Imports et modifications en masse dans HAL

Exemple des thèses de l'UL en partenariat avec l'INIST

Jean-François Lutz – Université de Lorraine
Alain Zasadzinski – INIST-CNRS
2es journées CasuHAL – Dijon – 01/06/2018

Pourquoi transférer des thèses vers HAL ?

- Dépôt numérique des thèses de doctorat adopté tôt dans les universités lorraines (2006 -2007).

  Extension aux thèses d'exercice en 2009-2010.

- Choix d'ORI-OAI en 2009 pour le signalement et la diffusion.

- Mise en production à l'automne 2010 (Pétale),


Version accessible
UNIVERSITÉ DE LORRAINE
PETALE
Publications Et Travaux Académiques de LorrainE
Accueil
Thèses de doctorat (9694)
Thèses d'exercice (3366)
Mémoires (1679)
Recherche sur tous les documents
Bienvenue sur PETALE
Pétale est la plateforme de diffusion en ligne des thèses et mémoires numériques de l'Université de Lorraine.
[En savoir plus]
Rechercher
Aide | Recherche avancée

Pourquoi transférer des thèses vers HAL ?

- En parallèle, politique de numérisation rétrospective des thèses au format papier :
  - 1990-2007 pour les thèses de doctorat
  - 2000-2009 pour les thèses d'exercice

- Signalement dans ORI-OAI également.

- Gestion des accès intranet / internet.

Pourquoi transférer des thèses vers HAL ?

- 2014 : le dossier « archive ouverte » est relancé.
- Analyse fonctionnelle et tests de 3 solutions locales... HAL rattrappé *in extremis* avec la v3.
- Choix de HAL au printemps 2015

Maintenir deux outils en parallèle ?

Préparer le basculement vers HAL

- Volumétrie relativement importante.
- Formats distincts mais mapping ABES TEF -> TEI.
- Deux problématiques :
  - l'import des thèses présentes dans ORI-OAI
  - le traitement des thèses courantes.

Préparer le basculement vers HAL

- Recours au service informatique interne pour l'import des thèses de doctorat et à l'INIST pour celui des thèses d'exercice

- Mise en place d'imports semestriels Sudoc -> HAL par l'INIST pour les thèses d'exercice courantes.

Work Flow dépôt initial

1. Mapping TEF vers TEI-HAL
2. Transformation via scripts shell + perl + librairie XML interne (INIST)
3. Insertion des éléments spécifiques au portail
   - Type de document (MEM pour les thèses d'exercice)
   - idStruct (AuréHAL) de l'établissement
   - Type d'accès (internet ou intranet)
   - Domaines (thèses d'exercice médecine, pharmacie, dentaire)
4. Parsing XML (xmllint)
5. Création des zip pour dépôt SWORD (TEI-XML avec ou sans pdf selon type d'accès)
6. Exécution du script d'import SWORD

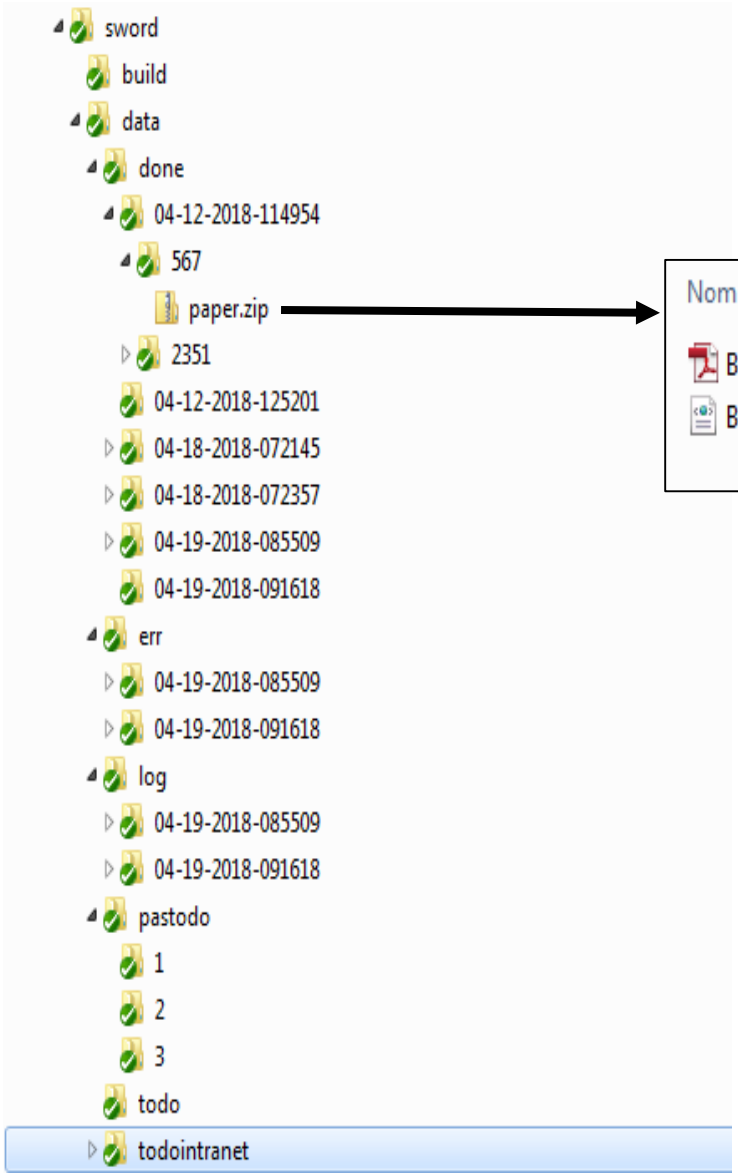
Work Flow correction en masse de métadonnées

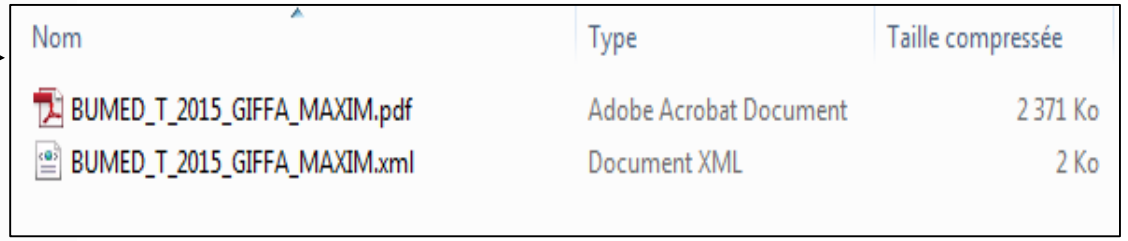
1. Correction des métadonnées dans la TEI-HAL avec suppression du champ « editionStmt »
2. Parsing XML (xmllint)
3. Récupération via les API de HAL des métadonnées correspondant aux références à corriger
4. Mapping entre les références initiales et celles récupérées dans HAL afin de récupérer l'identifiant HAL (docid) correspondant :
   - alignement titre+auteur (avec appauvrissement casse, suppression diacritiques et ponctuation)
5. Création des répertoires paper.zip pour dépôt SWORD (avec la TEI corrigée et un fichier txt contenant le docid de la notice)
6. Exécution du script de correction SWORD

Scripts de dépôt et de correction SWORD

```
#!/bin/bash

export PATH=$PATH:/usr/local/bin/

#-- smtp mail

SMTP="smtp://smtpout.intra.inist.fr:25"

#-- le user et mot de passe de connexion

user_pwd="*******:******"

#-- le mail pour reception cptr
#~ mail=florence.join@inist.fr
#~ mail=alain.zasadzinski@inist.fr
mail=colette.orange@inist.fr
#-- pour generer un dépot avec pdf  paper.zip et l'envoyer mettre zip="true"
#-- pour generer une correction de dépot :
        #~ zip="false"
        #~ maj="true"
        #-- l'url du dépôt
        #~ url=https://api-preprod.archives-ouvertes.fr/sword/hal/
url=https://api.archives-ouvertes.fr/sword/univ-lorraine/
        #~ url=https://api.archives-ouvertes.fr/sword/hal/

"# -- commande qui traite les données sous $HOME/sword/data/todo/
docker-compose run --rm --user=$UID sword -u $user_pwd -m $mail -p $url -l $zip -v $maj -j $SMTP
```

Organisation des répertoires de dépôt





Exemples de fichiers TEF et TEI-XML

Fichier TEF avec pdf
Fichier TEI-XML avec pdf

Fichier TEF sans pdf (accès intranet)
Fichier TEI-XML sans pdf

Notice WoS en TEI-XML pour dépôt avec PDF

Fichier TEI_XML pour correction HAL
Fichier associé avec idHAL

Cas de dépôt en masse depuis le WoS : notices + PDF (1)

Principale difficulté :

Résolution de <u>toutes</u> les affiliations françaises par un id_struct

Déclaration des affiliations étrangères dans l'élément « back » de la TEI-HAL

<u>Création d'un dictionnaire de « formes de noms de labos » vers auréHAL-structure</u>

- Utilisation de auréHal-structure, Labintel et RNSR pour création d'une table combinant un maximum de forme de noms
- Mapping entre RNSR, Labintel et auréHal-structure

Récupération des id_struct via API auréHAL (valid / old / incoming)

<u>Résolution des affiliations :</u>

- Appauvrissement des affiliations : casse, ponctuation et diacritiques
- Table à 3 ou 4 colonnes de toutes les formes de noms de labos possibles
- Double/triple alignements de chaque affiliation avec la table (script perl) :
  - forme du nom du labo / ville(s) ou tutelle(s)
  - type de labo / numéro /ville(s) ou tutelle(s)
  - Génération de l'id_struct (ou du RNSR ou Labintel …) et des chaînes de caractères reconnues dans la table pour validation/vérification
  - Détection des ambiguités (ex : id-struct multiples pour même affiliation)
  - gestion par dates de publication en relation avec évolution du labo
  - validation manuelle des ambiguités sous excel par documentaliste

Cas de dépôt en masse depuis le WoS : notices + PDF (2)

A fournir :

Cle ut avec pour chacune :

- pdf
- id_struct des auteurs des labos concernés pour chaque affiliation concernée
- type de documents
- domaines scientifiques HAL

Pour le dépôt en masse via sword :

- Téléchargement du corpus à partir du WoS et transformation en TEI-HAL
- Ajout des éléments ci-dessus dans la TEI-HAL

Attention doublons potentiels :

- Hal détecte les doublons sur la base du DOI
- Si besoin d'en faire plus, nous contacter (ex : dédoublonnage de documents sans DOI avant dépôt en masse dans HAL)