Establishing of a Slovenian Open Access Infrastructure: a Technical Point of View

Milan Ojsteršek (Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia)
Janez Brezovnik (Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia)
Mojca Kotar (University Office of Library Services, University of Ljubljana, Ljubljana, Slovenia)
Marko Ferme (Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia)
Goran Hrovat (Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia)
Albin Bregant (Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia)
Mladen Borovič (Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia)

Citation:

Milan Ojsteršek, Janez Brezovnik, Mojca Kotar, Marko Ferme, Goran Hrovat, Albin Bregant, Mladen Borovič, (2014) "Establishing of a Slovenian open access infrastructure: a technical point of view", Program: electronic library and information systems, Vol. 48 Iss: 4, pp.394 - 412

Structured Abstract:

Purpose – This paper presents a technical perspective when implementing the Slovenian open access infrastructure that consists of four institutional repositories and a national portal that aggregates content from the repositories in order to provide a common search engine, recommendations of similar documents, and similar text detection.

Design/methodology/approach – During the project, the necessary legal background and processes for mandatory submissions of final study works, research publications and research data were established, as well as processes for data exchange between the institutional repositories and the national portal, and processes for similar text detection.

Findings – The consortium consisted of four Slovenian universities that significantly differ in size, organisation, and workflows. It was anticipated that exactly the same legal background and software would be used for the four repositories. It turned out that complete unification was impossible due to the differences.

Practical implication – The national open access infrastructure will improve the visibility of Slovenian research organisations. It supports the compliance with the funders' open access mandates. The established infrastructure enables the depositing and archiving of approximately eighty percent of the peer-reviewed scientific publications that are annually published by Slovenian researchers. At the same time, the majority of final study works from Slovenian higher education institutions are available in full-text format.

Originality/value – This paper describes a technical perspective for setting up a national open access infrastructure, which has not been described in the literature previously.

Keywords: national open access infrastructure, institutional repositories, open access, plagiarism detection, recommendation system

Article Classification: Research paper

1. Introduction

The possibilities for publishing research results have greatly improved since the development of the Internet. Search engines can provide links to publications within the research areas in less than a second. In the early 1980s, research organisations around the world had realized that they could improve their recognition by publishing their works on the Internet. Initially the majority of institutions published the final works of studies, research articles and technical reports on their websites or FTP servers, but later found that an institutional repository would be required to enable searching, browsing, and archiving (Kim, 2011). As the publications would

be accessible to a wider array of stakeholders, science would evolve faster and knowledge would no longer be a monopoly of the wealthy (Suber, 2012). In addition, students and mentors would be aware of open access to their final works of studies, and thus strive to improve their work. The open availability of final study works in digital format would also allow the detecting of similar contents.

A national network of repositories increases the visibility and impact of national research activities and provides additional services for visitors such as federated searches across all involved repositories and the archiving of publications. A more detailed presentation of this is given in »Recommendations for Implementation of Open Access in Denmark« [1] and in the recommendations of the European Commission for the EU Member States on establishing national open access policies and infrastructures [2]. In Europe, a few such national infrastructures are already established. These are the Dutch NARCIS [3], the Irish RIAN [4], the Norwegian NORA [5], the German OA-Netzwerk [6], the Greek openarchives.gr [7], the Italian PLEIADI [8], the Polish »Poland Digital Libraries Federation« [9], the Portuguese RCAPP [10], the Spanish RECOLECTA [11] and the Swedish DiVA [12]. These portals aggregate metadata and allow searches across all repositories included within their own infrastructures. In 2008, the formation of such an infrastructure took place in Hungary (Karacsony, 2013). At the time of writing this article, they hadn't as yet established a national aggregator and common search engine.

The Slovenian open access infrastructure, presented in this paper, consists of institutional repositories (IR) for each Slovenian university and a national portal (NP) that aggregates content from the repositories. This open access infrastructure has similar functionalities as the previously cited national open access infrastructures. A national portal aggregates metadata and full texts of research publications, final study works and research data from institutional repositories. It enables federated search, plagiarism detection, and recommendation of similar documents among institutional repositories. Institutional repositories enable the submissions of final works of study, research publications and research data, Web 2.0 functionalities (social bookmarking, RSS, JSON, JavaScript API, and the ranking of and commenting on publications), archiving, searching, taxonomy browsing, OAI-PMH and statistics (about archived publications, their authors and usage of metadata and content). This infrastructure does not yet support the enhanced publications, which are supported in [3], however this is planned for the future. It is also planned to offer the use of metadata for citation purposes in different citation styles (APA, Harvard, IEEE, etc.) and their exporting to different reference management tools (Zotero, Endnote, RefMan, RefWorks), as can be seen in [3] and [7]. Additionally, each item, deposited within one of the four repositories that currently comprise the Slovenian national open access infrastructure, is checked against similar content (plagiarism detection) within the existing corpus of texts. Amongst the previously cited national open access infrastructures, only the German OA-Netzwerk [6] supports the detection of similar documents. The Slovenian national infrastructure is the only one that has a recommendation system that provides recommendations across all institutional repositories. These recommendations consist of titles from within the same IR and titles from other IR's and participating digital archives (Digital Library of Slovenia [13], VideoLectures.NET [14], and DKMORS [15]). The authors believe that the Slovenian open access infrastructure has some distinct advantages such as: all four institutional repositories use the same custom-built software, which is integrated with information and authentication systems of the universities, with the COBISS.SI national bibliographic system [16], the SICRIS national current research information system [17] and the national portal Open Science Slovenia. The national portal and the four repositories are also available from mobile applications for Android, Windows Phone and iOS devices. These features have been unavailable in any other of the previously cited national infrastructures.

Prior to the implementation of this project, there were already some open access repositories and digital archives in Slovenia that contained the final study works and research publications. These contents were, in most cases, inserted by librarians and not by the authors. Most institutions did not define adequate processes and legal background that would enable authors to deposit their publications within the repositories. An exception was the Digital Library of the University of Maribor, where such a process had already been implemented, thus enabling students to submit their final study works to the digital library. As it was impossible to simultaneously search the contents from different repositories and digital archives and with an absence of processes and policies for the submission of content, the consortium of Slovenian universities has decided to sign up to a call from the Ministry of Education, Science and Sport and has received funding for the establishment of a national infrastructure of open access (from the Ministry and the European Regional Development Fund). This project started in February 2013. During its establishment valuable experiences were learned as described in the articles of numerous authors (Bevan, 2005; Barwick, 2007; Doctor and Ramachandran, 2008; Jantz and Wilson, 2008; Tripathi and Jeevan, 2011; Paul, 2012).

The national open access infrastructure is the underpinning element for the introduction of open access policies in any country (for EU Member States see the Commission Recommendation on access to and preservation of scientific information [2]). No open access policies were implemented in Slovenia prior to the establishment of the national open access infrastructure. The Slovenian Research Agency requires that final reports of co-financed projects are deposited into the Digital Library of Slovenia [13], as well as copies of co-financed scientific and professional journals. In 2013 Slovenia named a National Point of Reference for Scientific Information, who is responsible for coordinating the implementation of measures from the Commission Recommendation [2]. The European Commission expects that by 2014 the policies for open access to scientific articles and data will have been established in all Member States at all relevant levels [18]. Anecdotal evidence tells the authors that the discussions have already started in Slovenia about the formation of open access policies regarding research funding institutions and research performing institutions.

The aim of the infrastructure presented in this article, is to enable open access of the intellectual production of Slovenian universities to interested stakeholders in Slovenia and worldwide. This article focuses on a description of the Slovenian open access infrastructure and functionalities offered by the software in use. The second section of a paper presents the state of open access in Slovenia before the establishment of the national infrastructure. The architecture of the Slovenian infrastructure of open access is covered in the third section. The organisational structure and processes for the mandatory depositing of publications within

institutional repositories are described in the fourth section. The document similarity and plagiarism detection approach is presented in the fifth section. The sixth section describes the recommendation system, used in the repositories for providing content-based recommendations. The seventh section covers some conclusions, and guidance for further work.

2. Overview of the Slovenian open access repositories and document archives before the establishment of a national open access infrastructure

Prior to setting up the national infrastructure for open access, the sources were examined that could be incorporated within the infrastructure. Different Slovenian organisations were issuing over 40 openly accessible journals recorded in the Directory of Open Access Journals [19]. A few organisations were publishing openly accessible monographs (e.g. Educational Research Digital Library, which contained a few dozen works).

By 2012, Slovenian universities had already established several repositories and document archives that were mostly populated via mandatory submission of final study works and working copies of research publications. The Slovenian open access OAI-PMH repositories that were operating before the establishment of the national infrastructure for open access were *dLib.si* - »*Digitalna knjižnica Slovenije*« [13] (archived scientific articles from Slovenian scientific journals, final study works and project reports), DKUM - »*Digital Library of University of Maribor*« [20], *PeFprints* (archived documents from the Faculty of Education of the University of Ljubljana), *ePrints.FRI* (archived documents from the Faculty of Computer and Information Science of the University of Ljubljana), *DRUGG* - »*Digital repository UL FGG*« (archived documents from the Faculty of Civil and Geodetic Engineering of the University of Ljubljana). Other archives included *VideoLectures.NET* [14] (archived videos of lectures by lecturers from around the world), Journal of Management, Information Systems and Human Resources of the University of Maribor (archived papers from this journal), *DKMORS* - *Digital Library of the Ministry of Defence* [15] (archived publicly available publications from the Ministry of Defence) and *Social Science Data Archives* (archives research data) [21].

The two major non OAI-PMH compliant archives at the University of Ljubljana were the »*Faculty of Economics publications*«, and the »*Faculty of Social Sciences documents*«. At the University of Ljubljana, digital archives also existed for the faculties of Administration, Sports, Biotechnical, Pharmacy and Arts. At the University of Primorska there were archives at the Faculty of Management and the Faculty of Humanities. A digital archive was also established at the University of Nova Gorica.

3. Architecture of the Slovenian national open access infrastructure

Figure 1. A network diagram of the Slovenian open access infrastructure

Figure 1 presents a network diagram regarding the national open access infrastructure (Open Science Slovenia) [22]. Within the national infrastructure and along with the Digital Library of the University of Maribor [20], institutional repositories were established for the University of Ljubljana (RUL) [23], University of Primorska (RUP) [24], and the University of Nova Gorica (RUNG) [25]. In

addition to metadata of full-text publications from outside the repositories, RUL also contains full-text publications and metadata from ePrints.FRI, PeFprints, DRUGG and Social Science Data Archives (SSDA) [21]. Metadata is gathered via OAI-PMH servers. National portal Open Science Slovenia [22] aggregates content and metadata from all institutional repositories and partially from dLib.si [13], VideoLectures.NET [14] and DKMORS [15]. Additionally, SSDA, ePrints.FRI, PeFprints, DRUGG, dLib.si, DKMORS and VideoLectures.NET all have defined publication submission processes. Metadata can be gathered from these sources in XML and JSON formats. The Open Science Slovenia portal only takes metadata and full-texts of publications for the purposes of federated search, recommendations regarding publications and the detection of similar publications.

Institutional repositories store full-text versions of final study works, and research publications as well as other intellectual productions from universities (teaching materials, project reports, studies, monographs, research data ...). Institutional repositories are powered by the same software used to run the Digital Library of the University of Maribor (Brezovnik and Ojsteršek, 2011). The software has been substantially upgraded due to the need for new functionalities and regarding process support for student/staff submissions of publications.

The national infrastructure of open access exchanges metadata with the Slovenian bibliographic system COBISS.SI (Seljak and Seljak, 2002) via SRU/SRW services. COBISS.SI [22] consists of the local databases of libraries and a union catalogue that aggregates metadata from the local databases. Local databases belong to the libraries of individual academic institutions, however exceptions do exist where one database is linked to several academic institutions and where several databases are linked to a single academic institution. A part of COBISS.SI also contains authority control – a normalisation database of personal and corporate names (CONOR.SI), where information is stored about authors and their affiliations to organisations. CONOR.SI also holds records of the Slovenian Research Agency's identity numbers for researchers, which can be used for establishing connections with the SICRIS (Južnič et al., 2010). SICRIS holds data about researchers, research organisations, research groups and research projects, therefore being of service during the evaluations of researchers' performances. Researchers can have their publications catalogued in COBISS.SI and by doing so improve their grades in SICRIS, which thus affects their rankings when trying to obtain research funding grants or want to be promoted.

Institutional repositories consist of administrator and visitors interfaces. The administrator interface is intended for the Study Office personnel, librarians, and system administrators. The Study Office personnel overviews the submission of the final study works. Librarians catalogue the student and staff publications into COBISS.SI and transfer the catalogued metadata from COBISS.SI into institutional repositories. The administrator user interface enables the import of publication metadata using a COBISS.SI identification number. In this way, all metadata of catalogued publications within COBISS.SI can be transferred into institutional repositories. This is used for all publications in the digital format that were submitted to different archives before the establishment of institutional repositories, and are adequately copyrighted for the university to make them accessible through the university repository.

The user interface can be used by registered (students and university staff) and non-registered visitors (general public). Students and researchers can deposit their publications on the repository and review the content (metadata and similar documents found by the detector of similar documents). They also have the possibilities of using Web 2.0 functionalities (social bookmarking, ranking of and commenting on publications). The user interface for the general public is bilingual (Slovenian and English) and is available as a Web application and as native mobile applications for Android, iPhone and Windows Phone platforms. The interface is adjusted to persons with disabilities in regard to the WAI specifications [26]. Faculties can include some of the functionalities from the repositories using a JavaScript API, which is also used by mobile applications. Metadata can be exported in a variety of formats including RSS, JSON, and RDF.

Green open access – publishing in traditional or open scientific journals and submissions to institutional or subject repositories is encouraged by the European Commission. The Horizon 2020 funding programme determines mandatory open access to all publications from co-financed EU projects. The European Commission co-financed the European infrastructure of open access for research publications (projects OpenAIRE (2009-2012) and OpenAIREplus (2011-2014)). Therefore, the established institutional repositories include OAI-PMH servers that return OpenAIRE compatible XML [27], thus enabling OpenAIRE servers to gathered metadata about Slovenian publications that were produced as the results of projects financed by the EU.

4. Organisational structure and processes for the mandatory depositing of publications in Slovenian institutional repositories

The ODUN project had a project manager and project team heads at each university. Each university project team head supervised a project group for policies and processes' definitions, and a project group for the information infrastructure. Each of the project group heads choose their own project co-workers. The project groups for policies and processes' definitions cooperated with offices, responsible for research or teaching at the respective universities. The same organisational structure is still in place today. Additionally, a group has been established consisting of representatives from all four universities. This group is responsible for the continued operation and development of the infrastructure. Each university also employs an IT specialist who is responsible for

technical maintenance and support of the IR. Questions and problems are handled by the content administrators at universities who, if necessary, contact the student offices or librarians at faculties or academies. The changes regarding libraries after establishing the national infrastructure included assistance to students and researchers in understanding the newly established open access environment (understanding processes regarding mandatory depositing of publications within institutional repositories, plagiarism prevention and detection, preparation of final study works according to university templates, etc.) and to assist researchers in clearing the copyrights issues. Actual operational activities, based on organisational changes, both at the national level and at individual repositories, are not fully determined as yet. Each university has also appointed a group for the advocacy and dissemination of information on open access to all stakeholders (university managements, researchers, students, librarians).

Within the project, legal background of final study work submissions were analysed in regard to their archiving within the institutional repositories and open access. Appropriate rules were newly created or revised. This included rules about a mandatory copy and instructions on the preparation of bachelor, masters and doctoral theses.

The process, which was established at universities, allows authors to submit their work into the institutional repository. After which, the submission is reviewed and catalogued into COBISS.SI by a librarian. Based on the experiences accumulated since 2008, it was found that a manual review of metadata is mandatory as in many cases the submissions are incomplete or contain spelling errors. In addition, only librarians can normalise those authors using the CONOR.SI authority database. After the publication metadata have been successfully catalogued within the COBISS.SI, they can be transferred to the institutional repository via SRU/SRW services.

The working group, which consisted of librarians, legal experts and IT professionals from all four Slovenian universities, suggested the submission processes of the final study works. Slovenia lacks a common university academic information system (UAIS), therefore each university provided its own variation of this process. Students at the universities of Maribor and of Nova Gorica submit their final study works into IR, which is filled with some metadata from UAIS. A sequence diagram of the final study work publications of the universities of Maribor and of Nova Gorica is presented in Figure 2. The submissions at the universities of Ljubljana and of Primorska take place at UAIS, the sequence of operations is otherwise the same as described for the universities of Maribor and of Nova Gorica.

Figure 2. A sequence diagram of final study work submission and publication at the universities of Maribor and of Nova Gorica

Figure 3. A sequence diagram of research item submission and publication

All four Slovenian universities have established the same process for research publications submissions, which is presented the sequence diagram in Figure 3. Researchers can submit articles, monographic chapters, monographs, conference papers, study materials, publications about patents, research data, and other types of publications. The types of publications are adapted according to the COBISS.SI typology, which is used by the Slovenian Research Agency for researchers' bibliographies evaluation. Once researchers are logged into the IR, they can submit new content as a whole, or they can use metadata from the catalogued record in COBISS.SI (optional request and replay on the beginning of the process). In the latter case, the IR takes care of the metadata transfer from COBISS.SI using the SRU/SRW protocol. In this case, researchers only need to provide the electronic version of their publications. A link to the SHERPA/RoMEO portal is also enabled to the publication authors, so they can check what type of access they can use depending on the publisher's copyright transfer agreement. During the insertions of names and surnames, suggestions from the CONOR.SI database are provided. These suggestions include the year of birth, if available in CONOR.SI, and researcher identifier from SICRIS, which simplifies the determination of the correct author. This can greatly simplify the librarian's work of cataloguing the publication in COBISS.SI. Authors can also determine the copyright holder and the type of access to the full-text publication. They can choose between immediate publication, closed access or delayed publication with embargo (these metadata are part of OpenAIRE compliance).

5. Similar text detection

Increases in copying from published texts without proper citation (i.e. plagiarism) have been noticed over the recent years. This problem has increased due to the growing number of documents available on the internet. The reason for improper citation is either lack of knowledge or deliberate copying of individual sentences or even chapters from the works of other authors (Alzahrani et al., 2012).

Finding similarities between documents is performed over two steps. The first step is document fingerprinting. Those documents that are similar to the document to be checked are found using algorithms which compare different document features. These features can be retrieved from the document itself (from chapters, paragraphs, sentences, fixed number of words) using hashing algorithms or using frequency analysis of words or phrases, which are obtained by semantic tagging. Features retrieved using hashing algorithms (Stein, 2007), are compared and the result is a list of documents and the percentages of coverage between features of document pairs. Vector-space model algorithms (e.g. LSA, BM25, Random indexing) and text classification algorithms (k-nearest neighbours, Naive Bayes, SVM) use frequency analysis of words or phrases (Alzahrani et al., 2012). The result is a set of candidate documents. This can be a fixed number of documents (e.g., first 50 documents ordered by similarity of a coverage percentage between document features). In the second step, called "pairwise feature-based exhaustive analysis", pairs of documents are checked using the longest common substrings. Several algorithms exist for this task, as described in a survey article (Navaro, 2001).

Document similarity detection (or plagiarism detection) is done within the national portal. The results of detections are available to students and staff. They can only see those documents where they had the roles of author, co-author or mentor. The detection is performed for each publication, submitted into the university repositories. Both coarse-checking (step 1) and fine-checking (step 2) analyses are supported by the custom-built software of the authors. The software does not check the similarities between images.

Software for coarse-checking texts returns similar sentences that are longer than forty characters (example of a result is shown on figure 4 - above). It enables coarse-checking due to its use of upgraded TextProc (Brezovnik and Ojsteršek, 2011). The language infrastructure used during the plagiarism detection process, consists of a morphological dictionary for the Slovenian language, which contains of approximately 8,000,000 word forms and 320,000 lemmas. Wikipedia labels from article titles from Slovenian, English and German Wikipedia are also used. They were extracted from Dbpedia (Morsey et al., 2012). A domain-specific semantic dictionary was created using keywords from publications metadata within the Open Science Slovenia portal. Sentences, marked by the coarse-checking software as similar, are undoubtedly the same in both texts. They differ only if the authors used synonyms, used a different grammatical person or used filler words (e.g. »therefore«, »however«). The software detects similarity even if the word order is changed or if any of the words are misspelled. The Coarse-checking algorithm (Brezovnik and Ojsteršek, 2011) which has been subsequently updated, first converts the text into UTF-8 format, eliminates extra whitespaces and new line characters (CR, LF). Then it splits the content into sentences. Words from these sentences are then lemmatised. Common words (e.g. »and«, »or«, »that«, etc.) are filtered out and all the remaining words are sorted alphabetically. This step also carries out spelling corrections using a morphological dictionary and POS tagger. In order to correct spelling errors the "Symmetric Delete Spelling Correction Algorithm" [28] is used. After the lemmatisation, the algorithm normalises those synonyms stored within the dictionary and can be transformed into single forms without changing the semantic meanings. A good example of this are the normalisations of the verbs »to present«, »to describe« and »to show«, which are synonyms in most cases. The algorithm then calculates hashes for these sentences. Finally, it compares the hashes of sentences from other documents within the corpus of documents, calculates the coverage of hashes for document pairs and provides a list of similar documents for every document within the corpus.

<p>Poročilo</p> <p>Levi dokument: Primer 2- plagiat.docx Desni dokument: Primer 1- plagiat.docx Pokritost v levem: 50,00 % Pokritost v desnem: 49,85 %</p> <p>Design/methodology/approach – During the project, the necessary legal background and processes for mandatory submissions of final study works, research publications and research data were established (343 znakov)</p>	<p>Purpose – The paper describes a technical perspective of implementation of the Slovenian open access infrastructure, which consists of four institutional repositories and a national portal that gathers content from the repositories in order to provide a common search engine, recommendation of similar documents, and similar text detection.</p> <p>Design/methodology/approach – During the project, the necessary legal background and processes for mandatory submissions of final study works, research publications and research data were established as well as processes for data exchange between the institutional repositories and the national portal, and processes for similar text detection.</p>	<p>Purpose – The paper presents a technical perspective of implementation of the Slovenian open access infrastructure, which consists of four institutional repositories and a national portal that aggregates content from the repositories in order to provide a common search engine, recommendation of similar documents, and similar text detection.</p> <p>Design/methodology/approach – During the project, the necessary legal background and processes for mandatory submissions of final study works, research publications and research data were established as well as processes for data exchange between the institutional repositories and the national portal, and processes for similar text detection.</p>
<p>Poročilo fine primerjave</p> <p>Levi dokument: Primer 2- plagiat.docx Desni dokument: Primer 1- plagiat.docx Pokritost v levem: 97,24 % Pokritost v desnem: 96,96 %</p> <p>s content from the repositories in order to provide a common search engine, recommendation of similar documents, and similar text detection. Design/methodology/approach – During the project, the necessary legal background and processes for mandatory submissions of final study works, research publications and research data were established as well as processes for data exchange between the institutional repositories and the national portal, and processes for similar text detection.</p> <p>s a technical perspective of implementation of the Slovenian open access infrastructure, which consists of four institutional repositories and a national portal that (165 znakov)</p>	<p>Purpose – The paper describes a technical perspective of implementation of the Slovenian open access infrastructure, which consists of four institutional repositories and a national portal that gathers content from the repositories in order to provide a common search engine, recommendation of similar documents, and similar text detection.</p> <p>Design/methodology/approach – During the project, the necessary legal background and processes for mandatory submissions of final study works, research publications and research data were established as well as processes for data exchange between the institutional repositories and the national portal, and processes for similar text detection.</p>	<p>Purpose – The paper presents a technical perspective of implementation of the Slovenian open access infrastructure, which consists of four institutional repositories and a national portal that aggregates content from the repositories in order to provide a common search engine, recommendation of similar documents, and similar text detection.</p> <p>Design/methodology/approach – During the project, the necessary legal background and processes for mandatory submissions of final study works, research publications and research data were established as well as processes for data exchange between the institutional repositories and the national portal, and processes for similar text detection.</p>

Figure 4. An example of coarse-checking (above) and fine-checking (below) outputs between two texts

Those documents that are found to be more than 1 % similar to the reviewed document during the coarse-checking step become candidates for entering into the fine-checking step. If the number of these candidates is less than 50, the rest of the similar documents are retrieved using the BM25 ranking function (Robertson et al., 2004). The fine-checking algorithm (example of a result is shown in Figure 4 – below) finds the longest common sub sequences between two texts. Kärkkäinen's algorithm (Kärkkäinen et al., 2009) is used for finding common sub-sequences greater than 14 characters. Another algorithm (Ferme and Ojsteršek, 2011) is used in cases where the pairs of documents have more than 60 % of coarse-checking coverage of hashes. This algorithm has a time complexity of nearly $O(N)$ if the pairs of documents are very similar. No spelling corrections are carried out for those texts using 'as-is'. If two documents are fine-checked (Figure 4 - below), the software marks those phrases or parts of sentences that are the same within both documents. A reviewer's task is to determine whether a sentence or a paragraph has been copied. Some sentences can be semantically identical as a whole or in part if the author has paraphrased the copied content. Similarity constants

(1% of coarse-checking coverage, 50 candidate documents for fine-checking, sentence length of 40 characters, subsequence length of more than 14 characters, etc.) are selected on the basis of experiences from examination plagiarism cases.

6. Recommendation system

The main goal of recommendation systems is to provide the visitors of the institutional repository with contents covering their interests. There are several approaches to recommendations, mostly divided into two categories. The first category of approaches operates exclusively over visitor's activities (Su and Khoshgoftaar, 2009). It uses the known activities of a group of visitors to make recommendations or predictions of the unknown activities for other visitors. Amongst the most widely used recommendation algorithms are collaborative filtering, binary vector approaches and SlopeOne. The second category of recommendation algorithms work solely with content, whilst the user activities are not of such importance but can be used to further weight the results during ranking. Similar text search falls within the area of text classification. Some of the algorithms that enable similar content recommendations are BM25, k-nearest neighbours, and latent semantic analysis. In addition, recommendation systems can differentiate between whether recommendations are calculated in real-time (memory based recommendation), or are pre-calculated (model based recommendation), and even a hybrid approach can be used (Bobadilla et al., 2013).

The objective of the recommendation system is to enable the visitors of institutional repositories to find similar documents after clicking onto a document within the institutional repository. Partial duplicates are omitted from the recommendations. Partial duplicates are determined using similar sentence and substring detection. If two documents have a sentence-based or substring-based coverage value of more than 60 %, they are marked as partial duplicates. The recommendation software includes content-based document recommendation that uses the BM25 ranking function and utilises additional weights during ranking (Borovič, 2012). Firstly, the metadata for each publication is obtained (authors, title, keywords, and abstract). Secondly, the metadata and the full-text of the publication are lemmatised. By utilising Wikipedia articles, a semantic tagging process (Burjek, 2011) of metadata and full-text of publications is used during the third step. Term frequencies (TF) and inverse document frequencies (IDF) are calculated for each publication during the fourth step. TF and IDF weights are determined from semantically tagged metadata and full-text. Similarity with other publications is then calculated using a BM25 ranking function suggested by Robertson, Zaragoza and Taylor (Robertson et al., 2004). Those pairs of documents with similarity values of 0 are discarded, as they are dissimilar. The result is a list of similar document pairs, which is then stored within the database. The recommendation threshold is set depending on the BM25 values of the documents on the recommendation list. Thus, the recommendation of similar publications is a result of selecting the top five highest ranking publications that exceed the threshold on the list, as ordered by the BM25 value. Additionally, other criteria such as the issue year, number of downloads, number of views and average rating, are used during the ranking process. A recommendation list could also be empty if the system cannot find any similar publications. The essential task is to maintain the database with up-to-date similarities when new documents are added to the system.

Content-based recommendation within the Slovenian national open access infrastructure is carried out at the national portal. On viewing a document within the institutional repository, the national portal sends a list of similar documents. Two different kinds of recommendation lists are sent. The first consists of similar documents within the institutional repository. The second shows similar content from across all other repositories including dLib.si, VideoLectures.NET and DKMORS. Both recommendation lists are cached for each document and stored within a database for enabling real-time responses.

7. Conclusions and future work

The open accessibility of Slovenian research publications and data will increase the visibility and impact of Slovenian researchers and research organisations. It is expected that more opportunities for cooperation with universities worldwide will appear in higher education and research activities. Slovenian and foreign companies will be able to use the results of publicly funded research from Slovenian research organisations, therefore improving the efficiency of public funding. The national open access infrastructure will also influence the Slovenian research policy. The Ministry of Education, Science and Sport and the Slovenian Research Agency are initiating discussions on the adoption of open access policies, aligned with the Horizon 2020 mandate on open access.

The ranking of repository contents in popular search engines will be improved by listing the repositories in the open access registries (ROAR, OpenDOAR, BASE, DART-Europe) and WORDCAT. The increased number of incoming links after the submission of new publications into the repositories will have an impact on the access statistics for university domains. Consequently, the Webometrics ranks of Slovenian universities might improve.

At all four universities, the legal backgrounds have been prepared or revised to support a mandatory process of a final study works and research publications submission and plagiarism detection process. Using the plagiarism detector, the number of those wishing to attain academic achievements in a non-ethical way is reduced. Students and researchers are encouraged to produce new, original ideas and to cite previously published ideas accordingly. Open access to the final works of studies influences the quality of peer reviewing the theses by their supervisors and other faculty staff involved during the research process. The qualities of the final works have increased since the universities have established theses submission processes. Plagiarism is being detected and prevented, thus ensuring better technical, citation and linguistic qualities of theses.

National open access infrastructure enables permanent access to knowledge from anywhere, via any mobile, embedded device or a web browser. Anyone can benefit from the results of education and research, especially those who have limited possibilities to access the information society benefits (people with disabilities, the elderly and other social groups that have limited access to e-services due to physical or other barriers).

Compared to establishing several repositories for individual faculties or academies, the implemented solution with a single repository for each university is more effective from the financial, time wise, and human resource perspectives. The links between the repositories and information systems of universities are operational and a national portal has been developed. In addition, the appropriate compatibility and interoperability is available for the direct participation of repositories and the portal in international initiatives.

The Slovenian national open access infrastructure is designed according to the latest international guidelines for repositories and national open access infrastructures, as well as to the recommendations of the European Commission regarding open access to the results of publicly funded research. The repositories of the universities meet the open access requirements of the EU funders.

Since the beginning of its establishment (15th October 2013), the number of publications submitted has increased on average by 1,142 per month. On 31st March 2014, the national infrastructure had 115,412 documents available. Between 24th March and 31st March 2014 (a week before the revision of this paper), 123,233 metadata records have been viewed and 30,896 files have been downloaded from the four institutional repositories. Between the 1st March and 31st March 2014 (a week before the revision of this paper), 669,820 publication metadata were viewed and 113,335 files have been downloaded. Up to the 31st March 2014 the mobile applications had been downloaded by 184 visitors (iPhone: 92, Android: 16, Windows Phone: 76). The DKUM web server statistics show that Android is used by 2.9 % of visitors, while iOS is used by 2.6 % of visitors. Between 1st March and 31st March 2014, the national portal search function was used by 4,828 visitors which generated 44,304 searches during that time. The national portal will need to be promoted to the visitors of institutional repositories. The institutional repositories are more used since the researchers and students are more familiar with them. The recommendation system is being frequently used, due to the fact that visitors tend to look for similar publications. Between 1st March and 31st March 2014, on average 2.63 recommended publications were viewed. The statistics in use discard the most frequent requests from all known web crawlers, which are stored in our crawler database.

Following on from this project, the universities intend to expand the initial infrastructure and integrate the remaining higher education and research institutions within the national open access infrastructure. An additional repository will be set-up for organisations without their own repositories. The main future activities of the Slovenian national open access infrastructure will be:

- inclusion of new features for registered visitors such as personalisation of content in search results, personal bookmarks, repairing of OCR generated text, etc.,
- enabling display of publication metadata in different citation styles (APA, Harvard, IEEE, etc.) or export to different reference management tools (Zotero, Endnote, RefMan, RefWorks),
- establishment of digital preservation of publications,
- enabling enhanced publications,
- including metadata of research datasets within the national portal from the national infrastructure of open research data (when it will be built),
- publishing metadata as linked open data (firstly, the data will be linked with the UDC taxonomy and info:eu-repo ontology [29]; furthermore, linkages with other taxonomies and ontologies will be done, such as Eurovoc, Agrovoc, MESH, and Dbpedia (Morsey et al., 2012)),
- enabling the partially automatic extraction of keywords and publication classification using known taxonomies (CERIF, ACM, UDC, ISCED, LCH, DDC, Eurovoc, Agrovoc, MESH...),
- enabling the extraction of additional metadata and knowledge in publications stored within the repositories and
- enabling publication linkage, archived in national infrastructure, according to time, spatial or semantic features.

Notes

1. Recommendations for Implementation of Open Access in Denmark, available at: <http://fivu.dk/en/publications/2011/files-2011/recommendations-for-implementation-of-open-access-in-denmark-final-report-from-the-open-access-committee.pdf> (accessed 17th January 2014).
2. Commission Recommendation of 17. 7. 2012 on access to and preservation of scientific information, available at: http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf (accessed 17th January 2014).
3. NARCIS: the gateway to scholarly information in the Netherlands, available at: <http://www.narcis.nl/> (accessed 17th January 2014).
4. Rian.ie: pathways to Irish Research, available at: <http://rian.ie/> (accessed 17th January 2014).
5. NORA: Norwegian Open Research Archives, available at: <http://www.ub.uio.no/nora/search.html?siteLanguage=eng> (accessed 17th January 2014).
6. OA-Netzwerk: German open access infrastructure, available at: <http://oansuche.open-access.net/oansearch/> (accessed 17th January 2014).

7. Openarchives.gr: Greek open access infrastructure, available at: <http://www.openarchives.gr/> (accessed 17th January 2014).
8. PLEIADI: Portale per la Letteratura scientifica Elettronica Italiana su Archivi aperti e Depositi Istituzionali, available at: <http://www.openarchives.it/pleiadi/> (accessed 17th January 2014).
9. Digital Libraries Federation - Federacja bibliotek cyfrowych, available at: <http://fbc.pionier.net.pl/owoc> (accessed 17th January 2014).
10. RCAAP: Repositório Científico de Acesso Aberto de Portugal, available at: <http://www.rcaap.pt/> (accessed 17th January 2014).
11. RECOLECTA: Recolector de Ciencia Abierta, available at: <http://www.recolecta.net/buscador/index2.jsp> (accessed 17th January 2014).
12. DiVA: Digitala Vetenskapliga Arkivet, available at: <http://www.diva-portal.org/smash/search.jsf> (accessed 17th January 2014).
13. Digital Library of Slovenia (dLib.si), available at: <http://www.dlib.si/?&language=eng> (accessed 17th January 2014).
14. VideoLectures.NET, available at: <http://videolectures.net/> (accessed 17th January 2014).
15. DKMORS – Digital Library of the Ministry of Defence, available at: <http://dk.mors.si/info/> (accessed 17th January 2014).
16. COBISS.SI – Slovenian Cooperative Online Bibliographic System and Services, available at: http://www.cobiss.si/cobiss_eng.html (accessed 17th January 2014).
17. Slovenian Current Research Information System (SICRIS), available at <http://www.sicris.si/default.aspx?lang=eng> (accessed 17th January 2014).
18. Towards better access to scientific information: boosting the benefits of public investments in research, available at: http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf (accessed 1st April 2014).
19. Directory of Open Access Journals, available at: <http://www.doaj.org/> (accessed 17th January 2014).
20. Digital Library of University of Maribor, available at: <http://dkum.uni-mb.si/lskanje.php> (accessed 17th January 2014).
21. Social Science Data Archives (SSDA), available at: <http://www.adp.fdv.uni-lj.si/eng/> (accessed 17th January 2014).
22. Open Science Slovenia: Access to Knowledge from Slovenian Research Organisations, available at: <http://www.openscience.si/default.aspx> (accessed 17th January 2014).
23. Repository of the University of Ljubljana, available at: <http://repozitorij.uni-lj.si/info/index.php/eng> (accessed 17th January 2014).
24. Repository of the University of Primorska, available at: <http://odun.univerza.si/info/index.php/eng/> (accessed 17th January 2014).
25. Repository of the University of Nova Gorica, available at: <http://repozitorij.ung.si/info/index.php/slo/> (accessed 17th January 2014).
26. Web Accessibility Initiative, available at: <http://www.w3.org/WAI/> (accessed 17th January 2014).
27. OpenAIRE Guidelines: For Literature repositories, available at: https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_Literature_repositories (accessed 17th January 2014).
28. 1000x Faster Spelling Correction algorithm, available at: <http://blog.faroo.com/2012/06/07/improved-edit-distance-based-spelling-correction/> (accessed 17th January 2014).
29. Description of "info:eu-repo" ontology, available at: <http://wiki.surf.nl/display/standards/info-eu-repo> (accessed 17th January 2014).

References

- Alzahrani, S. M., Salim, N. and Abraham, A. (2012), "Understanding plagiarism linguistic patterns, textual features, and detection methods", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 42 No. 2, pp. 133-149.
- Barwick, J. (2007), "Building an institutional repository at Loughborough University: some experiences", *Program: electronic library and information systems*, Vol. 41 No. 2, pp.113-123.
- Bevan, S. J. (2005), "Electronic thesis development at Cranfield University", *Program: electronic library and information systems*, Vol. 39 No. 2, pp. 100-111.
- Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A. (2013), "Recommender systems survey", *Knowledge-Based Systems*, Vol. 46 No. 7, pp. 109-132, available at: <http://dx.doi.org/10.1016/j.knosys.2013.03.012> (accessed 17th January 2014).
- Brezovnik, J. and Ojsteršek, M. (2011), "Advanced features of Digital library of University of Maribor". *International journal of education and information technologies*, 2011, Vol. 5 No. 1, pp. 34-41, available at: <http://www.naun.org/main/NAUN/educationinformation/19-520.pdf> (accessed 17th January 2014).

- Burjek, M. (2011), *Wikification of content in digital library of University of Maribor*, BSc thesis, University of Maribor, Slovenia, in Slovene language, available at: <http://dkum.uni-mb.si/lzpisGradiva.php?id=20570> (accessed 17th January 2014).
- Borovič, M. (2012), *A document recommendation system and content-based recommendation quality analysis with different text processing methods applied*, BSc thesis, University of Maribor, Slovenia, in Slovene language, available at: <http://dkum.uni-mb.si/lzpisGradiva.php?id=37811> (accessed 17th January 2014).
- Doctor, G. and Ramachandran, S. (2008), "Considerations for implementing an institutional repository at a business school in India", *International Journal of Information Management*, Vol. 28 No. 5, pp. 346-354, available at: <http://dx.doi.org/10.1016/j.ijinfomgt.2007.12.001> (accessed 17th January 2014).
- Ferre, M. and Ojsteršek, M. (2011), "Text analysis with sequence matching", *International Journal of Computing*, Vol. 5 No. 2, pp. 235-242, available at: <http://www.naun.org/main/NAUN/computers/20-234.pdf> (accessed 17th January 2014).
- Jantz, R. C. and Wilson, M. C. (2008), "Institutional Repositories: Faculty Deposits, Marketing, and the Reform of Scholarly Communication", *The Journal of Academic Librarianship*, Vol. 34 No. 3, pp. 186-195, available at: <http://dx.doi.org/10.1016/j.acalib.2008.03.014> (accessed 17th January 2014).
- Južnič, P., Pečlin, S., Žaucer, M., Mandelj, T., Pušnik, M., Demšar, F. (2010), "Scientometric indicators: peer-review, bibliometric methods and conflict of interests", *Scientometrics*, Vol. 85 No. 2, pp. 429-441. Available at: <http://link.springer.com/article/10.1007%2Fs11192-010-0230-8#page-1> (accessed 17th January 2014).
- Karacsony, G. (2013), "HUNOR: The Collaboration of Hungarian Open Access Repositories", *Procedia - Social and Behavioral Sciences*, Vol. 73, pp. 57-61, available at: <http://dx.doi.org/10.1016/j.sbspro.2013.02.019> (accessed 17th January 2014).
- Kärkkäinen, J., Manzini, G. and Puglisi, S. (2009), "Permuted Longest-Common-Prefix Array", in *Proceedings of the Combinatorial Pattern Matching*, Lecture Notes in Computer Science, Vol. 5577, Springer-Verlag, Berlin Heidelberg, pp. 181 – 19, available at: http://link.springer.com/chapter/10.1007%2F978-3-642-02441-2_17#page-1 (accessed 17th January 2014).
- Kim, J. (2011), "Motivations of Faculty Self-archiving in Institutional Repositories", *The Journal of Academic Librarianship*, Vol. 37 No. 3, pp. 246-254, available at: <http://dx.doi.org/10.1016/j.acalib.2011.02.017> (accessed 17th January 2014).
- Morsey, M., Lehmann, J., Auer, S., Stadler, C., and Hellmann, S. (2012), "Dbpedia and the live extraction of structured data from Wikipedia", *Program: electronic library and information systems*, Vol. 46 No. 2, pp. 157-181, available at: <http://www.emeraldinsight.com/journals.htm?articleid=17030684> (accessed 17th January 2014).
- Navarro, G. (2001), "A guided tour to approximate string matching", *ACM computing surveys*, Vol. 33 No. 1, pp. 31-88.
- Paul S. (2012), "Institutional Repositories: Benefits and incentives", *The International Information & Library Review*, Vol. 44 No. 4, pp. 194-201, available at: <http://dx.doi.org/10.1016/j.iilr.2012.10.003> (accessed 17th January 2014).
- Robertson, S., Zaragoza, H. and Taylor, M. (2004), "Simple BM25 extension to multiple weighted fields", in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM, New York, pp. 42-49.
- Seljak M. and Seljak T. (2002), "The development of the COBISS system and services in Slovenia", *Program: electronic library and information systems*, Vol. 36 No. 2, pp. 89-98.
- Stein, B. (2007), "Principles of hash-based text retrieval", in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, pp. 527-534.
- Suber, P. (2012), *Open access*, MIT Press, Cambridge, Massachusetts, available at: [http://cyber.law.harvard.edu/hoap/Open_Access_\(the_book\)](http://cyber.law.harvard.edu/hoap/Open_Access_(the_book)) (accessed 17th January 2014).
- Su, X. and Khoshgoftaar, T. M. (2009), "A survey of collaborative filtering techniques". *Advances in artificial intelligence*, Vol. 2009 No. 1, pp. 1-19, available at: <http://dx.doi.org/10.1155/2009/421425> (accessed 17th January 2014).
- Tripathi, M. and Jeevan, V.K.J. (2011), "An Evaluation of Digital Libraries and Institutional Repositories in India", *The Journal of Academic Librarianship*, Vol. 37 No. 6, pp. 543-545, available at: <http://dx.doi.org/10.1016/j.acalib.2011.08.012> (accessed 17th January 2014).