Katarzyna Zamlynska; Lukasz Bolikowski; Tomasz Rosiek

Migration of the Mathematical Collection of Polish Virtual Library of Science to the YADDA Platform

In: Petr Sojka (ed.): Towards Digital Mathematics Library. Birmingham, United Kingdom, July 27th, 2008. Masaryk University, Brno, 2008. pp. 127--130.

Persistent URL: http://dml.cz/dmlcz/702538

Migration of the Mathematical Collection of Polish Virtual Library of Science to the YADDA Platform

Katarzyna Zamlynska, Lukasz Bolikowski, and Tomasz Rosiek

Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw, ul Pawinskiego 5a, 02-106 Warsaw, Poland
E-mail: kzaml@icm.edu.pl, bolo@icm.edu.pl, tzr@icm.edu.pl
URL: http://www.icm.edu.pl/

Abstract. YADDA framework facilitates information exchange between digital document repositories. YaddaWeb, its web-based interface, provides browse and search functionalities. Content providers use DeskLight application to add or modify metadata and content. Internally, YADDA contains flexible repository aggregation mechanisms, multiple hierarchy support and full-text indexing capabilities. YADDA framework is an excellent solution for Open Access paradigm of content exchange. Migration of the Mathematical Collection of Polish Virtual Library of Science to YADDA is currently being prepared.

Key words: digital documents archive, mathematical publications repository, metadata indexing, YADDA, Polish Virtual Library of Science

1 Introduction

The Polish Virtual Library of Mathematics [1] project has started 8 years ago. The idea was to create an open access repository of Polish mathematical papers. The major part of our collection contains both journals and books edited by Polish Academy of Science. The oldest one was published in 1888, and less then 20% of the material is "born-digital". The number of articles exceeds 13,000.

At present, the virtual library is based on a simple, non-scalable backend. The plan for the nearest future is to migrate the mathematical collection to a new flexible architecture created by the ICM — called YADDA — which allows to store wide variety of contents and index them in many ways. This paper presents detailed information about the benefits of YADDA.

2 YADDA Framework Essentials

Nowadays, looking for a scientific literature in the Internet is quite complicated. The number of publications, pre-prints or peer-reviewed journals' articles increases. Researchers have to visit individual author pages, institutional

Petr Sojka (editor): DML 2008, Towards Digital Mathematics Library, pp. 127–130.
 ISBN 978-80-210-4658-0

repositories or journals databases. Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) [2] — a protocol suitable for the exchange of information about digital objects is a useful tool for simplifying this process. It is especially useful in the context of open access journals archives. However, this model is not appropriate for the academic literature databases with licensed-restricted access. Seamless integration of open access and access restricted contents in a single platform is the main innovation of YADDA software.

YADDA document repository framework allows to store and index metadata and content. It is an open architecture, aggregating content from autonomous repositories, which provide a common set of functionalities. Recently, selected modules of YADDA infrastructure were integrated into DRIVER (Digital Repository Infrastructure Vision for European Research [3,4]). The metadata format used in YADDA is flexible enough to adapt to particular needs of participating repositories, in particular it supports multiple document hierarchies (journal-volume-article, series-book-chapter, etc.). A variety of information may be stored together with a document object: publication information, change history, access control list, etc. Thus, YADDA is an excellent framework both for Open Access initiatives, as well as other business models.

Management of distributed data can be obtained on several levels. It is possible to incorporate content from large repositories, for example BazTech archives or ICM libraries [5]. It is also possible to import a single journal, e.g. Polish Journal of Veterinary Science, or even a single preprint harvested from author's personal web page.

3 Browse and Search Capabilities of Web Interface

Contents of YADDA are accessible to users through YaddaWeb [5] — a web interface. The service offers two basic modes of document search: browsing and searching. Access control mechanisms allow authorized users to access licensed content. Like most academic literature databases, YADDA offers keyword search and full-text search. Users can enter queries using intuitive and powerful CQL (Common Query Language).

YADDA search engine is still under development. In the near future the ranking algorithm will be enriched by a number of interesting features. For example, the impact of a document, measured by the number of references to it will be taken into account, keyword similarity heuristics will allow to find documents containing not only the keywords specified by user, but also similar ones.

Other upcoming features, not directly connected with search engine, include author disambiguation, which will allow to identify publications of the same author not only by comparing surnames, but also affiliations, e-mails, and articles' categories. Full-text similarity analysis will display links to potentially interesting documents related to the currently viewed one.

4 Metadata Creation

In general, the creation of metadata is a costly business. It can be done almost automatically for born-digital documents. However, in the Polish Virtual Library the major part of objects is scanned — adding metadata should be done manually. These digitalized papers written quite long time ago, usually in foreign languages (most of them in French), do not follow modern paper format (e.g Banach work: [6]). There are no abstracts or keywords. To create quality metadata we need professionals, who are both French speakers and mathematicians, and can describe content properly, or correct OCR results.

YADDA contains DeskLight [7], a user-friendly application, the content provider the possibility to add and manage the metadata manually. It is also possible to easily edit and change records which are already in YADDA repository. It does not require an advanced knowledge about XML format. DeskLight allows to deal directly with a live repository or work on an off-line copy of the data, which is synchronized with the repository from time to time. In the near future, it will be also possible to add mathematical equations to description, store them in metadata and display in YaddaWeb interface. Both LaTeX and MathML mark-up could be used.

5 YADDA Architecture Details

YADDA was developed under Apache Licence v 2.0 [8], and due to integration with DRIVER project was released for Mozilla Public Licence [9]. Actually YADDA platform used in ICM consist of three main elements: YADDA repository, YaddaWeb — search and browse repository through internet, DeskLight — input, edit and correct meta data in repository. From technical point of view, YADDA repository is a distributed collection of independent services, which can be integrated (both locally and via many kinds of network protocols) in particular applications. The most essential of these services are:

- Catalogue and Editor — storage of metadata. We elaborated BWMETA format for metadata, which is based on Dublin Core concept.
- Storage and Archive — storage of content.
- Authentication and Authorization — access control.
- Search — full text indexing and searching.
- Browse — handling relations between objects in a way similar to relational databases.
- Process manager — service which allows to define and run data-processing tasks.

All of them have precisely specified interfaces, so it's possible to create one's own implementation. However, prototype implementations of the services (except Authentication and Authorization) are currently available. Right now it is possible to handle with up to 500,000 objects. Our future plans are: enlarge number of objects up to 20,000,000, implement the Authentication

and Authorization module and integrate it with YaddaWeb and DeskLight, improve search results by adding services like citation analysis, similarity of keyword.

6 Conclusions

Migration of Polish Virtual Library of Mathematics to YADDA platform will be beneficial in several ways. Using this digital repository infrastructure will give the mathematical society a better access to the archived resources, more adequate search results ranking, better annotation of stored documents. Thus, the Polish Virtual Library of Mathematics powered by YADDA would fulfil all the requirements of a modern digital library.

References

1. The Polish Virtual Library of Science — Mathematical Collection. http://matwbn.icm.edu.pl.
2. Open Access Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html.
3. Digital Repository Infrastructure Vision for European Research. http://www.driver-repository.eu.
4. http://tiki.icm.edu.pl/tiki-index.php?page=Driver.
5. YaddaWeb yadda.icm.edu.pl.
6. Banach, S.: Sur une classe de fonctions continues. Fundamenta Mathematicae, 8, 166–172 (1926).
7. DeskLight http://tiki.icm.edu.pl/tiki-index.php?page=DeskLight.
8. Apache Licence v 2.0 http://www.apache.org/licenses/LICENSE-2.0.
9. Mozilla Public Licence http://www.mozilla.org/MPL/MPL-1.1.html.