COMPETITIVENESS AND INNOVATION FRAMEWORK PROGRAMME

THEME [CIP-ICT-PSP.2009.2.4]
[Open access to scientific information]

Grant agreement for: CIP-Pilot actions

Annex I - "Description of Work"

Project acronym: EuDML

Project full title: " The European Digital Mathematics Library "

Grant agreement no: 250503

Date of preparation of Annex I (latest version): 2009-12-16

Date of approval of Annex I by Commission: 2009-12-16

A1:
Our project

Project Number [1]	250503	Project Acronym [2]	EuDML

One form per project

General information			
Project title [3]	The European Digital Mathematics Library		
Starting date [4]	01/02/2010		
Duration in months [5]	36		
Call (part) identifier [6]	CIP-ICT-PSP-2009-3		
Objective most relevant to your topic [7]	CIP-ICT-PSP.2009.2.4: Open access to scientific information		
Free keywords [8]	mathematical research, distributed archiving, one-stop easy open access, innovative semantic interlinking of mathematical content, accessible mathematical content		

Abstract [9]

In the light of mathematicians reliance on their discipline's rich published heritage and the key role of mathematics in enabling other scientific disciplines, the European Digital Mathematics Library strives to make the significant corpus of mathematics scholarship published in Europe available online, in the form of an authoritative and enduring digital collection, developed and curated by a network of institutions.
National efforts have led to the digitisation of large quantities of mathematical literature, primarily by partners in this project. Publishers produce new material that needs to be archived safely over the long term, made more visible, usable, and interoperable with the legacy corpus on which it settles. In EuDML, these partners will join together with leading technology providers in constructing the Europe-wide interconnections between their collections to create a document network as integrated and trans-national as the discipline of mathematics itself. They will future-proof their work by providing the organisational and technical infrastructure to accommodate new collections and mathematically rich metadata formats, and will work towards truly open access for the whole European Community to this foundational resource, thereby retaining Europe's leadership in the provision, accessibility and exploitation of electronic mathematical content.
EuDML will design and build a collaborative digital library service that will collate the currently distributed content by the diversity of providers. This will be achieved by implementing a single access platform for heterogeneous and multilingual collections. The network of documents will be constructed by merging and augmenting the information available about each document from each collection, and matching documents and references across the entire combined library. In return for this added value, the rights holders agree to a moving wall policy to secure eventual open access to their full texts.

Approved by EC on 2009-12-16

A2:
List of Beneficiaries

Project Number [1]	250503	Project Acronym [2]	EuDML

List of Beneficiaries

No	Name	Short name	Country	Project entry month[10]	Project exit month
1	INSTITUTO SUPERIOR TECNICO	IST	Portugal	1	36
2	UNIVERSITE JOSEPH FOURIER GRENOBLE 1	UJF/CMD	France	1	36
3	THE UNIVERSITY OF BIRMINGHAM	UB	United Kingdom	1	36
4	FACHINFORMATIONSZENTRUM KARLSRUHE GESELLSCHAFT FUR WISSENSCHAFTLICH-TECHNISCHE INFORMATION GMBH	FIZ	Germany	1	36
5	Masarykova univerzita	MU	Czech Republic	1	36
6	UNIWERSYTET WARSZAWSKI	ICM	Poland	1	36
7	AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS	CSIC	Spain	1	36
8	EDITION DIFFUSION PRESSE SCIENCES	EDPS	France	1	36
9	UNIVERSIDADE DE SANTIAGO DE COMPOSTELA	USC	Spain	1	36
10	INSTITUTE OF MATHEMATICS AND INFORMATICS OF THE BULGARIAN ACADEMY OF SCIENCE	IMI-BAS	Bulgaria	1	36
11	MATEMATICKY USTAV AV CR V.V.I.	IMAS	Czech Republic	1	36
12	IONIAN UNIVERSITY	IU	Greece	1	36
13	MADE MEDIA LTD	MML	United Kingdom	1	36
14	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	CNRS/CMD	France	1	36

Approved by EC on 2009-12-16

A3:
Budget breakdown

Project Number [1]	250503	Project Acronym [2]	EuDML

One Form per Project

Participant number in this project	Participant short name	Personnel costs	Sub contracting	Other direct costs	Indirect Costs		Total costs	Max EC Contribution	Requested EC contribution
					Cost model (a)	Value			
1	IST	368,267.00	0.00	18,300.00	SFR	110,480.00	497,047.00	248,523.00	248,523.00
2	UJF/CMD	262,474.00	0.00	10,500.00	SFR	78,742.00	351,716.00	175,858.00	175,858.00
3	UB	275,508.00	0.00	9,300.00	SFR	82,652.00	367,460.00	183,730.00	183,730.00
4	FIZ	369,084.00	0.00	9,300.00	SFR	110,724.00	489,108.00	244,554.00	244,554.00
5	MU	234,638.00	0.00	9,300.00	SFR	70,390.00	314,328.00	157,164.00	157,164.00
6	ICM	365,200.00	0.00	13,075.00	SFR	109,560.00	487,835.00	243,917.00	243,917.00
7	CSIC	36,368.00	0.00	9,300.00	SFR	10,910.00	56,578.00	28,289.00	28,289.00
8	EDPS	44,000.00	0.00	4,200.00	SFR	13,200.00	61,400.00	30,700.00	30,700.00
9	USC	53,808.00	0.00	9,300.00	SFR	16,142.00	79,250.00	39,625.00	39,625.00
10	IMI-BAS	68,200.00	0.00	9,300.00	SFR	20,460.00	97,960.00	48,980.00	48,980.00
11	IMAS	65,656.00	0.00	9,300.00	SFR	19,696.00	94,652.00	47,326.00	47,326.00
12	IU	58,500.00	0.00	9,300.00	SFR	17,550.00	85,350.00	42,675.00	42,675.00
13	MML	110,154.00	0.00	3,300.00	SFR	33,046.00	146,500.00	73,250.00	73,250.00
14	CNRS/CMD	52,495.00	0.00	2,575.00		15,748.00	70,818.00	35,409.00	35,409.00
TOTAL		2,364,352.00	0.00	126,350.00		709,300.00	3,200,002.00	1,600,000.00	1,600,000.00

(a) AIC : Actual indirect costs , SFR : Standard flat rate

Approved by EC on 2009-12-16

COMPETITIVENESS AND INNOVATION FRAMEWORK PROGRAMME
ICT Policy Support Programme (ICT PSP)

Digital Libraries

ICT PSP call identifier : CIP-ICT-PSP-2009-3
ICT PSP Theme/objective identifier: 2.4 Open access to scientific information

Grant agreement for: PILOT TYPE B

Annex I - "Description of Work"

Project acronym: EuDML
Project full title: The European Digital Mathematics Library
Grant agreement no.: 250503

Date of preparation of Annex I (latest version): 23/11/2009 23:24
Date of approval of Annex I by Commission: *(to be completed by Commission*)

PROJECT PROFILE

Information on the proposed service/solution

Description of the issue and proposed service/solution

Mathematics is a basic science for a wide range of scientific disciplines. There are fundamental applications of mathematical knowledge in almost every area of the natural and social sciences and the humanities. New technological developments and innovations are often based on mathematical results years or decades old. While mathematics is probably the most affordable science in the sense that it doesn't need expensive research infrastructure in order to be developed at the highest levels, it is entirely dependant upon accessibility to the published, validated literature, which sets standards and records trusted results. In many European locations, while a high level of mathematical training is available, it is difficult for scholars and their students to compete with more fortunate universities because the libraries have gaps in their collections. Making our mathematical heritage easily available from everywhere will considerably augment its usability, and increase the competitiveness and productivity of scholars from all over Europe and the emerging countries.

The availability of electronic publications and retro-digitised material leads to significant improvements of the conditions for research. Creating a common infrastructure for searching and interacting within the deeply interlinked network of digital mathematical content will make mathematics readily available for all users of this resource, which in turn will be much more useful.

The publishers, whatever their business model is, expect to have a safe archival back-end so that they don't have to maintain their back catalogues indefinitely, and thus agree to transfer their content and to licence eventual open access to it. The projected service of a distributed collection of digital content in Europe may serve as a model for similar activities in other disciplines, on both technical and organisational sides.

Finally, the project satisfies the demand for reliable and long term availability of mathematical research output. Cared for locally, at participants' sites, and visible and usable globally.

Target users and their needs

The foremost target users are students and scholars whose work relies on the validity and availability of previous original mathematical knowledge. Working mathematicians have, of course, a prominent place among this group, as they produce and validate this knowledge, relying heavily on the pre-existing literature. They need fast, reliable and permanent access to the literature in their working environment (e.g. links to original full texts for the references they need, easy and fast discovery of references related to a first approximate match). Using a single well-organised resource with such a critical mass will insure maximum visibility, hence timely access from any location in the world. Together with the network of institutional repositories for instant dissemination of ongoing research, this will be the foundation of the mathematics-based emerging eScience.

Working scientists, scholars and engineers often produce new knowledge or products based on mathematical results that can be very old: such users need reliable access and innovative discovery strategies as they are not necessarily in an academic environment that provides dedicated tools and help from information professionals.

Librarians and Information managers will also be important users. The EuDML service will save them time, money, and resources in tracking down and obtaining mathematical documents and in managing their own collections of mathematical content.

Other privileged users are professors and their students, who will enjoy a long-standing reliable system with sophisticated tools available for search and exploration, with permanent URIs.

But we should also mention the science-friendly public at large, who will enjoy accessing Europe's extraordinary cultural heritage in the field of mathematics, reading the great achievements of the past in the author's own words through expected interoperability with Europeana.

Usage

The starting point for a user looking for a mathematical reference is a familiar search engine. Those relevant for mathematical content will be made aware of EuDML by pushing metadata into their

harvesting or indexing systems.
Once users have reached the EuDML website, which might be directly if a persistent EuDML URI was followed, they can navigate by browsing the collections, following links to related items (same author, text citing or cited by the given one, similar subject, similar mathematical content, etc.). They could also be guided by tips or additional keywords left by other users, and leave their own annotations as well. User profiles might also be derived in order to rank the query results depending on the user's mathematical background. Linked resources, such as Zentralblatt by partner FIZ can also be used to explore further with other methods, bringing the user back when a reference to an interesting EuDML item is finally found.
All item pages will provide a link to the associated full text, 95% of which point to a partner's repository serving the file under open access. The remaining 5% will be hosted at their publisher's platform, possibly charging for access.

Technology

The core of the system is built through metadata harvesting from the content providers (OAI-PMH with a dedicated fine-grain schema), for aggregation in a large database. The main component for querying the content will be web services on top of the YADDA system. A collection of open source software developed inside the consortium will be deployed to enrich the aggregated metadata, e.g. for matching and linking related items, create a lively interactive web site, perform structured mathematical OCR. All metadata available to or generated by the project, including item's full texts, will be made searchable. As much of the full texts as possible will be converted to structured XML with MathML representation of formulas and English metadata in order to pave the way to more efficient and user-friendly mathematical document retrieval.

Content

The content consists of published texts holding mathematical knowledge that has been validated through a scientific editorial process, so that they can serve for further reference in future works based on this mathematical knowledge. However, we shall favour integrating comprehensive collections over selecting each item: the above criterium has to be understood in a broad sense. The content has been digitised from paper by partners, provided to partners as born digital content from its publisher, or directly provided by publishers partnering in the project. It is typically open access within 5 years of its publication date, which results in less than 5% of the whole being subject to any form of restricted access, and that only temporarily.

Sustainability

The project aims at fostering European integration of existing services, and implementing a quantum jump in usability and deployment of the foreseen innovations to a wide corpus which is already produced and maintained without extra needs for funding. All the content holders have the commitment to keep their collections alive and used to their best potential. Once the pilot system is up and running, it will be their natural mission to perform the routine maintenance. Most of our partners are university or public bodies dedicated to providing quality scientific information to researchers and students. Those who are commercial expect a net return on investment thanks to the visibility and good image obtained through partnering in this project. Users are not expected to pay extra costs during or after the completion of Community funding.

Ownership

The actual content, and hardware, is owned by those bringing it into the project. The project doesn't change this fact. The consortium agreement will rule the way the IPR on software and full texts are shared and exploited among the partners. The rationale is that during the project, all relevant files and technologies will be freely exchangeable among consortium members. The service, as an abstract entity, will be owned by the consortium, which will be turned into a persistent body. The components and data generated in the course of the project will be freely usable by the partners who will be able to exploit them after the project end.

Section B1. Relevance

B1.1. Project objectives

In the light of mathematicians' reliance on their discipline's rich published heritage and the key role of mathematics in enabling other scientific disciplines, the European Digital Mathematics Library strives to make the significant corpus of mathematics scholarship published in Europe available online, in the form of an authoritative and enduring digital collection, developed and curated by a network of institutions. Whether a researcher needs to follow a subtle pyramid of reasoning through a chain of related articles, an engineer needs to find results related to a particular concept, or a school project studies the history of a specific mathematical issue, there is a common need for an integrated, interconnected gateway to the body of preserved mathematical literature.

At a national level, digitisation programmes have devoted enormous effort over the last decade to construct impressive digital repositories of the mathematical literature published in Europe. Moreover, all recent mathematical texts have been published digitally, which asserts that a huge part of the core mathematical knowledge produced or published in Europe now exists in digital form. However, this corpus is not as accessible and usable at it could or should be. This is mainly due to a lack of coordination among stakeholders, so that existing national repositories are not interoperable or interlinked and publishers do not routinely transfer their digital production to libraries for securing an alternative long term archiving and access provision to their output.

References in an article in one repository cannot usually be followed to find the target article in a different repository. Search facilities are of varying capabilities and many items are lacking the metadata essential to be able to find and exploit them. For instance, as most of the existing metadata is derived from the actual text printed in an item (either through keyboarding, OCR, or extraction from digital production source files), and as a large part of the historical corpus is not written in English, many very important items are left invisible to current queries. Ironically, the mathematical formalism can be viewed as a truly international, natural language-neutral idiom, but it is excluded from most metadata schemes and not exploited as it could be.

In EuDML, the partners, many of whom were the primary agents in their respective national digitisation programmes, and four of whom act as publishers of new material, will join together with leading technology providers in constructing the Europe-wide interconnections between their collections to create a document network as integrated and trans-national as the discipline of mathematics itself. More than that, they will future-proof their work by providing the organisational and technical infrastructure to accommodate new collections and mathematically rich metadata formats, and will work towards truly open access for the whole European Community to this hugely important resource, thereby retaining Europe's leadership in the provision, accessibility and exploitation of electronic mathematical content.

To this end, EuDML will design and build a collaborative digital library service that will collate the currently distributed content by the partnering content providers (national digital mathematics libraries, and publishing platforms). This will be achieved by implementing a single access portal for heterogeneous and multilingual collections, on top of a rich metadata repository. The network of documents will be constructed by merging and augmenting the information available about each document from each collection, and matching documents and references across the entire combined library.

In order to exploit full synergies of the aggregated content, relevant elements such as authors, bibliographic references and mathematical concepts will be singled out and linked to matching items in the collections; similar mechanisms will be provided as public web-services so that end-users or external mathematical resources will be able to discover and link to EuDML items. This way, EuDML will be a new player in the European (and, in general, international) emerging landscape of scientific information discovery services, enabled for reuse in new added-value chains (such as in mashups).

In order to overcome some limitations of the available metadata for some of the integrated collections, a set of dedicated tools will be packaged in order to generate metadata (structured textual OCR, mathematical OCR, keyword extraction, subject classification, bibliographic linking and citation, etc.).

Another set of tools will improve, to the extent that can be reached within this project's duration and resources, the accessibility to the corpus for visually impaired users. Born digital content will be converted to MathML and Daisy and made usable to Braille readers and text-to-speech engines; retrodigitised content will follow a similar path after evaluation of the quality of recognition obtained by current solutions such as the Japanese Infty Reader. In any case, the accessibility tools will prove useful to the overall goal of the project,

as they will at least provide machine-readable approximations of the actual mathematical content, thus make possible new ways of interlinking, associating, and thus querying and navigating the collections.

These technologies have previously been developed and applied in the context of individual collections, or on smaller test beds by technological partners of the project, but have never before been extended to many of the collections that will compose EuDML. In return for this added value and accessibility to their collections, the rights holders agree to a moving wall policy to secure eventual open access to their full texts. Thus EuDML will deliver a truly open, sustainable and innovative framework for access and exploitation of Europe's rich heritage of mathematics.

Due to the sophistication of their facilities, the scale of the combined archives that they operate over, and their interoperability with other resources such as Europeana, the services provided will effect a paradigmatic shift for users of mathematical literature:

- EuDML will serve as a proxy to existing portals that refer to mathematical content, including those aimed at professionals, the educational sector, or the larger public. Some of these portals that have a low coverage of European mathematics will be able to import and refer to this content easily. Even professional services as the Zentralblatt or US' MathSciNet cannot cope with the existing diversity of sources and do not link to a large number of relevant items because of the lack of a suitable interoperability mechanism.
- The availability of a rich, machine-readable version that can be fed to general or specialised Web search engine provides very high visibility and will boost access. For instance, a Google search on "géométrie algébrique et géométrie analytique" (resp. "Bolzano mathematical analysis") gives a first match at NUMDAM (resp. DML-CZ). Extending this recognition facility to the EuDML's full content will be a significant step for European mathematics.
- Once a user has reached the EuDML portal, a number of features will help locate related items, refine queries, and support quickly retrieving the most relevant documents. The portal will also make it possible to leave an annotation on some items (like adding a keyword). As an example, a user could add the keywords "prime number theorem" to Hadamard's seminal paper. This paper is written, in French, in such a modest manner that it bears no explicit reference to its main arithmetic result: getting to that conclusion from machine analysis of full texts (including those recent ones that refer to it) is possibly not entirely out of reach, but much more involved! Allowing end-user tagging and annotation of such texts will enrich the collection, and the accessibility and usability thereof, beyond what current machine analysis technology is capable of.

Starting with a core of public digitisation and digital libraries projects joined by few academic publishers, the EuDML consortium will act under the supervision of an external multidisciplinary scientific advisory board formed under the auspices of the European Mathematical Society. It will exploit existing and emerging standards to integrate the content available at each partner's site, and provide guidelines for integration of further partners, as well as for worldwide cooperation. It will serve as a forum where scientists, learned societies, librarians, publishers and technology and service providers will have the opportunity to share their visions, and design a powerful environment, together with effective strategies and policies for preserving and accessing mathematical references over the long term. It will advocate its results far beyond its initial boundaries with the clear goal of being as inclusive as possible and eventually turning the pilot implementation of a useful service into a sustaining infrastructure. The project will reuse technology available and controlled by the technology providing partners, and the resulting services will be supported by a sustainable business model, striving for comprehensiveness of contents, high quality integration and cost efficiency.

This project will put the European mathematical community at the leading edge of the global drive toward a World Digital Mathematical Library. It will help maintain Europe's foremost position in mathematical research. One of the prominent tools needed in order to perform mathematics-based research in the digital age is a trusted repository of mathematical knowledge, one click away from the researchers. This has been clearly stated by the communication "Digital Mathematical Library: a Vision for the Future" endorsed on the 26th August 2006 by the General Assembly of the International Mathematical Union. Further, it has already been recognised at a national level, as can be seen by the development of nationwide mathematical archives by many member states. We believe it is now time to integrate these archives at the European level and provide a major new component to the European Digital Library. We also believe that allowing new discovery paths for high-quality mathematical content from the very rich European heritage, as well as from

leading-edge research, can raise awareness on this field of European excellence among citizens and students, and lower barriers to accessing its great achievements.

In summary, and explicitly in the context of the objectives of the ICT Policy Support Programme, Objective 2.4: "Open Access to Scientific Information", EuDML will greatly facilitate access to a European-scale collection of digital mathematical content, thereby promoting and supporting its use and exploitation. It will do so by considerably enhancing the quality of that content by capturing new metadata and augmenting, refining and merging existing metadata, as well as providing sophisticated added-value services on the content. The consortium and the organisational structures created by the project will reinforce cooperation between the digital content stakeholders — not just for the duration of the project, but continuing into the foreseeable future. Finally, since the search mechanisms, interfaces, metadata and digital content are all in many languages and must work across many cultural divides, we are confident that EuDML will also tackle multilingual and multicultural barriers, opening widely access and navigation to a currently fragmented, non-interoperable set of isolated repositories.

B1.2 EU and national dimension

The stakes for the European presence at the top level of excellence in this field are high. In this period of rapid changes in the supply chain of high-quality scientific information, the choice for Europe is either to consolidate and expand existing European services, allowing a fair coverage of its scientific production, or to leave the field open to other competing developments (produced mainly in the US).

A firmly established European facility like EuDML will form the basis for future developments, where full control will not be delegated to another continent. It will provide both competition and different traditions to services emanating from non European sources.

The core mathematical knowledge, without which current science can not be understood, has been produced and stored in Europe and disseminated across many countries and languages. It became truly international at the end of the 19th century, as exemplified by the birth dates of the national mathematical societies (Bohemia: 1862, United Kingdom: 1865, France: 1872, USA: 1888, Germany: 1890, etc.). The first International Congress of Mathematicians was held in Zürich in 1897, with 197 members from 15 European countries plus 7 members from the USA. The International Mathematical Union was formed in 1920. The European mathematical community still maintains a top level position for mathematical research. Notice, e.g., that more than half of the 20 Fields Medals awarded since 1986 went to mathematicians born in Europe. Since the Abel Prize was created in 2002 as the mathematical counterpart to the Nobel Prize, 4 laureates out of 8 were from Europe. Thanks to outcomes of this project, the rich European mathematical heritage and current output will be available from anywhere, not only at much lower cost, but also in a form that is easier to discover, use and exploit for academia, industry and the general public. Moreover, EuDML will be an extraordinary contribution to Europeana – The European Digital Library.

This library will have a tremendous impact on the way mathematical research is conducted, streamlining the process of research and enabling new discoveries not otherwise achievable. It is necessary in order to defend the leading position of European mathematical research in the world.

The European DML will also constitute a contribution of the European Union to the support of researchers and other users of the mathematical literature in regions facing special challenges, such as Eastern Europe, candidate countries for the European Union or developing countries. The goal of keeping the resources available freely will enable these researchers and users to easily get the access to mathematical material that is so necessary for scientific, economic and industrial progress.

The project is designed to eventually integrate DML projects of those countries in Europe which support the global effort. It will stimulate additional countries to join in the creation of a global digital mathematical library. It will also be committed to persuading European publishers to cooperate with the library, licensing it for open access with a reasonable moving wall licensing policy.

The participation of Zentralblatt MATH, the most comprehensive international information service in mathematics, which indexes most of this content, certifies that it will benefit from state-of-the-art metadata enrichment and the highest visibility to the professional users. The European dimension is also assured by the collaboration of many independent journals published by small publishers, learned societies, or mathematics departments from all over Europe — through the European Mathematical Society's contributed Electronic Library of Mathematics, MathDoc's CEDRAM publishing platform, and 4 journals edited by the French Society of Applied Mathematics, through the medium-sized commercial publisher EDP Sciences. In this way we ensure that we are not assembling a fossil archive, but a living repository reflecting Europe's

current dynamism in the production of the mathematical knowledge. The whole project will be *scalable* so that technology as well as policies can eventually accommodate all European stakeholders, especially commercial publishers, most of whom do not partner in the project since Day 1 but will be approached during the lifetime of the project, so as to integrate all relevant content in the service.

B1.3. Maturity of the technical solution

Central service infrastructure (content management, resource discovery and communication layers)

An important part of the technical infrastructure will be based on the YADDA software suite (http://yaddainfo.icm.edu.pl/), originally developed by ICM, one of the project partners. Written in Java, YADDA, initially intended as Elsevier's ScienceServer replacement, has since developed into a mature open source, service-oriented distributed repositories integration and provisioning platform. The core services of YADDA, including the index service and the store service, have been selected as the central components of the European DRIVER infrastructure (Digital Repository Infrastructure Vision for European Research), over other solutions, for their performance, robustness and flexibility. Having the European research repositories infrastructure rely on YADDA technology is also a guarantee of its constant development and future improvements. Besides DRIVER, the YADDA suite has proven to perform excellent in a number of different environments, both distributed and centralized:

- Its performance is confirmed by it being used as a provisioning platform for the Polish National Virtual Library of Science initiative, successfully serving a heavily used, multi-million document, full text repository to Polish research and scientific communities.
- Its editing and remote repositories components are being used by a consortium of Polish technical university libraries, proving its scalability and suitability for work in a distributed environment.

Since 2009, the Polish Mathematical Collection, as hosted by ICM, is also provided on the YADDA platform, proving its suitability for that type of content. “AGRO” and “BAZHUM” – the bibliographic databases of Polish natural sciences and humanities, respectively, also are using YADDA as their technology platform, and PSJC (Polish Scientific Journals Contents) database is planning a migration this year.

Technical components

YADDA's service oriented architecture is built around several core components, developed separately but acting together in close orchestration. The most important YADDA services include:

- Index and Search Service – providing high performance indexing of local or remote content (both metadata and fulltext);
- Browse Service – a structural information browsing tool with multidimensional presentation and faceted browsing capabilities;
- Store Service – allowing for intelligent storing of local copies of the original content together with their context information, for use by the Index Service and by automated analytical tools that rely on fast access to the full texts along with the metadata (like the citation discovery and extraction tools or semantic similarity search)

YaddaWeb is a powerful and flexible web front-end, capable of performing simple operations locally, thus offloading the more central and “heavier” services. A single YaddaWeb instance is able to present different user interface profiles, and different skins, depending on the service being accessed and the identity of the user. YaddaWeb has user personalisation and session management facilities.
DeskLight is a separate content editing and maintenance application, providing a set of synchronised (or standalone) remote repositories and allowing for simultaneous collaboration of different metadata editors.
Besides the core services, a number of specialized application modules are available or being developed, including extensive service-to-service and user authentication and authorization mechanisms, and a content categorisation module.
YADDA's flexibility is enhanced by its full multilingual support, support for different types of content (including journal articles and books, but also multimedia, data or compound objects), its rich and extensible internal metadata model, and its powerful object-oriented data structures, representing not only the contents but also the authors, different roles, publishers, and the relations between them.
It is planned that, within the EuDML project, YADDA will provide important functionalities for resource discovery, as well as selected analytical modules and the baseline WWW front-end.

The lower level infrastructure for metadata aggregation will be based on OAI-PMH harvesting. This is well understood and well mastered by virtually all partners in this project, and particularly by the IST, who is the

developer of REPOX, the OAI-PMH framework currently deployed in TEL and to be extended for Europeana.
The Association analyser component will rely on linking technology already deployed in projects such as NUMDAM by CMD, DIGMAP (http://www.digmap.eu) and TELplus by IST, or DRIVER by ICM.
The Metadata enhancer component will use technology initiated by MU in the course of the DML-CZ project, including XML/MathML generation from TIFF images with a workflow using FineReader and Infty Reader, by CMD for generating MathML from LaTeX source using the open source converter Tralics (also adapted by MU), and by UB for extracting MathML representation of formulas from born digital PDF documents.
The Annotation and Accessibility components will use state-of-the art standard tools and open source software internally developed by UB, CMD and MML, backed up with standard YADDA features.
Finally, all the technological partners in the project have a deep knowledge and understanding of the actual web technology and emerging scenarios on interoperability, providing an assurance that the final results will be open and reusable in other emerging scenarios (for example, IST was the partner of the National Library of Portugal that pioneered the interoperability of PORBASE, the national union catalogue, with Google Scholar).

Section B2. Impact

B2.1a. Target outcomes and expected impact

The main outcome of the project will be to unify, into a common gateway, a distributed digital repository of mathematical reference documents. This will provide end users, i.e. researchers with a mathematical background, with a one-stop on-line site for resource discovery and seamless access to a much larger corpus than was ever possible before. Discovery will be made much more powerful thanks to the central search facility, the ability to deeply link content from heterogeneous suppliers, and mathematics-aware enhancements to the metadata making the whole content more accessible, notably to machine learning technologies, which will allow new bridges to be built across traditional barriers such as discipline, language, terminology and notation.

Content providers will enjoy new services such as metadata capture, augmentation and merging, document and reference matching and cross-repository document linking. The unified central access to such a large and integrated federation of libraries will increase both the visibility and usage of each individual collection, supporting the content holders in their commitment to open access to their content, following a short embargo period.

The services will rely on metadata to be provided by the partners and associated organizations, as well as dedicated enhancements specifically derived from this project. Access will be based on an infrastructure for the management and resolution of persistent identifiers for all the works, to be built during the project. Common authority and interlinking structures will be also developed.

The services will be accessible to humans through a web portal with many innovative discovery features, and to other machine services through a set of common digital libraries' interoperability protocols (namely Z39.50, SRU and OAI-PMH with an optional dedicated schema). A metadata registry will make it possible to easily integrate new data providers with new metadata schemas, reusing the aggregated metadata in any new required schema.

The partners in this pilot will bring together the dispersed European heritage of digital mathematical literature in a virtual collection. In its most basic aspect, it will already be an invaluable tool for turning a reference to a mathematical result into a link to its actual write-up, links that users will follow to enjoy more open access. But we will go far beyond this basic infrastructure: we will take advantage of a range of powerful solutions to problems of interoperability that have been deployed at different sites, such as the interlinking of items in their local repositories, mathematical formula searching and analysis of subject similarity. The project will pool and exploit these existing tools and standards to make the repositories interoperable across the borders of collections, languages and countries. Combining the repositories and their individual facilities makes it possible to surpass the mere sum of each project's features, and tackle issues such as multilingual math-aware services, which are not currently dealt with in a systematic manner. EuDML will improve the visibility of Europe's world class, digital content in the field of mathematics.

The scale of the project makes it more robust, because leading-edge techniques to overcome each issue will be provided by its partners; a highly diverse group including librarians and digital library specialists, publishers and structured documents specialists, mathematicians and mathematical knowledge management specialists, professional information service and document engineering specialists, and computer engineering specialists.

Most of the digital content available to the project is already freely available, although much of it is currently barely accessible due to the lack of satisfactory searching and linking facilities and the varying quality and detail of its metadata. However, the project's contributions are not only on the technical side: The collaboration of a variety of stakeholders — scientists, librarians and publishers — in the interest of building the reference digital mathematical library of the future is meant to ensure open access to the reference corpus over the very long term. Very recent content cannot systematically be forced into open access immediately without putting at risk some economically fragile publishers. But, as some of these publishers do produce scientific content of high importance to future generations, it is in the interest of science at large that their content be properly referenced and safely archived by a reliable third party. It might even be the only way to ensure that it will be still preserved and eventually openly accessible to the researchers of tomorrow. In this sense, we are fully aligned with the objectives of making more scientific output open access, in a sustainable way, by insisting on the long term, which is the relevant time scale for mathematics literature. In this pilot, one commercial publisher (EDP Sciences) will contribute its mathematical journals and release them as open

access with a five years moving wall, others (notably Elsevier and Springer Verlag) will do it through content, mostly digitised, that is already available in some of the partnering digital libraries. In the long term, the success of the project will obviously need to convince more publishers to provide their content. We will leverage on EDP Sciences' success story and EuDML success indicators to gain a wide acceptance among all stakeholders. We also need to be backed up by research funding organizations in Europe. Many of them are already involved in the national projects we will integrate, and will support our steps further. Moreover, the European Science Foundation expressed its interest in this endeavour when it organized a meeting in Santiago de Compostela on the necessity of a "European Virtual library of mathematics" (March 2009).

EuDML will make historical and comparative analysis, or even serendipitous discovery, of the development and achievements of European mathematics much easier for the European citizen. It will also enlarge considerably the overall user base for those resources, by lowering the barriers to discovery by means of search and navigation tools that are much more intuitive and efficient for non-specialists. This is also true for non-specialists in neighbouring fields who currently find it difficult, or at least discouragingly time-consuming, to navigate through sources dispersed on the Internet, with heterogeneous metadata sets, language support and coverage. In this respect, offering a central interoperable access point with uniform standards should prove particularly fruitful for interdisciplinary research.

Raising the awareness and the image of the great mathematical knowledge produced in Europe over past centuries can only have a positive impact on the inclination of European pupils and students for the scientific curriculum.

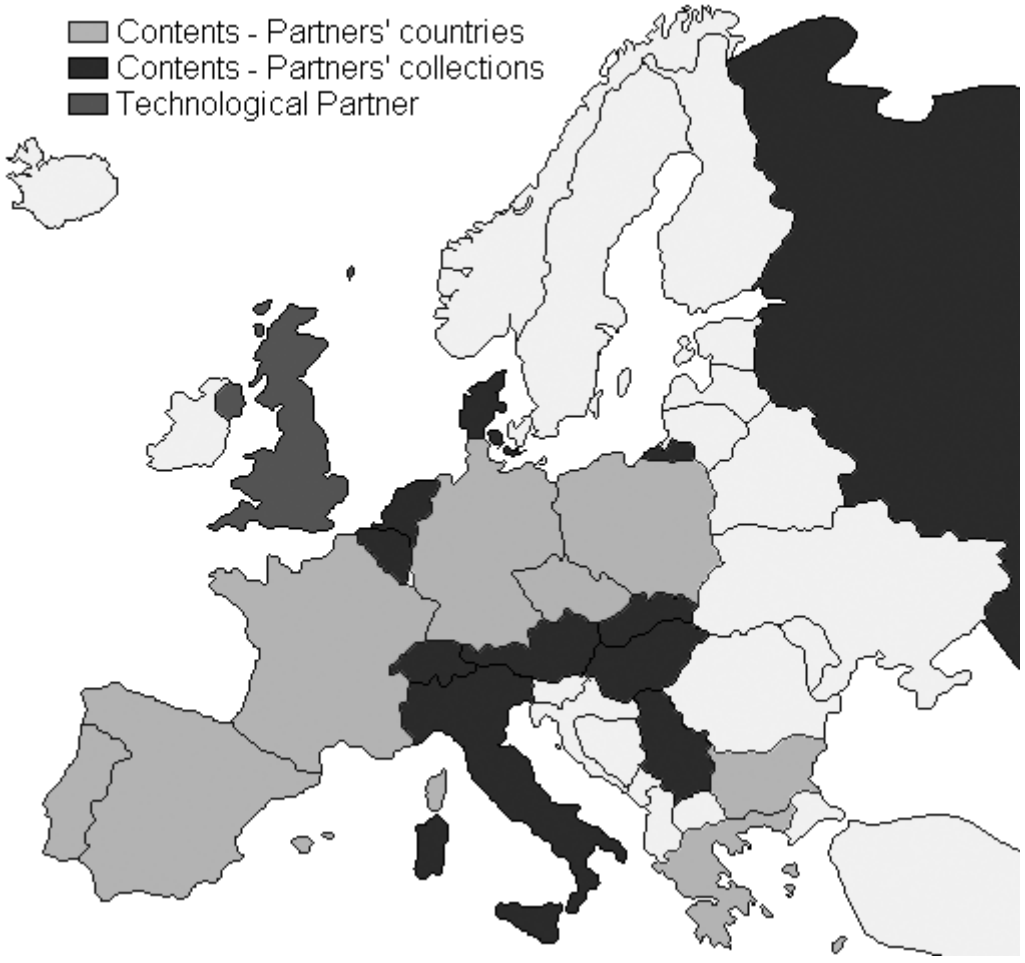


Figure 1: The European dimension of EuDML.

B2.1b. Underlying content

The following heterogeneous digital libraries represent state-of-the-art content we will integrate in EuDML (these collections contain a critical mass of the available content in Europe, of more than 2.5 million pages from 130,000 original items at the time of writing):

1. The collection *Mathematica* at State and University Library Göttingen produced by the DFG-funded project Electronic Research Archive for Mathematics (ERAM) with more than 500 books and 42 journals contains, in total, ca. 1.1 million freely available pages. In particular, the archive contains several important and first class mathematical journals, like *Mathematische Annalen* (1:1869-306:1996), *Inventiones mathematicæ* (1:1966-123:1996) and *Journal für die reine und angewandte Mathematik* (Crelle's Journal, 1:1826-493:1997) covering their "print only" period. Collections extend to journals from Switzerland, central Europe and South America. The collection of books (monographs and multivolume works) contains, for example, the collected works of Carl Friedrich Gauss, the Habilitation thesis of Bernhard Riemann, and the Ph.D. thesis of Emmy Noether.
2. The ICM provides free online access to its mathematics resources in the Biblioteka Wirtualna Nauki (Polish Virtual Library of Science) with 39 books and 10 journals, a total of over 400 volumes (over 170,000 freely available pages). The repository contains journals such as *Fundamenta Mathematicæ* (1:1920-165:2000), *Acta Arithmetica* (1:1935-51:1988,60:1991-95:2000), *Studia Matematica* (1:1929-143:2000) and *Colloquium Mathematicum* (1:1947-11:1963,62:1991-67:1994), as well as fundamental books of great mathematicians including Stefan Banach, Waclaw Sierpinski, or Casimir Kuratowski. ICM's resources comprise a highly multilingual collection, including publications in English, Esperanto, French, German, Polish, and Russian.
3. The Cellule MathDoc at University of Grenoble runs the NUMDAM and CEDRAM programmes, which are widely acclaimed for their mathematically oriented user friendly interface. NUMDAM is a mathematical digitisation programme among whose collections (ca. 700,000 pages) are renowned journals such as *Annales scientifiques de l'Ecole Normale Supérieure* (1864-2007), *Annales de l'Institut Henri-Poincaré* (1930-2000), and *Publications mathématiques de l'Institut des Hautes Etudes Scientifiques* (1959-2007), as well as two journals from Italy, *Compositio Mathematica* from Netherlands, and reference seminars such as those chaired by Cartan, Chevalley, Leray, Schwartz and Bourbaki. Main languages are French, English, Italian, and German. In 2008, NUMDAM started to acquire born-digital content directly from 3 publishers, accounting today for more than 100,000 pages. The CEDRAM platform is also an enabling infrastructure for independent and society journal publishing, which outputs ca. 5,000 pages per year for 5 journals and 3 seminars. All metadata is in XML, with mathematical formulas captured as MathML. Moreover, MathDoc, being associated with the Bibliothèque nationale de France, maintains an article-level catalogue of some journals and collected works digitised at Gallica, as a way to enhance access and usability to these important resources.
4. The CSIC, through the Institute IEDCYT (formerly CINDOC), runs the DML-E program under the supervision of the CEMAT consortium, which is the Spanish organization of mathematical societies. It contains 21 journals from Spain and Latin America (about 75,000 pages from 4,300 articles since 1975, both retrodigitised and born digital). It contains journals such as *Revista Matemática Iberoamericana, Collectanea Mathematica, Publicacions Matemàtiques* or *Revista Matemática Complutense*.
5. The Institute of Mathematics AS CR in Prague leads the ongoing DML-CZ project, under the technical management of Masaryk University in Brno, with a current provision of 9,400 digitised articles from 6 journals, totalling 97,000 pages. The language coverage is wide: English 69 %, German 10 %, Czech 8 %, Russian 8 %, French 5 %. It includes a collection of writings of Bernard Bolzano and monographs of other Czech mathematicians. By the end of 2009 the DML-CZ will hold more then 200,000 pages.
6. The European Mathematical Society, in cooperation with the Fachinformationszentrum (FIZ) Karlsruhe, has set up an important library of more than 80 open access research journals in mathematics (Electronic library of mathematics: http://www.emis.de/ELibM.html). Metadata for these works is scarce, so the facilities of a project such as EuDML are necessary to provide any hope of integrating them into a accessible digital library. The fact that the LaTeX sources are available will ease the integration.
7. EDP Sciences is a medium-sized publishing company, and a subsidiary of learned societies, which works closely with the scientific world. The editorial activities of the company cover astrophysics, applied and fundamental physics, mathematics, electronics, materials sciences, life sciences, and medical fields. It offers a specialist publication platform. The quality of its contents is validated by international scientific committees. It will contribute the commercial publisher point of view to the project, with the output of the 4 electronic journals in the ESAIM series, edited by the French Applied Mathematics Society

(SMAI). Their archives are digitised by the NUMDAM programme, which will be integrated in the EuDML platform for indexing and preservation.

8. *Portugaliae Mathematica* is a major reference journal of the Portuguese Mathematical Society. The National Library of Portugal digitised the collection from 1937 to 1993 (1,347 works in 14,828 images). The issues from 1993 were already published in digital format, and all the articles are described and indexed in detailed metadata. All the collection will be available to EuDML, under a close cooperation agreement between the Portuguese Mathematical Society and the IST.
9. *Zentralblatt MATH*, produced at Fachinformationszentrum (FIZ) Karlsruhe, is the world's most complete and longest running abstracting and reviewing service in pure and applied mathematics. The Zentralblatt MATH Database contains more than 2.8 million entries drawn from more than 4,600 serials and journals and covers the period from 1868 until present following the recent integration of the Jahrbuch database (JFM). The entries are classified according to the Mathematics Subject Classification Scheme (MSC 2000). Together with *Mathematical Reviews*, a service from the American Mathematical Society, it is a fundamental information service which has been widely Europeanized thanks to the FP5 LIMES project. It is a pre-existing infrastructure that will be exploited to some extent in order to overcome the high heterogeneity of the resources to be collated in this project, and that will benefit a great deal from its outcome, as it will enhance its operation with the possibility to actually refer to a much wider proportion of the texts there reviewed. Zentralblatt has recently converted all its TeX metadata into MathML.
10. RusDML (Russian Digital Mathematics Library) is part of a global effort to make all mathematical literature digitally available to mathematicians around the world. The focus of the RusDML project, as its name suggests, is Russian-language literature. The first stage of this project, which is approaching completion, is to digitize the most important Russian-language journals from 1866 to the present. The RusDML project, which is supported by the DFG (Deutsche Forschungsgemeinschaft), involves three German partners: the Technical University Berlin, Göttingen State and University Library, and the Technical Information Library Hannover. The Russian partner is the National Public Academic Technical Library in Moscow.
11. DML-Bulgaria will be an example of a yet-to-begin digitisation project, at the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences in Sofia. The overall estimation of publications of Bulgarian mathematicians is 90,000 pages. The main Bulgarian mathematical journals to be included in DML-Bulgaria are *Annuaire de l'Université de Sofia* (1906-2000, 14,255 pages), *Comptes rendus de l'Académie Bulgare des Sciences* (1948-2002, 64,290 pages of which 8,468 pages of mathematical papers), *Journal of Theoretical and Applied Mechanics* (1970-2002, 12,673 pages), *Mathematica Balkanica* (1971-2001, 8,852 pages), *Proceedings of the Annual Conference of the Union of the Bulgarian Mathematicians* (1972-2002, 14,456 pages), *Serdica Mathematical Journal* (1975-2002, 10,456 pages), *Pliska Studia Mathematica Bulgarica* (1977-2000, 2,127 pages), *Serdica Journal of Computing* (2007-, 3,000 pages) *Physico-Mathematical Journal* (1958-1993, 8,000 pages) and the mathematical monographs of BAS (3,197 pages). We expect EuDML to provide a fundamental boost to this initiative.
12. IU – Ionion University, maintains a research bibliographical database on the Web, Greek-language journals and conference proceedings in the field of mathematical research and education. The service provides a single point of access, as also searching according to semantic criteria, browsing according to predefined subject headings and linking every article to its corresponding review. The main mathematical journals included in the service are *Plus Mathematics, Mediterranean Journal for Research in Mathematics Education, THEMES in Education, Quantum, Diastasi, Cahiers de didactique des mathématiques, Theaititos*. In addition we have also digitized a number of historic archives on mathematical topics, and material from Greek educational institutions.

All these collections contain published validated mathematical texts of high importance to current and future generation of users of mathematics, and will continue to be enriched and extended. What we aim at is to provide a uniform European digital library from all these sources (as, for example, TEL already did for national libraries) but also to offer new added-value services.

It should be stressed that mathematics is to be understood in a very inclusive way throughout this text, as borders with other sciences, such as, to name a few, physics, economics and social sciences, were not so strong in the past, and are still crossed very often today, which is another indicator of the vitality and impact of mathematical research. As a consequence, most of our collections contain texts that would be qualified today as belonging to the aforementioned disciplines (e.g., Poincaré, Pasteur, Einstein).

Quantity and Quality of the Content								
Provider	**Type**	**Quantity & Definition**	**Format & Quality**	**IPR**	**Current Use**	**Existing Metadata**	**Language**	**Additional comments**
CMD	Journal articles 1810-2006	750,000 pages 43,000 articles	PDF/DjVu 600 dpi bitonal + hidden text 10% from born digital	Moving wall licence to post on www.numdam.org From author/publisher	>120,000 unique visitors/month about 1 million articles downloads/year	Author, title, Bib. Ref., Abstract, OCRed full text MR/Zbl IDs, Linked biblios	French (65 %) English (29%) Italian (5.6%)	NUMDAM Retrodigitised collections (50 serials)
	Journal articles 2001-2007	35,000 pages 1,400 articles	PDF born digital	0-5 years moving wall	10,000 visitors/month about 70,000 downloads/year	Idem + MSC (LaTeX/MathML metadata)	French (35%), English (65%)	CEDRAM publishing platform
CMD/BNF	Journal articles 1836-1934	40,000 pages 2,000 articles	PDF 300 dpi monochrome image	Public domain	1,000 unique visitors/month	Author, title, Bib. Ref. (LaTeX)	French	Gallica-Math: JMPA (Gallica content, CMD metadata)
	Collected works	40,000 pages 2,000 articles/chapters	PDF 300 dpi monochrome image	Public domain	1,000 unique visitors/month	Author, title, Bib. Ref. (LaTeX)	French, Latin	
ICM	Journal articles 1888-2000	124,500 pages 9,700 articles	PDF - scan monochrome	Rights obtained from the publisher, Freely available	1,300,000 articles/year 60,000 unique visitors/month	Author, title, volume, pages, editor	English 71%, French 15%, German 7%, Polish 5%, Russian 2%, Italian	Possible to make full metadata for 930 articles
	Journal articles 1920-1927	3,300 pages 275 articles	PDF -scan monochrome	Rights from publisher. Copyright expired		Author, title, volume, pages, editor, description	French	Cataloguing in process
	Journal articles 1991-2000	26,500 pages 1,880 articles	PDF born digital	Freely available		Author, title, volume, pages, editor	English 90%, French 7%, German, Russian	Possible to build full metadata for 930 articles
	Journal articles 2001-2007	4,850 pages 360 articles	PDF born digital	open licensed from the publisher	120,000 articles per year 7,000 unique visitors/month	Author, title, editor, volume, pages, abstract, full text, references	English	
	Books 1924-1979	13,250 pages 690 chapters	PDF scan monochrome	various, freely available	350,000 articles per year 20,000 unique visitors/month	Not yet	Polish 45%, English 15%, French 26%, German 13%, Russian	Possible to create metadata: author, title, book, editor
EDPS	Journal articles 2001-2007	22,300 pages 1,200 articles	PDF born digital	5 years moving wall		Author, title, Bib. Ref., Abstract, DOI, Linked biblios (LaTeX/XML)	French (2 %), English (98%)	ESAIM series
IST	Journal articles	More than 2,000 articles (Portugalae Mathematica)	TIFF/PDF 600 dpi mainly bitonal	Freely available, with a 5 years moving wall	Collection available in the National Digital Library and in TEL	Dublin Core, UNIMARC, TEL Profile	Portuguese, English, French	Fully indexed with MCS

SUBGoe RusDML	Journal articles	13,000 articles, 185,000 pages	TIFF / PDF 600 dpi, bitonal	Freely available	not fully active yet	Author, Title, Year, Publisher and/or Place	Russian	In progress, additional pages in preparation
SUBGoe Mathematica (ERAM/JFM)	Journal articles 1777 - 2001	50,000 articles, 950,000 pages	TIFF / PDF 600 dpi mainly bitonal	Freely available	about 50,000 unique visitors and 10,000 articles downloads/month	Author, Title, Publisher, Bib. Ref.	German, English, French and other languages	800,000 additional pages in the next two years (mostly monographs up to 1900)
	Books 1596 - 1993	500 books 53,000 pages						
IMAS	Journal articles 1951-2008	83,000 pages 7,600 articles	PDF 600 DPI bitonal	Mowing wall licence to post on dml.cz obtained from publisher; partly freely available	Not fully active yet	Author, title, Bib. Ref., MR/Zbl IDs Linked biblios	English, Russian, German, French, Italian, Czech, Slovak	107,000 additional pages by the end of 2009
	Journal articles 1992-2008	21,000 pages 1,900 articles	PDF 600 born digital	Mowing wall licence to post on dml.cz obtained from publisher	Not fully active yet	Author, title, Bib. Ref., MR/Zbl IDs Linked biblios	English, Russian, German, French, Italian	28,000 additional pages by the end of 2009
	Books 1810-1981	4,400 pages 24 volumes	PDF 600 DPI bitonal	Rights obtained from publisher and author or Copyright expired	Not fully active yet	Author, title, MR/Zbl IDs Linked biblios	German, English, Czech	
	Conference proceedings	3,500 pages 830 articles	PDF 600 DPI bitonal	Licence to post on dml.cz obtained from publisher	Not fully active yet	Author, title, MR/Zbl IDs	English, Russian, German, French	10,000 additional pages by the end of 2009
CSIC	Journal articles 1948-2008	4,300 articles 75,000 pages	PDF 600 DPI Bitonal and born digital	Rights obtained from the publisher	800 unique visitors/month	Author, title, Bib. Ref., Abstract MR/Zbl IDs Linked biblios	English 76%, Spanish 16% French 6% Catalan 1.5%	Additional pages are in preparation
IU	Journals and Conference proceedings 1850-today	>15,000 articles	PDF, PS	Freely available	Not fully active yet	Title, author, abstract, keywords	Greek English	
FIZ	References and Reviews from ZMATH (1868 - today)	2.8 Mio entries	TeX, HTML, PDF, txt	Copyright by FIZ, Springer Verlag, the Heidelberger Academy of Sciences	8.5 Mio requests p.a.	Author, title, abstract / review, classification, keywords, language,, year, etc. Converted to MathML	English, German, French	World wide coverage of mathematical publication
	Journal articles 1990 - now	33,500 articles	PDF, PS	freely available	10,000 unique visitors/month	Author title, abstract classification, keyword source, language, publication year, etc.	English and other European languages	The electronic library of the EMS
Summary	**> 2,600,000 pages > 170,000 bibliographic items > 2.800.000 reviews**			**95% open access**		**Heterogeneous metadata, but the focused scope will make it possible to create added value with it.**	**English, French, German, Polish, Russian, Czech, Slovak, Spanish, Italian, Portuguese, Catalan, Latin**	

IPR issues

Our clear policy, which will be advocated throughout the course of the project and beyond its founding participants, is that all digital mathematical literature should be preserved through a network of academic libraries enforcing eventual open access to everyone. This is the policy endorsed by the International Mathematical Union as described in its August 2002 communication: "Best current practices: Recommendations on electronic information communication". Given the nature and usefulness of the mathematical literature as a long-term reference, we feel that enabling eventual free access to it is more important to the scientific community than instant open access to a subset. A large part of our retrodigitised collections is in the public domain. A substantial part is too recent to be in the public domain, but each partner has taken care that authors and publishers have agreed to a moving wall policy, which means that any text becomes eventually freely accessible, with a delay that is agreed upon by the stakeholders. A typical value widely used currently is 5 years, which seems to achieve a reasonable balance between availability of the texts and incentive not to cancel subscriptions to small academic publishers who often could not afford open access. Notice that half the citations in papers published today point to references published more than 10 years ago. Thanks to this moving wall policy, many articles which are not in the public domain, nor published in some flavour of open access, will be freely accessible through the EuDML portal. On the other hand, the work with publishers for integrating new content into the library will provide only searchable metadata in the first place, paying back the publishers by giving them a share of the enhanced visibility achieved through our various innovative discovery tools. As a counterpart, the full texts will be made freely available at the end of the moving wall. As we believe the library we intend to build is the model that should be developed and extended to the whole of mathematics — and possibly beyond — it is an important feature that the model for transferring new content into it be well accepted by all stakeholders, and leave enough room for non open access commercial publishers to cooperate. Given this, and given that an already substantial part of new material in our partner's collections is open access, we estimate that around 5% of the aggregated collections described in this document will not be accessible at the beginning of the project, but will be within the next 5 years.

Concerning software and tools developed in this project; every tool developed to handle a specific collection is bound to that collection's formats and owners. Such tools can use the background management system used by participants and are not expected to be shared outside of this project. On the other hand, the formats and standards developed in this project will be public and even advertised as we intend to set the first blocks of a universal system. We will release, as open source, all the tools that exploit these formats, that provide interoperability, mining, matching, and that could be of interest to other communities.

Multilingual and/or multicultural aspects

Using the multilingual mathematical knowledge already developed by our partners and colleagues, we will support many European languages as well. The already digitised source material of the project partners cover a long tradition of mathematical research in Europe, which was published in the authors' own languages as well as in those international languages amongst which the most frequent are English, French, German, and Italian (most of the important mathematical journals published today still accept articles in these languages). Thus, the digitised source material is multilingual in essence, and more so as we go back in history.

It is thus an essential feature of the EuDML project to address the difficult question of providing access to articles based on their subject, or scientific meaning, rather than on their language (after all, Mathematics can itself be viewed as a formal language, so that discovery resources associated with this language will perform equally well on our whole content). This is a major departure from standard search engines such as Google or Yahoo and reflects a true paradigm shift for end users.

Three main directions will be investigated, they will be addressed in the definition of the metadata profile that will be developed at the start of the project, and room for future improvements will be reserved during the life-cycle of the project. Moreover, given that the available metadata is quite heterogeneous among the contributed collections, we will first experiment with the richer subset in order to test our methods, and then extend these techniques to the remaining collections that could be upgraded during the course of the project, using some of the automated enhancements foreseen below:

- The first and most straightforward strategy is to use existing thesauri of mathematical keywords, and use translation lists so that a query in one language can return documents in any other. This work will take its roots in the successful Jahrbuch project, where mathematicians added such English keywords or subject classification codes to old articles. Other sources of English keywords attached to texts without English metadata can be generated from the Mathematical Subject Classification or other less specific subject classifications, such as the UNESCO thesaurus and (possibly OCR generated) full text analysis of mathematical content based on MathML formulae and similarity with other, already classified, texts in the database. Moreover, users will be able to suggest themselves more up-to-date descriptions and keywords attached to any EuDML item, thanks to our annotation component.
- The second direction is to consider interlinking as a powerful access tool to the mathematical resources regardless of their language, but rather according to their subject and/or their scientific importance. The most basic interlinking example consists of adding to the metadata describing a document a link to its review in one of the reviewing databases (Jahrbuch, Zentralblatt MATH, Math. Reviews). A user can thus be given the MSC code of the article, and search for similar articles based on their subject using the advanced discovery and interlinking tools already available at the reviewing databases websites. The interlinking infrastructure deployed in the project will also allow the exploitation of links to and from other, related resources, such as citations from reviews and from subsequent works. This can be used to lend weight and significance to resources in a manner analogous to Google's PageRank algorithm. Such webs of citations provide a language-neutral way to find resources across multiple languages. The most fundamental works available in EuDML will have many links pointing to them regardless of the language in which they were written. Also, this will make it possible to design, for each work, its "social network", which will make it possible to offer a powerful scenario for resource discovery by serendipity (especially relevant for students and historians).
- The third strategy is based on the observation that the mathematical structures contained in mathematical texts are highly significant, and behave like a formal language which will allow developing powerful automated agents for those corpora for which we will have the mathematical content encoded in a semi-structured format. Mathematical knowledge management techniques will be solicited to assess its novel technologies such as mathematical OCR, XML/MathML full-text generation from (La)TeX source files or PDF, formula representation and searching, and mathematical similarity metrics.

Critical mass

It is estimated that about a third of the entire European mathematical heritage from 19th century onward exists in some digital form, and probably more than half of the core journals, summing up to about 10 million pages, most of which not freely accessible. The collections brought to this project sum up to over 2.6 million pages, which gives a clear picture of their weight in the current landscape.

Important bodies of mathematical literature are also digitally available in the rest of the world (JSTOR, project Euclid in USA, e.g.) but, to date, there have been no successful attempts to structure the efforts at a trans-national level.

A one-year (2002-2003) planning project coordinated by Cornell University Library was funded by the U.S. National Science Foundation (NSF) "toward the establishment of a comprehensive, international, distributed collection of digital information and published knowledge in mathematics", with a steering committee that was mostly European. Most of the conclusions from that group are still valid today: the need for standardization and coordination, the identification of intellectual property rights and conflict of interests among stakeholders as principal inhibitors. In 2005, the Moore Foundation considered funding a gigantic World DML, but faced the same inhibiting factors.

We believe that this situation won't be resolved, and that therefore that scientists will still lack the necessary infrastructure for handling the reference mathematical corpus, unless a core group of stakeholders takes up the challenge. This core group must have a critical mass in content, know-how, and a sufficient organisational diversity. They must take a pro-active approach and set the networking infrastructure, standards and policies so that we can further build on the current state of the art and aim at comprehensiveness in content while expanding geographically.

We claim that our consortium represents this critical mass, and will have enough authority, especially thanks to the support of the European Mathematical Society and EuDML's strong relationships to national societies

and the International Mathematical Union, to shape the future of the World Digital Mathematics Library with EuDML as its root.

Publishers linked to EuDML

Today, there is still a large diversity of publishers active in the mathematical field. They range from very large generalist publishers in the STM area (like Springer or Elsevier group) to small groups publishing a single series or journal (a mathematical department at some university, e.g.). In between are medium sized academic publishers and a large number of mathematical societies across the world. Some of these academic publishers only deal with the scientific editing of the content they publish, outsourcing publishing business to private companies, while others take care of the whole process internally. It has to be stressed that publishers of the highest scientific quality and reputation are to be found in each of the above categories.

While Springer-Verlag expressed interest in the DML initiative since its inception, and launched the EMANI initiative in cooperation with few libraries, they eventually changed direction when they merged with Kluwer and opted for running by themselves a privately owned, profit-driven generalist library integrated into their publishing platform, in a similar fashion to Elsevier. It was not possible to reach an agreement with these very big publishers during the time of preparation of this project, but we hope to be able to convince them of the benefits for them of our independent library model during the course of the project, notably by highlighting measured benefits for publishers partnering in EuDML. Notice however that they are linked to the project through content contributed by some of its partners: GDZ has some Springer and Birkhaüser books and journals, NUMDAM has 3 Elsevier journals and 1 Springer journal for which the publishers agreed to transfer their born-digital content one year only after publication, etc.

EDP Science, a medium-size STM publisher, is partnering in EuDML with the aim of contributing all its mathematical journals. A large number of small publishers are partnering as well through many EuDML partners who will contribute recent content from these publishers. CEDRAM at MathDoc (CNRS+UJF/CMD) will contribute 7 journals and seminars published by laboratories in French universities, DML-E at CSIC, DML-CZ at MU and IMAS will contribute content recently published by mathematical societies (in fact most of their journals are updated continuously by their publishers). Finally, the ELibM service of the European Mathematical Society, which is run by FIZ, provides the electronic edition of more than 80 journals from a wide variety of publishers across Europe and beyond, most of which will be contributed in real time to the project.

B2.2. Long term viability

The consortium intends to achieve sustainability of the EuDML services after the end of the project. The project efforts are aligned with the strategic long-term goals of the partner institutions and organisations that the results of the project are intended to serve. All the memory institutions are committed, in order to preserve and make accessible the mathematical heritage they care for, to acquire either past content through digitisation or new content through transfer of digital files from their publishers, and licensing for eventual open access.

The principal aims of sustainable EuDML services will be:

- to work toward comprehensiveness, service integration, and cost efficiency of the EuDML services,
- to assist in exploiting the benefits of networking for integration of digital library services such as data sharing and improvement,
- to advance cooperation with commercial partners, demonstrating benefits of cooperation for their wealth, advocating the necessity of a distributed preservation plan for their precious output,
- to create a non-profit service in the interests of the mathematics user community.

The exploitation plan will encourage synergies, accelerate wider adoption and overcome barriers to exploitation by lowering entry barriers for new information providers such as libraries or publishers. Wide dissemination of results is planned as is exchange of experiences across borders and scientific sectors and participation in co-ordination frameworks. The main objective is to improve access to large collections of literature for researchers, users in application areas (industry applications like e.g. mathematics of finance, public key cryptosystems, simulation, etc.), other professionals, and all others interested in mathematics.

In order to create a sustainable service from the EuDML project, there arises the need to account for:

- an organizational and legal framework, which will take its roots in the EuDML consortium and partners and institutions associated during its lifetime;
- balancing costs and potential sources of revenue of running the EuDML services;
- a common framework for dealing with IPR and copyright issues.

The strategic impact of the EuDML project is in the area of improvement of competitiveness of scientific/cultural sector, especially with regard to recent digitisation initiatives from the U.S.

The work plan for dealing with these issues is detailed in the description of work package WP2 (policies, exploitation, and dissemination).

B.2.3. Wider deployment and use

The currently digitised mathematical literature is not currently known or used to the scale that is possible and useful. This has to be changed by publicizing the EuDML effort and its achievements. This will be done in the form of publications, presentations at formal and informal meetings and discussions with stakeholders, conferences, and in several mathematics research institutions. This information will be disseminated and linked through the EuDML network, and project website. Good information dissemination on the project progress will be a key issue and will be done in accordance with the pioneering role of the project. The partnership with the European Mathematical Society will provide opportunities to publicize the project's outcome to the scientific communities at various events such as showing the available historical resources on the topic of a conference, disseminating poster exhibitions through the ERCOM network of scientific meeting centres across Europe and beyond, organizing poster sessions, or training sessions for researchers in parallel to the conference's main stream.

During the course of the project, two workshops will be important dissemination milestones: the first, after the initial 6 months phase of the project is completed, will bring together all content providers in the project to discuss the findings of the preliminary analysis of the collections, and the technical orientation to be pursued. This information will then be disseminated by the participants of the workshop. The second one, to be held one year later, will be open more widely, in particular to those stakeholders that might join the project, and present the preliminary version of the portal and the findings of assessments and user surveys.

Technically, awareness will be increased through interoperability with services such as TEL or the Europeana portal, where temporary exhibitions about mathematicians, or history of scientific subjects, e.g., can be used as teasers to the general public, with links to the original source texts powered with EuDML technology for linking and associating EuDML content with mathematical content.

The eased ability of linking toward European mathematical references will also give it much more visibility. We will pursue a pro-active linking strategy, by contacting mathematical resources with an important user base that refer to mathematical content and providing them with links for their references (examples are the MacTutor historical archive, the Mathematics genealogy project, the Wikipedia mathematical pages, PlanetMath, Wolfram's MathWorld).

Many of the technologies integrated within this project in order to generate MathML versions of the metadata, and to use this as a way to find new paths in the literature, make sense for the much wider corpus of scientific and technical texts. After having been assessed within this project, they will be offered to other communities.

Finally, we believe that EuDML also will be a very relevant content aggregator for future integration in Europeana (IST will be the technical liaison partner, so assuring a perfect alignment between the technological infrastructure to be developed and deployed for EuDML and the equivalent efforts also in place for Europeana).

Section B3. Implementation

B3.1. Consortium and key personnel

The EuDML project is a project of the European mathematical scientific community and mathematical subject information providers to propagate a joint forum for the great variety of different European digital libraries activities. It is in the nature of the project that there are partners in the project realising the practical and technical part of the intended EuDML service as well as partners acting as stakeholders and representatives of the mathematical scientific community. Moreover, projects like Göttingen's GDZ are active in the EuDML network without applying for financial support within this project. Starting with an active team that also represents a variety of organisations, the project is open for cooperation with other initiatives and is intended for further expansion to progressively cover increasingly more and more of the European mathematics corpus.

Partner No. 1 (Project coordinator): Instituto Superior Técnico: Computer Science Department (IST)

The Instituto Superior Técnico (IST), with nearly 8,000 students and 600 Professors with a PhD, is the most important school of the Lisbon Technical University and the oldest and largest Portuguese school of engineering. Its mission is to contribute to the development of society by promoting higher education of outstanding quality in the areas of Engineering, Science and Technology, at undergraduate and postgraduate levels, and by carrying out research and development activities in accordance with the highest international standards.

José Borbinha, **EuDML's general project coordinator**, is Professor of the Computer Science and Engineering Department of IST and the leader of the Information Systems Group at the associated laboratory INESC-ID. His main area of research is information systems in general and Digital Libraries in particular (architectures, preservation, metadata, digitisation and interoperability). He was director for I&D for the National Library of Portugal, is the chair of the IEEE Technical Committee on Digital Libraries, and has been involved in several projects related with TEL and Europeana.

Partner No. 2: Cellule MathDoc at Université Joseph Fourier Grenoble (UJF/CMD)

Cellule MathDoc is a "joint service unit" both of the Centre National de la Recherche Scientifique and of Université Joseph Fourier (Grenoble). It is devoted to mathematical documentation and in charge of the NUMDAM program where a large number of mathematical journals have been digitized and made available at http://www.numdam.org. Many of the features expected to be implemented at the European level by this project are already running there (we can stress reference linking to reviewing databases and to full texts, acquisition of born digital content from some publishers, among whom those collaborating with CEDRAM, Springer SBM and Elsevier Science). However, it lacks uniform multilingual and math-specific features which should be better tackled at the European level. For the purpose of this project, CMD is split up in two parts: UJF/CMD for UJF staff at CMD, which will take over the major part of the work to be achieved, and CNRS/CMD, partner No. 14, for CNRS staff at CMD whose involvement will be mandatory in some specific tasks.

Thierry Bouche, EuDML's scientific coordinator, is mathematician, maître de conférences at University Joseph Fourier (Grenoble). Member of the electronic publication committee (EPC) of the European Mathematical Society (EMS), co-chair of the digitisation standards committee of the WDML (IMU committee), in charge of NUMDAM, CEDRAM, Gallica-Math, mini-DML at MathDoc. Invited speaker and member of the programme committees of various conferences in the field of mathematical digital libraries.

Claude Goutorbe is engineer in computer science. Highly qualified specialist in Information Retrieval, databases, digital documentation, and web applications. He has developed several major information systems such as the database manager and web interface for Zentralblatt-MATH, NUMDAM, CEDRAM and mini-DML.

Partner No 3: University of Birmingham: School of Computer Science (UB)

The Birmingham School of Computer Science is an internationally leading institution with particular strengths in mathematical foundations of computer science, artificial intelligence and natural language processing. The School's Automated Reasoning group has worked on issues pertaining to electronic representation of intuitive mathematical concepts and the development of mathematical ontologies and was involved in the EU-network on Mathematical Knowledge Management (No. IST-2001-37057). The School's Scientific Document Analysis Group has been working on dedicated OCR systems for scientific documents and in particular on techniques for mathematical formula recognition and on extraction of mathematical notation from existing electronic documents. The Natural Language Processing group has a particular focus on the understanding of semantic and pragmatic meaning in text and have expertise in information retrieval and probabilistic techniques such as unsupervised topic-wise clustering of large document collections, textual entailment and document summarisation.

Mark Lee: Member of the School's Natural Language Processing group since 1998 and a lecturer since 2000. He works on the automated understanding of natural language text and in particular the modelling of contextual and pragmatic meaning. He also has interests in the development of shallow text processing techniques for applications in areas such as education and he has a research background in corpus linguistics and the data-driven analysis of natural language in large bodies of text. He is a co-investigator on several national EPSRC grants with focus on corpus study and implementation.

Alan Sexton: Lecturer in the School of Computer Science since 1991. With Volker Sorge, he is a co-leader of the Scientific Document Analysis Group and was an investigator in the EU-network MKM (No. IST-2001-37057). He has worked in databases for more than 15 years. Notably, he developed a novel approach to recognising fine distinctions in the components of single characters, which is currently in commercial use by the font foundry Bitstream. This work has since lead into optical character recognition and mathematical formula recognition.

Volker Sorge: Lecturer in the School of Computer Science since 2002. Member of the automated reasoning group and together with Alan Sexton he is a co-leader of the Scientific Document Analysis Group. He has worked on the integration of heterogeneous mathematical software systems and recently on the question of how mathematical knowledge can be adequately managed and retrieved, and has developed ontologies to sensibly structure mathematical concepts for the automated generation and understanding of mathematical textbook languages. He was an investigator in the EU-IHP Network CALCULEMUS (No. HPRN-CT-2000-00102), the EU-network MKM (No. IST-2001-37057) and has coordinated Birmingham's activities in the EU network of excellence CoLogNET (No. IST-2001-33123).

Partner No. 4: Fachinformationszentrum Karlsruhe / Zentralblatt MATH (FIZ)

Fachinformationszentrum Karlsruhe, shortly FIZ Karlsruhe, is an international scientific service institution, dedicated to providing information services and solutions for both information management and knowledge transfer in science and industry. Its activities focus on the development of e-science solutions and the provision of a worldwide unique collection of databases through its online service STN International <http://www.stn-international.de/> (The Scientific & Technical Information Network). For almost three decades, FIZ Karlsruhe has been offering its high-quality, value-added information services to scientists from all over the world involved in academic and industrial R&D, as well as to decision makers in business and administration.

The department mathematics and computer science of FIZ Karlruhe produces and provides various databases in mathematics and computer science. Zentralblatt MATH is also responsible for essential parts of the portal of the European Mathematical Society, among which ELibM, the largest freely accessible archive of mathematical journals. The database Zentralblatt MATH (ZMATH) contains more than 2.7 million entries drawn from about 3,500 journals and 1,100 serials from 1868 to present. It is the longest-running abstracting and reviewing service in mathematics covering the entire spectrum of mathematics inclusive applications and computer science, mechanics, physics.

Bibliographic databases are the native entry point to search and find mathematical publications. It is a permanent task to improve the quality, the functionalities and information retrieval of our database.

The department of mathematics and computer science of FIZ Karlsruhe has successfully coordinated the EU projects EULER (1998-2002) and the LIMES (2000-2004).

Prof. Dr. Bernd Wegner: Chief editor of Zentralblatt since more than 30 year and emeritus of the Technical University of Berlin. He is actively engaged in the conceptual and organisational development of electronic information and communication in mathematics since many years. He steered several activities and projects in electronic information and communication in the European Union and Germany.

Michael Jost: Long experience in the scientific information systems area, member of the scientific staff at FIZ Karlsruhe, coordinator for development and projects of Zentralblatt MATH, manager of the European Mathematical Information System (EMIS portal site), manager of the EULER project.

Dr. Wolfram Sperber: has worked in several information and communication projects of the mathematical community, project coordinator of Math-Net project and Math&Industry project, coordinating some community-based initiatives in information and communication, e.g. the special interest group information and communication of the Deutsche Mathematiker-Vereinigung and the Information&Communication Initiative of learned societies in Germany.

Partner No 5: Masaryk University Brno: Faculty of Informatics (MU)

Masaryk University in Brno with 30,000 students and 3,000 employees is the second largest university in the Czech Republic. With 9 faculties and a new campus under construction it is steadily growing and strengthening its research potential. Its Faculty of Informatics and Institute of Computer Science set standards of Computer Science technologies and services both locally and internationally. MU is publisher of the journal *Archivum Mathematicum*.

Miroslav Bartošek: Specialist in digital libraries and archives. Head of Library and Information Centre of the Masaryk Univerisity, Brno. Principal investigator or participant in several projects (university research plan "Digital Libraries", ERCIM Technical reports DL, Digital Library of Photographs at Masaryk University, WebArchiv - Infrastructure for Archiving the Czech Web). Participant in the DML-CZ project.

Petr Sojka: Researcher and lecturer (computer typesetting and typography, processing of natural language, text information systems). Technological solutions in the digitisation project for the Jan Otto Encyclopaedia. Technical supervisor of the DML-CZ digitisation program. Chairman of the DML 2008 workshop held July 2008, Birmingham, UK. Participant in the DML-CZ project.

Partner No. 6: Interdisciplinary Centre for Mathematical and Computational Modelling (ICM)

Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University (ICM), is a research centre in computational sciences, focusing in the areas of mathematical, natural and computational sciences, as well as networking and informational technology. It operates the leading scientific supercomputing centre in Poland. The centre has multi-year experience in provision of large scale Internet information and data services since 1994. The early experiences include setting up and maintenance of one of the largest software repositories in Europe (SunSITE in 1995), creation of the first and most popular Polish search engine in partnership with Infoseek Corp. (in 1996) and co-development of the currently most popular Polish search engine (Netsprint). ICM has also had extensive experience in Internet content replication techniques, being a founder and co-organizing a series of yearly International Web Caching and Content Replication conferences, playing an active role in TERENA's caching and indexing task forces, and establishing a countrywide caching hierarchy in Poland in 1998. Since 1995 ICM is running a multi-terabyte national Virtual Library programme (with over 10,000 full text articles being downloaded daily) serving hundreds of scientific and research institutions in Poland. As a subproject, the countrywide Internet library catalogue is being developed, as well as a number of comprehensive databases of Polish academic and research journals. The Virtual Library includes a publicly accessible extensive collection of Polish mathematical journals. ICM coordinates countrywide licensing programs for access to bibliographical databases and for remote software licenses. Different databases and content repositories are being gradually integrated with the use of a custom built, high-performance resource indexing broker system. ICM is one of the partners of the Digital Repository Infrastructure for European Research (DRIVER) project, developing and providing the IT technology for DRIVER core infrastructure, including the indexing service, browsing service, authorisation and authentication service.

Marek Niezgodka: Director of ICM at Warsaw University; since 1989 professor at Warsaw University, has initiated and manages the process of shaping the Polish virtual library of sciences; has managed numerous R&D projects; is a member of various international advisory and editorial boards in the field of applied and computational mathematics; is very active in the field of international cooperation, being an expert of the European Commission, acting as referee for German DFG SFB (Sonderforschungsbereiche) programmes in high-performance computing and modelling areas.

Wojtek Sylwestrzak: Director for Information Technology at ICM, Warsaw University, Poland; currently responsible for the Polish national Virtual Library of Science project. His past experience involves deployment of a number of milestone Polish Internet services since 1994, including data repositories, large scale search engines and distributed systems. His current interests include scalable search systems and distributed data architectures, and he leads the Polish team of the DRIVER Digital Repository Infrastructure for European Research project.

Partner No 7: Instituto de Estudios Documentales sobre Ciencia y Tecnología — IEDCYT (CSIC)

The IEDCYT of the Spanish Council for Scientific Research is a public organisation devoted to encouraging scientific information of quality in every field of knowledge. Its main objectives are to develop research projects in the field of scientific information and documentation, to train specialists and users in information technologies, and to promote the dissemination of the Spanish scientific output. It maintains one of the most important Spanish library of scientific and technical journals, a document delivery centre, a technical and industrial translation service, and several databases and portals.

Rosa de la Viesca: Awarded a higher degree ("Licenciatura") in Physics from the Universidad Complutense de Madrid in 1965, she has always worked in the field of scientific information, participated in several EU projects, was promoted to director of CINDOC (Centro de Información y Documentación Cientifica, now YEDCIT) institute belonging to the National Research Council of Spain (CSIC) in 1983, and since 1998, when she left the management of CINDOC. She was the Spanish representative on the committees of the EU Programmes IMPACT and INFO 2000, has served as vice-chairman of EUSIDIC, Secretary of the FID/II Committee and vice-president of the Intergovernmental Council of the PGI under UNESCO.

Elena Fernández: Has a degree in Chemical engineering. Her professional life other than in Physical Chemistry has been in the documentation field creating bibliographic databases and, lately, in the Internet media. She has created several vertical portals in different issues, and she participates in several European projects. She has published on thesauri, Internet Resources, Electronic publishing, etc.

Ramón Rodriguez: Has a PhD (Universidad de Sevilla, 1984) in Biology. Presently is a Staff Scientist at the Instituto de Estudios Documentales sobre Ciencia y Tecnología, IEDCYT (formerly CINDOC). His previous research interests and career has developed in experimental Biology. He is in charge of several initiatives dealing with scientific communication through the internet, such as Revistas CSIC (http://revistas.csic.es) and e-Revistas (http://erevistas.csic.es). His research interest is mostly focused on the analysis of cybermetrics and 'infobibliometrics' aspects of scientific publishing on the Internet.

Partner No 8: EDP Sciences (EDPS)

EDP Sciences (Édition Diffusion Presse Sciences), a subsidiary of learned societies, works closely with the scientific world. It is involved in the communication and dissemination of science to specialist audiences (researchers, engineers, students etc.) and non-specialist audiences alike (general public, decision makers, teachers…). EDP Sciences produces and publishes international journals, books and Internet sites with a predominantly scientific or technical content (astrophysics, applied and fundamental physics, mathematics, electronics, materials sciences, life sciences, and the medical field).

EDP Sciences offers a specialist publication platform with contents whose quality is validated by international scientific committees and has developed an expertise recognised in the field of processing and distribution of scientific information for the 50 journals online indexed in the main international databases.

EDP Sciences is a founding member of the DOI Foundation as well as a CrossRef Consortium member.

The computer science department of EDP Sciences has developed a journal production workflow based on LaTeX and XML technologies. It set up a platform for online publishing of scientific journals and conference

proceedings. Tools have been developed there in order that the published collections be interoperable with the main scientific databases and reviewing or indexing services. A tool recently developed in the framework of the French e-thesis project “Cyberthèse” is Lxir, an open source software for conversion from LaTeX source to XML/MathML format.

Marie-Louise Chaix: Project officer at the computer science department. Member of the board of the association GUTenberg (French user’s group of TeX).

Jean-Paul Jorda: Coordinator of the developer team at the computer science department. He teaches on structured documents at university Paris X Nanterre (Master Documents électroniques et flux d'information).

Partner No 9: University of Santiago de Compostela: Institute of Mathematics (USC)

The University of Santiago de Compostela is among the five main universities in Spain, with more than 30.000 students, 2200 teachers and 1000 administrative and service staff people. USC offers more than 60 official degrees and incorporates 300 research groups in 80 buildings covering one million square meters in two campuses. The Institute of Mathematics has participated in European projects like LIMES (Large Infrastructures in Mathematics) and MACSI-Net (Mathematics, Computing and Simulation for Industry).

Enrique Macias-Virgós: Doctor in Mathematics, he was the dean of the Faculty of Mathematics for seven years and vice-president of the Spanish Mathematical Society RSME. Presently he is the chairman of the Spanish Commission on Electronic Information and Communication, depending from the IMU Spanish Committee (CEMAT). He is also a member of the Electronic Publishing Committee of the European Mathematical Society. He is one of the organizers of the Spanish digitization project DML-E.

Felipe Gago: Doctor in Mathematics, is a staff member of the Spanish Mathematical Royal Society and participated in the LIMES project (Large Infrastructures in Mathematics, Enhanced Services).

Partner No 10: Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (IMI-BAS)

The institute is the main mathematical institution in Bulgaria. It was established in 1947. Since than it has been a leading Bulgarian centre for research and training of highly qualified specialists and exercising an efficient, long-range, consistent policy related to the fundamental trends in the development of mathematics, computer science and information technologies. IMI has a total staff of 238 (175 researchers, including 105 full and associate professors) in 21 departments.

The Institute has the richest mathematical library in the country and is publisher of the two of the main Bulgarian mathematical journals. Besides various activities related to digital born mathematical texts, at the institute a core group has been created working on digitisation of cultural heritage and dealing with digitisation of historical monuments and manuscripts. These activities have led to the establishment of the new “Digitisation of Scientific Heritage” department. The institute disposes of the adequate infrastructure, a leading role in the Bulgarian mathematical community, as well as the experience on several digitisation projects.

Radoslav D. Pavlov: Associate Professor at IMI-BAS (1979), PhD (Informatics, 1977), Deputy director of IMI-BAS, Head of the Department of Computational Linguistics at IMI – BAS (1984-), Vice-President of Alliance for Strategies and Development of Information Society (Bulgaria). His major research interests are in the fields of Human Language Technologies, Information Society Technologies, Knowledge Technologies and Management, Semantic WEB services, Semantic Information Processing, Digital Libraries and Content Management Systems, Algorithmics. In the latest years he was the site leader of 9 national and international projects.

Julian P. Revalski: Since 2001 full professor at the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences, 2000 – 2003 co-responsible involved into several international projects, 2001 – 2003 responsible for a team of about 25 persons from IMI-BAS and Sofia University that digitised 13 volumes of the "Jahrbuch über die Fortschritte der Mathematik" within a cooperation with SUB Göttingen and the Technical University Berlin, since 2003 member of the Advisory Committee of the Committee on Electronic Information and Communication of the International Mathematical Union (CEIC).

Peter L. Stanchev: Full Professor at IMI-BAS (2001), DSc (1998), Chair of the Information Research Department (since 2005), Professor at Kettering University (Since 2002). His major research interests are in the fields of Image Processing, Multimedia Database Systems, Data Base Systems, Information Systems, Expert Systems, Fuzzy Sets and Systems, Decision Making and Web Technologies. Member of IEEE, IFIP, the Association for Computing Machinery, the International Society for Computers and Their Applications, the International Association of Science and Technology for Development.

Partner No 11: Matematický ústav AV ČR, v. v. i. (IMAS)

Matematický ústav AV ČR, v. v. i. (Institute of Mathematics AS CR, public research institution) is a research institute in the Academy of Sciences of the Czech Republic. Its principal activities are scientific research in mathematics and its applications, and providing infrastructure for research. Since 1996, the IMAS has been running the Prague editorial unit for Zentralblatt MATH. It also coordinates development of the Czech Digital Mathematics Library (DML-CZ). The pilot part of the DML-CZ will be launched by the end of 2009, it will contain the major part of relevant mathematical literature ever published in the Czech lands. IMAS as the operator of the DML-CZ will provide its content to the EuDML and contribute the OCR facilities and tools for retrieval and enhancement of metadata. IMAS is also the publisher of three research journals included in the DML-CZ.

Jiří Rákosník: Mathematician, Deputy Director of IMAS and vice-president of the Czech Mathematical Society. Founder and head of the Prague editorial unit for Zentralblatt MATH and the initiator and coordinator of the project DML-CZ with the responsibility for handling metadata from mathematician's point of view. Member of the Electronic Publication Committee of the European Mathematical Society.

Partner No 12: Ionian University: Department of Informatics (IU)

The Department of Informatics was founded by the Ministry of National Education and Religious Affairs in 2004 and its scope covers Theoretical as well as Applied Informatics. It is located in the city of Corfu, Greece. The mission of the department is to advance scientific and research activities not already covered by existing university departments throughout Greece. As a result, its function is complementary to that of similar departments in Greece and is clearly focused on novel applications in the areas of Information Systems and Humanistic Informatics. Information Systems, on the one hand, is a crucial factor regarding production, services and management in enterprises nowadays. Humanistic Informatics, on the other hand, addresses the needs for education and research within the area of Informatics, and tones in with the nature of the Ionian University.

Professor Vassileios Chrissikopoulos is the Head of the Department of Informatics and Vice-Rector of the Ionian University. He holds a Diploma in Mathematics from Aristoteleio University of Thessaloniki, Greece, an MSc from the University of Warwick, an MSc from the University of Chelsea and a PhD from Royal Holloway College of London. He was Professor at the Department of Archival and Library Sciences of the Ionian University and his research interests include mathematical digital libraries.

Dr. Panagiotis Vlamos is an Assistant Professor at the Department of Informatics of the Ionian University. He holds a Diploma in Mathematics from the University of Athens and a Ph.D. in Mathematics from the Department of Mathematics, National Technical University of Athens. His research interests include mathematical modeling, pattern recognition, discrete mathematics and bioinformatics. He is the scientific responsible of the Hellenic Mathematical Society in the LIMES project and the editor in chief of the Hellenic Mathematical Bibliographic Database.

Dr Kostas Anagnostou is an Adjunct Lecturer at the Department of Informatics. He holds a BEng in Computer Engineering and Informatics from the University of Patras, an MSc in Information Systems Engineering from University of Manchester and a PhD in Computer Science from the University of Warwick. His research interests include videogame technologies, videogames in education, computer graphics, and image processing algorithms.

Eleni Christopoulou is an Adjunct Lecturer at the Department of Informatics of the Ionian University. She holds a Diploma and an MSc from Department of Computer Engineering and Informatics, University of Patras. Her thesis entitled "An evaluation framework for ontology languages and building tools" and she has

participated in a number of EU funded research projects working on ontology-based representation of data and context. Her research interests include ontologies, semantic web and context-aware systems.

Partner No. 13: Made Media Ltd (MML)

Made Media Ltd. is a digital media agency with a focus on user experience design, database-driven web applications and social media. The company employs ten full-time staff including user interface designers, website developers and server technicians. Made Media's staff are experts in the open-source LAMP and Ruby-on-rails MVC frameworks and take a specific interest in content management, web standards and interoperable web services. Recent projects include the BAFTA award winning 'Embarrassing Bodies' website for Channel 4, a collaborative document repository with dublin-core metadata for the Association for European Transport, and a collaborative technologies database for Euroscan, a network of European institutions researching new medical technologies.

Jake Grimley has been working in new media since graduating with a BSc in Physics in 1996. He has delivered new media projects for clients including the BBC, Channel 4, Discovery Channel, Goodyear, BMW and McCann Erickson. A 'creative developer', Jake held the Creative Director's position at his first company, before moving into a Developer's role, and setting up Made Media Ltd in 2003. He has developed open source software projects, including a PHP implementation of the Active Record ORM pattern, but now consults on the overall user experience design for his clients.

Stefan Lewandowski is a creative web entrepreneur with a keen interest in social media and new models of interaction. Stef has been working the web since 1999, founding the award-winning digital media agency 3form, producing projects for clients including Vivienne Westwood, the V&A and Lastminute.com. Stef has since set-up web 2.0 projects including 'Odadeo' the community site for fathers and 'Help me Investigate' the web's answer to local investigative journalism. Stef is a proponent of the Ruby on Rails MVC stack, and open web services like Openid, Opensocial and REST APIs. He brings a deep understanding of the cutting-edge of web development.

Partner No. 14: Centre National de la Recherche Scientifique / Cellule MathDoc (CNRS/CMD)

Cellule MathDoc is a joint service unit both of the Centre National de la Recherche Scientifique and of Université Joseph Fourier (Grenoble). It is devoted to mathematical documentation and in charge of the NUMDAM program, and other projects relevant to EuDML which have been set up by UJF and CNRS personnel: CEDRAM, mini-DML, Gallica-MATH, Bourbaki archive, etc. Main CNRS staff at CMD for this project are:

Yves Laurent is mathematician, directeur de recherches at Centre national de la recherche scientifique (CNRS) and director of Cellule MathDoc, member of the steering committee of EMANI, member of the coordination committee of Zentralblatt MATH.

Catherine Barbe-Zoppis is senior engineer in computer science specialised in databases, metadata and web services. She has developed several critical conversion, and quality insurance tools used in NUMDAM’s production.

Associated organization No. 1: European Mathematical Society (EMS)

The project will be undertaken under the auspices of the EMS representing mathematical societies in Europe. The society will grant use of the Electronic Library of Mathematics (ElibM project) which is supervised by its electronic publication committee and will nominate the chair of the EuDML scientific advisory board.

Associated organization No. 2: Göttingen University–State and University Library Göttingen (SUBGoe)

The State and University Library Göttingen is one of the largest University and Research Libraries in Germany and has a strong focus on ancient and modern mathematics. It has been involved into several national, international, and particularly European research projects dealing with digitisation, metadata, interoperability, and web services. In particular, SUB Göttingen is a partner in the Electronic Mathematical Archiving Network Initiative (EMANI)—an international collaboration with libraries, Zentralblatt MATH and

Springer publishing house. SUB Göttingen's digitisation centre (GDZ) was a key partner in the DFG supported ERAM project which digitised the *Jahrbuch* and many mathematical journals of that period. It holds the largest European collection of digitised mathematical books and journal articles.

Associated project: RusDML (Russian Digital Mathematics Library)

Participants of the EuDML project were also involved in the RusDML (Russian Digital Mathematics Library, http://www.rusdml.de/) project. This is a Russian-German project for establishing a digital archive of Russian mathematical publications (expressing a huge tradition of mathematics) and also part of the global effort to make all mathematical literature digitally available to mathematicians around the world. The project's experiences with respect to languages and scripts will provide a valuable basis for the integration of additional European languages.

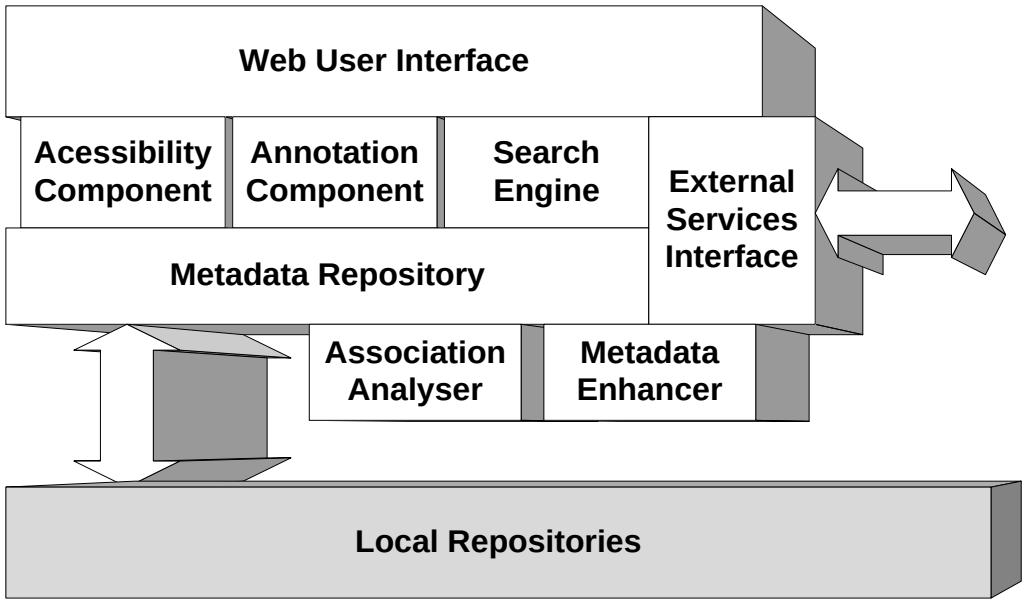
B3.2a. Chosen approach

Introduction and general description

There will be two type of networking activities. One is technical, consisting mostly in setting up the interoperability of the various resources that we shall integrate (standardisation of metadata). The other is tight networking among the partners, dealing with policy, exploitation plan, advocacy and awareness; it will be monitored by the scientific advisory board. See work packages 2 and 3.

There will be one central service activity, which will provide a one-stop access to the whole resource. It will start with basic services for exploring the collections — manually as well as automatically — and will be enhanced during the project's lifetime.

As soon as the project starts, we will implement novel techniques in order to enhance the performance of the service through augmentation of the available metadata when it does not meet the state-of-the-art, and associating related resources. Notice that, as existing mathematical writings cannot be trustfully represented by text-based formats, we consider throughout this text textual, i.e. XML (possibly with MathML formulas) versions of an item's full text a metadata, the raw data being typically a graphical format such as PDF. We will use small test beds from some of the projects to assess and integrate these add-ons to the core system. We will deploy them widely by end of year 1.



Pilot System Architecture

The diagram above presents the consortium's current architectural view of the intended pilot system, subject to review and refinement during the early stages of the project. Each box represents a component of the system, the block arrows represent distributed data flow within the system. Different local repositories can be, but are not required to be, replicated. However, the rest of the system is shared and/or replicated.

At the core is the **Metadata Repository** , which manages a replica of all the metadata for all the items in the different local repositories. It does not contain actual object (PDF) items, which remain in the local repositories (in this way we avoid having to deal with IPR issues over these, leaving the control of that to the local repositories). The Metadata Repository manages the metadata updates from the local repositories. It will support adapting the different forms of metadata that each provider has about their items to the common format required by the EuDML. To support those processes, the Metadata Repository also will include, in its internal architecture, a Metadata Registry.

The **Search Engine** provides search and item identifier resolution facilities. The Metadata Repository mediates with the local repositories to fetch actual items for the higher levels of the system.

Some items in the local repositories may not yet be freely accessible, but may need enhancement of their metadata. Hence a facility is planned for content providers to invoke the **Metadata Enhancer** directly on their privileged access items via the **External Services Interface**, to enhance the metadata before uploading.

At the bottom is the collection of **Local Repositories**. Each content provider maintains its own collection of items, which can be supplied on demand. Such requests are via identifiers, i.e. the local repositories are not required to provide search or browse facilities, but merely to manage their collections of items, provide access via identifiers and manage right of access issues corresponding the to moving wall open access policy adopted by the consortium.

The **Metadata Enhancer** is a component that can identify items for which it might be able to enhance the existing metadata, including fetching the item and analyse it (e.g. using current OCR, math formula recognition, keyword extraction, signature file construction, or bibliography analysis technologies). This component will actually consist of a range of different tools, that can be improved and extended over time, and that can handle different aspects of individual item metadata enhancement.

The **Association Analyser** works on metadata for sets of different items, rather than, as in the case of the Metadata Enhancer, on single items. It identifies related items from the metadata records and updates the metadata accordingly so that various links between documents, or links to relevant external resources, can be recorded in the Metadata Repository.

The **Annotation Component** provides mechanisms to attach new material to individual items in the repositories and maintain this new material. It is envisaged that this will support community interactivity with the collection by allowing users to add, view and update their own reviews, tutorials, comments, or recommendations to each item.

The **Accessibility Component** provides support for enhancing accessibility of items, if required, before presentation to end users. For visually impaired or dyslexic users, this will involve speech synthesis for text or for mathematical formulae (via MathML annotation), large print re-formatting or OCR facilities to make scanned image items accessible to Braille readers. Automatic translation facilities will be provided for texts in a language not understood by the user.

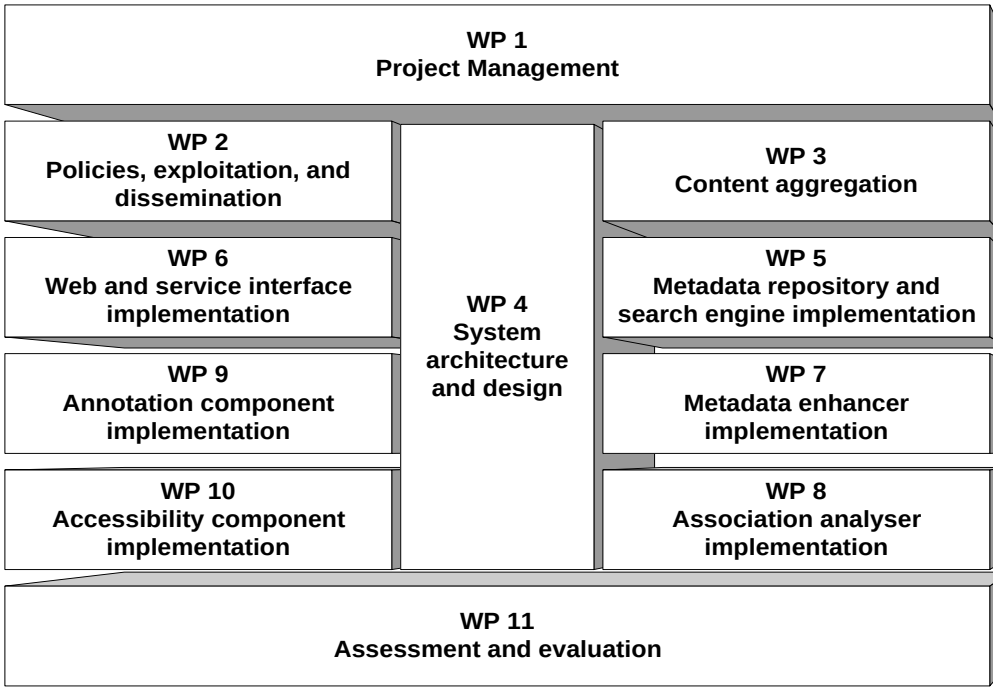
The **Web User Interface** and the **External Services Interface** provide public access to the system both via web browsers and by web services. This comprises bidirectional interoperability with external resources such as the Math Genealogy, MacTutor History of Mathematics, Mathworld, Wikipedia, EqWorld and other mathematical and science related web resources. Interoperability with Europeana will be also considered at this level.

Work package overview

In order to make the distributed content more **exploitable**, we shall aggregate a central metadata repository and endow each item with a persistent identifier and an associated resolution system pointing back to the

original resource. For this purpose we will reuse the framework developed in the scope of the project TELplus, developed by our partner IST.

All the added-value services that will match the work programme objectives for our specific content will thus be built on top of this basic platform. This is handled by WP3 ("Content Aggregation"), where we will define a common metadata schema, and export from each participating catalogue to this format. This work package will, as soon as possible, start to produce the data on which a first instance of the services will be designed (WP4 – "System architecture and design"), and built (WP5 – "Metadata repository and search engine implementation"). These activities will have a second round after assessment of the implemented prototype and will have achieved their principal task when all content partners will be able to export their catalogue to the agreed format which will be exploitable in the central system. WP3 will continue until the end of the project, serving as technical assistance to the content providers regarding their metadata and as an entry point to the project for newcomers.



In parallel, tight networking activities regarding longer term policy and organizational matters, as well as dissemination of results, will be performed in WP2 ("Policies, exploitation, and dissemination"). This activity extends to all stakeholders (projects partners, users, publishers as prospective content providers, policy makers, etc.). It will take advantage of the variety of stakeholders represented by our consortium's partners, aiming at defining a reasonable archiving policy securing **initial open access** to most of the mathematical scholarly content and **eventual open access** to the remainder after a delay of a few years.

For **availability** and **convenience**, EuDML is accessed via a web interface for human users, and a web service interface for tools and systems (WP6). For better **community involvement** and **enhanced interactivity**, we add Web 2.0 features of user annotation (WP9) and interoperability.

To make our content even more **accessible**, we provide explicit support, using the latest technologies, for visually impaired or dyslexic users as well as automatic language translation support in WP10.

Since the existing metadata is heterogeneous, often sparse and monolingual, we cannot meet the expectations of the work programme in terms of user functionalities unless we use all existing technologies at hand in order to augment the metadata to a minimal level of quality among all integrated collections. This will be an iterative, ongoing process using matching techniques to get some more metadata from available sources, OCR when applicable, exploitation of mathematical content, identification of citations, and generation of internal links between components of the project.

To make our content more **usable**, we will exploit the fact that these collections' content is heavily mathematical in nature, so that mathematical knowledge management techniques can be applied to overcome language barriers and connect various items related by their subject, supporting a new **paradigm** of

mathematical literature search and discovery that surpasses the pure text based and “Google-style” search paradigm prevalent today. These tasks will be performed jointly as WP7 and WP8, to provide support for enhancing metadata and identifying interesting links between items. They will both start at the beginning of the project, first tuning the existing technology provided to the project by its partners on a restricted set of collections which are already eligible to the workflow. After this, each work package will exploit the output of the other in order to consolidate the quality of the contributed metadata and attain production quality of the whole service before completion of the project.

The first work package will concern the management activity of the project during its whole life cycle. WP11 will monitor the technical quality of the project, and evaluate its output compared to the expectations, using various indicators developed for the purpose and user surveys.