Numérisation et archivage pérenne à la bibliothèque Cujas: un retour d'expérience

Noëlle Balley,
Chef du département des Monographies
Journées du CTLes, 20 juin 2013

Rappels sur le programme de numérisation

- Un projet mené (presque) entièrement en interne
- Un choix technique initial qui a conditionné tout le projet
- Un choix documentaire qui s’appuie sur le concours d’enseignants chercheurs
- 436 titres, 1 500 volumes, 893 000 pages

L’organisation du travail

- Numérisation interne:
  - Sur le scanner de la bibliothèque pour les ouvrages postérieurs à 1830
  - Océrisation et relectures partielles sur les parties structurantes
- Numérisation externalisée:
  - Pour les ouvrages antérieurs à 1830
  - Mode image, saisie fine des tables des matières

Les tâches

1. Saisie des tables des matières et bordereau d’état
2. Préparation matérielle des documents
3. Catalogage Sudoc
4. Indexation (facettes)
5. Numérisation (interne ou externe) -> TIFF
6. Contrôle qualité, réfections d’images
7. Génération des JPeg

Les tâches (2)

8. Océrisation (interne) et relectures
9. Conversion en PDF
10. Transformation des TDM en signets
11. Mise en ligne
12. Archivage

Un volume = 20 heures de travail au moins

- **Numérisation**
  - ✓ Numérisation — 1 heure 30 (au moins)
  - ✓ Traitement des images — 30 minutes
  - ✓ Contrôle qualité et réfection des images — 4 heures
  - ✓ Océrisation — 1 heure
- **Traitement documentaire**
  - ✓ Saisie des tables des matières — 2 heures 30 (moyenne)
  - ✓ Relectures et corrections OCR intégrales — 20 heures au moins

  OU
  - ✓ Relectures et corrections ciblées OCR — 5 heures au moins
  - ✓ Notice Dublin Core — 30 mn
  - ✓ Indexation — 1 heure
  - ✓ Contrôle final et validation — 2 heures
- **Traitement informatique**
  - ✓ Mise en ligne — 30 mn
  - ✓ Archivage — 30 mn + 1 heure au moins

Les acteurs

- Côté monographies :
  - 1 Chef de projet
  - 1 Bib Ass (contrôle qualité, tables des matières, Dublin Core)
  - 1 vacataire (numérisation, contrôle qualité, TDM, OCR et relectures)
  - Vacataires étudiants en renfort ponctuel (masters histoire du droit)

Les acteurs (2)

- Côté Traitement documentaire :
  - 1 Bib Ass (Catalogage), 1 conservateur (Indexation)
- Côté Informatique :
  - 1 ingénieur de recherche (développements, architecture informatique, sécurité…)
  - 1 PRCE (Métadonnées, archivage pérenne)
  - 1 technicien informatique (maintenance scanner, logiciels, serveurs)

L’archivage pérenne

L’archivage pérenne

- Inscrit dans les objectifs comme une évidence dès le début du projet
- … mais reste une notion vague jusqu’en 2010
- Les priorités jusqu’en 2011 : construire le projet et **mettre en ligne** en **respectant les échéances** fixées par la direction…
- D’où une certaine souplesse sur les petites erreurs qui ne bloquent pas la mise en ligne.
- **L’archivage oblige à corriger toutes les petites imperfections tolérées jusque là.**

« Pourquoi archiver? »

Puisque les documents existent sur papier…

Ce n’est pas l’information en elle-même qu’on archive, mais **le fait qu’elle soit numérisée** et **la valeur ajoutée par le travail intellectuel de l’équipe**.

.

« Archiver quoi? »

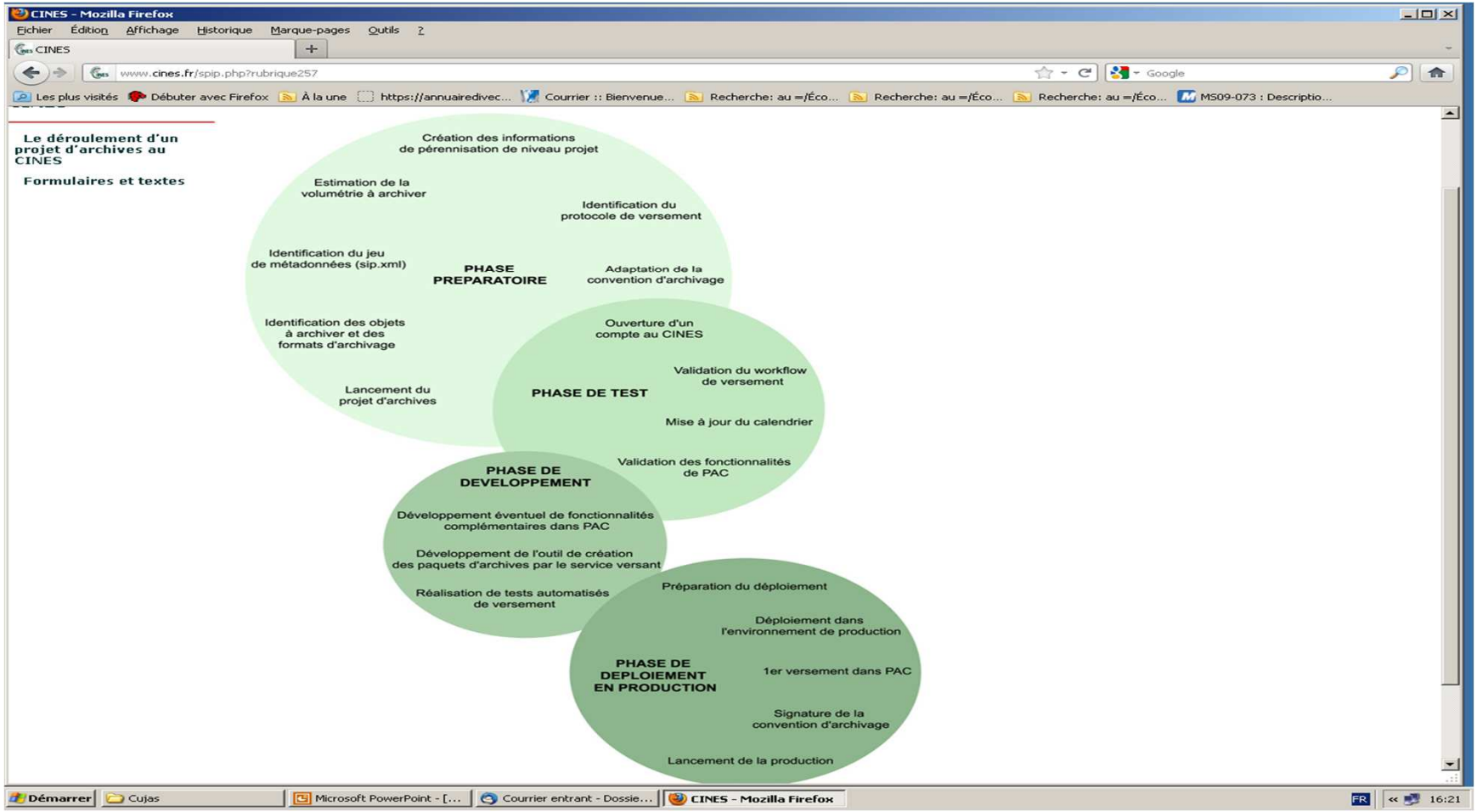
On pérennise la **possibilité de restructurer** les documents, et non la structuration elle-même (fichiers de diffusion)

On archive **les briques et le plan, pas la maison.**

La gestion du projet

- Contacts informels avec le CINES dès 2008
- 2009: évaluation des besoins dans le cadre du contrat quadriennal 2010-2013 : 20.55 TO à archiver sur 5 ans, demande budgétaire de 102 750 euros
- Réunion de lancement le 1er octobre 2009
- Signature de la convention le 10 mai 2010
- Réunions téléphoniques et échanges par mail avec le CINES
- Comptes rendus réguliers en interne par mail, par écrit et par oral
- Entrée en production le 22 février 2011 à 10 h 18
- Réunions annuelles de bilan

Les étapes du projet d'après le site du CINES



Les étapes côté Cujas

- 1. Convention avec le CINES signée le 10 mai 2010
- 2. Définition des « paquets d’archives » et des informations à préserver
- 3. Création du ppdi: 6 mois, 8 versions
- 4. Six mois de tests, développements et automatisation des « routines »
- 5. En production depuis février 2011.

1. La convention et ses annexes

- Annexe 1 : Conditions générales de service du CINES.
- Annexe 2 : Politique d'archivage du CINES.
- Annexe 3 : Liste des formats de données archivables acceptés par le service d'archives.
- Annexe 4 : Spécifications fonctionnelles et techniques de la plateforme d'archivage pérenne du CINES, incluant la liste des normes et standards mis en œuvre.
- Annexe 5 : Composition du comité de liaison.
- Annexe 6 : Liste des informations essentielles à préserver dans les archives et composition du comité d'experts, référent dans le cadre d'une migration de format.
- Annexe 7 : Identification de la communauté-cible et de sa base de connaissance.
- Annexe 8 : Volumétrie maximale et prix.
- Annexe 9 : Politique de communicabilité.

Evaluation de la volumétrie à archiver (convention annexe 8)

Année	Docs créés	Cumul docs créés	Estimation volume	Cumul estimation
2011	170	170	1 TO	1 TO
2012	150	320	3 TO	4 TO
2013	150	470	4.5 TO	8,5 TO

2. Identification des objets numériques à archiver

Liste des formats acceptés sur le site du CINES

L'unité d'archivage = le livre physique

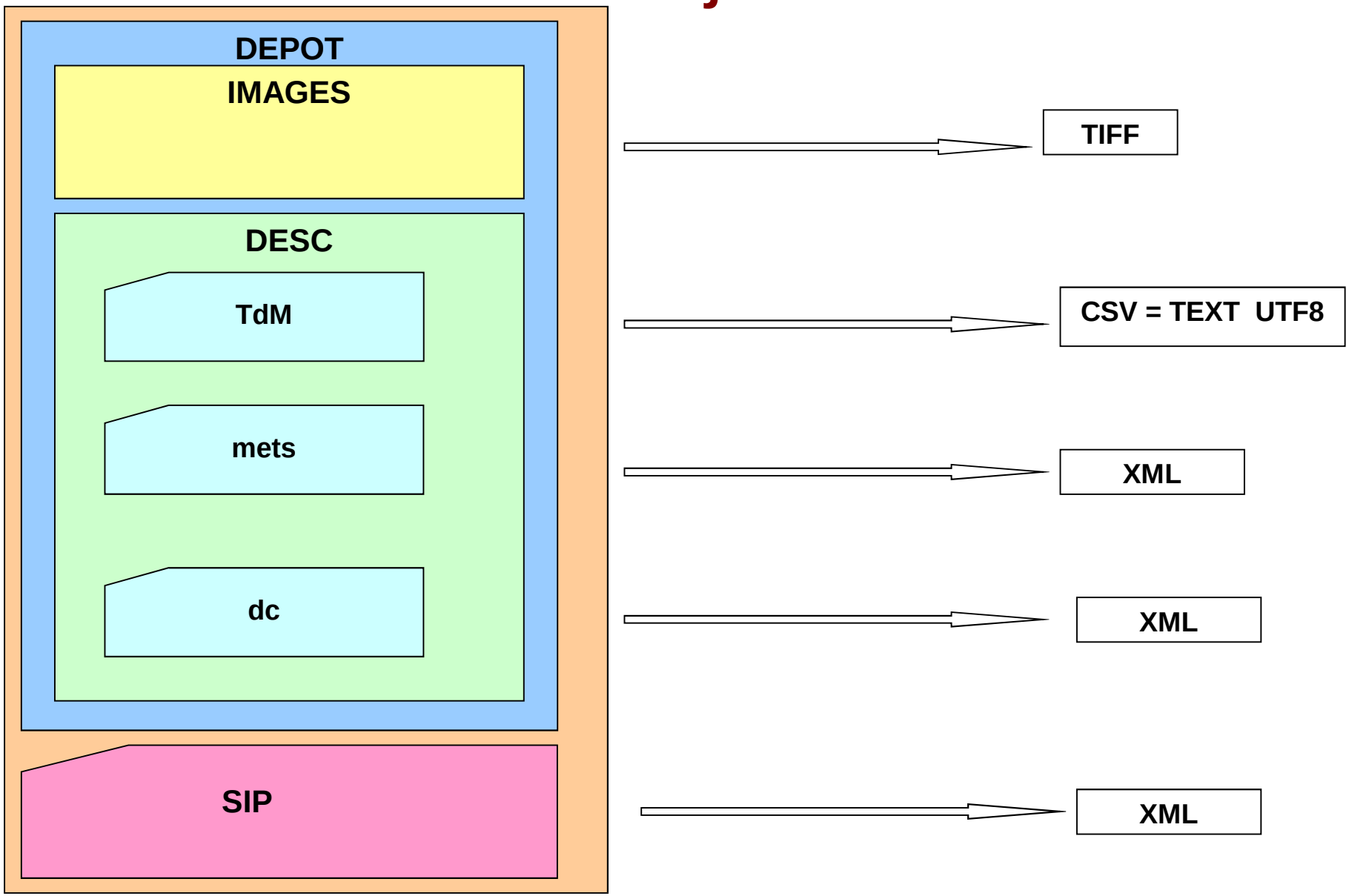
Le dialogue permanent avec le CINES nous a aidés à revoir les ambitions à la baisse…

Avantages d'un choix minimaliste: faisabilité, coûts

Choix de formats standardisés, non propriétaires (csv, plutôt qu'Excel…)

Importance de l'UTF8

Identification des objets à archiver



3. Création du PPDI

- Project Preservation Description Information : explications sur le site du CINES
- Document xml qui décrit le projet de numérisation dans tous ses aspects
- Une culture archivistique, déroutante pour les bibliothécaires
- Occasion de relire rétrospectivement son projet
- Chaque modification du projet implique une mise à jour du PPDI

La structure du PPDI

1. Nature des fonds
2. Service versant
3. Eléments techniques (circuits de production)
4. Contenu et objectifs de l’archivage
5. Droits et accès
6. Structure de l’unité d’archivage
7. Règles de saisie de la notice Dublin Core

4. Les tests

- Création de jeux de test
- Obligation de « nettoyer » les documents déjà en ligne:
  - Vérification de la structure et du contenu des dossiers
  - Renommage des fichiers et des dossiers
  - Révision des tables des matières
- **Les imperfections tolérées au début du travail doivent impérativement être corrigées, sous peine de rejet par le serveur du CINES**

Création des fichiers SIP

- SIP = Schéma xml, défini par le CINES et lié à ses procédures d’archivage
- Description fine des règles de contenu et de saisie
- Eléments descriptifs, techniques, liste des fichiers et « empreinte » de chaque fichier
- Un non respect des consignes de saisie entraîne le rejet du paquet d’archivage

Le fichier SIP

- <?xml version="1.0" encoding="UTF-8"?>
- <pac xmlns="http://www.cines.fr/pac/sip" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.cines.fr/pac/sip http://www.cines.fr/pac/sip.xsd">
- <**DocDC**>
- <title>Das Florentiner Rechtsbuch, ein System römischen Privatrechts / aus der Glossatorenzeit, aus einer Florentiner Handschrift, zum ersten Mal herausgegeben und eingeleitet von Dr. Max Conrat (Cohn)</title>
- <creator>Cohn, Max (1848-1911)</creator>
- <subject>Droit romain - Réception</subject>
- <description>impression normale</description>
- <description>1 vol. (LVIII-118 p.) ; 20 cm</description>
- <description>Document numérisé avec OCR par la bibliothèque Cujas</description>
- <description>[Le Livre de Florence, traité de droit privé romain du temps des glossateurs]</description>
- <description>Cujas.30.466</description>
- <publisher>Weidmann, Officine. Berlin</publisher>
- <date>1882</date>
- <type>Text</type>
- <type>monographie imprimée</type>
- <format>image/tiff</format>
- <source>Bibl. Interuniversitaire Cujas 0605668496</source>
- <source>Bibl. Interuniversitaire Cujas 30.466</source>
- <language>ger</language>
- <rights>Domaine public. Voir mentions légales</rights>
- </**DocDC**>
- <**DocMeta**>
- <authenticite>oui</authenticite>
- <dureeConservation>P1000Y</dureeConservation>
- <identifiantDocProducteur>0605668496</identifiantDocProducteur>
- <noteDocument>pas de note</noteDocument>
- <serviceVersant>Bibliothèque interuniversitaire Cujas</serviceVersant>
- <structureDocument>structure non arborescente</structureDocument>
- <version>version 0</version>
- </**DocMeta**>
- <**FichMeta**>
- <compression>pas de compression</compression>
- <encodage>UTF-8</encodage>
- <formatFichier>XML</formatFichier>
- <nomFichier>DESC/0605668496.dc.xml</nomFichier>
- <**empreinteOri type="MD5">d2b0ffcd9119078e1398d4a2eefa05be</empreinteOri>**
- </**FichMeta**>
- <**FichMeta**>
- <compression>pas de compression</compression>
- <encodage>UTF-8</encodage>
- <formatFichier>XML</formatFichier>
- <nomFichier>DESC/0605668496.mets.xml</nomFichier>
- <empreinteOri type="MD5">de04d1cf3b93f247bc44fee639cecc3b</empreinteOri>
- </**FichMeta**>

Création des fichiers METS

- Hypothèse prestation: 14 000 euros pour 12 jours de travail
- Décision de travailler d'abord en interne
- Création manuelle de fichiers mets « basiques », sans la Structmap, pour les tests
- Automatisation pour les versements
- Les balises pointent vers des fichiers extérieurs chaque fois que possible pour alléger le travail
- Développement d'un outil automatisant la création du fichier mets (sans Structmap): dialogue permanent informaticien / bibliothécaire sur les données à récupérer et leur emplacement
- Automatisation du fichier mets « basique »
- En cours : intégration de la Structmap dans les fichiers METS (nécessaire en raison de la mise en place d'une visionneuse page à page dans l'interface de consultation)

Le fichier mets (sans la Structmap)


```
<?xml version="1.0" encoding="UTF-8" ?>
- <mets xmlns="http://www.loc.gov/METS/" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd">
- <metsHdr>
- <agent ROLE="CREATOR">
  <name />
  <note />
  </agent>
  <altRecordID />
  <metsDocumentID />
  </metsHdr>
- <dmdSec ID="Cujas0605668496">
  <mdRef LOCTYPE="URN" MDTYPE="DC" LABEL="DC" MIMETYPE="text/xml" XPTR="./DESC/0605668496.dc.xml" />
  </dmdSec>
- <amdSec>
- <techMD ID="TECH0605668496">
- <mdWrap MDTYPE="DC">
- <xmlData>
  <Description>Document numérisé avec OCR par la bibliothèque Cujas</Description>
  </xmlData>
  </mdWrap>
  </techMD>
- <rightsMD ID="RIGHT0605668496">
- <mdWrap MDTYPE="DC">
- <xmlData>
  <rights>Domaine public. Voir mentions légales</rights>
  </xmlData>
  </mdWrap>
  </rightsMD>
  </amdSec>
- <fileSec>
- <fileGrp USE="IMAGES_MASTER">
- <file ID="TIFF_0001" MIMETYPE="image/tiff">
  <FLocat LOCTYPE="URL" xlink:href="./IMAGES/0605668496_0001.tif" />
  </file>
- <file ID="TIFF_0002" MIMETYPE="image/tiff">
  <FLocat LOCTYPE="URL" xlink:href="./IMAGES/0605668496_0002.tif" />
  </file>
- <file ID="TIFF_0003" MIMETYPE="image/tiff">
  <FLocat LOCTYPE="URL" xlink:href="./IMAGES/0605668496_0003.tif" />
  </file>
- <file ID="TIFF_0004" MIMETYPE="image/tiff">
  <FLocat LOCTYPE="URL" xlink:href="./IMAGES/0605668496_0004.tif" />
```

Un exemple de dialogue bibliothécaire / informaticien

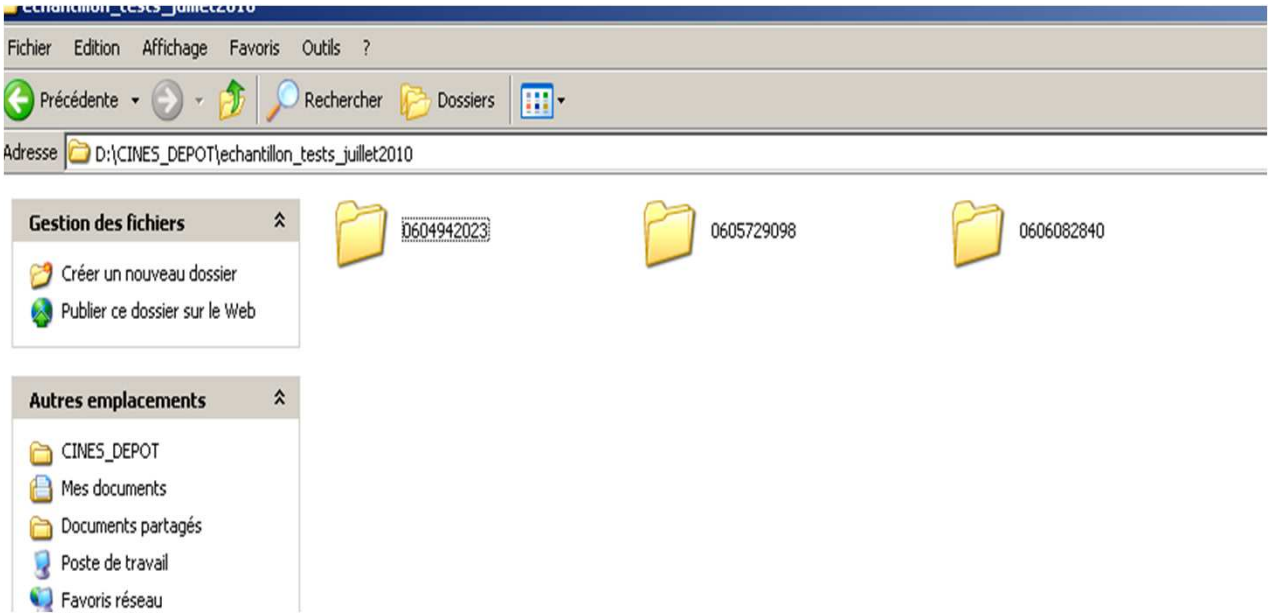
- **Fichier SIP**
- 1) Récupération de données bibliographiques issues de la notice Dublin core pour la balise <DocDC>

- <xsd:sequence><xsd:element ref="title" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="creator" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="subject" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="description" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="publisher" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="contributor" minOccurs="0" maxOccurs="unbounded"/><xsd:element ref="date" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="type" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="format" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="source" minOccurs="0" maxOccurs="unbounded"/><xsd:element ref="language" minOccurs="1" maxOccurs="unbounded"/><xsd:element ref="relation" minOccurs="0" maxOccurs="unbounded"/><xsd:element ref="coverage" minOccurs="0" maxOccurs="unbounded"/><xsd:element ref="rights" minOccurs="1" maxOccurs="unbounded"/> </xsd:sequence>
- **ATTENTION :**
- il faut respecter l'ordre des balises du sip.xsd
- la date doit être inscrite sous la forme AAAA-MM-JJ
- ne pas récupérer l'identifiant
- déqualifier les balises typées (voir tableau 1)
- 2)Récupération de données pour la balise <DocMeta><xsd:sequence><!-- la métadonnée authenticite est dépréciée--><xsd:element ref="authenticite" minOccurs="0" maxOccurs="1"/><xsd:element ref="dureeConservation" minOccurs="1" maxOccurs="1"/><xsd:element ref="identifiantDocProducteur" minOccurs="1" maxOccurs="1"/> <xsd:element ref="docRelation" minOccurs="0" maxOccurs="unbounded"/> <xsd:element ref="noteDocument" minOccurs="0" maxOccurs="1"/><xsd:element ref="serviceVersant" minOccurs="1" maxOccurs="1"/><xsd:element ref="structureDocument" minOccurs="0" maxOccurs="1"/><xsd:element ref="version" minOccurs="0" maxOccurs="1"/><xsd:element ref="versionPrecedente" minOccurs="0" maxOccurs="1"/></xsd:sequence>
- **A RECUPERER**
- <identifiantDocProducteur> = Récupérer l'identifiant du dossier à archiver = code-barres
- <serviceVersant> = Bibliothèque Interuniversitaire Cujas<version> = quand 1er versement, mettre version0

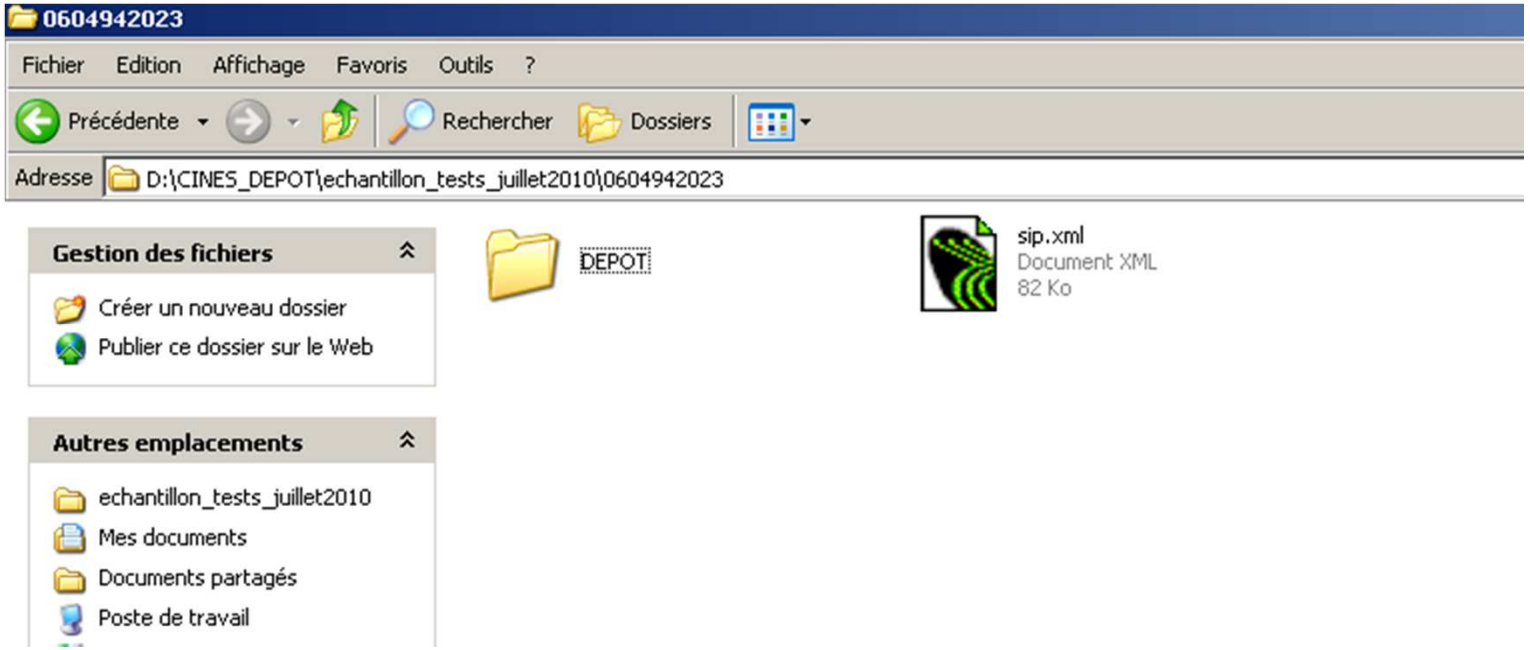
6. Constitution des paquets d’archives

Composition d’un paquet d’archivage (AIP)

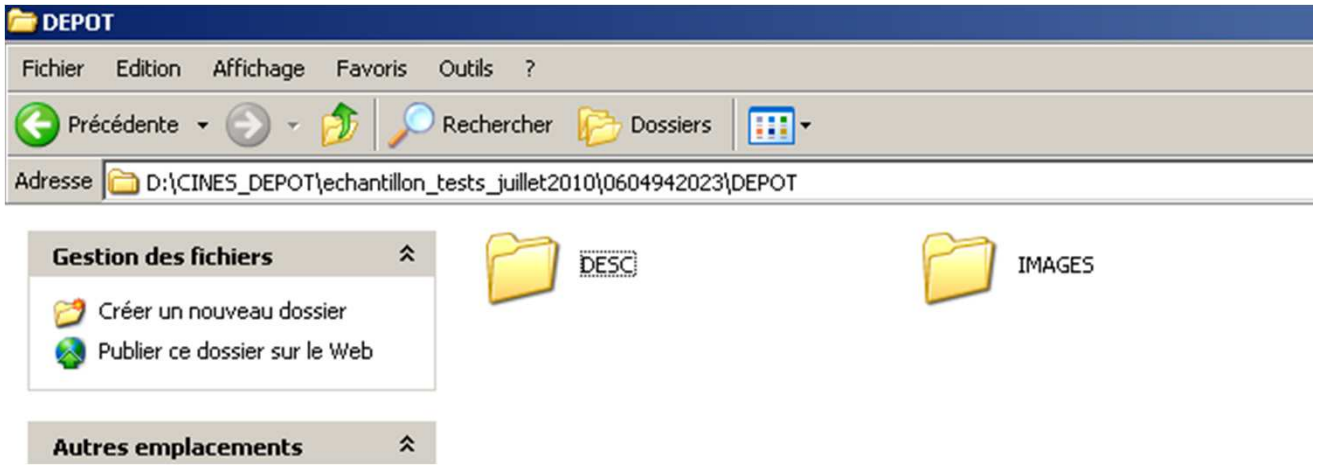
- *NB : 1 AIP = 1 volume physique*
- **Les différents paquets d’archive sont identifiés par le numéro de code-barres**



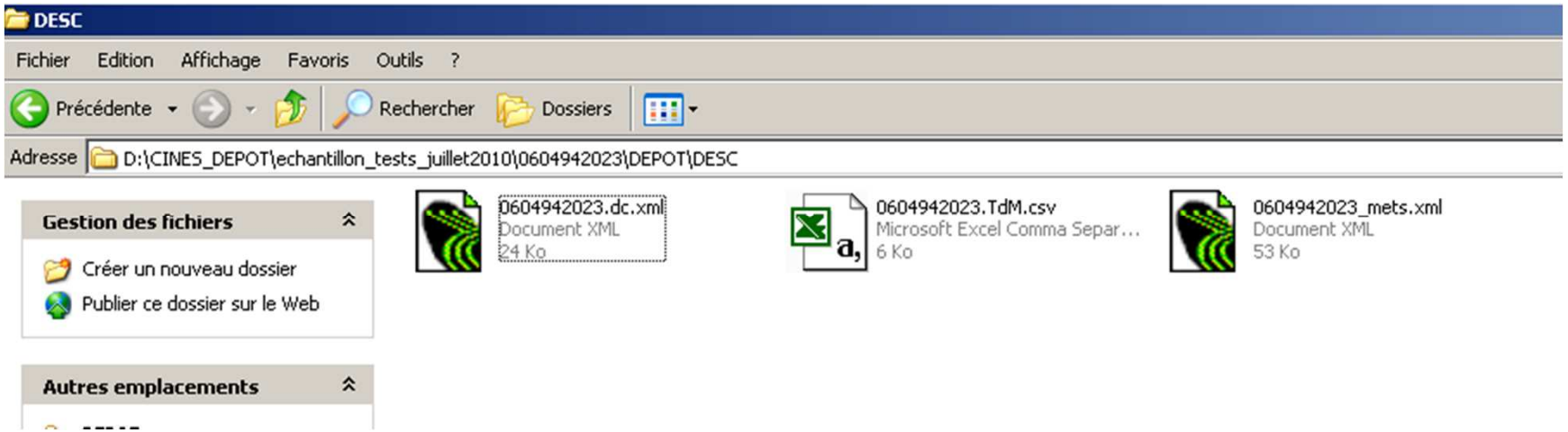
La composition du paquet d’archivage : répertoire dépôt et fichier sip.xml



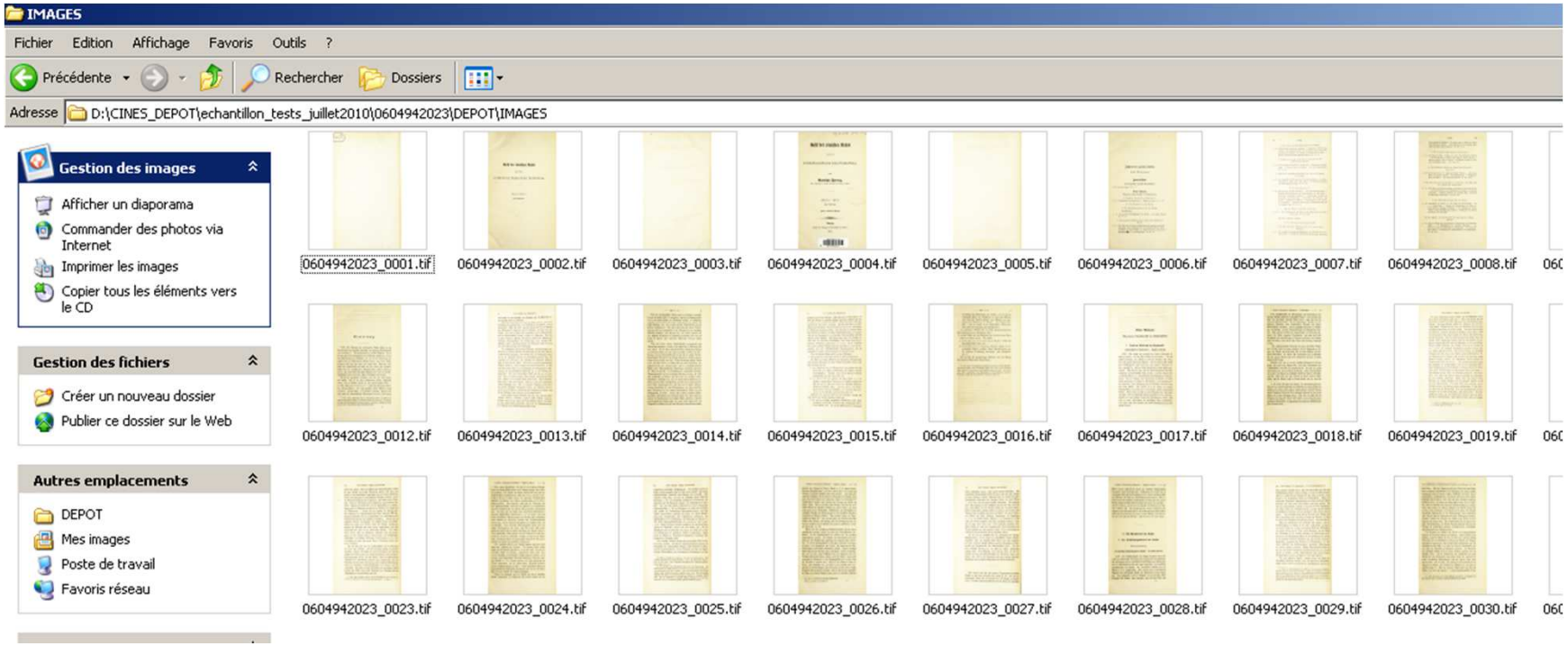
Composition du répertoire DEPOT : sous-répertoire DESC qui contient tous les fichiers de métadonnées et sous-répertoire IMAGES qui contient tous les fichiers master tif



Composition du fichier DESC : fichier Dublin Core (xml) + table des matières (csv utf8) + fichier mets (xml)



Composition du répertoire IMAGES : tous les fichiers tiff



Une fois en production…

- Rythme de versement: 6 livres par semaine au 1er semestre 2011, puis au fil de l'eau en fonction du rythme de numérisation
- Deux commandes Java pour
  - Constituer les paquets d'archives (10 mn env.)
  - Lancer la procédure d'envoi par FTP (30 mn de temps humain, 1 heure minimum pour l'envoi des fichiers)
- Formation d'une deuxième personne en doublon

Actuellement :

- Début 2013, tous les documents déjà en ligne sont archivés
- Fin 2013, 3 TO archivés
- Quelques échecs liés à
  - Des problèmes de connexion
  - De toutes petites erreurs dans une Dublin Core ou un fichier Excel
  - Des raisons non identifiées…

A venir…

- Intégration des documents numérisés par un prestataire extérieur
- Mais le cahier des charges a été rédigé avant la finalisation des procédures d’archivage…
- Mise au point nécessaire avec le prestataire au lancement du marché et à chaque contrôle qualité.

Le prix à payer

- Les libertés prises au début du projet ont obligé à un très gros travail de vérification et de réfection avant archivage: un an de travail sur les données mises en ligne depuis 2008, avec interruption du travail de numérisation interne pendant 1 an.
- Il a fallu refaire TOUT ce qui avait été mal fait au départ…
- Chaque changement majeur implique de reprendre l’archivage

Les enseignements

- Un dossier beaucoup plus lourd qu'on ne le pensait au début.
- Le PPDI oblige à relire en profondeur les objectifs et les méthodes de son programme.
- Ambitions revues à la baisse sur les formats à archiver, à la hausse sur la rigueur des procédures.
- Nécessite des compétences d'informatique documentaire (solides notions de xml, connaissance approfondie du Dublin Core) ET de développeur
- Dialogue permanent **indispensable** entre l'informaticien et le bibliothécaire
- Une chance exceptionnelle : la proximité informaticien / bibliothécaire au sein du même service

Pour le reste de l’équipe:

- L’archivage a des répercussions sur toute la chaîne et sur toute l’équipe.
- Nécessite une très grande rigueur dans TOUTES les étapes de la numérisation.
- AUCUNE ERREUR N’EST TOLEREE !
- C’est l’occasion de faire un contrôle qualité exigeant, avec une vraie motivation…

Pour des prestations externalisées de numérisation

- Nécessité d’un cahier des charges extrêmement précis et exigeant.
- Nécessité d’un contrôle qualité impitoyable.
- Ne pas hésiter à rejeter tout ce qui n’est pas conforme aux exigences pour obliger le prestataire à faire preuve d’une rigueur absolue.
- Les prestataires aussi découvrent les exigences de l’archivage…

Quelques conseils d’amie…

- Contacter son tiers archiveur et son service informatique le plus tôt possible
- Intégrer les procédures d’archivage d’emblée dans le projet
- Assurer le dialogue le plus étroit possible avec les informaticiens
- Apprendre à parler le xml dans le texte
- Exiger dès le départ, et maintenir, la plus grande rigueur dans les nommages, la constitution des dossiers, la saisie des notices et des tables des matières
- Faire un contrôle qualité rigoureux
- Ne pas utiliser de formats propriétaires, ou prévoir des conversions en formats libres
- Utiliser dès le départ l’UTF8

LA leçon à retenir :

- Un archivage pérenne réussi doit être conçu comme une démarche de projet dès le démarrage de la réflexion sur la numérisation.
- **La procédure d’archivage est la check list finale (nécessaire, indispensable, et en principe, suffisante…) de votre contrôle qualité.**

Merci de votre attention

noelle.balley@univ-paris1.fr