Exercices : thème 2 - Question 5

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

La numérisation est une forme d'acquisition d'information. On parle de dématérialisation, à savoir du passage du papier au numérique. De l'information, les organisations, les particuliers ou encore le web en accumulent, accumulent, accumulent… L'on comprend dès lors pourquoi il devient très vite essentiel de classer les informations de manière au moins à pouvoir s'y retrouver. Aussi, nous allons tâcher de comprendre les problématiques liées en outre à la prolifération de documents et d'informations. Si bien qu'à l'issue de ce cours, vous devez être capable :

- *D'identifier les différentes opérations de numérisations, d'indexation et de classification de contenus ;*
- *De repérer les contraintes liées à ces opérations ;*
- *De faire la différence entre : contenu, format et présentation d'un document ;*
- *D'identifier les sources d'informations pertinentes, en particulier sur le web, et d'utiliser efficacement un moteur de recherche en vous appropriant son fonctionnement général.*

Exercice 1 : la GED

Numérisation	La numérisation de l'information consiste à créer une représentation numérique (binaire ou textuelle) d'une information non numérique et à la conserver sur support informatique, typiquement dans un ou plusieurs fichiers.
Document électronique	Un document électronique est bien souvent un fichier informatique rédigé dans un format normalisé. Ainsi, il convient de faire la distinction entre : • La structure du document, son contenu, à savoir le document à proprement parler ; • La forme du document, sa représentation visuelle, 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).
Cycle de vie des documents	La vie d'un document suit plusieurs étapes : • La création ou l'acquisition : consiste à numériser ou produire un document et à le stocker sur support numérique ; • La révision : un document peut faire l'objet de révisions, à savoir de modifications. L'on a parfois 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 ; • La destruction : lorsqu'un document ou son archivage est devenu inutile, il peut finalement être détruit.
GED	L'acronyme GED signifie : Gestion(naire) Electronique de Documents. La GED consiste en un ensemble de méthodes et de technologies matérielles et

	logicielles offrant des fonctionnalités permettant d'assurer le cycle de vie des documents. Exemple de fonctionnalités : • Travail collaboratif sur un document ; • Automatisation de la production de document (exemple : rapport en format PDF) ; • Suivi des versions (versioning) ; • Classification et/ou indexation de documents ; • Sauvegarde et/ou archivage de documents ; • Etc.

Questions :

1. Documents numériques

Le terme « numérique » vient du latin *numerus* (nombre, multitude) et signifie « représentation par nombres ». Le nom « analogique » provient du mot grec *analogos* signifiant « qui est en rapport avec, proportionnel ».

Le numérique (*digital* en anglais) est une représentation de l'information par un nombre fini de valeurs discrètes (qui sont, au final, codifiables en binaires, à savoir au moyen de deux états, 0 et 1, et donc traitables par les machines électroniques que sont les ordinateurs).

L'analogique est une représentation d'une grandeur physique par une fonction continue. Pour mesurer cette grandeur, il faut une interface analogique (caméra, micro, etc.). La mesure peut ensuite être éventuellement numérisée par un convertisseur analogique/numérique (CAN). Par exemple, pour la copie d'une image :

- L'analogique consistera à essayer de reproduire à l'identique de ce qui est observé (avec un risque de déperdition d'information).
- Le numérique consistera à isoler chaque point de l'image (pixeliser) et à le caractériser (position, couleur) ce qui facilite sa reproduction à l'identique.

Document	**Analogique**	**Numérique**	
		Numérisé	Création native
Compte rendu saisi sur traitement de texte	☐	☐	☐
Rapport sur traitement de texte issu d'une reconnaissance vocale	☐	☐	☐
Photographique prise par un appareil numérique	☐	☐	☐
Diaporama créé sous PowerPoint	☐	☐	☐
Photocopie papier d'un cours	☐	☐	☐
PDF issu d'un cours scanné	☐	☐	☐
Captation vidéo en MPEG d'une conférence	☐	☐	☐
Fichier de données d'un capteur sismique sous tableur	☐	☐	☐

Température lue sur un thermomètre à mercure	☐	☐	☐

2. Indexation de documents

2.1. *Qu'est-ce qu'un index dans un document ?*

……
……
……

2.2. *Une indexation de tous les mots d'un texte est-elle pertinente ? Proposer une solution alternative.*

……
……
……

La recherche de termes dans un document PDF s'effectue par défaut par balayage de l'intégralité du texte du début à la fin (ce qui peut prendre du temps pour un long document).

Un index peut être constitué au préalable : il recensera les mots pertinents du document et leur emplacement dans une base de données. Les recherches s'effectueront ensuite dans cet index (plus court que le document) et seront donc plus rapides.

2.3. *A titre d'exemple, on souhaite établir l'index des termes des deux paragraphes ci-dessus. Lister les termes à indexer.*

……
……
……

2.4. *Le moteur de recherche Google indexe des images selon différents critères (essayer Google, puis Images et enfin Outils de recherche). Indiquer pour chacun de ces critères s'il est fondé sur des métadonnées (indexation textuelle) ou sur le contenu graphique de l'image (analysé automatiquement).*

Taille :		Couleur :	
Type :		Période :	
Droit d'usage :			

3. Conservation de documents

3.1. *La corédactrice du site de Madame Tulipe a procédé à des modifications du contenu de certains articles publiés sur le site en question. Comment appelle-t-on ce type d'opérations ?*

……

3.2. *Madame Tulipe s'aperçoit qu'il advient régulièrement que des erreurs soient introduites au fil des modifications. Malheureusement, le site internet ne permet aucun « retour en arrière ». Proposer une solution à Madame Tulipe.*

……..

3.3. *En vous servant du site www.service-public.fr, répondre à la question suivante : au sein d'une entreprise, quelle est la durée d'archivage minimale des documents suivants ?*

Relevés de comptes bancaires :	
Factures* :	
Bulletins de paie :	

*une facture est une pièce comptable.

Exercice 2 : l'indexation

Lorsque la volumétrie d'information augmente, il importe de classer / classifier les informations ou les documents. Citons quelques procédés de la vie quotidienne :

La classification thématique	On croise de nombreux sites et portails opérants une classification thématique. Un CDI ou encore une bibliothèque opère une classification thématique. Un site de vente en ligne (e-commerce) opère souvent une classification thématique (par types de produits).
Les marque-pages	L'on utilise aussi bien des marque-pages papier que numériques. On les utilise par exemples pour les « favoris » dans le navigateur (sous-entendu, les pages favorites).
La syndication de contenus	La syndication de contenu consiste à s'abonner à un fil d'actualités afin de pouvoir récupérer les « news » (exemple : les flux RSS).
La taxinomie	La taxinomie ou taxonomie consiste à associer des contenus à des mots-clefs, parfois appelés tags.

En matière de web, les internautes ont recours aux moteurs de recherche généralistes (Google essentiellement, Bing également) ou spécialisés (Youtube par exemple) afin de rechercher les documents qui les intéressent : pages web, images, vidéos, etc. Dès lors, le rôle d'un moteur de recherche est de proposer des contenus pertinents au regard des mots saisis par l'utilisateur.

Indexation	L'indexation de page web mais encore de documents PDF voire d'images consiste à analyser les documents afin de les associer à des mots-clefs. Il s'agit de compléter un index afin, à partir de mots-clefs, de pouvoir trouver les contenus correspondant. Pour indexer les contenus du web, les moteurs de recherche possède des robots qui parcourent continuellement la toile, procédé qu'on appelle le *crawling*.
Référencement naturel	Certains procédés non publicitaires favorisent la visibilité des pages web et 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.
Métadonnées	Les métadonnées sont les données connexes d'un document. Elles ne font directement 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.
Optimisation pour les moteurs de recherche	L'optimisation pour les moteurs de recherche, communément appelée SEO pour *Search Engine Optimization*, consiste dans l'ensemble des procédés permettant d'améliorer la visibilité des pages web et autres contenus dans les résultats de recherche.

Questions :

1. La référencement pas si naturel

PageRank

L'algorithme PageRank (PR) évalue la popularité d'un site internet par note de 1 à 10 (en échelle logarithmique : passer de 1 à 2 n'équivaut pas à passer de 8 à 9). Cette technologie a été développée en 1996-1997 à l'Université de Stanford et a donné lieu à la création de Google en 1998 (par Larry Page et Sergey Brin), qui en a fait son principal outil de classement des résultats de son moteur de recherche.

1.1. *Comparer le PR de sites mondialement connus (google lui-même) ou à audience plus restreinte sur :* www.pagerank.fr *ou* www.calcul-pagerank.fr.

……
……
……

Selon le PR, ce sont principalement les liens (directs ou indirects) pointant vers un site qui font sa popularité.

1.2. *Le critère du PageRank est-il qualitatif ou quantitatif ?*

……

1.3. *Identifier une manipulation possible du PageRank, i.e. un moyen de fausser le PageRank d'un site.*

……
……

Trust Rank

L'algorithme Trust Rank (« indice de confiance ») développé par deux chercheurs de l'université de Stanford et un de Yahoo!, note les sites entre 0 (équivalent à du spam) et 1 (site de confiance, tels que les sites gouvernementaux, les sites de référence, par exemple le W3C, ...).

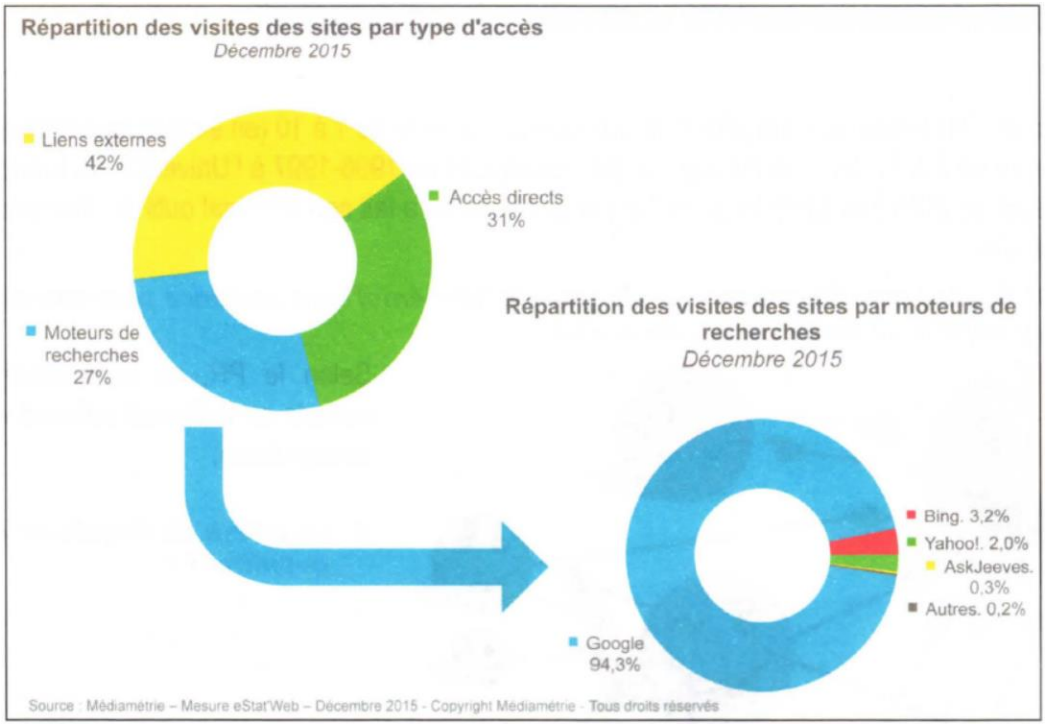
Présenté en 2004, il a aussi été adopté par Google en 2005 (déposant même une marque TrustRank) pour le combiner avec le Page Rank.

Panda et Pingouin

Le référencement est dit naturel lorsqu'il n'est pas issu d'un accord commercial entre le moteur de recherche et le site proposé (comme c'est le cas pour les liens commerciaux). Les moteurs de recherche admettent, voire encouragent, l'optimisation des sites en vue de leur référencement par les moteurs de recherche (SEO, *Search Engine Optimization*) car cela doit permettre de faire ressortir des contenus pertinents et de qualité. Par contre, ils luttent contre les manipulations de leur algorithme de recherche.

À partir de 2010, Google est confronté à des sites qui parviennent à se placer « haut » dans les résultats sans apporter de réelle valeur ajoutée car il s'agit de « fermes de contenus » qui agrègent, souvent sans autorisation, d'autres contenus, tout en les agrémentant de publicité : comparateurs de prix, annuaires, sites d'affiliation pour des bons de réduction ...

En février 2011, la mise à jour Google Panda fait chuter leur trafic jusqu'à 80 %. Là où Panda s'attaque aux contenus (notamment dupliqués), Google Pingouin traque, à partir d'avril 2012, les *backlinks* (liens pointant vers des sites) artificiels et le « bourrage » de mots-clés.



Le référencement étant une « guerre », certains peuvent créer des liens artificiels vers le site d'un

concurrent afin que celui-ci soit présumé coupable et déréférencé (« *negative SEO* »). Google doit donc maintenant proposer à un site de désavouer un lien qui pointe vers lui.

1.4. *Quel est l'enjeu du référencement par un moteur de recherche pour un site ?*

……
……
……

Référencement social

En 2009, Google abandonne le *Trust Rank* au profit d'un nouveau brevet, *Search Result Ranking Based On Trust*, qui tient compte des annotations des internautes en tant que vote de confiance, un *Person Rank*. C'est une approche en lien avec les réseaux sociaux : le « like » remplace le « link ». Et le « like » devient un critère de popularité d'un site internet.

1.5. *Quel est l'intérêt du référencement social ? Quel est le problème si l'on se cantonne à prendre en compte les « like » ?*

……
……
……

Search Engine Optimization

Dans le cadre de l'indexation de pages web, les moteurs de recherche ne valorisent pas tous les contenus textuels de la même manière. En raison de leur situation dans une page web, certains mots (mots-clefs) sont présumés avoir plus d'importance et être plus pertinents ! C'est au développeur à placer non seulement les bons mots-clefs mais encore les mots-clefs en bon endroit.

En particulier, les moteurs de recherche valorisent tout particulièrement les trois métadonnées :

Métadonnée	Mise en œuvre
L'URL de la page web	Rien ou réécriture d'URL.
Le titre de la page web	<title>Le titre de la page web</title>
La description de la page web	<meta name="description" content="La description de la page web" />

1.6. *Dans un document PDF comme dans une page web, quels sont les contenus textuels qui ont logiquement plus d'importance ?*

……
……
……

Finalement, pour faciliter la vie des systèmes d'indexation des moteurs de recherche, un site doit normalement posséder un ou plusieurs *sitemap*. Le *sitemap* est un document XML.

1.7. *Accéder au sitemap de l'Elysée (www.elysee.fr/sitemap.xml). Que contient un sitemap et à quoi cela peut-il bien servir aux moteurs de recherche ?*

……
……
……

2. Le référencement payant

AdWords	AdSense
Principe : payer pour être référencé sur le moteur de recherche Google ou des sites partenaires (AdSense) en achetant des mots-clés.	*Principe :* être payé pour afficher des publicités sur son site.

2.1. *Calculer les revenus d'un site en hypothèse basse et en hypothèse haute avec les données suivantes :*

• Chaque clic rapporte 5 à 50 centimes (selon le type de produits).
• 0,1 à 0,2% des visiteurs cliquent.
• Le site enregistrera 2 000 à 4 000 visiteurs par mois.

……
……
……
……

2.2. *Qu'est-ce que le CPC ?*

……
……

Exercice 3 : l'interopérabilité

Interopérabilité	Les logiciels produits en général des fichiers dans un format donné. Un format est une manière d'organiser et de stocker les données. 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.

	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.

Questions :

RSS signifie « Really Simple Syndication » (souscription vraiment simple) ou « Rich Site Summary » (sommaire développé de site).

Un flux RSS (ou fil RSS ou canal RSS) informe automatiquement ses abonnés des dernières nouveautés d'un site sans qu'ils aient besoin de se rendre sur le site. Pour l'éditeur du site, cela doit permettre de fidéliser les visiteurs et d'augmenter le trafic.

L'internaute s'abonne, de façon anonyme et gratuite, au flux dans son navigateur, courrieleur ou un agrégateur de flux client lourd (exemple : WebBulle) ou léger (exemple : NetVibes).

Le flux RSS fait appel à un fichier XML qui stocke notamment le titre et un résumé des nouveaux contenus ainsi que des liens directs vers l'intégralité de ces contenus. Voici le fichier du flux RSS d'un site de lycée :

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<rss version="2.0">
  <channel>
    <title>Lycée Jules Verne, Limours</title>
    <link>http://www.lyc-verne-limours.ac-versailles.fr</link>
    <description>Nouveautés du site du lycée Jules Verne de Limours</description>
    <item>
      <title>L'agenda du mois</title>
      <link>http://www.lyc-verne-limours.ac-versailles.fr/site/Arbo/Events/EvAgendaMois.php</link>
      <pubDate>Mon, 01 Feb 2013 00:00:00 GMT</pubDate>
      <description>Les événements du mois.</description>
    </item>
    <item>
      <title>Association sportive
      <link>http://www.lyc-verne-limours.ac-versailles.fr/site/Arbo/Events/AS/EvAS.php</link>
      <pubDate>Wed, 08 Jul 2012 00:00:00 GMT</pubDate>
      <description>Présentation de l'association sportive (AS) du lycée.</description>
    </item>
    <item>
      <title>Bilan d'étape du projet Développement durable</title>
      <link>http://www.lyc-verne-limours.ac-versailles.fr/site/Arbo/Educo/Projets/GalerieDD.php</link>
      <pubDate>Sat, 06 Jun 2012 00:00:00 GMT</pubDate>
      <description>Cette nouvelle page présente un bilan d'étape du projet.</description>
    </item>
  </channel>
</rss>
```

1. *Quelle information le prologue de ce document nous donne-t-il ?*

……

2. *Quel est l'élément racine de ce document ?*

……

3. *Ce document est-il bien formé ? Justifier.*

……

4. *Écrire l'arborescence de ce document (signaler d'un+ les éléments pouvant se répéter).*

……
……
……
……
……
……
……

Une nouveauté est ajoutée aujourd'hui au flux RSS : elle a pour titre « Nouvelle page d'accueil » et sa description est « Mise en ligne d'une nouvelle photo du lycée ».

5. *Écrire les lignes à insérer au fichier XML (en précisant l'emplacement).*

……
……
……
……
……

Question complémentaire : au moyen de la table ASCII fournie dans le cours, traduisez le texte « Animaux fantastiques » en binaire.