Identification of Scholarly Papers and Authors by Connecting Databases

Kensuke Baba[1], Masao Mori[2], Eisuke Ito[3]
[1]Library
Kyushu University
Japan

[2]Institutional Research Office
Kyushu University
Japan

[3] Research Institute for Information Technology
Kyushu University
Japan
{baba@lib, mori@ir, itou@ccg}.kyushu-u.ac.jp

ABSTRACT*:Repositories are being popular as places for publication of research outputs. To make more efficient use of scholarly information on the internet, repositories are required to cooperate with other databases. One of the essential processes of the cooperation is identification of scholarly papers and their authors. The straightforward approach is string matching of the title and authors' name, however this approach cannot always solve the difficulties by basic clerical errors and same names. This paper proposes a method to compensate for the inaccuracy of the identification by connecting different databases. The main idea of the method is that different metadata of a scholarly paper is linked by the authors themselves, therefore the correspondence is guaranteed by the authors.*

Keywords: Web database, Repository, Scholarly paper, Identification

Received: 12 March 2011, Revised 18 April 2011, Accepted 28 April 2011

1. Introduction

The number of digital contents on the internet is rapidly increasing. Especially, for scholarly information, electronic journals and repositories [12] are being popular as places for publication of research outputs. The metadata of a scholarly paper (that is, the information about the title, the author(s), and so on) is usually archived in plural databases severally, and therefore the metadata has some variations. For some papers, in addition to the metadata, the full-text is archived in plural databases and it has some versions such as the published version and pre-/post-print. The scholarly papers should be organized to make more efficient use for the users of the information.

In order to organize scholarly papers, it is not practical that an authority should manage all the papers. A feasible solution is cooperation of databases and advanced search functions thereby. For the solution, we have to make clear the relation on scholarly papers. The first step is \identification of scholarly papers", that is, to link the variations of the metadata of each paper. The versions of each paper can be managed by processing this step in detail. As the second step, one of the simplest organizations is classification with respect respect to the authors. The classification requires \identification of authors". As the result of these identifications, the metadata should have IDs which correspond to real papers and authors, respectively.

The straightforward approach of the identification of scholarly papers and authors is string matching [10] by the title and the authors' name. Some variations of the title can be identified by approximate string matching [11]. As for authors, the accuracy can be improved by matching of extra information such as he affiliation. However, this approach cannot always solve the difficulties by basic clerical errors and same names. If we have enough data for the identification, machine learning and rule based approach such as [4] are possible solutions. Another approach of a different quality is confirmation by the authors themselves. For example, the problems are solved by adding unified IDs for scholarly papers (such as DOI) and authors (such as the ID for membership of an association) to the metadata when the paper is registered. However, it is difficult to popularize unified IDs in advance, and moreover this solution cannot be applied to the papers which are already archived. The main idea of our solution is that the confirmation by the authors is realized by a cooperation of databases.

In this paper, we are trying to solve two problems in practical systems as a case study. Kyushu University has the researcher database DHJS (Kyushu University Academic Staff Educational and Research Activities Database, \Daigaku Hyoka Joho System" in Japanese) [1] and the repository QIR (Kyu(Q)shu University Institutional Repository) [2]. One of the problems we tackle is about identification of scholarly papers. DHJS has the metadata of scholarly papers which are produced by the researchers in the university. The number of the registered metadata is about 86,000 as of June 2011, however it is estimated that at most about 20% are duplicate data. The other problem is about identification of authors. In QIR, a search of an author is operated by the naive string matching on the metadata, therefore the search cannot recognize any same name. The previous problems are solved by the following cooperation of the systems. By connecting the metadata in DHJS to the full-text in QIR,

- The first problem is solved since the identification of any paper is operated in QIR by handwork,
- The second problem is solved since a user authentication is required in DHJS for registration of metadata.

The number of the institutions who have own repository in the world is about 2,300 as of August 2011 [3], and most of the institutions are considered to have the same problem. In this paper, the situation of the practical systems in Kyushu University are shown in detail, and the problem and its solution are described formally. Therefore, the proposed idea is applicable to other institutions.

2. Problem

This section describes the current situation of two databases, DHJS and QIR, and then formalize the problems we tackle.

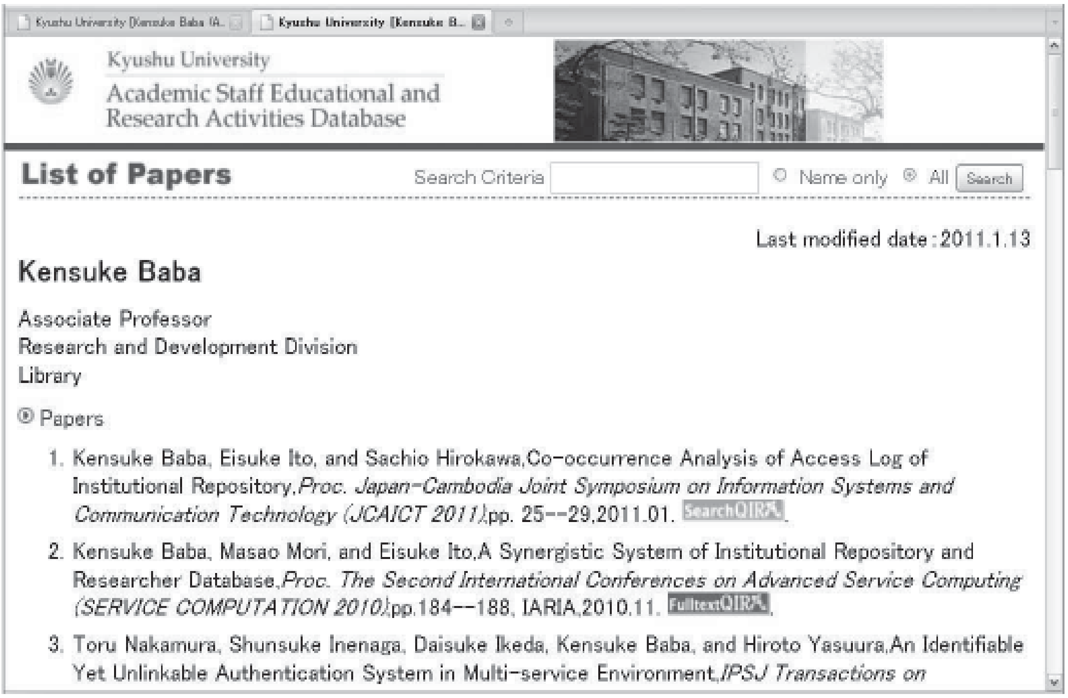
2.1 DHJS

DHJS is the researcher database of Kyushu University. DHJS has various kinds of data about the researchers in the university, for example, the posts, their research interests, and the scholarly papers they produced. The number of the researchers in the university is about 2,100 as of April 2011, and any researcher has a duty to register their research activities includes the metadata of scholarly papers into DHJS. DHJS consists of the two subsystems, the data-entry system and the viewer system. The data-entry system supports researchers to register their research activities to DHJS and equips a user (that is, a researcher) identification by a password. The viewer system shows the research activities registered in DHJS by the data-entry system. Figure. 1 is an example of the list of the metadata of scholarly papers shown on DHJS. The icons in the figure are mentioned in the following section.

The number of the metadata of papers registered in DHJS is about 86,000 as of January 2011. If a paper was written by plural authors in Kyushu University, the metadata of the paper might be registered by each authors severally. We practically estimated the ratio of the duplicate data in DHJS by calculating the edit distance [13] between the titles of the papers. Figure. 2 is the result of the calculation for a department with about 15,000 pieces of metadata. The horizontal axis shows the number of pairs and the vertical axis the edit distance which is formalized by the length of the longer title. The number of the pairs whose edit distance is less than 0.1 is about 3,000, that is, the number of the duplicate data is at most about 3,000 (and 1,500 if we assume that 4 pieces of metadata are registered for a single paper on average)[1]. Therefore, at most about 20% of the metadata are estimated to be duplicate. There was no significant difference of the ratio for every departments. By identification of the duplicate data, we can make more efficient use of the database, for example, we should be able to refer coauthors' data in DHJS from the metadata.

2.2 QIR

QIR is the institutional repository operated by Kyushu University Library. In general, institutional repository archives the



Kyushu University [Kensuke Baba (A...
Kyushu University [Kensuke B...

Kyushu University
Academic Staff Educational and
Research Activities Database

List of Papers

Search Criteria ○ Name only ● All Search

Last modified date : 2011.1.13

Kensuke Baba

Associate Professor
Research and Development Division
Library

● Papers

1. Kensuke Baba, Eisuke Ito, and Sachio Hirokawa,Co-occurrence Analysis of Access Log of Institutional Repository,*Proc. Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT 2011)*,pp. 25--29,2011.01. SearchQIR.
2. Kensuke Baba, Masao Mori, and Eisuke Ito,A Synergistic System of Institutional Repository and Researcher Database,*Proc. The Second International Conferences on Advanced Service Computing (SERVICE COMPUTATION 2010)*,pp.184--188, IARIA,2010.11. FulltextQIR.
3. Toru Nakamura, Shunsuke Inenaga, Daisuke Ikeda, Kensuke Baba, and Hiroto Yasuura,An Identifiable Yet Unlinkable Authentication System in Multi-service Environment,*IPSJ Transactions on*

Figure 1. The Web image of the list of scholarly papers in DHJS

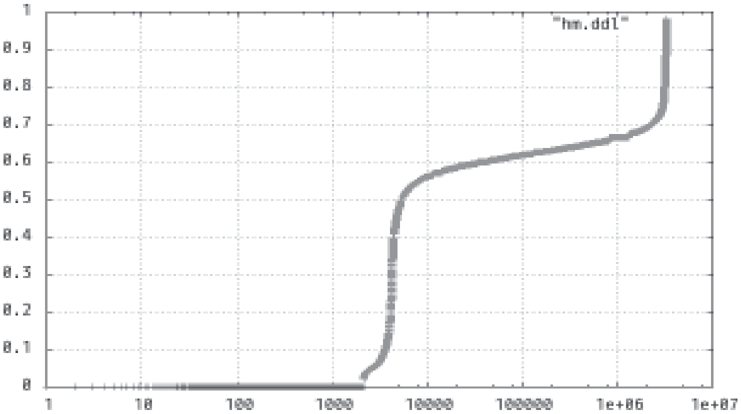


Figure 2. The edit distances between the all possible pairs of the 14,599 titles in DHJS for a department, where the horizontal axis shows the number of pairs and the vertical axis the edit distance which is formalized by the length of the longer title

full-text of each paper in addition to its metadata. The total number of the items (papers, slides, and so on) in QIR is about 17,000 as of August 2011. The registration of items to QIR are operated by staff in Kyushu University Library, and therefore the confliction of items are checked by handwork at the time. Figure. 3 is an example of the metadata of an item in QIR. The name of each author is linked to the profile page of the author, however the page is just the result of the naive string matching of the name for the items in QIR.The problem of the same name cannot be ignored.Actually, in the 2,136 professors (including associate and assistant ones) of Kyushu University, there exist

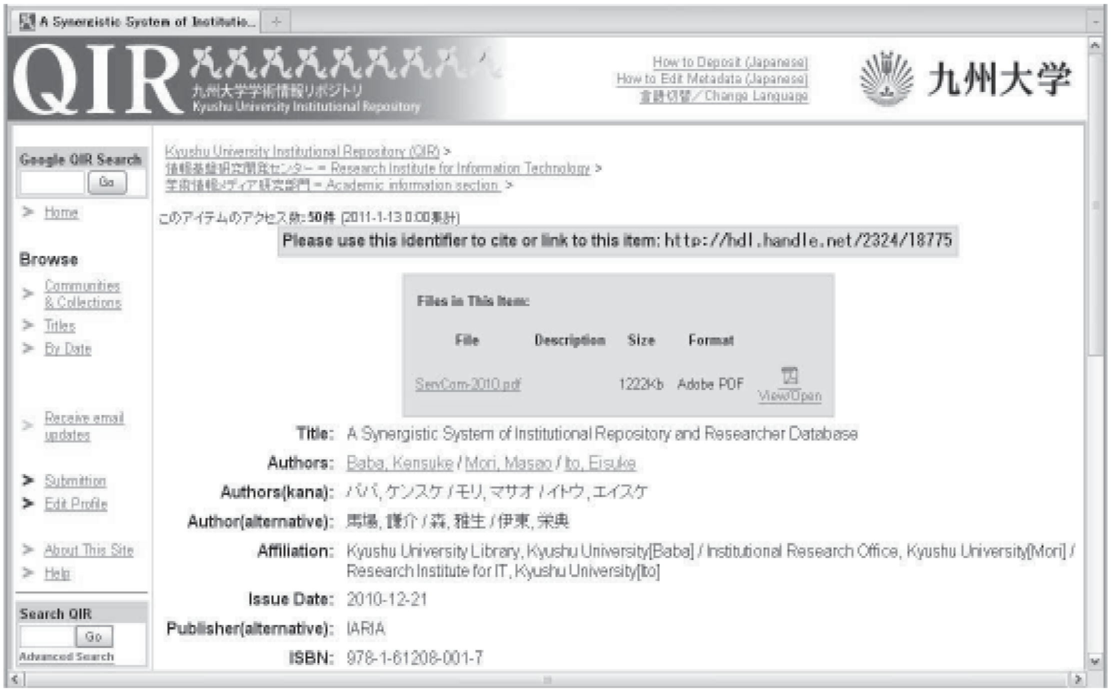
- 10 pairs (20 persons) of the same given name and family name.

[1] The number of the duplicate data is defined to be the gap between the number of the metadata and the number of the distinct papers. If we assume that the duplicate data is made by n authors for each paper, then nC_2 pairs are counted for each paper and the number of the duplicate data is n 1 for each paper. Therefore, the number of the duplicate data is 2m=n for the number m of the counted pairs.

- 186 groups (488 persons) of the same initial of the given name and family name (for example, there exist 5 researchers of name “M. Tanaka”)

- 352 groups (1,255 persons) of the same family name (for example, there exist 22 researchers of name “Tanaka”).

Moreover, in addition to the professors in Kyushu University, a lot of students and researchers in other institutions are included as co-authors of the papers in QIR.


A Synergistic System of Institutio...
QIR
九州大学学術情報リポジトリ
Kyushu University Institutional Repository
How to Deposit (Japanese)
How to Edit Metadata (Japanese)
言語切替／Change Language
九州大学
Google QIR Search
Go
Home
Browse
Communities & Collections
Titles
By Date
Receive email updates
Submition
Edit Profile
About This Site
Help
Search QIR
Go
Advanced Search
View/Open

Kyushu University Institutional Repository (QIR) >
情報基盤研究開発センター = Research Institute for Information Technology >
学術情報メディア研究部門 = Academic information section >

このアイテムのアクセス数: 50件 (2011-1-13 0:00集計)

Please use this identifier to cite or link to this item: http://hdl.handle.net/2324/18775

Files in This Item:

File	Description	Size	Format
SerrCom-2010.pdf		1222Kb	Adobe PDF

Title: A Synergistic System of Institutional Repository and Researcher Database
Authors: Baba, Kensuke / Mori, Masao / Ito, Eisuke
Authors(kana): ババ, ケンスケ / モリ, マサオ / イトウ, エイスケ
Author(alternative): 馬場, 謙介 / 森, 雅生 / 伊東, 栄典
Affiliation: Kyushu University Library, Kyushu University[Baba] / Institutional Research Office, Kyushu University[Mori] / Research Institute for IT, Kyushu University[Ito]
Issue Date: 2010-12-21
Publisher(alternative): IARIA
ISBN: 978-1-61208-001-7

Figure 3. The Web image of the metadata of an item in QIR

2.3 Formalization

The problems in the previous subsections are formalized.Since we are focusing on scholarly papers and their authors, we consider the set P of the real scholarly papers and the set A of the real authors. We define the *metadata* of a scholarly paper to be a pair of a string (the *title*) and a non-empty set of strings (the *author*). The problems are defined to be, for a given set M of metadata, to find the functions $f : M \rightarrow P$ and $g : M \rightarrow 2^A$ which represent the correspondence of the metadata to real papers and authors, respectively. In other words, it is to put indexes in P and A on the title and the author of each metadata.

In this paper, the input sets of metadata are M_D and M_Q for the metadata in DHJS and QIR, respectively. Then, the following assumptions are given for the problem by the current situations in the previous subsections. As mentioned in Subsection 2.1, the metadata in DHJS are registered by one of the authors and the registration requires a user authentication. Therefore, it can be regarded that we have the function $g_1 : M_D \rightarrow A$ such that $g_1(d) \in g(d)$ for any $d \in M_D$. As to Subsection 2.2, the correspondence between the metadata in QIR and the full-text is guaranteed by the check of the staff in the Library. Therefore, we have the function $f_1 : M_Q \rightarrow P$ such that $f_1(q) = f(q)$ for any $q \in M_Q$.

3. Solution

3.1 Main Idea

The main idea of our solution is, in terms of the formalization in Subsection 2.3, to find the function $h : M_D \rightarrow M_Q$ which satisfies the following conditions:

1. $f_1(h(d)) = f(d)$ for any $d \in M_D$,

2. $g_1(h^{-1}(q)) = g(q)$ for any $q \in M_Q$, and

3. $g(q) = g(d)$ for any $q \in M_Q$ and any $d \in h^{-1}(q)$,

where $h^{-1}(q) = \{d \in M_D \mid h(d) = q\}$ and $g_1(S) = \{g_1(s) \mid s \in S\}$ for a set $S \subseteq M_D$. This situation is illustrated in Figure. 4. In this example, $P = \{p_1, p_2\}$, $A = \{a_1, a_2, a_3\}$, $M_Q = \{q_1, q2\}$, and $M_D = \{d_1, d_2, d_3\}$, and then the functions f and g are as the following table.

x	q_1	q_2	d_1	d_2	d_3
$f(x)$	p_1	p_2	p_1	p_2	p_2
$g(x)$	$\{a_1\}$	$\{a_2, a_3\}$	$\{a_1\}$	$\{a_2, a_3\}$	$\{a_2, a_3\}$

As the assumptions, we have the functions f_1 and g_1 such that $f_1(q_i) = p_i$ for $i = 1,2$ and $g_1(d_i) = a_i$ for $i = 1, 2,3$ (the left-hand in Figure. 4). Then, h should be $h(d_1) = q_1$, $h(d_2) = q_2$, and $h(d_3) = q_2$. For the conditions 1 and 2, by the h we have $h^{-1} \circ f_1$ and $h_1 \circ g_1$ such that $h \circ f_1(d1) = p_1$, $h \circ f_1(d_2) = p_2$, $h \circ f_1(d_3) = p_2$, $h^{-1} \circ g_1(q_1) = \{a_1\}$, and $h_1 \circ g_1(q2) = \{a_2, a_3\}$. Additionally, for the condition 3, we have $h \circ h^{-1} \circ g_1$ such that $h \quad h^{-1} \circ g_1(d_1) = \{a_1\}$, $h \circ h^{-1} \circ g_1(d_2) = \{a_2, a_3\}$, and $h \circ h^{-1} \circ g_1(d_3) = \{a_2,a_3\}$.

In the sense of the practical system, the condition 1 is clearly satisfied by linking the metadata of a paper in DHJS to the metadata of the paper in QIR. The condition 2 is satisfied by the previous link if any author of the papers in QIR register the metadata of the papers. It is also clear that the condition 3 is satisfied by the link since the author(s) of a paper in DHJS is same as the paper in QIR.

3.2 Implementation

We have already developed a system which links the metadata in DHJS to the full-text in QIR [7, 9]. In Figure. 1, the dark colored icon “FulltextQIR” is connected to the corresponding full-text in QIR. Researchers put icons on the list in the data-entry system of DHJS, and link them to the full-text by themselves. (The other light-colored icon “SearchQIR” means that the metadata is not linked yet.) Therefore, the correspondence between the metadata in DHJS and the metadata in QIR is guaranteed by a check of the author instead of string matching. Namely, the function h can be realized by this link system.

The following is the outline of the implementation.The ID of any paper in QIR is attached to the metadata in DHJS by this link, which realizes the condition 1. Additionally, since the metadata in DHJS has the ID of the author who registered the metadata, we can put the ID to the corresponding metadata in QIR by the link, which realizes the condition 2. By returning the author IDs from QIR to DHJS after the IDs for all the authors are attached, also the metadata in DHJS can have the IDs of the authors, which is for the condition 3.

A problem in the implementation is about the dataflow between the databases. At the second phase in the outline, the ID of the author is sent from DHJS to QIR. However, the ID has to be returned from QIR to DHJS at the third phase. In general, such circulation of data is not suitable for managing databases. A precise policy should be defined for this dataflow.

3.3 Consideration

We are observing the number of the links in the developed system. As of June 2011, the number of metadata in DHJS is 86,255, and the number of the metadata which are linked to the full-text in QIR is 446. Hence the ratio of the linked metadata in DHJS is only 0.5 %. The number of the “discrete” full-texts in QIR which are linked from DHJS is 419 (that is, there exist 27 of duplicated metadata by co-authors). The ratio of the linked full-text in QIR is about 2 %. The number of the researchers who linked their metadata to the full-text at least one is 58.

To achieve the effectiveness of the system, in addition to the ratios of the linked data in the databases, the basic number of the full-texts archived in QIR should be increased. For this problem, we are developing a system to encourage researchers to register their papers to QIR by showing the result of access log analyses in QIR [6, 8]. Additionally, we are also developing a system which supports registration of scholarly papers to IR for researchers and repository managers [5].

4. Conclusion

A method to compensate for the inaccuracy of identification of scholarly papers and authors on the metadata in separated databases was proposed. We formalized the problem in the practical systems and proposed the solution in terms of the formalization. We also showed the outline of the implementation based on the idea of the proposed solution.

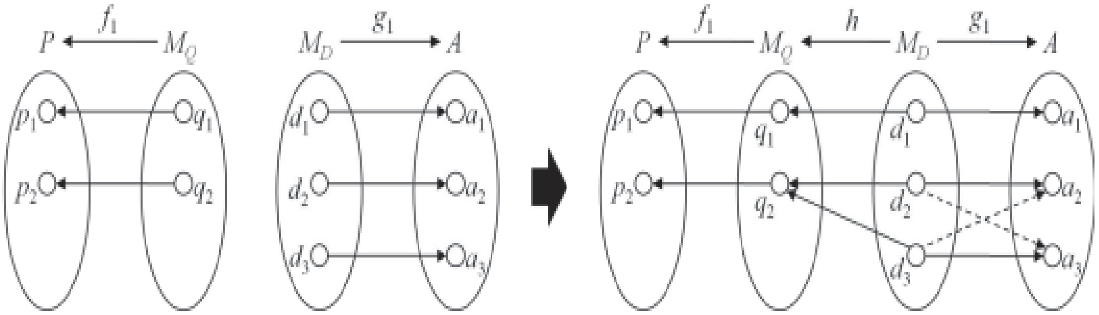


Figure 4. The relation between f_1, g_1, and *h* for an example (*P, A,M_Q ,M_D*). By *h*, we can compensate the relation of the dotted arrows

References

[1] Kyushu University Academic Staff Educational and Research Activities Database. http:// hyoka.ofc.kyushu-u.ac.jp/search/ index e.html,[accessed 8 Aug, 2011].

[2] QIR: Kyushu University Institutional Repository. https://qir.kyushu-u.ac.jp/dspace/, [accessed 8 Aug, 2011].

[3] ROAR: Registry of Open Access Repositories. http://roar.eprints.org/, [accessed 8 Aug, 2011].

[4] Afzal, M, T., Maurer, H., Balke, W, T., Kulathuramaiyer, N. (2010). Rule based autonomous citation mining with TIERL, *Journal of Digital Information Management*, 8 (3) 196-204.

[5] Baba, K., Hoshiko, N., Kudo, N., Yoshimatsu, N., Ito, E. (2011). Semi-automated paper-registration system for institutional repository. *In: The Third International Conference on Awareness Science and Technology*.

[6] Baba, K., Ito, E., Hirokawa, S. (2011). Co-occurrence analysis of access log of institutional repository, *In*: *Japan-Cambodia Joint Symposium on Information Systems and Communication Technology*, p. 25-29.

[7] Baba, K., Mori, M., Ito, E. (2010). A synergistic system of institutional repository and researcher database. *In: The Second International Conferences on Advanced Service Computing*, p. 184-188. IARIA.

[8] Baba,K., Mori,M., Ito, E., Hirokawa, S. (2011). A feedback system on institutional repository, *In: The Third International Conference on Resource Intensive Applications and Services*, p. 37-42. IARIA.

[9] Baba, K., Tanaka, T., Ishita, E., Mori, M., Ito, E., Hirokawa, S. (2011). Evaluation of link system between repository and researcher database. *In* : *International Conference on Asia-Pacific Digital Libraries*. Springer.

[10] Cormen, T, H., Leiserson, C, E., Rivest, R, L. (2001). *Introduction to Algorithms, Second Edition*. MIT Press.

[11] Crochemore, M., Rytter, W. (1994). Text Algorithms. Oxford University Press.

[12] Suber, P. (2011). Open access overview. Open Access News, 2007. http://www.earlham.edu/~peters/ fos/overview.htm, accessed 8 Aug.

[13] Wagner, R, A., Fischer, M, A. (1974). The string-tostring correction problem. *Journal of the ACM*, 21(1) 168-173.