SÉANCE DOCUMENTATION : INDEXATION ET RECHERCHE DE DOCUMENTS

Objectif : Comprendre la notion d'indexation (définitions et méthodes) et le fonctionnement des moteurs de recherche web et BU. Evaluer la fiabilité des documents obtenus lors d'une recherche. Connaître les normes de citation des ressources documentaires (livres, articles, sites web…)

A - Indexation

L'**indexation** des documents analogiques (ex. livres papier) ou numériques est le **préalable** indispensable à toute **recherche** suivant des critères (mots-clés entre autres), de documents pertinents pour notre besoin. Historiquement, seuls les documents de bibliothèques étaient systématiquement indexés par des bibliothécaires, à l'aide de fiches puis de bases de données, permettant ensuite de retrouver un livre ou autre document. Avec l'informatique puis Internet, la problématique de l'indexation s'est généralisée d'une part à nos documents personnels et d'autre part au web.
Les catalogues de bibliothèque font l'objet d'une indexation en partie encore réalisée *par des humains*, professionnels (bibliothécaires ou documentalistes)(*cf. infra*)
Le système d'exploitation (O.S.) d'un ordinateur indexe *automatiquement* les fichiers qu'il contient, ce qui rend possible la recherche de fichiers dans l'index, à l'aide du moteur interne à cet O.S.
Les moteurs web indexent eux aussi *automatiquement* une partie du web (*cf. infra*)
La recherche d'un document numérique va reposer sur plusieurs éléments importants : les éléments significatifs du document/fichier, et son indexation dans une base de données (catalogue, système de fichiers de l'ordinateur, ou index d'un moteur pour le web). Il est important de noter que **l'indexation est un processus très gourmand en ressources de calcul** (ordinateur ralenti, nombreux serveurs nécessaires aux moteurs web), aussi bien dans la réalisation que dans le stockage de l'index qui peut être volumineux (l'index de Google fait la taille du web).

L'INDEXATION D'UN DOCUMENT PORTE SUR SA STRUCTURE ET SUR SES MÉTADONNÉES

Qu'est-ce qu'un document ?

En **sciences de l'information**, un ***document*** est un **support** -papier, électronique ou autre- sur lequel sont consignées des **données** qui peuvent être utilisées pour consultation, étude ou preuve.
En **informatique** et en particulier sur Internet, on utilise le terme plus général de ***ressource***, qui est un fichier ou une base de données constitués d'éléments d'information structurés et intelligibles stockés sur un support électronique (*in silico*), et créés au moyen d'un logiciel.

Données et métadonnées

Le **contenu** d'un document sera une ***information primaire*** en sciences de l'information ou *des* ***données*** en informatique. Le contenu du document est lui-même *structuré* : il contient des titres, des chapitres qui facilitent la lecture du document. Cette hiérarchisation permet au lecteur de repérer les concepts importants : mots de niveau "titre", mots mis en évidence par une mise en forme (gras, souligné), mots répétés plusieurs fois. Dans le cas d'un document numérique, il faut toutefois noter que les niveaux de titre doivent être indiqués par le créateur, par exemple en utilisant les *styles de titre* dans un traitement de texte.
À ce contenu, s'ajoutent des **informations sur le document** lui-même (ou **propriétés**) : elles seront nommées ***informations secondaires*** en sciences de l'information et ***métadonnées*** en informatique *("***données sur les données***")*.
Les métadonnées existent pour tous les types de documents. Par exemple, les métadonnées suivantes sont associées à chaque ouvrage (*i.e.* livre papier ou numérique) :

- Son numéro ISBN (International Standard Book Number) unique d'identification
- Le ou les auteurs

- L'éditeur, la collection, l'année d'édition
- Le résumé
- Les mots-clés qui caractérisent le contenu de l'ouvrage.

Des **métadonnées** sont également associées à des documents numériques. Elles sont en général regroupées dans les *propriétés* du fichier. Par exemple, un document texte, créé sous LibreOffice ou MS Office, a pour métadonnées le nombre de pages, de caractères, le propriétaire du logiciel, le format (traduction du fichier en binaire spécifique au logiciel utilisé), la version du document, sa date de création, la version du logiciel utilisé, les mots-clés caractéristiques du contenu du document… Les métadonnées d'une photographie sont la date de création, le type d'appareil photo, la résolution utilisée, la sensibilité ISO, la vitesse de prise de vue, l'ouverture ou diaphragme, l'auteur (si renseigné), la taille de la photographie et parfois même les données de géolocalisation (lieu de prise de la photographie).
Ces informations secondaires ou métadonnées peuvent être intégrées, dans le cas d'un document textuel comme un livre, article… sur les premières pages ou la couverture et/ou dans une zone non affichée du fichier (cf. l'exemple des logiciels de Traitement de texte) ; et dans le cas des documents non textuels comme des photographies, dans le fichier. Lorsqu'elles sont saisies automatiquement par le dispositif (logiciel, appareil photo…), ces métadonnées ne sont pas toujours connues du rédacteur. Elles peuvent être lues et affichées par certains logiciels ou appli (ex. logiciels de cataloguage de photos ou de lecture de fichiers musicaux).
Sur le web, les éléments de contenu importants des pages web, mis en évidence (titres, mots mis en forme, mots répétés) et les métadonnées des pages web (balises <meta>) sont mémorisés dans *un index* par les moteurs de recherche qui s'appuieront sur cette base pour répondre à la requête d'un internaute.

L'INDEXATION RELIE LE DOCUMENT À DES MOTS-CLÉS.

L'indexation d'un document, opération réalisée automatiquement par des programmes consiste à **associer un ensemble de mots-clés à un document**. Tous les outils de recherche reposent sur l'utilisation de **mots-clés** qui peuvent être des **mots** ou des **expressions**.

L'indexation par des humains

Cas des Bibliothèques :
L'indexation pour la recherche dans les **catalogues de bibliothèques** et les **bases de données documentaires** est l'œuvre de **documentalistes** spécialisés. Les mots-clés utilisés sont pris dans une **liste contrôlée** (un **thésaurus**) afin d'assurer que tous les ouvrages ou tous les articles relevant d'une même thématique soient associés à un même ensemble de mots-clés. Le choix de la casse (majuscule/minuscule, caractères accentués ou non) est contrôlé. Les catalogues de bibliothèque, maintenant en ligne, sont des bases de données décrivant les documents papier ou numériques.
Cas du web :
Sur le web existe aussi une indexation par des humains, les internautes, **non professionnels.**
Aux premiers temps du web existaient des annuaires publics - classifications hiérarchiques, par catégories thématiques, de sites web, avec une indexation parfois contributive.
Sont ensuite apparus des services en ligne d'**indexation,** de **stockage et partage de signets** (signets = URLs de sites web "intéressants"), de **veille** pouvant être **coopératifs** comme les services gratuits mais commerciaux **diigo.com, Netvibes** (dashboard en ligne) ou bien **Zotero** (libre et gratuit). Ils sont basés sur un principe contributif (Web 2.0) et conduisent à une indexation « collective » dans laquelle c'est la recommandation des internautes qui permet de trouver des sites pertinents. Chaque internaute crée son propre répertoire (annuaire personnel) de sites web qu'il publie sur un site dédié. Il peut commenter ses choix, adjoindre des mots-clés (on parle de **tags** ou **d'étiquettes***)* en **langage naturel**. Il peut également partager ses trouvailles au sein de groupes, ce qui revient à construire des annuaires collectifs de sites web.
Enfin s'est développé le **microblogging** ex. Twitter qui constitue un outil de **veille collaborative** :

les sites "intéressants" ayant été sélectionnés par des humains, ils sont très pertinents, et les mots-clés, marqués par le caractère # (hashtag), permettent de retrouver les *tweets* correspondants. Cependant, Twitter ne permet pas à l'internaute de stocker de façon structurée et réutilisable facilement les URL des sites trouvés et/ou les tweets.
Avec ces outils collaboratifs et donc une indexation par les internautes, basée sur le **langage naturel**, les mots-clés utilisés dépendent de l'expertise de l'indexeur pour le domaine concerné, et ne sont pas choisis suivant de règles strictes notamment pour le choix du genre (féminin/masculin), singulier/pluriel, la casse, etc. Des documents relevant d'une même thématique ne seront pas obligatoirement associés aux mêmes mots-clés car il n'existe pas de thésaurus de mots-clés, ce peut être problématique. Une indexation par mots-clés choisis par les internautes s'appelle une ***folksonomie*** (= folk taxonomy)

L'indexation automatique par les robots des moteurs de recherche : moteurs web

L'indexation par des programmes automatiques ou robots (*crawlers, spiders*) diffère de l'indexation humaine par deux aspects : les robots indexent des **pages web (**contrairement aux annuaires ou listes de sites web créés par des humains qui indexent en général des sites complets) et les robots **ne sélectionnent pas** les pages web sur leur qualité…
Lors de l'indexation, une première opération automatique consiste à éliminer les « mots vides » ou « mots-outils » (articles, conjonction de coordination…) de la page. Une seconde opération consiste à « lemmatiser » le corpus, c'est-à-dire à supprimer accord, conjugaison, etc. Ensuite, seuls les mots les plus fréquents sont conservés comme mots significatifs.
Les robots indexent également les **métadonnées** contenues dans l'en-tête de la page web (non visible par l'internaute) lorsqu'elles existent : le titre de la page web (balise <title> en html), l'auteur, la description, les mots-clés libres affectés à la page par l'auteur. Ils indexent aussi les métadonnées associées aux autres documents (media : images, son, vidéos…) par exemple les données techniques EXIF (Extended Image Format) des images. Enfin, ils indexent automatiquement les mots significatifs des textes.
Les moteurs de recherche appliquent des formules de pondération pour affecter un poids plus élevé sur certains mots. Par exemple, Google affecte un poids plus élevé si le mot est dans le titre du document (balises <H1>, <H2> etc.> en html), en début de document ou en majuscules.
Les robots n'indexent pas :

- Les pages accessibles par abonnement (journaux, bases de données…)
- Les éléments dynamiques des pages : les robots pourront indexer le formulaire de recherche mais pas le résultat d'une recherche. Par exemple, les moteurs de recherche n'indexent pas les horaires de la SNCF, ni les ouvrages d'une bibliothèque, car ces informations sont enregistrées dans une base de données.
- Les sites dont les webmasters ont interdit l'indexation (balise <meta name="robots" content="noindex"> en html)
- Les robots se déplacent de lien en lien pour découvrir de nouvelles pages. Ceci a pour conséquence qu'une page orpheline (sans lien) ne sera pas indexée automatiquement.
- Les documents trop volumineux.

On appelle **WEB invisible** les ressources du WEB non indexées par les robots. Or, c'est la "partie immergée de l'iceberg" !

B - Affichage des résultats d'une recherche

LES HUMAINS PUBLIENT DES CATALOGUES, LISTES OU ANNUAIRES, FIABLES VOIRE COMPLETS.

Bibliothèques : après avoir indexé les ouvrages ou les revues, les documentalistes vont créer une fiche descriptive du document (la **notice** des catalogues de bibliothèques) contenant les mots-clés de référence et les métadonnées. Cette notice sera ensuite intégrée au catalogue, qui est une base de données, dotée d'une interface utilisateur permettant une recherche par moteur interne.

Sites de partages de signets : ils permettent à l'internaute de partager de façon pérenne et structurée les URL avec une classification thématique (ex. Zotero) parfois hiérarchisée, ou des tags ; souvent ils sont dotés de moteurs internes également.
Twitter : partage instantané de liens, recherche par mots-clés possible, mais pas optimale, seule la recherche sélective par personne ou par hashtag # est possible.
(historiquement, les annuaires de recherche, qui sont des classifications thématiques hiérarchisées de sites web, permettaient d'accéder à des sites en explorant l'arborescence, et/ou avec un moteur)

LES MOTEURS DE RECHERCHE AFFICHENT LES RÉSULTATS « LES PLUS PERTINENTS » MAIS PAS LES PLUS FIABLES.

La méthode de tri des résultats des moteurs de recherche est un compromis entre la fourniture des résultats les plus « pertinents » et le temps de la recherche. L'algorithme de recherche et classement (*algorithme de ranking*) est le secret de fabrication du moteur, protégé par un brevet non publié.

Le tri par pertinence s'appuie sur l'importance du mot.

L'affichage des réponses à la requête par mots-clés de l'internaute résulte d'un *calcul de score* pour chaque URL, basé sur quatre critères :

- Le poids d'un mot dans l'index
- La densité du mot dans le document : si deux documents comportent un mot de même fréquence, le document le plus court sera favorisé (et donc placé avant le document le plus long).
- Le poids du mot dans la base de recherche : les moteurs de recherche présentent en premier les URL contenant les mots les moins usités (selon le principe que l'utilisation d'un mot peu utilisé est plus précis donc le document a des chances d'être plus pertinent). Les mots les moins fréquents sont favorisés, d'où l'importance d'utiliser des synonymes.
- Un document contenant une expression identique à celle de la requête sera favorisé. Le choix de l'ordre des mots est important.

Le tri par popularité s'appuie sur l'intelligence collective.

Les moteurs de recherche tentent de fournir les sites les plus en adéquation avec les requêtes postées par les internautes. Par exemple, pour classer ses résultats, GOOGLE se base :

- sur la popularité de la source (PageRank) : analyse des liens pointant vers la source.
- sur l'analyse des recherches des internautes
  Étude statistique des liens cliqués par les internautes à partir d'une requête identique
- sur l'achat des mots-clés par les entreprises (référencement payant)
- sur un « indice de confiance » (TrustRank) mal connu. Il s'appuierait entre autres sur l'analyse du nom de domaine, le nombre de pages du site, l'audience du site.

Plus récemment, GOOGLE et QWANT s'appuient sur l'évolution technique du WEB (vers le **WEB sémantique**) pour proposer des liens vers des sites connexes à la requête, c'est-à-dire qui n'ont pas les mots-clés exacts de la requête postée par l'internaute mais dont le sujet reste proche. Le WEB sémantique ou le WEB des données (*web of data*) constitue la prochaine évolution du WEB qui engendrera une évolution des modes d'interrogation des moteurs.

Le référencement payant : les entreprises achètent des mots-clés

Lorsqu'une entreprise désire que son URL soit affichée bien classée sur les pages de résultats d'un moteur, elle « achète » des mots-clés. Lorsque ces mots-clés sont cherchés par un internaute, l'URL sera affichée sur la page de résultats. Les URL « commerciales » se distinguent des autres URL par une couleur de fond différente, et un positionnement différent dans la page (obligation légale).

Éthique des moteurs de recherche

Les moteurs de recherche sont des outils extrêmement puissants et apparemment gratuits mais à quel "prix" (caché) pour nos données personnelles ? Le développement de ces outils nécessite des financements importants et donc le **modèle économique** des sociétés qui développent ces outils se

pose. Ainsi les géants Google ou Microsoft (moteur Bing) ont choisi la gratuité d'usage de leur moteur mais, en contre-partie, utilisent les données des internautes (traces de navigation, création de profils-types) pour proposer aux entreprises une plateforme de publicité ciblée. **"Si le service est gratuit, c'est que vous n'êtes pas le client, mais la marchandise qui est vendue"**

Des tentatives de création de moteurs plus éthiques apparaissent, sans que les auteurs aient la puissance financière des géants du WEB :

- **DuckDuckGo** : ce moteur de recherche américain qui utilise les données de *crowdsourcing* (production collaborative de données) pour améliorer la pertinence des résultats proposés. Il n'enregistre aucune donnée privée (adresse IP, signature du navigateur…).
- **Qwant** : moteur de recherche européen qui n'enregistre pas les traces des internautes, ni les données privées. Il est financé pour partie par des fonds publics. Qwant se différencie par la présentation des résultats. Et il recherche des informations dans les médias sociaux.

C - Conseils pour une recherche documentaire

MÉTHODE

1) ouvrir le sujet (rechercher les définitions des mots du sujets, consulter une encyclopédie)
2) définir les mots-clefs de recherche (citer ces mots-clés dans le rapport documentaire)
3) Synthétiser l'information recueillie en fonction du public visé

USAGE D'UN MOTEUR DE RECHERCHE (EX : GOOGLE)

Pour saisir sa requête, on peut utiliser l'interface de *recherche simple,* et la préciser avec des "opérateurs", ou l'interface de *recherche avancée* (ex. avec Google : après la recherche simple, utiliser la barre d'outils, ou directement google.fr/advanced_search)

Opérateurs booléens :

- **« »** : les guillemets permettent de rechercher une expression exacte. *« Le blog du Modérateur »* présente les sites où les mots *Le blog du Modérateur* sont présents, uniquement dans cet ordre.
- **–** : le signe *moins* permet d'exclure un terme. La requête *astuces recherche -Google* permet de connaître les pages contenant *astuces* et *recherche*, mais exclut celles qui contiennent le terme *Google*.
- .. : deux nombres séparés par deux points permettent de recherche tous les nombres de la plage spécifiée. *Smartphone 200..400 euros* liste les téléphones compris entre 200 et 400 euros.
- **AND** : exclut les pages ne contenant pas les termes spécifiés. *Blog AND Modérateur* présente les sites contenant ces deux termes, mais pas ceux en contenant un seul.
- ***** : l'astérisque est souvent utilisé pour connaître l'intégralité d'une phrase ou d'une expression. *Qui vole * vole * *permet de retrouver l'expression *qui vole un œuf vole un bœuf.*
- **OR** : l'opérateur permet de rechercher un terme, ou un autre.

Opérateurs avancés les plus utiles :

- **site:** permet de rechercher les pages web d'un site spécifique. *site:blogdumoderateur.com* liste uniquement les pages web du blog du Modérateur.
- **filetype:** limite la recherche au type de fichier spécifié. *filetype:PDF* permet de rechercher uniquement les documents PDF, *Filetype:epub* des livres numériques, etc.

ÉVALUER LES DOCUMENTS TROUVÉS

Identifier le type de source documentaire :

1) sources traditionnelles (agences de presse, grandes organisations, associations et sociétés savantes, publications scientifiques…)
2) source émergeant de structures contributives plus ou moins formelles (avec ou sans règle de

validation de l'information, ex : l'encyclopédie collaborative Wikipedia a des règles d'écriture et des règles de validation comme le fait de citer les sources lorsqu'on écrit un article)
3) sources informelles (blogs, pages personnelles, RS etc)

Déterminer la fiabilité et la pertinence du document :

Fiabilité : **RIB** (Bon signe → Mauvais signe)

R = Réputation (auteurs ou institut reconnu ; vérification par les pairs → auteurs inconnus)

I = Intention (Fait et dialectique → Propagande et publicité)

B = Bibliographie (Précise et fouillée ; réputation des références → Bibliographie absente)

Pertinence du document pour ma recherche : **PAF**

P = Public (Expert ---> Grand public)

A = Adéquation (Thème englobant ; Exactement le sujet ; Plus spécifique)

F = Fraîcheur de l'information (Date de mise à jour récente → Absence de date ou document très anciens)

CITATIONS : RÉDACTION DES RÉFÉRENCES BIBLIOGRAPHIQUES

Pourquoi citer ses sources ?

- pour attester de la fiabilité de son propre travail
- pour attribuer une idée originale ou découverte à son auteur, par honnêteté et souci de justice
- pour permettre au lecteur d'aller plus loin en consultant lui-même les sources

Comment citer ses sources ?

Quand on utilise des ressources, que ce soit pour en citer des extraits ou pour y faire référence, il faut dresser la liste des **références bibliographiques** en respectant certaines règles. Le plus souvent la liste des références bibliographiques est classée en fonction de l'ordre alphabétique des noms d'auteur, puis pour un même auteur par ordre chronologique des publications. **Les références obéissent à des normes de présentation** : elles doivent contenir des éléments précis, placés dans un ordre défini, caractérisés par une typographie particulière et reliés par une ponctuation standard. Les modèles proposés ici s'appuient sur la norme internationale **ISO 690:1987**, complétée par la norme **ISO 690-2:1998** en ce qui concerne les références de documents sur supports électroniques.

Règles générales de présentation

Qui ? 'NOM' de l'auteur de préférence en majuscules suivi d'une virgule, 'Prénom' en minuscules (sauf la première lettre en majuscule) suivi d'un point (autant que possible, indiquer le prénom en entier).

Quoi ? 'Titre du document', 'sous-titre' (s'il y a lieu), le tout en italique, mention de l'édition (lorsque disponible, ne jamais mentionner pour la 1ère édition). NB : titre et sous-titre sont séparés par " : "

Où ? Quand ? 'Lieu d'édition' suivi de ' : ' et 'Nom de l'éditeur' suivi d'une virgule, 'date d'édition' suivi d'un point. 'Nom de collection' (facultatif), 'numéro' dans la collection (facultatif). S'il s'agit d'une ressource en ligne, il faut également préciser l'adresse complète de la ressource (URL), ainsi que la dernière date de consultation de la ressource.

Combien ? 'Nombre de pages' suivi de ' p.', ce nombre peut être précédé, s'il y a lieu, de la mention du nombre de feuillets ou de planches (en chiffres romains)

Numéro normalisé (ISBN) : tous les ouvrages publiés depuis 1972 reçoivent un numéro d'identification unique appelé ISBN (international standard book number). Ce numéro doit être reproduit à la fin d'une référence d'ouvrage, de chapitre d'ouvrage.

D - Outils spécialisés de recherche

RECHERCHE D'IMAGE

Les recherches d'images peuvent s'effectuer :
-soit avec un moteur de recherche textuel, qui va rechercher les mots-clés dans les méta-données des fichiers images, dans la balise HTML <ALT=" "> (destinée aux malvoyants), ou dans le texte qui accompagne l'image (mais les algorithmes indexent mal les images, malgré des progrès...)
-soit avec un moteur de **recherche d'images par comparaison** entre une image de référence, préalablement téléchargée ou dont on fournit l'URL, et des images accessibles sur le Web. Pour cela, il faut utiliser un moteur comme Google Images ou une autre banque de données d'images (voir plus bas). L'icône de l'appareil photo présent dans la barre de recherche permet de télécharger l'image de référence ou d'indiquer le lien. Le moteur proposera alors des images similaires.
Il existe également des logiciels d'identification d'images spécialisés– couplés à des banques d'images – comme Pl@ntNet, spécialisée dans les images de végétaux. Ce logiciel compare une photo de plante préalablement téléchargée aux photos contenues dans sa base de données et propose ensuite une identification de la plante. La base de données est créée sur la base du *crowdsourcing.*

EXEMPLE DE BANQUES D'IMAGES (VERIFIER LES DROITS)

- **Wikimedia Commons** (libre) : http://commons.wikimedia.org/wiki/Accueil
- **Flikr** (en partie sous CC) dont The Commons : http://www.flickr.com/commons (images fournies par les musées, les bibliothèques publiques, les agences gouvernementales, etc.)
- **Pl@ntNet** (sans but lucratif) : https://plantnet.org/

CARTOGRAPHIE

- **GoogleMap** est le service de cartographie en ligne de Google. Il a été lancé en 2004 aux Etats-Unis et en 2006 en France. Les données collectées sont la propriété de Google. En 2013, Google a racheté Waze, logiciel GPS et Google Maps s'est enrichi d'une fonction GPS, notamment sur les terminaux mobiles.
- **OpenstreetMap** est un service de cartographie en ligne libre et contributif, c'est-à-dire qu'il se base sur le travail coopératif des internautes dans le but de proposer une alternative à Google. Il a été mis en route en juillet 2004 au University College de Londres. OpenstreetMap France est une association loi 1901. Les données collectées par l'association sont accessibles à tous sous licence libre OBdl et permet à tous de réutiliser ces données.
- **Geoportail** est le portail de cartes géographiques en ligne et images satellites de l'IGN (Institut Geographique National), concernant la France donc (web et appli).

De nombreux appareils photos et téléphones enregistrent les coordonnées GPS dans les photos lors de l'enregistrement des photos, pour peu que le propriétaire ait donné le droit d'accès des fonctions GPS de sa machine à l'appareil photo. Dans ce cas, les méta-données techniques sont enregistrées dans les données EXIF (Extended Image File Format) de la photographie. Les données EXIF sont accessibles via les propriétés du fichier image. Certains logiciels de traitement d'images comme les Irfanview permettent d'afficher les informations (Image > Information > EXIF Info) et de lancer la géolocalisation du lieu dans Openstreetmap, Googlemaps, etc.

RECHERCHE DE LIEU

La recherche de lieux s'effectue soit avec des **coordonnées GPS**, soit par **adresses**, si celle-ci existe dans la base de données des logiciels. La recherche sur les noms de ville (par exemple : Vierzon) renvoie une ou plusieurs propositions. Le lien « Où est-ce ? » dans le champ de recherche d'OpenStreetMap renvoie les coordonnées GPS en format décimal.
Rechercher une coordonnée GPS : entrer le point en coordonnées décimales séparées par une virgule : latitude, longitude. Le séparateur de décimales est un point (ex : 11.3).
Attention, il existe deux type de coordonnées GPS :

- le format DMS (Degré, Minute, Seconde)
- le format DD (Décimal). Les minutes et les secondes y sont converties en heure.

L'orientation Ouest et Sud sont comptées négativement et l'orientation Est et Nord positivement.

E - Le droit des auteurs et des œuvres

LA PROTECTION DES ŒUVRES

Contrairement aux brevets qui concernent les inventions, le droit d'auteur et le copyright concernent les "**œuvres de l'esprit**" : œuvres littéraires et artistiques (ex. musique, films ou vidéos, photographies), mais aussi logiciels, jeux vidéos, design…
Concernant la protection des œuvres, il faut être conscient que les réglementations et les usages diffèrent d'un pays à l'autre. En France, c'est le **droit d'auteur** qui protège les œuvres de l'esprit. Il se compose du droit moral et de droits patrimoniaux.
Le droit moral reconnaît la paternité de l'œuvre (l'identité de son auteur) et protège son intégrité. Ce droit est perpétuel et incessible. Le droit moral est constitué du droit de paternité (obligation de citer l'auteur du document), du droit de divulgation (l'auteur a le choix de diffuser son œuvre ou non et d'en choisir les conditions), du droit de repentir (l'auteur peut modifier son œuvre même après qu'elle ait été diffusée), du droit de retrait (l'auteur a le droit de demander l'arrêt de la diffusion de son œuvre) et du droit de respect de l'intégrité de l'œuvre (l'auteur peut exiger que l'œuvre ne soit pas diffusée par morceaux).
Les droits patrimoniaux (communément appelés droits d'exploitation) concernent **le droit de reproduction** (imprimer, copier et numériser) et **le droit de représentation** (communiquer et diffuser). Ils permettent à l'auteur (ou à ses héritiers) d'être rémunéré pour chaque utilisation de l'œuvre. Ces droits peuvent être cédés à un tiers "ayant-droit" (un éditeur, par exemple) gratuitement ou avec une contrepartie financière. Ces droits perdurent 70 ans après la mort de l'auteur au bénéfice des héritiers ou autres ayants-droit.
Les **œuvres du domaine** public peuvent être utilisées gratuitement à condition de citer l'auteur (respect du droit moral). Il s'agit des œuvres dont les droits patrimoniaux ont expiré ; des œuvres volontairement placées dans le domaine public par leurs auteurs, des œuvres non originales (lois, textes réglementaires, etc.).
Il existe des exceptions au droit d'auteur : le **droit de courte citation**, **l'exception pédagogique et de recherche** en font partie (usage d'œuvres protégées en cours, ou usage pour des travaux de recherche), et quelques autres cas.
Dans les pays anglo-saxons, c'est le **copyright** qui protège les œuvres de l'esprit. Le copyright ne concerne que les droits patrimoniaux. En France, les mentions "Copyright", © ou "Tous droits réservés" n'ont aucune valeur juridique. Elles ont seulement un rôle informatif permettant d'identifier la personne à contacter pour demander l'autorisation d'exploitation.

Un peu d'histoire

La protection juridique de ces œuvres n'a pas toujours existé : le copyright date de 1710, en Angleterre (*Statute of Anne*) et le droit d'auteur français de la révolution française (1791-1792) à l'initiative du dramaturge Beaumarchais. A l'époque, ces régimes juridiques avaient été conçus pour permettre à l'auteur de vivre de ses œuvres sans être spolié par les "exploitants" (libraires-éditeurs, directeurs de théâtre…) et la protection était limitée (en Angleterre, 14 ans après publication, en France, 10 ans après la mort de l'auteur) mais la durée n'a fait que s'allonger ensuite… au détriment de l'arrivée de l'œuvre dans le Domaine public.

LES LICENCES DES RESSOURCES NUMÉRIQUES

Toute ressource du web français est soumise au droit d'auteur (qui relève de la loi). Dans certains cas, est associée à l'œuvre une **licence** (*i.e.* un **contrat d'utilisation**) facilitant ainsi sa réutilisation : il suffit de se référer aux termes de cette licence, sans avoir à demander d'autorisation (parfois payante) à l'auteur ou aux ayants-droits. De fait c'est l'auteur, titulaire du droit d'auteur par la loi, qui

peut *choisir* d'associer une licence à son œuvre.
Une licence formalise les conditions d'utilisation et de distribution d'une œuvre. Il existe différentes catégories de licences. Les **licences "libres"** permettent d'utiliser l'œuvre, de la redistribuer, de la modifier pour créer une œuvre dérivée, et de redistribuer l'œuvre dérivée, y compris commercialement. Cette licence peut imposer que toute copie ou œuvre dérivée soit diffusée avec la même licence : clause de **copyleft** (un jeu de mots sur copyright) ou "partage dans les mêmes conditions". L'avantage du copyleft est que les contributions apportées par les uns et les autres profitent au plus grand nombre d'utilisateurs, sans que personne ne puisse s'approprier l'œuvre. Les licences de logiciels libres, la licence de documentation libre de GNU (GFDL) ou certaines licences Creative Commons sont des licences libres. Les licences **"ouvertes"** ou de **"libre diffusion"** accordent aussi à l'utilisateur le droit d'utiliser et rediffuser l'œuvre - au moins à titre gratuit- mais peuvent imposer des **restrictions** : interdiction de modifier l'œuvre ou d'en faire une exploitation commerciale. Certaines licences Creative Commons sont des licences ouvertes.

LES 6 LICENCES CREATIVE COMMONS

Creative Commons signifie **"Communs créatifs"**, un concept qui s'applique aux œuvres de l'esprit et s'inscrit dans la lignée de la défense des "communs" ou "biens communs", un mouvement politique et philosophique à la fois ancien et en plein essor.
Inventées en 2002 par le juriste américain Lawrence Lessig pour "décompresser" le système du copyright, le système des 6 licences Creative Commons permet à l'auteur de spécifier par avance quels droits il concède à l'utilisateur. La paternité de l'auteur doit toujours être citée. En associant une licence Creative Commons à une œuvre, l'auteur choisit de céder par avance à tous, de façon non exclusive, ses droits patrimoniaux -au moins pour des usages gratuits. Par contre, l'auteur peut définir un certain nombre d'options, comme autoriser ou pas les usages commerciaux, autoriser ou pas les modifications. Pour associer une licence, il suffit d'intégrer le logo précisant les options choisies. Le texte juridique formel de chaque licence CC est publié sur son site web par l'organisation américaine Creative Commons, et traduit en français sur CC France. A l'inverse, la licence associée à une œuvre consultée ou téléchargée nous renseigne sur les droits que l'auteur nous accorde. Une 7e licence, CC 0, est la plus proche possible de la notion de "domaine public".

CC – BY / Attribution
L'utilisateur final doit citer l'auteur lorsqu'il utilise l'œuvre. Tous les usages (y compris avec modifications, et commerciaux) sont autorisés.

CC – BY– SA / Attribution – Partage dans les mêmes conditions
L'utilisateur final doit citer l'auteur. Il est tenu de partager l'œuvre qu'il a créé à partir de la première œuvre dans les mêmes conditions.

CC – BY - ND / Attribution – Pas de modification
L'utilisateur final doit citer l'auteur. Il ne peut pas modifier l'œuvre et doit la diffuser dans son intégralité.

CC – BY – NC – SA / Attribution – Pas d'utilisation commerciale
L'utilisateur final doit citer l'auteur. Il est interdit de commercialiser l'œuvre ou toute création issue de celle-ci..

CC – BY – NC – SA / Attribution – Pas d'utilisation commerciale – Partage dans les mêmes conditions.
L'utilisateur final doit citer l'auteur. Il est tenu de partager l'œuvre qu'il a créé à partir de la première œuvre dans les mêmes conditions. Il est interdit de commercialiser l'œuvre ou toute création issue de celle-ci.

CC – BY – NC-ND / Attribution – Pas d'utilisation commerciale -Pas de modification
L'utilisateur final doit citer l'auteur. Il ne peut pas modifier l'œuvre et doit la diffuser dans son intégralité. Il est interdit de commercialiser l'œuvre ou toute création issue de celle-ci.

CC0 (proche du Domaine Public)
En France, l'utilisateur final doit citer l'auteur (pas aux USA). La CC0 permet au titulaire des droits d'auteur de renoncer par choix au maximum légal de ses droits au profit du public.

LE "COPIER-COLLER" ET LE PLAGIAT

Pour faire au plus simple, le **plagiat** est le fait de recopier à l'identique ou en la modifiant légèrement tout ou partie de l'œuvre d'une personne, et de s'en attribuer la paternité. C'est inacceptable éthiquement puisque le plagiaire n'a pas fourni de travail original contrairement à l'auteur, qui se trouve alors lésé. Dans le droit français, le plagiat relève de la **contre-façon**, qui est un délit, c'est-à-dire qu'un plagiat peut être puni d'une amende allant jusqu'à 500 000 € et/ou d'une peine de 5 ans de prison. Dans le monde académique (université-recherche), le plagiat ou le copier-coller est un fléau pour la communauté scientifique et la crédibilité des résultats publiés. En effet, chaque publication scientifique doit être originale et doit citer correctement les travaux antérieurs sur le sujet, avec le nom de leurs auteurs.
Au niveau des études universitaires, le copier-coller de travaux présentés comme siens relève de la **tricherie**, comme le fait de copier sur un voisin en examen. Si l'acte de copier-coller est avéré, le cas relève du conseil de discipline de l'université : l'étudiant peut être frappé d'interdiction de passage de tout diplôme national pendant 5 ans.
Du fait de la multiplicité des cas de fraudes, plusieurs entreprises ont développé des solutions logicielles qualifiées **"anti-plagiat"**. En fait, il s'agit de logiciels qui fonctionnent un peu sur le modèle des moteurs de recherche. Ces logiciels lemmatisent le texte soumis, c'est-à-dire qu'ils déterminent les mots-clés les plus pertinents, puis comparent cette liste de mots-clés avec des listes disponibles dans leur base et complètent cette comparaison avec une recherche sur le web. Ils rendent ensuite comme réponse un **% de similitude** entre le travail soumis et les diverses ressources possibles ou probables trouvées par le logiciel.

Si on est de bonne foi et qu'on veut être sûr de ne pas plagier : bonnes pratiques

Il faut concevoir et rédiger un travail original ! Certes on peut emprunter certaines idées originales à un auteur mais il faut le créditer de ces idées. On peut citer *entre guillemets* un *court extrait* en spécifiant son *auteur*. Il faut plus globalement citer ses sources dans la *bibliographie* à la fin.