Cours : thème 2 - Question 5

Question 5 : La numérisation suffit-elle à valoriser l'information ?

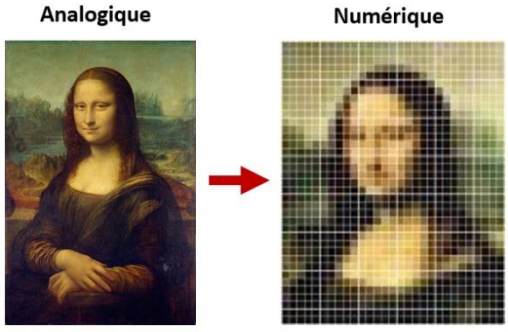
Notions abordées :
- Gestion électronique de documents (GED) : acquisition/mémorisation, publication/révision, indexation/classification, sauvegarde/archivage ;
- Classification de l'information et métadonnées ;
- Classification de l'information et visibilité sur internet : référencement naturel/payant ;
- Interopérabilité et standardisation des échanges : fichier/format, syndication de contenu et flux RSS, XML, JSON…

1. Numérisation

1.1. Qu'est-ce que la numérisation ?

Numérisation	La numérisation de l'information consiste à créer une représentation binaire d'une information non binaire et à la conserver sur support informatique, typiquement dans un ou plusieurs fichiers.

En quelque sorte, numériser, c'est transcrire la réalité sous forme de « zéros » et de « uns », c'est-à-dire sous forme binaire. On parle de valeurs discrètes, discrètes parce qu'il y a un nombre fini (=limité) de valeurs possibles : 0 ou 1.



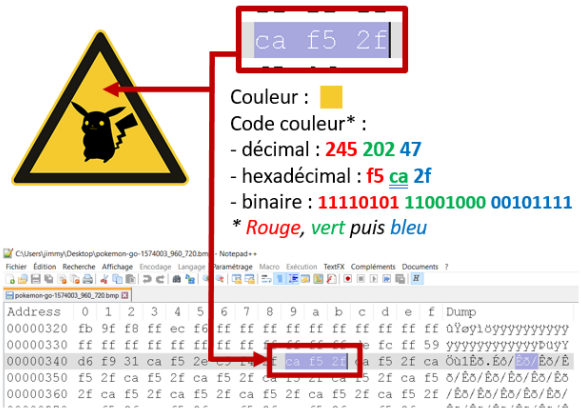
Problème ! La réalité n'est pas faite de 0 et de 1… Le son, la lumière, la température… Ces informations peuvent prendre une infinité de valeurs ! Elles sont analogiques. On parle de valeurs continues. Pourtant, un ordinateur, ou plus généralement un équipement informatique, ne sait conserver de l'information que sous forme numérique, sous forme binaire.

Exemple des images :

- D'un point de vue analogique, une image réelle est constituée d'une infinité de points. En effet, une image est un signal lumineux continu.
- D'un point de vue numérique, une image est un tableau (matrice) constitué d'un nombre limité de points (pixels) ayant chacun une couleur.
- Numériser une image va consister à capter l'image sous forme lumineuse et à enregistrer son équivalent numérique : une matrice de pixels.

1.2. Comment numériser ?

Document électronique	Un document électronique est bien souvent un fichier informatique (voire plusieurs) rédigé dans un format standard. Ainsi, il convient de faire la distinction entre : • L'information que contient le document, c'est-à-dire son contenu, à savoir le document à proprement parler ; • La représentation visuelle du document, à savoir le rendu qu'on obtient en ouvrant le document dans un logiciel (exemples : un document PDF ouvert sous Adobe Acrobat Reader ou encore une image ouverte sous Paint).
Fichier	Un fichier est une séquence d'octets à laquelle on a attribué un nom, le nom du fichier. La taille du fichier est le nombre d'octets qui le constitue. On distingue souvent les fichiers textes et les fichiers binaires. *Rappel : 1 octet = 8 bits. 1 bit = un 1 ou un 0.*
Format	Un format de fichier, c'est une manière d'organiser un fichier, autrement dit une façon de structurer l'information. On associe souvent une extension de fichier à un format : format XML (fichier « .xml »), format PNG (fichier image « .png »), format PDF (fichier « .pdf »)…



Exemple d'une image (format BMP) :

- Une image est un fichier binaire.
- Si on l'ouvre dans un logiciel permettant de visualiser l'image, tout va bien ! On obtient une belle image.
- Si on l'ouvre sou Notepad++, avec un plugin approprié, on visualise le réel contenu de l'image, tout de suite moins sympathique. Si on y regarde de plus près, on trouve des codes couleurs !

LSB \ MSB	0	1	2	3	4	5	6	7
	000	001	010	011	100	101	110	111
0	0000	NUL	DLE	SP	0	@	P	`	p
1	0001	SOH	DC1	!	1	A	Q	a	q
2	0010	STX	DC2	"	2	B	R	b	r
3	0011	ETX	DC3	#	3	C	S	c	s
4	0100	EOT	DC4	$	4	D	T	d	t
5	0101	ENQ	NAK	%	5	E	U	e	u
6	0110	ACK	SYN	&	6	F	V	f	v
7	0111	BEL	ETB	'	7	G	W	g	w
8	1000	BS	CAN	(	8	H	X	h	x
9	1001	HT	EM	)	9	I	Y	i	y
A	1010	LF	SUB	*	:	J	Z	j	z
B	1011	VT	ESC	+	;	K	[	k	{
C	1100	FF	FS	,	<	L	\	l	|
D	1101	CR	GS	-	=	M	]	m	}
E	1110	SO	RS	.	>	N	^	n	~
F	1111	SI	US	/	?	O	_	o	DEL

Exemple des fichiers texte (format ASCII) :

- Un fichier texte, c'est un fichier dont le contenu binaire correspond exactement à du texte, c'est-à-dire qu'on peut traduire les octets du fichier en caractères.
- Pour ce faire, on utilise une table de correspondance appelée jeu de caractères (exemple : ASCII, ANSI, UTF8, etc.). On dit que le texte a un encodage (exemple : encodage UTF8).
- A droite, une table ASCII (l'encodage le plus simple, sur 7 bits). Par exemple :
  - le caractère « espace » sera codé : 0000 010
  - le caractère « B » sera codé : 0001 100
  - le caractère « j » sera codé : 1010 111

Les fichiers XML ou encore HTML sont des fichiers textes. C'est pourquoi vous pouvez facilement en lire le contenu sous NotePad++ par exemple. Le XML comme le HTML sont des formats. Nombreux sont par ailleurs les logiciels qui définissent leur propre format de document (exemples : Word et ses documents DOCX, OpenOffice et ses document ODT, Excel et ses documents XLSX, etc.).

1.3. Pourquoi numériser ?

Au sein d'une organisation, on utilise de multiples documents : contrats, devis, factures, courriers… Les documents prolifèrent ! Les informatiser, c'est entre autres réaliser des économies de papiers. Mais pas seulement.

Dans une organisation comme sur internet ou à la télévision, de l'information, l'on en a partout ! Il importe de pouvoir s'y retrouver, c'est-à-dire de pouvoir trouver l'information et les documents dont on a besoin quand on en a besoin. Il importe également de pas perdre d'informations.

C'est là où la numérisation prend tout son sens ! Car la numérisation n'a d'intérêt que si les informations numérisées sont classées, si elles sont faciles d'accès ou encore si elles ne risquent pas d'être perdues. En effet, ce qu'on espère, c'est finalement gagner du temps. Et le seul fait de numériser ne suffit pas. Ce n'est que le premier pas. Il faut encore valoriser l'information numérisée.

C'est dans ce contexte que nous allons nous intéresser à la gestion électronique de documents mais encore au référencement.

2. Gestion électronique de documents

2.1. La gestion de documents

Les documents électroniques naissent, vivent et meurent. C'est le cycle de vie des documents.

Cycle de vie des documents	
Naissance	La naissance d'un document intervient essentiellement de deux manières différentes : • La création : consiste à produire un document numérique à partir d'un équipement informatique (numérique vers numérique) ; • L'acquisition : consiste à numériser un document et à le stocker sur support informatique (analogique vers numérique).
Vie	Plusieurs étapes sont susceptibles d'intervenir au cours de la vie d'un document : • La révision : un document peut faire l'objet de révisions, à savoir de modifications. L'on a parfois même besoin de conserver les versions consécutives d'un document ; • La publication ou la diffusion : un document n'a d'intérêt que s'il peut être consulté. La publication consiste en la mise à disposition du document. • La sauvegarde : un document peut être perdu. Afin d'éviter sa perte, l'on peut mettre en place un système de sauvegarde (exemple : duplication) ; • L'archivage : pour des raisons par exemple légales, on peut être amené à archiver des documents, c'est-à-dire à stocker et donc conserver le document sur le long terme ;
Mort	La vie d'un document s'achève par la destruction. Lorsqu'un document et son archivage sont devenus inutiles, le document peut finalement être supprimé.

On utilise souvent le terme de gestion électronique de documents, abrégé GED. Il existe même des logiciels spécialisés dans la gestion du cycle de vie des documents. Un tel logiciel est qualifié de gestionnaire électronique de document, également abrégé GED.

GED	L'acronyme GED signifie : Gestion(naire) Electronique de Documents. La GED consiste en un ensemble méthodes et technologies matérielles et logicielles offrant des fonctionnalités permettant d'assurer le cycle de vie des documents. Exemple de fonctionnalités : • Travail collaboratif sur un document (exemple : Google Document) ; • Partage de documents (exemples : DropBox, OneDrive…) ; • Automatisation de la production de document (exemple : rapport en format PDF) ; • Suivi des versions, appelé *versioning* (exemples : SVN, Git) ; • Classification et/ou indexation de documents ; • Sauvegarde et/ou archivage de documents ; • Etc.

2.2. La classification de documents

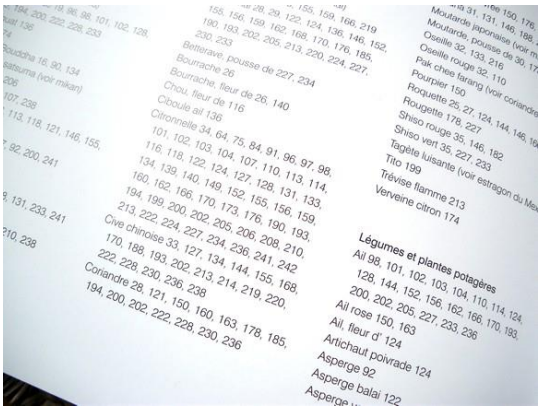
On a communément besoin de classer les informations, de les catégoriser, afin de les retrouver plus facilement et surtout plus rapidement. Il existe plusieurs modes de classification. En voici quelques-uns :

Utilisation de **Marque-pages**	Le principe du marque-page nous vient des livres. Il permet initialement d'identifier une page pour y accéder plus vite ultérieurement. Au sein d'un navigateur, sur le même principe, un marque-page permet d'identifier des pages web afin d'y accéder plus facilement ultérieurement.
Utilisation de **Post-it**	Un post-it, papier ou numérique, sert plus ou moins de memento. Il permet de stocker des informations importantes pour y avoir accès tout de suite.
Classification **Thématique**	Une classification thématique, c'est une classification par catégories, par thèmes. Par exemple, sur un site e-commerce, les produits sont souvent classés par thèmes : pantalons, chaussures, chemises, vestes… Une recherche par thème permet alors de trouver plus vite ce qu'on cherche.
Utilisation d'une **Taxinomie et de tags**	La taxinomie ou taxonomie est une méthode de classification de l'information. Elle consiste souvent à utiliser des tags, encore appelés marqueurs ou étiquettes. Un tag est un mot-clef. On va associer des mots-clefs aux documents pour faciliter le travail des moteurs recherche. Et lorsqu'on va fournir des mots-clefs à un moteur de recherche, il va chercher les documents utilisant ces mêmes mots-clefs.
Recours à la **Syndication de contenu**	La syndication de contenu est liée à la recherche d'actualités récentes. C'est un procédé consistant à mettre en place un flux d'actualités régulièrement renouvelé. On peut alors s'inscrire à ce flux, ce qu'on appelle la syndication. La source d'information va régulièrement renouveler son flux d'actualités et les personnes syndiquées recevront les nouvelles actualités. Un format de flux de syndication d'usage courant est le flux RSS. C'est un format XML.

Afin de rechercher facilement l'information, on a souvent recours à des moteurs de recherche amenés à

classifier l'information sans l'intervention des utilisateurs.

Moteur de recherche	Un moteur de recherche est un outil logiciel qui, en réponse à une recherche, va tenter de fournir les résultats les plus pertinents possibles. Autrement dit, un moteur de recherche va tenter, à partir de mots-clefs donnés, de retourner les documents ou informations correspondant le mieux aux mots-clefs fournis. Il fournit ce qu'on appelle des résultats de recherche. Le site de Google (*www.google.com*) est bien entendu un exemple de moteur de recherche. Il en existe bien d'autres et vous en utilisez d'ailleurs bien d'autres, et pas seulement sur internet.
Indexation	L'indexation est un procédé consistant à classer les documents typiquement en fonction de mots-clefs. En lisant les documents, on va construire ce qu'on appelle un index ou encore une table d'index ou d'indexation. L'objectif de l'indexation est de construire ce fameux index. En fouillant dans celui-ci, un algorithme (le moteur de recherche) va pouvoir, à partir des mots-clefs fournis par un utilisateur, retourner à ce dernier les documents ou informations répondant le mieux à sa recherche, c'est-à-dire les résultats correspondant le mieux aux mots-clefs qu'il a saisis.
Métadonnées	Les métadonnées sont des données connexes d'un document. Elles ne font pas partie du contenu du document mais viennent apporter des informations complémentaires telles que : • Mots-clefs ou tags facilitant la classification d'un document ; • Auteur ou copyright ; • Date de création, version, etc.



Exemple d'index de livre :

- Certains livres intègrent un index ;
- L'index d'un livre va associer à un mot-clef, voire à une expression clef, le n° des pages où l'on peut retrouver ce mot ou cette expression clef ;
- Un tel index est destiné à trouver plus vite l'information ;
- Par exemple, l'index à gauche nous indique que le mot « bourrache » peut être trouvé page 26.

Sur internet, c'est finalement le même principe, mais automatisé. Le moteur de recherche ne va pas indexer des numéros de pages mais des liens (URL) vers des pages web, vers des images, vers des documents PDF, bref, vers des ressources web. Le moteur de recherche va non seulement indexer ces ressources mais encore les classer par ordre de pertinence.

Important ! Il faut bien comprendre qu'un moteur de recherche, pour proposer des résultats de recherche, doit d'abord indexer les résultats. Sinon, il ne pourra pas les proposer… C'est ainsi que Google dispose de robots, c'est-à-dire de logiciels autonomes, parcourant continuellement la toile afin d'indexer les pages web, les images ou encore les documents PDF disponibles sur internet. Les robots de ce type sont appelés des crawlers.

3. Classification de l'information sur internet

3.1. L'information sur le web

Internet est peuplé d'informations. On parle souvent de ressources web : pages web, images, vidéos… Et tous les propriétaires de site(s) internet rêvent de voir leur(s) site(s) en tête des résultats de recherche. De fait, un site internet n'a d'intérêt que s'il est visible dans les résultats de recherche. Plus qu'être visible, il est même préférable pour un site internet qu'il apparaisse vite dans les résultats de recherche. En effet, plus un site apparaît « haut » dans les résultats, plus il a de chance d'être visité par les internautes.

Mais, un moteur de recherche comme Google indexe-t'il réellement l'information tout seul ? Bien, non ! On peut influer sur la visibilité d'un site internet.

Référencement ou **Optimisation pour les moteurs de recherche**	Le référencement est l'ensemble des procédés permettant d'améliorer la visibilité d'un site internet et de ses pages web (et autres contenus) dans les résultats de recherche des moteurs de recherche. On parle encore d'optimisation pour les moteurs de recherche, abrégée SEO pour *Search Engine Optimization*.

Remarque ! On notera que plus de 90% des internautes utilisent Google. On ne se préoccupe donc en pratique quasiment que du référencement sur Google.

3.2. Le référencement naturel et le référencement payant

Référencement naturel	Certains procédés non publicitaires favorisent la visibilité des pages web ou autres contenus dans les résultats de recherche. L'ensemble de ces procédés est qualifié de référencement naturel. Ils gravitent autour de deux axes : • La qualité des contenus (qualité des pages web en outre) jugées sur la bases de nombreux critères ; • La popularité des contenus (en particulier, la popularité du nom de domaine associé). On parle de *e-reputation*.
Référencement payant	Le référencement non naturel, ou plus simplement référencement payant, consiste tout bonnement à payer pour être vu. Parmi les outils de référencement payant bien connus, on retrouve : les campagne Google AdWords, les campagnes Facebook ou encore les campagnes Linkedin. De fait, la publicité est la plus grosse source de revenus des moteurs de recherche. En effet, ces derniers tirent souvent l'essentiel de leurs revenus de la diffusion des publicités des annonceurs*. C'est ainsi qu'en 2012, respectivement 69% et 27% des revenus de Google provenait d'AdWords et d'AdSense. * Un annonceur est une entreprise ou organisation qui diffuse de la publicité en vue de se faire connaître.

3.3. Les critères de référencement

Afin de classer les pages web en fonction de leur pertinence au regard d'une recherche, des algorithmes doivent être utilisés. Ainsi, Google utilise par exemple les algorithme Google Panda, Google Pingouin ou encore PageRank. Ces algorithmes sont en quelque sorte destinés à noter les pages web. Il existe ainsi des critères de pertinence. Nous en citerons quelques-uns.

Critères de popularité	La popularité ou e-reputation est évaluée sur la base de critères tels que : • Le nombre de liens directs ou indirects pointant vers une page web. Ces liens sont qualifiés de *backlinks*. L'algorithme PageRank vise par exemple à calculer une note de 1 à 10 en fonction du nombre de *backlinks* ; • L'activité sur les réseaux sociaux. Par exemple, plus les articles d'un site sont publiés sur les réseaux sociaux et reçoivent de commentaires ou encore de l*ike*, plus le site est supposé être populaire ; • Les avis des internautes ; • Etc.
Critères de qualité des contenus	La qualité du contenu des pages web est aussi importante, voire parfois plus encore parfois, que leur popularité. La qualité du contenu est évaluée en outre sur la base de critères tels que : • Les métadonnées de la page web doivent être « propres », en particulier : - l'url de la page : https://www.domaine.fr/ma/page/web - le titre de la page : <title>Le titre de la page web</title> - la description de la page : <meta name="description" content="La description" /> • La page web doit être bien structurée. Ses sous-titres doivent en outre être clairs et « propres » : - Question 5 : La numérisation suffit-elle à valoriser l'information ? - Numérisation - Qu'est-ce que la numérisation ? … - Gestion électronique de documents … • Les mots importants doivent être mis en valeur (en gras, en italique) : - référencement ou optimisation pour les […] - moteur de recherche

4. Echange d'informations et de documents

4.1. Problématique et solution

De même que les organisations s'échangent beaucoup d'informations, entre autres des documents électroniques, les logiciels ont souvent besoin de communiquer entre eux, à savoir de s'échanger des informations (exemple : envoi d'un mail à partir d'une boîte mail Outlook reçu par une boîte mail Gmail).

Par ailleurs, pour se comprendre, deux personnes ont besoin de discuter ensemble dans un langage commun. Pareillement, deux logiciels, pour se comprendre, doivent utiliser un langage commun. Nous

avons vu, au travers de la question 6 du programme (réseaux informatiques), que logiciels et matériels utilisent des protocoles pour s'échanger des informations en réseau. Toutefois, il faut encore que les logiciels puissent comprendre et traiter l'information qu'ils s'échangent. C'est ainsi qu'apparaît la notion d'interopérabilité.

Interopérabilité	Les logiciels sont souvent amenés à communiquer entre eux. Pour ce faire, ils doivent pouvoir s'échanger des informations dans un format donné. En informatique, l'interopérabilité est la capacité qu'ont deux systèmes informatiques à pouvoir échanger des données entre eux. Afin de faciliter les échanges et donc l'interopérabilité des systèmes, des formats standards ont été définis. En particulier, les langages XML et JSON sont des langages de description de données destinés à servir de formats intermédiaires dans le cadre d'échanges entre systèmes.

4.2. Langage XML

Le XML n'est pas un langage de programmation ! Il s'agit d'un format. Comme le JSON, il permet de structurer de l'information, de décrire des données.

Structure d'un document XML		
1	`<?xml version="1.0" encoding="UTF-8" ?>`	Prologue précisant la version XML et le jeu de caractères
2	`<catalogue>`	Elément racine (unique, englobant tous les éléments)
3	`  <manuel parution="2017">`	Balise ouvrante de l'élément *manuel*
4	`    <titre>Systèmes d'information de …</titre>`	Couple de balises *titre* imbriqué dans l'élément *manuel*
5	`    <niveau>Terminale</ niveau >`	Couple de balises *niveau* dans l'élément *manuel*
6	`    <section>STMG</ section >`	Couple de balises *section* dans l'élément *manuel*
7	`    <auteurs>`	Elément *auteurs* contenant 0 à N élément *auteur*
8	`      <auteur prenom="François" nom="DUREL" />`	Elément *auteur* avec attributs *prenom* et *nom*
9	`      <auteur prenom="Michèle" nom="ROY" />`	Les balises *auteur* sont auto-fermantes
10	`    </auteurs>`	Balise fermante de l'élément *auteurs*.
11	`    <!—manuel à recommander -->`	Commentaire non lu par les logiciels
12	`  </manuel>`	Balise fermante de l'élément manuel.
13	`  <manuel>`	Balise ouvrante d'un second élément *manuel*
…	`    …`	
…	`  </manuel>`	Balise fermant du second élément *manuel*
…	`  …`	
…	`</catalogue>`	Fermeture de l'élément racine et fin du fichier XML.

Le XML offre des avantages en matière de fiabilité :

- Un document XML doit obligatoirement être bien formé, c'est-à-dire que tous les éléments ouverts doivent être fermés et tous les éléments doivent être correctement imbriqués.
- Un document XML peut respecter un schéma défini au moyen d'un XSD ou d'une DTD. Un schéma XML définit les balises et attributs pouvant être utilisés par un document XML et la manière dont les éléments peuvent et/ou doivent être imbriqués les uns dans les autres ;
- Un document XML est dit valide s'il respecte le schéma qu'il s'est engagé à respecter.

Le HTML est par exemple un format de XML. En effet, le HTML est bel et bien du XML. Il a son propre schéma. Et le schéma actuel du HTML est celui du HTML5.