Fiche référence

FR.1.0 - Août 2008

Ministère de la Culture et de la Communication

MOTEUR DE RECHERCHE AVANCÉE POUR LE SITE CULTURE.GOUV.FR

Contexte

Le site culture.gouv.fr est un site majoritairement statique composé de pages HTML et de documents bureautiques (PDF, RTF, documents Word, Excel…).

Le site propose énormément de contenus mais ne dispose pas d'un moteur de recherche, ce qui rend l'information difficilement accessible.

Par ailleurs, le contenu est dispersé sur plusieurs sites et il n'y a pas de points d'entrée offrant une vision globale des contenus.

Le Ministère de la Culture et de la Communication souhaite rapidement mettre à disposition des internautes un moteur de recherche, avant de travailler sur une refonte profonde du site.

Points clés du projet

Proposer un moteur de recherche centralisé permettant de rechercher parmi l'ensemble des sites référencés par le Ministère de la Culture et de la Communication.

La majorité de ces sites étant statiques, il est nécessaire de parcourir les sites pour récupérer leur contenu, aussi bien les pages HTML que les documents.

Proposer des fonctionnalités avancées au niveau de la recherche : recherche approximative, rapprochement grammatical…

Etre capable de maîtriser le contenu indexé : fréquence d'indexation, importance relative des pages, suppression de pages de l'index...

Briques technologiques

- Nutch
- Solr
- PHP

Description et phasage

- Utilisation de Nutch comme crawler (récupération des contenus) et de Solr pour indexer les contenus et traiter les requêtes de recherche .
- Développement d'un composant Nutch et de divers plugins permettant de faire le lien entre Nutch et Solr.
- Développement d'une interface de consultation du moteur en PHP et échanges avec Solr via des appels JSON

14 rue Gaillon – 75002 Paris-France
Tél. : +33 (0)1 42 682 800 – Fax : +33 (0)1 40 075 486
97, rue Racine – 69100 Villeurbanne – France
Tél. : +33 (0)4 26 689 500 – Fax : +33 (0)4 26 689 501
www.openwide.fr – contact@openwide.fr