Toolset for entity and semantic associations - Final release

Petr Sojka, Mark Lee, Radim Rehurek, Radim Hatlapatka, Maroš Kucbel, Thierry Bouche, Claude Goutorbe, Romeo Anghelache, Krzysztof Wojciechowski

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

ICTPSP
ICT POLICY SUPPORT PROGRAMME
part of the Competitiveness and Innovation Framework Programme CIP

DEMO

Project Acronym: EuDML

Grant Agreement number: 250503

Project Title: The European Digital Mathematics Library

D8.4: Toolset for Entity and Semantic Associations – Final Release

Revision: 1.0 as of 8th February 2013

Authors:

Petr Sojka	Masaryk University, MU
Mark Lee	University of Birmingham, UB

Contributors:

Radim Řehůřek	Masaryk University, MU
Radim Hatlapatka	Masaryk University, MU
Maroš Kucbel	Masaryk University, MU
Thierry Bouche	Université Joseph-Fourier, CMD/UJF
Claude Goutorbe	Université Joseph-Fourier, CMD/UJF
Romeo Anghelache	FIZ Karlsruhe, FIZ
Krzysztof Wojciechowski	University of Warsaw, ICM

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	✓
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	31st May, 2012	Petr Sojka	MU	First version as placeholder for partner's input.
0.2	22nd Jan, 2013	Mark Lee	UB	Attempt of first draft.
0.3	30th Jan, 2013	Petr Sojka	MU	Most gensim stuff added, new structure, summary.
0.4	31th Jan, 2013	Petr Sojka R. Řehůřek	MU	Language independent gensim.
0.5	31th Jan, 2013	T. Bouche C. Goutorbe	UJF/CMD	Citation matching description added.
0.6	1st Feb, 2013	Mark Lee	UB	English tweaked, stats updated.
0.7	4th Feb, 2013	T. Bouche R. Anghelache	UJF/CMD FIZ	Improved matching description, added report on ZBMath.
0.8	5th Feb, 2013	Petr Sojka K. Wojciechowski	MU ICM	Similarity section closed, citation comparison shortened.
0.9	5th Feb, 2013	Petr Sojka	MU	Version for internal review.
0.91	8th Feb, 2013	Petr Sojka	MU	Info about scored similarity.
0.92	8th Feb, 2013	Alan Sexton	UB	Review. Phrasing and English correction.
1.0	8th Feb, 2013	Petr Sojka	MU	Final version.

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

Contents

Executive Summary

In this document we describe the final release of the toolset for entity and semantic associations, integrating two versions (language dependent and language independent) of Unsupervised Document Similarity implemented by MU (using GENSIM tool) and Citation Indexing, Resolution and Matching (UJF/CMD). We give a brief description of tools, the rationale behind decisions made, and provide elementary evaluation.

Tools are integrated in the main project result, EuDML website, and they deliver the needed functionality for exploratory searching and browsing the collected documents. EuDML users and content providers thus benefit from millions of algorithmically generated similarity and citation links, developed using state of the art machine learning and matching methods.

"It is often safer to be in chains than to be free."
Franz Kafka

1 Introduction

There are relations, associations and chains thereof among documents in every digital library. Some of them are already explicit in the paper metadata, like reference lists with hypertext links to Zbl or Mathematical Reviews databases, but most types of associations are implicit and hidden in the free text. Mathematical works depend heavily on previous results: typically the network of literature with rich links between individual documents gives full proof and understanding of published knowledge.

Some links are explicitly provided by metadata or can be inferred by subject classification (for example, MSC codes) but often such associations are either missing or incomplete. In addition, the corpus of mathematical literature is constantly growing which makes the ability to associate newly added documents with older documents in the collection highly desirable. Consequently, we developed and integrated tools in EuDML that allow us to automatically set up a network structure associating related works within a collection of documents.

This report describes the final version of the toolset for automatic semantic association between documents within and outside of the EuDML collection. In D8.3 [7] we have already made a value release of a toolset of services which facilitate the discovery of [semantic] associations between documents within our collection. The toolset focused on two key techniques:

Similarity Clustering collates articles with similar topic and content. It is achieved either by automatically classifying documents according to a pre-defined scheme (e.g. MSC) or clustering documents based on co-occurring terms.

Citation Indexing aims to create a network of documents within the collection by automatic parsing and linking of citations.

The final toolkit consists of two implementations for each functionality. These are:

1. Semantic Similarity
   - gensim (MU) [12] is a software library implementing e.g. unsupervised clustering algorithm which uses various machine learning methods to measure semantic similarity based on word co-occurrences (distributional semantics).
   - Yadda Similarity Service (ICM) a service based on the Lucene open source information retrieval software library [11].
2. Citation Interlinking
   - Yadda Citation Interlinking Service, based on the Lucene open source information retrieval software library [11].
   - UJF Citation Matcher, a citation resolution service based on [6].

Evaluation of the value toolkit demonstrated that for each functionality, the service relying on technology from project partners outperformed the service based on the open source Lucene software libraries already present in Yadda. Therefore, our final release of the toolset relies on just two services—gensim (MU) and the UJF Citation Matcher. Both services have been considerably extended to deal with the particular issues noted during

evaluation of the previous versions of the toolset in the EuDML library. The former is described in Section 2 and the latter in Section 3.

2 Similarity Services

In a modern digital library, offering similarity services for *exploratory search* and browsing capabilities is a must. Instead of keyword based or word based search, there are methods which try to capture word meaning, allowing such services and methods to be termed *semantic* [8]. Similarity metrics are not based just on terms (or word stems), but on *topics*, weighted sets of semantically similar bags of words. Using machine learning methods, these topics can be computed automatically from the word sets representing documents [9].

The similarity service then delivers the functionality of finding ordered set of documents semantically similar to the given one. Exact features of documents which are considered during semantic similarity search depend on implementation—in our case we had to cope with specifics of mathematical formulae, as they often constitute more than half the tokens in the full text of a mathematical paper, and to the best of our knowledge no similarity metric currently supported takes math representation into account. Leading methods for document similarity are based on statistics of a bag-of-word document representation (vector space model) called *distributional semantics*.

2.1 Distributional Semantics

The GENSIM library has been developed using the latest methods for the computation of distributional semantics [5]. GENSIM contains several automated algorithms for deriving semantic representation from plain text. To consider and evaluate them users can compare the performance between Latent Dirichlet Allocation (LDA) [3], Latent Semantic Analysis (LSA) [4] and plain TF-IDF in subdemos 1–3 of the GENSIM Demo at `http://aura.fi.muni.cz:8080`, by means of a drop-down list. The first two methods compute a higher-level, semantic similarity, the last one only measures (weighted) word overlap.

Semantic properties of LSA and LDA come from exploiting word co-occurrence within documents. In a training corpus of documents, words that tend to appear together are taken to be semantically related and soft-clustered together ("soft" because a word belongs to each cluster with a weight or probability, not as a binary decision). Each such cluster of words describes one topic. Obtaining the clustering automatically and efficiently is a major challenge; GENSIM is a leading framework in scalable model training, and has already been used in several dozen of projects and applications.

Given a trained LSA or LDA model, any text document can be described by how much it belongs to each topic. This gives a higher level (more abstract) representation of the document's contents—now even two documents that do not share any words in common can be evaluated as closely similar. This is a strict departure from an exact keyword overlap as popularized in search engines with boolean keyword searches.

2.2 Semantic Indexing and Search

We have implemented several state-of-the-art/leading-edge methods, and evaluated them in previous deliverables [9, 7]. We have realized that performance depends on many features, representation choices of documents, machine learning method parameters, all of which have to be considered during the implementation of a semantic similarity search.

One aspect of indexing, which may vary between implementations, is the ability to add documents in an incremental fashion: incremental indexing means that document collections may be added to the service without the need to re-index all the documents added before. The availability of incremental indexing—*on-line processing*—is an important feature considered during the comparison of different implementations. For efficient processing, on-line similarity processing is a must and is used in EuDML similarity implementation(s).

In its core, the similarity service exposes two operations, similarity *indexing* and *search*:

- **Similarity indexing** – All documents must be preprocessed (tokenized, stop-listed, weighted, ...) before being made available to the similarity search. Indexing is done only once for each document and can therefore be considered a setup process for the service. Indexing is incremental (on-line), so the whole similarity matrix does not have to be recomputed when a new document arrives.
- **Similarity search** – The core functionality of semantic similarity search is to find similar documents to a particular document. Given a set of documents which are indexed (i.e. the EuDML collection), the similarity service returns an ordered list of the most similar documents to the specified one. It should be noted that different similarity metrics are defined (in fact, each implementation of the service is a realization of a particular metric) and could be combined.

After giving a general overview of the service, in the next three sections we focus on three implementations in the toolkit: two based on the GENSIM library and a Yadda similarity service implementation. Although not all three are meant to be used in parallel, they are all in place in the EuDML system.

2.3 Language Dependent GENSIM Similarity

The goal of this tool is to assist humans in data exploration via similarity browsing. GENSIM [5] is a software framework for modelling semantic similarity of text documents. Within the context of digital libraries, it allows querying for "similar documents", with similarity based purely on document contents (i.e. on plain text, possibly enhanced by metadata).

To demonstrate the capabilities of GENSIM, several web demos have been produced which showcase possible scenarios (similarity demos):

Proof-of-concept Demo 1 of GENSIM possibilities. Four subdemos available at `http://aura.fi.muni.cz:8080` demonstrate GENSIM functions to answer questions like:

1. For a given article, what are its ten most similar articles in the library?
2. Given two articles, how similar are they?
3. What are the pairs of the most similar articles across the entire collection (plagiarism candidates)?

4. Given an article, what are the topics/'keywords' covered by this article (data exploration)?

5. What is the expected query performance over a real-world sized collection of scientific documents?

MIaS4gensim Demo 2 `http://aura.fi.muni.cz:8889` This Demo showcases the most and the least common document terms, plus the most prominent words for LDA and LSA topics. From the results it is clear that math formulae appear frequently in topics, and are thus important to consider in math-aware document similarity computations. As a result, EuDML's similarity handling should take math into account when building a paper's bag-of-word list, during tokenization and preprocessing.

EuDML integration Demo 3 shows the integration of GENSIM in the final version of the EuDML system. The deployed GENSIM similarity system can be found at `http://eudml.org`, where it is used to compute similarities for more than 100,000 documents, on-the-fly. There are screen shots of the user interface in Deliverable D6.5 [15].

The first two demo URLs use data from the MREC collection of mathematical texts [10] (a snapshot of 434,894 full-text articles from ARXMLIV, originally from ARXIV) to verify the framework's usage and scalability on documents typical for EuDML (scientific papers with a lot of math). Current statistics of similarity computations show that the similarity matrix is computed for almost quarter million documents in EuDML.

To integrate GENSIM (written in Python) into the Java-based EuDML, we run GENSIM as a client/server architecture. The server is responsible for creating, updating, querying and maintaining the similarity index, while the client (Java) issues similarity and update requests. All communication between the two happens over TCP/IP.

We also implemented transaction handling for server modifications: all operations that mutate a server's state are done over a copy, so that the server is free to handle "live requests" without any down-time. At the end of a transaction, the copy is either committed (i.e., requests start being served from the modified index) or rolled back (all changes made in the transaction are discarded). In this way, the GENSIM similarity server achieves high availability even during index updates.

The server is designed to handle indexing and querying large data efficiently. It uses virtual memory to handle datasets that cannot fit in RAM. With datasets larger than the physical memory (i.e., more than tens of millions of documents, the exact number depending on a particular server's specification), the server continues to serve requests gracefully, albeit more slowly because of the virtual memory swapping.

A similarity index is created *separately* for every language, which means a user receives similar papers only in the language of the original paper. Indexes for languages containing less than 1,000 documents employ a simpler, TF-IDF (log entropy) model, compared to languages with more documents that use full LSA with 400 dimensions (topics). The threshold of 1,000 was chosen based on our observation that full statistical methods require a substantial amount of training data to produce meaningful results.

Preprocessing uses standard tokenization, considering only alphabetical terms coming from each article's full-text.

2.4 Language Independent Similarity by GENSIM

The rich metadata (in particular, mathematical formulae) provided by EuDML prompted a second, enhanced version of the GENSIM similarity server. Here, all languages are indexed under a single common index, so that a query returns the most relevant documents regardless of their language.

Apart from full-texts, this version also takes into account each article's MSC codes, keywords, mathematical formulae, its title and description. All of these parts receive different (customizable) weights which mandate their relative importance during subsequent similarity computations.

Some of these parts, such as MSC codes, MSC descriptions in English, paper keywords and their translations into English, are either language-independent or available in multiple languages, allowing (partial) similarity matching between documents written in different languages. We have evaluated also Google translations of full texts, but these are costly and thus not sustainable unless Google would donate their translation service for the project. Having similar papers in different languages (via interlingual terms as MSC descriptions and codes, translated keywords, and math 'pronounced' in English) was the main reason behind using a single unified, language-independent semantic indexing.

The extended, richer document structure is being passed on to the server via JSON format, in a way that is fully backward-compatible with the existing APIs and could be used even in parallel with language dependent GENSIM similarity. The created *similarity document* used by GENSIM algorithms contains not only metadata in various languages, but full text split into plain text and math formulae. Math is passed separately in several formats together with other metadata allowing special weighting for each type of data because they express paper semantics in a different way. The actual math formats provided are:

- Set of MathML formulae.
- Concatenation of MathML formulae transformed to text using a new *MathML-to-text converter* [13]. The MathML elements converted to text are only in English version, which serves in this way as a kind of *interlingua* language.

MathML-to-text converter processes XML files that contain one or more MathML blocks and converts each such block into plain text format. Converted text is equivalent to original MathML, operations and symbols are written out in words, so you can read mathematical formulae like a novel.

Input file is parsed using streaming API for XML. MathML block is then transformed into simplified DOM model. Based on MathML type, presentation or content, slightly different method is used. If both types are present, content MathML takes precedence, because presentation MathML can be ambiguous and unclear.

The GENSIM server also supports transactions and very large, out-of-memory indexes in the same way as the language-dependent version described in the previous section.

2.5 Yadda Similarity

The Yadda similarity service was implemented over the Lucene full-text search engine [11] and its "more like this" search functionality. Indexing consists of building an inverted index of documents, i.e. a mapping between words (terms) and the documents which con-

tain them together with some additional statistics (for instance the number of occurrences in each document). The inverted index is used for similarity searches in the following fashion:

- A given document's important features are determined during search. For full-text search, any 'feature' is just a word which can be considered specific or highly significant for the given document. Such words are chosen using term frequency/inverse document frequency (TF-IDF) statistics for each word.
- The most specific/significant words of the document are used to construct a boolean OR full-text query—standard full-text search and its "term vector model" is used to determine set of documents which match the query in the best way (documents which are most similar to the words in the query). Documents found during full-text search are returned as similarity results.

2.6 Similarity as a Service

Another issue left to be resolved recently was what method (API) should be employed for the export of similarity results. The export of similarity results via service is demanded by some data providers like DML-CZ, because this will allow them to use a much richer base for full text similarity computations than they now do[1]. Scores have been added to the interface to allow using EuDML similarity on data provider sites. Moreover, the *Linked data* approach is becoming popular and exposing links and relations with EuDML ids could result in increased incoming traffic,

The service for accessing similarity results as a XML file is available under the address `http://eudml.org/api/rest/similarItems` with a demonstrator at `http://project.eudml.org/api-tester/similarItems`. More detailed description can be found in [14, Section 4.3].

2.7 Evaluation of Similarity Services

A comparison of the two language dependent similarity services was conducted by randomly selecting six documents (two English, two German and two French articles) and comparing the top five most similar articles returned by GENSIM (language dependent variant) and the Yadda similarity service. In general, we have found GENSIM to return more subject specific research articles rather than articles consisting of a generic overview of a wider subject. In addition, even 'language dependent' GENSIM appears to cope better with different languages. Disappointingly, one of the two French language articles and both of the German language articles produced no similar articles according the Yadda similarity service, as described in D8.2 [9]. For more details, see [7]. As a result, only GENSIM has been deployed in EuDML system since version 1.3. It is sufficiently robust even though for some articles there are no full-texts available, and thus GENSIM is using, for these cases, only basic metadata and MSC tokens for bag-of-words semantic paper representation.

As noted during evaluation in D11.3 [1], multi-linguality has remained an issue. While the content provided in EuDML includes articles in a significant number of languages, there is not necessarily a large number of articles for every language. Similarity

1. Only NUMDAM papers have been used for similarity in DML-CZ so far.

lists then consisted only of papers in the same language, which causes problems for minority languages as every paper in a minority language was deemed similar to the others in the same language.

On the other hand, it is seen that similarity gives usable results even in the cases of papers without full texts at the disposal for internal EuDML use.

At the time of writing, a thorough comparison of both versions of GENSIM has not finished yet. Propagating both (or even all three) similarity metrics to the end user interface might confuse them, based on the experience of DML-CZ and seconded by other EuDML partners. We expect that in the near future our evaluation will confirm better semantic and language independent behaviour of language independent GENSIM similarity metrics taking into account also formulae, and that EuDML will switch it as the main and only one. The interfaces of both versions are the same so switching will be trivial.

3 Linking and Matching Tools

Most mathematical works explicitly refer to other works via references/citations which are used within the text and then listed as part of the bibliography. Therefore, via the citations, it is possible to locate a particular document within a rich network of other (related) documents. Since many documents in the EuDML collection are likely to contain bibliographic references, and many of these references are likely to point to (other) documents within the EuDML collection, we use *Bibliographic Reference Matching* as the primary means to build up a relationship network between articles. The goal of bibliographic reference matching is to assign to a bibliographic reference an identifier of the referenced document.

3.1 UJF Citation Matcher

As described in D8.1 [8], D8.2 [9] and [6], the UJF Citation Matcher provides a robust method for resolving (incomplete or possibly incorrect) citations to a particular document or identifier.

The UJF/CMD citation extraction and matching algorithm does not attempt to perform citation parsing or citation field tagging prior to trying to find matching citations. Instead, citations are viewed simply as strings of characters with no attempt to parse a structure to the citation string. As argued in [6] this avoids several problems:

1. Even if citation parsing is successful, individual fields often remain coded in different ways and cannot be compared using exact comparison methods.
2. Errors in parsing can result in the complete failure of the matching process.
3. The parsing process is both costly and error prone.

For these reasons, the UJF/CMD approach relies on just a shallow analysis of the input string and relies on a number of string similarity evaluation methods and ad hoc heuristics which work reasonably well in practice in the context of mathematical databases. In particular the matcher relies on numerical information in the citation string to resolve the citation. This approach proves to be relatively resistant to multi-lingual and typesetting issues.

The initial version of this tool was essentially targeted at the matching of journal articles. During the course of the project, algorithms for book matching have been designed and implemented.

The service is now fully integrated within EuDML. It is used internally by the EuDML processing system in order to establish links between EuDML documents—see Section 3.2 for an evaluation of matching results.

In addition two services are available to external users

- At `https://eudml.org/refsLookup`, an interactive lookup where the user inputs a bibliographic citation (as a string) and gets back near matches when they are found.
- At `http://eudml.org/api/rest/batchref` a tool that can be used to process a batch of reference strings and get back EuDML identifiers when they are found.

3.2 Evaluation of Linking Services

As reported in D8.3 [7] the value release of the Association toolkit consisted of two services for citation matching and indexing: the UJF Citation Matcher evaluated below, and the Yadda Citation Interlinking Service based on the open source Lucene toolkit [11] described in more detail in D8.2 [9].

Evaluation of the two prototype services revealed that the UJF Citation Matcher provided a more robust service compared to Yadda. [7]. The UJF Citation Matcher is capable of good coverage of citations due to its reliance on robust methods which allow it to deal with issues related to typographic mistakes and differing conventions on referencing.

However, coverage is still an on-going concern and future work is required to improve both the accuracy and coverage of the citation indexing service. Currently, the total number of items in EuDML is 225,809 (after deduplication). Of these, 52,156 documents have their references recorded in the metadata. Following extraction, this resulted in 656,651 individual reference strings. Using the lookup tool developed at UJF/CMD we were able to match 99,282 reference strings to EuDML documents. Of these, the vast majority (98,000) resolve to articles in journals and proceedings and just 1,282 to books.

Coverage is better at the document level. Out of the 52,156 documents, 34,480 documents contain at least one matching reference. This figure can be broken down to 34,298 articles, 332 books.

Conversely, 38,391 documents can be matched to at least one reference. Again, this can be broken down to 38,059 articles and 332 books. Some references strings have been matched more than once and thus require further disambiguation (possibly manually).

EuDML collections are expected to cover about 6% of the whole mathematical reference corpus (estimated to be above 3.5 million items as of 2012). However, the EuDML corpus has some specificities:

- it contains few books, which are the most cited items. Most of them from the 19th century up to the first half of 20th century.
- it contains a very strong collection of European journals going back to the beginning of 19th century, with many fundamental works heavily referenced.

- the relative coverage of important and long-lasting journal articles is better in the early period when Europe was the center of the mathematical world, and decreasing with time, as a fast growing number of articles have been published elsewhere.
- some content partners have bibliographical references for their journal articles, but most have not. Moreover, formalized separate reference lists are a rather recent practice. To get most of the reference lookup, and create a denser citation network, we would need to extract citation information from the running text or scanned PDFs from all contributed items. Unfortunately, this data could not be produced within this project.

For these reasons it is fair to expect a poor matching rate for book citations, but a higher one for journal articles as citations from real-world articles are not a random sample but typically point to older important works, which is precisely the kind of content we have in EuDML (we see that a book is cited 3.8 times on average, while an article is cited 2.5 times). From the above statistics, we infer that the 15% success rate of matching citations within EuDML proves that we succeeded in assembling a corpus of reference documents as at least 15% of the citations in EuDML are matched to a EuDML item though EuDML represents only 6% of the existing published documents in mathematics (a given cited item might be counted multiple times here, which is a feature of this analysis).

To give a comparison with a flat and comprehensive collection of references covering 1931-current mathematical output worldwide, FIZ ran a series of queries to the Batch Ref service for over 3.2 million references in Zentralblatt, out of which 170,273 matches were obtained. This is a 5.3% match rate, which means that almost all EuDML items have been matched to a Zentralblatt reference, as we have 177,000 EuDML items with a publication year greater than 1931, which amounts to roughly 5.7% of the 3 million items in Zentralblatt MATH for the period 1931–2012. As EuDML has been using the Zentralblatt lookup (based on the same technology, but fine-tuned in a different manner), we will have data to conduct a thorough review of matching accuracy during the next weeks.

4 Conclusions

This document reports the final release of the toolset for entity and semantic associations which consists of two functionalities—integrating Unsupervised Document Similarity implemented by MU (using GENSIM tool) and Citation Indexing and Matching (as provided by UJF/CMD). The toolset consists of stable technologies integrated within EuDML and allows users explore the rich content of the library with computed similarity and citation links, as demonstrated on EuDML site `http://eudml.org` and on stand-alone demos. The developed tools have been evaluated and helped to reach the indicators for the evaluation set in the DoW, as reported in D11.4 [2] and are in everyday use in the EuDML system.

References

[1] Romeo Anghelache, Michael Jost, Brigitte Bidegaray-Fesquet, Thierry Bouche, Thomas Fischer, Hans-Karl Hummel, Ioannis Karydis, Klaus Kiermeier, Yves Laurent, Helena Mihaljevic-Brandt, Radoslav Pavlov, Gilberto Pedrosa, Aleksandar Perovic, Lucia Santamaria Lara, Olaf Teschke, and Krzysztof Wojciechowski. EuDML assessment and evaluation — First report,

April 2012. Deliverable D11.3, revision 1.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.

[2] Romeo Anghelache, Jiří Rákosník, Jarmila Štruncová, Jiří Veselý Alexander Nowiński, Krzysztof Wojciechowski, Jean-Luc Archimbaud, Brigitte Bidegaray-Fesquet, Yves Laurent, Thierry Bouche, Julien Puydt, Thomas Fischer, Michael Jost, Klaus Kiermeier, Lucia Santamaria Lara, Helena Mihaljevic-Brandt, Aleksandar Perovic, Olaf Teschke, Radoslav Pavlov, Georgi Simeonov, and Petr Sojka. EuDML assessment and evaluation – Final report, January 2013. Deliverable D11.4 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[5] gensim. `http://mir.fi.muni.cz/gensim/`.

[6] Claude Goutorbe. Document Interlinking in a Digital Math Library. In Petr Sojka, editor, *Proceedings of DML 2008*, pages 85–94, Birmingham, UK, July 2008. Masaryk University. `http://www.fi.muni.cz/~sojka/dml-2008-program.xhtml`.

[7] Mark Lee, Petr Sojka, and Radim Řehůřek. Toolset for Entity and Semantic Associations – Value Release, May 2012. Deliverable D8.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, `https://project.eudml.eu/sites/default/files/D8.3-v1.0.pdf`.

[8] Mark Lee, Petr Sojka, Volker Sorge, Josef Baker, Wojtek Hury, and Łukasz Bolikowski. Association Analyzer Implementation: State of the Art, November 2010. Deliverable D8.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, `https://project.eudml.eu/sites/default/files/D8.1-v1.pdf`.

[9] Mark Lee, Petr Sojka, Volker Sorge, Wojtek Hury, Łukasz Bolikowski, and Radim Řehůřek. Toolset for Entity and Semantic Associations – Initial Release, May 2011. Deliverable D8.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, `https://project.eudml.eu/sites/default/files/D8.2-v1.pdf`.

[10] Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec. Web Interface and Collection for Mathematical Retrieval. In Petr Sojka and Thierry Bouche, editors, *Proceedings of DML 2011*, pages 77–84, Bertinoro, Italy, July 2011. Masaryk University. `http://hdl.handle.net/10338.dmlcz/702604`.

[11] Apache Lucene. `http://lucene.apache.org/`.

[12] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`, software available at `http://nlp.fi.muni.cz/projekty/gensim`.

[13] Petr Sojka, Michal Růžička, Maroš Kucbel, and Martin Jarmar. Accessibility Issues in Digital Mathematical Libraries. In *Proceedings of the Conference Universal Learning Design*. Masaryk University, Brno, February 2013. 8 pages, to appear.

[14] Krzysztof Wojciechowski and Aleksander Nowiński. Web and Service Interface Implementation – Service Interface, 2012. Deliverable D6.4 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.

[15] Krzysztof Wojciechowski, Aleksander Nowiński, Jake Grimley, and Martin Líška. Public User Interface – Final Release, January 2013. Deliverable D6.5 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.