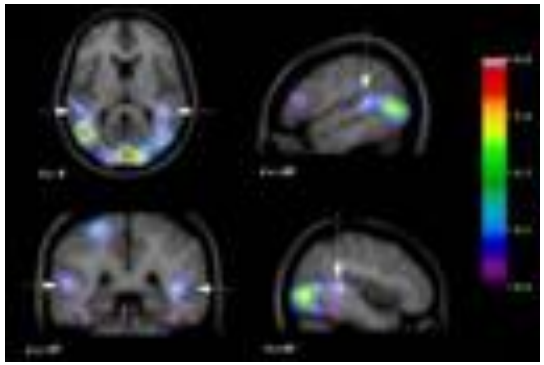
INSTITUT FRANÇAIS DE L'ÉDUCATION
RECHERCHE

ReVEA
Ressources vivantes pour l'enseignement et l'apprentissage

Thématiques, types de fichiers et fréquentation de sites web



ÉCOLE NORMALE SUPÉRIEURE DE LYON

Le besoin

ReVEA
Ressources vivantes pour l'enseignement et l'apprentissage

- Suivre l'utilisation des sites webs de ressources utilisés par les enseignants
  - Quels types de documents (texte seul, texte avec images, image, vidéo)
  - Sur quelles thématiques scientifiques
  - Document récent, actualisé ou plus ancien
  - Document « populaire » (souvent téléchargé)
  - Consultation ciblée ou plus globale
  - Moyen d'arriver à la ressource (moteur de recherche ou navigation sur le site)
  - Y a t il des influences géographiques

Données liées au document

ReVEA
Ressources vivantes pour l'enseignement et l'apprentissage

- URL d'accès au document
- Son titre
- Ses dates (création, mise en ligne, modification)
- Sa nature (html, image, pdf, excel, word)
- Présence d'éléments multimédia (son, vidéo)
- Sa thématique scientifique (tectonique, etc.)
- Ses auteurs
- Mots clés associés à la page

Données liées à la fréquentation

ReVEA
Ressources vivantes pour l'enseignement et l'apprentissage

- URL d'accès au document
- Adresse IP anonymisée de l'utilisateur
- Geolocalisation de l'utilisateur
- Date de la consultation
- Nombre de requêtes mensuelles sur la page
- Requête initiale ayant amené sur la page
- Moteur de recherche ayant amené sur la page

La réalisation technique

- 3 sites ciblés en SVT : Planet-Terre (PT), Acces et Planet-Vie (Jussieu Vie)
- Différentes natures technique des sites
  - Données et descriptions XML en base de données interrogeables de façon uniforme (Planet-Terre)
  - Descriptions XML et données sous Plone en bases de données hétérogènes (Acces)
  - Descriptions XML, données externes (Planet-Vie)
- Autres natures possibles
  - Aucune description, données en base (Plone, Mysql)
  - Aucune description, site externe

Sources des données

ReVEA
Ressources vivantes pour l'enseignement et l'apprentissage

- Base eXist des métadonnées LOM-FR (les 3 sites)
- Base eXist des données Docbook DOA (PT)
- Logs Apache des sites web (PT et Acces hébergés à l'ENS de Lyon)
- Métadonnées pour les sites Plone
- Système de fichiers pour des sites statiques

Traitements des logs Apache

ReVEA
Ressources vivantes pour l'enseignement et l'apprentissage

- Anonymisation et Geolocalisation
  - Géolocalisation en premier
  - premier prototype d'outil qui géolocalise avec une précision Ville à partir d'une IP
  - Anonymisation par renumérotation à partir de 0
  - Anonymisation par calcul non réversible sur le dernier octet de l'adresse IP
- Extraction d'autres informations des logs Apache
  - Date de consultation
  - Point d'entrée de la visite et origine (referer)
  - Nature, présence d'images dans une page

Traitements des données XML

ReVEA
Ressources vivantes pour l'enseignement et l'apprentissage

- Création d'une seconde base de données alimentée par l'analyse des bases eXist
  - Titre
  - Dates (création, mise à jour)
  - Nature (mise en avant sur le site ou réelle)
  - Thématique scientifique (Unisciel, Dewey)
  - Auteurs
  - Présence de sons, images, vidéos
  - Mots clés associés à la page

Exemple de ligne de log

- 195.83.134.20
- -
- -
- [06/May/2011:11:42:46 +0200]
- "GET /EducTice/logo_ife.png HTTP/1.1"
- 200
- 5555
- "http://eductice.ens-lyon.fr/EducTice"
- "Mozilla/5.0 (Windows; U; Windows NT 5.1; fr; rv:1.9.2.17) Gecko/20110420 Firefox/3.6.17 (.NET CLR 3.5.30729)"

Format de log combined

- "%h %l %u %t %r %>s %b %{Referer}i %{User-Agent}i"
  - %h : adresse IP du client
  - %l : un tiret “-”
  - %u : nom de l’utilisateur si connu ou “-”
  - %t : heure d’arrivée de la requête
  - %r : première ligne de la requête : GET suivi de l’URL et du protocole
  - %s : statut de la requête, 200 si pas d’erreur
  - %b : taille du retour en octets (taille du fichier envoyé) ou “-” si rien
  - %Referer : URL d’origine de la requête
  - %User-Agent : type (détaillé) du navigateur