Architectures et technologies WEB

Introduction aux technologies WEB

V01.00 17/08/2016 - MAURICE

Plan de la séance

- Chapitre 8
  - Le Référencement

Déroulement du cours

- *Chapitre 1 : Préambule*
- *Chapitre 1 : Les technologies WEB*
- *Chapitre 2 : La Programmation cliente*
- *Chapitre 3 : Le modèle MVC / Les templates*
- Chapitre 4 : La sécurité
- *Chapitre 5 : Les bases de PHP / Le Contrôleur*
- *Chapitre 6 : PHP 2ème partie / ORM*
- *Chapitre 7 : SOA / ROA*
- **Chapitre 8 : Le référencement**

Vocabulaire

- **SERP** (*Search Engine Result Pages*) désigne les résultats de recherche
- **SEO** (*Search Engine Optimization*) désigne l'ensemble des techniques permettant d'améliorer la visibilité d'un site web
- **Soumission** (en anglais submission) consistant à faire connaître le site auprès des outils de recherche ;
- **Positionnement** (en anglais ranking) consistant à positionner les pages d'un site en bonne position dans les pages de résultat pour certains mots-clés
- **SEO White hat** désigne les référenceurs respectant scrupuleusement les consignes des moteurs de recherche aux webmasters, dans l'espoir d'obtenir un référencement durable en jouant avec les règles du jeu
- **SEO Black hat** designe les référenceurs adoptant des techniques contraires aux consignes des moteurs de recherche, dans le but d'obtenir un gain rapide sur des pages à fort potentiel de monétisation, mais avec un fort risque de déclassement

Vocabulaire

- **NetLinking** désigne le fait d'obtenir des liens externes pointant vers son site web car cela augmente d'une part le trafic et la notoriété de son site
- le **crawl** (en français *exploration*) designe l'exploration de votre site par les robots d'exploration des moteurs de recherche
- lien **NoFollow** designe un lien indiqué aux moteurs de recherche comme à ne pas suivre

Bonnes Pratiques HTML

- Ajouter le Doctype en première ligne
- Fixer l'Encodage en fonction des caractères utilisés dans vos pages (UTF-8, ISO8859-1)
- Code HTML de structure limité, réduire l'utilisation des balises neutres,
- Placer les Styles et les Javascript dans des ressources externes
- Les Documents HTML doivent être bien formé
- Utilisez un Balisage sémantique,
- Utilisez les microdata HTML5

Bonnes Pratiques HTML

- Créez des URL adaptées, reformatez les URL,
- Utilisez des balises META décrivant précisément le contenu de la page,
- Evitez les tables, les frames,
- Proscrire les balises dédiées à la présentation,
- Rédigez un contenu original et attractif,
- Prenez soin des corps de texte pour qu'ils soient lisibles par les moteurs

Bonnes Pratiques HTML

- Choisissez bien votre nom de domaine
- Ajoutez un titre spécifique à chaque en correspondance avec son contenu
- Définissez les attributs title et alt des balises A et IMG
- Utilisez les balises h1,..,h6 pour organisez vos documents.
- Utilisez les balises et pour mettre en exergue des expressions que vous jugez importantes, et qui constituent par extension l'intérêt de votre page web.
- **Balisez sémantiquement votre contenu**

Bonnes Pratiques

- Privilégiez le contenu, insister sur les mots clés à mettre en avant
- Évitez la répétition des titres, des paragraphes dans les différentes page de votre site.
- Évitez les appels AJAX
- Développez votre réseau et vos liens entrants.
- Soumettez votre site web dans les moteurs de recherches.
- Utilisez des outils de suivi et de mesure d'audience. Par exemple Google Analytics

Le Contenu

- Le contenu de votre page web est considéré comme son intérêt principal.
  Vous **DEVEZ** absolument travailler correctement le début de votre article, pour qu'il soit accrocheur.
- **La règle d'or des 600 px :**

Tout doit être dit dans les 600 premiers pixels de votre page web.

Les mots clés

- **Établir une liste de mots-clés pour son site**
- Il faut bien réfléchir aux mots-clés que vous souhaitez pour votre site : un bon mot-clé est un mot-clé qui correspond à deux facteurs.
  - **Il doit être beaucoup recherché sur Google**
  - **Il doit bien correspondre à votre site**
- Les zones suivantes devront être riches en mots-clés :
  - vos balises <title> ;
  - vos liens hypertextes mots-clés ;
  - balise <h1> ;
  - balises <h2> à <h6>.

Les mots clés

- Il existe des outils permettant de comparer le volume de recherche d'un mot-clé par rapport à un autre et donnant des suggestions : http://www.google.com/trends/?hl=fr

- Enfin, il existe des sites permettant de connaître les mots-clés des sites concurrents : SEMRush.com

Organisation des liens Internes

- Bien définir l'intitulé des liens
- Bien organiser les liens internes de votre site
- les liens constituent un enjeu très important dans votre stratégie de référencement.
- Maitrisez Le Google Juice ! (ou Link Juice)
- Vos fichiers doivent absolument être nommés correctement,
- Utilisez les tirets comme séparateurs les URL, evitez underscores « _ »
- Évitez les erreurs 404 !
- Récrivez vos URL

Développez vos liens externes

- **Définition d'un backlink**
- **Un lien entrant : c**e lien doit être fait « en dur » avec la balise de lien HTML <a Un backlink n'est donc rien d'autre qu'un lien hypertexte qu'un site A fait vers un site B.
- **Eviter les backlinks à éviter** issus de sites qui sont blacklistés par Google,
- Netlinking : inciter les autres à créer des liens
- **Le link ninja : consiste a insérer ces liens dans des Forum**
- Le link baiting repose en fait sur votre contenu. Le tout est de vous débrouiller pour produire un contenu de grande qualité qui va susciter de l'intérêt chez les internautes
- Le référencement grâce aux annuaires
- Exploitez intelligemment les réseaux sociaux

Balises META

- Balises META
  - Les META Tags sont des balises non affichées à insérer en début de document HTML afin de décrire finement le document.
  - Etant donné l'usage abusif des métas constaté dans un nombre important de sites web, les moteurs utilisent de moins en moins ces informations lors de l'indexation des pages.
  - La balise meta "keywords" a ainsi été officiellement abandonnée par Google

Balises META

- META description
  - –La balise meta description permet d'ajouter une description décrivant la page, sans les afficher aux visiteurs
  - –Il est conseillé d'utiliser le codage HTML pour les caractères accentués et de ne pas dépasser une vingtaine de mots clés.

```
<meta name="description" content="description de la page" />
```

Ne dépassez pas 160 – 165 caractères, espace compris, pour être sur que la description s'affiche entièrement dans la page de résultat des moteurs de recherche

- Rédigez 1 description unique par page.
- Utilisez la même description sur deux ou plus de pages vous coutera cher puisque les moteurs de recherche considéreront que vous avez dupliqué vos pages, ce qui est interdit en référencement.
- **L'outil gratuit Yakaferci permet d'analyser votre balise meta description**

Balises META

- META robots
  - –Elle permet de décrire le comportement du robot vis-à-vis de la page,
  - –indiquer si la page doit être indexée ou non et si le robot est autorisé à suivre les liens.
  - –Par défaut l'absence de balise robots indique que le robot peut indexer la page et suivre les liens qu'elle contient.
  - –La balise robots peut prendre les valeurs suivantes :
    - •index, follow : cette instruction revient à ne pas mettre de balise robots puisque c'est le comportement par défaut.
    - •noindex, follow : le robot ne doit pas indexer la page (toutefois le robot peut revenir régulièrement pour voir s'il existe de nouveaux liens)
    - •index, nofollow : le robot ne doit pas suivre les liens de la page (par contre le robot peut indexer la page)
    - •noindex, nofollow : le robot ne doit plus indexer la page, ni suivre les liens. Ceci se traduira par une baisse drastique de la fréquence de visite de la page par les robots.

```
<meta name="robots" content="noindex,nofollow"/>
```

Les Adworks

- Créez vos campagnes de publicité avec Google Adwords
- Il s'agit du programme de publicité de Google qui « complète » Adsense. Celui-ci est destiné aux annonceurs !
- **Adwords, un programme publicitaire**
- Avant toute chose, vous devez savoir que Google Adwords est un programme publicitaire mis en place par Google. Google est une entreprise qui vit en partie grâce à la publicité... Google Adwords occupe une place tellement importante dans l'univers de Google qu'il représente sa principale source de revenus.

Lynx

- Testez votre site avec Lynx
  - Lynx est un navigateur en mode texte, fonctionnant dans une fenêtre DOS.
  - Lynx est outil permettant d'évaluer l'accessibilité technique d'un site web, ou sa capacité à être indexé correctement par les moteurs de recherche
  - Lynx ne sait en effet afficher que du texte HTML, des élements de formulaires et des hyperliens.
  - Les images, les tables, les frames, les scripts ainsi que Flash et les éléments multimédias sont tous ignorés (les cookies peuvent être acceptés ou non).
  - Son comportement est donc très proche de celui d'un robot d'indexation, ou d'un système de navigation pour personnes aveugles (synthèse vocale, plage braille).

PageRank

- Inventé par Larry Page en 1997
- Indicateur permettant de classer les pages.
- Principe de base est d'attribuer à chaque page une valeur proportionnelle au nombre de fois que passerait par cette page un utilisateur parcourant le web.
- Selon le brevet Google, ces critères sont :
  - les liens entrants et sortants ;
  - les ancres ;
  - le trafic associé à la page ;
  - le comportement des internautes : le choix de la page dans les résultats ;
  - le nom de domaine ;
  - l'hébergement.
  - les sites

PageRank

- Google détecte et sanctionne les campagnes massives de liens artificiels
- Google intègre des critères qualitatifs à l'analyse des liens (sémantique, confiance , comportement des utilisateurs).
- TrustRank vient de Yahoo, il accorde des valeurs de confiance aux sites gouvernementaux, et aux sites de référence

Sitemap

- Le plan de votre site, ou *sitemap* est un fichier listant toutes les pages qui existent sur votre site.
- Le sitemap est une solution pour simplifier et accélérer le travail des crawler :
  - si votre maillage interne n'est pas bien effectué,
  - si votre site est très récent, peu de liens amènent vers les différentes pages

Sitemap

- un fichier sitemap est un fichier XML qui liste les URL d'un site web de façon à favoriser l'exploration du site par les moteurs de recherche.
- Un fichier sitemap XML doit :
  - Débuter par une balise d'ouverture <urlset> et terminer par une balise de fermeture </urlset>
  - Spécifier l'espace de nom (standard de protocole) dans la balise <urlset>
  - Inclure pour chaque URL une entrée <url> en tant que balise XML parent
  - Inclure une entrée enfant <loc> pour chaque balise parent <url>
  - le fichier XML doit être enregistré en UTF-8.
- Un sitemap ne peut lister qu'au maximum 50 000 URL et la taille du fichier XML ne doit pas dépasser 10 Mo (10 485 760 octets).
- Toutes les URL listées dans un fichier sitemap XML doivent provenir du même hôte,

Sitemap

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://mon-domaine.fr/</loc>
    <lastmod>2013-12-15</lastmod>
    <changefreq>daily</changefreq>
    <priority>1</priority>
  </url>
  <url>
    <loc>http://mon-domaine.fr/page-a.html</loc>
    <lastmod>2013-12-15</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
</urlset>
```

ROBOT

Nom du moteur	User-Agent
Google	GoogleBot
	Mediapartners-Google
	Googlebot-Image
	Googlebot-Mobile
	Google Wireless Transcoder
	AdsBot-Google
Bing	bingbot
	MsnBot
Yahoo	Slurp
Yandex	Yandex
Orange/Voila	VoilaBot
Cuil	Twikler

Robots.txt

- Contient des commandes à destination des robots d'indexation des moteurs de recherche afin de leur préciser les pages qui peuvent ou ne peuvent pas être indexées
- Tout moteur de recherche commence l'exploration d'un site web en cherchant le fichier *robots.txt* à la racine du site.

Robots.txt

- Le fichier *robots.txt* (écrit en minuscules et au pluriel) est un fichier ASCII se trouvant à la racine du site et pouvant contenir les commandes suivantes :
  - **User-Agent:** permet de préciser le robot concerné par les directives suivantes. La valeur ***** signifie "tous les moteurs de recherche".
  - **Disallow:** permet d'indiquer les pages à exclure de l'indexation. Chaque page ou chemin à exclure doit être sur une ligne à part et doit commencer par /. La valeur **/** seule signifie "toutes les pages du site".

Robots.txt

Exclusion de toutes les pages :

```
User-Agent: *
Disallow: /
```

Exclusion d'aucune page (équivalent à l'absence de fichier *robots.txt*, toutes les pages sont visitées) :

```
User-Agent: *
Disallow:
```

Autorisation d'un seul robot :

```
User-Agent: nomDuRobot
Disallow :
User-Agent: *
Disallow: /
```

Robots.txt

```
Exclusion d'un robot :

User-Agent: NomDuRobot
Disallow: /
User-Agent: *
Disallow:

Exclusion d'une page :

User-Agent: *
Disallow: /repertoire/chemin/page.html

Exclusion de plusieurs page :

User-Agent: *
Disallow: /repertoire/chemin/page.html
Disallow: /repertoire/chemin/page2.html
Disallow: /repertoire/chemin/page3.html

Exclusion de toutes les pages d'un répertoire et ses sous-dossiers :

User-Agent: * Disallow: /repertoire/
```

Les pénalités

- Le cloaking (en français dissimulation) est une technique proscrite par les moteurs de recherche, consistant à générer un contenu HTML différent selon qu'il s'agit d'un visiteur ou d'un moteur de recherche.
- En effet il est possible de repérer les robots des moteurs de recherche par la présence d'un champ User-Agent spécifique dans les requêtes HTTP qu'ils envoient et de leur fournir un contenu différent, comportant des mots clés en surnombre, non affichés aux visiteurs.
- Si l'utilisation de cette technique est constatée par un moteur de recherche, le moteur de recherche peut donner une pénalité, qui se traduira par une dégradation du positionnement du site, voire une éviction !

Les pénalités

- Le content spinning est une technique consistant à générer aléatoirement différents textes en utilisant des paraphrases.
- Cette technique est principalement utilisée par les spécialistes du référencement black hat pour créer un grand nombre de contenus unique sur la base d'un même texte de base.
- Les moteurs de recherche ont tendance à ignorer le contenu dupliqué, c'est-à-dire un même contenu existant sur différentes pages et n'en retiennent qu'une version.
- Il est à noter que cette technique est proscrite par les moteurs de recherche, notamment Google, qui indique clairement que le contenu généré automatique, c'est-à-dire tout contenu généré par des programmes, doit être bloqué par un fichier robots.txt afin de ne pas être référençable.

Rewriting d'URL

- **L' URL Rewriting - APACHE mod_rewrite**
- MOD_Rewrite du serveur Apache permet de réécrire les URL à la volée (appelé URL Rewriting).
- Ce module utilise des expressions régulières.
- Ouvrir le fichier **httpd.conf** dans le dossier Apache et de décommenter les 2 lignes suivantes en enlevant le # :

```
LoadModule rewrite_module modules/mod_rewrite.so
AddModule mod_rewrite.c
```

Rewriting d'URL

- Créez un fichier .htaccess dans les répertoires ciblés par le rewriting d'URL
- **`RewriteEngine on`**

  Cette instruction active la réécriture d'URL. Elle devra toujours être mise dans le fichier htaccess.

Rewriting d'URL

- **Réécrire des URL dynamiques**
- La réécriture des URL permet de présenter à l'internaute une url plus mnémotechnique, facilitant dans la foulée son indexation par les moteurs de recherche qui ne laisseront plus sur le côté des pages dynamiques avec de multiples paramètres.

```
http://www.monsite.fr/index.php?module=detail&ref=livre1
```

Rewriting d'URL

- http://www.monsite.fr/detail_livre1.html
- Deviendra : http://www.monsite.fr/index.php?module=detail&param=livre1

```
RewriteEngine on
RewriteRule ([a-z]+)_([a-z]+)\.html$ /index.php?module=$1&ref=$2
```

- http://www.monsite.fr/index.php?module=catalogue&action=12&param=1
  deviendra : http://www.monsite.fr/catalogue/12/1

```
RewriteEngine on
RewriteRule ([a-z]+)/([0-9]+)/([0-9]+).\html$
/index.php?module=$1&action=$2param=$3
```

Rewriting d'URL

- [L] Ce (drapeau) signifie que la règle est la dernière à être appliquée et que le module ne doit plus tenter de réécrire cette chaîne.

Les outils

- Google Webmaster Tools
- Bing Webmaster Tools
- Google Tendances des recherches
- ÜberSuggest - Suggestion de mots-clés
- Trouver les liens pointant vers une page
- Connaître le pagerank d'une page web
- http://www.yakaferci.com

Les références

- http://www.commentcamarche.net/contents/web/referencement.php3

Exercice

- Réécrivez les URL de votre site