Techniques de diffusion du droit

Module 5: Acquisition de documents

Ernst Perpignand

ernst@lexum.umontreal.ca

Soumission électronique

- La soumission électronique se fait suivant un protocole d'échange sur lequel les intervenants s'entendent
  - Moyen de communication
  - Format des documents
  - Bordereau

Moyen de communication

- Réseau
  - Email
    - Un programme utilisant une librairie permettant d'accéder à un compte de courrier électronique
    - JavaMail API
  - FTP

Moyen de communication

- HTTP
  - Interface Web de soumission
    - Page JSP, PHP
  - Téléchargement de site
    - Librairie permettant la navigation d'un site web
    - Java: HttpClient

Borderau de communication

- Format tabulé
  - simple
  - Peut s'éditer dans un classeur (Excell)
  - Convient mal aux valeurs multiples
  - Peu extensible
- Format XML
  - Extensible
  - Convient aux valeurs multiples
  - Un peu plus compliqué

Envoi

- En général, il est préférable de rassembler les documents et le bordereau dans un envoi
  - Disquette
  - Répertoire
  - Fichier compressé

Interface web de soumission

- Peut constituer un outil apprécié des éditeurs s'ils ont accès à Internet
- Pas nécessaire
- Peut être coûteux à développer
  - Rémunération des experts
- Technologies
  - Java (JSP, java beans)
  - PHP

Téléchargement de sites partenaires

- Solutions ad hoc
  - Concevoir un programme pour
    - Effectuer une connexion web sur le site partenaire
    - Extraire les liens vers les documents intéressants
    - Librairie java HttpClient
- Fragile au changement de l'organisation présentationnelle du site partenaire

Extraction des méta information

- Écrire des programmes pour lire le contenu en format texte ou HTML afin d'en extraire les infos
  - Expressions régulières (Perl, Java)
    - Référence\s*:\s*\d{2}-\d{2,4}
    - Trouve la chaîne : Référence : 24-1987/CC

Saisie manuelle

- Documents
  - Saisie simple
    - Precision 99%
  - Saisie double
    - Deux personnes saisissent séparément
    - Le texte est ensuite comparé par un logiciel qui surligne les différences (diff)
    - Une troisième personne fait la correction
    - Précision 99.95%

Saisie manuelle

- Livre (2500 à 3000 caractères par page)
  - 99.95% de précision : 1 à 2 erreurs par page
  - 99.995% (saisie triple): 1 erreur à toutes les sept pages
- Journal (6500 à 7500 caractères par page)
  - 99.95% de précision : 3 ou plus erreurs par page
  - 99.995% (saisie triple): 1 erreur à toutes les trois pages

Numérisation

- Permet d'obtenir un fichier électronique à partir d'un document papier
- La fiabilité des numériseurs a beaucoup augmenté, mais ils n'arrivent à reconnaître que 95% des caractères
  - Validation par un éditeur
  - On peut les utiliser pour la saisie initiale