L’import XML-TEI dans HAL

Service IES – Pôle Archives Ouvertes Inria

Mai 2016

Sommaire

1. Rappel des fondamentaux
   1.1. Qu'est-ce que la TEI ?
   1.2. Quelques éléments pour démarrer

2. Consolidation et validation du fichier XML
   2.1. Les pré-requis
   2.2. Les données génériques
   2.3. Les données par type de document
   2.4. Valider son fichier TEI-HAL

3. La phase d'import
   3.1. Import par l'interface web
   3.2. Exemple de message d'erreur
   3.3. Import par requête URL (cURL)
   3.4. Création d'une nouvelle version

4. Postproduction des dépôts : et ensuite ?
   4.1. Les dernières étapes

1

Rappel des fondamentaux

1.1. Qu’est-ce que la TEI ?

- Text Encoding Initiative : Basé sur XML, langage de marquage qui permet d’échanger des données textuelles informatisées.
- LA TEI est un format pivot standard choisi et utilisé pour l’import/export dans HAL
- La TEI est adaptée pour HAL (métadonnées spécifiques à HAL comme l’audience)

1.2. Quelques éléments pour démarrer

- Pour tout import dans HAL, il est nécessaire de posséder un compte HAL (login + mdp)

- L'import Sword concerne les documents éligibles dans HAL uniquement !

- L'import Sword permet de déposer dans un portail de HAL ou dans HAL générique un seul document décrit dans un fichier xml au format TEI/HAL avec éventuellement des fichiers associés

- Quelques portails ont des types de document particulier. Pour Hal-Inria, il est possible de déposer des mémoires d'étudiant (HAL-inria+)

- Les fichiers XML-TEI déposés doivent être conformes au schéma de HAL (disponible ici : https://api.archives-ouvertes.fr/documents/aofr-sword.xsd)

2

Consolidation et validation du fichier XML

2.1. Pré-requis

- Travailler à partir d'un fichier .xml déjà existant (à partir d'un export Zotero ou autre)
- Une seul document par fichier

Pour vous aider à rendre vos fichiers .xml valides pour l'import Sword dans HAL au format XML-TEI, l'ensemble des métadonnées possibles sont décrites à l'adresse suivante :

- http://api.archives-ouvertes.fr/docs/sword
(documentation sur l'import Sword avec quelques exemples par types de document au format XML)

- http://api.archives-ouvertes.fr/documents/all.xml
(fichier xml annoté qui permet, tous types de documents confondus, de faire la corespondances entre les métadonnées de HAL et les informations à renseigner dans le XML-TEI)

- https://github.com/CCSDForge/HAL/tree/master/Sword
(documentation complète en anglais de l'import Sword dans HAL ainsi que un exemple de xml-TEI pour chaque type de document)

2.2. Les données génériques

Structure globale d'un fichier TEI-HAL (il est nécessaire de respecter l'ordre des blocs) :

```
<TEI xmlns=http://www.tei-c.org/ns/1.0 //Partie récapitulative des url de référence pour l'import Sword
   xmlns:hal="http://hal.archives-ouvertes.fr/"
   xsi:schemaLocation="http://www.tei-c.org/ns/1.0 https://api.archives-ouvertes.fr/documents/aofr-sword.xsd">
<text> //Partie principale
    <body>
      <listBibl>
        <biblFull> //Référence bibliographique structurée
          <titleStmt> //Mention des financements
          <editionStmt> // Zone des fichier(s)
          <publicationStmt> //Licence Creative Commons choisie
          <seriesStmt> //Collections
          <notesStmt> //Mention des notes
          <sourceDesc> //Description bibliographique de la ressource (titres, auteurs et affiliations)
          <profileDesc> //Complément descriptif de la ressource (mots-clés, résumé, domaine, type de document)
        </biblFull>
      </listBibl>
    </body>
    <back> //Description des nouvelles entrées dans AuréHAL (auteurs, laboratoires, projets)
        <listOrg type="laboratories"> //Définition de nouvelles structures (équipes, laboratoire, institution)
        <listOrg type="projects"> //Définition de nouveaux projets (ANR, européens)
   </back>
  </text>
</TEI>
```

2.2 Données concernant le fichier

NB : Pour l'import d'une notice seule, ne pas intégrer cette partie !

- *text/body/listBibl/biblFull/editionStmt*

```
<editionStmt>
   <edition>
      <date type="whenWritten">2016</date>
      <ref type="file" subtype="author" n="1" target="ftp://ftp.ccsd.cnrs.fr/test.pdf">
         <date notBefore="2017-01-01"/>
      </ref>
   </edition>
</editionStmt>
```

- <ref type= http://api-preprod.archives-ouvertes.fr/ref/metadataList/?q=metaName_s:fileType&fl=*&wt=xml (file, fichier source ou données supplémentaires/annexes)
- <ref subtype= http://api-preprod.archives-ouvertes.fr/ref/metadataList/?q=metaName_s:fileSource&fl=*&wt=xml (fichier auteur, fichier avec autorisation de l'éditeur, fichier financé par l'éditeur, …)
- <ref n= (1 pour le fichier principal, 0 sinon)

- 2 possibilités :
  - Importer une notice avec un fichier et le lien vers l'URL (serveur ftp HAL notamment https://hal.inria.fr/user/ftp)
  - Importer un .zip avec le fichier .xml et le fichier .pdf

2.2. Renseigner les auteurs

text/body/listBibl/biblFull/sourceDesc/biblStruct/analytic/author

- Renseigner un auteur existant dans AuréHAL (mentionner l'IDHAL ou le HALAuthorID et les affiliations si nécessaire)

```
<author role="aut">
  <persName>
    <forename type="first">Laurent</forename>
    <surname>Romary</surname>
  </persName>
  <idno type="idhal">laurentromary</idno> OU <idno type="halAuthorId">49567</idno>
  <affiliation ref="#struct-95237"/>
</author>
```

- Créer un nouvel auteur

```
<author role="aut">
  <persName>
    <forename type="first">Prénom</forename>
    <surname>Nom</surname>
  </persName>
  <email>prenom.nom@structure.fr</email>
  <affiliation ref="#localStruct-1"/> OU <affiliation ref="#struct-95237"/>
</author>
```

2.2. Renseigner les structures pour AuréHAL

- Créer une nouvelle structure n'existant pas dans HAL (partie <back>) : *Text/back*

```
<listOrg type="laboratories">
        <org type="laboratory" xml:id="localStruct-1">
          <orgName>Mon nouveau laboratoire</orgName>
          <orgName type="acronym">MNL</orgName>
          <desc>
            <address>
      <addrLine>adresse du laboratoire</addrLine>
      <country key="FR"/>
      </address>
            <ref type="url">http://ccsd.cnrs.fr</ref>
          </desc>
          <listRelation>
            <relation active="#localStruct-2"/>
            <relation active="#struct-300009"/>
          </listRelation>
        </org>
```

```
<org type="institution" xml:id="localStruct-2">
          <orgName>Institut Test</orgName>
          <desc>
            <address>
      <country key="FR"/>
      </address>
          </desc>
        </org>
```

- Lors de l'ajout d'un laboratoire ou d'une équipe de recherche, il faut obligatoirement renseigner une structure de rattachement de type institution (la tutelle d'une structure laboratoire ne peut être un autre laboratoire)

- Code de la langue sur deux lettres ISO 639.1

2.2. Renseigner les mentions de titre

text/body/listBibl/biblFull/sourceDesc/biblStruct/analytic/title

- Obligation de mentionner au moins un titre dans au moins une langue (au choix)
- Le sous-titre est optionnel

```
<title xml:lang="en">this is my conference paper title</title>
 <title xml:lang="fr">ceci est mon titre de communication dans un congrès</title>
 <title xml:lang="en" type="sub">my subtitle in english</title>
 <title xml:lang="fr" type="sub">mon sous-titre en français</title>
```

2.2. Renseigner les données de description complémentaire de la ressource

text/body/listBibl/biblFull/sourceDesc/profileDesc

- Zone dédiée aux métadonnées relatives à la langue du document, mots-clés, domaines et classifications, type de document (COMM, ART, COUV…), résumé et collaborations.

```
<profileDesc>
   <langUsage>
      <language ident="en"/>
   </langUsage>
   <textClass>
      <keywords scheme="author">
         <term xml:lang="en">keywords_2_en</term>
         <term xml:lang="en">keywords_en</term>
         <term xml:lang="fr">keywords_fr</term>
         <term xml:lang="fr">keywords_2_fr</term>
      </keywords>
      <classCode scheme="classification">PACS</classCode>
      <classCode scheme="halDomain" n="info"></classCode>
      <classCode scheme="halDomain" n="info.eiah"></classCode>
      <classCode scheme="halTypology" n="COMM"></classCode>
   </textClass>
   <abstract xml:lang="en">resume_en</abstract>
   <abstract xml:lang="fr">resume_fr</abstract>
   <particDesc>
      <org type="consortium">Collaboration/Projet</org>
      <org type="consortium">Collaboration/Projet 2</org>
   </particDesc>
</profileDesc>
```

Liste domaines de HAL : http://api.archives-ouvertes.fr/ref/domain/?q=text:*&wt=xml&fl&rows=1000

2.3. Les données par type de documents

- Certaines métadonnées sont obligatoires selon les documents (ex. : directeur de thèse pour le document « thèse »).

Liste des types de documents (https://api.archives-ouvertes.fr/ref/doctype?q*&wt=xml&instance_s=inria ou instance_s=all pour tous les portails)

- ART = Article dans une revue
- COMM = Communication dans un congrès
- POSTER = Poster
- OUV = Ouvrage
- COUV = Chapitre d'ouvrage
- DIROUV = Direction d'ouvrage
- PATENT = Brevet
- OTHER = Autre type de publication
- PREPRINT = Pré-publication, document de travail
- REPORT = Rapport
- THESE = Thèse
- HDR = HDR

- MEM = Mémoire d'étudiant (HAL-Inria+ uniquement)

Se référer à la documentation spécifique et aux fichiers d'exemples associés sous la page Github du CCSD dédiée à l'import Sword :

https://github.com/CCSDForge/HAL/tree/master/Sword

2.3. Les données par type de documents

text/body/listBibl/biblFull/notesStmt

La zone de notes contient des métadonnées selon les types de documents (exemple : type proceedings que pour le type COMM) :

```
<notesStmt>
   <note type="commentary">Commentaire</note>
   <note type="description">description</note>
   <note type="audience" n="2"/>
   <note type="invited" n="0"/>
   <note type="popular" n="0"/>
   <note type="peer" n="1"/>
   <note type="proceedings" n="1"/>
</notesStmt>
```

Signification des valeurs n=""
note type="audience" n="" http://api.archives-ouvertes.fr/ref/metadataList/?q=metaName_s:audience&fl=*&wt=xml
note type="invited" n="" http://api.archives-ouvertes.fr/ref/metadataList/?q=metaName_s:invitedCommunication&fl=*&wt=xml
note type="popular" n="" http://api.archives-ouvertes.fr/ref/metadataList/?q=metaName_s:popularLevel&fl=*&wt=xml
note type="peer" n="" http://api.archives-ouvertes.fr/ref/metadataList/?q=metaName_s:peerReviewing&fl=*&wt=xml
note type="proceedings" n="" http://api.archives-ouvertes.fr/ref/metadataList/?q=metaName_s:proceedings&fl=*&wt=xml

2.3. Les données par type de documents

text/body/listBibl/biblFull/biblStruct/monogr

Exemple de champs pour le type de document communication sur la partie <monogr> :

```
<monogr>
  <idno type="isbn">978-1-4577-0538-0</idno>
  <idno type="localRef">Référence interne</idno>
  <idno type="localRef">Référence interne 2</idno>
  <title level="m" type="main">titre ouvrage</title>
  <meeting>
     <title>Mon titre de conférence</title>
     <date type="start">2016-01-22</date>
     <date type="end">2016-01-27</date>
     <settlement>Prague</settlement>
     <country key="CZ"/>
  </meeting>
 </monogr
  <imprint>
     <publisher>Éditeur commercial</publisher>
     <pubPlace>Lieu de publication</pubPlace>
     <biblScope unit="serie">Titre du volume</biblScope>
     <biblScope unit="volume">Volume</biblScope>
     <biblScope unit="issue">1</biblScope>
     <biblScope unit="pp">2</biblScope>
     <date type="datePub">2016-02-03</date>
  </imprint>
```

2.4. Valider son fichier TEI-HAL

- Valider son fichier avec le schéma XML de HAL via un éditeur XML (exemple Oxygen, XML Spy, etc.)

- Le schéma XML de validation est disponible à l'adresse suivante : https://api.archives-ouvertes.fr/documents/aofr-sword.xsd

- A noter : le schéma est plus « light » que HAL, aussi un fichier validé par le schéma peut ne pas passer lors de l'import (par exemple métadonnée obligatoire oubliée pour un type de document particulier)

  Exemple : le résumé français/anglais des thèses => un seul résumé est validé par le schéma mais ne passe pas à l'import car il est obligatoire de renseigner dans deux langues (dont l'anglais).

- L'en-tête du fichier XML-TEI

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:hal="http://hal.archives-ouvertes.fr/"
  xsi:schemaLocation="http://www.tei-c.org/ns/1.0 https://api.archives-ouvertes.fr/documents/aofr-sword.xsd">
```

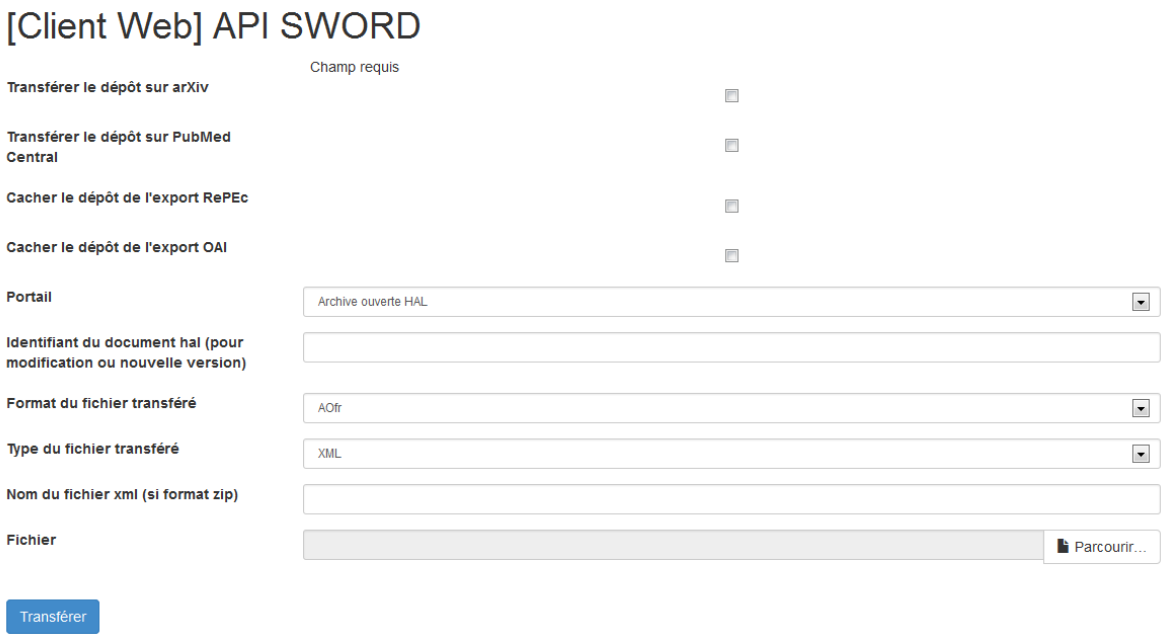
3

La phase d’import

3.1. Import par l’interface web

L’interface de dépôt par formulaire web :
https://api.archives-ouvertes.fr/sword/upload/
https://api-preprod.archives-ouvertes.fr/sword/upload/ (bac à sable pour les tests)

Le login et mot de passe HAL du compte sont demandés pour l’accès à l’url.



Dépôt avec fichier, il est indispensable que le fichier de métadonnées, meta.xml pointe vers le document grâce à la mention : target="DOC.pdf". L'archive .zip doit donc contenir meta.xml et DOC.pdf.

3.1. Import par l’interface web

Dépôt archivé dans HAL : **hal-01035764**

Retour SWORD


```
        <?xml version="1.0" encoding="utf-8"?>
<entry xmlns="http://www.w3.org/2005/Atom" xmlns:sword="http://purl.org/net/sword/terms/" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:hal="http://hal.archi
ves-ouvertes.fr/">
  <title>Accepted media deposit to HAL</title>
  <id>hal-01035764</id>
  <hal:password>5&amp;tc6uu</hal:password>
  <hal:version>1</hal:version>
  <updated>2016-05-25T11:38:43+02:00</updated>
  <summary>A media deposit was stored in the HAL workspace</summary>
  <sword:treatment>stored in HAL workspace</sword:treatment>
  <sword:userAgent>HAL SWORD API Server</sword:userAgent>
  <source>
    <generator uri="https://api-preprod.archives-ouvertes.fr/sword" version="1.0">hal@ccsd.cnrs.fr</generator>
  </source>
  <link rel="alternate" href="https://halv3-preprod.archives-ouvertes.fr/hal-01035764"/>
</entry>
```


Si le dépôt est réussi, l’identifiant hal est mentionné en bas de la page ainsi que le mot de passe, sinon un message d’erreur est retourné.

Ce qui est valable pour HAL en général l’est pour l’import Sword également (exemple : transfert vers arXiv et métadonnées obligatoires)

3.2. Exemple de message d’erreur

Exemples de retour :

Erreur de dépôt :

```
string(105) "{"meta":{"identifier":{"regexNotMatch":"'My DOI number' n'est pas un DOI valide, par exemple : 10.xxx"}}}"
```

Retour SWORD


```
        <?xml version="1.0" encoding="utf-8"?>
<sword:error xmlns:sword="http://purl.org/net/sword/error/" xmlns="http://www.w3.org/2005/Atom" href="http://purl.org/net/sword/error/ErrorBadRequest">
  <title>ERROR</title>
  <updated>2016-05-10T11:12:59+02:00</updated>
  <author>
    <name>HAL SWORD API Server</name>
  </author>
  <source>
    <generator uri="https://api-preprod.archives-ouvertes.fr/sword" version="1.0">hal@ccsd.cnrs.fr</generator>
  </source>
  <summary>Some parameters sent with the request were not understood</summary>
  <sword:treatment>processing failed</sword:treatment>
  <sword:verboseDescription>{"meta":{"identifier":{"regexNotMatch":"'My DOI number' n'est pas un DOI valide, par exemple : 10.xxx"}}}</sword:verboseDescription
>
  <link rel="alternate" href="https://api-preprod.archives-ouvertes.fr" type="text/html"/>
</sword:error>
```

3.3. Import par requête URL (cURL)

- Dépôt d'un .zip

Exemple :
curl -v -u test_ws:test https://api-preprod.archives-ouvertes.fr/sword/hal/ -H "Packaging:http://purl.org/net/sword-types/AOfr" -X POST -H "Content-Type:application/zip" -H "Content-Disposition: attachment; filename=comm.xml" --data-binary @depot.zip

la réponse en cas de succès est "201 Created" avec identifiant HAL et mot de passe.

- Dépôt d'une notice

Exemple :
curl -X POST -d @meta.xml -v -u test_ws:test https://api-preprod.archives-ouvertes.fr/sword/hal/ -H "Packaging:http://purl.org/net/sword-types/AOfr" -H "Content-Type:text/xml"

La réponse en cas de succès est "202 Accepted" avec identifiant HAL et mot de passe.

- A noter qu'en utilisant l'en-tête HTTP On-Behalf-Of, il est possible de donner la liste des propriétaires du dépôt

3.4. Création d’une nouvelle version

Pour créer une nouvelle version d’un dépôt, il est possible de le faire par une requête curl ou via le formulaire web:

- Via le formulaire web il suffit de renseigner l’identifiant du dépôt dans le champ « Identifiant du document hal (pour modification ou nouvelle version) »
- Autres actions possibles que le dépôt via curl : DELETE (dans le cas d’une notice à supprimer) et PUT pour une modification (soit création d’une nouvelle version ou modification des métadonnées).

Exemple nouvelle version : requête HTTP **PUT** sur l'URL https://api.archives-ouvertes.fr/sword/*identifiant ressourcevversion* en incluant dans le contenu HTTP le XML dans le format annoncé dans l'entête **Packaging**. Il est possible de s'assurer de l'intégrité du contenu envoyé en fournissant sa signature **md5** dans l'entête HTTP **Content-MD5**.

En cas de succès, la réponse du serveur est "200 OK" et un contenu décrivant une entrée SWORD.

4

Postproduction des dépôts : et ensuite ?

4.1. Les dernières étapes

- Les notices sont en ligne immédiatement après dépôt (temps d'indexation Solr), ceux avec fichier ont une phase de modération (quelques jours)
- Les entrées dans la partie <back> du document (création de nouvelles entrées auréHAL), sont en état « à valider » dans AuréHAL, qu'il convient de faire valider par un gestionnaire HAL. Pour plus de simplicité, il est préférable de nous contacter en amont des dépôts pour créer et valider ces entrées.

Merci.

Avez-vous des questions ?

archive-ouverte@inria.fr