EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY

ICTPSP

DELIVERABLE

Project Acronym: EuDML

Grant Agreement number: 250503

Project Title: The European Digital Mathematics Library

Deliverable 5.2 - The EuDML Search and Browsing Service

Revision: 1.0

Authors:

Aleksander Nowiński (ICM)

Tomasz Rosiek (ICM)

Michał Politowski (ICM)

<table>
<tr><td colspan="3">Project co-funded by the European Commission within the ICT Policy Support Programme</td></tr>
<tr><td colspan="3">Dissemination Level</td></tr>
<tr><td>P</td><td>Public</td><td></td></tr>
<tr><td>C</td><td>Confidential, only for members of the consortium and the Commission Services</td><td>PP</td></tr>
</table>

Tuesday, 1 March 2011

Revision History

Revision	Date	Author	Organisation	Description
1.0	25/02/2011	A.N	ICM	First consolidated version
1.0c	27.02.2011	Wojtek	ICM	copyedit
1.0cc	28/02/11	M.P.	ICM	copyedit
1.1	28/02/11	A. N.	ICM	Component names updated

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Document Index

1. Introduction

The purpose of this document is to provide a description of the first live demonstration of the EuDML software. The demo includes the basic orchestration of the EuDML core services, and demonstrates the basic functionality of the EuDML digital library.

The demonstration site was presented to the consortium members on the EuDML plenary meeting, on 31 of January 2011 in Madrid. The demo system is composed of the core system services (storage, search and browse services) and an adapted version of YADDA web user interface. This web user interface will be a base for a future development in WP6. The system operates on metadata in NLM format, provided by WP3 and offers basic digital library functionality, including search, browse and presentation of the metadata.

1.1. Accessing the demo site

The demo site may be accessed and tested on the web, at the following URL:

http://demo.eudml.eu/demo/

It is recommended to use Firefox 3.+ browser, as other browsers are not fully tested for the demo purpose. There are no access restrictions, it is open to the world.

2. Technical overview

The demo system architecture is presented on Diagram 1:Demo system architecture and data flow

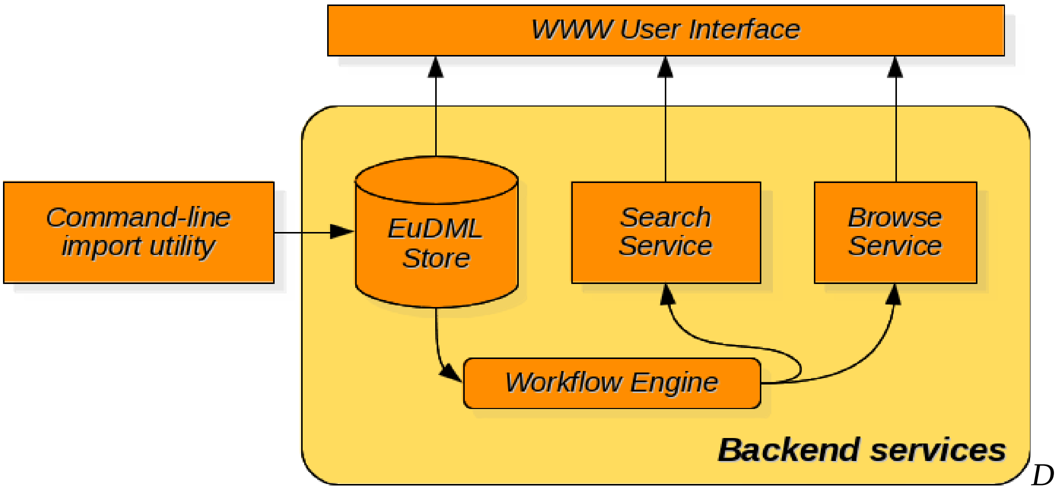


Diagram 1: Demo system architecture and data flow

The arrows in the diagram represent the data flow in the system. The demo system is composed of two parts: the document repository (Backend services) and the WWW User Interface (web UI). Currently the web UI accesses the repository with HttpInvoke protocol. The repository provides basic services: browsing, searching and data access. It has also an engine to process data from the storage and use it to build search service indexes and browse component relations.

This architecture will be a base for a whole solution, and when new services will be added into the EuDML ecosystem, they will enrich and extend the backend component.

2.1. Storage Service

All metadata files in the system are stored in the EuDML Storage service. This service is currently implemented on top of the YADDA Metadata Directory and YADDA Archive services. On the system level, files are stored in a Postgres SQL database (metadata and smaller files) and in a filesystem (larger files). Service stores data in a form of compound records. For each entity (article, journal...), a record exists, containing:

- original (source) metadata (possible multiple files)
- EuDML metadata (NLM) converted and normalized from source, known as 'Base NLM'
- optionally enhanced EuDML metadata (result of enhancement process)
- content files downloaded from provider
- plain text files for full text search (not yet indexed)

Within EuDML storage, the objects are identified with EuDML identifiers in URI form. Each record is composed of several parts of two kinds: metadata of the object and contents of the object (including plaintext for indexing etc.). Each part has its internal identifier (partId) inside the record, so a pair of EuDML record identifier and partId locates specific object in the store. The storage is the main source of all data in the system. It is used whenever data is presented to the user, both as a record page and as a snippet in a search result page. It is also used by the processor component to create data for the indexes of the search and browse services.

2.2. Search service

The search service has a well defined role in this presentation, which is providing expected search capabilities to the system. The current version of the system uses old, but well-tested YADDA Search Service, based on the Lucene search engine. Currently, a new service, which is based on SOLR technology is already prepared, but has not been tested sufficiently. Therefore it was decided that it will not be integrated into this demo stage, as it would bring unnecessary risk to integrate a number of new components at once. Together with the SOLR engine, a mathematical indexing and formula search developed by MU will be integrated later on.

2.3. Browse service

The browse service is responsible for browsing large, organised collections of data. In the demo it is responsible for browsing the list of journals and is also used to render a publication tree on a journal page. As the interface will be enriched, more browse applications will appear at later stages (such as browsing topics, authors etc.).

The browse service is technically a service which allows navigation over 'relations'. A relation has a form of a single database table. The service is backed by a Postgres SQL database and it is filled with tuples derived from data from the EuDML store by the workflow engine component. The service is optimised for paging (which is appropriate for web presentation) and is stateless, so users may have multiple browsing sessions opened without problems of timeouts or limits of the server resources.

2.4. Workflow Engine

The workflow engine is not visible directly. It is the component responsible for all iterative type activities within the system, including indexing, building browse relations and application of enhancers to data. It is a new generation of the workflow service from the YADDA services suite, which was developed for EuDML and is used here for the first time. It is based on Spring Integration framework, which allows to process entries in asynchronous workflows. Within the workflow engine multiple processes are defined, separate for basic data processing (like indexing) and for the enhancement processes.

The performance of about 30 elements per second has been achieved for the indexing and relation building flow, which results in the total processing time of the demo site below 30 minutes. As this operation is required only once after import (and sometimes on update), its performance is currently considered to be sufficient but will be re-evaluated for the future production service.

Additionally, an enhancement process was defined for testing purposes, downloading PDF content files from the partners' sites and attempting to extract the citations from them. However, since all metadata to be used in the demo are already of better quality than could be possibly achieved with these automatic enhancers (as manually added citations are already present), it is not used in the presentation, as the automatic enhancers are obviously expected to enhance the data, not to to degrade it. The purpose of setting up the process was to prove, that all enhancement techniques developed in WP7 may be added into the system in a near future.

2.5. User Interface

The User Interface of the demo is currently a version of a new YADDA web UI, yaddaweb-lite, adapted to the EuDML demo specific requirements. It is a modular web interface based on Spring MVC and Tiles technology, and will be a base for building the EuDML user interface in the future, within WP6. Currently used interface skin (which is customized by CSS) is adopted CSS from original UI, with project colour scheme and logo added. It expected to be used only for this demo purposes, and subsequently replaced by the skin to be developed in WP6.

3. User Interface overview

3.1. General

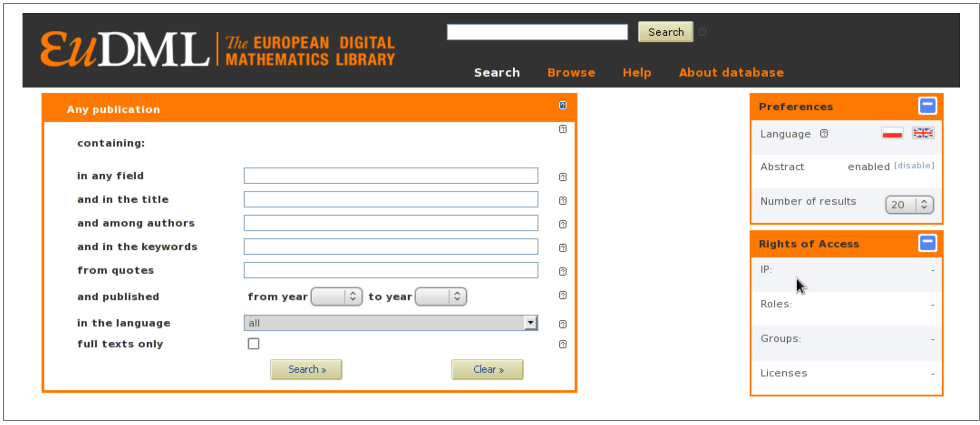
The user interface of the demo has three main parts: the main area (occupying the central and the left part of the page), the top navigation bar and side panels. In the main area, information is presented to the user. The top bar serves to navigate easily between main parts of the application, perform immediate search or get help. The help set is not available, as the interface is subject to change. The 'About database' button redirects to the project web page.

On the right side there are two panels. The first one is the panel to adjust the basic preferences (number of entries on search or browse page, language and default abstract visibility). The second one is responsible for current user license information and is currently not used.

The user interface, what is very important nowadays, supports multi-tab and multi-window browsing with no side effects. The URLs in the demo are not yet in the final form, and especially object URLs are subject to change, and should by no means be considered final.

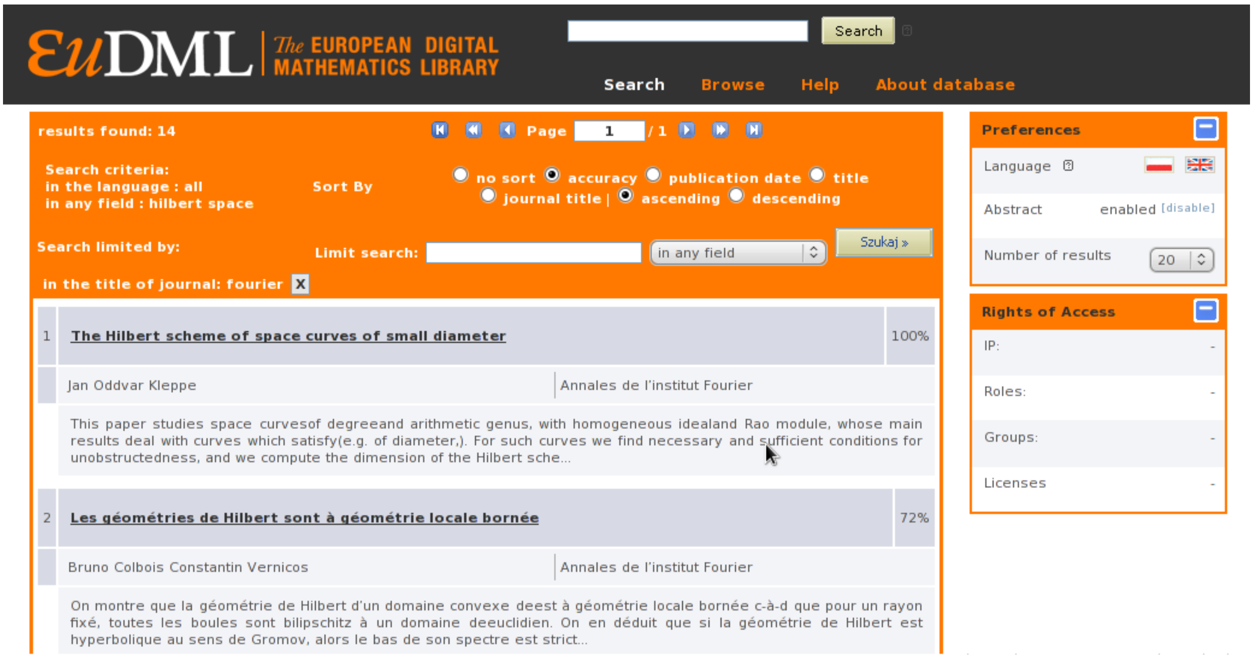
3.2. Home screen

The home screen of the demo is a search screen, allowing to start searching database easily. This is displayed in screenshot 1. This is the page where a user is directed, when he arrives at the EuDML page for the first time. As D6.1 demonstrates, the users are usually most interested in searching over the database collection, so it is a natural starting point for a visitor. Various fields may be used to perform the search, including the title, author name, keywords etc. All the query terms are naturally combined with boolean AND operator.



Screenshot 1: Home screen

When a user submits a search query, search results are displayed, together with brief snippets and information about the accuracy of the results. The user can then refine the search by adding more filter queries and sort results accordingly to his needs. The number of search results presented on the screen may be adjusted in the upper right panel. Search highlighting is not yet supported, nor are other more complex search features, which will be available when the SOLR search service is integrated. An example of search results is presented in screenshot 2. They are search results for the term 'Hilbert space', then refined by the 'fourier' term required in a journal title.



results found: 14

Search criteria:
in the language : all
in any field : hilbert space

Search limited by:
in the title of journal: fourier X

1 The Hilbert scheme of space curves of small diameter 100%

Jan Oddvar Kleppe | Annales de l'institut Fourier

This paper studies space curvesof degreeand arithmetic genus, with homogeneous idealand Rao module, whose main results deal with curves which satisfy(e.g. of diameter,). For such curves we find necessary and sufficient conditions for unobstructedness, and we compute the dimension of the Hilbert sche...

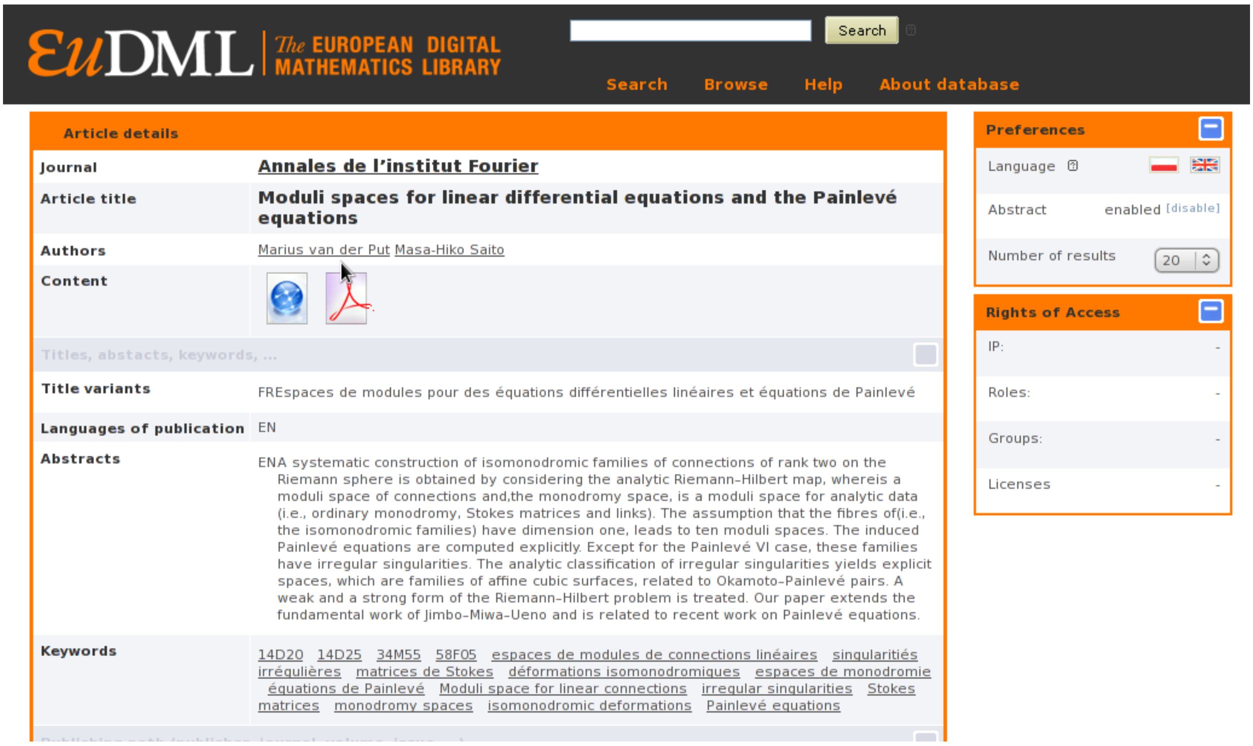
2 Les géométries de Hilbert sont à géométrie locale bornée 72%

Bruno Colbois Constantin Vernicos | Annales de l'institut Fourier

On montre que la géométrie de Hilbert d'un domaine convexe deest à géométrie locale bornée c-à-d que pour un rayon fixé, toutes les boules sont bilipschitz à un domaine deeuclidien. On en déduit que si la géométrie de Hilbert est hyperbolique au sens de Gromov, alors le bas de son spectre est strict...

Screenshot 2: Refined search results

The user may open the article's record, to view its metadata and optionally download it from the original source. This screen is visible in screenshot 5. The two icons in the content section represent article at the provider site (the article page and a direct link to the article's contents). Copies of the articles' contents are stored in the EuDML database for the internal enhancement processes, but are never served to the users directly from EuDML database. Some metadata are still not presented in the best possible way, for example keywords usually are mixed with categories or original identifiers are not visible despite being available within data.



Article details

Journal	Annales de l'institut Fourier
Article title	Moduli spaces for linear differential equations and the Painlevé equations
Authors	Marius van der Put Masa-Hiko Saito
Content	
Titles, abstacts, keywords, ...	
Title variants	FREspaces de modules pour des équations différentielles linéaires et équations de Painlevé
Languages of publication	EN
Abstracts	ENA systematic construction of isomonodromic families of connections of rank two on the Riemann sphere is obtained by considering the analytic Riemann-Hilbert map, whereis a moduli space of connections and,the monodromy space, is a moduli space for analytic data (i.e., ordinary monodromy, Stokes matrices and links). The assumption that the fibres of(i.e., the isomonodromic families) have dimension one, leads to ten moduli spaces. The induced Painlevé equations are computed explicitly. Except for the Painlevé VI case, these families have irregular singularities. The analytic classification of irregular singularities yields explicit spaces, which are families of affine cubic surfaces, related to Okamoto-Painlevé pairs. A weak and a strong form of the Riemann-Hilbert problem is treated. Our paper extends the fundamental work of Jimbo-Miwa-Ueno and is related to recent work on Painlevé equations.
Keywords	14D20 14D25 34M55 58F05 espaces de modules de connections linéaires singularitiés irréqulières matrices de Stokes déformations isomonodromiques espaces de monodromie équations de Painlevé Moduli space for linear connections irregular singularities Stokes matrices monodromy spaces isomonodromic deformations Painlevé equations

Preferences

Language

Abstract enabled [disable]

Number of results 20

Rights of Access

IP: -

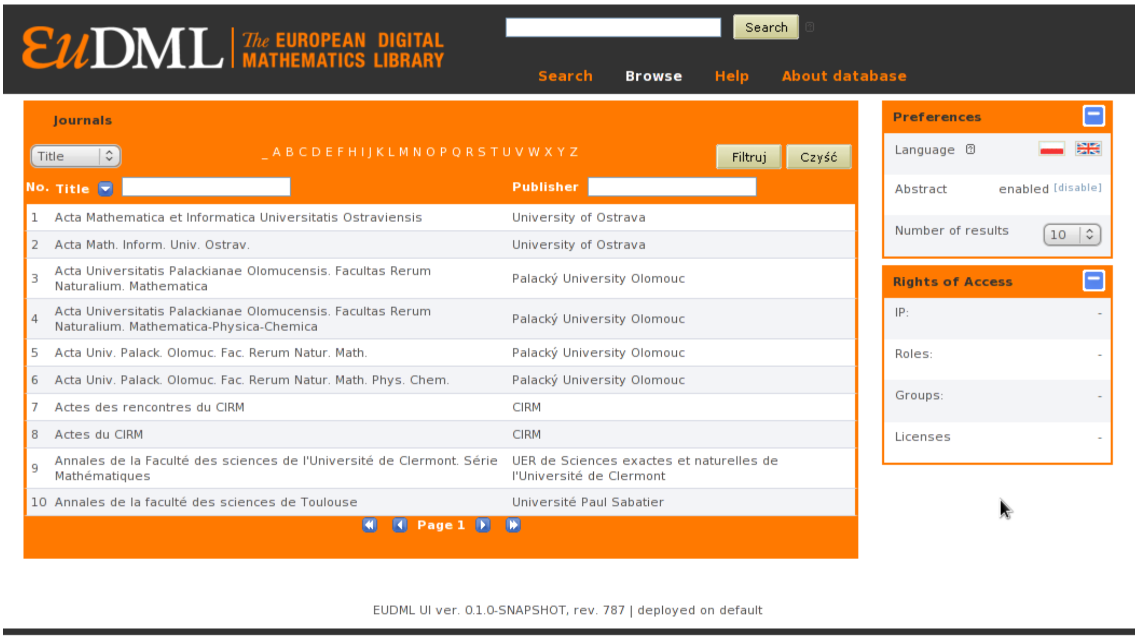
Roles: -

Groups: -

Licenses -

Screenshot 3: Article details

To demonstrate the browse capabilities a single browse view is used. It may be accessed from top menu 'browse' button, which brings a browse journal view, as displayed in screenshot 4. On the screen, the user may see the list of journals and navigate to a particular journal screen. All the journals present in database are listed in the browse journal screen. The user may apply a filter to limit the browse scope. Note that all journal titles (including title versions, like translations or shortcuts) appear on the list, so a user knowing only one version still will be able locate the desired journal.


EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY
Search
Search
Browse
Help
About database
Title
_ A B C D E F H I J K L M N O P Q R S T U V W X Y Z
Filtruj
Czyść
Page 1

Journals

No.	Title	Publisher
1	Acta Mathematica et Informatica Universitatis Ostraviensis	University of Ostrava
2	Acta Math. Inform. Univ. Ostrav.	University of Ostrava
3	Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica	Palacký University Olomouc
4	Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica-Physica-Chemica	Palacký University Olomouc
5	Acta Univ. Palack. Olomuc. Fac. Rerum Natur. Math.	Palacký University Olomouc
6	Acta Univ. Palack. Olomuc. Fac. Rerum Natur. Math. Phys. Chem.	Palacký University Olomouc
7	Actes des rencontres du CIRM	CIRM
8	Actes du CIRM	CIRM
9	Annales de la Faculté des sciences de l'Université de Clermont. Série Mathématiques	UER de Sciences exactes et naturelles de l'Université de Clermont
10	Annales de la faculté des sciences de Toulouse	Université Paul Sabatier

Preferences

Language

Abstract enabled [disable]

Number of results 10

Rights of Access

IP: -

Roles: -

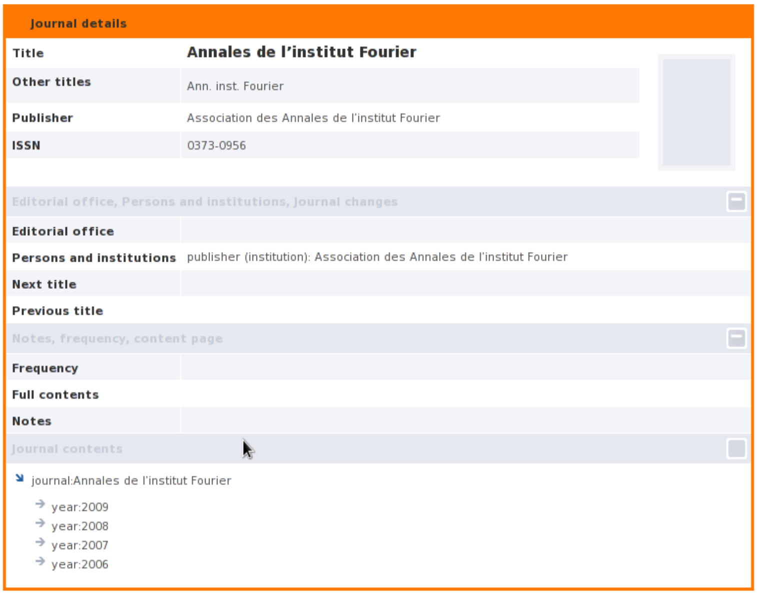
Groups: -

Licenses -

EUDML UI ver. 0.1.0-SNAPSHOT, rev. 787 | deployed on default

Screenshot 4: Browse journals screen

Selecting specific journal brings the journal page where all details of the journal are presented. It may be seen in screenshot 5. It contains information about the journal, like ISSN number(s), publishers and all other known information. It is possible to add a miniature of the journal cover, which improves the overall user perception.



Journal details

Title	Annales de l'institut Fourier
Other titles	Ann. inst. Fourier
Publisher	Association des Annales de l'institut Fourier
ISSN	0373-0956

Editorial office, Persons and institutions, Journal changes

Editorial office	
Persons and institutions	publisher (institution): Association des Annales de l'institut Fourier
Next title	
Previous title	

Notes, frequency, content page

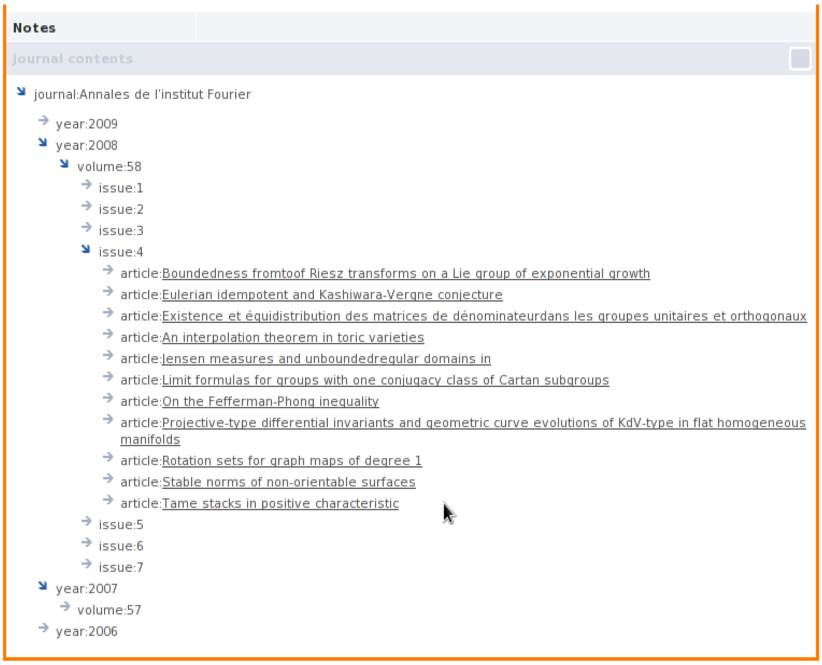
Frequency	
Full contents	
Notes	

Journal contents

- journal:Annales de l'institut Fourier
  - year:2009
  - year:2008
  - year:2007
  - year:2006

Screenshot 5: Journal details view

A special feature of the journal screen is a tree of the publication structure, which is also driven by information stored in the browse service. This tree is presented in screenshot 6. It allows to browse structured article collection easily, without adding unnecessary views.



Notes

Journal contents

- journal:Annales de l'institut Fourier
  - year:2009
  - year:2008
    - volume:58
      - issue:1
      - issue:2
      - issue:3
      - issue:4
        - article:Boundedness fromtoof Riesz transforms on a Lie group of exponential growth
        - article:Eulerian idempotent and Kashiwara-Vergne conjecture
        - article:Existence et équidistribution des matrices de dénominateurdans les groupes unitaires et orthogonaux
        - article:An interpolation theorem in toric varieties
        - article:Jensen measures and unboundedregular domains in
        - article:Limit formulas for groups with one conjugacy class of Cartan subgroups
        - article:On the Fefferman-Phong inequality
        - article:Projective-type differential invariants and geometric curve evolutions of KdV-type in flat homogeneous manifolds
        - article:Rotation sets for graph maps of degree 1
        - article:Stable norms of non-orientable surfaces
        - article:Tame stacks in positive characteristic
      - issue:5
      - issue:6
      - issue:7
  - year:2007
    - volume:57
  - year:2006

Screenshot 6: Journal contents tree

This concludes the brief overview of the user interface of the demo application.

4. Data in the demo site

The data in the demo site is only a part of all the data already prepared and analysed in the WP3. The demo contains only journal articles, no books nor book series were imported, and no dedicated book view was prepared.

4.1. Used data formats

All the data imported into the demo site was either in NLM format natively, or was converted off-line in WP3. As the EuDML store preserves not only the converted metadata, but the source metadata as well, the metadata was imported as the source part, and a trivial conversion (copy) was applied to obtain the metadata used. This is subject to change, as WP5 effort will lead to manage all data harvesting and conversion processes with the REPOX component.

4.2. Data Sources

The demo contains approximately 55,000 documents. This is only about one-fifth of all documents expected to be stored in the EuDML initially, but it gives a good view on the current performance and possible problems.

The data was collected from

- CEDRAM (The center for diffusion of academic mathematical journals),
- DML-CZ (Czech Digital Mathematics Library),
- DML-E (Spanish Digital Mathematics Library),
- ElibM (The Electronic Library of Mathematics),
- GDZ (Göttinger Digitalisierungszentrum),
- NUMDAM (Numérisation de documents anciens mathématiques),
- the *Portugaliae Mathematica* journal,

varies in quality, and has different time extent. This diversity allowed to already detect a number of issues, like the need of proper handling of journals with no ISSN number or unknown publisher.

5. Next steps

The demo is an important step, as it required a close and effective cooperation between different workpackages. Not all the available services were integrated for the demonstration purposes, but the integration work has started and is underway. The next steps after the demonstration will be done in various aspects, the most important being:

- integration of the metadata enhancers and starting the work on the real documents from the collection (WP7)
- integration of REPOX metadata harvesting and transformation engine with EuDML store (WP5)
- development of the user interface in its eventually designed form (WP6)
- integration of the other resources – books etc. (WP3)

These tasks are expected to be the main direction of the development until the next demo on month 18 of the project.