Tutoriel
UNESCO – TITAN
De la Numérisation à la Préservation :

Description de la préservation sur la base de la norme ISO/OAIS (Open Archival Information System) et AXIS OK

Document de travail pour une contribution technologique au Tutoriel UNESCO Mémoire du Monde 2020 axé sur la préservation.

Version V2020_03_03

Objectifs

L'objectif de ce document est d'offrir une « ligne éditoriale » technologique pour la rédaction du **TUTORIEL UNESCO 2020** – Mémoire du monde.

Pour des raisons de volume, la rédaction a été divisée en trois parties :

- **Partie 1** : L'introduction des concepts liés à la numérisation/ préservation qui traite de l'aspect « représentation » à partir du modèle OAIS (structuration et signification des documents).
- **Partie 2** : la problématique de l'interopérabilité, de l'échange entre des systèmes d'information hétérogènes, le développement de certains concepts comme la substance, la gestion original/copie dans un univers numérique et les évolutions souhaitables du modèle OAIS pour couvrir l'introduction de la persistance dans l'opérationnel, la représentation numérique du signifié des « Bases de connaissance » associées communautés; les représentations des liens pertinents entre les '*Designated Communities*' ; la gestion de Configuration ([Historique / Systémique / Culturelle / Spatio-temporelle], … , les accès en perception (humaine ou technique) tant à l'acquisition qu'à la jouissance ; les accès aux outils de traitement assurant le « Business » et la « Persistance ».
- **Partie 3** : sélection, recommandations générales, description des éléments pour la mise en œuvre d'une solution, …

Objectifs

Cette première partie « De la numérisation à la préservation » est basée sur l'importance des « informations de représentation contenues dans le modèle ISO/OAIS » (à partir d'un document de travail de J-L Blanchart).

- Une description des processus de codage des essences (texte, audio, photo, vidéo) avec une description du contexte technologique, une brève description des technologies mise en œuvre et enfin une liste des normes/standards appliqués ! Les différents processus doivent pouvoir introduire la nomenclature des informations de représentation.
- Une description des processus nécessaire à la reproduction, le partage, le contrôle, l'archivage, la migration et la destruction des documents. La préservation est un synonyme de protection, de sauvegarde, elle prend donc en charge en plus de la représentation (la conversion numérique analogique) la signification des documents numérisés!
- La prise en compte des traitements effectués lors de la numérisation en vue de faciliter l'archivage d'un document numérique (accrochage des données de pérennisation).

Cette première partie devrait permettre aux différents participants d'interpréter aisément le schéma ardu de l'OAIS qui figure sur la slide suivante :

Le diagramme du modèle d'information de l'OAIS :

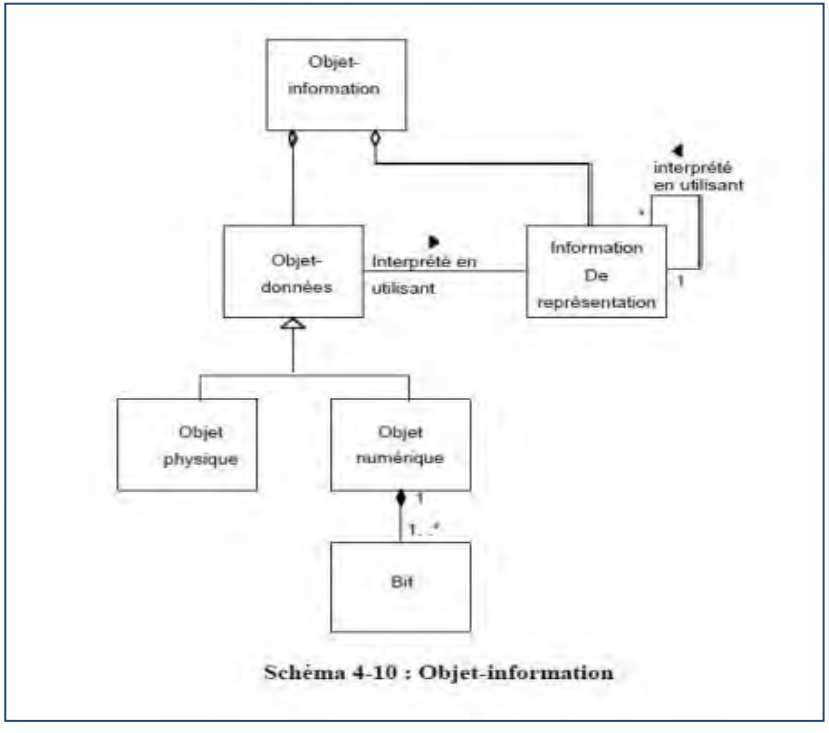


Schéma 4-10 : Objet-information

Objet information : Objet-données avec ses Informations de représentation

Information de représentation : les données accompagnée d'une description (de structure et sémantique) permettant d'interpréter cette chaîne de bits en vue d'offrir un objet d'information en perception à l'usager d'un système d'information

Il faut à donc la fois préserver les données (sur des supports adéquats) , les information de représentation (et les applications qui ont générés ces données) et enfin créer une base de connaissance pour générer les liens entre les données et leur(s) signification(s).

Préface

Ce Tutoriel **UNESCO 2020** « Mémoire du Monde sur la préservation s'inscrit dans le contexte :

- de la publication de la brochure « Normes et lignes directrices techniques et organisationnelles pour les initiatives de numérisation des patrimoines culturels soutenues par la CFWB» publiée en 2009 et du **PEP'S**..
- du domaine **« Information & Communication » de l'UNESCO** : construire des sociétés du savoir globales et locales comme clés du développement et de la paix.
- du **« Patrimoine Documentaire »** au sens où l'UNESCO en fait une dimension essentielle à la fois sociale, sociétale, économique, scientifique, et culturelle.

La problématique de la préservation n'est qu'un des aspects de la problématique de **la conservation du Patrimoine documentaire** qui comprend également la sélection, l' appropriation (en particulier la numérisation), et l'accès. La première partie de ce tutoriel va couvrir la préservation et partiellement l'appropriation. La sélection sera traitée dans la troisième partie dans le cadre des recommandations aux pouvoirs publics.

Préface

Le cycle de vie des documents ne doit donc pas abandonner l'idée que la conservation a comme objectif de permettre à des humains de jouir en perception de documents et d' informations non documentaires dans l'espace et dans le temps.

L'introduction d'une technologie nouvelle se doit de prendre en compte l'ensemble des processus qui nécessitent la mise en œuvre de moyens matériels, logiciels, juridiques et humains pour la **numérisation** (le traitement) et la gestion dans l'espace (l'échange) et le temps (la pérennité) de la **préservation** : édition, indexation, conservation, partage et retrait des documents numérisés.

Dans le cadre d'une évolution technologique, il est nécessaire d'insérer la numérisation dans la préservation ! Ce sont **deux faces interdépendantes** au sein d'un même processus ! En effet tous les traitements effectués lors d'un archivage d'un document numérique nécessitent l'existence de fonctions permettant l'accrochage des données de pérennisation (documentation du contenu, déclaration d'existence, transcription textuelle, traductions, information de représentation, base de connaissance) avec les fichiers essences (texte, audio/vidéo, photo, 3D…).

La numérisation : la médiation informatique

la numérisation est un processus qui transforme des données textuelles/graphiques, iconiques, sonores/musicales, audiovisuelles, 3D, …) ou d'un signal électrique en données numériques (fichiers/dossiers) que des dispositifs informatiques numériques pourront traiter, rendre accessibles et exploiter en ligne ! La numérisation traite exclusivement l'aspect « représentation » des documents.

Objet photo en perception

Analyse
Filtrage
Code – Table de codage
Echantillonnage
Quantification
Métadonnées

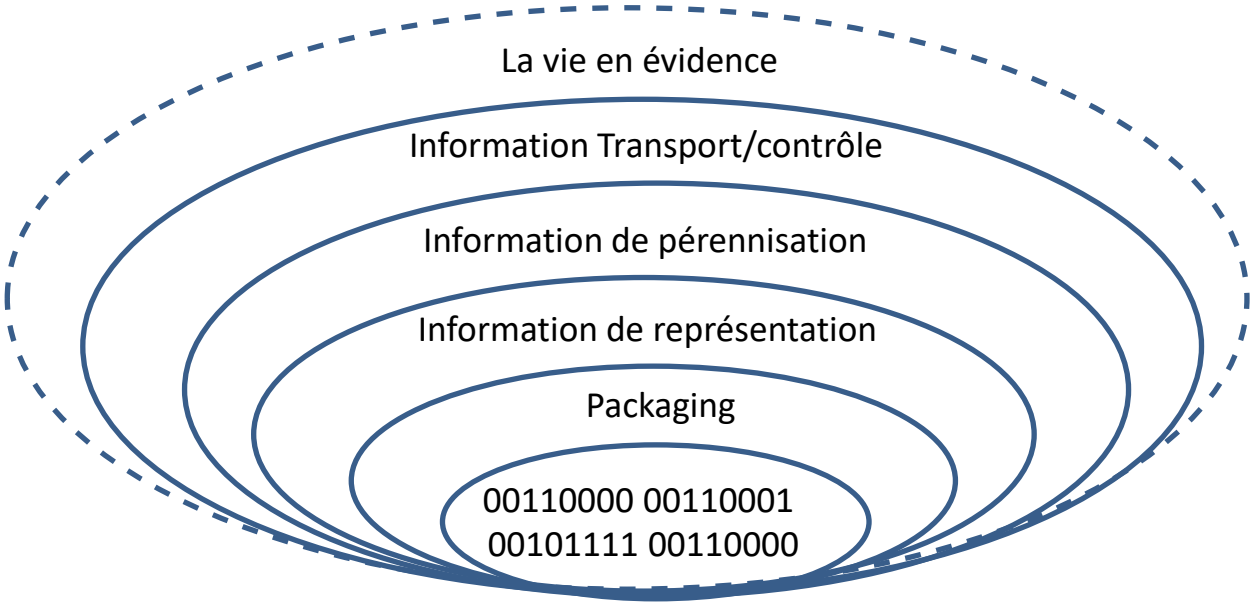
Information de représentation

00110010 00110000 00110000 00110001
00110010 00110000 00110000 00110001
00101111 00110000 00111001 00101111
00110001 00110001 00110010 00110000
00110000 00110001 00101111 00110000
00111001 00101111 00110001 00110001
00101111 00110000 00111001 00101111
00110001 00110001 00110010 00110000
00110000 00110001 00101111 00110000
00111001 00101111 00110001 00110001
……

Fichier

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 récréer l'objet avec des bits manquant, un fichier corrompu ou corriger des erreurs de transmission (redondance de 15% pour CD audio, 50% pour le format BETA NUM de chez SONY). Enfin la numérisation génère le "code génétique" d'un document. La connaissance de cette suite de valeurs, permet de **reproduire un document à l'identique**, le concept d'original s'évanouit à condition de ne pas coder avec pertes (compression numérique).

La préservation : la connexion à la vie réelle



La préservation : toutes les données numériques doivent être organisées, emballées et entreposées adéquatement pour assurer le stockage, la reproduction, l'indexation, le partage, le contrôle, l'archivage, la migration et la destruction). La préservation implique la transformation dynamique, la durée, le long terme, elle prend donc en charge à la fois la représentation (la conversion numérique analogique), la signification des documents et la connexion avec la vie en évidence !

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées et les (hyper) liens
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 Les différentes encapsulations (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.4 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 4.1 le traitement du texte
    - 4.2 Le traitement du son (temps)
    - 4.3 Le traitement des images (espace)
    - 4.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 La représentation de la substance : le modèle d'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

1. Création et définition d'un document

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.

Un document est porteur **de signes, de symboles, de conventions sémiotiques, linguistiques, de codes, de tables de codage, …** exprimant l'intention sur des supports (les tablettes cunéiformes, les hiéroglyphes, papyrus, parchemin, papier, vinyle ...)

Avec le temps, les humains ont façonné des outils complémentaires 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 (les bandes magnétiques ou optiques, DVD, clé USB, ...).

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 pariétale. Un **univers numérique** requiert l'aide d'une application pour, par exemple, rendre perceptible un fichier codé en MP3 ou en MPEG2, un site web.

1. Définition de la substance d'un document

D'une façon générale le mot **substance** signifie désigne la matière qui constitue un objet (minéral, végétal; mais cela peut également signifier ce qu'il y a d'essentiel dans un document, dans un acte la « substantifique moelle » (F. Rabelais).

En ontologie, le mot substance désigne ce qu'il y a de permanent dans toute mutation des objet représentés par des signes. Une **oeuvre** n'est pas constituée d'un ensemble de signes, mais est considérée comme véhiculant une intention, un contenu.

Dans ce Tutoriel nous allons lier ce concept de substance à celui d'**œuvre** [**work**]. Cette formulation permet de distinguer le contenu conceptuel d'un document et les signes (représentation) qui véhiculent ce contenu conceptuel. Elle va faciliter le fait que le concept d'œuvre fédère l'ensemble de ses représentations symboliques (lien vers des personnes/agent, des rôles, des droits, ...

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

2.0 TIC : Représentation - donnée

Représentation : le fait de rendre perceptible (un objet, une chose abstraite) au moyen d'un ensemble de signes (idéogrammes, images, sons, ...) mais aussi la modélisation formalisée d'un objet dont on peut déclarer l'existence.

中 國

Un **idéogramme** est un symbole graphique représentant un mot ou une idée utilisé dans certaines langues vivantes (comme le chinois et le japonais) ou anciennes (comme les hiéroglyphes de l'Égypte antique). Les idéogrammes contiennent dans leur forme une partie du sens du mot qu'ils servent à transcrire.

1847

Les chiffres sont également des idéogrammes

la Chine
zhōng guó

Un **phonogramme** est un caractère écrit qui est la transcription arbitraire d'un son. Le phonogramme n'a pas de rapport au sens : il sert à transcrire un son quel que soit le sens auquel renverra ce son.

Un **pictogramme** est une représentation graphique, un dessin figuratif stylisé ayant fonction de signe. Utilisé dans l'art rupestre (dessins peints), il sert à la signalétique, il constitue une alternative à la signalisation multilingue, il permet de réduire le volume de signes inscrits sur un panneau d'affichage.

Données : une représentation formalisée de l'information (faits, de concepts ou d'instructions) présentée suivant un formalisme adapté à la communication, à la mémorisation, à l'interprétation ou le traitement par des humains ou par des moyens automatiques.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

2.1 TIC : données vs information : les chiffres

Un **chiffre** 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		𓏺	𓏺𓏺	𓏺𓏺𓏺	𓏺𓏺 𓏺𓏺	𓏺𓏺 𓏺 𓏺𓏺	𓏺𓏺 𓏺 𓏺𓏺 𓏺	𓏺𓏺 𓏺𓏺 𓏺𓏺 𓏺	𓏺𓏺 𓏺𓏺 𓏺𓏺 𓏺𓏺	𓏺𓏺 𓏺 𓏺𓏺 𓏺𓏺 𓏺	𓎆
Latin romain		I	II	III	IV	V	VI	VII	VIII	IX	X
Sinogrammes	〇	一	二	三	四	五	六	七	八	九	十
Hébreu		א	ב	ג	ד	ה	ו	ז	ח	ט	י
Araméen		𐡀	𐡁	𐡂	𐡃	𐡄	𐡅	𐡆	𐡇	𐡈	𐡉
Grec		Αα	Ββ	Γγ	Δδ	Εε	Ϛϛ	Ζζ	Ηη	Θθ	Ιι
Devanagari	०	१	२	३	४	५	६	७	८	९	
Brahmi	𑁦	𑁧	𑁨	𑁩	𑁪	𑁫	𑁬	𑁭	𑁮	𑁯	
Abjad		١	٢	٣	٤	٥	٦	٧	٨	٩	
Arabe *orientale*	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	

https://fr.wikipedia.org/wiki/Chiffre

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;
- **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).

2.1 TIC : données vs information : les nombres

Les **nombres** sont des codes composés de chiffres et représentent un classement, une quantité ou une valeur. Un nombre est écrit comme une séquence d'un ou plusieurs chiffres qui peuvent être de différentes longueurs.
La représentation du concept de «nombre» varie selon les cultures, mais se réfère toujours exactement à la même abstraction.

Les nombres représentent un certain nombre d'informations (code postal, carte d'identité, numéro de mobile, etc.) et/ou indiquent une place dans une série ou classement (liste, numéro de rue, numéro atomique, etc.).
Les chiffres de "un" à "neuf" ont sans doute été inventés en Inde avant notre ère. Ils apparaissent dans des inscriptions de Nana Ghât au 3e siècle avant J.-C . La numération de position avec **l'invention du zéro** (qui était un point à l'origine), a été inventée au cours du 5e siècle.

La propagation de cette numération est passée par le monde arabe au 9ème siècle : "Livre de l'addition et de la soustraction d'après le calcul des Indiens" avant de pénétrer l'occident chrétien au 10ème siècle (via le Pape Sylvestre).

2.1 TIC : les fondements de la linguistique

Selon Ferdinand de Saussure (1857-1913) l'objet de la linguistique est la langue qui doit être étudiée en tant que système de signes articulant chacun un signifiant et un signifié :

Signifié → **Connotation** : Par exemple : date anniversaire Attentat New-York

Signifié → **Dénotation** : Année/Mois/ Jour

Signifiant → **Représentation** : 2001/09/11

Signifiant → **Numérisation** :

00110010 00110000
00110000 00110001
00101111 00110000
00111001 00101111
00110001 00110001

2.1 TIC : les fondements de la linguistique

- **Le signifiant** = la représentation de la forme et de l'aspect matériel du signe (= insignifiant !), c'est la partie sensible du signe, appréhendable par les sens, en tant qu'enveloppe matérielle permettant d'accéder au signifié.
- **Le signifié** = la représentation mentale du concept immatériel et intelligible, l'idée associée au signe.
- La **relation** qui unit le Signifié et le Signifiant (représentée par la ligne droite) est **arbitraire** : un même concept peut être associé à des signifiants différents (selon les langues), un signifiant peut avoir différents signifiés ! Le signifiant associé à un concept donné (ou l'inverse) s'impose à la communauté linguistique (un locuteur ne peut décider de le modifier arbitrairement).

Un même signifiant peut correspondre à des signifiés différents dans les cas d'**homonymie**, la **synonymie** montre qu'il est courant de donner des signifiants différents à des significations équivalentes et la **polysémie** atteste de l'existence d'un même signifiant pour plusieurs sens apparentés ... C'est donc au **contexte** qu'il revient de valider le **sens** !

Information : la signification que l'humain assigne aux données au moyen de conventions. Lors de l'échange, elle peut être représentée par des données.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

2.2 Les métadonnées et les hyperliens

C’est dans le cadre de la description de ressources sur Internet dans les années 1990 que le terme métadonnée (en anglais : metadata) est apparu et s'est ensuite généralisé sous une définition floue (une donnée sur une donnée).

La généralisation des langages à balises (<chevron>) de type **GML** (Generalized Markup Language) comme **SGML (**Standard Generalized Markup Language)**, HTML (**Hypertext Markup Language)**, XML**) (Extensible Markup Language) qui vont permettre de fusionner d’une part les fiches, les notices, d’autre par les contenus dans une même enveloppe et qui vont donner au terme une signification.
Dès 1994, le Web 1.0 propose l’usage de métadonnées lors de la création du W3C. Les métadonnées entrent dans le cœur de l'architecture des systèmes d’information.

Méta@donnée [Metadata] du préfixe grec meta et du latin data "informations") : c’est une donnée dont la couche Meta sert à définir ou décrire le signifié de la donnée pour un humain ou une machine.

Elles sont dorénavant liées et définies dans l’espace du web sémantique (voir Resource Description Framework [RDF]; RDF Schema et Ontology Web Language (OWL).

2.2 Les métadonnées et les hyperliens

Dans l'Internet, **un lien** est un mécanisme inséré dans un texte, un bouton ou une image sur lesquels il est possible de cliquer afin de naviguer entre des sites et des pages de la Toile ! Dans les applications, les liens sont souvent utilisés pour insérer des notes de bas de page ou de fin, des commentaires des bibliographies, des références à des index ou glossaires.

Inventé par Ted Nelson en 1965 (dans le cadre du projet Xanadu) un **hyperlien** ou lien **hypertexte**, est une référence dans un *système « hypertexte »* permettant de passer automatiquement d'une page d'un document consulté à un autre document. Les liens hypertextes sont soit "externes" ou "internes" selon leur cible ou destination. Les liens internes décrivent l'organisation d'un site Web, et offrent ainsi aux moteurs de recherche de mieux comprendre leur structure et l'importance donnée à certaines pages.

Les liens sont, par défaut, écrits en bleu et soulignés. La norme **XLink** spécifie l'écriture d'hyperliens dans les langages basés sur XML. XLink permet d'établir des liens entre plus de deux ressources, d'associer des métadonnées, et de créer des liens hors des ressources liées. Un **wikilien** conduit à une page qui n'existe pas encore (sur Wikipédia en rouge).

Un lien peut rompre pour plusieurs raisons. L'explication la plus courante est que la page web n'existe plus, ce qui conduit à l'affiche d'une erreur HTTP 404 (la page est introuvable).

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

2.3 TIC : communication : l’échange de la signification

La **Communication** est un processus par lequel les organismes vivants définissent et partagent de la signification. La « Communication » nécessite un expéditeur, un message et un récepteur et un même « univers linguistique et sémiotique ». Voici un aperçu sommaire sur les facteurs constitutifs de tout acte de communication :

- Le **message** lui-même; le **destinateur** qui envoie un message et le **destinataire** qui est censé recevoir le message;
- Pour être opérant, le message requiert d'abord **un contexte** auquel il renvoie (c'est ce qu'on appelle aussi, dans une terminologie quelque peu ambiguë, le "référent", contexte saisissable par le destinataire, et qui est soit verbal, soit susceptible d'être verbalisé;
- Le message requiert **un code**, commun au destinateur et au destinataire (ou, en d'autres termes, à l'encodeur et au décodeur du message);
- Le message requiert un **contact**, un **canal physique** et une **connexion psychologique** entre le destinateur et le destinataire, contact qui leur permet d'établir et de maintenir la communication ;

2.3 TIC : communication : l'échange de la signification

- **Langage** : un moyen de communication avec un ensemble de signes (vocaux, gestuels, graphiques, tactiles, olfactifs, etc.) doté d'une sémantique, et le plus souvent d'une syntaxe.
- **Langue** : un système de signes linguistiques, vocaux ou graphiques ou gestuels, qui permet la communication entre les individus, avec une syntaxe et une grammaire. La langue est composée de signes qui unissent chacun un signifié (concept) et un signifiant (représentation) variable dans l'espace et dans le temps, des unités discrètes en perception qui doivent être identifiées par l'analyse et qui définissent une combinatoire.
- **Signification** : en linguistique, c'est le sens d'une expression (mot, phrase, énoncé, document etc.), c'est-à-dire l'idée qui y est associée (dénotation : le sens littéral du signifiant, connotations : l'ensemble des sens figurés potentiels ou dans un contexte donné).
- **Sémantique** : l'étude de la signification, du signifié des symboles et expressions. La sémantique traite des langages. Dans sa dimension textuelle elle prend en compte l'étude du texte (composé de phrases comportant un sujet, un verbe, un complément, ...).

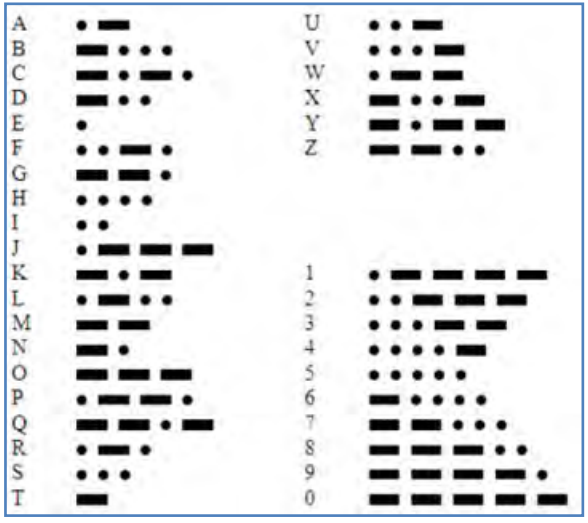
Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

2.4 TIC : codage vs transport

Les fondations technologiques des moyens de télécommunications modernes (signaux, duplex, multiplexages, commutateur, etc) naissent au début du XIX siècle. Transmettre de l'information via des signaux électriques de façon fiable devient une nécessité. Inventé en 1832 pour la télégraphie, le code Morse international est considéré comme le précurseur des communications numériques.



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), 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.

2.4 TIC : codage vs transport : le modèle OSI

En 1977, l'ISO publie une norme, **référence ISO/OSI 7498**, globalement intitulée «Modèle basique de référence pour l'interconnexion des systèmes ouverts» destinée à gérer et organiser les échanges de données entre des systèmes informatiques indépendants via 7 couches normalisées.

Les **trois couches inférieures** sont orientées infrastructure et sont souvent fournies par un système d'exploitation et par le matériel. Elles sont normalement transparentes pour les paquets de données à transporter, alors que **les couches supérieures**, qui sont orientées application, ne le sont pas nécessairement, notamment au niveau de la couche « présentation ».

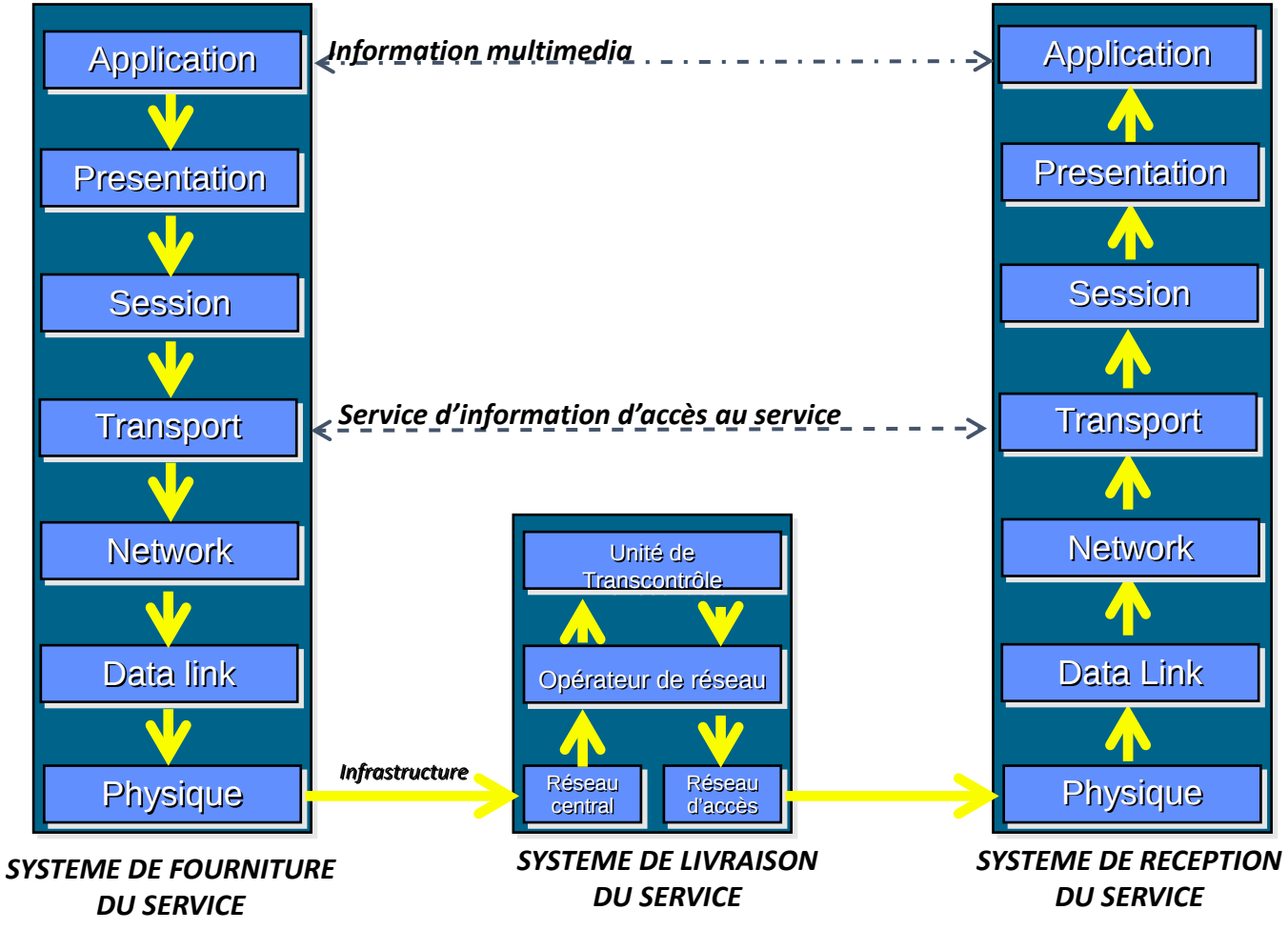
L'intention est de favoriser une **interopérabilité des systèmes** sans se préoccuper du contenu informatif transporté par la connexion. Cette approche est extrêmement porteuse par sa méthodologie qui définit le rôle fonctionnel de chaque couche, des interfaces entre chaque couche, et des protocoles assurant le dialogue entre les systèmes distincts au niveau de chaque couche. Cette approche fait que chaque couche est orthogonale, c'est-à-dire indépendante l'une de l'autre. Le contenu fonctionnel de chaque émetteur ou récepteur de la couche peut provenir de fournisseurs et de technologies indépendantes.

2.4 OSI : Open Systems Interconnection

Flux fonctionnel

Les paquets passent de couches en couches. A chaque couche, des informations de formatage et d'adressage sont ajoutées au paquet. Les paquets sont transformés en trames, et ce sont les trames qui circulent sur le réseau. Arrivées à destination les trames sont retransformées en paquets par filtrage des couches de « l'ordinateur récepteur ».

Flux réel d'information



2.4 TIC : le protocole du modèle OSI :

Un **protocole** décrit une série d'étapes à suivre pour permettre une communication harmonieuse et fiable entre des ordinateurs ou périphériques connectés en réseau. Les protocoles « orientés connexion » fonctionnent avec des accusés de réception (pour permettre une gestion des pertes des paquets et les erreurs de transmission).

Le système ne se préoccupe pas de savoir comment le bloc fonctionnel est réalisé, les seules choses qui comptent ce sont le respect des interfaces (vers le bas et vers le haut), le respect du protocole et le respect de la fonctionnalité attendue.

Le protocole OSI décrit la façon dont les couches successives se transmettent et encapsulent les données au niveau de la machine émettrice. Chaque couche ajoute des informations **(<En-tête> + <Données> + </Queue>)** au paquet de données de la couche précédente : l'adresse de l'expéditeur et l'adresse cible du destinataire, des informations d'horloge pour synchroniser la transmission, des informations d'interface (couche APPLICATION), de formatage (couche PRESENTATION), de connexion (couche SESSION), de séquence (couche TRANSPORT), de routage et d'adressage (couche RESEAU) et de transmission (couche PHYSIQUE).

2.4 TIC : le protocole du modèle OSI :

Au niveau de la machine réceptrice, chaque couche lit l'en-tête et le supprime avant de transmettre vers la couche supérieure. Ainsi à la réception, le message est dans son état originel.
À chaque niveau, l'emballage des données change de label : car on lui ajoute un en-tête, ainsi les appellations changent suivant les couches : au niveau de la couche application le paquet de données est un « message », il est encapsulé sous forme de « segment » par la couche transport, une fois encapsulé par la couche transport il prend le nom de « paquet » dans la couche réseau et enfin on parle de « trame » au niveau de la couche liaison (de 64 à1518 octets) et de signal au niveau de la couche physique.

En fonction des exigences de la transmission, des caractéristiques des réseaux et des couches physiques, d'un stockage intermédiaire, le système assure le transcodage et la fabrication des trames en fonction de la méthode d'accès au réseau, la division des messages en trames de bits bruts ou leur regroupement, le contrôle CRC (Cyclical Redundancy Check) des erreurs dans la transmission d'un paquet.

Dans l'internet actuel, c'est un modèle 4 couches TCP/IP (Application, Transport, Internet, Accès réseau), plus souple, qui a été adopté par le marché. Toutes les fonctions des quatre couches supérieures sont considérées comme faisant partie intégrante du protocole applicatif.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

2.5 TIC : l'Intelligence Artificielle vs la connaissance

Le tsunami de l'Intelligence Artificielle est omniprésent dans l'univers de la recherche informatique. Le «machine learning» et le « deep learning » sont des technologies en devenir mais d'un strict point de vue de la science, elles ne sont aujourd'hui ni autonomes, ni auto-apprenantes (ce que le terme d'intelligence implique) !

Cependant, certains développements sont prometteurs sur certaines applications d'analyses de données relatives à des documents textuels (reconnaissance d'entités nommées) et iconographiques (extraction de zones d'intérêts dans des macro-blocs de pixels).

Sur les 30 dernières années la technologie informatique a fortement progressé dans deux domaines :

- la **capacité de produire** d'importants **volumes de données** via de multiples applications. Chaque application constitue un domaine propriétaire (un silo qui contrôle et gère ses propres ressources) et qui n'offre que peu de capacité de dialogue avec d'autres systèmes !
- la **capacité de transporter** ces données à bas prix via des réseaux comme l'Internet ainsi que l'interfaçage entre une application et l'utilisateur (IHM : Interface Homme Machine)

2.5 TIC : l'Intelligence Artificielle va la connaissance

Ce qu'il faut développer, ce sont **des technologies favorisant la structuration explicite et l'échange** de ces données entre des systèmes informatiques hétérogènes (cfr l'interopérabilité) !

De nombreux efforts ont été déployés par les institutions culturelles pour normaliser les métadonnées descriptives, visant à homogénéiser leur structure et à améliorer leur interopérabilité pour leur publication.

Avec l'influence du **Web sémantique**, ces ensembles de métadonnées ont évolué vers des réseaux de connaissances en vue de bénéficier de la complémentarité des différents référentiels.

Le stockage et la structuration données constituent un enjeu majeur pour toutes les institutions/entreprises. L'adoption massive des services cloud et l'augmentation gigantesque du volume des données créées, stockées, analysées, nécessitent des systèmes de plus en plus performants pour prendre des décisions basées sur ces volumes colossaux de données.

2.5 TIC : l'Intelligence Artificielle vs la connaissance

Dans ce cadre l'interconnexion de technologies combinant de l'intelligence humaine et artificielle doit permettre aux différents intervenants de bénéficier du meilleur des technologies actuellement disponibles. Disposer d'une modélisation conceptuelle apte à jouer un rôle de référentiel conceptuel est une nécessité pour la description des contenus de documents ou de ressources! Ce sont les ontologies qui facilitent une compréhension partagée [a shared understanding*] d'un domaine d'intérêt à la fois pour des humains et des machines.

Une ontologie décrit les concepts et les rapports qui sont importants dans un domaine particulier, fournissant un vocabulaire pour ce domaine aussi bien que des spécifications automatisées de la signification des termes utilisés dans le vocabulaire. Les Ontologies traitent de taxonomies et de classifications, schémas de base de données, et de théories entièrement axiomatisées.

Ces dernières années, des ontologies ont été introduites dans beaucoup de communautés scientifiques de manière à partager, réutiliser et traiter la connaissance d'un domaine spécifique.

*Article publié en 1996 par M. Uschold et M. Gruninger.

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 fragment** : un document publié par éléments formant chacun un sous-ensemble
- L**'interopérabilité** est la capacité que possède un système, dont les interfaces sont explicites, à fonctionner et échanger avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre. L'interopérabilité est une nécessité dans le domaine de l'archivage. En effet, le demandeur de services et le fournisseur de services doivent pouvoir interagir dans l'espace et dans le temps en employant des représentations de données structurées et normalisées.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

3.1 Numérisation des media : le rôle de l'applicatif

Logiciel et application sont des outils informatiques qui ont pour but d'aider un utilisateur à exécuter une action (paramétrages ou programmation).

- **Le Logiciel** est constitué d'un ensemble de programmes, procédés et règles ainsi que de la documentation, regroupés sous forme de fichiers dans une mémoire, destinés à être installés sur des PC et fonctionnant avec des applications intitulées *.exe.
- **Application** : depuis la généralisation des périphériques mobiles, des nouveaux logiciels plus agiles ont été créés pour s'accommoder aux récents développements des périphériques et portent le nom d'Application.

Un Logiciel couvre donc un champ sémantique plus large que l'application facilite l'action d'un utilisateur pour une action précise.

- **Système d'exploitation** : c'est le logiciel de base (souvent appelé OS — de l'anglais Operating System) qui assure la communication entre le processeur, les périphériques et l'utilisateur. Il permet de gérer le matériel et les autres logiciels. Windows 10, Mac OS, Linux, Android, iOS. … sont les systèmes d'exploitation les plus connus et les plus utilisés. Chaque système d'exploitation a un impact sur l'utilisation des logiciels. En effet, certains logiciels ne peuvent être utilisés que sur Windows ou Android. Cela impacte également la migration de données dans l'espace (l'échange) et surtout dans le temps.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

3.2 Définition d'un fichier :

- Un **fichier** informatique est l'unité de stockage de données communes à tout ordinateur. Toutes les applications génèrent des données structurées dans un fichier fermé, et l'OS de l'ordinateur écrit, lit, stocke, importe et exporte des fichiers numériques. Un fichier informatique constitue donc un ensemble de données sous format binaire réunies sous un même label, enregistrées sur un support de stockage permanent (mémoire de masse : un disque dur, un CD-ROM, une mémoire flash ou une bande magnétique), et manipulées comme un tout.
- La **syntaxe des titres des fichiers** comporte un label (généré par l'usager) et une "extension" courte figurant après le point séparateur :
  - le titre (label) proprement dit : les lettres de A à Z et les chiffres de 0 à 9
  - l'extension facultative (en général 3 caractères) qui synthétise l'application utilisée.
- **Les formats de fichiers** :
  - Les formats de fichiers décrivent l'organisation et dépendent des processus de codage/décodage à appliquer aux données du dit fichier pour la relecture.
  - Les fichiers sont souvent désignés en fonction du codage (et donc du décodage). Par exemple, *.jpg est une image JPEG, *.doc est un fichier texte codé par l'application traitement de texte « Word » de Microsoft et windows7.exe est une application exécutable dans l'OS Windows.

3.2 Définition d’un dossier

Chaque outil dispose d’une zone de stockage, intitulée **Dossier**, dans laquelle sont rangés les différents Fichiers que l'on va trouver sur un périphérique numérique. Ces dossiers, également appelés "répertoires", sont créés sur le disque dur lorsque le système d'exploitation et les applications sont installés. Pour ce faire, il existe un dossier racine qui est à la source de tous les autres (arborescence), symbolisé par le signe \ (backslash sous Windows ou / (slash) sous Linux, et automatiquement créé lors du formatage du disque dur de stockage.

Le bureau de l'utilisateur est en réalité un dossier. Certains fichiers sont parfois des dossiers. Bien qu'ils ne soient pas identifiés comme tels, un seul fichier peut en réalité être un dossier.

Un dossier contient un ou plusieurs fichiers (ou peut être vide) avec juste un titre. Contrairement aux fichiers, un dossier peut contenir à la fois des fichiers et d'autres dossiers.

L’encapsulation des données en mode fichier/dossier structure l’interface graphique d’un PC et organise l’enregistrement des données. Il existe d’autres formes d’encapsulation comme le ZIP (qui diminue la taille), METS, MXF, … et bien sûr les bases de données.

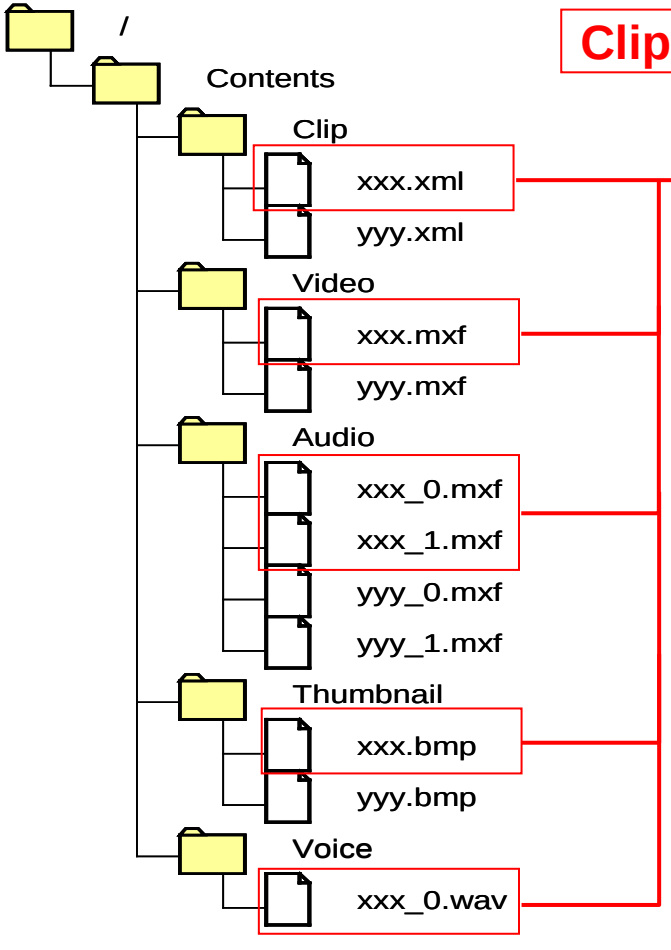
3.2 L’encapsulation des données :

L'encapsulation de données est le processus qui ajoute aux données des informations d'en-tête de protocole supplémentaires avant leur transmission. Pour les encapsulations, on utilise en général des termes comme wrapping ou boxing qui décrivent le concept "d'emballage" d'un élément relativement "primitif".
Les données enveloppées à chaque couche portent le nom de **PDU** (Protocol Data Unit) et contiennent deux choses : la donnée en elle-même et l’en-tête spécifique à cette couche ainsi que les en-têtes des couches précédentes. L'encapsulation permet de définir des droits d'accès aux données : publique, privée ou protégée.

METS (Metadata Encoding and Transmission Standard) est une norme de codage des métadonnées proposée par la LOC (Library of Congress) et fournissant un format de document XML pour la gestion des objets de bibliothèque numérique dans un référentiel et l'échange de ces objets entre les référentiels (ou entre les référentiels et leurs utilisateurs).
METS définit une structure hiérarchique :

- <structMap> code la hiérarchie comme une série imbriquée d'éléments .
- chaque contient des informations d'attribut spécifiant le type de division
- des pointeurs METS (<mptr>) pointeur de fichier (<fptr>) peuvent identifier le contenu correspondant à une .
- la section <fileSec> spécifie des fichiers, des groupes de fichiers ou des emplacements spécifiques dans un fichier.

3.2 L'encapsulation: l'exemple du MXF



Un exemple d'encapsulation de « Fichiers/Dossiers » issu du monde de l'audiovisuel : le **MXF (M**aterial e**X**change **F**ormat) OP ATOM est un conteneur défini par la SMPTE et utilisé par l'industrie audiovisuelle pour structurer les données (fichiers vidéo, audio, imagettes, ...). D'autres OP «modèles opérationnels» peuvent contenir ou référencer plusieurs matériaux, tout comme une EdL (Edit list : chronologie d'un programme de montage vidéo).

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

3.3 La saga des met@data !!!!

Depuis la nuit des temps, les bibliothèques, les archives ou plus tardivement les médiathèques ont acquis une longue pratique de la description et de l'indexation des contenus et des documents qu'ils manipulent. Les fiches cartonnées imaginées par Paul Otlet ont été normalisées en 1954 (sous la référence **ISBD** : International Standard Bibliographic Description).

L'informatisation de ces notices bibliographiques date des années 60 et elles ont été normalisées par l'ISO (le format **MARC** [**MA**chine-**R**eadable **C**ataloging] utilisant la norme ISO 2709).

Les formats MARC sont implémentés dans les logiciels SIGB (Système Intégré de Gestion de Bibliothèque). Elles assurent la gestion interne des ressources documentaires et, du côté usagers, elles permettent d'optimiser la recherche et la localisation des documents.

Ces formats MARC ont une structuration fine et détaillée sur deux niveaux (zone et sous-zone) et ils ont été adoptés pour gérer et localiser des documents électroniques.

3.3 La saga des met@data !!!!

En 1988, la **LOC** (Bibliothèque du Congrès des États-Unis), proposait un protocole client-serveur (le Z39.50) pour rechercher à travers un réseau informatique des informations dans des bases de données et interroger simultanément plusieurs catalogues.

Afin de rendre les bases de données bibliographiques plus accessibles aux usagers aussi bien dans ou hors les murs des bibliothèques la LOC à conçu **« BIBFRAME »** dans les années 2010 pour remplacer le format MARC en se basant sur le web de données.

Les métadonnées sont **des données contextualisées**. Elles doivent répondre aux interrogations tant humaines que machines aux questions « qui, quoi, où, pourquoi, quand et comment » 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.

3.3 La saga des met@data !!!!

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.
Types de métadonnées :

- Les métadonnées **descriptives** : elles 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 métadonnées **techniques** : elles décrivent le format d'un fichier, la structure d'un jeu de données et les informations liées au stockage.
- Les métadonnées **opérationnelles** : elles 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 métadonnées **business** : elles 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, ...
- (caractère performatif .....)

3.3 La saga des met@data !!!!

La gestion des métadonnées 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 données entre plateformes.
- pour les utilisateurs : elles permettent des recherche/navigation plus performante.

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

Il existe de nombreux modèles de métadonnées qui peuvent être écrits selon plusieurs standards : syntaxe " meta " HTML et Dublin Core, XML, DTD EAD (Encoding archival description), TEI (Text Encoding Initiative), RDF (Resource Description Framework), etc.

Voir : http://preservationtutorial.library.cornell.edu/tutorial-french/metadata/table5-1.html

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.4 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

3.4 Les métadonnées ancillaires des fichiers :

Le fait d'inclure les propriétés d'un document (via un modèle de métadonnées) 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.

3.4 Les métadonnées ancillaires des fichiers :

Les métadonnées 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, et
- **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.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

4.1 Transformation numérique : le traitement de textes

La numérisation d'un texte est sa transformation en une suite de caractères prises 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.

4.1 Transformation numérique : le traitement de textes

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 alphanumérique et des instructions de fonctions.

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 complet est illustré dans la figure qui suit.

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]	56	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	[SYNCHRONUS 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	¨				

ASCII American Standard Code for Information Interchange

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

numéro 48, le chiffre 0, valeur 60, soit 110000 en binaire,

numéro 65, la lettre A majuscule, valeur 101, soit 1000001 en binaire).

Les caractères de numéro 0 à 31 et le 127 correspondent à des commandes de contrôle du terminal informatique. Le numéro 32 est l'espacement, 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 !

4.1 Normalisation du traitement de textes

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 UTP (déclinée en **UTP-8** ; **UTP-16** et **UTP-32**). Non seulement les caractères 'occidentaux' sont représentables mais aussi les caractères arabes, sanscrit, cyrillique, les idéogrammes (chinois, japonais, ...).

4.1 Traitement de texte : la structuration fichier/dossier

À partir de Microsoft Office 2007, **Microsoft Office** utilise les formats de fichier en XML, tels que *.docx, *.xlsx et *.pptx. En réalité, ce sont des ZIP (un format de fichier permettant l'archivage et la compression de données sans perte de qualité) contenant des dossiers et des fichiers. Le format **Open XML** utilise la technologie de compression zip pour le stockage de documents, ce qui permet de réduire le coût d'utilisation des fichiers par courrier électronique, par réseaux et sur Internet ... Inutile d'installer un utilitaire zip pour ouvrir ou fermer des fichiers dans Office. Les fichiers sont structurés de manière modulaire permettant de conserver différents composants de données dans le fichier séparés les uns des autres. Ce qui permet d'ouvrir les fichiers même si un composant au sein du fichier (par exemple, un graphique ou un tableau) est endommagé.
Le logiciel de traitement de texte **OpenOffice** (ou StarOffice de Sun Microsystems) crée des fichiers avec une extension *.odt. Ce logiciel, est similaire à d'autres formats de fichiers documents texte. C'est le fichier XML contenu dans un wrapper ZIP qui structure un fichier *.odt. Ces fichiers peuvent contenir du texte, des images. Des tableaux ainsi que divers objets numériques généralement utilisés pour divers modèles de documents peuvent également être insérés ou intégrés dans les documents enregistrés au format *.odt.
Sur le **Mac**, une application a une extension APP, et ce qui semble n'être un seul fichier est en fait un dossier (voir fichier APP).

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

4.2 La transformation numérique des media

La transformation de l'audio, de la photographie ou de la vidéo comporte en général au moins trois phases :

- **une phase d'échantillonnage** où des caractéristiques d'un document (par exemple, le temps pour l'audio, l'espace pour la photo, ...) sont explorées à intervalles temps régulières ;
- **une phase de quantification** qui attribue une valeur représentant l'objet aux points d'échantillonnage (résolution pour le son, pixel pour l'image)
- **une phase de numérisation** qui confère à chacune de ces valeurs une valeur numérique (par exemple sous forme binaire) similaire au système de codage du texte !

L'ensemble de ces traitements effectués par un appareil de transformation numérique (Numériseur) constituent un fichier de données successives qui va stocker l'ensemble des données numérisées relatives à un document, un objet, un son, ...

La transformation binaire (transformation des signes interprétables par l'ordinateur) est la primitive du langage de l'ordinateur et le format de fichier (typé e.g. txt, docx, Jpg, mp4, Wav, ...) une forme d'organisation du fichier en fonction de codage d'un objet (texte/ASCII, image/JPEG, video MPEG/,...).

4.2 Codage – décodage - Transcodage

- **Le Codage** : après une phase de numérisation, il est possible d'appliquer un codage visant à réduire le nombre de bits (0 ou 1) présents dans cette suite de nombre.
  - ❑ Le Codage peut être défini sans perte (ZIP, TAR, WAV, RAW, NEF, DPX, ...)
  - ❑ Ou encore avec perte (MP3, JPEG, MPEG, HEVC, ...) mais en fonction de paramètres appliqués avec ou sans perception de cette perte d'information.
- **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).
- **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. Le transcodage d'une vidéo codée en mode entrelacé (une trame paire, une impaire) en progressif (pour un affichage sur un écran informatique) est susceptible de générer un effet de peigne ! Un transcodage d'un fichier généré avec perte ne peut regénérer le format d'origine. Le transcodage est souvent synonyme de pertes (notamment destructeur pour les métadonnées ancillaires).

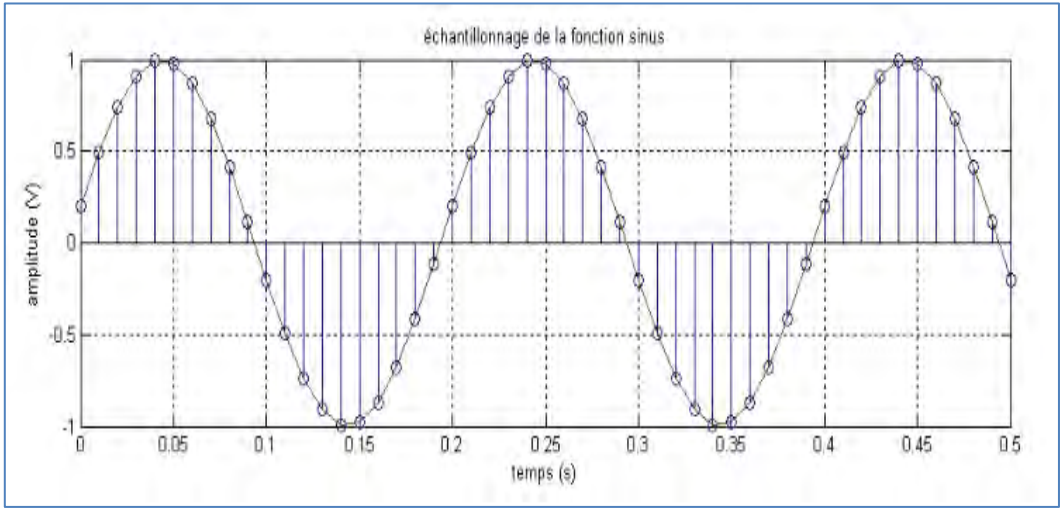
Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

4.2 La transformation numérique du son

Dans un codage numérique de l'audio, la forme d'onde d'origine est divisée en instantanés individuels appelés « échantillons » :



- La fréquence d'échantillonnage correspond au nombre d'instantanés d'un signal audio pris par seconde. La fréquence d'échantillonnage doit être supérieure au double de la fréquence maximale à reproduire.
- La résolution détermine la plage de fréquences d'un fichier audio (nombre de valeurs d'amplitude).

Plus la résolution (nombre de bits/échantillon) et la fréquence d'échantillonnage sont élevées, plus le profil de la forme d'onde numérique sera proche de celle de la forme d'onde analogique d'origine, ce qui augmente la dynamique et la fidélité tout en réduisant le bruit de fond.

4.2 La transformation numérique du son

De nombreux formats audio sont basés sur la norme **RIFF** (Resource Interchange File Format). Mis au point par Microsoft et IBM en 1991, le format de fichier **WAV**, est constitué d'un petit en-tête indiquant le taux d'échantillonnage et la résolution en bits, puis d'une longue série de nombres, un pour chaque échantillon. Par exemple, à raison de 44 100 échantillons par seconde et de 16 bits par échantillon, un fichier mono nécessite un espace de 86 Ko par seconde, soit environ 5 Mo par minute.

L'**UER/EBU** (European Broadcasting Union) a défini en 1997 une extension broadcast du format WAVE à l'usage des professionnels, le **BWF** « Broadcast Wave Format » permettant notamment l'ajout de métadonnées « broadcast » comme le time-code, des informations d'identification, ou encore de mesure audio.

De nombreux formats audio existent : **MP3** (MPEG-2 Audio Layer III), **Ogg**, **AIFF** (Audio Interchange File Format), **CAF** (Core Audio Format), **CDA** (Compact Disc Audio), **FLAC** (Free Lossless Audio Codec) , **RAW** (Real Audio Wrapper).

Le type de format correspond à l'extension du fichier (c.-à-d. les lettres qui se trouvent après le point dans le nom du fichier, par exemple .mp3, .wav, .ogg, .wma).

4.2 Exemple : le codage son vs la substance

« Angèle « a sorti un album intitulé « BROL » de 12 pour un total d'écoute de 41 minutes :

1. Il peut être acheté sous la forme d'un « Boitier CD-Audio » (une boite plastique contenant un disque optique, deux couvertures papier face/dos et une pochette). Les données sur le disque sont codées suivant le standard industriel https://fr.wikipedia.org/wiki/Red_Book. Basé sur un échantillonnage à 44,10 kHz à 16 bits, il n'a pas de structure « fichier ». Il est muni de codes correcteurs d'erreur et d'un système d'inclusion de métadonnées culturelles et techniques. Volume : environ 380 Mo pour 41 minutes).
2. Génération d'un fichier *.wav du contenu musical : un codage à la fréquence d'échantillonnage de 44,1 kHz, se fait sans perte de substance, ni de réduction de taille. La taille du fichier *.wav est de 415 Mo [109 % de la source].
3. Le codage en*.bwf (pour inclure des métadonnées supplémentaires : time code, informations de qualité, de traçabilité, play-list, ..) génère 428 Mo [113% de la source].
4. Le codage en *.flac génére une représentation « sans-perte » de substance de taille de 250 Mo [66% de la source].
5. La génération d'un fichier MP3, en qualité 'moyenne', entraîne une compression des données d'un facteur 8 (soit 49 Mo [13% de la source], avec évidemment une perte de substance.

Un CD-A permet de représenter 79,8 minutes maximum sur 740 Mo : soit 9,273 Mo par minutes

4.2 Exemple : le codage son vs la substance

La génération d'un fichier *.wav ou *.bwf au départ du CD-A « BROL », avec une fréquence d'échantillonnage audio autre que 44,1 kHz procure un résultat dont la qualité est directement affectée par la qualité du générateur et par la fréquence d'échantillonnage choisie.
Une qualité professionnelle est usuellement définie par un lissage par interpolation d'un facteur de 8 au moins; puis par la génération d'une représentation caractérisée par 24 bits par échantillon à 96 kHz. Cela signifie 327% plus de bits par minute.
Toute transformation de ce type, si elle est bien faite conduit à une perte de substance très minime, tout en apportant des avantages indirects nombreux. Par contre, si le lissage est mal fait, la qualité obtenue peut s'avérer désastreuse.

Du point de vue des studios d'enregistrement et de production, la substance de référence est définie par le format choisi : souvent de 32 bits par échantillon à 192 kHz.

Dans ce cas, le CD-A représente une perte de substance raisonnable car considérée comme acceptable pour une écoute privée ; a fortiori, le MP3 est considéré comme acceptable pour une écoute dans un environnement bruyant ou pour des auditeurs peu exigeants.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

4.3 La transformation numérique des images :

Le premier scanner numérique date de 1957 ! Mais il faut attendre la fin des années 80 et le développement du **CCD** (Charged Coupled Device) afin de disposer d'un capteur photographique. Ce composant électronique photosensible convertit un flux lumineux en un signal électrique analogique. Ce signal est ensuite amplifié, puis échantillonné et enfin numérisé par un convertisseur analogique-numérique en vue d'obtenir une représentation numérique de l'ensemble des points (Pixels) de l'image et une quantification.

Le signal de base issu du CCD est fragmenté en points appelés « pixels », abréviation de « picture » (image) et de « el(ement) » (« élément »). L'ensemble des pixels, reliés entre eux, constituent la résolution de l'image. Plus il y a de pixels, en horizontal et vertical, meilleure est la résolution de l'image.
Un pixel est calculé par l'assemblage électronique de trois "points": un rouge + un vert + un bleu. La couleur est produite par l'addition de ces 3 couleurs à des intensités variables, ce qui permet de produire toute une palette de couleurs différentes.

Le premier appareil photo numérique aurait été développé par KODAK (mais qui n'a jamais développé le concept) https://fr.wikipedia.org/wiki/Appareil_photographique_numerique)

Le premier appareil photo numérique **Model 1** a été présenté en 1991 par la société Dycam lors du CeBIT à Hanovre. A l'époque les prises de vue étaient en noir et blanc avec une résolution minimaliste de 376 x 284, mais cela générait déjà un volume de 854.272 bits par photo

4.3 La transformation numérique des images :

Le codage d'une image noir ou blanc est aisé ! Pour scanner un texte, le pixel ne peut uniquement prendre que les valeurs noir ou blanc et donc 0 et 1.
Le codage de la couleur est réalisé sur trois octets, afin de représenter la gradation de la valeur d'une composante couleur de 0 à 256. Ces trois valeurs codent généralement la couleur dans l'espace RVB.
Le nombre de couleurs différentes pouvant être ainsi représenté est de 256 × 256 × 256 possibilités, soit environ 16,7 millions de couleurs.
Avec le temps des quantification sur 10 ou 12 bits sont devenues courantes !
Dans les logiciels d'édition d'image, dans l'incrustation vidéo, il est intéressant de séparer les images et les fonds. C'est également important pour pouvoir représenter des images en réflection ou en transmission lumineuse. Pour que ce soit possible, il faut indiquer, à chaque pixel, le degré de transparence. L'attribution d'une valeur de transparence, sur un octet, permet tous les mélanges. Cette valeur est généralement appelée α (alpha). Le rendu d'un pixel s'obtient par multiplication et addition, récursive, des valeurs des couches des codages sur 24 bits (Paint : 256 valeurs x 4) ou sur 32 bits (pour les logiciels Gimp, Photoshop).

4.3 La transformation numérique des images :

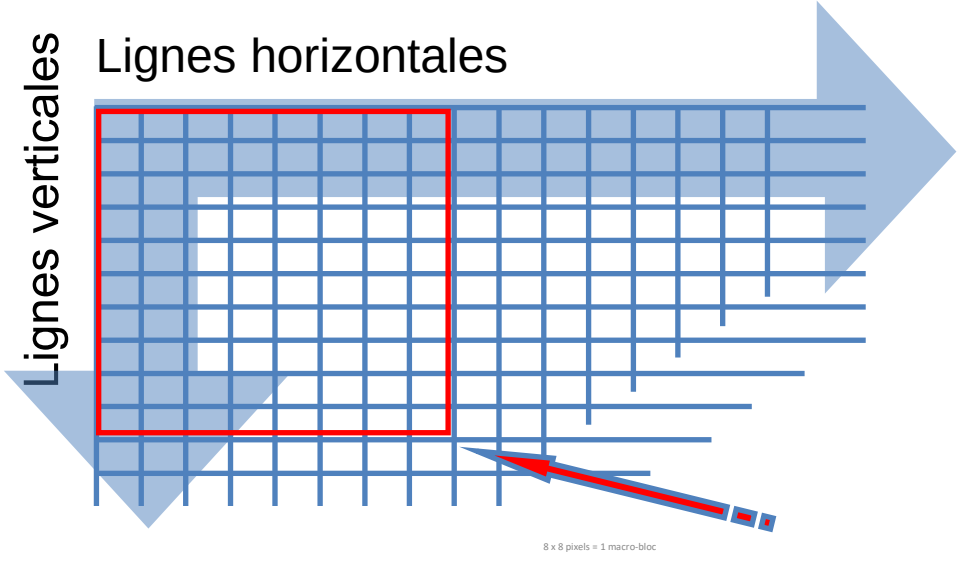
Il existe deux systèmes de codage des couleurs : primaires **RVB** (Red Green Blue en anglais) et **YUV** basé sur les trois longueurs d'ondes auxquelles l'œil humain est réceptif : surtout le Vert, puis le rouge, et enfin le bleu.

- dans **le système RVB**, on utilise la synthèse additive des trois couleurs : l'addition des trois couleurs donne du blanc, le rouge et le vert donnent du jaune, le vert et le bleu du cyan et le bleu et le rouge du magenta. La couleur résultante dépend de l'intensité de chaque couleur primaire. Ce système RVB est utilisé par les capteurs CCD et Cmos qui équipent les appareils photo et les caméscopes numériques.
- le **système de codage YUV** est créé depuis une source RVB. Il est codé en trois composantes : la première (Y) représente la luminance (informations de luminosité) tandis que les deux autres (U et V) sont des données de chrominance (informations de couleur). Chaque composant est formaté en fonction d'**une table de codage**. Exemple : la luminance (Y) est égale à l'addition de 0,58 V + 0,29 R + 0,11 B (soit 60 % de vert)

A l'origine ce codage YUV permettait d'économiser de la bande passante par rapport au codage RVB (qui transportait les informations de luminance pour chaque canal de couleur) ainsi que d'assurer la compatibilité entre les différents récepteurs analogiques de la télévision. Durant les recherches sur les technologies numérique il est apparu que ce modèle colorimétrique permettait en plus de réduire la taille d'une image via une compression. Comme l'œil humain est plus sensible à la luminance (définition) qu'à la chrominance, ce codage permet de dégrader plus fortement la chrominance d'une image tout en assurant un niveau de qualité.

4.3 Numérisation des media : le codage des images

A titre d'exemple, pour réduire le volume de données à stocker ou à transporter, le codage numérique des images JPEG utilise le principe de la transformation en cosinus discrète (DCT).

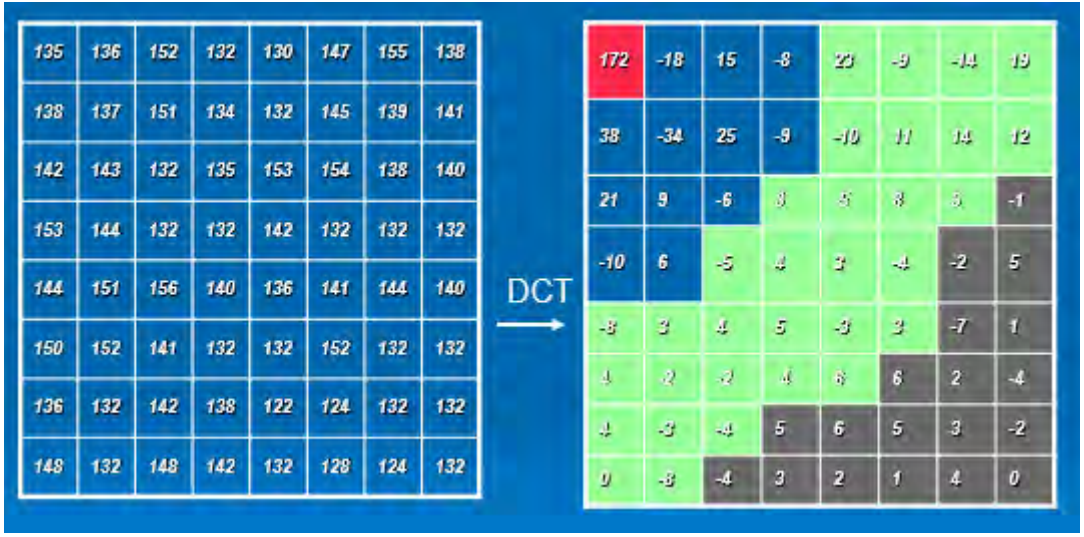


Afin de mettre en évidence des zones à forte redondance spatiale, l'image est découpée en blocs de 8 pixels sur 8 pixels. Ce concept de macro-bloc est la base de calcul pour la DCT. La taille de 8 X 8 pixels était un bon compromis à l'époque car elle faisait appel à des puissances de calcul raisonnables.

Afin de procéder à une étude de la redondance spatiale (la similarité entre les pixels adjacents aussi bien en horizontal qu'en vertical) la transformée en cosinus discrète DCT permet de passer d'une représentation spatiale de l'information (chaque nombre représentant la valeur de l'intensité d'un pixel] à une représentation fréquentielle de l'Information (un coefficient représentant des variations d'intensité lumineuses ou colorimétriques(via la table de codage Y'CbCr qui est utilisé pour les images JPEG).

4.3 Numérisation des media : le codage des images

À chacun de ces macro-blocs est ensuite appliquée cette transformation qui remplace les données des 8 x 8 pixels par 64 autres coefficients ! On s'aperçoit ainsi que les 64 pixels de base ne portent pas tous une information de même importance :



La DCT applique des coefficients afin de réduire les écarts de niveaux dans les hautes fréquences où l'œil est moins sensible. Cette réorganisation s'effectue par ordre croissant, les détails les plus fins étant situés en bas et à droite de la matrice, la première valeur en haut à gauche, représentant la valeur calculée de la matrice.

Les pixels de base en haut à gauche sont essentiels, et ne pas les prendre en compte se traduirait par une nette différence entre un macro-bloc source et un macro-bloc reconstitué. En revanche, négliger les valeurs qui figurent en bas à droite n'influent pas énormément sur l'aspect du macro-bloc reconstitué.

4.3 La transformation numérique des images :

La compression JPEG ne sélectionne qu'une partie bien choisie de l'information traduite par les coefficients de la DCT. Ce sont des critères d'efficacité qui déterminent le fait que certains coefficients sont en partie ou complètement négligés et qui génèrent une compression numérique avec plus ou moins de pertes (surtout en chrominance).

En 1992, l'ISO a normalisé le **JPEG** (fruit du Joint Photographic) qui a défini et normalisé e format de codage des images numériques.

JPEG définit deux classes de processus de compression :

- **sans pertes ou compression réversible** : I n'y a pas de pertes d'information et il est donc possible de revenir aux valeurs originales de l'image. Elle fait l'objet d'une **norme spécifique appelée JPEG-LS.**
- **avec pertes ou compression irréversible** : c'est le JPEG « classique » qui permet des taux de compression de 3 à 100 (réduction du débit binaire à la sortie du codage et donc du volume de données à stocker). Ce genre de codage utilise de la DCT (Discrete Cosine Transform) sur des macro-blocs afin de permettre la suppression des coefficients nuls ou proches de zéro (et donc réduire le volume de données à conserver).

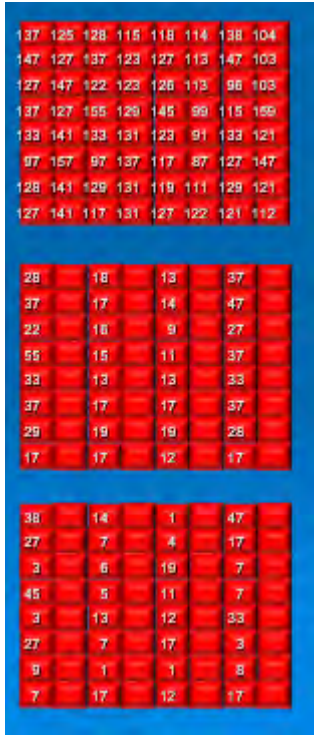
Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

4.4 Codage vidéo : l'échantillonnage des couleurs

Si d'une manière générale la vision humaine présente une sensibilité moindre aux couleurs qu'à la définition, c'est évidemment encore plus vrai dans le cadre d'images animées ! En vidéo il est donc possible de conserver moins de données de chrominance que de luminance. Il est donc possible de sous-échantillonner la chrominance, ce qui correspond à un besoin de réduire le poids du signal pour l'enregistrement sur disque ou carte compact flash.



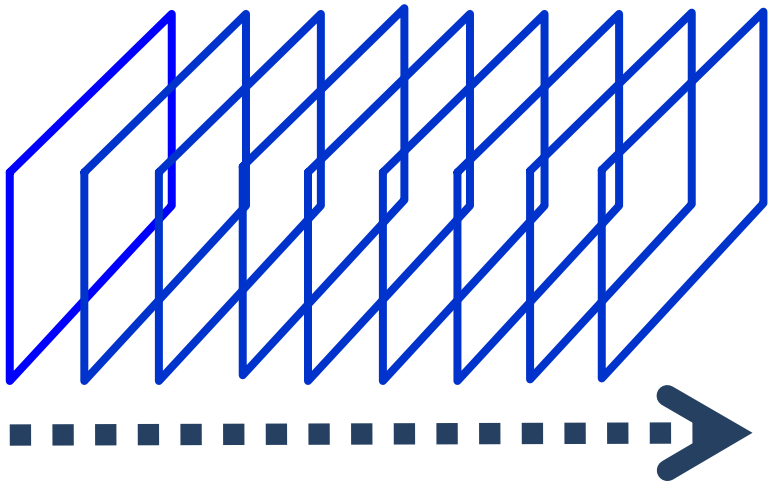
Pour les besoins de l'échantillonnage, le signal est découpé en tranches temporelles de 4 pixels de référence. Dans ce cas l'échantillonnage est réputé 4.4.4, soit une information complète en luminance et en chrominance.

Le sous-échantillonnage de la chrominance est une méthode de réduction de volume de données contenues dans des images numériques. Dans tous les cas, l'information de luminance sera entièrement conservée car elle est indispensable pour une bonne perception de l'image enregistrée par notre œil et comporte donc les 4 valeurs captées. Pour l'échantillonnage de la chrominance, seule la moitié des valeurs seront conservées (en mode 4.2.2 : voir illustration) ou le quart (4.2.0.)

4.4 Codage vidéo : Compression intra – inter image

La **compression** vidéo est une méthode de **compression** de données, qui consiste à réduire le volume d'une seule et même image (codage **intra-image identique au JPEG**) ou des pixels entre images voisines (utilisant des techniques **inter-image**).

La compression vidéo intra-image (prédiction spatiale) traite chaque image séparément. L'image est divisée en zones (macros-blocks), qui sont en général de 8 pixels sur 8 pixels pour un traitement DCT.



Avantage :

- Pas de temps de traitement : la DCT travaille chaque bloc de 8 X 8 pixels indépendamment.

Désavantage :

- La réduction de débit est faible

Formats utilisant la compression intra-image : DV-AVI, Mjpeg, AVC-intra, DVCPROHD.

La compression temporelle (prédictions inter-images) s'effectue par GOP (Group of Pictures).

https://www.gypsevideo.fr/tout-savoir-sur-la-video/les-techniques/codecs

4.4 Codage vidéo : le traitement de l'AV : le MPEG 2

Les réunions à l'ISO du groupe MPEG (Moving Picture Experts Group : des spécialistes provenant de l'industrie de l'électronique des composants, de l'informatique et des télécommunications) ont démarré en 1988 dans le but de développer une première norme ISO MPEG-1 pour des applications de stockage audio/vidéo du type Vidéo CD.

En novembre 1994, l'ISO a publié la norme MPEG-2 (Moving Picture Experts Group) faisant suite à MPEG-1.

MPEG-2 définit tous les aspects du codage source de l'image et du son et ainsi que transport (TS : Transport stream) à travers des réseaux de télévision numérique ou la production de DVD (PS Packet Stream).

La caractéristique de la famille MPEG est basée sur l'usage du GOP (Group of Pictures). Cette caractéristique engendre une latence au codage/décodage (plusieurs secondes dépendant de la longueur de GOP).

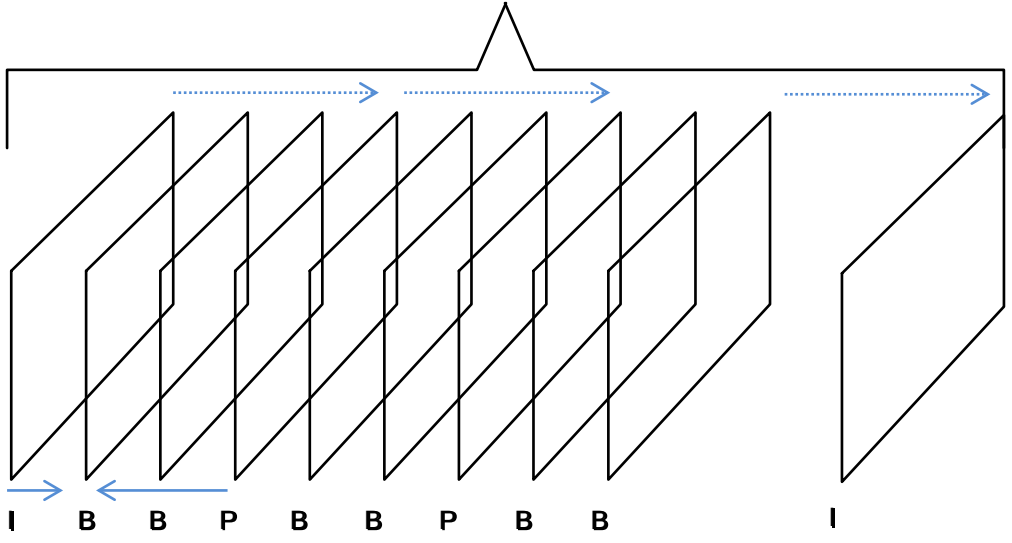
4.4 Codage vidéo : le traitement de l'AV : le MPEG 2

Le film, la vidéo combinent des dimensions à la fois spatiales (images) qui évoluent dans le temps (un nombre d'images par seconde reproduisant un mouvement).

I Intracoded frame

P Forward/predicted frame

B Bi-directional/interpolated frame

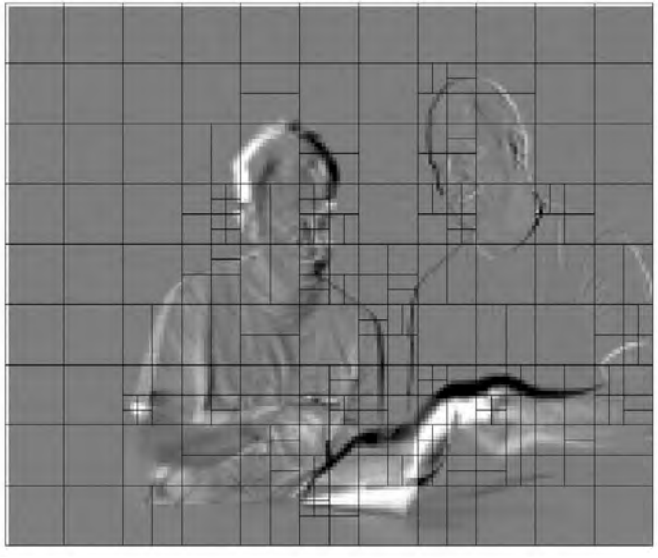


Group of pictures GOP : 12 - 16

Afin d'optimiser le codage des images, le codec découpe le flux en séquences de l'ordre de 12 à 16 images (ou plus Gop : Group of Pictures) en déterminant un image mère (I = I Frame codée en mode intra) et des images P (prédites ; calculée par différence avec les images précédentes) ou B (calculée en fonction des images qui précédent ou suivent) afin de compresser au maximum le débit à transmettre ou à stocker.

4.4 Le traitement de l'AV : du MPEG-2 au MPEG-4

Dans l'évolution MPEG-4 AVC, des macro-blocs de tailles plus petites (4 X 4) et plus grandes (16 X 16) du fait des puissances de calcul disponibles depuis, plus importante.



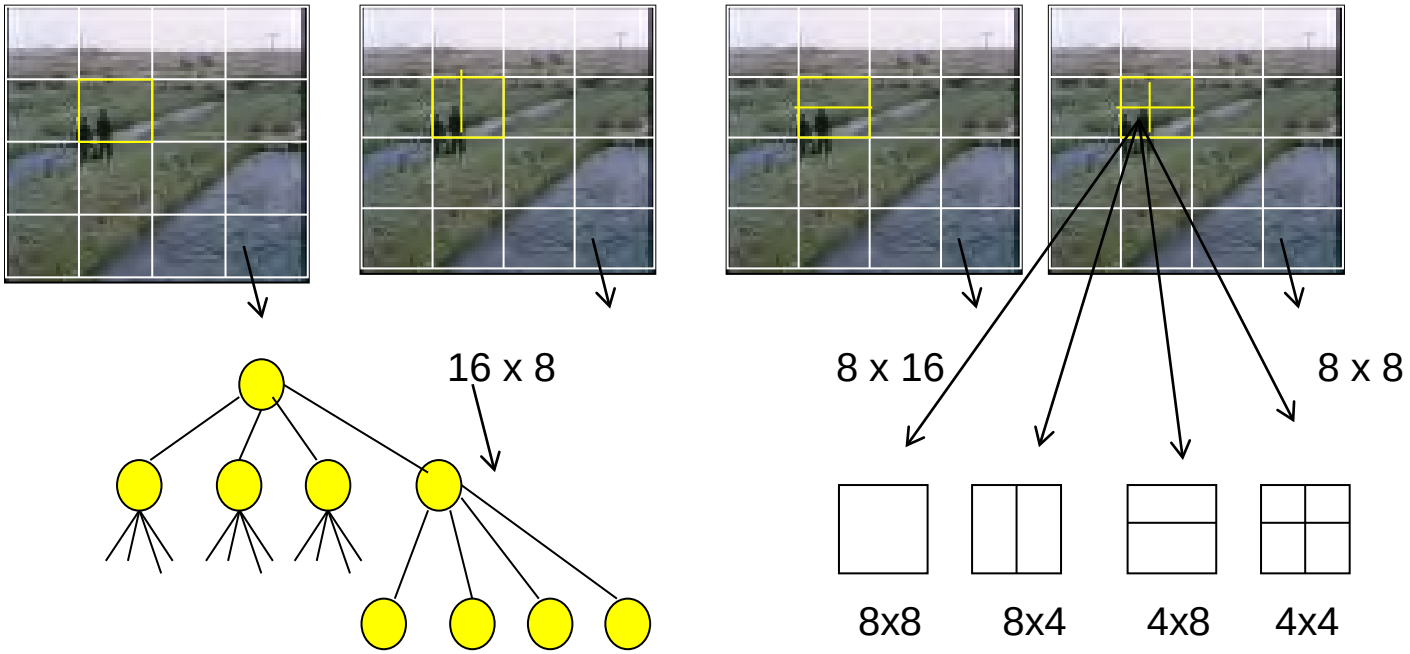
From Iain E.G. Richardson : H.264 and MPEG-4, WIley, 2003

Par rapport aux macro-blocs de taille fixe, cette méthode présente l'intérêt de mieux s'adapter au contenu, en traitant très différemment les aplats des zones les plus détaillées. Elle est en outre nettement plus propice au traitement parallélisé, tel qu'avec un Graphics Processing Unit (GPU) et offre plus de possibilité pour des extractions sémantiques.

4.4 Le codage vidéo : la prédiction du mouvement

Ces différents macro-blocs constitueront la base pour le calcul de l'estimation du mouvement déterminant le vecteur de mouvement. En MPEG-4, qui permet de coder des objets, il existe trois types de VOP (Vidéo Object Plane); des I-VOP (Intra-coded) sans compensation de mouvement, des P-VOP I (Predicted) prédits à partir des derniers I-VOP ou P-VOP, et des B-VOP (Bidirectionnaly predicted) prédits à partir des plus récents I-VOP ou P-VOP antérieur et/ou postérieur qui contribuent efficacement à la réduction de données à stocker et à transmettre.

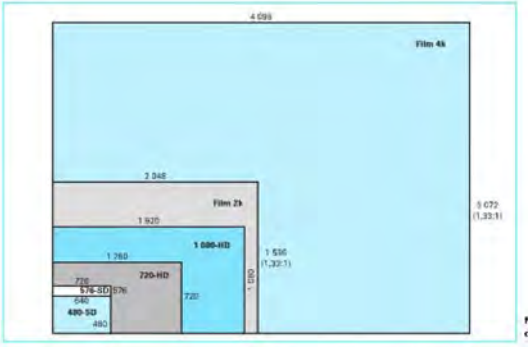
Estimation de prédiction de mouvements en MPEG 4/H.264



Tree-structured motion compensation

Finest partition is a complete quad-tree : all blocs of 4x4 pixels

4.4 Codage vidéo : la naissance de la HD



L'histoire de la représentation par l'image par la photographie, le cinéma, la télévision est déjà longue et ses usages multiples. La taille moyenne des écrans TV était de 12'' (30cm) en 1950 en format 4/3; elle est aujourd'hui de 50'' (127 cm) en format 16/9. Elle a donc quadruplé en 60 ans.

Le passage d'une diffusion analogique à une transmission numérique généralisée a modifié le marché du brun ! En une dizaine d'années, les écrans plats ont complètement remplacé les bons vieux postes de télévision, avec des nouvelles normes « HD Ready », puis « Full HD ». L'image en Haute Définition repose sur une résolution d'écran de 720p (définition de 1280 x 720 pixels), le Full HD sur 1080p (1920 x 1080 pixels). HD et Full HD reposent sur la norme de codage vidéo MPEG-4 AVC ou H.264/AVC, que l'on retrouve également dans le format des disques Blu-ray. Le cinéma numérique a adopté le 2K avec un codage reposant sur le JPEG2K.

4.4 Le codage de la vidéo : le futur de l'Ultra HD

Aujourd'hui ce sont les téléviseurs 4K (Ultra HD) sont les « futurs » nouveaux rois de la TV haut de gamme avec une résolution d'image numérique de 3840 × 2160 pixels, soit environ 8 millions de pixels. Cela signifie que l'image est plus riche en détails, … et offre ainsi une expérience immersive sans devoir augmenter la distance entre l'écran et le spectateur.

Place donc à la norme H.265/HEVC (High Efficiency Video Coding), capable de supporter jusqu'à une définition 8K (8.192 x 4.320 pixels). Grâce à un processeur plus puissant et des algorithmes complexes, le HEVC inaugure une nouvelle méthode de découpage des trames «Coded Tree Unit» variable (trois tailles) et sur trois niveaux avec des gains de compression deux fois supérieurs.

Depuis quelques années, la télévision japonaise (NHK) présente les évolutions d'un projet dénommé « Super Hi Vision », qui propose d'augmenter la résolution à 7680 pixels sur 4320 lignes, soit une définition 16 fois supérieure à celle de la télévision Full HD. Conçue pour être visionnée à une distance de 0,75 de la hauteur de l'écran et pour remplir 110° du champ visuel, cette expérience procure un réalisme incroyable !

Par ailleurs, la BBC a démontré récemment que le système de balayage progressif à 50 images/sec n'était pas suffisant pour éliminer tous les artéfacts dans les mouvements rapides. Elle propose d'y remédier en augmentant la fréquence d'image à 300 images progressives/sec.

4.4 Autre format : le codage DWT

D’autres normes de codage sont apparues qui elles sont intra uniquement (JPEG2K ou JPEG-XS) basés sur une autre technologie (Ondelettes) :

Le JPEG 2K qui utilise la DWT(Transformée en Ondelettes Discrète) pour le codage des images ! Le codec cible « des zones d'intérêt » qui offrent plusieurs mécanismes pour prendre en charge l'accès aléatoire spatial ou l'accès à une zone d'intérêt à divers degrés de granularité. Ces normes présentent une caractéristique qui permet un décodage hiérarchique (en fonction de l’application sélectionnée on n’est pas obligé de tout décoder par exemple si la qualité de reproduction n’est pas importante).

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

5.1 La modélisation d’un document (objet d’information)

Dans le cadre d’une **préservation long terme (« deep preservation »),** il est nécessaire de disposer d’une modélisation des données d’un document pour construire le système d'information sous-jacent. La conceptualisation traite de la catégorisation et des relations (d’inclusion, « partie-tout » ou d’état et de changement d’état) entre les éléments d’un système.

Le **Modèle de référence ISO/OAIS** est particulièrement adapté à la problématique de l'archivage numérique, même s'il ne préjuge pas de la nature des objets à archiver.

Le modèle d’information de l’OAIS s'appuie sur le formalisme UML (Unified Modeling Language) de représentation des objets du domaine et plus spécifiquement sur ce qu'on appelle les diagrammes de classe. Une classe décrit le comportement et le type d'un ensemble d'objets partageant des propriétés. Une classe est un concept abstrait représentant des objets concrets.

Principes fondamentaux d'une (re-)présentation «en évidence» d'un fichier binaire :

Objet(s) physique(s)
ou
Objet (s) numérique(s)

Objet(s) numérique(s): objet(s) constitué(s) de séquences de bits

Informations qui traduisent un Objet-données dans un objet perceptible. Par exemple, la définition du code ASCII décrit comment une séquence de bits (un Objet-données) est convertie en caractères.

Objet-données avec son Information de représentation (et sa base de Connaissance)

5.1 La modélisation d'un document (objet d'information)

L'**Objet-données** peut aussi bien prendre la forme d'un objet physique que celle d'un objet numérique (c'est-à-dire une séquence de bits).

L'Information de représentation qui accompagne un objet numérique, ou séquence de bits, sert à donner une signification supplémentaire à cet objet. Son rôle est d'établir la correspondance entre les bits et des types de données généralement reconnus comme un caractère, un nombre entier, un nombre réel, ou des groupes de ces types de données. Elle relie ces types de données à des concepts signifiants de plus haut niveau, et décrit les relations potentiellement complexes entre les objets.

Toutes les données inclues dans l'objet données ne sont pas des informations de représentation liée à l'Objet-information comme des systèmes de redondance pour corriger les risques liés aux imperfections de codage de l'Objet Donnée (des codes de corrections d'erreur)!

Le modèle OAIS n'impose nullement que la définition des informations de représentation soit représentée en sémantique !

5.1 Information de représentation : un réseau de représentations !

- le diagramme indique que l'objet « Information de représentation » peut contenir des références à **d'autres informations de représentation**.
- En effet, l'information de représentation est elle-même un objet information possédant son propre objet numérique et l'information de représentation associée à la compréhension de cet objet (cela est indiqué de façon concise par l'association « interprété en utilisant »). Le jeu résultant d'objets est appelé **réseau de représentation.**
- Ce processus de récursivité peut se poursuivre jusqu'à ce que l'on aboutisse à des formes physiques. Par exemple, la norme ISO 646 sur le codage ACSII, lorsqu'elle est un objet numérique, est un document codé en ASCII, et sortir de cette boucle nécessite alors de stocker un document papier présentant cette norme.

Objet Information =
Objet données + information de (re-)présentation

Nous voyons apparaître au travers de cet exemple deux équations fondamentales de la préservation :

- Pour pérenniser l'information contenue dans un objet numérisé, **il est nécessaire, mais pas suffisant** de conserver cet Objet données !
- Pour la préservation, il est indispensable de conserver, avec cet objet données, un ensemble d'informations appelées **Information de (re-)présentation** qui nous permettra de passer des bits constituant l'objet numérique à la présentation du contenu informationnel de cet objet en évidence pour un humain.
- Pour être utile, une information de représentation doit aussi être en adéquation avec **la base de connaissance des utilisateurs** de l'archive (i.e. la « communauté d'utilisateurs cible » selon l'OAIS), tant actuels que futurs.

5.1 Information de structure et Information sémantique

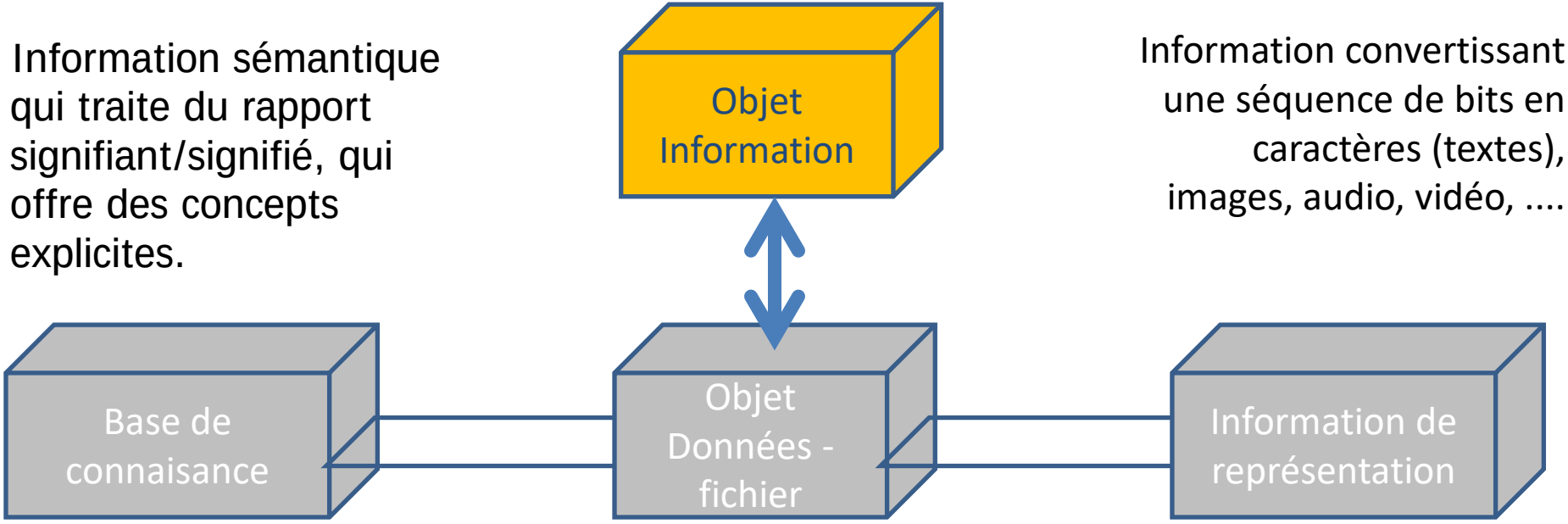
Il est utile de rappeler que l'Information sémantique associée à une ou plusieurs parties d'une information encodée numériquement est indépendante du format. Par exemple, la signification des nombres dans un fichier de données ne dépend pas de leur encodage en valeurs entières discrètes ou réelles IEEE; le sens des mots dans un document est indépendant du fait que le document soit un fichier Word ou un fichier PDF, etc. Afin de faciliter la gestion des informations de représentation, l'asbl Titan a proposé un modèle qui scinde :

- L'**information de structure** interprète les chaînes de bits sous la forme de types de données et d'autres structures de plus haut niveau (dans notre exemple, l'interprétation des chaînes de bits en caractères ASCII, et la structuration des caractères ASCII en séquences répétables). Elle inclut notamment la spécification du format de données et de l'environnement logiciel et matériel nécessaire pour accéder aux données.
- L'**information sémantique** fournit un cadre additionnel d'interprétation des structures de données. Dans notre exemple, l'information sémantique est donnée par la signification des champs d'information. La langue utilisée est aussi une information sémantique.

Objet information :

Objet-données + Information de représentation + Base de Connaissance

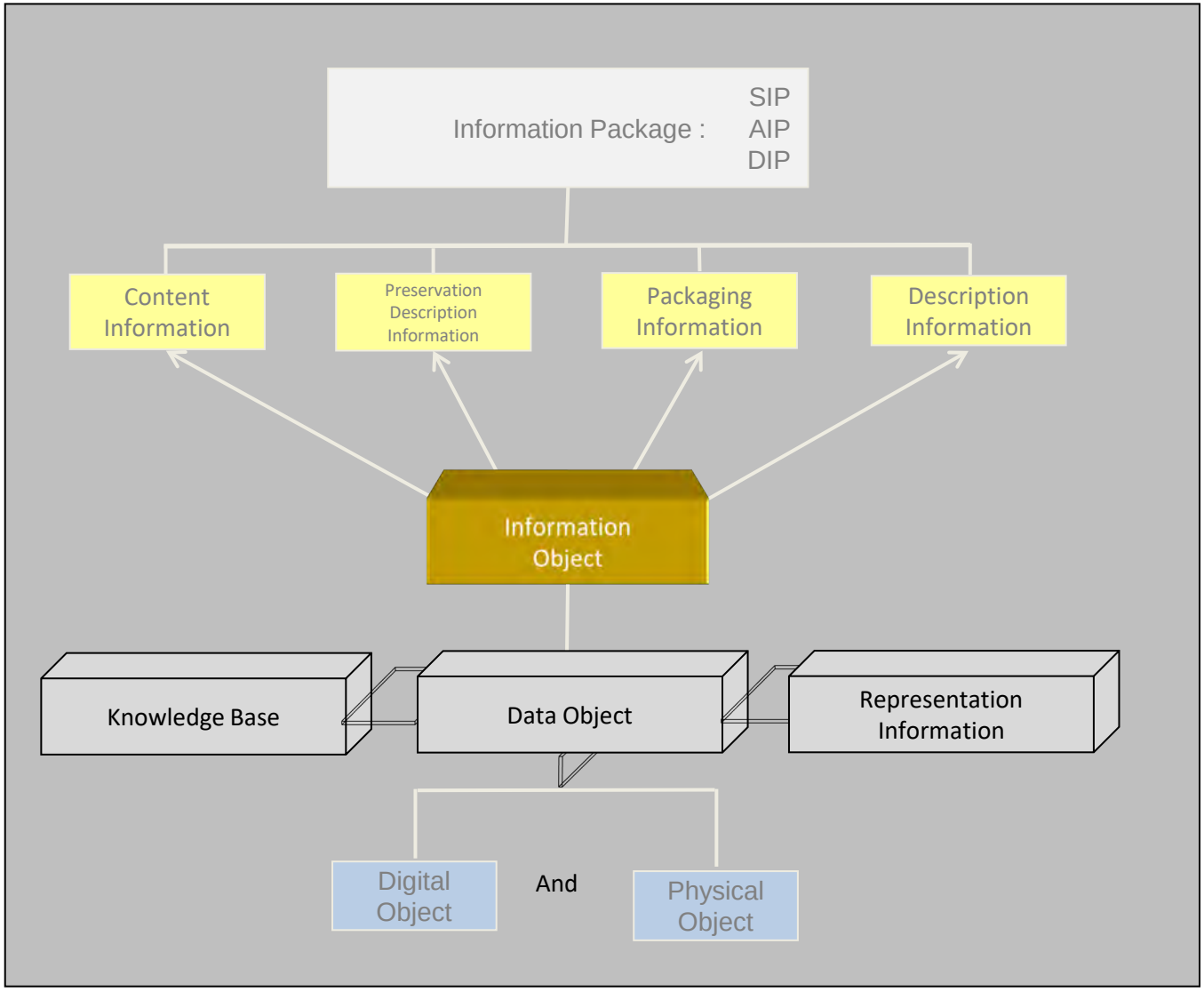
Information sémantique qui traite du rapport signifiant/signifié, qui offre des concepts explicites.



Information convertissant une séquence de bits en caractères (textes), images, audio, vidéo, ...

Objet(s) numérique(s): objet(s) constitué(s) de séquences de bits

PDI:
- Reference
- Context
- Provenance
- Fixity



5.1 Information sémantique et base de connaissance

Le contenu de **l'Information sémantique** peut être très diversifié et complexe. Cela peut comprendre la signification particulière associée à chacun des éléments de l'Information de structure, les opérations réalisables sur chaque type de données et leurs corrélations … C'est pour cette raison que l'asbl a proposé le concept de « Base de Connaissance ».

Une **base de connaissance** offre une cartographie des connaissances spécifiques à un domaine spécialisé donné, sous une forme exploitable par un ordinateur. Elle peut contenir des règles (dans ce cas, on parle de base de règles), ou des ontologies (qui expriment les propriétés et les relations que les termes utilisés partagent). Sur la base des règles ou des ontologies un moteur d'inférence - simulant les raisonnements déductifs logiques - peut être utilisé pour déduire de nouveaux faits. Contrairement à une base de données relationnelle (basée sur la syntaxe et la statistique), elle offre une navigation interactive dans un graphe de résultats validés … et cela indépendamment des langues !

Il faut à donc la fois préserver les données (sur des supports adéquats) , les informations de représentation (et les applications qui ont généré ces données) et enfin créer une base de connaissance pour générer les liens entre les données et leur(s) signification(s).

Tutoriel UNESCO 2020 part 1 : plan de la présentation

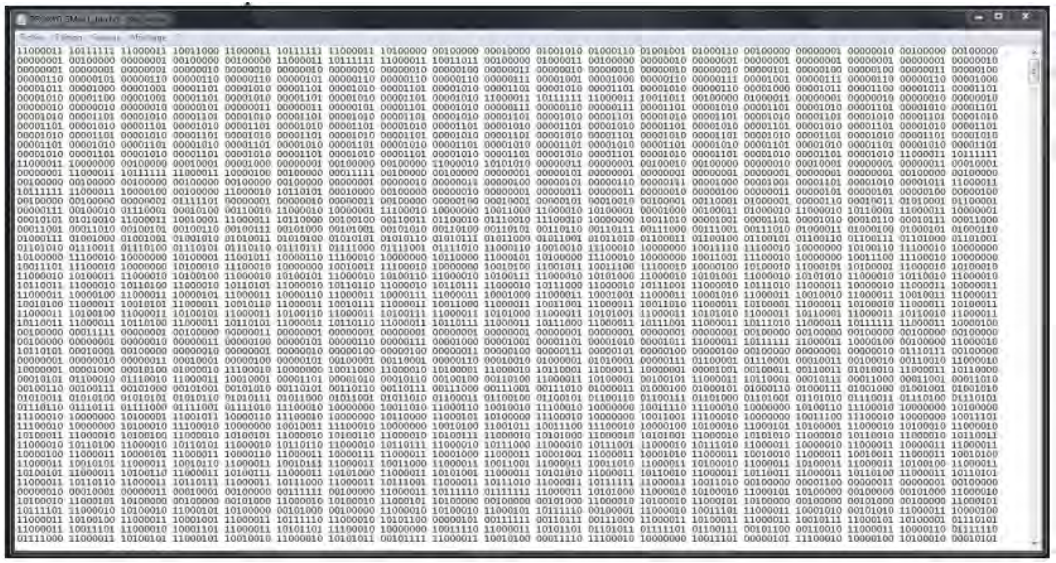
V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

5.2 De l'objet donnée à la représentation

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).

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



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 information supplémentaire qui sont appelées « **information de représentation** ».

5.2 De l'objet donnée à la 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

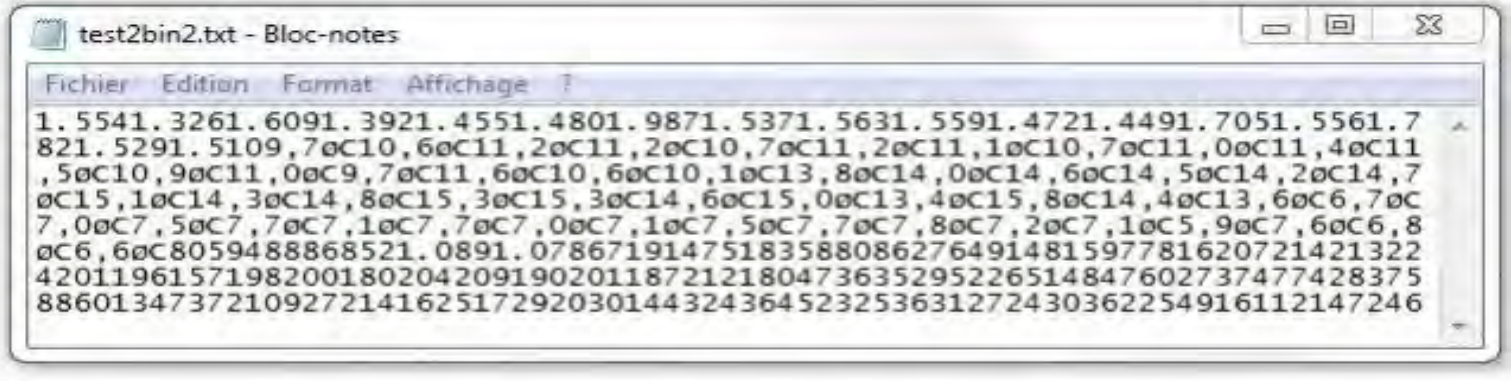
Pour permettre dans l'avenir à un internaute d'interpréter ce fichier,

- 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.

5.2 De l'objet donnée à la représentation

Pour interpréter le fichier, nous pouvons par exemple 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.

Nous pouvons maintenant visualiser le fichier avec le bloc-notes :

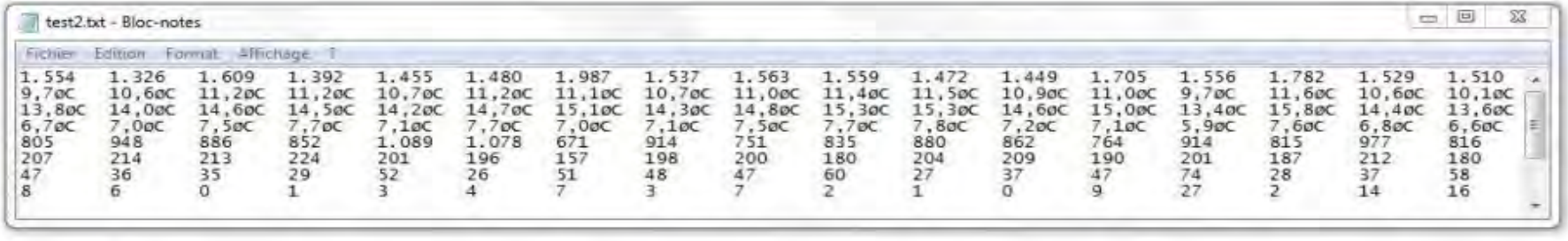


Mais nous sommes encore loin d'en avoir une représentation intelligible et percevoir le sens de tous ces caractères !

5.2 De l'objet donnée à la représentation

Pour ce faire, nous avons besoin d'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 (par exemple, uns séquence contient X champs d'informations contenant Y nombres entiers codés sur 6 octets).

Si cette information est disponible, 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
9,7øC	10,6øC	11,2øC	11,2øC	10,7øC	11,2øC	11,1øC	10,7øC	11,0øC	11,4øC	11,5øC	10,9øC	11,0øC	9,7øC	11,6øC	10,6øC	10,1øC
13,8øC	14,0øC	14,6øC	14,5øC	14,2øC	14,7øC	15,1øC	14,3øC	14,8øC	15,3øC	15,3øC	14,6øC	15,0øC	13,4øC	15,8øC	14,4øC	13,6øC
6,7øC	7,0øC	7,5øC	7,7øC	7,1øC	7,7øC	7,0øC	7,1øC	7,5øC	7,7øC	7,8øC	7,2øC	7,1øC	5,9øC	7,6øC	6,8øC	6,6øC
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

Cette représentation est déjà beaucoup plus claire (et elle a d'ailleurs la structure d'une feuille de tableur), mais il nous manque encore des informations capitales pour la compréhension de ces chiffres, à savoir la signification de chacun de ces champs.

5.2 De l'objet donnée à la représentation

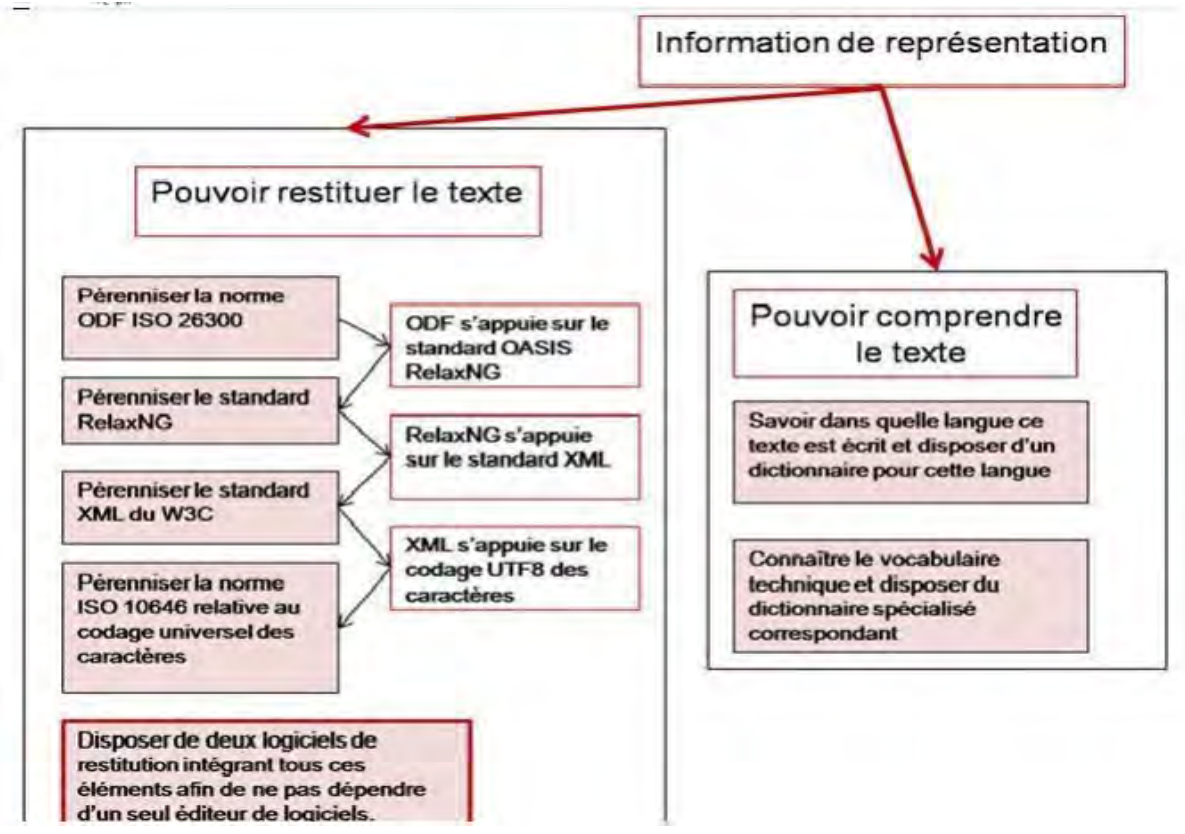
Si l'on explicite la signification de chaque champ par exemple dans un tableur excel, on obtient alors le résultat suivant :

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

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

Base HELB : Jean-Louis Blanchart

Information de représentation + Base de connaissance: Représenter le texte vs Comprendre le texte !



Cas d'un document technique stocké dans un fichier au format ODF

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

5.3 Extraction des données – structure des données

L'extraction de données est un processus d'exploration de vastes ensembles de données visant à obtenir des données « lisibles par machine », qui sont créées pour être traitées par un ordinateur, au lieu d'être présentées pour un utilisateur humain. L'extraction de données se fonde essentiellement sur des schémas. L'analyse des données est effectuée de plusieurs façons, notamment à l'aide de notions comme l'apprentissage machine, où des algorithmes adaptatifs complexes sont utilisés pour analyser artificiellement les données.

La structure des données est liée aux informations contenues, pas à la façon dont elles sont affichées. Des formats comme les documents Word, les pages HTML et les fichiers PDF s'intéressent plus à la mise en page visuelle des données. Certains formats sont plus facilement lisibles par machine, comme le CSV, le XLM, le JSON et les fichiers Excel.

Une fois les données collectées et stockées, l'étape suivante consiste à donner du sens aux données.

5.3 Les formats d'organisation des données :

Le **format d'organisation des données** est déterminé par le type d'opération requis ou par les algorithmes qui seront appliqués. Les types d'organisation possibles sont les suivantes :

- le mode **Tableau** qui stocke un ensemble de données afin de faciliter le calcul de leur emplacement ou leur extraction.
- Le mode **Pile** stocke les données en suivant l'ordre dans lequel les opérations seront appliquées.
- Le mode File : premier entré, premier sorti.
- Mode **liste chaînée** : chaque élément ou nœud contient une référence, ou un lien, vers l'élément suivant de la liste.
- Le mode **Arbre** : une organisation sous une forme hiérarchique abstraite. Chaque nœud peut contenir plusieurs sous-valeurs.
- Le mode **Tri** stocke des chaînes pouvant être représentées visuellement sous forme graphique.
- Le mode **Graphe** stocke les données de façon non linéaire dans des nœuds individuels et des liens. Les relations entre ces nœuds contiennent des propriétés (classer la pertinence ou le poids de ces relations) en vue de créer des listes chaînées. La relation entre deux nœuds va indiquer par exemple qu'un client d'un site de e-commerce (un nœud) achète régulièrement un type de produit particulier (un autre nœud). Ce produit est aussi acheté par d'autres clients, dont l'historique d'achat est également stocké en base. Une requête Graph va donc permettre de proposer cet historique d'achat au premier client, par le jeu des relations de la base.

5.3 La signification d'un document (base de connaissance)

Une **base de connaissance** offre une cartographie des connaissances spécifiques à un domaine spécialisé donné, sous une forme exploitable par un ordinateur. Elle peut contenir des règles (dans ce cas, on parle de base de règles), ou des ontologies (qui expriment les propriétés et les relations que les termes utilisés partagent).

Les **Ontologies** traitent de taxonomies et de classifications, schémas de base de données, et de théories entièrement axiomatisées. Une ontologie décrit les concepts et les rapports qui sont importants dans un domaine particulier, fournissant un vocabulaire pour ce domaine aussi bien que des spécifications automatisées de la signification des termes utilisés dans le vocabulaire.

Un **moteur d'ontologies** est un processus applicatif qui possède une dimension sémantique (définition et mise en relation de concepts) ; une dimension logique (validation de la création de relations entre concepts ou déduction de relations non explicites entre concepts) et enfin une dimension usage avec la construction d'un vocabulaire («outil», «utilisateur», «traitement»), la construction d'une syntaxe (sujet, verbe, complément) et d'une grammaire : «l'utilisateur» «accorde» «l'outil» pour réaliser «un traitement».

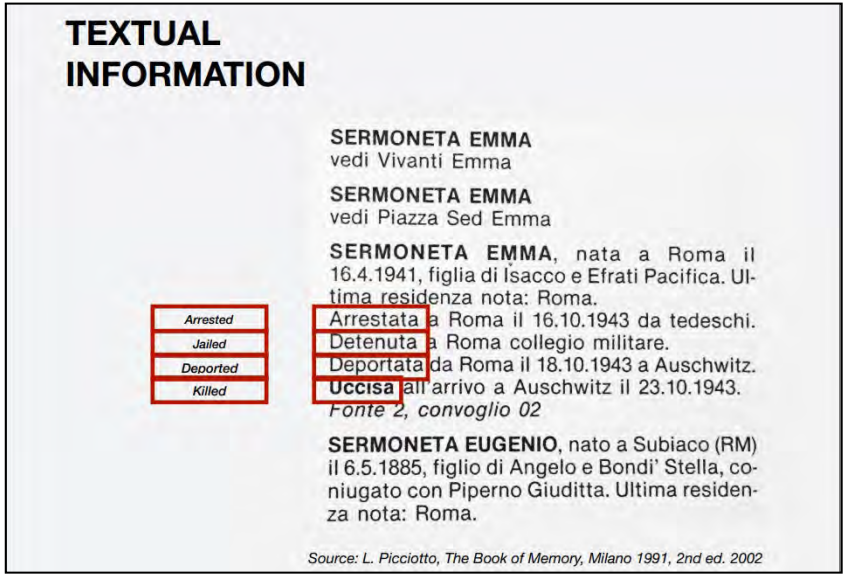
La finalité du moteur d'ontologie est de :
- **réduire la distance** entre le langage de la machine (logique) et le langage de l'utilisateur (lisible) entre des langages pratiqués dans différentes cultures (métier, région, temps, etc.) en intercalant entre l'une et l'autre un langage compréhensible par l'un et l'autre
- **améliorer** la qualité, l'efficacité et l'efficience des **échanges** entre les acteurs d'un projet via un socle sémantique commun

5.3 La signification d'un document (base de connaissance)

Le 22 novembre 2019, le projet ADOCHS organisait à la KBR une journée d'étude internationale sur la publication de données ouvertes, et en particulier sur la thématique de la SHOA en Italie).

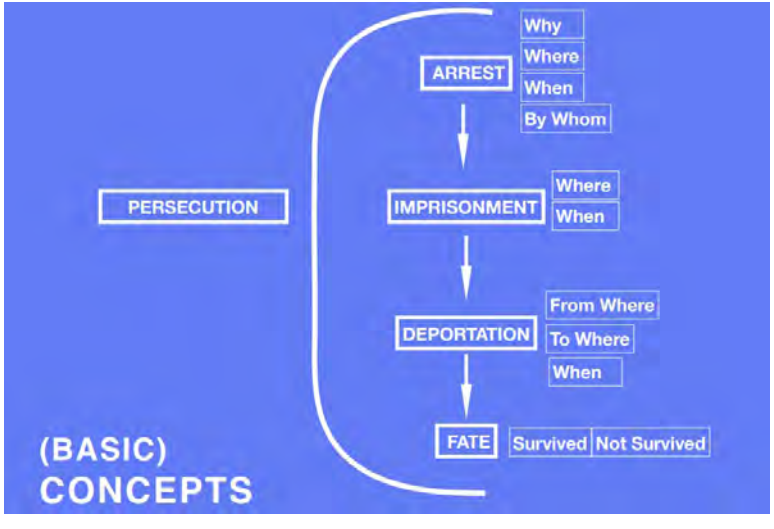
La présentation « SHOAH Ontology » constituait un bel exemple d'extraction de données en vue de donner du sens à ces données via une conceptualisation (concept de « persecution ») et une organisation en graphe.

5.3 KBR : « Linking the past » : the Shoah Ontology



Les données disponibles sur la persécution des Juifs en Italie pendant l'occupation nazie (1943-1945) : les noms, prénoms, date d'arrestations, les lieux de détention aux déportations vers les camps de concentration et d'extermination.

La description formelle de la persécution des Juifs en Italie pendant l'occupation nazie (1943-1945) : **les arrestations**, **les lieux de détention**, **les déportations** vers les camps de concentration et d'extermination et l'issue (survivant/non survivant).



5.3 La connaissance : classe – relation - propriété

Une **ontologie** comme celle de la SHOA s'efforce de représenter les choses avec aussi peu de différence structurelle et fonctionnelle que possible par rapport à la réalité sous-tendue. Cette manière de faire permet de préparer les définitions ontologiques formelles par des diagrammes, des figures qui prennent la forme de « Flow diagrams » où les noms exprimés à l'intérieur des symboles sont les noms des futures classes dans les ontologies. Les exemples traités en annexe seront exprimés comme des instances de classes représentant les réalités visées.

Une **classe** déclare des propriétés communes à un ensemble d'objets, des attributs représentant l'état des objets et des méthodes représentant leur comportement. Elle apparaît comme un moule ou une usine à partir desquels il est possible de créer d'autres objets ; dans ce cas, il s'agit d'une instance d'une classe (création d'un objet ayant les propriétés de la classe).

Une **relation** entre deux classes décrit les connexions structurelles (les liens) entre leurs instances.

Une **propriété** est une caractéristique structurelle. Dans le cas d'une classe, les propriétés sont constituées par les attributs que possède la classe.

5.3 La connaissance : classe – relation - propriété

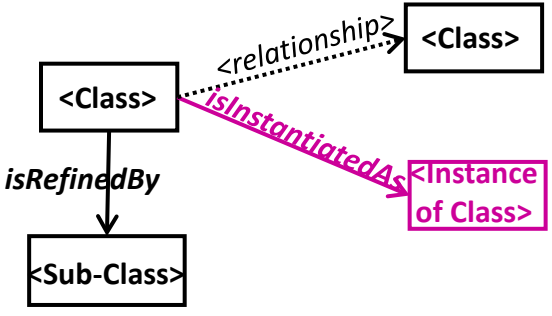
(1) : CLASSES

<Class>

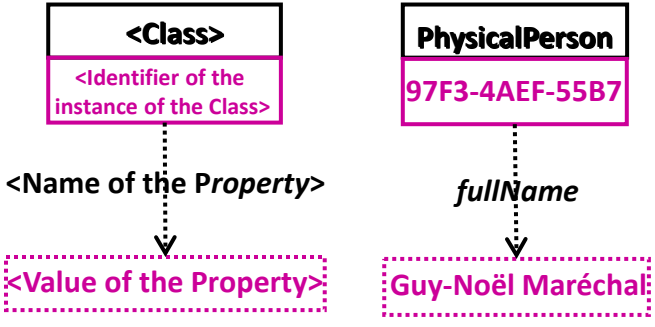
A Class is represented by a black rectangle. The string <Class> means that, at usage, the string has to be replaced by its name !
Example: <Class> becomes **PhysicalPerson**

PhysicalPerson

(2) RELATIONSHIP



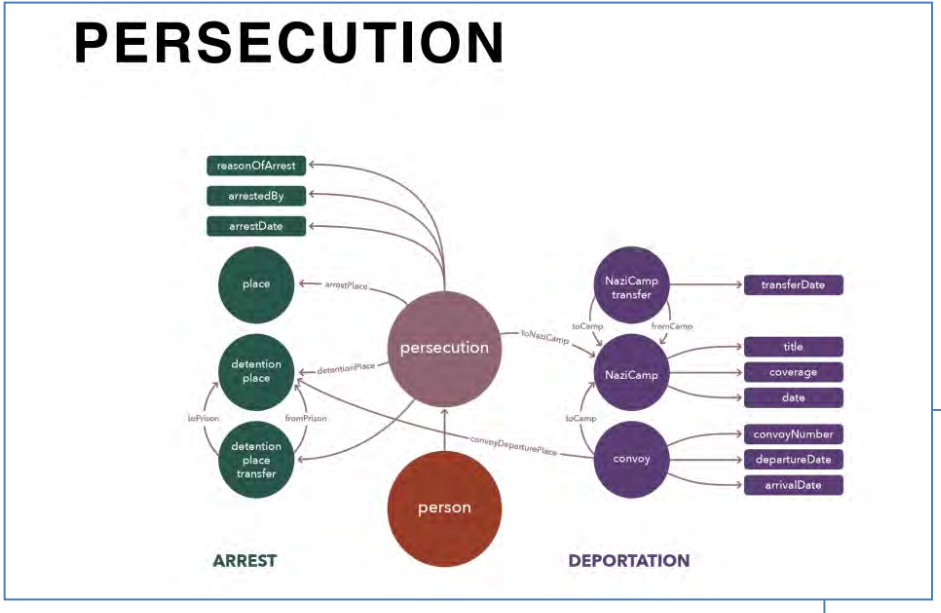
(3) PROPERTIES

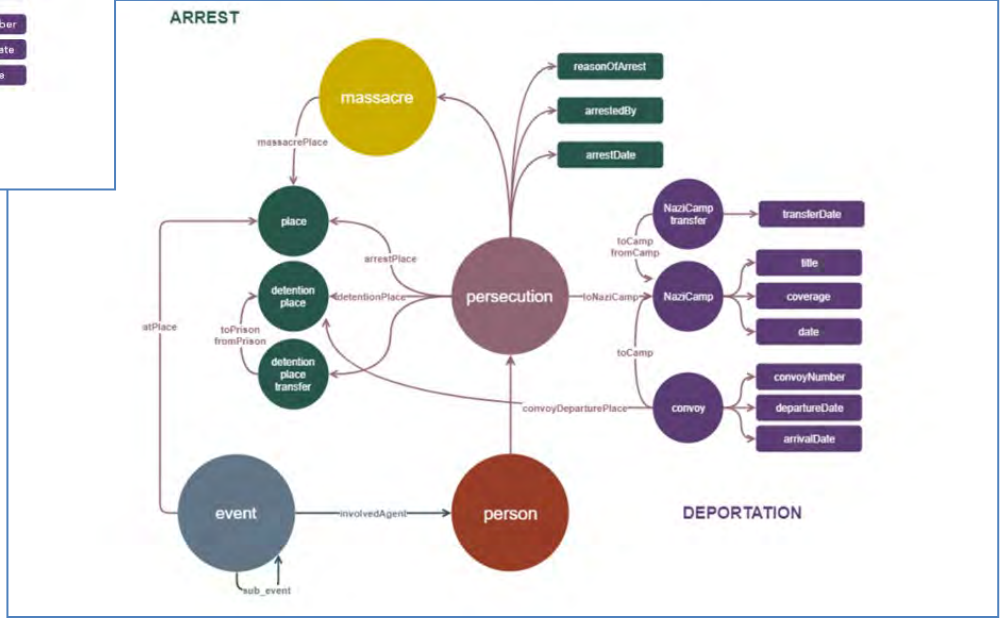


Properties are attached to instances of Classes through typed links. These typed links are represented by **dotted arrows** labelled with the '**Name of the Property**'

The **Values** of the Properties are represented in a **violet dotted rectangle**

5.3 KBR : « Linking the past » : the Shoah Ontology





Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

5.4 L’interopérabilité de migration.

Une **migration** de données fait référence au transfert de données entre différents types de formats de fichiers, bases de données et systèmes de stockage. Cependant, le «transfert» n'est pas le seul aspect de la méthodologie de migration de données. Comme les données sont diverses, hétérogènes, insignifiantes, … le processus de migration doit inclure des mappages et des transformations entre les données « source » et « cible ».

La migration de données peut être divisée en plusieurs types:

- 1. **Migration de base de données** : cela affecte simultanément le langage ou le protocole de données et modifie les données sans modifier le schéma.
- 2. **Migration d'applications** : chaque application possède un modèle de données unique et en plus les applications ne sont pas portables. L'introduction d’un middleware sémantique au cours du processus contribue à combler le fossé technologique.
- 3. **Migration de stockage** : le transfert de données d'un système de stockage à un autre, tel qu'un disque dur ou le cloud qui garantit une certaine évolutivité des applications ou services d'informations.

Les deux principaux obstacles auxquels sont périodiquement confrontés les archivistes sont la nécessité de faire migrer les fichiers numériques et l’obsolescence rapide du matériel utilisé pour les stocker.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

OAIS

Open Archival Information System

- Le Modèle de Référence **ISO/OAIS** a été élaboré par le Comité Consultatif pour les Systèmes de données spatiales (CCSDS : Consultative Committee for Space Data Systems) comme une contribution au Comité technique de l'ISO 20, sous-comité 13.
- C'est un cadre de réflexion pour la compréhension et l'application des concepts nécessaires à la préservation à long terme de l'information numérique (incluant l'évolution technologique)
- Première version de la norme publiée en 2002 en tant que norme ISO (ISO 14721:2003) : réalisée par la CCSDS avec le concours des archives nationales et des grandes bibliothèques
- La norme a été révisée en 2012 (ISO 14721:2012)

Pourquoi OAIS ?

OAIS est un modèle abstrait qui :

- définit **les concepts** indispensables à la compréhension et à l'analyse du problème de l'archivage numérique à long terme,
- propose **un référentiel terminologique commun** permettant aux diverses communautés concernées de dialoguer et de se comprendre indépendamment des vocabulaires spécifiques à leurs domaines respectifs,
- propose **un modèle d'information** (comment décrire les objets en vue de leur préservation ?), un modèle fonctionnel (quelle organisation mettre en place pour assurer cette activité de préservation ?), et des stratégies de préservation (quelles méthodes pour éviter l'obsolescence technologique ?)

C'est une base indispensable pour la définition, l'élaboration et la mise en œuvre de solutions technologiques et organisationnelles pour la préservation. C'est une approche générique d'un modèle d'archivage dans le cadre d'une communauté d'utilisateurs :

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

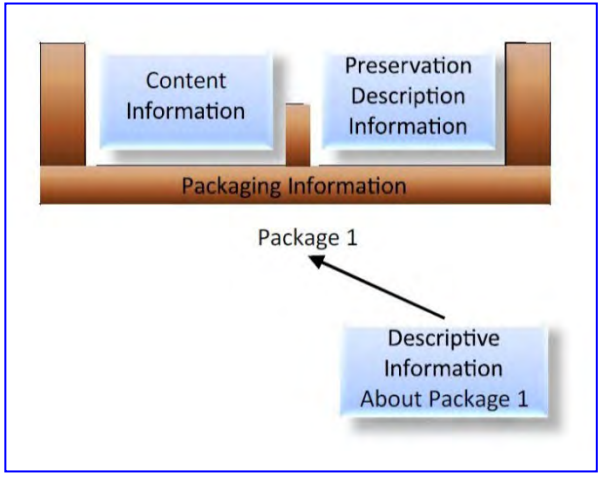
6.1 OAIS : le concept de PACKAGE SIP – AIP – DIP

La préservation nécessite un emballage des données ! Le « package » est un concept de conteneur conceptuel qui contient des ensembles de données et des métadonnées pertinentes. Les contenus d'un « package » peuvent être dispersés ou réduit en un seul objet numérique.

Pour l'emballage, OAIS définit des ensembles de données relatives :

- au contenu
- aux informations de description de la préservation du contenu (référence, contexte, provenance et fixité)

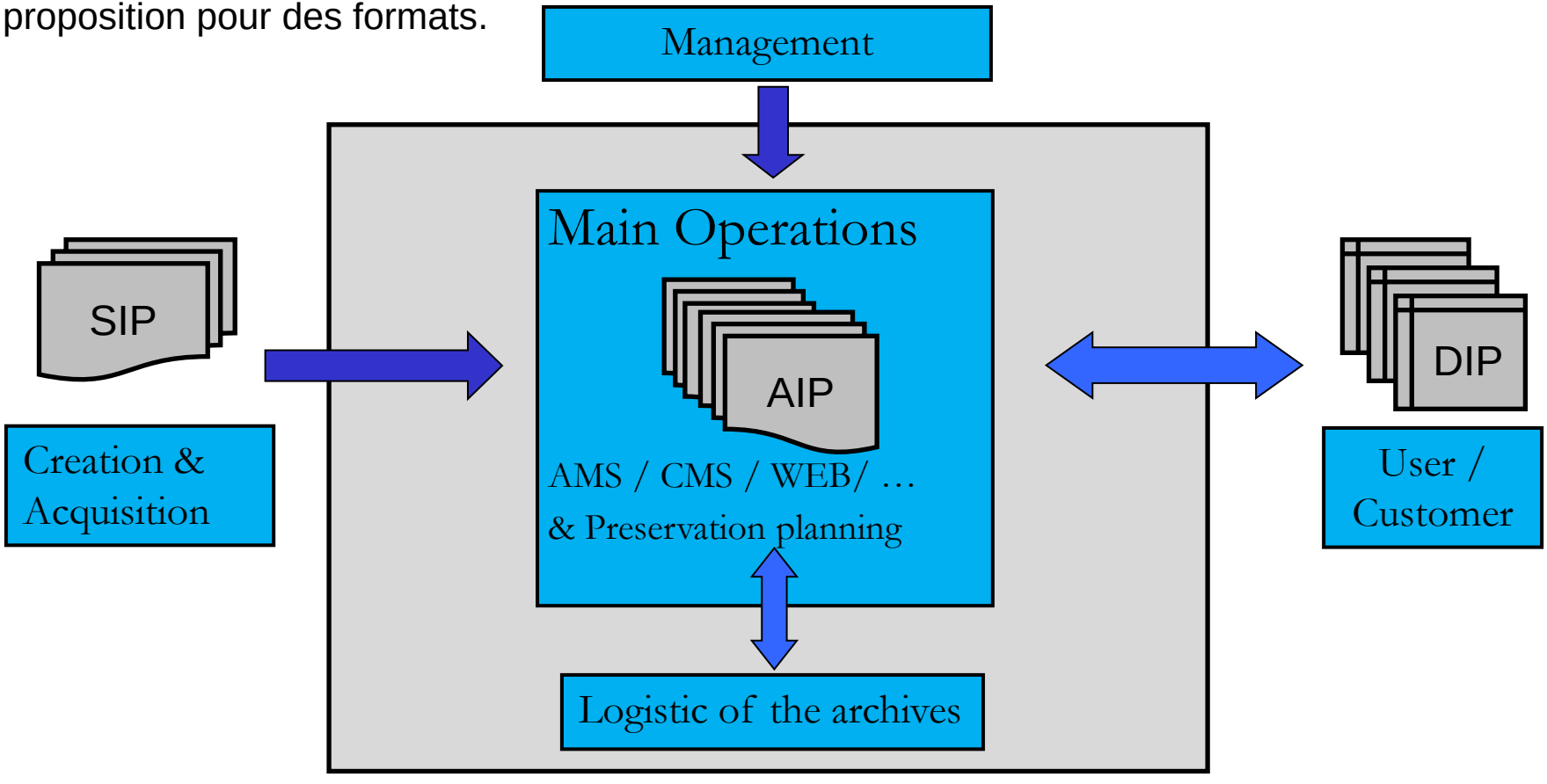
Ces deux éléments (données contenu + données de pérennisation) sont reliés entre eux par le biais de l'information d'empaquetage.



L'information de description (étiquette) fournit aux autres entités du système (notamment à l'entité «accès») toutes les données de gestion techniques et archivistiques nécessaires.

6.1 OAIS : formats pour la représentation des contenus :

Pour la partie « modelling », la norme ISO reconnait trois types de formats pour la représentation des contenus (SIP – AIP – DIP), par contre, il n'y a pas de formulation de proposition pour des formats.



6.1 Types de formats pour la représentation des contenus

- **SIP : « Submission Information Packages »:** les formats de présentation candidat à l'archivage. Il s'agit de formats les plus complets possibles que les applications sont capables de générer et où les objets sont définis de manière autonome. Ces SIP sont fournis par un 'producteur à l'importation dans un système.
- **AIP : « Archival Information Package »** : les formats de gestion de l'archivage : les SIP sont traités dans des modules d'ingestion, de validation et de structuration en vue de permettre la capacité de gérer la persistance au sein d'un système. C'est-à-dire que les AIP ont une vocation de gestion des évolutions des contenus archivés et doivent être suffisamment généraux pour être capables de générer des formats ciblés à la demande pour l'exportation.
- **DIP «Dissemination Information Packages »:** les formats d'export ciblés : ces représentations sont dites « exogènes ». Il s'agit de formats ciblés sur une communauté particulière 'designated community', ayant un objectif global défini. L'EBUCore en est l'exemple pertinent incontestable. Il est ciblé sur les besoins des diffuseurs visant à s'échanger des contenus exploitables en y incluant leur environnement.
- **P-DIP « Persistent Dissemination Information Package »** : un cas particulier notable, où la 'designated community' est un autre système d'archive.

6.1 OAIS : PDI : les informations de pérennisation

Un certain nombre **d'informations de pérennisation** (PDI), sont indispensables à la bonne compréhension de l'objet archivé dans le temps. Ces informations de pérennisation sont de plusieurs sortes :

- des informations de **provenance**, qui documentent l'historique du contenu d'information ;
- des informations de **contexte**, qui détaillent les liens entre le contenu d'information et son environnement ;
- des informations **d'identification**, qui permettent d'identifier sans équivoque le contenu d'information ;
- des informations **d'intégrité**, qui décrivent les mécanismes garantissant l'intégrité du contenu d'information.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

6.2 Modèle d'information : représentation de la substance :

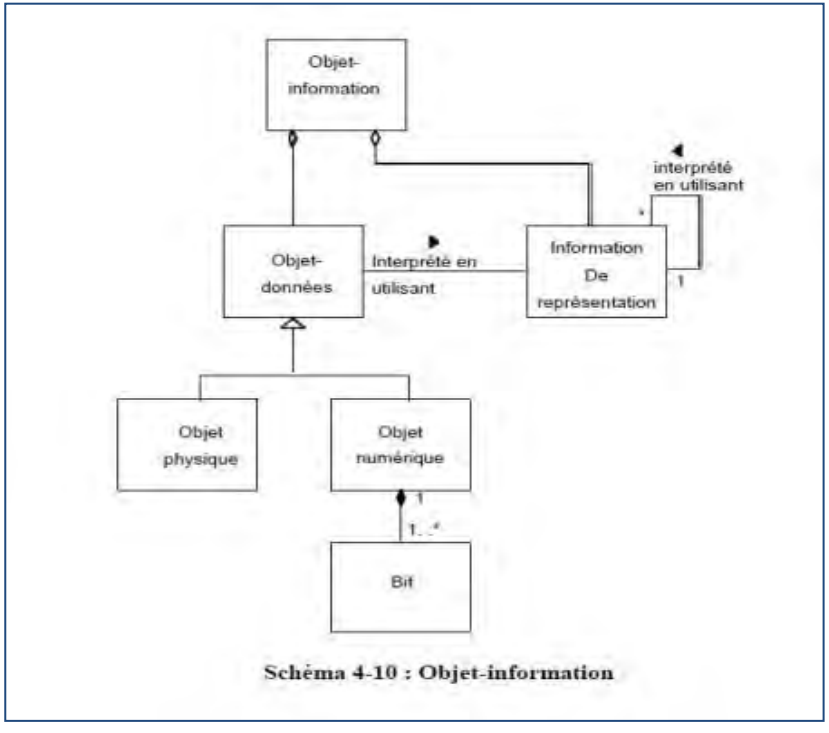


Schéma 4-10 : Objet-information

Objet information : Objet-données avec ses Informations de représentation

Information de représentation : les données accompagnée d'une description (de structure et sémantique) permettant d'interpréter cette chaîne de bits en vue d'offrir un objet d'information en perception à l'usager d'un système d'information

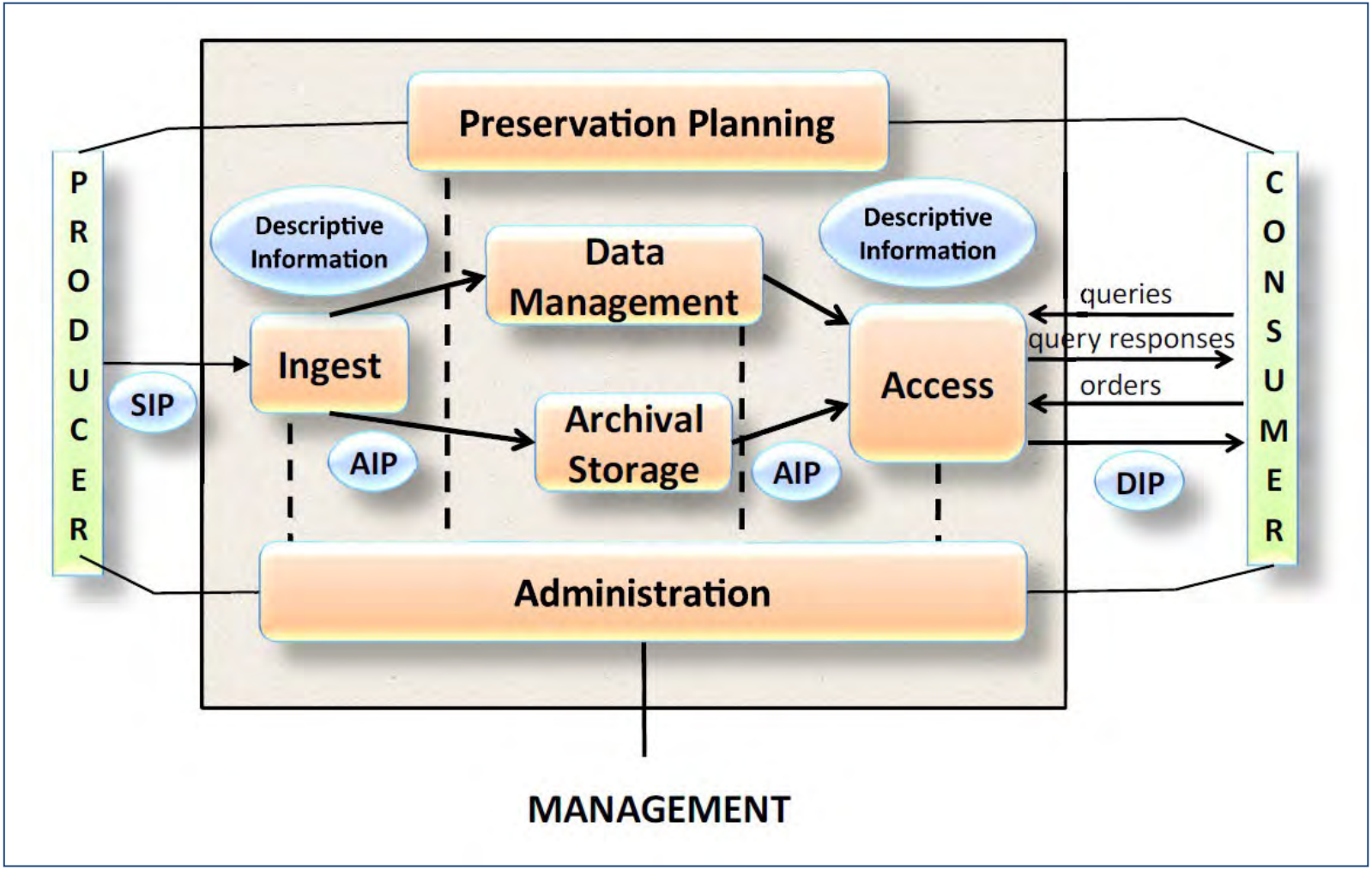
Il faut à donc la fois préserver les données (sur des supports adéquats) , les information de représentation (et les applications qui ont générés ces données) et enfin créer une base de connaissance pour générer les liens entre les données et leur(s) signification(s).

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

6.3 OAIS : le modèle fonctionnel



6.3 OAIS : Schema fonctionnel :

- L'entité **« entrées »** reçoit, contrôle et valide les objets à archiver par l'entité **« stockage »,** tandis que les informations nécessaires à leur description et à leur gestion dans le temps sont transmises à l'entité **« gestion des données »** :
  - L'entité **« stockage »** assure la conservation physique des objets archivés. Elle tient les objets archivés à la disposition de l'entité « accès ». Conformément aux règles établies par l'entité «administration », elle prend en charge la réalisation des copies multiples et le renouvellement des supports anciens.
  - L'entité **« gestion des données »** prend en charge la mise à jour de toutes les informations internes – base de données – nécessaires au système d'archivage. Elle fournit aux autres entités du système toutes les informations de gestion techniques et archivistiques nécessaires (et notamment à l'entité «accès »).
- L'entité **« planification de la pérennisation »** est la cellule de veille et de planification du système. Elle prépare et planifie ces évolutions. Elle gère également la « communauté d'utilisateurs cible » en vue de garantir que le service d'accès reste conforme aux attentes nouvelles des utilisateurs.
- L'entité **« administration »** assure la coordination générale du système, établit les règles internes et veille à la qualité globale du service rendu aux utilisateurs.
- L'entité **« accès »** regroupe tous les services qui sont en interface directe avec les utilisateurs. Outre les fonctions de contrôle d'accès, il s'agit principalement de permettre aux utilisateurs de rechercher dans le « catalogue » des objets archivés, et de leur fournir les objets dont ils passent commande.

Tutoriel UNESCO 2020 part 1 : plan de la présentation

V2020_02_29

1. Création, définition et substance d'un document
2. Concepts issus de la Science de l'information :
    - 2.1 La représentation (la donnée) vs la signification (information)
    - 2.2 Les métadonnées
    - 2.3 La communication (l'échange de la signification)
    - 2.4 Le codage de la représentation vs le transport
    - 2.5 L'intelligence Artificielle vs la connaissance
3. La formalisation informatique
    - 3.1 Le rôle de l'applicatif
    - 3.2 L'encapsulation (fichier/dossier, zip, ….)
    - 3.3 La saga des métadonnées
    - 3.3 Les métadonnées ancillaires de fichiers
4. La transformation numérique des essences
    - 5.1 le traitement du texte
    - 5.2 Le traitement du son (temps)
    - 5.3 Le traitement des images (espace)
    - 5.4 Le codage de la vidéo (intra – inter-image)
5. Les concepts de la préservation :
    - 5.1 La modélisation d'un document (objet d'information)
    - 5.2 Le codage/décodage numérique (informations de représentation)
    - 5.3 La signification d'un document (base de connaissance)
    - 5.4 La migration
6. Définition de la norme ISO/OAIS :
    - 6.1 Le concept d'emballage – Package (SIP – AIP – DIP)
    - 6.2 Le concept de représentation de l'information
    - 6.3 Le schéma fonctionnel
    - 6.4 Les limites du modèle conceptuel ISO/OAIS

6.4 OAIS : Les avantages et les limites du modèle conceptuel

Le modèle conceptuel exprimé pour les « Open Archival Information System », dans sa version 2012, couvre correctement les objectifs qui sont les siens :

- des systèmes couvrant l'archivage persistant d'un «SIP » confier par un 'Provider'
- l'organisation de l'accès à ces données à un 'Consumer' par un protocole de recherche dans des métadonnées,
- la livraison sous la forme d'un « DIP » « Package de données» ;
- le tout dans le contexte d'une 'Designated Community'.

Le modèle décrit également des système couvrant la persistance de la substance (une représentation particulière de l'Information) représentée dans chaque SIP.

Par contre :

- La gestion de l'OAIS est disjointe de celle du 'Provider' et de l'opérationnel.
- Le modèle OAIS ne se préoccupe pas de l'exploitation du DIP par 'Consumer'; même si le 'Consumer' et le 'Provider' peuvent être la même personne (réutilisation de la matière)
- OAIS ne couvre pas la persistance de «Documentary Heritages».
- Un ensemble sémantiquement cohérent de données structurées (une collection) ne peut être géré comme un tout : le système gère chaque « Package » de manière disjointe.
- Le système ne gère pas les perceptions en évidence : ni lors de la jouissance des DIP, ni lors de l'acquisition ou de l'authoring de contenu du SIP.

The END ...

Pour d’autres informations voir : www.titan.be