Putting old data into new system: Web-based catalog card image searching

Minami, Toshiro
University Library, Kyushu University

Kurita, Hidekazu
Department of Informatics, Faculty of Information Science and Electrical Engineering, Kyushu University

Arikawa, Setsuo
Department of Informatics, Faculty of Information Science and Electrical Engineering, Kyushu University

出版情報：Digital Libraries: Research and Practice, pp.141-148, 2000. IEEE
バージョン：

Putting Old Data into New System: Web-based Catalog Card Image Searching

Toshiro Minami
University Library
Kyushu University
Fukuoka 812-8581, JAPAN
minami@lib.kyushu-u.ac.jp

Hidekazu Kurita
Department of Informatics
Kyushu University
(Currently,
Toshiba Corporation)

Setsuo Arikawa
Department of Informatics
Kyushu University
Fukuoka 812-8581, JAPAN
arikawa@i.kyushu-u.ac.jp

Abstract

This paper proposes a new approach to solve the data inputting bottleneck problem for library catalog data, or metadata. The data have been provided with paper cards arranged in wooden boxes. A lot of efforts have been taken to digitize them in order to put these data to be machine-readable. However, despite such efforts, only a small amount of data has been digitized so far because the inputting is done manually. We solve this problem by using the catalog card images digitized by high-speed scanners. This approach has advantages such as: (1) We can deal with the electronic catalog data with remarkably reduced time and cost. (2) It enables the seamless integration of the image-based and keyword-based searches. (3) It boosts up the process of inputting of the catalog data itself.

1. Introduction

It is practically impossible to access to a specific book in a public or university libraries without searching facilities, because there are a huge amount of books there. Thus providing appropriate means for searching is indispensable for libraries. Considering the ease of access, books are arranged according to a classification criterion such as Dewey Decimal Classification method[1]. One or more catalog cards are made for each book that is collected in the library and they are arranged according to their author names, titles and other ways.

Thanks to the recent advancement of information technology(IT), the search can be easily carried out with the computers nowadays. Since computers are very good at doing simple jobs with high speed with accuracy, keyword search is coming to be the most popularly used search method. OPAC (Online Public Access Catalog) system is the one used in most libraries for keyword-based search of documents. So the metadata information about the new books of the libraries is registered in a machine-readable text form from the beginning so that they are accessible in the OPAC system.

Furthermore the network environment, especially the Internet accessing environment, becomes a boom these days, thus it is now an indispensable service for the libraries to give accesses of Web-based OPAC-search and reading electronic documents including articles of E-journals.

Considering these environmental changes caused by the "IT revolution" it is required to boost up the library systems towards the digitized libraries. However a big problem is lying in front of us. A huge amount of paper catalog cards are waiting to be inputted. In Kyushu University Library, for example, nearly 3,400 thousand of books are collected and cataloged, and about 700 thousand items out of them, i.e. only about 1/4, are accessible by the OPAC system. In Kyoto University, about 800 thousand items are accessible by the OPAC system out of about 5,700 thousand (about 1/7)[6].

Efforts have been made in order to overcome such a situation. NII(National Institute of Informatics) and many of the university libraries have been cooperating with inputting the data written or printed on their catalog cards. A database in NII holds all the metadata inputted by these libraries. Via network, the librarians in these libraries can check the information about the book at hand before they actually input the whole data about the book. If the data is already registered by some other libraries, they can download the data into its local catalog database. If the data has not been inputted by any other library, they type all the

0-7695-1022-1/01 $10.00 © 2001 IEEE

data from scratch. All the newly inputted data are also registered in the central database in NII as well as in the local database.

Despite of such efforts, inputting is still a laborious job and time-consuming. In Kyushu University, for example, about 60 thousand to 70 thousand books are registered annually in the normal pace. If we keep this pace, it would need more than two decades to finish the metadata inputting and have the complete library catalog database.

In this paper we propose a new method of inputting and searching of catalog data by digitizing the card data by using high-speed scanners and using the image data of the catalog cards. We can notably reduce the cost and time by this method. It has big potential to open up a new application field as well.

The rest of this paper is organized as follows. In Section 2, we present the basic idea of image-based catalog card search method. Then in Section 3, we describe how this idea is realized as a computer system. In Section 4, we demonstrate its potential usefulness by showing some possible extensions of this system. In Section 5, we compare our approach to other image-based catalog card search methods and clarify the differences of our system. And finally, in Section 6, we summarize our discussions and suggest some important future directions.

2. Using Catalog Card Images for Searching

As was mentioned in the previous section, a lot of catalog data printed or marked on the paper cards are still waiting to be inputted. More than two decades should be required to be finished if we do our job in the current pace.

In order to overcome this bottleneck problem, we came up with the idea of using images of the catalog cards instead of waiting for the end of manual inputting. In this way, we are able to have a complete library catalog database in a very short time. The imaging process must be very fast if we use a high-speed scanner. Once the database is constructed we can provide the searching service on the network environment so that the users can use the database from anywhere in the network.

Because of the limitations of the high-speed scanners we used, the resolution of the images are up to 300 pixels per inch. It is sufficient for ordinary use as is shown in Figure 1. This is not a sufficient one if we want to use an OCR(Optical Character Reader) software and create text data for the items printed or written on the catalog cards. However the low resolution is not a

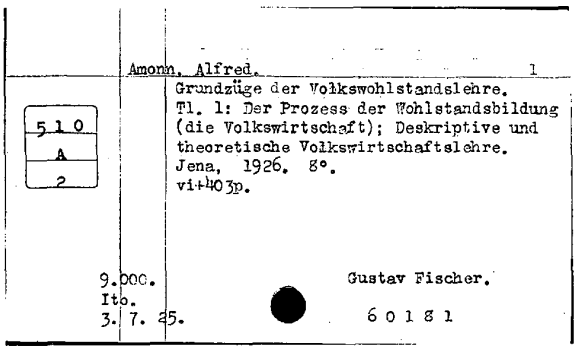

Figure 1. An Example Card Image

big problem. Considering that there are a lot of catalog cards that have hand-written characters including the hand-made corrections to the printed items, OCR softwares will not work very well either even if we can make card images with high resolution. So we have to take another way except to use the OCR softwares.

The sample card image in Figure 1 is one in moderate quality. Some cards are better than this one and some others are much worse than this with a lot of noises in their images like the one in Figure 2.

In the rest of this section we first show the usefulness of this method by pointing out some of the advantages, and then we show a couple of problems to overcome in order to have a good use of image catalog data.

2.1. Advantages

This method has some notable advantages. Taking the case of Kyushu University Library we compare this method with the ordinary method.

- Cost:
  The cost for manually inputting the data is about 700 yen per card in our estimation, whereas the scanning cost for the image is about 10 yen per card. Thus the total costs for about 1,600 thousand items are about 1 billion yen for manual inputting and 16 million yen for scanning.

- Speed:
  In terms of speed, about 70 thousand cards have been manually inputted annually, whereas more than 10 thousand cards can be scanned per day per one high-speed scanner. The differences of time and costs are enormous. In our experience in Kyushu University Library, it took only a couple of weeks to finish scanning the cards of the Faculty of Literature, which are nearly 380 thousand.

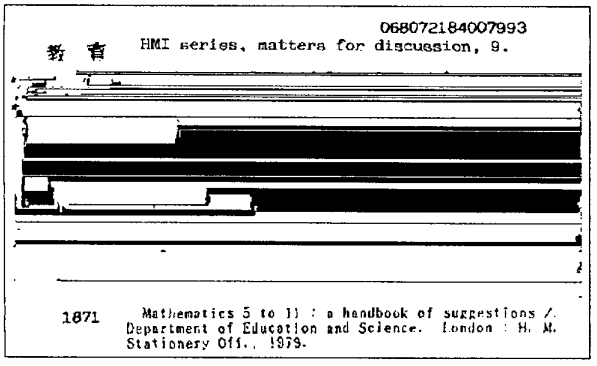

Figure 2. Card with "B-Noise"

- Service:
  Relating to the speed of inputting, all the data become machine-readable in a short time, therefore the database becomes complete. This is very important. Suppose you tried to find a book in a library database and failed. You cannot say that the book is not in the library from this failure if the database is not complete. The book could be there and its data may not be inputted to the database yet. You can say that the book is not there only after you search the traditional catalog card boxes for the book and fail. Databases should be complete as early as possible.

- Incremental Feature:
  It is quite easy to incrementally add new searching functions to the system. In the first step, only the card images are provided to the user, then the system gives a new service such as keyword-based search. We will discuss in more detail about this in Section 4.

We have developed a prototype system to search for the catalog card images[4]. The user can access to this system from a remote place with a Web browser. This makes the users easy to get the information about the books and documents of the library. They do not have to go to the library. They can check the information from their working place or even from home. We will describe in more detail about this system in next section, and will discuss its potential usefulness and importance in Section 4.

2.2. Problems

In order to develop such a system, we take up two major problems to be solved.

B-Noise

The first one is to check the quality of the image data. Figure 2 is an example card, which we call the, "bar code" (or "burst") noise, or B-noise for short. In order to detect such cards, we used the B-noise degree[3], which is defined as follows.

1. For each line of this image calculate the sum of the squared of the run length of (i.e. the number of consecutive) black pixels. Let us call the value $v(i)$ for the line i.

2. For each line calculate the weighted value $w(i) = (v(i) + (v(i-1) + v(i+1)/2)/2)$.

3. Calculate the average of $w(i)$ and make it the B-noise degree.

By using B-noise degree, we can detect inappropriate cards with satisfactory accuracy. From our experience, 10 thousand is the good threshold for detecting the card with B-noise. As examples, the card No.31 in Figure 8 has B-noise and the degree is 73,803, which is greater than 10 thousand so the # characters are attached as a mark to the item number of this card. The next card with No.32 has no B-noise and the degree is 391; less than 10 thousand.

Data Organization

The second problem is checking of correctness of the data organization[5]. The number of data is very large. For example, the number of catalog cards of Faculty of Literature is about 380 thousand, and 70 and 86 thousand for Faculty of Education and Science, respectively. We cannot check the correctness manually when the target items are so many. What we could do is to develop a checking program and check the consistency of the data organization. For example, it checked the tree structure of the image data, which is described in Section 3.2. It also checked if the names is appropriate, if the data type is appropriate, and so on.

However after such a laborious effort we still have chance to find some new kinds of data errors. It did happen in our case. For example, we happened to find a data which was scanned upside-down when we had a look of a search results. So we added the checking item of this type to the checking list of the program. We believe repeating such a process is the only way to check the validity of a huge amount of data. You cannot predict and check all the errors in advance.

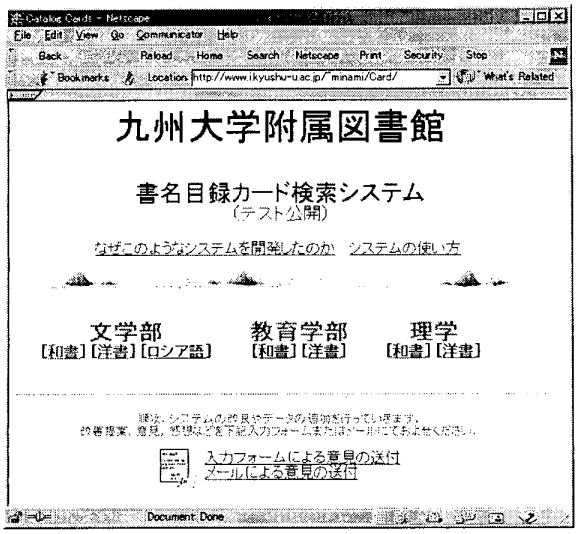

Figure 3. The Home Page of the System

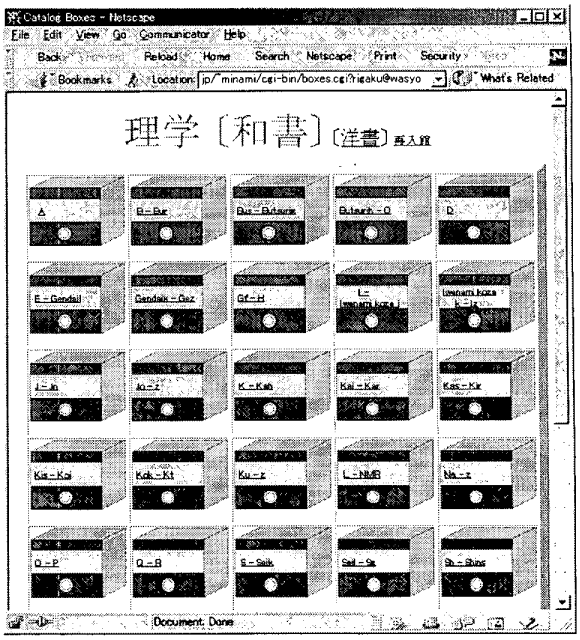

Figure 4. The Card Boxes

3. The Image Card Search System

In this section we describe the implemented system. First we take up the windows of the system and describe how the system is designed, especially putting emphasis on interfacing issues. Then we show the system organization and describe how it is implemented, including the data structure.

3.1. System Interface

The windows of the system consists of four types: the home page, the card boxes window, the box and card images window, and one card image window.

(1) Home Page

The home page of the system looks like Figure 3. From the top to the bottom are, the title of the Webpage(Kyushu University Library), the name of the system(The Book Catalog Card Search System)(Test Version)), two links to the explanations("Why We Have Developed this System" and "How to Use the System"), the names of the faculties of the data(Literature, Education and Science) together with the categorical classifications based on the written language(Japanese including Chinese and Korean, Foreign, and Russian for Faculty of Literatures), and the links to the pages for the reader's comments.

Generally the catalog cards of a faculty is divided into two kinds; the one written in Japanese and the one written in non-Japanese(or foreign languages). In the faculty of literature those written in Russian are stored separately so they are divided into three categories. The system has to deal with such varieties.

(2) Card Boxes Window

Choose the data category by clicking in the home page, the list of card boxes for the chosen category is presented as is shown in Figure 4. This is the page for those of the faculties relating to science that are written in Japanese. The first line shows these together with the links to foreign books and to the home page.

The arrangement of the card boxes is the same as the original card boxes so that the user can easily make matching with the real arrangement of the boxes in their memory and find the target box with ease. The labeling system is not always the same. For foreign books they are arranged in alphabetical order. For Japanese books, on the other hand, the labeling system varies from faculty to faculty. In this example the labels are arranged according to the alphabetical order with romanized descriptions. In the Faculty of Education, they use Kana(Japanese characters) for writing and take the Kana ordering for arrangement. The system is supposed to be adapted to such differences as well.

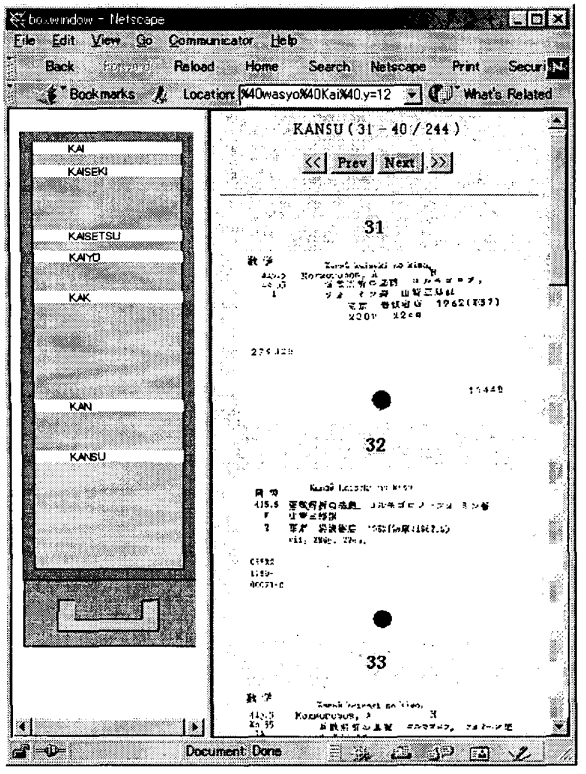

Figure 5. A Box and Card Images

(3) Box and Card Images Window

Inside of the card box something like in Figure 5 appears in a new window as the user clicks one of the boxes in Figure 4. In this example the box labeled with "Kai-Kar" is selected. The window is separated into two parts; the left part is for showing how the box is divided by guide cards, and in the right part the actual cards are listed.

In the left part the guide cards are arranged proportionally to the number of cards for the guide so that the user can easily estimate the place where the card he or she intends to find.

Estimating the position for the card having the intended title is much easier than expected. After a couple of clicking, the cards close enough to the intended card would appear in the right frame.

In the right part the cards are displayed in a block consists of 10 cards. The user scrolls and clicks the buttons for the next, previous one or two blocks away from the current cards. By these facilities, it is easy to locate the target card.

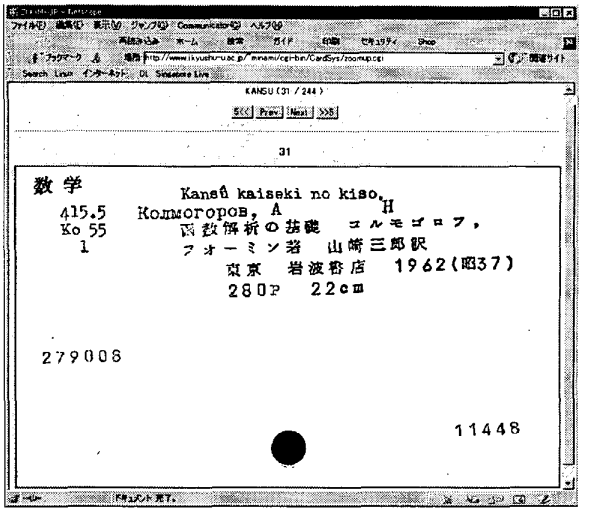

Figure 6. One-Card Image Window

(4) One-Card Image Window

One card image window(Figure 6) appears as you click on a card image so that you have the big card image for having a look in detail. It will close itself as the user clicks on the image.

3.2. System Organization

The system is constructed with a couple of Web pages written in HTML and several CGI(Common Gateway Interface) programs written in Perl. For example, the home page(Figure 3) is prepared as a HTML document and the pages that display the card images(such as Figure 4 and 5) are generated by Perl programs.

There are two kinds of data; the image data and information data. The former makes a tree structure of the form:

< faculty > / < category > / < box name > / < guide card name > / < image file > .

The information data also take the similar structure. The attribute values such as B-noise degree, the Japanese label for the boxes and other data including the meta-data such as the title, the authors, the keywords and etc, are stored as information data. The CGI programs search for the information data that are needed for forming the screen in the HTML format.

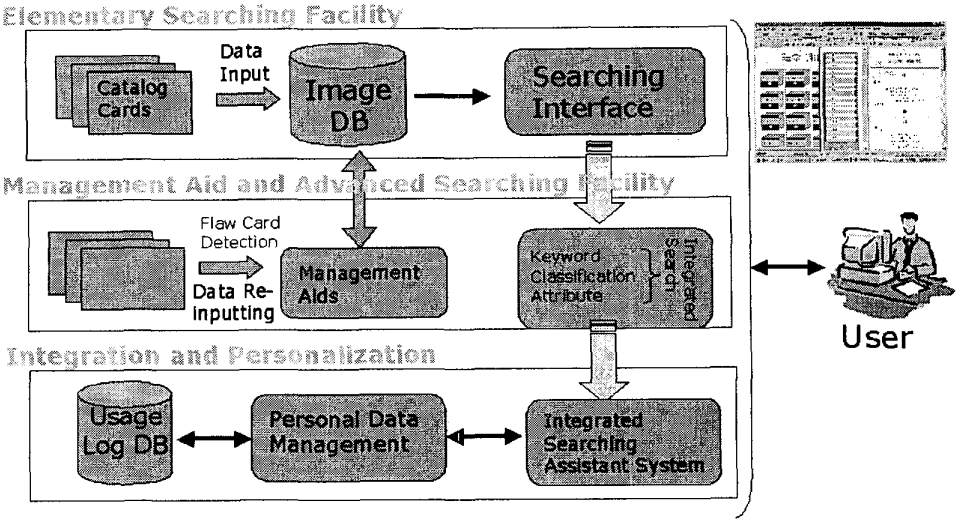

Figure 7. Advancement of the System

4. Extensions of the System

Figure 7 shows our intended plan for advancement of the system. It consists of three phases.

4.1. Elementary Searching Facility

In the first phase, elementary searching facility is implemented, which is the current status. In this phase the search is based on sorting according to the title. Usually the Japanese titles are sorted according to the Kana-order, and other titles are sorted with alphabetical order.

4.2. Management Aid and Advanced Searching Facility

In the next phase, the new features for management and some additional features for other searching facilities are added. Now we are working for these facilities.

Figure 8 is an example window of displaying both the card image and its related, or attached, information. In this example, the B-noise degree of the card, title, and authors are displayed. Note that the first card is a bad one with B-noise. The system successfully detects this.

By managing the image data and its attached data together, we can add some useful services.

- By using this form, the ordinary inputting jobs can be performed more efficiently. In the ordinary inputting the catalog information, cards should be processed one by one. Otherwise it would be easy to lose trace of which cards are ended and which are not. In this form, the correspondence of the card (image) and its corresponding input data is managed by the computer, thus you are free from worrying the loss of the correspondence.

- The next advantage of this is, relating to the previous item, it causes no problem even if you input part of the catalog data and do the rest later. By using this good feature, you can input the titles and authors first so that as many cards can be searched with keywords for the titles and the author names. Once the card image is found, the user can easily get whatever information about the card so far as it is written on the card. In this way, the keyword search facilities can be realized much faster than inputting in the ordinary way of inputting.

It would be easy to realize the keyword-based searching feature if the OCR(Optical Character Reader) is good enough for practical purposes. However its precision is still low. It varies from card to card according to the size of the characters, how old the paper is, and if the characters are printed or hand-written. In our ex-

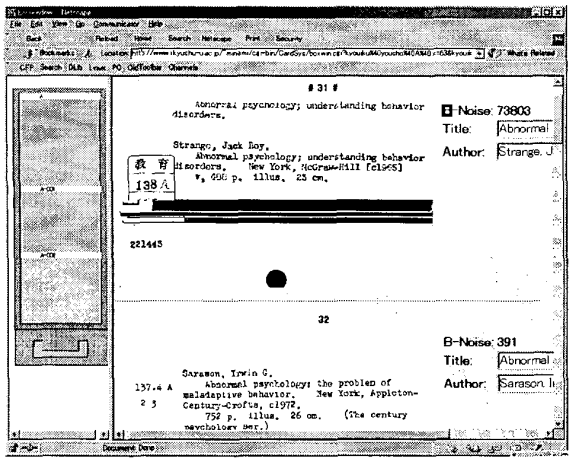

Figure 8. Card Image with Additional Information

periments, roughly speaking about 60 to 70% was the precision. Thus it is far away from practical. However some of the cards are well printed and are supposed to be read fairly well by the OCR program. They are the cards printed from the OPAC data.

We are developing a system that will find the correspondence between the OPAC and data and its printed card. So far, we have developed a program which recognize the card image created from the OPAC data[3]. As the next step we are going to develop a program that will apply an OCR program to the detected OPAC card image, then take matches to the OPAC database, and determine what data is printed on the card. By using this method we can identify the OPAC data image so that we can make the searching area narrower.

4.3. Integration and Personalization

There are two key features in this phase. The first one is the integration of this system to OPAC system. In the previous phase we are supposed to integrate the three styles of search mechanisms; i.e. keyword, classification, and attribute-based styles of search. So it becomes fairly easy to integrate the system with OPAC, which provides the keyword search facility to the library materials.

Another key feature is the personalization. The system collects the log-data of the users of the system so that each user is able to have the services adapted to their field of work, and various personal preferences. The system will be fully integrated with various styles of searching facilities with such personalization.

5. Related Work

The libraries of Keio University and Princeton University provide catalog card image searching facility. In the system of Keio University[2], it is provided as a part of the OPAC system for Chinese, Korean, Arabic and Russian documents. The image of the catalog data is attached to the catalog information expressed by the Romaji(Roman spelling) and Ping spelling. This is quite a good way for documents written in foreign languages because even if the system lacks the font data, the whole card data can be displayed as an image data so that users can recognize the original information. This method matches well with our policy that it is indispensable to integrate the keyword-based search method and image-based search method.

Princeton University Library provides a means of searching for catalog card images as supplementary catalog search[7]. In their system, the image catalog contains the record items cataloged before 1980. The ordinary OPAC system covers those items from 1980. This system is basically the same with our system, except we intend to integrate the keyword-based OPAC-like searching and classification-based image searching facilities.

The Library of Virginia[8] provides the image catalog data for the collections of the library. Some manuscripts are also provided in their images. This attempt is a good example of using catalog card images. However they are provided one by one so that the system is not sufficient for dealing with a huge amount of cards.

6. Concluding Remarks

In this paper we presented a new approach to digitizing the catalog data of libraries most of them are currently accessible only with searching paper catalog cards. In our method, firstly we scan all the cards with remarkably small cost and time, then we provide the Web-based catalog card searching service by using these catalog card images.

By taking this approach we have several benefits including:

- It is quite easy to provide the searching service on the Internet because the cost and time are very low. Roughly speaking it would take only one month for one million cards. In the current approach of manually inputting the card data, it would take more than two decades until providing the searching service for all the catalog data collected in the libraries. Once the complete database is constructed we can provide an

electronic service for searching for catalog data for the materials of the library, which is very important for database services.

- The card images are also useful to accelerate the inputting of the catalog data, because as in Figure 8 the librarians are able to input the catalog data by just seeing the card image and its inputted data at the same time. Further, it is possible to input part of the data, such as titles and authors, first and input the rest later on. By taking this way, it is quite easy to input all such data that are useful for keyword searching.

- It is possible to integrate the classification-based search and the keyword-based search. The former one is the method taken in the current system and was described in Section 3. The users will find more and more books in the keyword-based searching facility as more and more catalog data are stored as coding data.

For the further research topics, we would list up the following issues:

- Assisting facilities for management of the system including the database for the card images and their related information,

- Integrating different types of searching method such as keyword-based, classification-based, and attribute-based search,

- Personalizing the information finding service so that the users can get the results according to their research field, interests, personal preference, and so on.

References

[1] Dewey Decimal Classification:
`http://www.oclc.org/oclc/fp/index.htm`

[2] Keio University Library:
`http://catalog.lib.keio.ac.jp/ckabooks/` (in Japanese)

[3] Kurita, H.: Information Search System for the Imaged Library Catalog Cards, Master's Thesis, Department of Informatics, Kyushu University, 2000. (in Japanese)

[4] Kyushu University Library:
`http://www.lib.kyushu-u.ac.jp/`
(in Japanese)

[5] Matsukawa, S. and Minami, T.: Bottlenecks for Inputting the Library Catalog Card Images –Checking the Validity of Huge Amount of Data–, 2000. (in Japanese) (to be presented in the 19th Digital Library Workshop)

[6] Nagao, M.(Ed.): I.R.101: An Introduction To Information Retrieval, Kyoto University, 1999. (in Japanese)

[7] Princeton University Library:
`http://imagecatl.princeton.edu/ECC/`

[8] The Library of Virginia:
`http://image.vtls.com/collections/`