Advanced features of Digital library of University of Maribor

Janez Brezovnik, Milan Ojsteršek

***Abstract*—** Advanced features of digital library of University of Maribor are described in this paper. A short introduction describes some basic facts about the digital library and mentions its main purpose, but the main part of this paper is about features, that are mostly not found in other digital libraries. These features include integration with other information systems, plagiarism detection, informative and useful statistics about mentors and specific content extraction from documents, served by the digital library. We present existing functionality and describe some ideas for future development. A natural language processing framework, called TextProc, is also briefly mentioned, since it is used to perform plagiarism detection.

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

I. Introduction

DIGITAL library of University of Maribor is running approximately since autumn of 2008. Roughly at the same time, the University of Maribor published a revised regulation about final theses that now requires from all students to provide electronic version of their final thesis beside the printed one. Because of this, our digital library also enables students to upload their final theses. 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. This directly supports the main purpose of digital library of University of Maribor, which is collection and dissemination of final theses, created by students of all faculties of the University of Maribor. 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 at all times – all that is required is a web browser. Digital library also employs modern technologies like RSS (Really Simple Syndication), to inform users about newly published documents. Table 1 presents some statistics about number of published documents on October 2010, where column "New" means number of documents, added in the last month. The same table is also available on the first page of the digital library.

From technical point of view, the digital library of University of Maribor is a web application, build on top of the 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 I
NUMBER OF DOCUMENTS IN DKUM ON OCTOBER 2010

Document type	All	New
Bachelor theses	8313	371
Master theses	254	18
Doctoral theses	83	8
Other documents	50	0
All documents	8700	396

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.

In this paper we present 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 October 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.

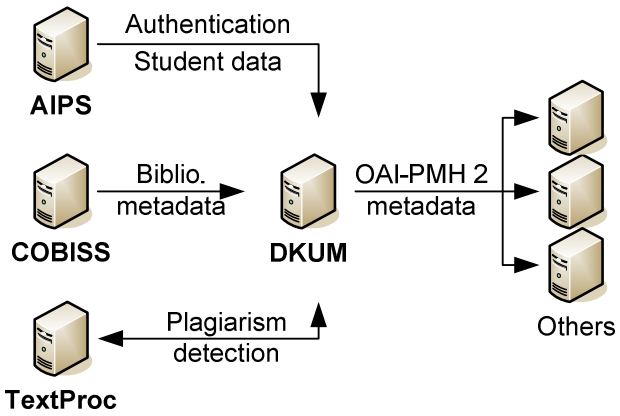

Fig. 1 DKUM integrated with others

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 (mostly English). 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 (preventing any changes), but not published.

Before documents are published, they are processed by local librarian. Each faculty has its own local library, so each faculty processes it 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 is 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. All that is required for an import or linking operation is the COBISS identifier of the bibliographic unit. DKUM can only perform read operations and is not allowed to make changes in COBISS.

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 repositories like DRIVER repository (Digital Repository Infrastructure Vision for European Research [5]), WorldCat [9], ROAR (Registry of Open Access Repositories [10]) and others. We implemented this protocol using an open source OAI PMH v2 data provider [6].

Currently we are working on a way to enable web sites to display new DKUM documents on their web pages. We will provide a small piece of JavaScript code that can be added to any web site and designed in any way they see fit. With minimal integration work, this will enable any site to display a personalized list of new DKUM documents on their pages. For instance, a faculty web site would so be able to display all new theses of this particular faculty. Any further clicks on this list will redirect to DKUM. Currently we will provide the data via existing RSS pages, which are limited in the kind of content they provide (only new documents by thesis type and organization). In the future the data will be provided using a web API [11] that will perform advanced search and will return its results in different formats, including JSON (JavaScript Object Notation) that is suitable for JavaScript

processing. With the introduction of a web API, we will also be able to extend the kind of information we provide to more than what RSS provides. For example, we will also provide lists of new thesis of a specific mentor that could be displayed on mentors personal home page or his laboratory home page. This will not only spread the word about DKUM, but will also enable the development of new web applications known as mashups [12], where DKUM will serve as a data source.

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 the given person is a mentor or co mentor per faculty;
- table with bar chart showing number of theses published by year and thus showing the activity of the mentor (shown on Fig. 2);
- all keywords associated with the given mentor, sorted by frequency (shown on Fig. 3). Indirectly this list shows all research areas a selected mentor is or was 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.

Fig. 2 shows the number of theses published, where person Milan Ojsteršek was the mentor. All data prior year 2008 is inaccurate and a warning is displayed to the user (not shown on Fig 2.). Numbers about master and doctoral theses are also not accurate.

Fig. 3 shows all the keywords of all theses of a given mentor. Keywords are sorted by frequency. List of keywords with frequency of 1 is long and was cut short for this paper.

Person: **Milan Ojsteršek**

Year	BSc	MSc	Ph.D.	Σ
2010	3	0	0	3
2009	3	**1**	0	4
2008	4	0	0	4
2007	2	0	0	2
2006	2	0	0	2
2005	**7**	0	0	**7**
2004	5	0	0	5
2003	4	0	0	4
2002	2	0	0	2
2001	3	0	0	3
2000	4	0	0	4
1999	2	0	0	2
1998	1	0	0	1
1997	1	0	0	1

Fig. 2 Number of theses published per year for a selected person; types of theses are (from left to right) bachelor, master and doctoral thesis

Person: **Milan Ojsteršek**

There are **192** keywords, that occur **243** times.
26 keywords (13.4% of all) are together repeated 75 times (30.86% of all).

Frequency Num.	%	Keywords
8x	10.67%	XML
6x	8%	web applications
5x	6.67%	internet
4x	5.33%	ASP, COM
3x	4%	XSL, natural language processing, world wide web, electronic exams, systems
2x	2.67%	protection of personal data, security, Web forms, ASP.NET, informatic Internet applications, e-education, search engines, notification service Slovenian language
1x		framework, Microsoft. Net, web application, personalize, cookies, filte framework, J2EE, ADO.NET, Linux, digital identities management, sing digital identity, electronic mail, parallel processing, distributed databa network, operating systems, servers, firewall...

Fig. 3: List of keywords associated with selected mentor

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 or applicative projects that require academic participation.

IV. Plagiarism Detection

A simplified definition of plagiarism is "copying of content from other authors and then calling it your own work". This is a kind of theft and is therefore illegal. With the rise of internet

and the availability of massive amounts of digital content, plagiarism is becoming a serious problem, since copying is extremely easily performed using copy/paste technique. This problem is also present in academia, especially in undergraduate studies, where students copy content from all kind of sources from the internet without properly referencing them. More about the problem of plagiarism in academia be read in [13][14] and the University of Maribor is not immune to it.

DKUM has the ability to perform plagiarism detection. Actually, DKUM doesn't do plagiarism detection by its 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). This is a generic plug-in and breaks the content up regardless of its meaning. Because of this, certain content is broken, that from a human perspective should not be, but this problem is then solved by the second plug-in.
- 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 plagiated and which are not; they only contain a similarity percentage, similar content and its position in documents (if requested). Similarity is computed as quotient of length of similar content and the entire content length, expressed as percentage. The value can be between 0 and 1, where zero means no similarity and 1 means exact, 100% copy.

Plagiarism detection reports 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. On the list, document pairs are displayed, sorted by similarity with most similar pairs at the top. If it is possible to determine which the source is and which is the copy (by publishing year), then this information is also shown. From this page, user can either display more data about a specific document or look at the detailed similarity report that looks the same as the second option.

Second option is via document record editing pages, as shown on Fig 4. Record editing is split into multiple pages, accessible via tabs. The last tab is "Plagiarism" and requires additional administrative permissions for access. On this page a detailed similarity report is displayed for a specific document, labeled "Document A". On the top of the page some basic data about this document is shown, like title, authors, mentor and how many similar documents there are. On the left side there is a list of documents that are similar with currently viewed document. This list is sorted by similarity (percentage of similar content) with most similar

documents at the top. This list can be filtered, displaying only copies or only source documents (originals). Clicking on any title in this list selects the "Document B". This fills the list on the right, showing all similar content in the order, as it appears in the document A. Content can be written exactly the same or slightly differently. Different contents are marked with read background and both versions are displayed for comparison.

Fig. 4: Detailed view on document similarity report

On Fig. 4 there are two samples of different content (the third one is not displayed entirely). In the first example there is a difference in the letter case; word "svetovne" is in the first sentence written in upper-case and in the second one in lower-case. In the second example there is a difference in the last two words of the sentences. In English, the first sentence ends with "lower prices" and the second one ends with "lower price".

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. We would also like to include other sources in Slovenian language, but those sources are only available as web pages. To include them in plagiarism, we require its content to be fetched, processed and stored in the plagiarism detection system's database. To achieve this we will probably require a web crawler, which will collect and also update this content on regular basis. One such crawler is described in [16]. This collected content, especially from Wikipedia, can also be used to enrich the content in DKUM (more about that later).

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 the following contents:

- table of content,
- table of figures,
- table of tables,
- all URL addresses, mentioned anywhere in the document,
- table of technical terms and acronyms,
- table of equations,
- list of references.

The mentioned software is capable to extract even more (like actual figures as images), but for now only above listed content is used in DKUM. These contents will be visible to end users as part of metadata display and will also be used for other purposes. The extraction software is, like TextProc, deployed separately and is integrated with DKUM via web services.

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 document 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. 5 (the tab "Secondary language" is selected). Document title and authors are always displayed at the top regardless of the selected tab, so that the user always knows what document he is looking at. Note the link "More about this mentor…" at the top: clicking on this link opens the page with mentor statistics.

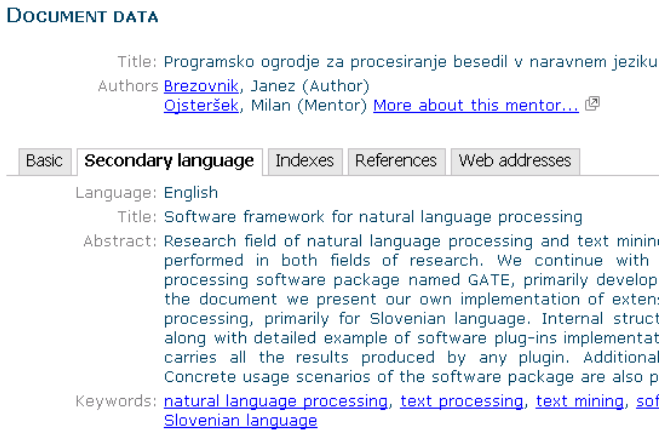

Fig. 5 Metadata display with tabs; content in foreign language is shown

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 the file and reading it. On Fig. 6 the extracted references from a document are shown. If there are web addresses in the references, then they are made clickable. Currently the feature of content extraction is still in the development and it is not jet publicly available (Fig. 6 was made on the development version of DKUM).

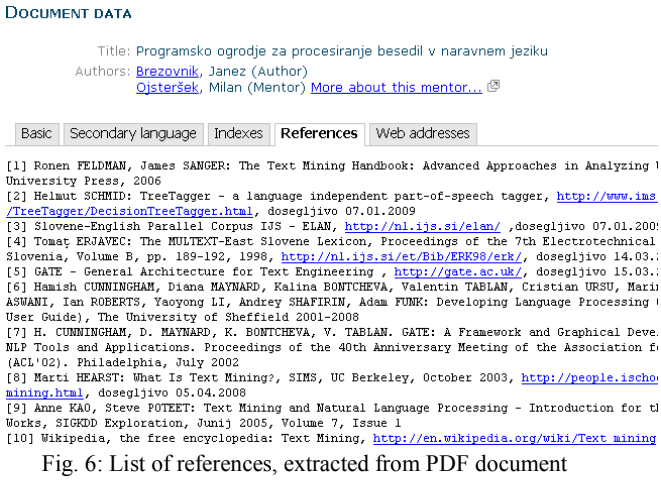

Fig. 6: List of references, extracted from PDF document

In the future, the extracted references will also be used for the following:

- If a reference refers to a document in the DKUM, then that reference (text of the reference) will be appropriately marked and will be clickable, thus displaying the data of the referred document.
- Counting references and displaying a list of referred documents from DKUM.
- Displaying a list of documents that share the same references (or any subset of them).
- Displaying a list of similar documents (in the sense that they discuss the same topic, not in the sense of plagiarism). Similarity can be determined in many ways using different data about the documents, like keywords, content and also references. If two documents are using the same references, they might be similar. We already did some research in this area and a prototype of document recommendation system exists for DKUM. When references are properly extracted, they can be used in the recommendation algorithm and hopefully improve the recommendations.
- For detection of citations in the process of plagiarism detection. It is legal to copy small pieces of the content from documents of other authors, if the copied content is properly marked and if a reference to the original source is added. Since citations are not plagiarism, they must be detected and marked in similarity reports. This is fundamental functionality in the plagiarism detection system, but still not implemented in our own. The citations won't be removed from similarity reports though; someone may be interested to view, what exactly was cited. It may happen, that the author adds a citation, but actually copies more content than is marked as citation.

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. In advanced search users can already specify which data to search for. A new data category will be added for searching equations, technical terms and acronyms, and search results will contain the searched data (actual equations or acronym explanations), not documents. The same could also be accessible via simple search using operators, like Google operator "define". Index of references could also be searchable. Search results would include all documents that use the searched reference.

VI. Future Development

There are many ideas for improvement of DKUM, from more basic to advanced, later ones involving natural language processing. Many ideas were already presented in previous chapters. Some of them presented here are drawn from [15].

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. Library of University of Maribor, who is also a member of DKUM, has several handwritten scripts that were digitized and they would like to publish those scripts as images in DKUM. Currently DKUM is only able to publish these images as a single big download and obviously this is not a good way to publish images. We are already planning a change in the way content is available in DKUM,

depending on the content type. In the case of a scanned document, available only as a set of images, those should be accessible as an image gallery. Each image should include tags (as part of metadata) that are searchable via simple or advanced search.

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 shown 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. In the chapter about integration we mentioned a way to show contents from DKUM on any other web page. The same solution will enable to show documents of a public shelve on any web page, for instance on a personal web page of a user, on professor's web page or on the web page of a specific course.

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, since there is no registration functionality and because we currently don't track anonymous users. For a time we were using an open source web analytics software called Piwik [17], but there were problems with logging custom data and the software itself was in early stages of development, so we temporally disabled this software. At the time of writing this paper, Piwik has already reached the version 1.0, so we will give it another try.

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, Eurovoc [7] 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 [8]. 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

The existing and future capabilities of DKUM are described in this paper. Main focus is on advanced features, not in the sense that they are hard to implement, but in the sense that they

are not often found in other digital libraries. Such features raise the bar of what digital libraries are capable of. 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 regarding natural language processing, since those documents are well structured and written in formal Slovenian language. They are also a credible source of knowledge, since final theses are reviewed by mentors, who are experts in their fields of research. This data will help in development of generic, language dependent natural processing features, which will be used in future research and applied projects. With these new features we will integrate TextProc with DKUM even further, making the knowledge in final theses from DKUM available also in other forms, not just as documents and not just for humans, but also for mashup applications, using data from DKUM.

REFERENCES

[1] COBISS: "About COBISS", http://www.cobiss.net/about_COBISS_Net.htm, visited on July 2010.
[2] "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification", NISO Press, Bethesda, Maryland, U.S.A., ISSN: 1041-5653, http://www.loc.gov/z3950/agency/Z39-50-2003.pdf, visited on July 2010.
[3] "Open Archives Initiative", http://www.openarchives.org/, visited on July 2010.
[4] "Open Arhive Initivative Protocol for Metadata Harvesting", http://www.openarchives.org/pmh/, visited on July 2010.
[5] Digital Repository Infrastructure Vision for European Research, http://www.driver-repository.eu/, visited on July 2010.
[6] H. Stamerjohanns: "Phpoai2 Data Provider", Carl von Ossietzky Universität Oldenburg, Institute for Science Networking, http://physnet.uni-oldenburg.de/oai/, visited on July 2010.
[7] "Eurovoc thesaurus", http://europa.eu/eurovoc/, visited on July 2010.
[8] I. Čeh, M. Ojsteršek, "Developing a Question Answering System for the Slovene Language", WSEAS Transaction on Information science and applications, Issue 9, Vol. 6, 2009.
[9] "WorldCat.org: The World's Largest Library Catalog", http://www.worldcat.org/, visited on September 2010
[10] "ROAR: Registry of Open Access Repositories", http://roar.eprints.org/, visited on September 2010
[11] Wikipedia: "Web API", http://en.wikipedia.org/wiki/Web_api, visited on September 2010
[12] Wikipedia: "Mashup (web application hybrid), http://en.wikipedia.org/wiki/Mashup_%28web_application_hybrid%29, visited on September 2010
[13] S. Carmen Cismas: "Anti-Plagiarism Strategies for Environment Engineering Students", Recent Advances in Energy & Environment, Proceedings of the 5th IASME / WSEAS International Conference on Energy & Environment (EE'10), 2010
[14] Z. Mahmood: "Students' Understanding of Plagiarism and Collusion and Recommendations for Academics" WSEAS Transactions on Information science and applications, Issue 8, Volume 6, August 2009
[15] B. Horvat, M. Ojsteršek: "Towards the Novel Classification Schemes in Digital Libraries", WSEAS transactions on information science and applications, December 2006, Volume 3, Issue 12
[16] S. Pohorec, M. Verlič, M. Zorman: "Local Search Engine with Global Content based on Domain Specific Knowledge", WSEAS Transaction on Information science and applications, Issue 1, Volume 6, January 2009
[17] "Piwik – Open source web analytics", http://piwik.org/, visited on October 2010

Janez Brezovnik holds a master's degree in computer science from Faculty of Electrical Engineering and Computer Science, University of Maribor since 2009. His major fields of study are web technologies and natural language processing.

He has 6 years of experience as a developer of web applications like content management systems, identity management systems, digital libraries and natural language processing software, the later as part of master thesis. His current job (since 2009) is teacher's assistant for computer science at Faculty of Electrical Engineering and Computer Science at University of Maribor, Smetanova ulica 17, 2000 Maribor, Slovenia. Previous research interest included identity management systems and web technologies, but now focuses on natural language processing and digital libraries.