EuDML

Le projet EuDML

The European Digital Mathematics Library

Thierry Bouche

Cellule MathDoc & institut Fourier,
Université de Grenoble

L'IST au prisme de l'Europe
Journées FréDoc 2011
Bordeaux
11 octobre 2011

Plan

La documentation en mathématiques

Enjeux spécifiques

- La documentation mathématique *validée* ne se périme pas (Euler 1999)
- Les résulats anciens ne sont pas remplacés par les nouveaux : ils sont leur fondation (Richelot 2004)
- Elle est valide comme un *tout*, qui forme un vaste réseau (Corona bug)
- Elle est utile pour d'autres sciences, de façon *asynchrone* (Weber crypto)

$\Longrightarrow$ Elle doit donc être soigneusement validée, rangée, indexée et conservée (GDZ Spr. Zbl MR)

$\Longrightarrow$ Elle doit rester accessible sur le très long terme (Galois 1828)

La documentation en mathématiques

Enjeux spécifiques

- La documentation mathématique *validée* ne se périme pas (Euler 1999)
- Les résulats anciens ne sont pas remplacés par les nouveaux : ils sont leur fondation (Richelot 2004)
- Elle est valide comme un *tout*, qui forme un vaste réseau (Corona bug)
- Elle est utile pour d'autres sciences, de façon *asynchrone* (Weber crypto)

$\Longrightarrow$ Elle doit donc être soigneusement validée, rangée, indexée et conservée (GDZ Spr. Zbl MR)

$\Longrightarrow$ Elle doit rester accessible sur le très long terme (Galois 1828)

La documentation en mathématiques

La bibliothèque de référence

Nous avons donc besoin d'une bibliothèque

- exhaustive
- à jour
- bien rangée
- grande ouverte
- facile d'accès pour les non-mathématiciens

Papier OK ? (bibliothèques, prêt inter., fourniture de documents, catalogues fusionnés, bases de données MR/ZM...)

Électronique Un rêve... (WDML : NSF, IMU, EMS, Moore, EMANI...)

$\longrightarrow$ De nombreux projets de numérisation depuis l'an 2000
Quelques projets d'acquisition native
(ELibM, ERAM, NUMDAM, DML locales, etc.)

$\longrightarrow$ EuDML premier projet (pilote) d'intégration international

La documentation en mathématiques

La bibliothèque de référence

Nous avons donc besoin d'une bibliothèque

- exhaustive
- à jour
- bien rangée
- grande ouverte
- facile d'accès pour les non-mathématiciens

Papier OK ? (bibliothèques, prêt inter., fourniture de documents, catalogues fusionnés, bases de données MR/ZM...)

Électronique Un rêve... (WDML : NSF, IMU, EMS, Moore, EMANI...)

$\Longrightarrow$ De nombreux projets de numérisation depuis l'an 2000
Quelques projets d'acquisition native
(ELibM, ERAM, NUMDAM, DML locales, etc.)

$\Longrightarrow$ EuDML premier projet (pilote) d'intégration international

La documentation en mathématiques

La bibliothèque de référence

Nous avons donc besoin d'une bibliothèque

- exhaustive
- à jour
- bien rangée
- grande ouverte
- facile d'accès pour les non-mathématiciens

Papier OK ? (bibliothèques, prêt inter., fourniture de documents, catalogues fusionnés, bases de données MR/ZM...)

Électronique Un rêve... (WDML : NSF, IMU, EMS, Moore, EMANI...)

$\Longrightarrow$ De nombreux projets de numérisation depuis l'an 2000
Quelques projets d'acquisition native
(ELibM, ERAM, NUMDAM, DML locales, etc.)

$\Longrightarrow$ **EuDML** premier projet (pilote) d'intégration international

La documentation en mathématiques

Échelle de temps

- Prépublications instantanées (labos, arXiv/HAL, courriel, pages perso)
- Délais de publication assez longs : 1-2 ans
- Publication à fins de prestige, carrière et d'attribution
  Fournit une version de référence pour les travaux à venir
- Seulement 50 % des articles cités aujourd'hui
  sont parus il y a moins de 10 ans
- Environ 25 % des articles cités aujourd'hui
  sont parus il y a plus de 20 ans

La documentation en mathématiques

Dimension modeste, forte croissance

Une estimation de la taille du corpus mathématique publié dans la tradition occidentale depuis Euclide :

- 3 millions de textes couvrant $<$ 100 millions de pages
- 100 000 nouveaux textes paraissent chaque année
- 80% articles de revues, 10% chapitres dans des ouvrages collectifs, 10% livres
- $<$ 10% parus avant 1900
- $>$ 80% parus après 1950

La documentation en mathématiques

Une grande variété d'acteurs

Grande diversité éditoriale, pas de modèle économique dominant

- Environ 600 revues vivantes dédiées à la recherche mathématique (dont une vingtaine en France)
- 2000 périodiques comportant des articles de maths
- Importance des livres
- De nombreux éditeurs de taille modeste font un travail scientifique de premier plan (laboratoires, sociétés savantes, PME...)
- Les publications de laboratoires sont souvent en accès libre
- Les structures privées préfèrent souvent assurer la pérennité de leurs services en limitant le libre accès (embargo partiel ou total)

The European Digital Mathematics Library

CIP-ICT-PSP.2009.2.4 Open access to scientific information

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

The European Digital Mathematics Library

Vision EuDML (2008)

La bibliothèque numérique de mathématiques devrait s'efforcer de réunir un corpus mathématique **aussi vaste que possible** pour

- le **préserver** à très long terme,
- le rendre **disponible en ligne**
- en accès **libre à terme**,
- sous la forme d'un collection **de référence**,
- **alimentée** en continu par les nouveautés des éditeurs,
- **valorisée** par des outils de recherche et d'interopérabilité sophistiqués,
- développée et entretenue par un réseau d'**institutions**

$\Longrightarrow$ EuDML, projet pilote CIP 2010-2013

The European Digital Mathematics Library

Vision EuDML (2008)

La bibliothèque numérique de mathématiques devrait s'efforcer de réunir un corpus mathématique **aussi vaste que possible** pour

- le **préserver** à très long terme,
- le rendre **disponible en ligne**
- en accès **libre à terme**,
- sous la forme d'un collection **de référence**,
- **alimentée** en continu par les nouveautés des éditeurs,
- **valorisée** par des outils de recherche et d'interopérabilité sophistiqués,
- développée et entretenue par un réseau d'**institutions**

$\Longrightarrow$ **EuDML**, projet pilote CIP 2010-2013

Le projet EuDML

Fiche d'identité

EuDML Implémentation pilote (orientée utilisateur final) d'un guichet d'accès unique au contenu mathématique fourni par 11 institutions,
avec des fonctions innovantes de recherche, accessibilité, multilinguisme, navigation et interactivité

Consortium $12 + 1^2$ participants européens, $1 + 1^2$ partenaires associés
Portugal (1), Royaume-Uni (2), Espagne (2), France (3), Allemagne (2), Pologne (1), République Tchèque (2), Grèce (1), Bulgarie (1)

Profil 3 années (01/02/2010-31/01/2013), 487 PM,
Financement max CE : 1,6 M€

Contenu **250 revues ; 235 000 textes ; 2 600 000 pages**

The European Digital Mathematics Library

Un portail global

Un point d'accès unifié

Pour les utilisateurs Un portail web personnalisable permettant de feuilleter, fouiller, naviguer les collections

Pour les machines Des services pour transformer les références en liens

Bénéfices attendus

- Fouille de textes mathématiques facilitée
- Plus de visibilité pour un corpus éclaté
- Un seul service pour lier les références
- Valeur ajoutée aux articles nouveaux : les références pointent *quelque part*

The European Digital Mathematics Library

Un portail global

Un point d'accès unifié

Pour les utilisateurs Un portail web personnalisable permettant de feuilleter, fouiller, naviguer les collections

Pour les machines Des services pour transformer les références en liens

Bénéfices attendus

- Fouille de textes mathématiques facilitée
- Plus de visibilité pour un corpus éclaté
- Un seul service pour lier les références
- Valeur ajoutée aux articles nouveaux : les références pointent *quelque part*

The European Digital Mathematics Library

Une archive répartie

Un réseau d'institutions

- Démultiplier l'impact des projets DML européens
- Une sorte de dépot légal volontaire pour les textes mathématiques
- Une archive répartie indépendante des textes intégraux
- Des institutions scientifiques pérennes sans but lucratif pour assurer l'entretien et la préservation à long terme des collections

Bénéfices attendus

- Les contenus sont pris en charge par des tiers pour l'intérêt public
- Disponibilité du corpus sur le long terme
- Les producteurs de contenus n'ont pas à se préoccuper de l'archivage pérenne
- Les éditeurs font leur métier, les bibliothécaires le leur
  (édition : susciter, sélectionner, produire les meilleurs textes
  bibliothèques : sélectionner, acquérir, organiser, indexer, donner accès)

The European Digital Mathematics Library

Une archive répartie

Un réseau d'institutions

- Démultiplier l'impact des projets DML européens
- Une sorte de dépot légal volontaire pour les textes mathématiques
- Une archive répartie indépendante des textes intégraux
- Des institutions scientifiques pérennes sans but lucratif pour assurer l'entretien et la préservation à long terme des collections

Bénéfices attendus

- Les contenus sont pris en charge par des tiers pour l'intérêt public
- Disponibilité du corpus sur le long terme
- Les producteurs de contenus n'ont pas à se préoccuper de l'archivage pérenne
- Les éditeurs font leur métier, les bibliothécaires le leur
  (édition : susciter, sélectionner, produire les meilleurs textes
  bibliothèques : sélectionner, acquérir, organiser, indexer, donner accès)

The European Digital Mathematics Library

Libre accès à terme

Le créneau mobile

- Lorsqu'un éditeur a complété sa production, une copie archivable (métadonnées et tectes intégraux) est fournie à l'institution *ad hoc*
- Ces contenus sont validés et enregistrés
- Les nouveaux textes sont indexés et apparaissent dans les résultats de recherches
- L'accès aux textes intégraux se fait sur le site de l'éditeur sous son contrôle
- À l'issue du créneau mobile, la copie locale devient librement accessible

Bénéfices attendus

- Les textes intégraux sont archivés par des tiers pérennes
- Meilleures visibilité et navigabilité du corpus, y compris récent
- La quantité de textes de référence en libre accès augmente
- La mathématique établie, mère de toutes les sciences, à la portée de tous !

The European Digital Mathematics Library

Libre accès à terme

Le créneau mobile

- Lorsqu'un éditeur a complété sa production, une copie archivable (métadonnées et tectes intégraux) est fournie à l'institution *ad hoc*
- Ces contenus sont validés et enregistrés
- Les nouveaux textes sont indexés et apparaissent dans les résultats de recherches
- L'accès aux textes intégraux se fait sur le site de l'éditeur sous son contrôle
- À l'issue du créneau mobile, la copie locale devient librement accessible

Bénéfices attendus

- Les textes intégraux sont archivés par des tiers pérennes
- Meilleures visibilité et navigabilité du corpus, y compris récent
- La quantité de textes de référence en libre accès augmente
- La mathématique établie, mère de toutes les sciences, à la portée de tous !

The European Digital Mathematics Library

Innovation

Nous intégrons des nouvelles technologies

- Métadonnées MathML (OCR, conversions LaTeX, extraction PDF)
- Recherche de formules
- Mathématiques accessibles
- Relations sémantiques
- Similarité et classification des textes

Bénéfices attendus

- Progrès en gestion des savoirs mathématiques
- Nouvelles modalités de découverte et de navigation
- Banc d'essai pour de nouveaux modes d'interaction avec le corpus
- Éprouver des outils de production réutilisables
- Retourner des métadonnées améliorées aux fournisseurs de contenu

The European Digital Mathematics Library

Innovation

Nous intégrons des nouvelles technologies

- Métadonnées MathML (OCR, conversions LaTeX, extraction PDF)
- Recherche de formules
- Mathématiques accessibles
- Relations sémantiques
- Similarité et classification des textes

Bénéfices attendus

- Progrès en gestion des savoirs mathématiques
- Nouvelles modalités de découverte et de navigation
- Banc d'essai pour de nouveaux modes d'interaction avec le corpus
- Éprouver des outils de production réutilisables
- Retourner des métadonnées améliorées aux fournisseurs de contenu

Les contenus EuDML

Synthèse

Collections **225 périodiques et séries, 235 000 textes, 2 600 000 pages**

Allemagne	ERAM/JFM, GDZ, ELibM (85 000 textes)
France	Gallica-Math, NUMDAM, CEDRAM, EDPS, TEL (50 000 textes)
Rép. Tchèque	DML-CZ (27 000 textes)
Russie	RusDML (17 000 textes)
Pologne	DML-PL (14 000 textes)
Grèce	HDML (2 400 textes)
Espagne	DML-E (6 400 textes)
Portugal	SPM/BNP (2 000 textes)
Bulgarie	BulDML (450 textes)

Rétro BNP/SPM/IST, DML-CZ, DML-E, DML-PL, Gallica, GDZ, HDML, NUMDAM, RusDML

Natif BulDML, CEDRAM, DML-CZ, DML-E, DML-PL, EDPS, ELibM, NUMDAM

Les contenus EuDML

Sélection

Processus En cascade : Projet → institution → collections

Critères Textes mathématiques **publiés** et **validés** scientifiquement, destinés à servir de **référence**

- Pour être éligible, il faut une paire (**texte intégral** [PDF], **métadonnées** [XML]) archivée par l'une des institutions partenaires

Items Un *item* EuDML est l'unité logique pertinente pour l'utilisateur Une **monographie**, un **volume**, une **œuvre** en plusieurs tomes, un **article** de revue, une **contribution** dans un livre collectif, une **communication** publiée dans des actes

- À ce jour : **235 000 items** dans **12 collections** (185 000 articles, 45 000 chapitres et contributions, 2 500 livres, 300 œuvres en plusieurs tomes)

Les contenus EuDML

Détenteurs du copyright

Domaine public	Quelques revues, la plupart des livres
Public	50 universités, académies, instituts, laboratoires
Fondations	Compositio Mathematica, quelques ASBL
Sociétés	20 sociétés savantes
Éditeurs	45 revues

Birkhäuser	5 revues (GDZ)
EDPS	7 revues (5 à jour dans NUMDAM)
Elsevier	5 revues, 1 à jour (NUMDAM)
de Gruyter	2 revues (GDZ)
Heldermann	6 revues (5 à jour dans ELibM)
Hindawi	12 revues (à jour dans ELibM)
Noordhoff	1 revue (NUMDAM)
AK Peters	1 revue (ELibM)
Springer	2 périodiques (NUMDAM, 1 revue à jour $\rightarrow$ 2007)
	9 revues (GDZ)

Métadonnées EuDML

Le babel des formats

Les métadonnées ont des structures et des niveaux de détail très variables

SQL	Base de données maison : DML-E
DTD maison	MathDoc, FIZ, IST, HDML
DTD standard	DC, Dspace, minidml, METS, NLM

Nous avons basé le format EuDML v 1.0 sur
NLM Journal Archiving and Interchange Tag Suite
pour le stockage et l'échange des métadonnées

Métadonnées EuDML

Schéma EuDML

NLM Journal Archiving and Interchange Tag Suite

Pour

- Largement testé et exploité (EDPS, PubMed Central, JSTOR...)
- Standard NISO
- Précis et flexible (données structurées *et* plates)
- Support MathML et *alternatives*
- Description du contenu de périodiques, livres, collections
- Permet de stocker l'information de tous les partenaires, et extensible

Contre

- Nécessite un "application profile"
- Conçu pour les textes intégraux
- Pas tous les types de documents prévus (chapitre dans un livre édité...)

$\Longrightarrow$ EuDML schema, v. 1.0

- Trois types de documents comme conteneurs : **article**, **book**, **mbook**
- Déviation minimale par rapport aux DTD NLM standard
- "Best practices recommendation"

Résultats

Résultats de la première période

- Analyse des contenus, définition d'un format d'échange
- Conférence à Prague en octobre 2010 pour discuter avec des partenaires potentiels (Springer, LMS, . . .) sur la base
  - Libre accès à terme (créneau mobile)
  - Nouveautés fournies par les éditeurs et indexées rapidement
  - Archivage partagé dans un réseau de bibliothèques numériques de référence
- Constition d'une base de données avec **235 000 références**
  environ 1/5 du contenu numérique existant (1/3 de l'existant « DML »)
- Une série d'outils destinés à améliorer l'indexation, l'accès
  et la visibilité de ce corpus spécifiquement mathématique
- Première démo publique d'un prototype à demi fonctionnel

This being partially funded by the CIP
COMPETITIVENESS AND INNOVATION FRAMEWORK PROGRAMME
(ICT PSP Digital Libraries, project # 250.503)

We will *deliver* a truly open, sustainable and *innovative* framework for *access* and exploitation of Europe's rich heritage of *mathematics*.

Thierry BOUCHE
Institut Fourier & Cellule MathDoc, Grenoble
MathDoc *director*
EuDML *scientific coordinator*
EMS Electronic Publishing Committee
CICM Steering Committee
IMU Committee on Electronic Information and Communication