



PROJET EXQI-LIBRE

la qualité des documents et des données non structurées

Contexte

Actuellement, les données structurées d'une entreprise croissent annuellement de 15 à 20% tandis que les données non structurées croissent de 50 à 200%. Cette croissance est due en partie à la numérisation des contenus des documents, au développement des échanges de type Web 2.0 et aux publications des communautés scientifiques ou techniques.

La qualité de l'information des documents textuels et des images est devenu un sujet de plus en plus crucial en particulier lorsqu'il s'agit de migrer ces documents d'un système d'information à un autre ou lorsqu'il s'agit de lire et d'interpréter les documents ou encore d'améliorer la performance des moteurs de recherche.

Problèmes

Matériel		
code	libellé	référence documentaire
MC70-W	assist. num. prof. MC70	HP1D-2009-01834-FR.doc



Métadonnées
date de création
01-01-2000

CARACTERISTIQUES EDA

Conception légère et robuste

Résiste aux tests de chutes et de chocs, avec une antenne intégrée et une protection IP-54.

Interopérabilité accrue

Grâce à Windows® Mobile 6.0, avec une sécurité évoluée, une plate-forme de développement souple et un système de messagerie mobile amélioré.

Rétro-compatibilité

Utilise des accessoires **MC70 EDA...**

La première difficulté liée aux documents textuels numériques apparaît à la lecture de ceux par des outils de traitement automatique du langage. Un format exotique, des caractères « pirates », une orthographe déficiente et les outils de traitement automatique ne peuvent remplir leurs fonctions correctement.

La migration d'un système documentaire, ou non, vers un autre soulève d'autres problèmes de natures différentes : formats de documents hétérogènes, doublons de documents sous des références différentes, incohérence des entités nommées (noms de personnes, d'organismes, matériels, locaux, dates, ...), consolidation des fiches d'identité, identification et nommage des documents ...

Ces mêmes types de problèmes sont prendre en considération pour définir des index pertinents permettant de retrouver la bonne information et pour affiner la restitution des résultats des moteurs de recherche.

Questions

- Quels critères de qualité ou indicateurs de confiance peut-on définir sur des documents ou des données non structurés dans les situations de traitement automatique du langage, de la migration, de la recherche d'information ?
- Comment établir un diagnostic de qualité sur des documents ou des données non structurés ?
- Comment améliorer la qualité des documents ou des données non structurés ?

Contact : projets@exqi.asso.fr

