INSTITUTIONAL REPOSITORY DRIVEN BY ACCESS RIGHTS AS A PART OF PLAGIARISM DETECTION SYSTEMS

Daniel Jakubík, Šimon Suchomel, Luboš Lunter, Michal Brandejs

Abstract: Masaryk University (MU) has developed an institutional repository with plagiarism detection service as an extension of the university's information system (IS). The repository enables various options of storing research and eventually publishes it in accordance with copyrights. Setting the access mode is managed by approval process, which is enabled by the repository. Therefore, the university had to set the rules and processes for proposing and approving the access modes in order to be able to set the proper access rights. The article advocates the hypothesis that the implementation of the university repository must focus not only on technical tasks, but also on methodological tasks. The paper describes both tasks and also the benefits of institutional repository driven by access rights deployment, where some files can be hidden for common users. Our approach is based on the idea that even the inaccessible files are usable in limited access mode and valuable sources for plagiarism detection tools and related services.

Key words: institutional repository; plagiarism detection; open access

1 Introduction

Institutional repositories are becoming more and more widely accepted components for preserving and disseminating accumulated data and knowledge in the form of scientific papers, proceedings, and less conventional genres like software or art-related works. As such they have a potential to be an important part of digital scholarly communication, and thus an accelerator of scientific progress. According to Hanlon and Ramirez (Ramirez and Hanlon, 2011), a key barrier to fulfill this potential lies in an insufficient intellectual property management. Our research aimed to establish institutional repository with embedded rights management and security mechanisms that balance a protection of property rights owners (e.g. publisher) and utilization of non-public works.

The purpose of this paper is to provide an overview of development of access rights driven repository with decision making features for access management. A general philosophy behind our solution is that the data should be maintained directly from the authors that can also participate on the rights management process. Therefore, a heavy emphasis is placed on simplicity and automated checking.

2 Background

IS MU has been under continuously development since 1999. It is focused on administration of university education, e-learning and other related areas of academic life. The significant objective of our research lies in plagiarism detection and e-learning applications to enhance academic integrity. (Lunter et al., 2013)

In 2004, Masaryk University provided its students and teachers an electronic archive of theses together with the plagiarism detection tool, which helped them to identify textually reused parts in theses. Since 2004 all students of MU are required to upload their thesis and metadata into the information system. Subsequently, a thesis is published in accordance with access rights. At that time, demand for a similar service arose at other universities as well, which represented an impetus for the joint project Theses.cz[1] and Odevzdej.cz[2]. The project has started in 2008 and involved MU as well as 16 other universities. Nowadays, there are 41 institutions involved.

After signing the Berlin Declaration in 2013 (Berlin Declaration, 2003), the development of the university repository was commenced. During its development, the experience from aforementioned projects were utilized, in order to build the repository on similar technologies. The repository was designed as an extension of registry of publications and results of scientific activities in the IS MU. Original purpose of the registry was to collect only metadata for scientific publications and the mandatory reporting. Although, the full texts are not required for government reporting, the archive of scientific production is beneficial for the university.

Main Benefits:

Plagiarism detection—Scientific papers are potential source of cheating and violation of academic integrity. Consequently, the extension of the database for similarity searching naturally increases the probability of a successful text reuse detection. In particular, papers with limited access are included when computing document similarities for plagiarism detection.

Visibility and citations elevation—As Olsbo shows (Olsbo, 2013), there could be a connection between the Internet visibility, ranking and the relative citation impact of universities in different countries. These relationships can be traced back to the effectiveness of the open access publishing, self-archiving and Open Access policies of the countries and the universities. (Hitchcock, 2013) The impact of Open Access model on the citation is described in the case study of Koler-Povh et al. (Koler-Povh et al., 2012)

Long term preservation—University's repository is the option how to ensure long term preservation of academic outputs. Even though that individual requirements for repositories may differ between institutions, a long-term preservation of their intellectual heritage is generally essential. (Conway et al., 2011)

E-learning—The advantage of integrated repository in university information system is possibility of accessing the papers, which could not be public due to license agreements.

For the universities, the main objectives of digital libraries lie in storing and disseminating the peer-reviewed knowledge and information of high standard. Naturally, there is an effort to make most of the records open to public, but it is not possible to fully achieve the goal for various reasons, such as publisher's copyrights or a disclaimer. Therefore, a wide range of access rights were incorporated into the repository. It allows

[1] http://theses.cz

[2] http://odevzdej.cz

the authors to create different levels of access modes for different users for each file with respect to their agreement with a publisher.

3 Interconnected Text Reuse Detection

Plagiarism in academia is often referred as an academic dishonesty. It is a persistent moral offense and if it is brought to light later on, it also discredits the institution where it originated from, because it passed unnoticed and should have been detected and dealt with accordingly at the time of submission. Higher educational institutions are not usually fond of making such cases public, on the contrary, they try to conceal it and resolve the issue internally as much as it is possible (Weber-Wulff, 2014). Prevention and early detection are the best ways of solving plagiarism issues.

3.1 Prevention

The issue of plagiarism is addressed at Masaryk University now for many years. All faculties teach theirs courses concerning plagiarism working with texts and citing skills individually, but they are all obliged to follow the university's study and examination regulations. Students are required to sign a declaration of originality in their theses. Also, a supervisor mandatorily confirms the state of submitted thesis prior to its defense. The supervisor needs to evaluate the text before the confirmation. However, evaluating text reuse may be quite a difficult, tedious and time consuming process and so, when marking papers, such as theses or seminar works, an automated computer system facilitating the task of checking for plagiarism, has proved to be very helpful. The supervisor needs to understand the output of the software and to decide whether the detected similarities are actual plagiarism.

Plagiarism is not only cheating, it may appear in form of unintentional plagiarism such as omitting citations or quotations. Students are encouraged to use university's text reuse detection systems for themselves prior to handing in the work into the IS. In fact anyone can have their work checked using the Odevzdej.cz system. (Lunter et al., 2013) While Odevzdej.cz utilizes the interconnected database with other anti-plagiarism tools run by MU, it allows the students to retrieve the same results as the supervisor will obtain by checking their work in the IS. The user is informed about similarities via an e-mail report, which they can study as part of the formative feedback before handing in the final version of their paper. The formative feedback, with assistance from an automated text reuse detection system, has a positive impact on students' final submissions. (Davis and Carrol, 2009) The main goal of existence of text reuse detection systems is to improve the quality of students' works.

3.2 Detection

Plagiarism detection system which works on the basis of evaluating documents similarities must be aware of both the original document and the plagiarized document in order to report the similarity. It is up to the implemented algorithm and methodology of the system what similarity to detect. Modern plagiarism detection systems recognize similarities even between variously altered texts. The similarity is calculated based on specific document characteristics, which are collected and stored in the database for each document known to the system. Documents in university repositories are good candidates to be included in similarity evaluation of the anti-plagiarism system. Such academic works could be targeted for plagiarism, because they contain high quality texts. Importance in access right lies in their ability to include into similarity search also documents, which are inaccessible for the user. The output of the software differs in different usage scenarios. For example, if the user is not granted with required access rights to a similar document, the software displays information about the issue, thus the similarity detection remains comprehensive. In the output, there can be provided a contact at the author or at the administrator. On the other hand, an administrator or a supervisor will get access not only to the similarities, but also to the source document.

Having indexed documents in the system with proper access rights has also benefits for the authors. It is a kind of protection for the text where its originality and also its age can be easily proofed, even if the document have not been made public.

It is crucial for an antiplagiarism system to have documents base of high quality. The goal is achieved by the usage of different document resources beginning with theses, seminar works, and institutional papers from the repositories up to relevant documents from online sources. The process of online source retrieval is performed for each document entering the anti-plagiarism detection and the main purpose is to retrieve a relatively small subset of similar documents, which may have been plagiarized from, from the vast document corpus, which in real world is the Web. (Suchomel and Brandejs, 2014) Documents are looked up utilizing modern search engines, which possess the computational power to index the whole Web. The searches are executed based on multiple types of queries created based on textual characteristics of the input document and on automatically extracted keywords. Firstly, this leads to retrieving textually similar documents from the Web. Secondly, which is particularly beneficial when done based on academic papers in the institutional repository, it retrieves thematically related documents. The scientific papers in the repository represent variety of contemporary themes in academia. The keywords extracted from such papers provide an opportunity to retrieve more theme-related and current material from the Web. All the retrieved documents based on each input paper are indexed and consequently compared to all future suspicious documents during the plagiarism detection. As a result of that the system obtains related documents from online sources based on each newly added document into repository, which does not have to be publicly accessible. This supports coverage of the overall research theme for any future plagiarism detection of documents concerning that theme (Suchomel, 2017).

4 Plagiarism Detection in Limited Access Mode

Different types of files require different approaches of accessibility. Plagiarism detection systems need to process data irrespective of accessibility, however they cannot reveal any inaccessible text to the user. Many universities in the Czech Republic demanded plagiarism detection system, but only few of them were prepared to publish their theses online. Analogous challenge arouse when developing plagiarism detection for seminar papers, which cannot be published without the author's permission.

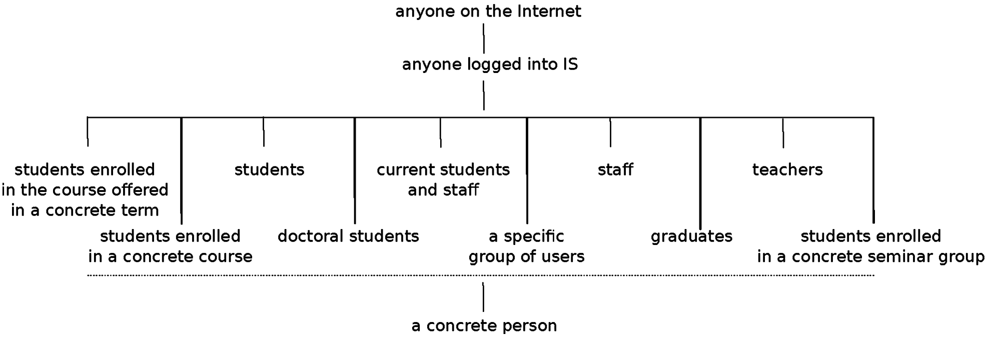
anyone on the Internet
anyone logged into IS
students enrolled in the course offered in a concrete term
students
current students and staff
staff
teachers
students enrolled in a concrete course
doctoral students
a specific group of users
graduates
students enrolled in a concrete seminar group
a concrete person

Figure 1. Hierarchy of access rights

Limited access of results published via a university repository is even more complex, as discussed in chapter 5.

4.1 Access Rights Approach in Repository Structure

In the repository, we distinguish three basic types of access rights, which are by their philosophy quite similar to access rights in UNIX systems. Right to read permits the user to enter the directory, dump its content, read the files and mark them as read. Right to upload permits the user to enter the directory, create new subdirectories and upload new files. Right to administer enables to perform all operations on a given file or a directory. Besides these traditional access rights, a right to view has been added. It was originally incorporated due to the embedded video player, but today it is also used by the embedded PDF reader. The view right allows to read the files, even though the user is not permitted to download the file. The right may be granted to one or more entities including groups of students, teachers, or all authenticated users. The whole hierarchy of possible entities can be seen in figure 1.

All files are stored in our database file system. A basic unit of this deposit is a node which can be both the directory and the file. Any node can contain any number of objects, with links to a network of resources or stored files, as shown in figure 2 This enables the system to store different formats (e.g. PDF, DOC, and TXT) of one file in different objects under one node. By binding access rights with a node and not with an object, the access to the attached paper does not depend on the format preferred by the user.

Since nodes can be nested, the access rights must reflect the whole path to the file. In the early stages of repository development, a hierarchy of system access rights was used. Consequently, in order to access the file it was necessary to check rights of every node on the path from the root directory to the given file. The result of this state was a serious slowdown of access rights checking. Therefore, the system was rebuilt to system of flat access rights in which every node keeps the rights of its parent directory. Disadvantage of this approach is in necessity to recalculate the rights of each descendant, when modifying permission of the directory.

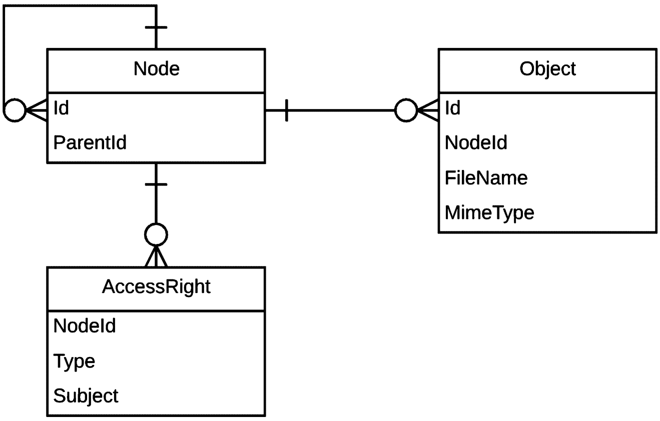

Figure 2. Node and objects relations

4.2 A Role Based Access Control

Behind each repository, there are number of roles and responsibilities that can be identified and that need to be fulfilled. The people who occupied these roles often need to access files in the repository. Therefore, it is desirable to grant access to files for these users regardless of permissions. A role based access control can ensure this. It determines access only for authenticated users based on the function which the user is allowed to perform within the system. The determination of membership and the allocation of transactions to a role must comply with organization-specific protection guidelines. In the repository, there are three main types of application access roles recognized:

- Repository management—the members of this role are academic authorities, who are responsible for decision making and sustainability issues.
- Repository administration—the members are employees authorized by a dean. They are responsible for authorization of access rights and for communication with the authors.
- System support—the members are responsible for providing a technical support.

Depending on their roles, users have different views of the data in the repository. For example, the repository management needs to display detailed statistics, reports and overviews of the repository content for each department in order to evaluate the research, but they do not need to manipulate with the records. Next, the repository administration needs to access all materials in order to assist the author with filling metadata and setting the access rights. The access roles are principally assigned to a specific department, resulting that the privileges given by a membership in a group are applied only to data associated with that department.

4.3 Access to the Content

Any digital library or institutional repository of substantial size relies on software-based search engines to help the users to locate documents related to their information need. However, since there are lot of documents which are not permitted to be read by every user, the search engine should not allow them to be discovered. A method proposed by Kasprzak at al. (Kasprzak et al., 2010) was utilized for the purpose of handling access rights in enterprise full text search systems. The method is based on incorporating access rights descriptions into the inverted index in the form of "virtual tokens". Virtual tokens are special words which have no weight nor position within the document, therefore they are excluded from either the proximity or the exact phrase searches. To incorporate access right of a given document into inverted index, we can add a virtual token with the user group identification for each user group, whose members can read the document. Consequently, before query execution the query must be modified by following way:

$$a\ query\ AND\ (p : group_1\ OR \ldots OR\ p : group_n)$$

The prefix p : marks virtual token describing the access rights and $group_1$–$group_n$ denote all groups the user belongs to.

The approach based on virtual tokens has proven to be very useful. In our submission, the virtual tokens are used not only for access rights, but also for implementation of faced search, which is popular among librarians, because it enables to showcase metadata.

A research institution may produce many different types of research outcome, from experimental data through scientific publications and patents to software. Based on the type of the outcome it can be desirable to attach several types of an output, e.g. research data, presentation or preprint, to one record in the repository. In such cases, different access rights may be applied to both metadata and to each attachment. Therefore, they must be indexed separately. If necessary, they can be connected via unique hash, a virtual token with a unique value for a given group. In our case we used PURL of landing page for given record.

5 Approval Process

The dissemination of academic results via the repository is based on the assumption that the author knows best about the copyright of the result, which may include rights of third parties, and the author decides whether to include the paper into repository and under which circumstances. Nevertheless, the final responsibility lies mostly on university due to owning the property rights. In practice, this means that only the university as an author's employer can decide whether the work can be open to public use. Concerning institutional repositories, setting the publishing mode of a scientific result may pose a significant administration overhead for larger institutions which produce thousands of papers each year. This resulted in the establishment of a two-tier process of setting access rights.

The first tier represents the initial result submission settings entered by the author who has to determine the nature of the work, such as its type and whether it is job-related. The author can also set the access rights directly, if the work was not determined as job-related or the employer does not own the property rights for the given type of the publication. Otherwise, the employee is entitled to suggest the access rights, but the rights must be authorized by a delegate of the dean, which represents the second tier in the process of authorization. Delegates rely on a special application for accepting or rejecting suggested access rights. If the access rights are accepted, they are immediately applied and the file is published in accordance with them. In the other case, e.g. if the suggested access rights do not match the agreement with the publisher, or the author cannot substantiate the authors' consent with other possible co-authors to make the file public, the reasons for rejecting are automatically sent to the author. Consequently, they can decide about repeating the approval process and suggesting new access rights, or let the decision be made by the authorized employee.

In addition to this functionality the application also provides support for setting-up the policies of approval process. This includes choosing the types of scientific outputs for which the author can set the rights directly, which can be set globally or separately for each department. The application also holds the whole approval history, which helps to resolve disputes. On top of that, the data from this evidence are used for recommending appropriate rights. Our experience show that all parties can benefit from granting an appropriate rights at the time of submission. Therefore, the technical and methodological support such as incorporating the copyright information service for open access archiving provided by SHERPA / RoMEO (Flick et al., 2016), is emphasized in the system.

In terms of the repository content, the growth of the repository has been positive. Although, the authors are not forced to upload a full text, we are noticing a growing trend in number of repository submissions. At the time of writing this article, the repository included almost 7 thousands of publications containing the full text (about 18% of university production). Approximately one third of this number had to be authorized to accept the suggested rights. A key factor in this outstanding result is the wide range of access rights. In the repository, there are approximately one quarter of attached files open to the public. The rest of the attached files are accessible in restricted mode.

6 Conclusion

This paper has given an account of access rights driven repository implemented as part of university's information system, which extends the integrated plagiarism detection solution. The main goal of the study was to describe important aspects of the repository with included support of decision-making processes of access rights setting. This study has shown a solution based on two tier access rights management, complying with third parties rights. Taken together, this approach supports the employer to be able to control the access to the job-related works. The positive effect of the proposed solution is to promote the authors' motivation to upload even non-public research. Specifically, the added value is represented by plagiarism detection or utilization of not public content for e-learning.

The future work will focus especially on further motivation of the authors to submit the full texts into the repository. The content of the repository is continuously interconnected with more applications across the information system, which increases user comfort. Simultaneously, the central service, which is based on similar principles, was developed with cooperation of 23 educational institutions, which resulted in development of system Repozitar.cz (Jakubík et al., 2011). The Repozitar.cz is a nationwide repository service offering necessary technical, organizational, social and legal environment.

Literature

Berlin Declaration, 2003, *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, [online], available at: https://openaccess.mpg.de/Berlin-Declaration (accessed 22 March 2017)

Conway, E., Giaretta, D., Lambert, S., Matthews, B.. Curating scientific research data for the long term: a preservation analysis method in context. *International Journal of Digital Curation*, 6(2):38–52, 2011.

Davis, M., Carrol, J.. Formative feedback within plagiarism education: Is there a role for text-matching software? *International Journal for Educational Integrity* 5(2) December, 2009 pp. 58–70 ISSN 1833-2595

Flick, L., Norris, S. A. *Using SHERPA/RoMEO: Finding policies for self-archiving articles*. 2016.

Hitchcock, S.. *The effect of open access and downloads ('hits') on citation impact: a bibliography of studies*. 2013.

Jakubík, D., Lunter, L., Brandejs, M., Brandejsová, J., et al. A central repository of publication results, implemented as a part of systems for revealing plagiarism. *Seminář ke zpřístupňování šedé literatury 2011: 4. ročník semináře zaměřeného na problematiku...*, 4(1), 2011.

Kasprzak, J., Brandejs, M., Obšívac, T. Access rights in enterprise full-text search. In *Proc. ICEIS*, 2010.

Koler-Povh, T., Turk, G., Južnič, P. Does the open access business model have a significant impact on the citation of publications? case study in the field of civil engineering. In *5th Belgrade International Open Access Conference* 2012, page 89.

Lunter, L., Jakubík, D., Suchomel, Š., Brandejs, M. Inter-university cooperation on plagiarism detection systems in the Czech Republic. In J. Rybička (ed.) *Plagiarism Across Europe and Beyond*, pp. 216–224, Brno, 2013. Mendel University in Brno.

Olsbo, P. Does openness and open access policy relate to the success of universities? *Information Services & Use*. 33, 2, 87–91, Apr. 2013. ISSN: 01675265.

Ramirez, M., Hanlon, A. Asking for permission: A survey of copyright workflows for institutional repositories, portal: *Libraries and the Academy,* ll(2):683–702, 2011.

Suchomel, Š., Brandejs, M. Approaches for Candidate Document Retrieval. Information and Communication Systems (ICICS), 2014 *5th International Conference on*, Irbid, 2014, vol. 2014, April, p. 1–6. doi:10.1109/IACS.2014.6841959.

Suchomel, Š. *Source Retrieval for Text Reuse Detection*. Doctoral Dissertation, Faculty of informatics, Masaryk University, Brno 2017.

Weber-Wulff, D. False Feathers: *A Perspective on Academic Plagiarism.* Springer Science & Business, 2014

Copyright statement

Authors

Daniel Jakubík (daniel.jakubik@gmail.com), Šimon Suchomel, Luboš Lunter, Michal Brandejs, Masaryk University, Brno, Czech Republic