EuDML

Le projet EuDML

Thierry Bouche

Cellule MathDoc & institut Fourier,
Université de Grenoble

Documentation mathématique
CIRM, Luminy, 11 octobre 2010

Plan

1. La documentation mathématique
2. La documentation mathématique électronique
3. La bibliothèque numérique de mathématiques (WDML)
4. Le projet EuDML

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. . .)
Mais les formats de référence sont désormais numériques (et les chercheurs sont impatients !)

Électronique Un rêve. . . (WDML)

$\Longrightarrow$ De nombreux projets (numérisation) depuis l'an 2000 (ELibM, ERAM, NUMDAM, WDML, 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...)
Mais les formats de référence sont désormais numériques (et les chercheurs sont impatients !)

Électronique Un rêve... (WDML)
$\Longrightarrow$ De nombreux projets (numérisation) depuis l'an 2000 (ELibM, ERAM, NUMDAM, WDML, 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...)
Mais les formats de référence sont désormais numériques (et les chercheurs sont impatients !)

Électronique Un rêve... (WDML)

$\Longrightarrow$ De nombreux projets (numérisation) depuis l'an 2000 (ELibM, ERAM, NUMDAM, WDML, etc.)

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

La documentation en mathématiques

Échelle de temps

- Prépublications instantanées (labos, arXiv, 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
- Environ 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
- $<$ 20% parus avant 1900
- $>$ 50% 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)

La documentation en mathématiques

Étude de cas

Trois parmi les meilleures revues du monde

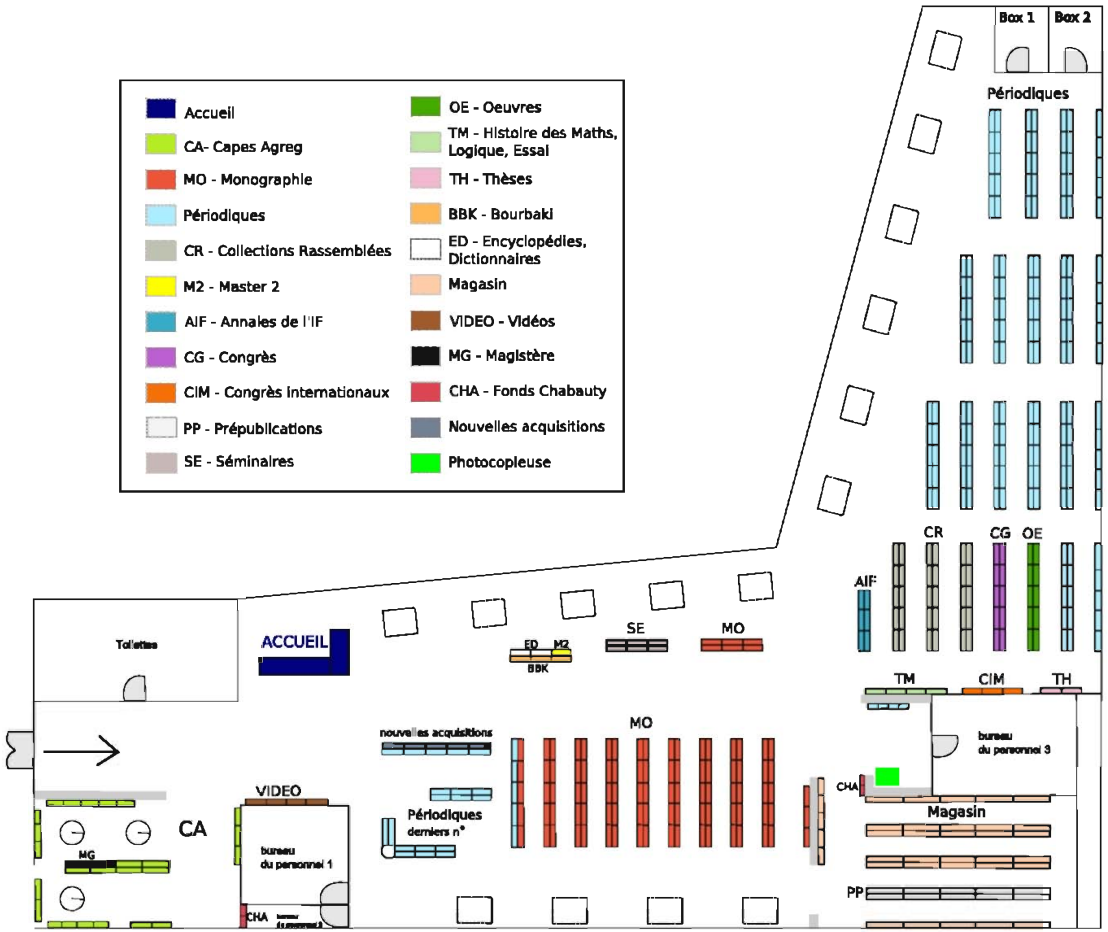
Annals of Math. revue publiée par un laboratoire (Princeton/IAS), successivement éditée en interne, en accès libre sur arXiv et EMIS, puis sur project Euclid, désormais archives chez JSTOR (après un embargo de 5 ans) et édition électronique MSP (payante mais pas chère)

Inventiones Math. revue commerciale Springer, archives numérisées libre d'accès sur GDZ (jusqu'en 1996)

Publ. Math. IHES revue publiée par une institution indépendante, éditée en interne, imprimée par les PUF au xx^{e} s., désormais fabriquée et diffusée par Springer, archives libres d'accès sur NUMDAM (après un embargo de 5 ans)

Une bibliothèque de mathématiques typique

(Institut Fourier, France)



Documentation mathématique & électronique

Les avantages du numérique

- Passer à l'électronique *devrait* être un atout formidable pour ouvrir de nouvelles voies de fouille dans le corpus mathématique et d'accès à celui-ci
- L'infrastructure nécessaire fournirait les fonctions de base de la bibliothèque de référence, enrichie des possibilités offertes par le numérique
- Soit :
  - Une archive globale (distribuée) stockant ce qui paraît dans le monde au fur et à mesure (par numérisation ou nativement numérique)
  - Un registre à jour de toutes les ressources disponibles
  - Des outils pour résoudre des références ou croiser les contenus de différentes bases de données
  - Une navigation sans frontières dans la totalité du corpus
  - Un accès instantané

Documentation mathématique & électronique

Les inconvénients du numérique

Le numérique n'a pas que des avantages pour les chercheurs et les bibliothécaires...

- Les grandes plateformes d'édition et les outils courants ne sont pas adaptés au contenu mathématique
- Un babel de « standards » pour la structuration des données, les interfaces utilisateurs, les formats de fichier, etc.
- Des nouvelles barrières d'accès apparaissent (copyright, licences, DRM...)
- Tendance à la concentration (paresse de utilisateurs et interopérabilité limitée, cf. Microsoft ou Google !)
- Les coûts augmentent
- On mesure la « valeur » d'une publication à partir d'indicateurs faciles à produire mais dont la signification reste à déterminer (nombre de téléchargements, nombre de liens ou de citations, « impact », etc.)

Documentation mathématique & électronique

Désordres

- De nombreux textes parus sur papier n'ont pas d'équivalent numérique, *mais*
- de nombreux textes numériques sont dupliqués sur plusieurs sites (dans des versions différentes ou non), *tandis que*
- de nombreuses collections sont découpées entre plusieurs sites, *et*
- les sites disparaissent ou changent d'adresse, de propriétaire ou de politique commerciale comme de chemise

$\Longrightarrow$ Gérer un accès exhaustif et à jour au corpus numérique dans son ensemble nécessite des ressources financières et humaines infinies

Documentation mathématique & électronique

Risques

- Une désintermédiation dans l'édition avec perte de qualité (auto-archivage et autres modalités d'auto-édition : pages perso, « personal collected works »...), de fiabilité (qui valide ? qui relit ?) et de pérennité (la version citée est-elle celle qui a été lue ?)
- Une privatisation de la fonction bibliothèque (projets de numérisation d'Elsevier, Springer, mais aussi bibliothèques numériques universitaires « opérées » par NUMILOG...)
- La multiplication d'archives privées, non interopérables, conçues comme des produits commerciaux qui seront abandonnés lorsqu'ils ne seront plus rentables
- Une très grande visibilité des catalogues, mais un accès aux contenus réservé aux plus riches

The Digital Mathematics Library

Vision

"In light of mathematicians' reliance on their discipline's rich published heritage and the key role of mathematics in enabling other scientific disciplines, the Digital Mathematics Library strives to make the entirety of past mathematics scholarship available online, at reasonable cost, in the form of an authoritative and enduring digital collection, developed and curated by a network of institutions."

NSF DML project, Cornell 2002, CEIC 2004, IMU 2006

Projets WDML

Historique

- Généralisation de l'édition électronique, puis numérisation à partir de 1997 (JSTOR, Gallica, ERAM/JFM, NUMDAM...)
- John Ewing. "Twenty Centuries of Mathematics : Digitizing and Disseminating the Past Mathematical Literature" (2000)
- Digital Mathematics Library. NSF planning project (2002-2003, Cornell University Library) "toward the establishment of a comprehensive, international, distributed collection of digital information and published knowledge in mathematics".
- Mathematical Knowledge Management meetings (2001–) + DML workshops (2008–) : technical challenges.
- EMS' expression of interest to the European Commission (2003)
  Proposals to EC programmes (2003–2008 : FP6, eContentplus...)
- AMS/MSRI proposal to the Moore foundation (2005)
- EMANI (2002-2007 : Springer + bibliothèques)
- Contre-attaque commerciale (Elsevier Backfiles, Springer Online Archives...)
- IMU support (2002–2006 : Vision, Best practices)
- EuDML (01/02/2010-31/01/2013)

Projets WDML

Collections

- **Ameriques** JSTOR (250 000 textes), project Euclid (110 000), CMS (4 000)
- **Asie** DML-JP (30 000 textes), China ? ?
- **Europe** EuDML+ (250 000 textes)
  - **Allemagne** ERAM/JFM, GDZ, ELibM (120 000 textes)
  - **France** Gallica-Math, NUMDAM, CEDRAM, TEL (50 000 textes)
  - **Grèce** HDML (8 000 textes)
  - **Pologne** ICM/BWM (13 000 textes)
  - **Portugal** SPM/BNP (2 000 textes)
  - **Espagne** DML-E (6 500 textes)
  - **Rép. Tchèque** DML-CZ (25 000 textes)
  - **Russie** RusDML (17 000 textes)
  - **Bulgarie** BulDML (270 textes)
  - **Serbie** 4 400 textes
  - **Suisse** SwissDML (5 000 textes)
- **Commercial** 700 000 textes ?
  - **PME** CUP 20 revues, OUP 30, Hindawi 18, Walter de Gruyter 13, Wiley 42, Taylor & Francis 58...
  - **Elsevier** 4 revues in NUMDAM, 63 in Backfiles, 100 alive (320 000)
  - **Springer** 14 revues in GDZ, 1+2 in NUMDAM, 120 in Online Archives, 179 alive (300 000)

The European Digital Mathematics Library

Vision corrigée (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 référencement sophistiqués,
- développée et entretenue par un réseau d'**institutions**

$\Longrightarrow$ EuDML, implémentation pilote d'un point d'accès unique au contenu de 12 partenaires européens

The European Digital Mathematics Library

Vision corrigée (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 référencement sophistiqués,
- développée et entretenue par un réseau d'**institutions**

$\Longrightarrow$ **EuDML**, implémentation pilote d'un point d'accès unique au contenu de 12 partenaires européens

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

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

Consortium $12 + 1^2$ participants européens, $1 + 1^2$ partenaires associés

Objectifs 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

Profil 3 années (01/02/2010-31/31/2013), 488 PM, coût global : 3,20 M€ (financé pour moitié par la CE).

Contenu 235 000 textes ; 2 600 000 pages

Rétronumérisé NUMDAM, Gallica, DML-PL, GDZ, SPM/BNP, HDML, DML-CZ, DML-E, RusDML.

Numérique natif BulDML, CEDRAM, DML-PL, EDPS, ELibM, DML-CZ, DML-E

The European Digital Mathematics Library

Consortium

IST	**Gestion & Coordination technique** Instituto Superior Técnico (Lisbonne, Portugal)
UJF/CMD	**Coordination scientifique** Université Joseph-Fourier : MathDoc (Grenoble)
CNRS/CMD	Centre national de la recherche scientifique : MathDoc (France)
UB	University of Birmingham : Computer Science Dpt. (Royaume Uni)
FIZ	Fachinformationszentrum : Zentralblatt (Karlsruhe, Allemagne)
MU	Masarykova univerzita : Informatique (Brno, République tchèque)
ICM	University of Warsaw : ICM (Pologne)
CSIC	Consejo superior de investigaciones cientificas : IEDCYT (Madrid, Espagne)
EDPS	Édition Diffusion Presse Sciences (Paris, France)
USC	Universidade de Santiago de Compostela : Instituto de Matemáticas (Espagne)
IMI-BAS	Institute of Mathematics and Informatics, BAS (Sofia, Bulgarie)
IMAS	Matematicky Ustav Av Cr V.V.I. (Prague, République tchèque)
IU	Ionian University : Informatics Dpt. (Corfou, Grèce)
MML	Made Media UK (Birmingham, Royaume Uni)
	~
EMS	European Mathematical Society
SUBGoe	Bibliothèque universitaire de Göttingen (Allemagne)

The European Digital Mathematics Library

Fonctionnalités attendues

- Fonctions de base pour la découverte et l'accès aux textes : feuilletage, recherche avancée, recherche de citations
- *batch lookup* pour transformer une référence en lien de façon massive (MathWorld, Wikipédia, éditeurs, etc.)
- Support spécifique pour les maths (affichage, recherche adaptée)
- Un max de liens (bases de données, citations, similitude...)
- Personnalisation (profils d'utilisateurs, espace de travail sauvegardé)
- Interactivité (communautés, commentaires partagés)
- Accessibilité, en particulier pour les malvoyants

The European Digital Mathematics Library

Collections

Partenaires EuDML : bibliothèques numériques

BulDML	3 revues
DML-CZ	11 revues, 6 série de conférences, 35 livres
DML-E	22 revues
DML-PL	10 revues, 4 série de livres
HDML	8 revues, 29 conférences, 20 livres
NUMDAM	30 revues, 29 séminaires, 270 Doct. Th., 1 série de monographies
SPM/BNP	1 revue

Partenaires EuDML : édition électronique

CEDRAM	10 prédiodiques
ELibM	91 revues
EDP Sciences	7 revues

The European Digital Mathematics Library

Collections

Partenaire un peu spécial

ZMATH 3 million reviews

Partenaires & collections associés à EuDML

Gallica-Math 1 revue, 98 livres

GDZ-Math 42 revues, 1531 monographies, 294 ouvrages en plusieurs tomes

RusDML 11 revues

The European Digital Mathematics Library

Collections à venir ?

Futurs associés d'EuDML ?

BDIM	1 revue *en cours*
eLib SANU	9 revues
eLib MATF	150 livres, 354 Doct. Th.
SwissDML	4 revues
TEL	1996 Ph. D. Th.
Gallica	800 livres
IMU	Actes des conférences internationales

The European Digital Mathematics Library

Détenteurs des droits

Domaine public quelques revues, la plupart des livres

Secteur public 50 Universités, centres de recherche, instituts, académies

Fondations Compositio Mathematica, quelques associations

Sociétés 20 sociétés mathématiques

Éditeurs

Birkhäuser	5 revues (GDZ)
EDPS	7 revues (5 mise à jour in NUMDAM)
Elsevier	5 revues, 1 mise à jour (NUMDAM)
de Gruyter	2 revues (GDZ)
Heldermann	6 revues (5 mise à jour in ELibM)
Hindawi	12 revues (up-to-date in ELibM)
Noordhoff	1 revue (NUMDAM)
AK Peters	1 revue (ELibM)
Springer	2 periodicals (NUMDAM, 1 revue mise à jour jusqu'en 2007)
	9 revues (GDZ)

Les contenus en chiffres

Définitions

Sélection Pyramidale : le projet sélectionne les institutions, qui sélectionnent les collections.

EuDML *item* Un texte éligible pour EuDML est un texte complet, traitant de mathématiques, produit par un processus éditorial, existant sous forme numérique.
Un livre, mémoire, article, comunication dans un livre édité…
Concrètement c'est la donnée d'une paire (texte intégral numérique [PDF], métadonnées [XML]) archivée par l'une des institutions partenaires

Synthèse **235 000 textes** (dont 1000 livres, 300 thèses) dans **13 collections** provenant de **11 institutions**

Les contenus en chiffres

Un babel de métadonnées

Critères qualitatifs

Basique Informations assez précises pour identifier un texte de façon unique (dans la littérature mondiale). Rend la recherche et le feuilletage possible.
Type*, référence bibliographique détaillée*, auteurs*, titre*, résumé, mot-clés, langue*, identifiant unique*, PURL*.

Avancé Tout ce qui est au-delà de basique et pertinent pour EuDML.
Traductions des titres, résumés ; translitérations des écritures non latines ; numéros MR/ZM ; MathML ; références citées...

Sousbasique Détails insuffisants
Référence bibliographique comme chaîne de caractères, pas de numéro de tome pour un journal, métadonnées sur le volume d'un journal ne décrivant pas chaque article...

Synthèse Les standards des collections sont à
55% avancés, 44% basiques, 1% sousbasiques
50% cherchables en plein texte

Les contenus en chiffres

Un babel de formats de métadonnées

Si la qualité des métadonnées est très variable, la forme aussi !

SQL	Bases de données sans format d'échange XML : DML-E, HDML
DTD maison	MathDoc, FIZ, IST
DTD standard	DC, Dspace, minidml, METS, NLM

Proposition en cours d'étude :
Utiliser **NLM Journal Archiving and Interchange Tag Suite**
pour l'échange et le stockage des métadonnées EuDML

EuDML metadata schema

NLM Journal Archiving and Interchange Tag Suite

- En production à grande échelle (EDPS, PubMed Central, JSTOR)
- En cours de normalisation (NISO)
- Supporte MathML (et *alternatives*)
- Précise et flexible
- Permet de traiter les articles de revues, les livres, et les collections de livres
- Peut transporter le plein-texte

Moyen de communication entre le système central et les partenaires : OAI-PMH

The European Digital Mathematics Library

Stratégie

- Renoncer à la vision initiale globale, pyramidale et centralisée
- Commencer avec un petit groupe de partenaires motivés, qui représentent une diversité suffisante pour anticiper l'avenir
- Proposer un modèle économique assez souple pour permettre à toutes les catégories d'éditeurs de s'y retrouver (créneau mobile)
- Mais ne référencer que les textes qui sont archivés par une institution fiable !

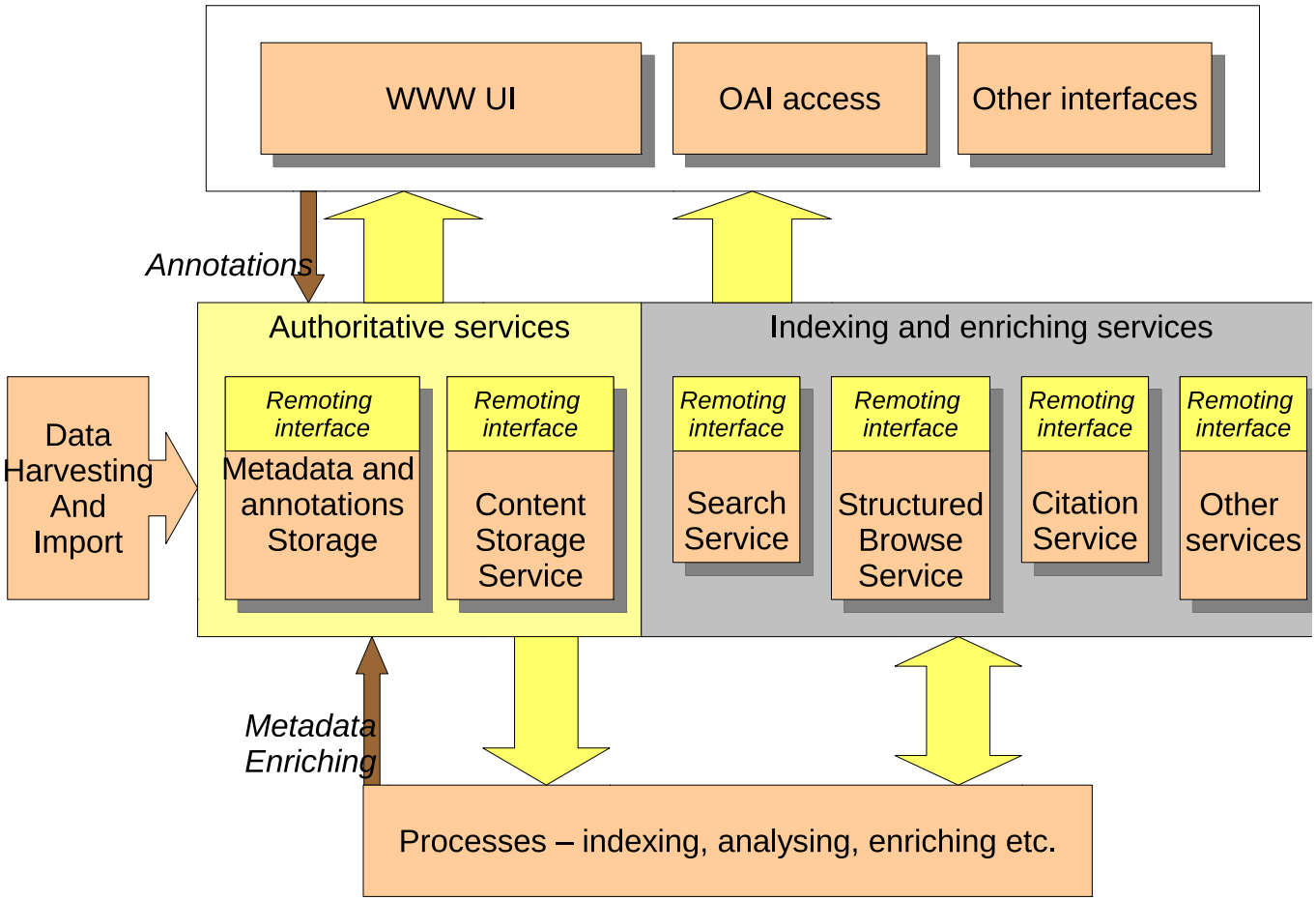
Projets WDML

Approche technologique

- Trouver un noyau commun de métadonnées qui permette des fonctionnalités riches sans fixer la barrière technologique trop haut
- Développer des chaînes de création automatique de métadonnées pour les collections trop dépouillées (GDZ, DML-CZ)
- Croiser toutes les informations disponibles sur un texte donné
- Utiliser le contenu mathématique pour contourner le multilinguisme des collections (et le monolinguisme des métadonnées)
- Dépasser les limites des formats graphiques : founir un accès au contenu scientifique plutôt qu'à des images

The European Digital Mathematics Library

Architecture



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*.