Archivage à Mathdoc

AG RNBM - 12/09/2022

Contenu sélectionné

- Livres et articles de Numdam et du centre Mersenne
- Données. Pour chaque article/livre :
  - PDF et DjVu (PDF/A prévu)
  - Matériel supplémentaire fournis par les auteurs
  - Images d’illustration
- Métadonnées. Pour chaque numéro/volume et livre :
  - XML au format JATS/BITS

Contenu sélectionné

- Sources des documents du centre Mersenne :
  - TeX (fichier TeX, bib, ...)
  - Images utilisées dans le PDF
- Sources des documents numérisés :
  - Images scannées (tif, jpg)
  - XML du prestataire
- Sources des documents nativement numériques acquis par Numdam :
  - XML fournis par l’éditeur

Contenu rangé

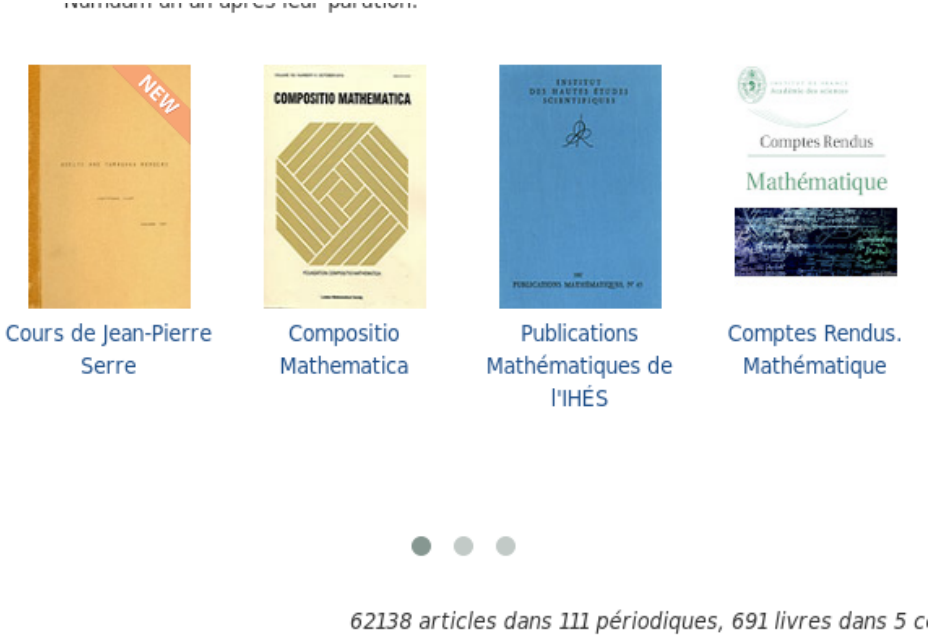
```
labbeo@ptf-tools:/mathdoc_archive$ ls ALCO
ALCO_2018__1_1  ALCO_2019__2_4  ALCO_2020__3_6  ALCO_2022__5_2
ALCO_2018__1_2  ALCO_2019__2_5  ALCO_2021__4_1  ALCO_2022__5_3
ALCO_2018__1_3  ALCO_2019__2_6  ALCO_2021__4_2  ALCO.jpg
ALCO_2018__1_4  ALCO_2020__3_1  ALCO_2021__4_3  ALCO_small.jpg
ALCO_2018__1_5  ALCO_2020__3_2  ALCO_2021__4_4  ALCO.xml
ALCO_2019__2_1  ALCO_2020__3_3  ALCO_2021__4_5
ALCO_2019__2_2  ALCO_2020__3_4  ALCO_2021__4_6
ALCO_2019__2_3  ALCO_2020__3_5  ALCO_2022__5_1
```

Contenu rangé

- AFST/
  - AFST.xml
  - AFST.jpg
  - AFST_2022_6_31_3/
    - AFST_2022_6_31_3.xml
    - AFST_2022_6_31_3_861_0/
      - AFST_2022_6_31_3_861_0.pdf
      - AFST_2022_6_31_3_861_0.djvu
      - src/tex/

Contenu mis à jour régulièrement

On repart de zéro à chaque itération de Numdam



numdam-pre → numdam.org

Sauvegarde de l’archive

- Sur les serveurs de Numdam, du centre Mersenne
- Sur les serveurs de Mathrice
  - Grenoble
  - Bordeaux
  - Lille

  eux-mêmes sauvegardés, surveillés
- Sur des disques durs externes

Création de l’archive

- Automatiquement, de façon contrôlé, par une application
- Outil de vérifications, écrit avec du code différent de celui qui sert à créer l’archive
  - Actuellement, vérification de la présence des fichiers

Utilisation de l’archive

- Par les développeurs pour recréer des sites web avec du contenu
  - Environ 1 jour pour un créer un site local (grâce à notre plateforme PTF de sites web)

OAIS

- Modèle abstrait
- Spécifie de manière très générale l'architecture logique et les fonctionnalités d'un système d'archivage

https://www.cines.fr/archivage/un-concept-des-problematiques/le-modele-de-reference-loais/

OAIS

« une information de structure qui explique la façon dont d'autres informations sont organisées (ex. : les tables de correspondance entre les noms de fichiers et les numéros de page pour un ouvrage numérisé) »

```
<article>
  <idart>AIF_1949__1__1_0</idart>
  <cphys>page0011.tif page0012.tif
page0013.tif page0014.tif</cphys>
</article>
```

LOCKSS

- v1 : début des années 2000
- v2.0-alpha: 04/2019
- v2.0-alpha5 : 12/2021.
  Pas de conversion prévue entre la v1 et la v2
  « However, an upcoming version of the classic LOCKSS system (v1) will contain an experimental tool to test the migration »

=> Pas de version stable avant 2-3 ans ?

LOCKSS

Moissonner les métadonnées

- Écriture de code XML et Java pour chaque éditeur
- CLOCKSS pour Mersenne : serveur FTP, puis site web local, moissonné par une instance LOCKSS

=> Même travail que Geodesic

LOCKSS

Comment accéder aux données ?

CLOCKSS : « steps is required to extract the content from the directories or archive files that were copied from the CLOCKSS box repository, and generate from them a web site »

http://documents.clockss.org/index.php/CLOCKSS:_Extracting_Triggered_Content

Archivage pérenne pour collections RNBM

Geodesic étendu à des collections sous licences

- Gestion des droits d’accès (déjà fait par le Cedram)
- Stockage des PDFs