

Migration et import de données

Avec Alfresco et l'ETL Talend

Marc Dutoo - Alfresco Meetup - 30 Avril 2009

Agenda

- Qui sommes-nous ?
- Problématiques de l'import de données
- La réponse ETL Connector & Talend
- Architecture
- Démonstration
- Questions

Qui sommes-nous ?

Open Wide – architecte Open Source

- ~ 90 employés sur Paris et Lyon, spin off de Thalès
- Portail, gestion documentaire, Business Intelligence...
- Offre de services pour les logiciels libres couvrant tout le cycle
 - Industrialisation de l'usage des composants Open Source, pour les grands comptes et les administrations

THALES



Marc Dutoo – architecte GED / SOA / BPM

- Conseil, formations, projets
- Responsable R&D (Eclipse JWT, OW2 Scarbo...)



Open Wide et Alfresco

Partenaire Gold depuis plus de 3 ans

- ☛ L'assurance de la qualité
- ☛ L'assurance de l'engagement de l'éditeur à nos côtés



De nombreux projets réalisés et en production



Une expertise de haut niveau de l'intégration et de la solution



Pour pouvoir gérer des documents, il faut qu'ils soient dans Alfresco !

- Une entreprise a toujours des documents antérieurs à l'introduction Alfresco
 - ▶ Ne serait-ce qu'un dossier partagé en réseau
- Et en sus des documents, souvent une organisation, une classification, voire des métadonnées métier
 - ▶ Fichiers d'index
 - ▶ Fichiers de métadonnées scannés (dématérialisation)
 - ▶ Métier : wiki, base de données, application...
- NB. On écarte la migration de version 😊

Cas d'usage 1 – Flexibilité et simplicité

« Peu » de documents et de méta-informations, classification antérieure non formalisée, besoin ponctuel

- Exemple : KM et Intranet de collectivités locales

Solution : Flexibilité et simplicité

- Action automatique avalant un fichier index
 - Les documents sont déposés dans Alfresco dans un espace technique
 - Puis un index global les référence et fournit classification et métadonnées
- Architecture
 - Action abstraite à étendre avec le code de classification métier
 - Autorise toute opération lors de la classification
- Adapté au partage de connaissances...

Cas d'usage 2 – Performance

Nombreux documents, import récurrent, classification formalisée

- Exemple : Miauton (CH, Bâtiment, 5000 factures / jour)

Solution : performance et évolutivité

- Import automatique des documents déposés dans un dossier sur le système de fichier, chacun avec son fichier de métadonnées
- Architecture
 - ▶ Polling threadé
 - ▶ Index : optimisation, amélioration du volume (avec Alfresco)
 - ▶ Evolutif : autres documents métier, scanner / OCR (Kofax, Readsoft...)
- Adapté à la dématérialisation !

Cas d'usage 3 – Versatilité

Beaucoup de documents, de métadonnées, plan de classement établi, récurrence...

- ▀ Flexibilité, mais pas de passer du temps sur du code jetable !
- ▀ Exemple : Habitat 76

Solution : versatilité

- ▀ Intégration d'outillage ETL (Extract, Transform, Load)
 - ▶ Principe de l'ETL : introspection des données (fichier, base...) et design de leurs flux d'intégration (« jobs »)
- ▀ Réalisation : ETLConnector pour Talend
 - ▶ En contribution sur la forge Alfresco (projet etlconnector)
 - ▶ Talend : ETL Open Source, leader mondial, basé sur Eclipse, français



ETL Connector – Présentation

Projet en collaboration

- ▶ entre Habitat 76 (besoin et fonctionnel), Open Wide (conception et développement), Talend (support)

L'utiliser

- ▶ Fourni avec Talend 3.1RC3
- ▶ Validé avec Alfresco 2.1 et 2.9b, erreurs mineurs en 3 Stable
- ▶ Tutorial : <http://nicolas.feray.free.fr/blog/?cat=5>
- ▶ Voir aussi forums francophones Alfresco et Talend
- ▶ Quelques bonnes pratiques...

Premier pas

- ▶ Installer ETL Connector dans Alfresco (3 jars dans WEB-INF/lib)
- ▶ Designer un job Talend avec tAlfrescoOutput en flux sortant
- ▶ Y configurer l'adresse web du serveur Alfresco, mapper au minimum la métadonnée cm:name, et exécuter !

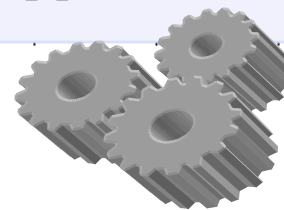
Démonstration : Import de quitus

- ❏ Fonctionnalités démontrées
 - ▶ Fichiers référencés et métadonnées décrites dans un fichier XML
 - ▶ Génération de l'arborescence de classification
 - ▶ Métadonnées métier (métamodèle spécifique)
 - ▶ Propriétés multivaluées
- ❏ Autres fonctionnalités
 - ▶ Modes de mise à jour
 - ▶ Positionnement des habilitations
 - ▶ Associations : parent-enfant et relations...
- ❏ Merci à Habitat 76 et Nicolas Feray pour le cas métier et le jeu de tests !

ETL Connector – Architecture - 1

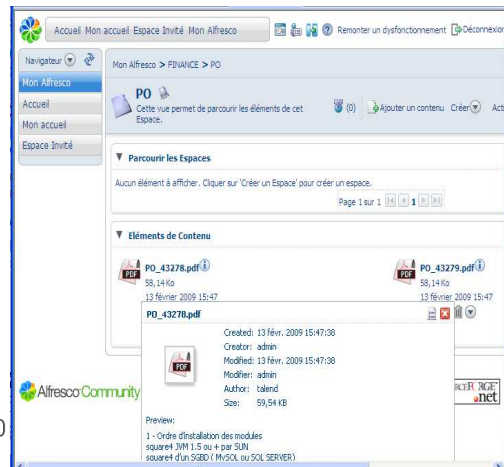
	A	B	C	D	E	F
1	file_name;destination_folder_name;source_path;author					
2	PO_43278.pdf;FINANCE\PO;file:D:/PO/customer1_2009-03-24.pdf;talend					
3	PO_43279.pdf;FINANCE\PO;file:D:/PO/customer2_2009-03-24.pdf;talend					

XLS, RDBMS...



Talend code generation

XML / REST



ETL Connector – Architecture - 2

- Reprend l'architecture d'import ACP d'Alfresco et ajoute
 - ▶ transactions sur l'import de chaque nœud,
 - ▶ trace des erreurs,
 - ▶ Identification par nom (et nom xpath),
 - ▶ modes de mise à jour...

- Interactions ETL – Alfresco en XML / REST
 - ▶ XML : ACP (Alfresco Content Package) !
 - ▶ Documents : référencés par des URLs (pointant sur le système de fichier local, ou monté en réseau, ou HTTP...)
 - ▶ Architecture : basée sur les Command d'Alfresco

- « Bon citoyen » dans l'architecture globale
 - ▶ reprise par Alfresco envisageable

▀ L'import de documents :

- ▶ De la reprise de documents à l'ETL
- ▶ Une solution versatile : ETL Connector pour Alfresco et Talend
- ▶ Plus d'informations : <http://forge.alfresco.com/projects/etlconnector>

▀ Au-delà...

- ▶ Open Wide pour vos intégrations Alfresco !
- ▶ Autres réalisations : interfaces web riches, collaboratif avancé, processus métier, ERP (SAP), intégration de portail, d'applications métier...
- ▶ Sessions de démonstrations

Questions – réponses...

**Merci de
votre
attention !**

Contact

marc.dutoo@openwide.fr

www.openwide.fr

