Acquisition des documents

Plan

I – Sources
II – Stratégies d'acquisition
III – Récupération des métadonnées

I – Sources

Contexte

- Structure de diffusion étatique
  - Accès plus facile et direct
  - Accès parfois restreint par des luttes internes
- Structure de diffusion non-étatique
  - Nécessité de convaincre, d'obtenir des autorisations

Administration publique

- Ministère de la Justice, Secrétariat Général du Gouvernement, etc
- Parfois intéressés, mais pas toujours
  - Dépend de leurs projets concurents
  - Dépend du désir de rentabiliser la diffusion du droit
- Compétences technologiques variables

Cours / tribunaux

- Président, juges, centre de documentation, greffe, etc
- Généralement très intéressés à diffuser leurs jugements
  - Pour faire connaître leur travail
  - Augmente leur notoriété
- Compétences technologiques limitées

Imprimeurs officiels

- Entreprises parapubliques
- Ne sont généralement pas intéressés à partager les documents
  - Y voient une perte de revenus potentiels
  - Perdent leur monopole acquis dans l'univers papier
- Compétences technologiques importantes

Support d'acquisition

- Papier
  - Nécessite de la numérisation et de l'OCR
  - Doit être transporté physiquement
  - Implique des coûts importants
- Support électronique (Cédéroms, disquettes, clefs USB)
  - Doit être transporté physiquement
  - Nécessite de la main d'oeuvre pour retirer les documents du support
  - Implique des coûts réduits

Support d'acquisition (suite)

- Réseau (intranet, Internet)
  - Réception peut être totalement automatisée
  - Rapidité d'exécution
  - N'implique aucun coût d'acquisition

Format des fichiers électroniques

- Fichiers textes
  - Perte de beaucoup d'information sur la structure du document
  - Conversion facile
- Fichiers de traitement de texte
  - Souvent le document original
  - Un mauvais emploi des outils entraîne des problèmes de conversion
  - Plusieurs logiciels différents existes, dont plusieurs sont propriétaires (Word, WordPerfect, OpenOffice)

Format des fichiers électroniques (suite)

- Languages balisés
  - Nécessite une compétence technologique importante au niveau de la source
  - Perte de contrôle sur l'étape de la conversion
  - Diffusion facile

II – Stratégies d'acquisition

Cueillette manuelle

- Avantages
  - Assure l'exhaustivité des collections même lorsque la source collabore difficilement
  - Réduits les tâches au niveau de la source
  - Facilite les bonnes relations avec la source
- Inconvénients
  - Occasionne des coûts de main-d'oeuvre importants
  - Dépendance face au format de fichier utilisé

Transmission des documents sources

- Avantages
  - Coûts de réception peu élevés
- Inconvénients
  - Nécessite un traitement manuel pour la diffusion
  - Dépendance face au format de fichier utilisé

Téléchargement de sites Web

- Avantages
  - Permet l'automatisation de la diffusion
  - N'implique aucune tâche supplémentaire pour la source
- Inconvénients
  - Coûts de développement élevés
  - Doit être adapté à chaque modification du site Web
  - Dépendance face au format de fichier utilisé

Formulaires de soumission Web

- Avantages
  - La source possède un grand contrôle sur la diffusion
  - Les connaissances technologiques nécessaires sont minimes pour la source
  - Automatisation de la diffusion
- Inconvénients
  - Coûts de développement élevés
  - Rend les traitements de masse difficiles

Logiciel institutionnel intégré

- Avantages
  - La source possède un grand contrôle sur la diffusion
  - Prise en compte de la totalité des métadonnées
  - Automatisation de la diffusion
- Inconvénients
  - Nécessite une expertise technologique à la source
  - Processus de développement logiciel doit être intégré
  - Doit être adapté à chaque source

III – Récupération des métadonnées

Contexte

- Les métadonnées sont nécessaires à la création de l'interface et au moteur de recherche
- Plus il y a de métadonnées, plus l'interface peut être précis
- Exemple pour un jugement
  - Date
  - Noms des parties
  - Numéro de dossier

Processus manuels

- Consiste à ouvrir chaque fichier pour en extraire l'information requise
- L'information est ensuite
  - Insérée dans une base de données
  - Insérée dans un fichier de langage balisé
- Cette technique est souvent indispensable lorsque les documents proviennent de plusieurs sources
  - Le manque d'uniformité limite l'automatisation

Processus automatisés

- Intégré avec l'acquisition du document
  - Téléchargement, formulaires Web, logiciels intégrés
  - Ne requiert aucune tâche supplémentaire
- Reconnaissance automatique
  - Développement d'un programme qui lit les documents, identifie les étiquettes, récupère l'information qui suit et l'insère dans la base de données ou le fichier balisé
  - Uniquement si les documents sont uniformes
  - Requiert tout de même une vérification manuelle