MDML: The Mathdoc Digital Mathematics Library

Alexandre Bouquet Thierry Bouche

Univ. Grenoble Alpes, CNRS, CMD, 38000 Grenoble, France

Abstract

Following the steps of previous projects such as EuDML, Mathdoc is launching its Digital Mathematics Library. Based on a reliable infrastructure made for Numdam, learning from previous projects, and relying on a network of institutions we trust, we aim to push the ball further for accessing mathematical content online. We focus for a start on the aggregation part, aiming to reach a critical mass of mathematical content by harvesting various sources: OJS instances, preprint repositories, and locals DMLs. We thus build a database of mathematical documents, linking back to the source's website for accessing content.

1 Introduction

With the global progress of technology, most of scientific papers are now online somewhere. In the beginning of the 2000's mathematicians started to dream of gathering all digital papers in a single database, pushing the services of a traditional library to a global scale, taking advantage of the digital paradigm: cataloguing and providing access to the mathematical knowledge of all times [3]. This would form the (Global) Digital Mathematics Library (DML).

Since then, a lot of projects emerged, contributing to this dream. The first step was the development of locals DMLs, often nationwide, bringing mathematical content from a country in the same place. Most of these had a digitisation project at their core, but many have also managed to arrange updates from publishers for born-digital content. Some of these projects are still actively growing while others are completed. The number of sources for digital mathematics has exploded but it is still difficult to locate a relevant item.

A first attempt at integration, meant as a proof-of-concept, was started at Mathdoc back in 2004, the mini-DML project [1]. This was followed up at a much larger scale by the EuDML project, partially funded by the European Commission during the years 2010-2013 [8]. This project has defined policy and standards to set up a Europewide network of technical and content partners.

Thus, aware of previous project's boundaries and successes, we started a new project, aiming at pushing the ball further (the *M* prefix in MDML can also be interpreted as *medium*, with an intended target of getting *Mega*).

This article presents the Mathdoc DML, the choices we made and the goals we set, and the software architecture and methods we adopted. It also announces the release of our first Web interface at `dml.mathdoc.fr`.

2 Goals and perimeter

Basically the goal is to make a big part of the mathematical corpus available from the same place, with the best possible metadata to facilitate searching, and interoperate with relevant infrastructures. The system is based on

- an OAI-PMH harvester to gather metadata;

In: C. Kaliszyk, E. Brady, J. Davenport, W.M. Farmer, A. Kohlhase, M. Kohlhase, D. Müller, K. Pąk, and C. Sacerdoti Coen (eds.): Joint Proceedings of the FMM and LML Workshops, Doctoral Program and Work in Progress at the Conference on Intelligent Computer Mathematics 2019 co-located with the 12th Conference on Intelligent Computer Mathematics (CICM 2019), Prague, Czech Republic, July 8–12, 2019, published at http://ceur-ws.org

- a periodic task orchestrator;
- the new Numdam platform [2] to provide the core XML parser, and the searching and browsing interface.

As the project started a few months ago, we intend to first reach a critical mass in indexed content with a highly usable interface, so the focus is currently on the aggregation part. Later on, we will take an incremental approach to improve our DML services.

Compared to mini-DML, which was a proof-of-concept, we try to use much more detailed metadata in order to offer a better user experience, we also started to harvest from much more sources, and to adapt to more common metadata schemas. Compared to EuDML, we use more or less the same metadata, we do not have yet an API, we won't try to revive some of its features. The main benefit of our work is to move ahead: an entirely new technology behind, a worldwide scope. Our main target audience is the working mathematician, always struggling to find a source for published references they gather from database searches, citations or colleagues' hints. We try to make it easy to search a still highly fragmented, heterogeneous corpus.

2.1 Choosing data source

The first thing to do for building a DML is to choose where the data comes from. There are quite a lot of available resources on the internet, and thus we must choose on which criterion we base our choice. We decided to stand on EuDML's shoulders, which means we intend to aggregate from local DMLs that ensure

- quality of the mathematical content;
- long-term reliability (well-maintained systems with persistent URLs);
- a usable OAI-PMH server delivering quality metadata (JATS/BITS if possible, or at least fine-grain enough to enable a decent browsing of collections).

We thus rely on a network of institutions we trust, with a common goal of archiving and broadcasting mathematical content, with sustainability rather than profit in mind.

Following on mini-DML, we decided to also include preprint servers such as arXiv or HAL, because they provide open access to a huge quantity of useful mathematics. It will be possible to filter search results so as to exclude preprints, when the user is looking for formally published material only[1].

In order to maximize the number of sources, we also started to ingest content from isolated journals published with Open Journal System (OJS), when we believe that they are backed up by a trusted institution, such as a learned society or a University library. OJS instances are now shipped with an OAI server, and a JATS plugin is available, so a lot of quality items are available through this method. The challenge here will be to draw an inventory of all relevant OJS instances, and to select which are eligible with our criteria.

EuDML has done a great job of ingesting data sources with still no support of OAI protocol to this day, thus we use EuDML's OAI server to retrieve some of its content, in order to take advantage of the former project rather than spending time reproducing what has been done. We avoid it when it's possible though, because we want regular updates from our chosen sources, and EuDML is currently stalled. We intend to harvest anew currently frozen sources when possible.

When we have defined the source, the next step is to import the data in our database. The goal here is not to store locally a copy of the full text of articles, but to facilitate the search and link to the source for accessing the content. We build thus more a catalog than a physical library, making the choice of the source more important as we rely on it to provide content.

Our first goal is thus to break the (quite artificial) "EU" barrier in EuDML. In this first round, we leave out the more fancy stuff as we focus on making more content visible.

2.2 Importing data

As outlined above, we select sources that support the OAI-PMH protocol [7]. This protocol makes it possible to retrieve easily metadata of mathematical items, with an explicit XML schema. EuDML set a flavor of JATS [6] as its internal format and, as other EuDML partners, we adopted it afterwords as our internal format for

[1] When arXiv or other preprint repositories will have explicit metadata for identifying postprints ("author accepted manuscript" or "version of record": content identical to the published version), these will be considered acceptable alternatives to publisher's version.

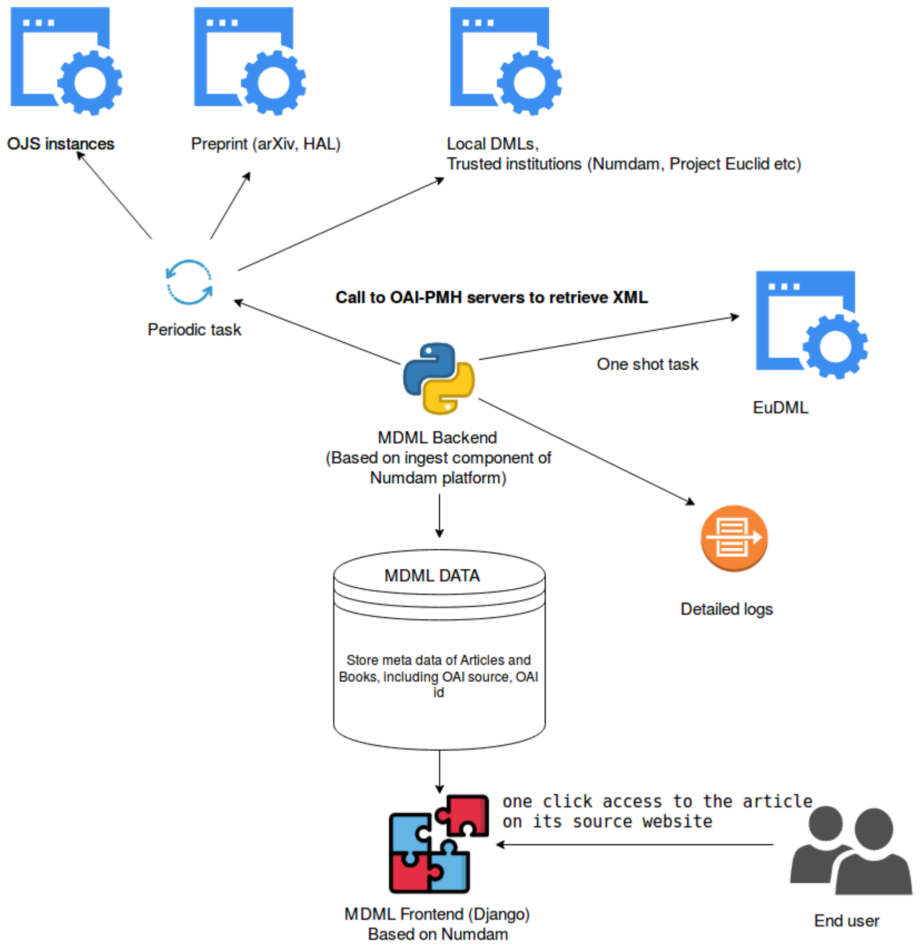

OJS instances
Preprint (arXiv, HAL)
Local DMLs,
Trusted institutions (Numdam, Project Euclid etc)
Periodic task
Call to OAI-PMH servers to retrieve XML
One shot task
EuDML
MDML Backend
(Based on ingest component of
Numdam platform)
Detailed logs
MDML DATA
Store meta data of Articles and Books, including OAI source, OAI id
one click access to the article
on its source website
MDML Frontend (Django)
Based on Numdam
End user

Figure 1: Workflow in MDML

document-oriented projects such as Numdam[2], Centre Mersenne[3], etc. We like it because it is very exhaustive and well-structured. We also like it because it is meant for the kind of content we deal with, including native support for MathML expressions, and the ability to encode up to the full text. However, we also import articles from Dublin Core when this is the only format available, because it represents a big part of resources available through OAI. However only basic metadata are retrieved with this format. We import books in BITS format [5].

As said above, this is the first step for importing data, we are prepared to support more formats and other protocols over time.

Once harvested as or converted to JATS and BITS, we clean somehow the metadata and ingest it in our platform, which is based on the one described in [2] which has now been used for large documents sets for different projects (Numdam, Centre Mersenne). Most of the database structure to store the metadata is based on the platform's existing one. This core system will continue to improve with the evolution of the multiple projects of Mathdoc including MDML.

An other big advantage of the OAI Protocol is the possibility to choose in which date interval we want to import the data, thus making the import of new data easy, and avoiding the cost of ingesting the same data again and again. Moreover, it allows us to set up automated regular update over all of our sources, bringing new content automatically on MDML. Although scheduled, this never really happened with EuDML harvester.

We also set up a log system to keep track of import, storing raw data and the source, making it possible to understand why it crashed, and how to fix it.

When we have stored the data, the next step is to present the data back to the users, and make it possible to browse the digital library, in the same way a user can wander in a physical library and browse printed volumes.

[2] http://www.numdam.org
[3] http://www.centre-mersenne.org

2.3 Browsing data

Searching among digital items can be done in different manner: searching authors, keywords, title, equations or browsing specific journals or books. To be able to propose an fielded search, we need to have thorough metadata. However, the most important thing is to have a lot of items available, if we really want added value.

The implementation of searching in MDML is based on the Numdam platform, thus benefiting from an already proven tool. The searching engine is getting better over time because its core is common for Numdam and all journals managed by Centre Mersenne.

3 Technical/implementation details

3.1 Backend

As the MDML project is tightly linked with the Numdam platform written in Python, it seemed obvious to use Python as well for MDML. To harvest and retrieve the XML from OAI sources, we use the great Sickle plugin[4]. We store different information about the source to harvest: OAI server url, OAI set, XML format, if it is a one shot ingestion or not (i.e. needs update or not), the type of provider and the last harvest date. Moreover, we also store what kind of processing we will do. The tricky part is that almost every source of data is different in some way, regardless of format used (JATS, Dublin Core, etc). Even if it's a minor difference, we need to have a system with a common processing and specify only the small part specific to each source. Thus we made a sort of multiplexer of XML parsers, depending on OAI server. Each source parser inherits from the common XML parser, and we override what's different.

Then we feed it to the Numdam based platform, with an additional layer to store OAI metadata such as OAI id and OAI source.

A recurrent task has been set up in the background with Celery[5], to check regularly for new data. As said earlier, OAI-PMH allows us to specify a date interval for harvesting, so the last harvest date is stored for every source of MDML, and we update it at each new harvest.

Detailed logs are stored if any of the items harvested failed to be ingested by the platform, including raw XML and source information, then allowing us to enhance our import tasks quickly, and to do the import again.

3.2 Frontend

The website is based on Django, same as Numdam and sites managed by Centre Mersenne. There is of course an additional layer as well to serve items on MDML website, but the core is common and can benefit from the work done by Mathdoc on Numdam. For instance, the platform natively supports a dual TeX/MathML description for mathematical content, with mathjax on board in order to present it correctly in most situations. The end goal here is a one click access to the article on its source website.

4 Conclusion and perspective

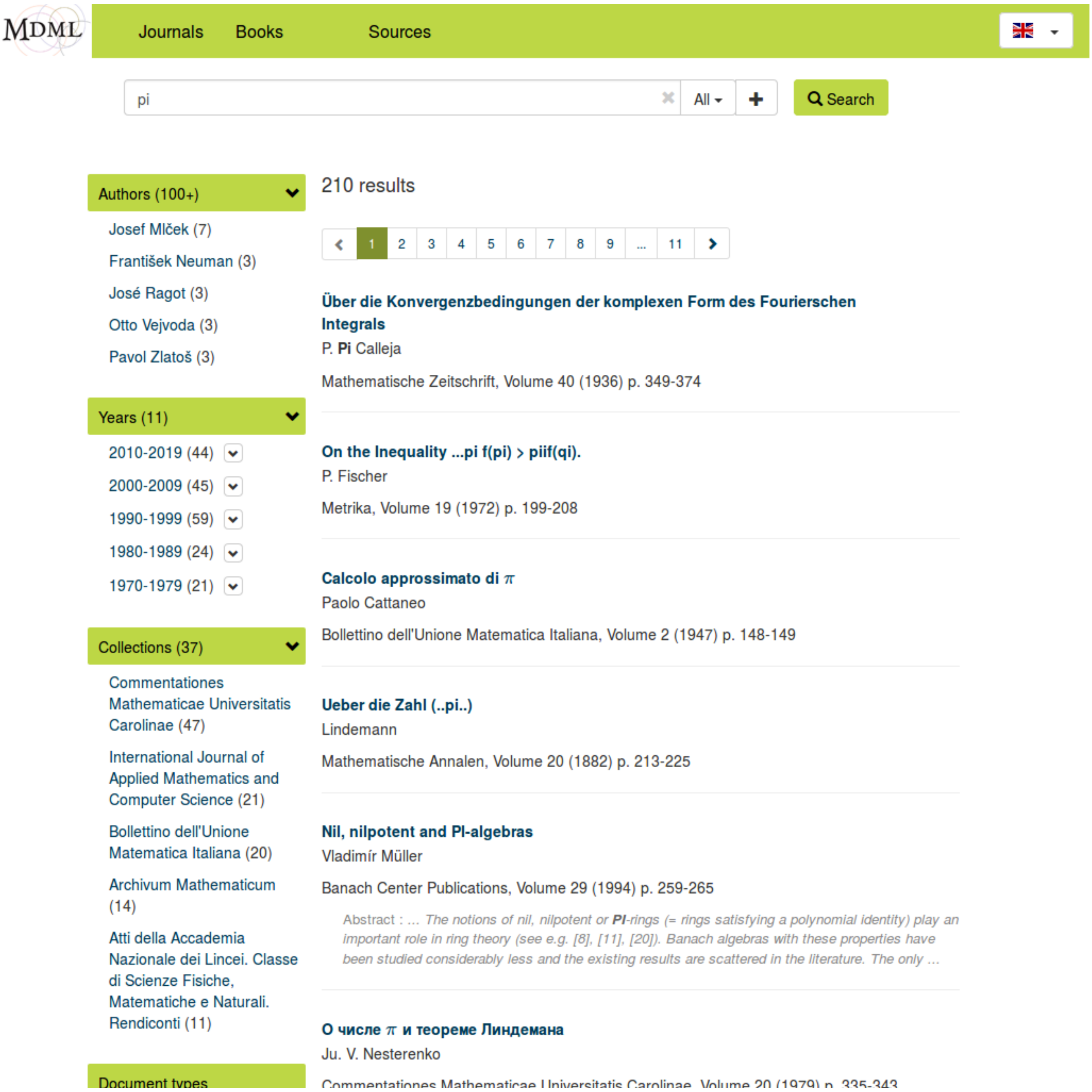
Based on the experience of Mathdoc and its various projects in the area of mathematical documents and metadata, and all previous DML projects, the ambition is that the Mathdoc DML be a significant step forward in terms of content covered towards the Global DML [4] supported by the IMU and the newly founded International Mathematical Knowledge Trust. We choose to have an incremental approach, and to set up a solid foundation based on the production-ready platform maintained by Mathdoc. The project will evolve over time, and there is still a lot of work to be done. The number of items will grow by itself as new content is published at the sources we harvest, and new sources will be regularly added. The quality of the search engine, browsing and metadata displayed will also improve over time, alongside Numdam and Centre Mersenne's websites. In the end, we hope to provide a DML with a great deal of items, and thorough metadata to be able to browse seamlessly mathematical content.

References

[1] Thierry Bouche. Introducing the mini-DML project. In Hans Becker, Kari Stange, and Bernd Wegner, editors, *New developments in electronic publishing, AMS/SMM Special Session, Houston, May 2004, ECM4 Satellite Conference, Stockholm, June 2004*, pages 19–29. FIZ Karlsruhe / Zentralblatt MATH, 2005.

[4] https://github.com/mloesch/sickle

[5] http://www.celeryproject.org/



Josef Mlček (7)
František Neuman (3)
José Ragot (3)
Otto Vejvoda (3)
Pavol Zlatoš (3)

2010-2019 (44)
2000-2009 (45)
1990-1999 (59)
1980-1989 (24)
1970-1979 (21)

Commentationes Mathematicae Universitatis Carolinae (47)
International Journal of Applied Mathematics and Computer Science (21)
Bollettino dell'Unione Matematica Italiana (20)
Archivum Mathematicum (14)
Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti (11)

210 results

‹ 1 2 3 4 5 6 7 8 9 ... 11 ›

Über die Konvergenzbedingungen der komplexen Form des Fourierschen Integrals
P. **Pi** Calleja
Mathematische Zeitschrift, Volume 40 (1936) p. 349-374

On the Inequality ...pi f(pi) > piif(qi).
P. Fischer
Metrika, Volume 19 (1972) p. 199-208

Calcolo approssimato di π
Paolo Cattaneo
Bollettino dell'Unione Matematica Italiana, Volume 2 (1947) p. 148-149

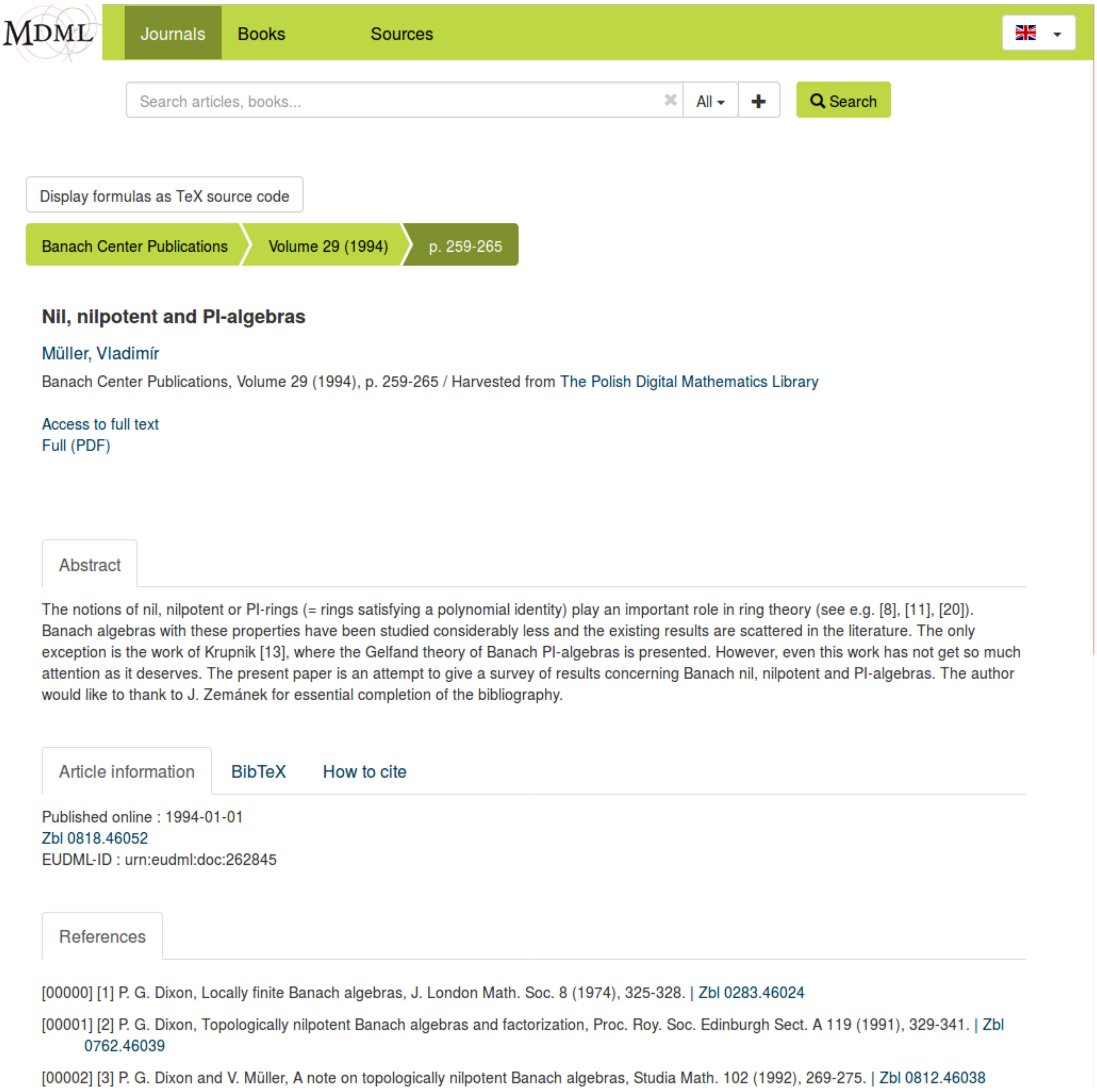
Ueber die Zahl (..pi..)
Lindemann
Mathematische Annalen, Volume 20 (1882) p. 213-225

Nil, nilpotent and PI-algebras
Vladimír Müller
Banach Center Publications, Volume 29 (1994) p. 259-265

Abstract : ... *The notions of nil, nilpotent or **PI**-rings (= rings satisfying a polynomial identity) play an important role in ring theory (see e.g. [8], [11], [20]). Banach algebras with these properties have been studied considerably less and the existing results are scattered in the literature. The only ...*

О числе π и теореме Линдемана
Ju. V. Nesterenko
Commentationes Mathematicae Universitatis Carolinae, Volume 20 (1979) p. 335-343

Figure 2: Searching for items in MDML


MDML
Journals
Books
Sources
Search articles, books...
All
Search
Display formulas as TeX source code
Banach Center Publications
Volume 29 (1994)
p. 259-265

Nil, nilpotent and PI-algebras

Müller, Vladimír

Banach Center Publications, Volume 29 (1994), p. 259-265 / Harvested from The Polish Digital Mathematics Library

Access to full text
Full (PDF)

Abstract

The notions of nil, nilpotent or PI-rings (= rings satisfying a polynomial identity) play an important role in ring theory (see e.g. [8], [11], [20]). Banach algebras with these properties have been studied considerably less and the existing results are scattered in the literature. The only exception is the work of Krupnik [13], where the Gelfand theory of Banach PI-algebras is presented. However, even this work has not get so much attention as it deserves. The present paper is an attempt to give a survey of results concerning Banach nil, nilpotent and PI-algebras. The author would like to thank to J. Zemánek for essential completion of the bibliography.

Article information | BibTeX | How to cite

Published online : 1994-01-01
Zbl 0818.46052
EUDML-ID : urn:eudml:doc:262845

References

[00000] [1] P. G. Dixon, Locally finite Banach algebras, J. London Math. Soc. 8 (1974), 325-328. | Zbl 0283.46024

[00001] [2] P. G. Dixon, Topologically nilpotent Banach algebras and factorization, Proc. Roy. Soc. Edinburgh Sect. A 119 (1991), 329-341. | Zbl 0762.46039

[00002] [3] P. G. Dixon and V. Müller, A note on topologically nilpotent Banach algebras, Studia Math. 102 (1992), 269-275. | Zbl 0812.46038

Figure 3: One article's details on MDML

[2] Thierry Bouche and Olivier Labbe. The new Numdam platform. In Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke, editors, *Intelligent Computer Mathematics Proceedings of the 10th International Conference, CICM 2017, Edinburgh, UK, July 17-21, 2017*, number 10383 in Lecture Notes in Computer Science, pages 70–82. Springer, 2017. also available at http://doi.org/10.5281/zenodo.581405xs.

[3] John Ewing. Twenty Centuries of Mathematics: Digitizing and Disseminating the Past Mathematical Literature. *Notices of the AMS*, 49(7):771–777, 08 2002.

[4] Patrick D. F. Ion and Stephen M. Watt. The Global Digital Mathematics Library and the International Mathematical Knowledge Trust. In Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke, editors, *Intelligent Computer Mathematics 10th International Conference, CICM 2017, Edinburgh, UK, 2017, Proceedings*. Springer, 2017.

[5] National Center for Biotechnology Information, U.S. National Library of Medicine. Book interchange tag set: JATS extension, version 2.0, February 2016. Full online documentation at https://jats.nlm.nih.gov/extensions/bits/.

[6] National Center for Biotechnology Information, U.S. National Library of Medicine. Journal archiving and interchange tag library, NISO JATS version 1.2, January 2019. Full online documentation at https://jats.nlm.nih.gov/index.html.

[7] Open Archives Initiative. Protocol for Metadata Harvesting. Documentation at http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

[8] Wojtek Sylwestrzak, José Borbinha, Thierry Bouche, Aleksander Nowiński, and Petr Sojka. EuDML—Towards the European Digital Mathematics Library. In Petr Sojka, editor, *Proceedings of DML 2010*, pages 11–24, Brno, July 2010. Masaryk University. http://dml.cz/dmlcz/702569.