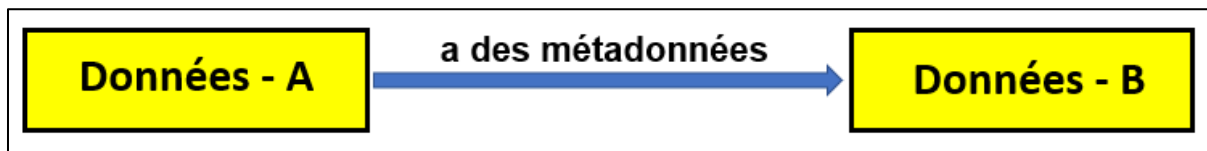


Figure 8-1 : La représentation des DONNÉES par des Metadonnées

Ce formalisme est illustré par la figure ci-dessus, sous le nom de « Donnée – B », qui sont les métadonnées de référence des « Données – A » ! Plus simplement dit, les « Données-B » sont les métadonnées des Données-A.

Ce formalisme :

- est la plupart du temps implicite ou renvoie à une norme technique ;
- peut-être explicite et lisible par les machines : c'est particulièrement le cas si les données sont exprimées dans des langages formels sémantiques, tel que OWL ;
- est dans certains cas tels que des Métadonnées (dites ancillaires) sont fusionnées dans les Données : par exemple, dans les fichiers JPEG le formalisme prévoit de pouvoir y intégrer la date de prise de vue, la position GPS, l'ouverture de l'optique, la vitesse de l'obturateur, la focale, la sensibilité ASA et autres. En revanche, indiquer le nom des personnes que l'on voit sur la photo et le lien vers leurs positions sur la photo ; ou indiquer le nom de l'évènement [Mariage de Julie et Marc] ... seront exprimés de manière externe au fichier JPEG.

- Représentation de l'Information Meta R et Meta S :

Les « Métadonnées » expriment les contextes et le cadre d'interprétation de « données ».

Toute « Donnée » s'entend en référence aux « Données » nécessaires à définir le formalisme utilisé pour représenter l'Information (Meta R : Formats / Codages / Protocoles) et Sémantique (Meta S).

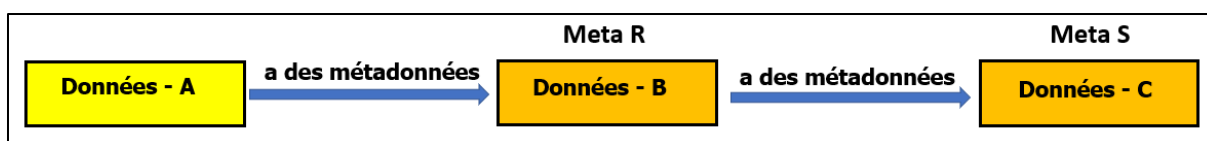


Figure 8-2 : La représentation des DONNÉES par des Metadonnées R et S

Ce formalisme est illustré par la figure ci-dessus, sous le nom de « Donnée – B », qui sont les métadonnées de représentation des « Données – A » et de « données C » qui sont l'expression de la signification des « données B ».

C'est dans le cadre de la description de ressources sur Internet dans les années 1990 que le terme métadonnée (metadata) est apparu et s'est ensuite généralisé sous une définition floue (une donnée sur une donnée.) Il a fallu bien longtemps avant que la définition suivante soit acceptée :

[Métadonnée] : à partir du préfixe grec méta et des données latines "informations" : c'est une donnée pour laquelle la méta couche procure un moyen de représenter le « signifiant » et le « signifié » des données pour une interprétation par la Machine. Ce Tutoriel utilise également le terme de données sémantiques pour les distinguer des données « tables/plates ».

- **Meta R** : Information de Représentation des signifiants et des structures : interprète les chaînes de bits comme des types de données et d'autres structures de niveau supérieur (par exemple, interpréter les chaînes de bits en caractères ASCII et structurer les caractères ASCII en séquences reproductibles). Il comprend notamment la spécification du format des données et de l'environnement logiciel et matériel nécessaire pour accéder aux données.

- **Meta S** : Utilise les Meta R, via des langages normalisés [RDF / OWL / SKOS ...], pour représenter les « Signifiés » : Information sémantique : fournit un cadre supplémentaire pour interpréter les « expressions » de données : Textuel – Structurel – Contextuel.

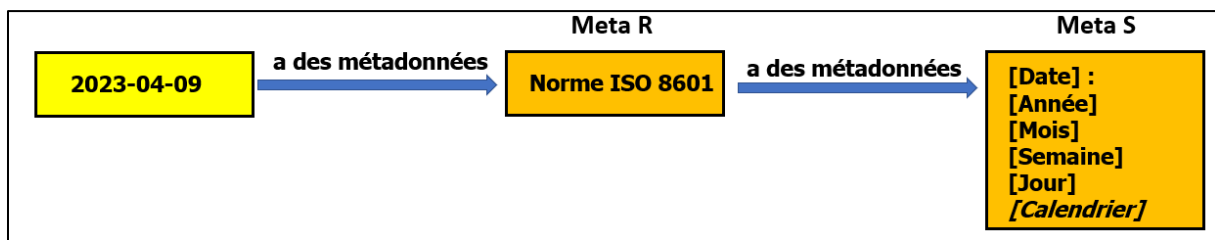


Figure 8-3 : La représentation des DONNÉES par des Metadonnées

Représentation selon ISO 8601 Plages de valeurs :

Année	(A)	AAAA, sur 4 chiffres ou sur 2 chiffres en format abrégé
Mois	(M)	MM, de 01 à 12
Semaine	(W)	WW, de 01 à 53
Jour	(J)	J, le jour de la semaine, de 1 à 7

Ce formalisme peut être explicite et lisible par les machines : c'est particulièrement le cas si les données sont exprimées dans des langages formels sémantiques, tel que [RDF / OWL / SKOS ...].

- Les « Métadonnées R et S » expriment les contextes et le cadre d'interprétation de « données » :

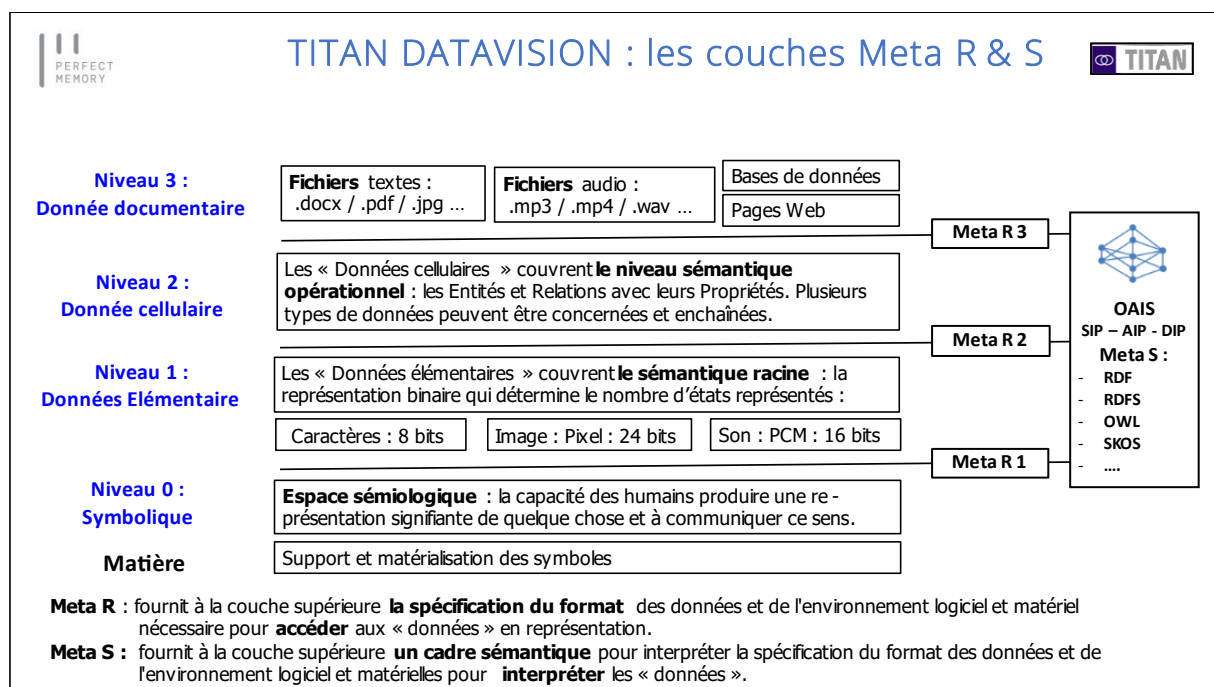


Figure 8-4 : DATAVISION : Meta R (Représentation – Meta S (Sémantique)

Ce formalisme peut aller jusqu'à expliciter les déclarations d'existence de « quelque chose », qu'il soit existant dans le monde réel physique, dans le monde des concepts ou représentation de « quelque chose » ! Dans ce dernier cas, cette représentation est implicitement déclaration d'existence de ce « quelque chose » et explicitement métadonnée de ce « quelque chose ».

A traiter :

Meta R & S : implémentation d'une solution compatible OAIS par construction)

9 Données – Métadonnées : exemples niveaux d’interopérabilité

- Exemple : de l’objet donnée à la représentation³ :

Cette partie se concentre sur l’ <objet données> (représentant un ou plusieurs fichiers), les <informations de représentation> (génération via des applications) et la contribution à la structuration d’un <objet d’information> (qui peut représenter des collections d’objets)⁴.

Un livre, une peinture, une photo, ..., présentent de l’information «en évidence», en perception mais quid de ceci :

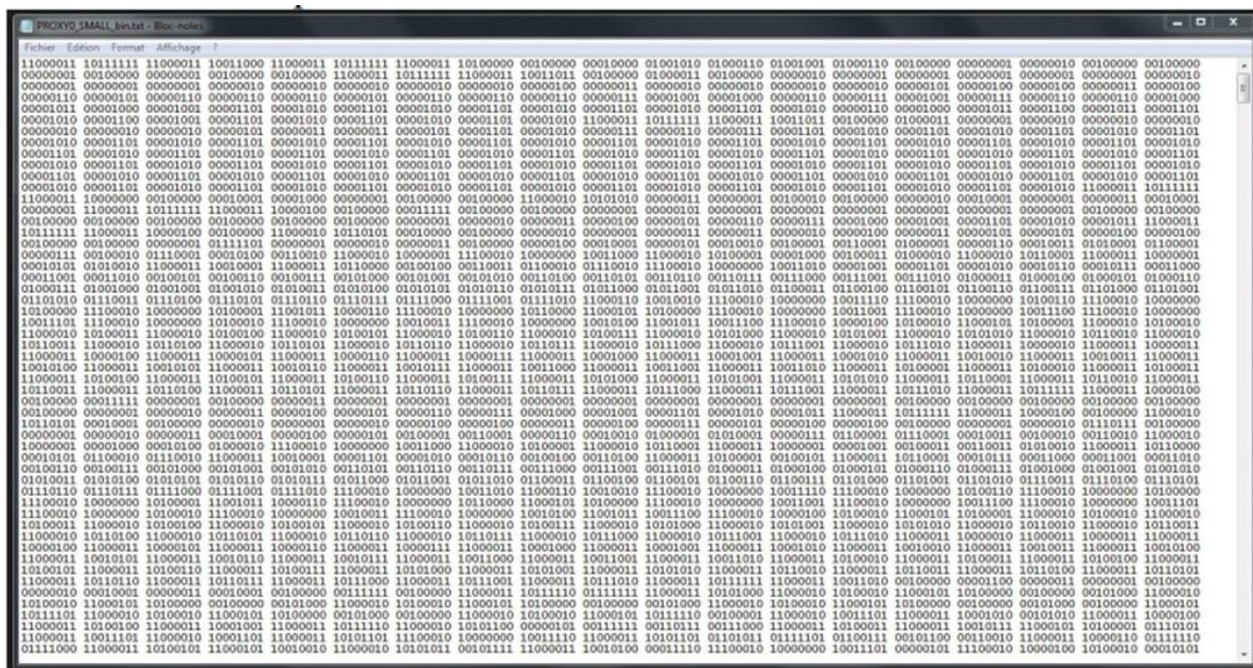


Figure 9-1 : un fichier à l’état brut

Un fichier se présente sous la forme d’une séquence de bits ! Il est clair que, sous cette forme, il est impossible d’interpréter la séquence de bits (d’avoir accès à l’objet-information) sans informations supplémentaires. Ce sont ces informations supplémentaires qui sont appelées « information de représentation ».

Le format d’un document est une partie de cette information de représentation.

Dans notre exemple le fichier contient des caractères codés selon la norme ISO 646 (codage ASCII) :

TAB (Tabulation horizontale)	09	00001001
LF (Line feed)	10	00001010
VT (Vertical tabulation)	11	00001011
CR (Carriage return)	13	00001101
Chiffre 1	49	00110001
Chiffre 2	50	00110010
Lettre A	65	00100001
Lettre a	97	01100001

Figure 9-2 : information de représentation : codé selon la norme ISO 646

³ Base HELB : Jean-Louis Blanchart

⁴ Suivant la codification de la norme OAIS :

Pour permettre dans l'avenir à un internaute d'interpréter ce fichier, le documentaliste doit :

- avoir la certitude que cette norme de codage sera préservée dans un espace accessible,
- ou alors il doit en assurer la conservation dans l'Archive.

Pour interpréter le fichier, il faut introduire la séquence de bits dans un convertisseur binaire / ASCII qui transforme chaque octet du fichier binaire en une représentation graphique correspondant au caractère codé dans cet octet. La visualisation du fichier se fait avec le bloc-notes :

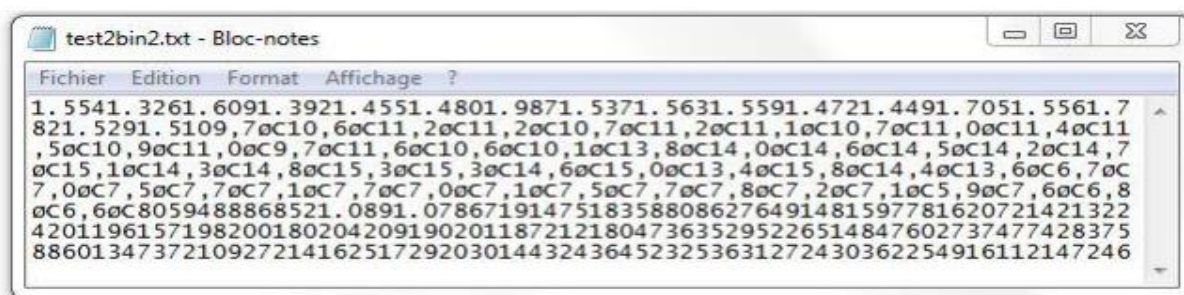


Figure 9-3 : représentation dans un convertisseur ASCII

Pour disposer d'une représentation intelligible et percevoir le sens de tous ces caractères, il faut des informations supplémentaires, par exemple le fait de savoir que notre fichier est constitué d'une répétition de séquences se terminant par le caractère « retour de chariot », chaque séquence ayant sa propre structure qui doit être explicitée (une séquence contient X champs d'informations contenant Y nombres entiers codés sur 6 octets). Si ces informations sont disponibles, alors l'application peut afficher le fichier sous une forme plus structurée :

test2.txt - Bloc-notes																
Fichier	Edition	Format	Affichage	?												
1.554	1.326	1.609	1.392	1.455	1.480	1.987	1.537	1.563	1.559	1.472	1.449	1.705	1.556	1.782	1.529	1.510
9,70C	10,60C	11,20C	11,20C	10,70C	11,20C	11,10C	10,70C	11,00C	11,40C	11,50C	10,90C	11,00C	9,70C	11,60C	10,60C	10,10C
13,80C	14,00C	14,60C	14,50C	14,20C	14,70C	15,10C	14,30C	14,80C	15,30C	15,30C	14,60C	15,00C	13,40C	15,80C	14,40C	13,60C
6,70C	7,00C	7,50C	7,70C	7,10C	7,70C	7,00C	7,10C	7,50C	7,70C	7,80C	7,20C	7,10C	5,90C	7,60C	6,80C	6,60C
805	948	886	852	1.089	1.078	671	914	751	835	880	862	764	914	815	977	816
207	214	213	224	201	196	157	198	200	180	204	209	190	201	187	212	180
47	36	35	29	52	26	51	48	47	60	27	37	47	74	28	37	58
8	6	0	1	3	4	7	3	7	2	1	0	9	27	2	14	16

Figure 9-4 : information de représentation : codé en table

Cette représentation qui affiche la structure d'une feuille de tableur est déjà beaucoup plus claire, mais il manque encore des informations capitales pour la compréhension de ces chiffres, à savoir la signification de chacun de ces champs !

Si l'on explicite la signification de chaque champ par exemple dans un tableur excel, on obtient alors le résultat suivant :

Variables mesurées	Valeurs normales (a)	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Durée de l'ensoleillement (en heures)	1.554	1.326	1.609	1.392	1.455	1.480	1.987	1.537	1.563	1.559	1.472	1.449	1.705	1.556	1.782	1.529	1.510
Température moyenne réelle (0-24h)	9,7°C	10,6°C	11,2°C	11,2°C	10,7°C	11,2°C	11,1°C	10,7°C	11,0°C	11,4°C	11,5°C	10,9°C	11,0°C	9,7°C	11,6°C	10,6°C	10,1°C
Température maximale moyenne	13,8°C	14,0°C	14,6°C	14,5°C	14,2°C	14,7°C	15,1°C	14,3°C	14,8°C	15,3°C	15,3°C	14,6°C	15,0°C	13,4°C	15,8°C	14,4°C	13,6°C
Température minimale moyenne	6,7°C	7,0°C	7,5°C	7,7°C	7,1°C	7,7°C	7,0°C	7,1°C	7,5°C	7,7°C	7,8°C	7,2°C	7,1°C	5,9°C	7,6°C	6,8°C	6,6°C
Total des précipitations (en mm)	805	948	886	852	1.089	1.078	671	914	751	835	880	862	764	914	815	977	816
Nombre de jours de précipitations (pluie >= 0.1 mm)	207	214	213	224	201	196	157	198	200	180	204	209	190	201	187	212	180
Nombre de jours de gel (min < 0°C)	47	36	35	29	52	26	51	48	47	60	27	37	47	74	28	37	58
Nombre de jours d'hiver (max < 0°C)	8	6	0	1	3	4	7	3	7	2	1	0	9	27	2	14	16
Nombre de jours d'été (max >= 25°C)	25	17	29	20	30	14	43	24	36	45	23	25	36	31	27	24	30
Nombre de jours de forte chaleur (max >= 30°C)	3	6	2	2	5	4	9	1	6	11	2	1	4	7	2	4	6

Figure 9-5 : le bilan climatologique belge de 1998 à 2013 publié par l'IRM.

L'objet-information est le bilan climatologique belge de 1998 à 2013 publié par l'IRM.

Meta S : Information de représentation + Base de connaissance :

Représenter le texte vs Comprendre le texte !

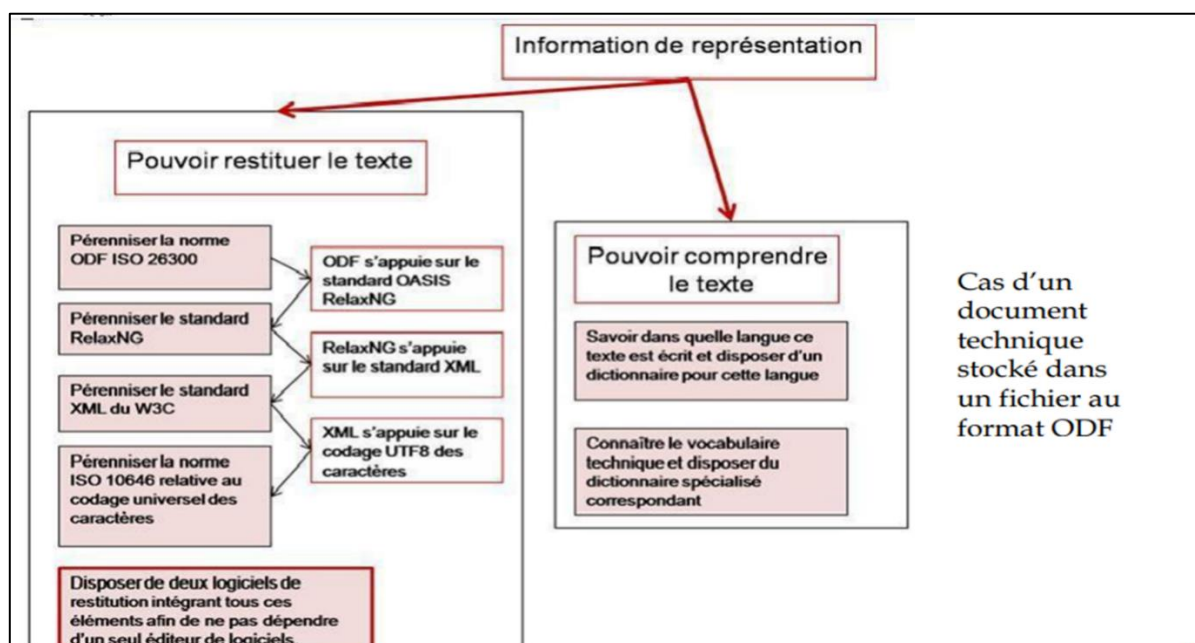


Figure 9-6 : les normes impliquées dans la représentation de ce fichier ODF

L'exemple ci-contre illustre à suffisance l'importance d'une normalisation au sein des couches Meta 1 Représentation et 2 Sémantique afin de garantir une pérennité de l'exploitation des données.

10 la structuration des données : modèle table vs graphe

La transformation d'information en données requiert des connaissances que traduisent le traitement et l'interprétation qu'en ont les individus.

La création de connaissances est un processus complexe qui peut être analysé à travers différents modèles. Ces modèles offrent des perspectives intéressantes sur la manière dont les connaissances sont créées et partagées au sein des organisations. Le processus de création de connaissances implique : le partage des connaissances tacites, la création de concepts, la justification des concepts, la construction d'un archétype et l'extension des connaissances.

Les textures, les dimensions et les corrélations entre les données peuvent être exprimées mathématiquement. La plupart des problèmes auxquels les entreprises/institutions font face peuvent être résolus à l'aide de modèles analytiques basés sur les mathématiques pures. De nombreux algorithmes de ML (machine learning) reposent sur l'algèbre linéaire.

Une table est un ensemble de données organisées sous forme d'un tableau où les colonnes correspondent à des catégories d'information. La notion de table est apparue dans les années 1970 chez IBM avec l'algèbre relationnelle qui est une théorie mathématique en relation avec la théorie des ensembles.

On distingue deux types de tables :

- les tables simples qui comprennent un nombre variable de colonnes avec une seule ligne de données,
- et les tables à double entrée qui comportent un nombre variable de colonnes et plusieurs lignes de données.

Chaque ligne est appelée un enregistrement et chaque colonne, un champ. Un enregistrement est un moyen cohérent et significatif de combiner des données sur un sujet. Un champ est un élément d'information unique (type d'élément qui apparaît dans chaque enregistrement).

Les colonnes représentent l'en-tête de la table et les lignes représentent le corps de la table. Les colonnes sont également appelées les attributs ou les champs. Les propriétés d'une table sont constituées par les colonnes (attributs) de cette table. Chaque ligne d'une colonne représente une valeur pour la colonne.

Les tableaux à double entrée sont utilisés pour mettre en relation des caractéristiques, des critères ou des propriétés d'objets ou d'éléments qui se croisent horizontalement (caractère en lignes) et verticalement (caractère en colonnes). Ils sont souvent utilisés pour observer, raisonner, et se repérer dans l'espace. Par exemple, ils peuvent être utilisés pour mettre en évidence les liens entre un objet et sa taille, sa couleur, sa position dans l'espace. Les tableaux à double entrée sont également utilisés pour classer et comparer des données en croisant lignes et colonnes.

Dans une représentation en table, les colonnes ne sont pas des « classes » et les lignes ne sont pas des « propriétés » de ces classes.

Les tables comportent des colonnes qui sont en fait des catégories et des lignes qui sont destinées à accueillir des enregistrements de valeurs relatifs à un terme :

TITAN asbl	Nom	Langue	Citoyen de	Origine	Date de Naissance
4BA-TRE	Belleflamme Alfred	D – FR – EN - NL	Belgique	Limbourgeoise / germanophone	1955-03-21
07JLB-SE	Blanchart Jean-Louis	FR – EN – NL	Belgique	Hainaut/ Picard	1960-01-18
5BF-ADM	Bodson Francis	FR – EN – NL	Belgique	Mosane / Wallonne	1947-07-31
1GM- ADM	Maréchal Guy-Noel	FR – NL - EN	Belgique	Mosane/ Wallonne	1937-12-25
6MP- ADM	Pirlot Michel	FR – EN	Belgique	Calestienne/ Picard	1950-04-18
ORRO- PRE	Roberts Roger	FR – D – NL - EN	Belgique	Limbourgeoise / Thioise	1950-04-09
3MLS-SEC	Schellings Marie- Laure	FR – EN	Belgique	Mosane / Wallonne	1963-05-03

Figure 10-1 : représentation de données en mode table

La figure suivante illustre la gestion des données des différents membres du CA de l'asbl Titan via une représentation en tableau sur la base de caractéristiques individuelles :

Le titre : composition du CA de l'asbl TITAN (Président, Secrétaire, Administrateurs et secrétariat)

- La source est inconnue.
- Les colonnes : il s'agit de différentes catégories qui ne sont pas homogènes. Le nom/prénom, , la (les) langues parlées, la nationalité, l'origine, la date de naissance.
- Les lignes : la liste établie sur la base d'un classement alphabétique latin du nom de famille suivi du prénom de la personne.

Cette représentation appelle de nombreux commentaires :

- Le tableau est rédigé dans la langue française (Belgique), Police Arial Corps 15, Noir ;
- Les noms/prénoms sont typiques d'un usage occidental (Nom de Famille, suivi d'un premier Prénom) ;
- Pour la langue parlée, le documentaliste a fait appel aux abréviations de la norme ISO-639-1 ;
- La dénomination du Pays est en langue française ;
- Pour l'origine, le documentaliste a fait appel à un catalogage régional et culturel ;
- Pour la date de naissance, elle est relative au calendrier Julie rédigée suivant la syntaxe de la norme ISO-8601 ;
- Le code de la première colonne relève d'une syntaxe implicite du documentaliste (SEC = secrétaire).

Enfin ce tableau a été généré par une application (PPT – Office 365 dans ce cas) ! Tout export dans une autre application va modifier les paramètres de cette représentation.

10.1 L'indexation : de la « Table » au « Tableau » :

En informatique, un [tableau] est une structure de données représentant une séquence finie d'éléments qui doivent être du même type et auxquels on peut accéder (position, indice, ...). Chaque case ne peut accueillir qu'une seule valeur et donc une chaîne comme FR -NL – D doit faire l'objet d'une colonne par langue ! De même la syntaxe « Nom – Prénom » ne peut faire l'objet d'aucune distinction par la suite du fait qu'elle figure dans une même colonne !

Un tableau est représenté en mémoire sous la forme de cellules contiguës dont la configuration est pré-existante aux enregistrements ! L'accès à chaque élément du tableau se fait à partir de l'adresse de base du tableau et de l'index de l'élément. De ce fait les opérations d'insertion ou de suppression d'élément sont quasi impossibles. Pour ce faire, il est nécessaire de copier tous les éléments du tableau original dans le nouveau tableau, puis de libérer l'espace mémoire alloué à l'ancien tableau.

Cette vision table (flat model) est au cœur de la base de données relationnelle. La modélisation sémantique d'une base de données relationnelle va permettre de construire un conteneur donnée/concept (signifiant/signifié) que la machine informatique va pouvoir exploiter en vue de pouvoir répondre plus intelligemment aux interactions avec l'utilisateur. C'est une transition entre une indexation implicite de données et une sémantique explicite qui se charge à la fois du traitement du sens des données, de leur représentation et de leurs relations effectives.

10.2 Les bases des données relationnelles :

Dans les bases de données relationnelles, une table est un ensemble de données organisées sous forme d'un tableau où les colonnes correspondent à des catégories d'information (une colonne peut stocker des numéros de téléphone, une autre des noms...) et les lignes à des enregistrements, également appelés entrées. Chaque table est l'implémentation physique d'une relation entre les différentes colonnes. Chaque correspondance est définie par une ligne de la table.

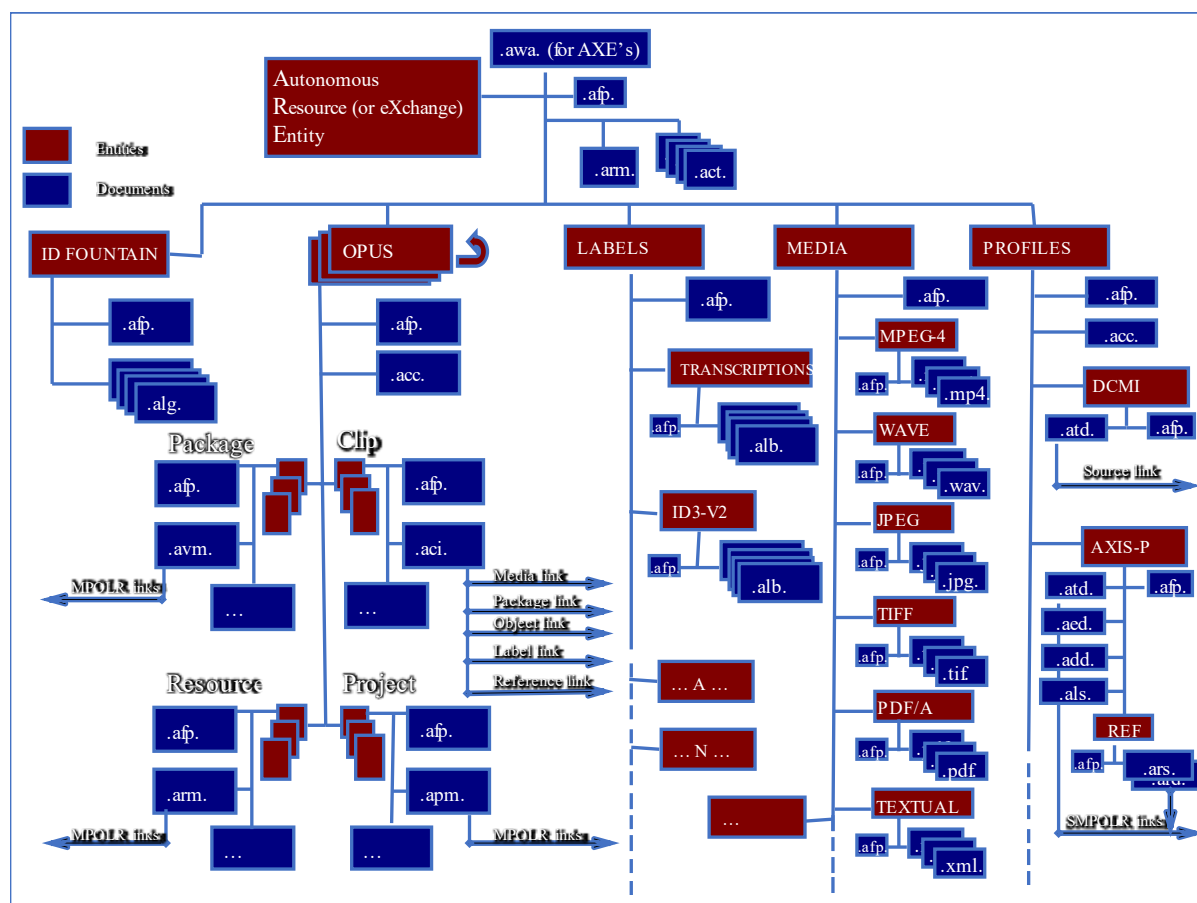


Figure 10-2 : description de la définition des tables d'une base de données

Lors de la conception d'une base de données relationnelle, il est important de clairement définir toutes les tables qui la composeront et les différentes associations qui les lient, de manière à pouvoir dresser le schéma conceptuel qui permettra de décrire l'organisation des données dans la base de données avant de la mettre informatiquement en place.

On distinguera les tables courantes qui caractérisent des entités (Tables : “ID Fountain”, “OPUS”, “LABELS”, “MEDIA”...) contenant des données, et les tables associatives qui assureront la liaison entre au moins deux tables courantes.

Il existe deux niveaux pour définir et manipuler les données d’une table dans une base de données relationnelle :

- un niveau de définition des données d’une table : qui permet de définir, lier, et contraindre les données via un langage de définition de données ;
- un niveau de manipulation des données d’une table : qui permet d’ajouter, supprimer, rechercher des données via un langage de manipulation de données.

10.3 Les bases de données orientées graphe

Les bases de données orientées graphe sont souvent utilisées pour stocker des données complexes et interconnectées dans des domaines tels que la bio-informatique, la finance (M-A : merge and acquisition), les outils de recherche ... et les réseaux sociaux. Avec certaines bases de données en graphe, la mise en œuvre représente des tables de manière efficace tant en taille de la base qu'en performance lors des accès !

La littérature technologique regorge de comparaisons entre ces deux modes de structuration !

Par exemple : les bases de données orientées graphe :

- représentent la complexité d’une représentation. De ce fait elles peuvent être plus difficiles à interroger que les bases de données relationnelles, car les requêtes peuvent nécessiter des connaissances plus avancées en matière de graphes ;
- exigent plus d’expertise sémantique pour les maintenir que les bases de données relationnelles ;
- sont peu efficaces pour stocker des données qui ont une structure tabulaire simple.

... mais tous ces discours ne tiennent pas la route à partir du moment où c’est l’interface homme machine qui prend en charge la complexité !

Aucun utilisateur de Google ne mesure à quel point les algorithmes du SEO sont complexes, le niveau de sophistication de Bing aujourd’hui avec des éléments empruntés à de l’Intelligence Artificielle (deep / machine learning = apprentissage machine) doit décoiffer !

10.4 La structuration des données en table vs en graphe :

La structuration des données en graphe et en table sont deux approches différentes pour stocker et organiser les données :

- les bases de données relationnelles utilisent des tables pour stocker les données,
- les bases de données orientées graphe utilisent des nœuds et des « arêtes » pour représenter les données.

Les bases de données orientées graphe sont souvent utilisées pour stocker des données complexes et interconnectées, tandis que les bases de données relationnelles sont souvent utilisées pour stocker des données structurées et tabulaires. Les bases de données orientées graphe sont plus flexibles et évolutives que les bases de données relationnelles.

En effet, les limites d'une structure « table » viennent de son avantage : une table étant représentée en mémoire sous la forme de cellules contiguës, les opérations d'insertion et de suppression d'élément sont impossibles, sauf si on crée une nouvelle table, de taille plus grande ou plus petite (selon l'opération).

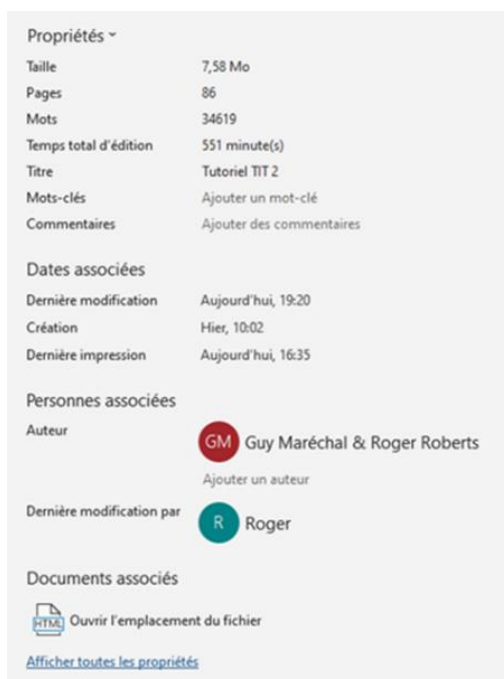
10.5 La structuration des données en table vs en graphe :

Les bases de données orientées graphe ont plusieurs avantages par rapport aux bases de données relationnelles. Les bases de données orientées graphe :

- sont très flexibles et peuvent facilement s'adapter à des structures de données complexes et changeantes ;
- sont très performantes pour les requêtes qui impliquent des relations complexes entre les données ;
- sont très évolutives et peuvent facilement gérer des ensembles de données très volumineux ;
- sont très efficaces pour stocker des données qui ont une structure hiérarchique ou en réseau ;
- offrent la cohérence : la capacité pour un système à refléter sur la copie d'une donnée les modifications intervenues sur d'autre copies de cette donnée. Cette notion est principalement utilisée dans trois domaines informatiques : les systèmes de fichiers, les bases de données, et les mémoires partagées ;
- offrent de l'inférence : un mécanisme qui permet à un compilateur ou un interpréteur de rechercher automatiquement les types associés à des expressions, sans qu'ils soient indiqués explicitement dans le code source. Il s'agit de construire des informations valables qui ne sont pas écrites explicitement dans un texte. Les inférences sont donc des raisonnements déduits ou induits à partir des informations du texte. Elles permettent la construction des grandes composantes macrostructurales et leur mise en relation souvent causale.

10.6 Les données ancillaires des fichiers :

Le fait d'inclure les propriétés d'un document (via un modèle de données) dans le fichier permet par la suite de les organiser, de les identifier et de les rechercher sur la base de leurs propriétés. Dans OFFICE (WINDOWS), il existe quatre types de propriétés de document :



- Propriétés standard : l'auteur, le titre et l'objet (sans contrainte sur les valeurs)
- Propriétés automatiquement mises à jour : la taille, les dates de sa création ou de sa dernière modification et les statistiques conservées par les programmes Office (le nombre de mots ou de caractères dans un document). Pas de possibilité de spécifier ou modifier ces propriétés automatiquement mises à jour. Ces propriétés permettent d'identifier ou de rechercher des documents (par exemple, rechercher tous les fichiers créés après le 3 août 2015 ou tous les fichiers modifiés hier).
- Propriétés personnalisées supplémentaires : une heure, une valeur numérique ou les valeurs des propriétés non personnalisées.
- Propriétés de bibliothèque de documents (un site web ou d'un dossier public). Lors de la création d'une bibliothèque de documents, il faut définir les propriétés et déterminer des règles pour leurs valeurs.

Lors de l'ajout de documents à la bibliothèque, il faut mettre à jour les propriétés qui sont parfois incorrectes. Par exemple, une bibliothèque de documents peut inviter l'utilisateur à renseigner des propriétés comme « Soumis par », « Date », « Catégorie » et « Description ». Toutes les propriétés requises de la bibliothèque de documents sont entourées de bordures rouges dans l'onglet Informations de Word 2016, Excel 2016 et PowerPoint 2016.

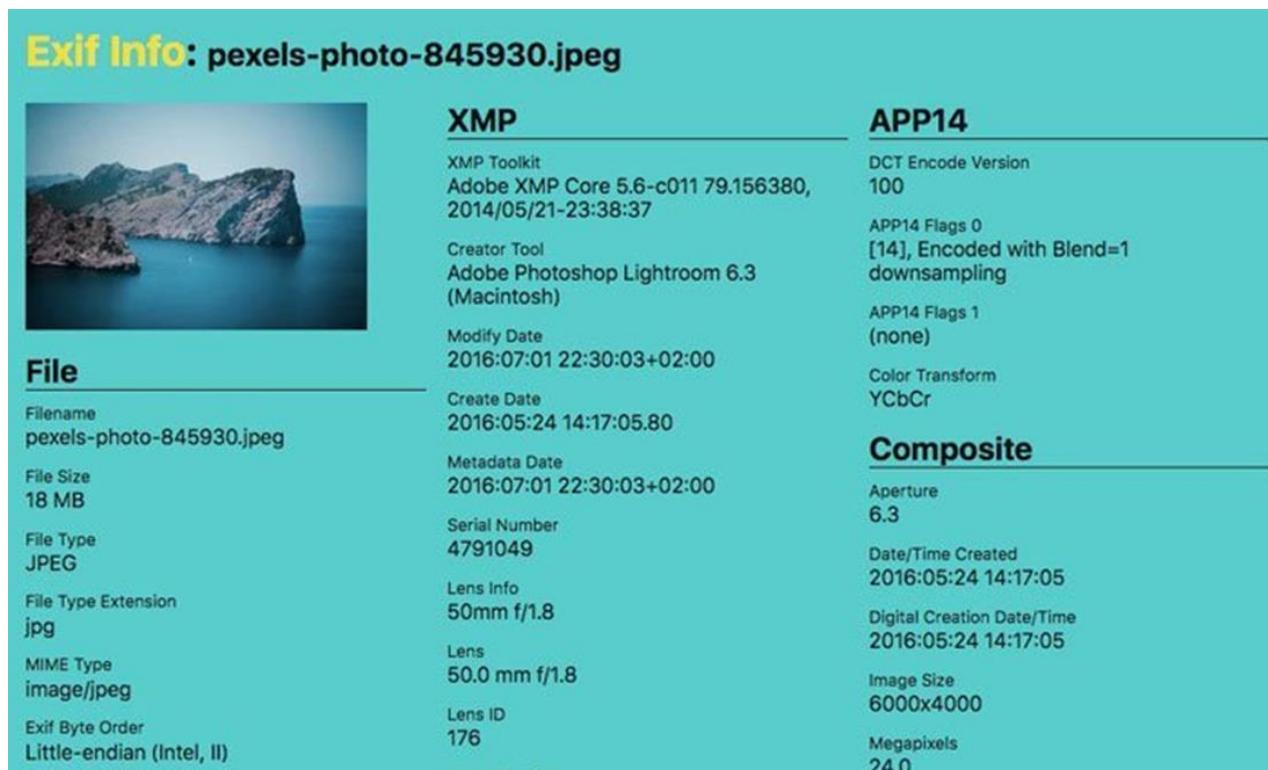


Figure 10-3 : données ancillaires d'une image (fichier) JPEG

Les données descriptives d'une photographie sont des données autres que celles constituant la représentation (les pixels) et les données techniques (vitesse d'obturation, diaphragme, lumière, sensibilité) et donnant des informations sur le contenu (sujet, date, lieu...), les droits (auteur, licence etc...), et des éléments techniques liés à la prise de vue (appareil photo, logiciel de retouche).

Elles sont stockées au sein du fichier (JPG, TIFF, PNG...).

Trois formats coexistent pour contenir ces données :

- **EXIF** (Exchangeable image file format),
- **IPTC** (Information Interchange Model) avec des "attributions" et droits, ...
- **XMP** (Extensible Metadata Platform) qui utilise du XML et qui est présent dans le PDF. XMP définit différentes méthodes pour stocker un document XML au sein même de fichiers JPEG, JPEG 2000, GIF, PNG, HTML, TIFF, Adobe Illustrator, PSD, PostScript, etc.

10.7 Classification des données :



Figure 10-4 : classification de différents types de données !

[donnée ancillaire] : donnée liée au processus de fabrication d'un fichier (les caractéristiques techniques d'une photo, la date, les coordonnées géographiques du lieu où elle a été prise).

[donnée structurée] : une donnée qui a été prédéfinie et formatée selon une structure précise (organisée sous forme d'une table de données) avant d'être placée dans un entrepôt de données (data warehouse).

[donnée non structurée] : une donnée stockée dans son format d'origine et non traitée avant son utilisation. Elles se présentent dans une multitude de formats de fichiers (e-mails, posts sur les réseaux sociaux, des chats, des données de capteurs IoT et des images satellites ...

[donnée semi-structurée] contiennent des balises ou d'autres marqueurs pour séparer les éléments sémantiques et appliquer des hiérarchies d'enregistrements et de champs dans les données. (également connu sous le nom de structure auto-descriptive). Dans les données semi-structurées, les entités appartenant à une même classe peuvent avoir des attributs différents même si elles sont regroupées, et l'ordre des attributs n'a pas d'importance.

La **complexité de l'échange de données** entre différents types dépend du format de données utilisé et des protocoles de communication employés.

En informatique, la sémantique est utilisée pour décrire le sens des données échangées entre différents types. La complexité de l'échange de données entre différents types peut être résolue en utilisant des langages de modélisation sémantique tels que RDF (Resource Description Framework) et OWL (Web Ontology Language). Ces langages permettent de décrire les données échangées entre différents types en utilisant des ontologies qui décrivent les concepts et les relations entre elles comme [Entité] – [Relation] – [Propriété] ...

[donnée sémantique] : donnée dont le formalisme utilisé dans leur représentation recouvre une définition du sens (jusqu'au niveau de détail choisi) attribué aux données et que ce formalisme est lui-même représenté par des données, via un langage informatique formel utilisant des primitives spécifiques à la logique et aux mathématiques.

11 La représentation et la signification :

La vision du [web sémantique] ouvre une dynamique nouvelle : le signifiant devient un [Identifiant] ! Le lien entre le signifiant et le signifié [Ressource] est qualifiable (une triplette qui définit un sujet, une relation et un objet) et les ontologies peuvent superposer de multiples relations entre ces « objets » en graphe ... c'est une avancée importante en termes de création de « connaissance » commune, interprétable par l'humain et la machine !

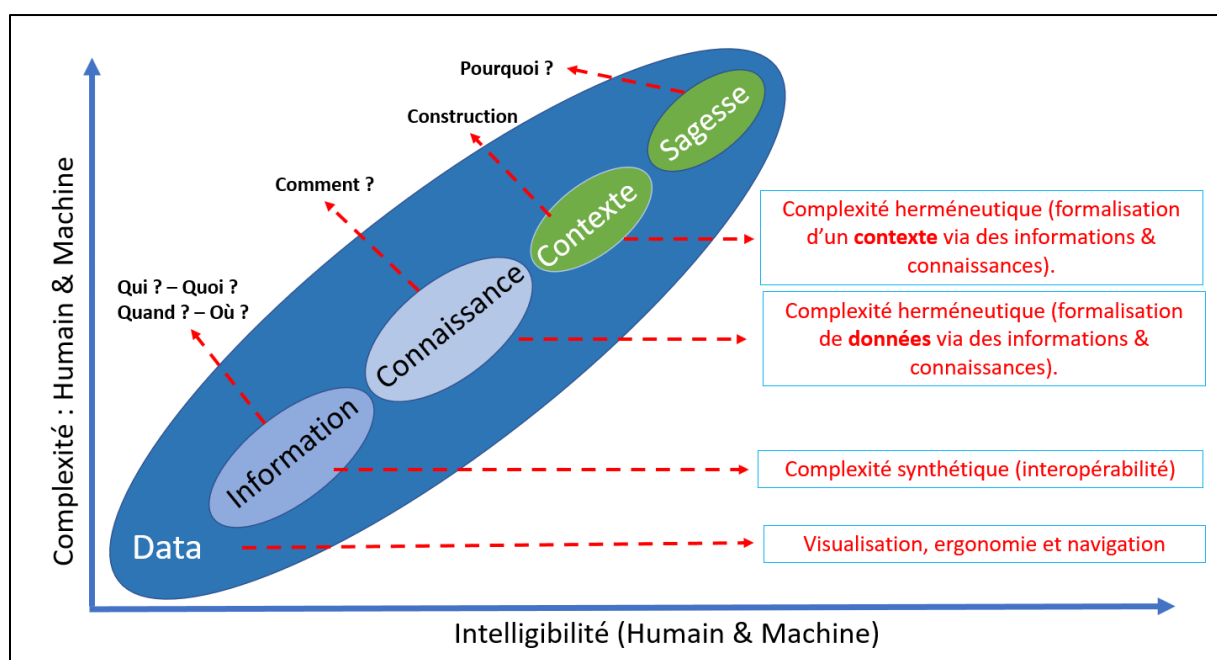


Figure 11-1 : TITAN Modèle Conceptuel : un processus intégré ?

C'est la **[Sagesse - WIDOM]** qui contribue à élaborer la couche information, celle-ci induit en retour du sens (un contexte).

[Information] : la signification que les humains assignent aux données au moyen de conventions appliquées aux données.

Cette information va être traitée au niveau d'une couche globale intitulée [Connaissance] :

[Connaissances - Knowledge] : Représentation de l'Information exprimée selon un formalisme. Le formalisme peut être de tout type et le support de la représentation peut être de tout type y compris la mémoire humaine !

C'est la couche « connaissance qui contribue à la mise en forme des données (via des langages naturels, informatiques, des formules mathématiques, des graphes, ..). C'est cette même couche « connaissance qui contribue à la spécialisation de données sémantiques (via des langages comme OWL, RDF, ..).

[Contextualisation - Insights] : la prise en compte du cadre de perception à travers lequel on émet ou on reçoit un contenu. L'irruption du Web sémantique et la généralisation des Cookies, la complétion automatique des champs d'un formulaire, la notion d'historique d'utilisation, .. ont contribué à la définition du contexte dans les interactions Homme - Machine.

[Structuration] : manière dont les unités linguistiques s'agencent entre elles en un système ordonné de règles qui décrivent à la fois les unités et les relations qu'elles entretiennent entre elles. La structuration doit être exprimée suivant un formalisme normalisé, lisible par les machines de traitement des données et compréhensibles par les humains : par exemple the « Ontology Web Language ».

Ce rôle «structurant» en tableau, graphe et arbre permet de trier, filtrer, stocker... et de rechercher et naviguer dans les silos de données.

12 Quelques définitions

12.1 Symbolique : la relation Concept - Symbole :

[Structuralisme] : publié en 1916, « le Cours de linguistique générale » de Ferdinand de Saussure propose d'appréhender la langue comme une structure (une entité de dépendances internes) : chaque entité se définit par les relations (d'équivalence ou d'opposition) qu'elle entretient avec les autres.

La **[sémiologie]** ou **[sémiotique]** est une méthode scientifique, qui est inspirée de la linguistique, et qui comprend l'étude des signes et des processus de signes, l'indication, la désignation, la ressemblance, l'analogie, l'allégorie, la métonymie, la métaphore, le symbolisme, la signification et la communication. Vaste domaine au point que certains scientifiques considèrent qu'il s'agit d'un espace « out of science », trop complexe que pour pouvoir être qualifié de « scientifique ».

La **[sémantique]** est une branche de la linguistique qui étudie les signifiés, ce dont on parle, ce que l'on veut transmettre par un énoncé, soit l'ensemble des processus concourant à la construction d'un sens dans la communication. Contrairement à la sémiologie, la sémantique du fait qu'elle repose sur la linguistique est reconnue comme science (au niveau de la phrase pour les puristes, pas au niveau textuel !

Les analyses [syntaxique] (les relations entre les unités linguistiques & la grammaire) et surtout [sémantique] (la construction du sens) ont généré de nombreux travaux ardues !

La sémantique joue le rôle d'une cartographie ; elle formalise les liens entre les données/informations.

Ce rôle « structurant » en tableau, graphe et arbre permet de trier, filtrer, stocker... et de rechercher et naviguer dans les silos de données. C'est un « Knowledge Graph » qui propulse Google, Facebook utilise Apache Giraph et Twitter a créé FlockDB, des outils construits pour assurer une grande évolutivité.

[Donnée] : une représentation formalisée d'information (faits, concepts ou d'instructions) exprimée suivant un formalisme adapté à la communication, à la mémorisation, à l'interprétation ou au traitement par des humains ou par des moyens automatiques.

12.2 Quelques définitions :

- **[Information]** : la signification que les humains assignent aux données au moyen de conventions appliquées aux données.
- **[Connaissances]** : Collection de Données (faits, concepts ou instructions) propre à servir de référence ou de contexte pour d'autres la fabrication de données textuelles,
- **[Communication]** : est un processus par lequel les organismes vivants définissent et partagent de la signification. La « Communication » nécessite un émetteur, un message, un récepteur et surtout un même « univers linguistique et sémiotique ».
- **[NS]** (Name Space - Entités Nommée) et xmlschemas permettent aux noms des éléments et des attributs dans un document XML d'être uniques.
- **[DNL]** : Documents Non Linéaires, les hypertextes, documents hypertextuels ou hypermédias qui ont vus le jour avec le développement depuis une vingtaine d'années des technologies de l'information et de la communication.
- Le **[paratexte]** est l'ensemble des éléments textuels et non textuels d'accompagnement d'un contenu. Il se compose du péri-texte (titre, préface, dédicace, etc.) et de l'épi-texte (présentation, publicité, etc ..). Le paratexte inclut des éléments visuels tels que des photos, des schémas, de la vidéo.

12.3 Archivage : quelques définitions

- **Archive** : organisation chargée de conserver l'information pour permettre à une communauté d'utilisateurs cible d'y accéder et de l'utiliser.
- **Long terme** : période suffisamment longue pour qu'il soit nécessaire de prendre en compte les changements technologiques, et notamment la gestion des nouveaux supports et formats de données ainsi que l'évolution de la communauté d'utilisateurs.
- Une **collection** est à la fois un regroupement d'objets en fonction d'une intention (un thème, une valeur documentaire, esthétique, le prix, la rareté, ...) et l'activité qui consiste à acheter, stocker et gérer l'accès à ces objets.
- Un **DOCUMENT** est la trace d'une action qui peut se manifester à l'homme. La trace peut se manifester directement (un livre...) ou via un outil (un fichier MP3...). Jusqu'à présent, les organisations se sont concentrées sur les métadonnées associées aux Documents et sur les processus d'extraction des métadonnées des Documents.
- Un **PATRIMOINE DOCUMENTAIRE** comprend ces documents uniques - ou groupes de documents - et les artefacts associés d'une valeur significative et durable pour un contexte désigné (une communauté, un pays ou pour l'humanité en général), et dont la détérioration ou la perte serait un appauvrissement nuisible !
- Un **PATRIMOINE DOCUMENTAIRE EXCEPTIONNEL** comprend le Patrimoine Documentaire qui est exceptionnel dans son contexte d'origine et/ou transcende son contexte d'origine et doit donc faire l'objet d'attention, d'aide et de régulation par les pouvoirs publics !

13 Conclusion :

Le réseau Internet, en tant que connecteur, constitue un vecteur majeur en termes de normalisation universelle !

- L'Internet repose sur de nombreuses technologies normalisées (protocole HTTP, IP ou le DNS).
- L'Internet connecte des équipements très variés, issus d'équipes différentes, et utilisant des logiciels développés de manière indépendante.

En général, le protocole est décrit par un texte, la spécification, et cette spécification est parfois formalisée dans une norme, lorsqu'elle a été développée et approuvée par un organisme de normalisation.

En revanche, la complexité de l'échange de données entre différents types dépend du format de données utilisé et des protocoles de communication employés.

En informatique, la sémantique est utilisée pour décrire le sens des données échangées entre différents types.

La complexité de l'échange de données entre différents types peut être résolue en utilisant des langages de modélisation sémantique tels que :

- RDF (Resource Description Framework),
- et OWL (Web Ontology Language).

Ces langages permettent de décrire les données échangées entre différents types en utilisant des ontologies qui décrivent les concepts et les relations entre elles comme [Entité] – [Relation] – [Propriété].

Rôle d'un "data scientist" :

Un data scientist est un professionnel qui s'occupe de la gestion et de l'analyse de données massives, communément appelées Big Data. Le but principal de ce métier est de donner du sens à ces données et d'en extraire de la valeur. Pour cela, le data scientist doit être un spécialiste des chiffres, des programmes informatiques et des statistiques qui sont traités. Il doit utiliser les services web pour traiter les données, ce qui le distingue du statisticien. Le data scientist aide l'entreprise/institution à prendre des décisions stratégiques et efficaces en concevant des modèles et algorithmes qui collectent, stockent, traitent et restituent les données. Les compétences requises pour ce métier sont la créativité, la rigueur, le sens de l'organisation et bien sûr la bonne maîtrise des chiffres.

De façon générale, un bon data scientist doit avoir des connaissances solides en mathématiques.

Deuxièmement, le data scientist doit être doué d'une forme de créativité technologique. Pour cause, il utilise la technologie pour explorer d'immenses ensembles de données et travailler avec des algorithmes complexes afin de résoudre des problèmes complexes. Pour ce faire, le scientifique des données doit être capable de coder, de créer des prototypes de solutions rapides, et de les intégrer aux des systèmes de données complexes. Parmi les principaux langages associés à la science des données, on compte :

- le SQL (Structured Query Language ou « langage de requêtes structurées ») est un langage informatique normalisé servant à exploiter des bases de données relationnelles. La partie langage de manipulation des données de SQL permet de rechercher, d'ajouter, de modifier ou de supprimer des données dans les bases de données relationnelles ;
- le Python : un langage de programmation qui peut s'utiliser dans de nombreux contextes et s'adapter à tout type d'utilisation grâce à des bibliothèques spécialisées. Il est cependant

particulièrement utilisé comme langage de script pour automatiser des tâches simples mais fastidieuses, comme un script qui récupérerait la météo sur Internet ou qui s'intégrerait dans un logiciel de conception assistée par ordinateur afin d'automatiser certains enchaînements d'actions répétitives ;

- le langage R : un langage de programmation et un logiciel libre destiné aux statistiques et à la science des données soutenu par la R Foundation for Statistical Computing. Il fait partie de la liste des paquets GNU3 et est écrit en C, Fortran et R.
- Le logiciel SAS traditionnel se compose d'un ensemble de modules permettant de répondre aux besoins suivants par la programmation :
 - création et gestion de bases de données ;
 - traitements analytiques des bases de données ;
 - création et diffusion de rapports de synthèse et de listing.

En périphérie, on retrouve également Java, Scala, et Julia. Toutefois, la seule connaissance de ces langages ne suffit pas.

Création et utilisation des métadonnées :

La connaissance et les compétences requises pour extraire et créer de la documentation utile pour permettre d'utiliser et d'interpréter correctement les données. Ce processus comprend la documentation des métadonnées, c'est-à-dire les définitions et descriptions sous-jacentes des données.

Visualisation des données

La connaissance et les compétences requises pour créer des tableaux, des graphiques et des diagrammes utiles qui offrent une présentation visuelle des données, y compris l'évaluation de l'efficacité des représentations visuelles (p. ex. l'utilisation du graphique adéquat) en tenant compte de leur exactitude pour éviter une fausse représentation des données.

Annexe 1 :

Les normes informatiques de codage et transcodage de caractères :

ASCII (American Standard Code for Information Interchange) est une norme informatique de codage de caractères apparue dans les années 1960. ASCII définit 128 codes à 7 bits, comprenant 95 caractères imprimables : les chiffres arabes de 0 à 9, les 26 lettres de l'alphabet latin en minuscules et en capitales, des symboles mathématiques et de ponctuation. ASCII suffit pour représenter les textes en anglais, mais il est trop limité pour les autres langues comportant des lettres accentuées.

Unicode est un système de codage de caractères utilisé par les ordinateurs pour le stockage et l'échange de données textuelles. Ce système fournit un numéro unique (point de code) pour chaque caractère des principaux systèmes d'écriture dans le monde.

« UTF-8 » est l'abréviation de « 8-Bit UCS Transformation Format » et désigne le codage de caractères le plus largement répandu sur le World Wide Web.

Unicode et UTF-8 sont des notions de natures différentes, qui ne peuvent pas être directement comparées. Unicode est un ensemble de caractères et UTF-8 est l'un des algorithmes utilisables pour les encoder en mémoire : tables par bloc. L'UTF8 est au service de l'UNICODE.

Retro compatibilité UNICODE – ASCII : le premier octet de l'UNICODE sert à encoder les caractères ASCII, donnant ainsi au jeu de caractères une totale compatibilité avec ASCII.

Un transcodage UNICODE en ASCII est destructeur, les caractères accentués attendus s'affichent sur deux caractères improbables au lieu d'un seul : <https://titan.be/fr/axis-csrm-les-huit-mod%C3%A8les> vs les-huit-modèles.