Abstract—Technical background of digital library of University of Maribor is described in this paper. We start with basic description of the library and its purpose, but the main focus of this paper is set on features, that are mostly not found in other digital libraries. Features like plagiarism detection, informative and useful statistics and specific content extraction are described. We present existing functionality and describe some ideas for future development.

Keywords—digital library, natural language processing, plagiarism detection, Slovenian language, text processing, University of Maribor

I. INTRODUCTION

Although never publicly launched, digital library of University of Maribor is running approximately since autumn of 2008. This silent launch somewhat coincides with the new regulation adopted by the university, that requires from all students to provide electronic version of their final thesis beside the printed one. For now, the main purpose of our digital library is collection and dissemination of final theses, created by students of all faculties of the University of Maribor. Students upload their final theses into digital library themselves as PDF files; after those documents are processed by the local librarian, they are publicly available via the digital library to everyone. Till now, diplomas, master and doctoral theses were accessible only as a single printed version at the local library of a specific faculty; now those documents are accessible to everyone from everywhere – all that is required is a web browser. With RSS, we inform users about newly published documents. Table 1 presents some statistics about number of published documents on July 2010, where column “New” means number of documents, added in the last month.

Technically, digital library of University of Maribor is a web application, build on LAMP platform (Linux, Apache, MySql and PHP). Its publicly available functionality includes simple search, advanced search and browsing though a simple manually made hierarchy. This hierarchy currently includes only two levels, types of theses on the first level and faculties on the second; documents are linked to the second level automatically. Hierarchical structure will be expanded in the future.

<table>
<thead>
<tr>
<th>Document type</th>
<th>All</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bachelor theses</td>
<td>7563</td>
<td>341</td>
</tr>
<tr>
<td>Master theses</td>
<td>221</td>
<td>12</td>
</tr>
<tr>
<td>Doctoral theses</td>
<td>74</td>
<td>5</td>
</tr>
<tr>
<td>Other documents</td>
<td>47</td>
<td>6</td>
</tr>
<tr>
<td>All documents</td>
<td>7905</td>
<td>364</td>
</tr>
</tbody>
</table>

In a way, digital library of University of Maribor contains all the knowledge of the whole university. The first goal of digital library is to simplify access to this knowledge with search and browse functions. But all this knowledge is hidden in texts, written in natural language. At the end, users (in our case mostly students) still have to read the texts to find what they were really looking for. The next goal is to extract this knowledge and use it to help our users even more – to find what they are searching faster or to deliver information in such a way, that users don’t need to actually read the documents.

This paper presents advanced features of digital library of University of Maribor, where some of these features represent out first steps towards this goal. Under advanced features we understand features that either assist to basic features of a digital library (cataloging of documents, search and retrieval) or introduce new features, never seen in digital libraries. We first describe integration of digital library with other systems, continue with certain specific statistics, describe plagiarism detection, based on documents in digital library, and end with targeted content extraction.

In the reminder of this paper, digital library of University of Maribor will be referred to with the acronym DKUM, which is Slovenian for “Digitalna knjižnica Univerze v Mariboru”. All statistic data mentioned in this paper show the state as it was on July 2010.
II. INTEGRATION

While DKUM is capable to perform its basic operations on its own, it is still integrated with several other information systems, as shown on Fig. 1. Let us start at the beginning, where documents are collected.

![Fig. 1 DKUM integrated with others](image)

As already mentioned, the majority of the corpus consists of final theses, created by graduate and postgraduate students of University of Maribor. Those documents are uploaded by students themselves. A student visits DKUM web page and logs in, where the first external information system is used. Authentication is handled by AIPS – academic information subsystem. AIPS contains all the data about all students of University of Maribor, including login data. Communication between DKUM and AIPS is done via web services. When logged-in, students visit DKUM subpages, intended for document upload, consisting of 3 steps. First, students must enter some basic information about their thesis, like title, abstract and keywords in their native and in one of foreign languages of their choice. They also enter their mentor and co-mentor (if they have one). In majority of cases however, some of this information is already available via AIPS and if so, then those fields are already filled out. Additionally, information about faculty and students is also read from AIPS and saved in DKUM for later use. Second step is the actual document upload. Students must upload a single PDF file, which must be identical to the printed version. They may also upload other files, but are not required to. When the first two steps are completed entirely, they can proceed with step three, where students can get a written statement of equality of both printed and electronic version of the document. This statement is a filled-out form in PDF format. Students need to print this statement, sign it and add it to the printed version. When the printed version of the thesis with the signed statement is delivered to the faculty, their electronic version is locked, but not published.

Before documents are published, they are processed by local librarian. Each faculty has its own local library, so each faculty processes its own documents, when the printed version arrives to the library. First, document is cataloged in COBISS – cooperative online bibliographic systems and services. This is an external information system. It is a network of local bibliographical databases, spanning over 400 local libraries in Slovenia, while the same system is also running in Serbia, Bosnia and Herzegovina, Montenegro, Macedonia and in process of implementation in several other countries [1]. Cataloging of student theses into COBISS is done regardless of DKUM’s existence. What is new however is that the COBISS record now also includes a direct link to the documents in DKUM. This way all new theses in COBISS also point to DKUM. After cataloging in COBISS is done, local librarian also checks the record in DKUM and connects it to the record in COBISS. After that, the DKUM record is published. This way, both records in DKUM and COBISS point to each other. With this, documents in DKUM are also searchable in COBISS, which expands the DKUM reach and makes its corpus available to even greater audience.

Integration with COBISS is done with the help of a data exchange protocol, Z39.50 [2]. In this regard, DKUM supports full import of COBISS record, where all metadata from COBISS are copied into DKUM, either as a new record or it overwrites an existing record. DKUM also supports linking, where only identifiers are copied, but the content remains unchanged.

DKUM also supports OAI-PMH v2 protocol (Open Archive Initiative [3] – Protocol for Metadata Harvesting [4]). This protocol is based on open standards (XML, HTTP and Dublin Core) and enables metadata exchange between libraries. Although this is a client – server protocol, DKUM currently supports only the server side of this protocol. This way DKUM can provide its metadata to other libraries, but not the other way around. Currently, this protocol is used to provide our metadata for the DRIVER repository (Digital Repository Infrastructure Vision for European Research [5]). We implemented this protocol using an open source OAI PMH v2 data provider [6].

III. INFORMATIVE AND USEFUL STATISTICS

DKUM has as set of publicly available subpages dedicated to statistics. Beside already mentioned statistic about number of documents and number of new documents per document type and faculty, statistics also include:

- top viewed records of the last week and month;
- top downloaded documents of the last week and month;
- comparison chart with number of document per faculty and thesis type;
- time chart with number of new document for selected faculty and thesis type;
- pie chart showing corpus size in terms of number of files by type (currently about 11584 files of all types, ¾ of it are PDF files) and file sizes by type (about 29,3 GB of files, almost ¾ of it are PDF files);
- yearly reports (by calendar year and academic year)
with number of documents and number of new
document per faculty and thesis type. They were
needed by librarians, but we made those reports
available to everyone;
• statistics about mentors.

Statistic about mentors is not just interesting, but may be
useful to students. This statistics show what mentors
(professors) are doing now and where doing in the past. This
information is extracted from keywords of documents, were a
given person is a mentor. This statistics includes:
• number of theses (bachelor, master and doctoral),
  where given person is a mentor or co mentor per
  faculty;
• table with bar chart showing number of theses
  published by year and thus the activity of the mentor
  (shown on Fig. 2);
• all keywords associated with the given mentor, sorted
  by frequency. Indirectly this list shows all research
  areas a selected mentor is active in;
• keyword appearance by year. This shows the mentors
  activity period in specific research area;
• a list of other persons that contributed to development
  of theses (other mentors or co mentors), sorted by
  frequency. It is shown who cooperated with the selected
  person.

The statistic about mentors could be extended even further.
We could create a special search, where students would search
for mentors depending on what a student would like to do for
his final work. From the mentor – co mentor relation a social
network could be drawn, maybe for the whole university. This
social network could show several things, like cliques (social
subnetwork), who is working in interdisciplinary fields
(working with people from other faculties), who doesn’t like to
work with others and so on. This could potentially be
interesting for professors, for instance when searching for
research partners or experts on specific research area. This
kind of search could also be useful for industry, doing research
projects.

IV. PLAGIARISM DETECTION

DKUM has the ability to perform plagiarism detection. Actually, DKUM
doesn’t do plagiarism detection by it self; it only displays the end results. Actual detection is done using
a natural language processing framework called TextProc. This
framework was developed by us independently of DKUM and
it uses software plug-ins to do its language processing. Each
plug-in performs a specific function from natural language
processing. These plug-ins can be then put together into
processes, that perform a higher natural language processing
function. One such process is plagiarism detection. Since most
documents in DKUM are in Slovene language, this process
and its plug-ins are specifically designed to process Slovene
language. The framework itself is language independent, since
language dependent functionality is hidden in plug-ins.
TextProc also has the ability to run those processes via web
application (for testing) and as a web service (for integration),
so DKUM communicates with TextProc via web service.
Below, the process of plagiarism detection in DKUM is
described.

First, DKUM converts uploaded documents into plain
text. This is done regardless of the plagiarism detection, since
plain text is also needed for DKUM’s search functionality. For
now, only published documents are converted into plain text,
since they represent the final version of the document. Next,
plain texts of these documents are uploaded to a separate
server with TextProc installed. Upload is done for each
document separately via web service that runs the first of two
TextProc processes for plagiarism detection. First process
prepares the document by running it through the plug-ins in
this order (each point is a plug-in):
• text is tokenized (broken into words).
• Certain tokens are merged back together by a given set
  of rules. For instance, a decimal number “3,14” is
  separated by the first plug-in. The second plug-in
determines that the comma is not a sentence separator,
  so it is merged into one word.
• All words are converted into lemma form (canonical or
dictionary form of the word). This part is very language
  specific, since Slovenian language is heavily inflected.
• Sentences and clauses are determined.
• Paragraphs are determined.
• Words in lemma form are merged into new sentences
  without redundant spaces, tabs or line feeds that may be
  present in the original text; only a single space
  character is used as word delimiter. Also, words are
  sorted alphabetically on the level of a sentence. This
  way, word order within sentences becomes irrelevant.
• Newly constructed sentences are hashed using a hash
  algorithm. Currently MD5 (Message-Digest algorithm
  5) is used; several variants of SHA algorithm (Secure
  Hash Algorithm) are already supported.
• Previous plug-in in called again; this time it hashes whole paragraphs.
• Documents and its hash values for sentences and paragraphs are stored in a database.

As already mentioned, this process is executed for each document. This process returns a document identification number, which is stored in DKUM for later use. This number is used, when DKUM calls the second web service, which runs the second TextProc process, consisting of only one plug-in. This plug-in does a simple database search, which returns all identical hash values that appear in given document and other document at the same time. A report in XML format is returned as a result. This report is then processed in DKUM and stored in DKUM’s database. Results are stored in such a way, that enables progressive plagiarism detection – reports for new documents automatically update existing reports of previously processed documents. Reports don’t tell which documents are plagiarized and which are not; they only contain a similarity percentage, similar content and its position in documents (if requested).

Plagiarism detection report can be viewed on administration pages of DKUM in two ways. First, we have a list of top similar pairs of documents for a given faculty or whole university of Maribor. Second option is via document record editing pages that are organized into tabs. One of these tabs shows detailed output of plagiarism detection report. On one side, there is a list of documents with similarities with currently viewed document. This list is sorted by similarity (percentage) and clicking a title of one of the documents shows all similar sentences or paragraphs. Contents, detected as similar but written slightly different (differences in character case, word order or word form with the same lemma) are marked and displayed for easy manual comparison. Both report views show, which document is the source and which is a copy, if it is determinable (by date of publishing).

Although plagiarism detection reports are visible on administration pages of DKUM, actual usage scenarios of them is not determined yet. There are some ideas, but no concrete usage scenarios. One idea is that mentors could check the report for documents of their students, before those documents are published. They could also check any document for similarity with DKUM corpus like homework and coursework assignments or any other documents that are not intended to be stored in DKUM. Such documents would be stored in DKUM temporarily, only as long as is required for detection operation to complete. The other idea is that students could check their own work as a part of upload process.

Till now, plagiarism detection is done only between documents in DKUM corpus. In future, other sources will be added. We will add Slovenian part of Wikipedia, which is commonly used by students as reference. We will also add theses and seminar works from other universities and research papers.

V. SPECIFIC CONTENT EXTRACTION

We are currently working on extraction of specific content from uploaded document. Special software is in development that takes a document (for instance a PDF file) and extracts table of content, figures index, index of tables, all URL addresses, index of technical terms and acronyms, index of equations, index of references and more. These contents will be visible to users as part of metadata display.

Search results in DKUM show a limited amount of information: title and abstract in primary language of the document (mostly Slovenian), author, publishing date, number of views, number of content file downloads and link to the file with full content. If users want to see all data about selected document, they have to click on the title. Because each thesis has its title, keywords and abstract also available in a foreign language, and because a record contains a lot of information, this information display is organized into several pages, accessible via tabs, as shown on Fig. 3 (the tab “Sekundarni jezik” is selected, meaning “Secondary language”).

All previously mentioned extracted content will also be available via tabs. This way, users will get a better impression of what the selected document is about, without actually opening a file. Some of the extracted content could also be used in other ways. For instance, extracted equations, technical terms and acronyms could be searchable via advanced search. User can specify which data to search for. A new data category includes all documents that use the searched reference. In metadata display of a document, users could also get a list of other documents that refer to currently viewed document.
VI. FUTURE DEVELOPMENT

There are many ideas for improvement of DKUM, from more basic to advanced, later ones involving natural language processing.

As first we want to extend the types of content we serve. Currently our corpus consists mainly of bachelor, master and doctoral theses. There were several requests to also include learning material, produced by teaching staff at the University of Maribor; these learning materials are already in digital form. We would also wish to include research papers, produced at our university.

In the introduction of this paper, we already mentioned a browsable hierarchy with documents, which we intend to expand. One possibility is automatic construction of hierarchy, based on other hierarchies. We already have the necessary software that can build this hierarchy based on UDC (Universal Decimal Classification). Also we wish to integrate other taxonomies (ACM, Eurovoc, DBLP, University of Maribor taxonomy, DBpedia...)

There are several upgrades planned that would improve user experience. First we have to take care for user registration and login. Currently the only users that are able to log in are students, librarians and some personal from education office. Registration for all users is required. Logged-in users would be able to customize certain aspect of DKUM, like default search values (home faculty, default document type, language), number of hits per page (search results are paged) and so on. We would introduce “personal book shelves”, where users could create their own lists of documents. Those could be private or public where public ones could be viewed by others. Personal shelves with the introduction of learning material could be potentially useful for teaching staff. They could collect useful learning material from DKUM and put them into public shelves (e.g. one shelf per course) and share these shelves with students.

There is also the idea of using the torrent protocol for downloading big files (like ISO images of optical discs) or several documents at once, like all documents form a personal bookshelf. Introduction of torrent protocol would potentially save some of our network bandwidth because of the distributed nature of this protocol. It would also help users with big downloads, since torrent clients are much better at handling download interruptions than browsers.

Currently DKUM is available only in Slovenian language, although content in foreign language is also collected. There are several theses, especially doctoral theses that are written entirely in English. If research papers are included in DKUM, the lack of multilingual support becomes even more problematic. There are also many exchange students at University of Maribor, that could use DKUM, but can’t because of the language. One of improvements of DKUM will be the implementation of multilingual capabilities. This way, DKUM will also be available in English, possibly also in German language.

More advanced ideas include improvements in areas of personalization, better plagiarism detection and knowledge extraction. Using data mining algorithms on usage logs of all users a content recommendation system could be performed. Current problem is the lack of usage logs for specific users.

There are also several ideas for plagiarism detection algorithm, like replacing all numbers, written as words into actual numbers, before hash is calculated. Actually, all numbers can be removed and replaced by a tag, representing any number. Often people copy some text and make small corrections, like changing word order (we already detect this) and changing numbers (removing or adding decimal numbers). From plagiarism detection perspective, actual numeric values are irrelevant. Certain words in Slovenian language can be present in a sentence but carry no actual meaning and may be added only to confuse plagiarism detection. These words can also be removed, before hash is calculated.

We build our own semantic lexicon, which is consisted from Slovenian Wordnet [7], Eurovoc [8] and domain specific dictionaries extracted from DKUM theses. We will use these lexicons for synonym normalization, which will improve plagiarism detection.

We will use semantic lexicon and other sources (Wikipedia, Dbpedia, documents from other Slovenian universities and national digital library) for linking DKUM documents to these sources.

Users of DKUM have possibilities to comment and rate DKUM documents. We will also enable users make semantic annotations to these documents. Annotations will be used for semantic tagging and in recommendations to other users.

With help of natural language processing techniques and semantic web we will try to extract not only content, but knowledge. Since DKUM contains knowledge of the University of Maribor, there is a lot of knowledge waiting to be discovered. Using this, we could build completely new ways of using not only DKUM, but all future digital libraries. Extracted knowledge could also be useful for other applications. One such application is a question answering system in our own development, which is already present as a part of DKUM web pages [9]. Currently, it is capable of answering basic questions regarding DKUM, but in the future it may also recommend documents on specific areas of research (via statistic) or even produce answers, based on knowledge, extracted from corpus of DKUM.

VII. CONCLUSION

In this article, the existing and future features of DKUM are described. Main focus is on advanced features, not often found in other digital libraries. Currently our focus is primarily on natural language processing capabilities, since we need those capabilities not only in DKUM, but also in other projects. For this purpose, TextProc framework is in further development; plagiarism detection in DKUM is its first practical application. Documents from DKUM are also useful for testing and development purposes, since those documents are well structured and written in formal Slovenian language. This data
will help in development of generic, language dependent natural processing features, which will be used in future research and applied projects.

REFERENCES