A comparison of mapping strategies from DDC to CLC*

Fang LI† & Yihua ZHANG
Shanghai Jiao Tong University Library, Shanghai 200240, China

Abstract

Purpose: This study aims to discuss the strategies for mapping from Dewey Decimal Classification (DDC) numbers to Chinese Library Classification (CLC) numbers based on co-occurrence mapping while minimizing manual intervention.

Design/methodology/approach: Several statistical tables were created based on frequency counts of the mapping relations with samples of USMARC records, which contain both DDC and CLC numbers. A manual table was created through direct mapping. In order to find reasonable mapping strategies, the mapping results were compared from three aspects including the sample size, the choice between one-to-one and one-to-multiple mapping relations, and the role of a manual mapping table.

Findings: Larger sample size provides more DDC numbers in the mapping table. The statistical table including one-to-multiple DDC-CLC relations provides a higher ratio of correct matches than that including only one-to-one relations. The manual mapping table cannot produce a better result than the statistical tables. Therefore, we should make full use of statistical mapping tables and avoid the time-consuming manual mapping as much as possible.

Research limitations: All the sample sizes were small. We did not consider DDC editions in our study. One-to-multiple DDC-CLC relations in the records were collected in the mapping table, but how to select one appropriate CLC number in the matching process needs to be further studied.

Practical implications: The ratio of correct matches based on the statistical mapping table came up to about 90% by CLC top-level classes and 76% by the second-level classes in our study. The statistical mapping table will be improved to realize the automatic classification of e-resources and shorten the cataloging cycle significantly.

Originality/value: The mapping results were investigated from different aspects in order to find suitable mapping strategies from DDC to CLC while minimizing manual intervention. The findings have facilitated the establishment of DDC-CLC mapping system for practical applications.

Keywords Dewey Decimal Classification (DDC); Chinese Library Classification (CLC); Co-occurrence mapping; Mapping strategies

* This work is jointly supported by the Foundation for Humanities and Social Sciences of the Chinese Ministry of Education (Grant No.: 11BTQ007) and Shanghai Society for Library Science (Grant No.:10BStX02).
† Corresponding author: Fang Li (E-mail: fl@lib.sjtu.edu.cn).
1 Introduction

Distributed digital libraries and online information resources have recently been growing rapidly. Information resources are indexed in many different ways, often with different subject vocabularies, and are organized according to different schemes. In order to provide better user experience of searching, browsing, and one-stop seamless services in digital libraries or subject information portals, it is significant to render different knowledge organization systems (KOS), such as controlled vocabularies and classification schemes, interoperable and establish interoperability models of different KOS with high reliability.

1.1 The latest researches

KOS is a general term referring to the tools that present the organized interpretation of knowledge structures\(^{[1]}\), including classification schemes, thesaurus, etc. Interoperability has been defined as the ability of two or more systems or components to exchange information and use the exchanged information without special effort on the part of either system\(^{[2]}\). To make different KOS compatible, many efforts have been concentrated on term or concept mapping. Manual mapping or direct mapping was the major method in the beginning. Since direct mapping is largely conducted by hand, in a labor-intensive and error-prone process, many researchers began to develop semi-automatic or automatic mapping systems\(^{[3]}\).

Projects using the mapping methods have been reported at various conferences and project websites since the early 1990s, such as Classification Web\(^{[4]}\), WebDewey\(^{[5]}\), the Unified Medical Language System (UMLS)\(^{[6]}\), the High-Level Thesaurus Project (HILT)\(^{[7]}\), Renardus\(^{[8]}\), the project of mapping between the Mathematics Subject Classification (MSC) and Schedule 510 (Mathematics) in DDC20\(^{[9]}\), and so on. These projects attempted to establish interoperability among KOS in different languages and with similar or different structures. With the development of computer-aided methods, some new methods and technologies used in achieving interoperability of KOS have been developed such as defining mapping relations based on co-occurrence, satellite and leaf node linking, switching languages, linking through a temporary union list, and linking through a thesaurus server protocol\(^{[10]}\). The development of computerized mapping systems and the innovation of some existing matching tools have also accelerated the automation of KOS mapping.

1.2 Major challenges and solutions in KOS mapping

In view of the projects for cross-language and cross-structure mapping, it is not surprising to find that interoperability can be achieved through a variety of approaches and mapping vocabularies is still a largely intellectual effort. Researchers summarized
the following problems associated with interoperability of KOS and presented possible solutions for the future research:

- Mapping multilingual vocabularies and integrating the views of different cultures are always the major problems when attempting to integrate or map KOS in different languages and involving different cultures. The conceptual model, link discovery tools, ontology alignment tools, term variants in the matching algorithm are used to solve real-world discovery problems in multicultural and multilingual contexts[11-14].
- Mapping KOS in different structures is an extremely challenging task because the source and target systems possess different structures and characteristics. Developing concordances or translating between a thesaurus and a classification or among various systems constructed on different principles will be a possible method[3,11-14].
- Mapping methods rely almost completely on intellectual efforts. Computer-aided techniques are increasingly used to achieve or improve interoperability. Both human mapping and computer-aided mapping will co-exist for a period of time[11-14].
- Due to the complexity of KOS mapping, some strategies have been developed and reported in the mapping efforts worldwide. For instance, Doan et al. reported that some concept nodes cannot be matched automatically because they are simply ambiguous. Their solution was to incorporate user interaction in the matching process[3].

1.3 Researches on mapping between DDC and CLC

In China, the library and information professionals have explored and employed various methods to establish mapping relations between DDC and CLC. Dai and Hou[15] tried to build an automatic mapping system between DDC and CLC by computing the class concept similarity. Liu et al.[16] put forward a new idea of mapping keywords in different indexed records’ databases to build a unified knowledge organization system automatically. Jia and Hao[17] studied on the direct mapping rules of combination categories between CLC and DDC. Our previous research[18] set out in a new direction, i.e. finding mapping relations between classification numbers of bibliographic records by designing an automatic mapping system based on a statistical mapping table and a manual one.

Based on the researches investigated, the following is a summary of some common problems currently[19]:

- Interoperability methods rely almost completely on intellectual efforts.
- The related projects have been limited to a certain range of disciplines.
Few systems have been successfully developed and they are also far from application.
Few studies are published on the construction of a comprehensive mapping system.
Indexed data in metadata records are not fully used in the mapping process.

1.4 The research purpose

Continuing our previous study of designing an automatic mapping system between DDC and CLC based on the co-occurrence mapping method, which is the first such an attempt in China, this study will focus on a discussion of our mapping strategies. Our study selects Schedule 620 (Engineering & allied operation) as an example, aiming at exploring an automatic discovery mechanism of mapping relations based on co-occurrence, and presenting a comparative analysis of different mapping strategies.

2 Design ideas of DDC-CLC mapping

As mentioned above, mapping is a popular method and also the key in the research projects of KOS interoperability. Zeng and Chan[1] and Vizine-Goetz et al.[20] divided mapping into direct mapping and co-occurrence mapping. Direct mapping is to establish equivalence between terms in different controlled vocabularies or between classification numbers. Co-occurrence mapping is to establish mappings by collecting co-occurrence frequency counts of terms or classification numbers from different schemes in the same bibliographic records.

Our research aims to study the strategies for mapping from DDC to CLC. DDC and CLC are two classification schemes with similar concept descriptions and hierarchical structures, yet in different languages. We attempt to establish DDC-CLC mapping tables mainly based on statistical co-occurrence mapping, as utilizing classification numbers rather than class concepts can avoid complex translation and comparison of the terms or concepts. However, the DDC numbers in the bibliographic records are unlikely to cover the complete classification scheme for the co-occurrence mapping. Therefore, DDC-CLC relations created manually based on direct mapping are supplementary to the statistical mapping tables.

Figure 1 illustrates the design idea of our DDC-CLC mapping system, which aims at automatic generation of CLC numbers via the existing DDC numbers in the bibliographic records. There are mapping process and matching process. The mapping process is to establish the DDC-CLC mapping table including getting a sample of bibliographic records, extracting DDC-CLC relations from the sample, selecting co-occurrence DDC-CLC relations to establish the statistical mapping
A comparison of mapping strategies from DDC to CLC

Fang LI & Yihua ZHANG
Research Paper

National Science Library,
Chinese Academy of Sciences
http://www.chinalibraries.net

51

Specific mapping strategies are necessary to implement the practical mapping process. For instance, we have to consider issues like the source of the records sample, the sample size, the methods of selecting reliable statistical DDC-CLC relations, and the necessity to add mapping relations manually, etc. In Fig. 1, the thickness of the arrows roughly represents the data volumes involved. In the matching process each DDC number will be provided with a mapped number by using the mapping table. We have made three assumptions of the mapping strategies in order to verify the feasibility of the mapping process.

Assumption 1: A larger sample size of bibliographic records will increase the mapping accuracy.

Assumption 2: A mapping table including one-to-multiple DDC-CLC relations will produce better results than a mapping table including only one-to-one DDC-CLC relations.

Assumption 3: Statistical mapping tables will provide more accurate mapping results than the manual mapping table.

3 Experiment and results

We did all the comparison of mapping strategies by taking Schedule 620 (Engineering & allied operation) as an example. A small sample and a large sample are used to...
establish different statistical mapping tables, and there is also an independent test sample to examine the mapping results based on statistical or manual mapping table. All of the three samples are composed of bibliographic records with both DDC and CLC numbers. Detailed description is as follows.

### 3.1 Small sample vs. large sample

Sample A is a small sample of 1,449 USMARC records provided by China Academic Library & Information System (CALIS) Union Catalog Center\(^c\). The records all contain DDC numbers in 620 and are distributed in 1,164 DDC-CLC relations, including 586 different DDC numbers. All the relations were sorted by DDC numbers in ascending order and by frequency in descending order as illustrated in Table 1. The DDC-CLC relations with the highest frequency were picked out for each DDC number to establish a statistical mapping table (called MTA). For example, the DDC number 621.395 was matched to the CLC numbers TN47, TP331, TN407, TN79, etc. As 621.395-TN47 had the highest frequency, TN47 would be the mapped CLC number for 621.395 in MTA. MTA totally consisted of 586 DDC-CLC relations.

<table>
<thead>
<tr>
<th>DDC numbers</th>
<th>CLC numbers</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>620</td>
<td>TB1</td>
<td>3</td>
</tr>
<tr>
<td>620</td>
<td>TB23</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>621.395</td>
<td>TN47</td>
<td>5</td>
</tr>
<tr>
<td>621.395</td>
<td>TP331</td>
<td>3</td>
</tr>
<tr>
<td>621.395</td>
<td>TN407</td>
<td>2</td>
</tr>
<tr>
<td>621.395</td>
<td>TN79</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Sample B is a larger sample including Sample A. Sample B contains more CALIS records and some records from Shanghai Jiao Tong University (SJTU) Library\(^d\) as well. It consists of 20,006 USMARC records and is distributed in 9,808 DDC-CLC relations including 2,534 different DDC numbers in 620. Similarly, another statistical mapping table (called MTB1) was established according to the highest frequency of the mapping relations and it contained 2,534 DDC-CLC relations.

\(^c\) China Academic Library and Information Systems (CALIS) is a national academic library consortium in China. Cataloguers from Chinese academic libraries copy records in CALIS Union Catalog Center database. Records are checked by cataloging specialists and quality control programs and thus considered as authority records.

\(^d\) SJTU Library is qualified to contribute its original records to CALIS Union Catalog Center database. Bibliographic records from SJTU Library are considered as qualified records.
A comparison of mapping strategies from DDC to CLC

Sample T consists of 1,515 Springer e-Book records\(^\circ\) with the DDC numbers in 620 and their corresponding CLC numbers. It is a test sample, independent of Sample A or Sample B, to compare the accuracy of the mapping results from the two statistical mapping tables, i.e. MT_\(A\) and MT_\(B1\).

Through the statistical mapping tables, a CLC number would be given to each Sample T record according to its DDC number. Then the given CLC numbers were compared with the original CLC numbers of Sample T to measure the mapping accuracy. To be brief, the following abbreviated forms are defined.

- A DDC_\(T\) number: A DDC number of a Sample T record.
- A CLC_\(T\) number: A CLC number of a Sample T record.
- A DDC_\(M\) number: A mapped DDC number, which can be the same number in the mapping table as the DDC_\(T\) number or a certain superordinate class number of the DDC_\(T\).
- A CLC_\(M\) number: A mapped CLC number given to the DDC_\(M\) number in the matching process.
- \(P_D\): The percentage of sample records, of which DDC_\(M\) numbers and DDC_\(T\) numbers are identical.
- \(P_1\): The percentage of sample records, of which CLC_\(M\) numbers and CLC_\(T\) numbers contain the same CLC top-level classes. For instance, CLC_\(M\) number TN47 and CLC_\(T\) number TP331 contain the same CLC top-level class T.
- \(P_2\): The percentage of sample records, of which CLC_\(M\) numbers and CLC_\(T\) numbers contain the same CLC second-level classes. For instance, CLC_\(M\) number TN47 and CLC_\(T\) number TN402 contain the same CLC second-level class TN.
- \(P_3\): The percentage of sample records, of which CLC_\(M\) numbers and CLC_\(T\) numbers contain the same CLC third-level classes. For instance, CLC_\(M\) number TN47 and CLC_\(T\) number TN402 contain the same CLC third-level class TN4.
- \(P_4\): The percentage of sample records, of which CLC_\(M\) numbers and CLC_\(T\) numbers contain the same CLC fourth-level classes. For instance, CLC_\(M\) number TN407 and CLC_\(T\) number TN402 contain the same CLC fourth-level class TN40.
- \(P_C\): The percentage of sample records, of which CLC_\(M\) numbers and CLC_\(T\) numbers are identical.

\(^\circ\) Springer e-book records are downloaded from Online Computer Library Center (OCLC) WorldCat, which is a global network of library content and services. WorldCat holds 271 million bibliographic records adhered to international standards.
$P_D$ actually serves as a coverage indicator, implying how many DDC_T numbers can be found in the mapping tables. $P_1$, $P_2$, $P_3$, $P_4$, and $P_C$ provide us with a general idea of mapping accuracy at different levels. If only correct CLC top-level classes are required for the mapping, the mapping accuracy is likely to be $P_1$. Similarly, if correct CLC second-level classes are required, the mapping accuracy is likely to be $P_2$. If completely correct CLC numbers are required, the mapping accuracy is likely to be $P_C$.

The testing process is illustrated in Fig. 2.

Figure 3 illustrates the percentages of the sample records in terms of different levels of similarities between the mapped numbers and the original numbers of Sample T. A total of 1,230 DDC_T numbers were found in MT_A, thus $P_D$ of MT_A was 81.2%. Other 284 DDC_T numbers needed to be shortened so that they could be found in MT_A and one DDC_T number could not be found. $P_D$ of MT_B1 was 93.3%, much higher than that of MT_A. It was the large sample that provided more DDC numbers in the mapping table. $P_1$ was 77% for MT_A, i.e. 77% of the mapped CLC numbers in the mapping table. $P_1$ of MT_B1 was 79.8%, slightly higher than that of MT_A. With the increase of the order of classes, the percentages of accuracy, i.e. those from $P_1$ to $P_C$ for both large and small samples, dropped significantly. Low proportion of the mapped CLC numbers was identical with the CLC_T numbers. Although Sample B’s size was ten times more than the size of Sample A, the mapping results of MT_B1 were not much better than those of MT_A. Nevertheless, the mapping accuracy of MT_B1 remained about 3 percentage points higher than that of MT_A. The first assumption was weakly verified that a larger sample size could lead to a bit better results.
3.2 One-to-one vs. one-to-multiple DDC-CLC relations

In the above comparison, even with a larger sample size, $P_1$ reached only nearly 80%. It was found that one-to-multiple DDC-CLC relations were the main reason why the rest 20% CLC numbers do not contain the same CLC top-level classes with the CLCT numbers. Therefore, rather than one-to-one DDC-CLC relations with the highest frequency, more DDC-CLC relations from Sample B were picked out to establish a mapping table.

A statistical mapping table (called MTB2) was created to contain DDC-CLC relations with the first and the second highest frequencies for each DDC number. Similarly, MTB3 was composed of DDC-CLC relations from the first to the third highest frequency for each DDC number, and MTB4 was composed of DDC-CLC relations from the first to the fourth highest frequency.

As there were one-to-multiple DDC-CLC relations in the mapping tables, the testing process should be modified (Fig. 4). All the mapped CLC numbers were compared with the CLCT numbers and the nearest CLC number would be selected. In other words, it is the best if the two CLC numbers are identical, otherwise the longer their same superordinate class numbers, the better. For instance, as for the DDC number 621.395, the CLC numbers, TN47, TP331, TN407, and TN79, were selected into MTB4 by frequency. If a CLCT number was TN47 for 621.395, obviously TN47 would be selected as the CLC number. If a CLC number was TN402, TN407 would be the nearest CLC number to TN402 in this case.

Figure 5 shows that there was no change in the percentages of MTB1. $P_0$ of all the four mapping tables was 93.3%. $P_1$ increased from 79.8% for MTB1 to 89.6% for MTB4. $P_3$, $P_4$ and $P_C$ increased about 15 percentage points from MTB1 to MTB4. It is illustrated that the percentages increased obviously from MTB1 to MTB2 and
National Science Library, Chinese Academy of Sciences

Research Paper

from MT_{B2} to MT_{B3}. Nevertheless, from MT_{B3} to MT_{B4}, they tended to be stable, as fewer DDC-CLC relations were added into the mapping table. Anyway, the second assumption was significantly verified.

$$\begin{array}{c|c|c|c|c|c|c}
\text{Is the DDC No. a section?} & \text{No} & \text{Yes} \\
\hline
\text{The mapped CLC No. is given.} & \text{No} & \text{Yes} \\
\hline
\text{All the mapped CLC numbers are compared with CLC}_1. & \text{No} & \text{The nearest mapped CLC No. is given.} \\
\end{array}$$

Fig. 4 Testing process with one-to-multiple DDC-CLC relations.

3.3 Statistical tables vs. manual mapping table

Before adding DDC-CLC relations manually to the statistical mapping tables, we needed to examine how many class numbers listed in the DDC schedules can be
covered by the record samples. Schedule 620 consists of more than 1,800 class numbers. The size of Sample B reaches 20,006 records, about 10 times the number of the 620 classes in the schedules. However, the statistical mapping tables from Sample B only contain 2,534 different DDC numbers in 620, of which 866 numbers are listed in the DDC schedules and the other 1,668 numbers are not listed in the schedules. Table 2 lists details of the two samples. As a large proportion of the DDC numbers listed in the schedules were not covered by the co-occurrence mapping relations in the bibliographic records, it was very likely that in the application of the mapping table a DDC number would fail to be matched in the statistical mapping table. In this case, the table needed to be supplemented by adding mapping relations manually.

<table>
<thead>
<tr>
<th>Sample</th>
<th>No. of records</th>
<th>No. of different DDC-CLC relations</th>
<th>No. of different DDC numbers</th>
<th>No. of DDC numbers listed in the schedules</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1,449</td>
<td>1,164</td>
<td>586</td>
<td>397</td>
</tr>
<tr>
<td>B</td>
<td>20,006</td>
<td>9,809</td>
<td>2,534</td>
<td>866</td>
</tr>
</tbody>
</table>

Then the accuracy of the mapping relations added manually was examined to determine if it was reasonable to utilize these relations to supplement the statistical tables. All the class numbers listed in Schedule 620 were given one or more recommended CLC numbers according to the class concepts to create a manual mapping table, called MTC. With one-to-multiple relations in the manual mapping table, its testing process was the same as that demonstrated in Fig. 4. The mapping results are illustrated in Fig. 6. $P_d$ of $MT_c$ was 82.8%, similar to that of $MT_A$. It seemed that more DDC numbers...
in the large sample were built or synthesized numbers that are not specifically listed in the schedules and they improved considerably the value of $P_D$. $P_1$ of MT$_C$ was 83.4%, between the $P_1$ values of MT$_{B1}$ and MT$_{B2}$. $P_2$, $P_3$, $P_4$ and $P_C$ of MT$_C$ were similar to those of MT$_{B1}$. The data verified the third assumption partially that the statistical mapping tables with one-to-multiple relations can provide better mapping results than the manual table.

4 Discussion

This study selected Schedule 620 as an example and discussed the mapping strategies from DDC to CLC based on the co-occurrence mapping method. We chose authoritative bibliographic records from CALIS and SJTU Library as the statistical sample and Springer e-book records as the test sample. Human-induced classifying discrepancy is thought to be eliminated to some extent in the selection of DDC-CLC relations from large records samples. It is supposed that different records samples should not cause large variances in mapping results.

A small sample and a large sample were first compared to investigate the effect of the sample size on the mapping results. One finding was that the larger sample produced a much higher coverage of DDC numbers in test sample records, more specifically, 81.2% for Sample A and 93.3% for Sample B. Nevertheless, Sample B covered less than half class numbers listed in the DDC schedules. The reason is that DDC specifies only a fraction of all possible class numbers but allows cataloguers to build, or synthesize, numbers for a great many subjects (or aspects of subjects) not listed in the schedules. This is accomplished by the addition of extensions derived from various tables or by following specific instructions for number building in particular classes[21]. Therefore, from a perspective of the application of the DDC-CLC mapping table, we should focus on the widely used DDC classes and do not need to provide CLC numbers for all the class numbers listed in the DDC schedules. However, the test sample size in our study was quite small. A larger test sample should be used to measure the coverage of the statistical mapping table in our future study. Another finding was that a larger sample could only produce a bit better mapping results, i.e. percentages of the accuracy remained about 3 percentage points higher. The main reason was that only the DDC-CLC relations with the highest frequency were picked out to create the statistical mapping table. In this case, the sample size played a limited role in increasing the accuracy of the mapping results.

Selection of DDC-CLC relations with the highest frequency for the statistical table was not suitable sometimes, as one DDC class is likely to be related to several CLC classes. Multiple DDC-CLC relations of Sample B were then picked out in
A comparison of mapping strategies from DDC to CLC

establishing several other mapping tables. It was indicated that the percentages of mapping accuracy obviously increased about 10 to 15 percentage points from one-to-one to one-to-three relations reserved in the mapping tables. Yet from one-to-three to one-to-four relations, the percentages rose very slightly. Thus, it is justified that up to three CLC numbers are reserved for one DDC number so as to reduce complexity in the subsequent matching process with reasonable mapping accuracies. How to select reliable DDC-CLC relations from the bibliographic records is the key to improve the accuracy.

The comparison between the statistical tables and manual mapping table demonstrated that it was reasonable to utilize manual mapping relations to supplement statistical ones. The coverage indicator of MTc was close to that of MTa and the percentages of mapping accuracy of MTc were approximately the same as that of MTb1. The statistical mapping tables with one-to-multiple relations, i.e. MTb2, MTb3, and MTb4, produced better results than the manual one. Therefore, statistical mapping rather than the time-consuming manual mapping should be utilized as much as possible.

5 Conclusion

This study discussed the mapping strategies that specified how to establish the statistical mapping tables to map from a DDC number to a CLC one while minimizing manual intervention. The three assumptions for the mapping strategies were verified more or less to improve the efficiency of the mapping process.

There were several limitations of the mapping system. First of all, the size of Sample B in our study was still very small. Less than half class numbers in Schedule 620 were covered by the statistical mapping tables. Thus, the sample size should be increased to provide more DDC numbers so as to reach a higher coverage in the mapping process. Secondly, only DDC-CLC relation frequencies were taken into account in establishing the statistical tables. The CLC numbers might be combined to obtain better DDC-CLC relations in order to increase the accuracy of the results. Thirdly, the test sample size was very small and a comprehensive test should be conducted. Fourthly, the impact of different DDC editions should be studied in our future research. Finally, although one-to-multiple DDC-CLC relations in the records were collected in creating the mapping table, how to select one appropriate CLC number in the subsequent matching process needs to be further studied.

In our future study, LCC and LCSH will be introduced into the selecting process. LCC-CLC relations or LCSH-CLC relations can be collected and be utilized to narrow our choices when dealing with one-to-multiple DDC-CLC relations.
References


A comparison of mapping strategies from DDC to CLC

Reference Citation Format

The format for citations in text and for bibliographic references follows GB/T 7714—2005. The citation should be ordered in number as it appears in the text of the submitted article.

• For journal article


• For book


• For proceedings


• For electronic journal article


• For electronic book


• For thesis or dissertation

Submission Guidelines

♦ Aims
Chinese Journal of Library and Information Science (CJLIS), being sponsored by the Chinese Academy of Sciences (CAS) and published quarterly by the National Science Library of CAS, is a scholarly journal in the field of library and information science (LIS). Its aim is to provide an international communication link between researchers, educators, administrators, and information professionals.

With the publication of the research results both from China and from other foreign countries, the Journal CJLIS strikes a balance between theory and practice. With its goal to provide an open forum for Chinese and international scholars in this field to exchange their research results, CJLIS also offers new possibilities in the advancement of Chinese library operations. The CJLIS tries to establish a platform for LIS students, researchers and library staff all over the world to engage in intellectual dialog and also to improve library services so as to promote even more quickened and substantial development of LIS in China.

♦ Scope
Striving toward academic excellence, innovation, and practicality, the CJLIS mainly includes research papers both on the theoretical as well as on the practical fronts in all aspects of the field. More specifically, it includes but not limited to informatics, library management, information technology application, knowledge organization system, knowledge management, archives, permanent preservation of library resources, LIS education, and so on.

♦ Refereeing Process
Articles and papers covering the topics or themes mentioned above, will be refereed through a double-blind peer review process.

♦ Editorial Advisory Board
The Editorial Board is composed of the nationally and internationally well-known scholars and researchers in the LIS field, the high quality of this Journal is thus reasonably assured.

♦ Manuscripts Categories
As the first English-language academic journal on LIS published in Mainland China, the CJLIS will take a proactive attitude to trace and report the prevailing hot issues in the field around the globe as well as the more serious scholarly communications. As such, the submitted manuscripts are classified into constant categories and unfixed categories. In the former category, research papers, library practice and progress reports are the essential components. In the latter, book reviews, biographical sketches, anecdotes, reminiscence of prominent librarians and brief communications will appear occasionally.

Research papers represent original research work or a comprehensive and in-depth analysis of a topic. More than 3,000 words are considered as a proper length for such manuscripts, with an abstract between 100 to 200 words.

Library practice covers the latest development and application in any segment of library field work and information service. The length of the manuscript is preferred to be more than 3,000 words, with an abstract between 100 to 200 words.

Progress reports reflect the projects result or research progress on the key topics of the library and information science. Submissions of articles to this section are expected to be comprehensive and analytical, which may deepen the understanding of the discussed issue and stimulate further researches on the topics, or give a new perspective on future technological applications. The manuscript length should be within 10,000 words, with an abstract between 100 to 200 words.

♦ Manuscripts Requirements
All papers can be submitted either in English or in Chinese (or both) with a double-line space. For the assurance that all the materials of the to-be-submitted are included, please check the following:

Title. Please give a brief biographical introduction to all contributing authors and their research background on a separated paper. For a better organization of the paper, please use the headings and subheadings.

Authors and affiliations. Please do not forget to write down the mailing address of each and every article contributor.

References. Be sure all the references used should be cited properly in both in-text and in bibliography. Particular attention should be paid to the proceedings. Do not forget adding the name(s) of editors of the compilation, as well as the name of the publishers. For the detailed information please request a copy of Reference Citation Format.

♦ Copyright
All submitted papers normally should not have been previously published nor be currently under consideration for publication elsewhere. For all the materials translated or obtained from other published resources, they should be properly acknowledged. All copyright problems should be cleared without any legal entanglements prior to the publication.

♦ Notes for Intending Submissions
A guide for authors and other relevant information, including submitting papers online, is available at the website of the Editorial Office of the CJLIS (http://www.chinalibraries.net). For any questions, you can e-mail the Office or directly to:

Prof. Zhang Xiaolin
Editor-in-Chief of CJLIS
The CJLIS Editorial Office
National Science Library, Chinese Academy of Sciences
No.33 Beishuan Xilu, Zhongguancun, Haidian District,
Beijing 100190, P.R. China
Tel: 86-10-82624454 or 86-10-82626611 ext. 6628
Fax: 86-10-82624454
E-mail: chinalibraries@mail.las.ac.cn
Website: http://www.chinalibraries.net

♦ Subscription
For single-copy subscription in China: RMB ¥ 200/issue. For subscription outside of China, US $ 199/Volume yearly (including air shipping)
CONTENTS

Research Papers

1 Defining an open access resource strategy for research libraries: Part III —The strategies and practices of National Science Library
Xiaolin ZHANG, Xiwen LIU, Lin LI, Yan ZENG & Li-Ping KU

12 A view on big data and its relation to Informetrics
Ronald ROUSSEAU

27 Person-specific named entity recognition using SVM with rich feature sets
Hui NIE

47 A comparison of mapping strategies from DDC to CLC
Fang LI & Yihua ZHANG

62 A study of information exchange through social networks in rural China
Ya LIU

76 Is it time for wider acceptance of e-textbooks? An examination of student reactions to e-textbooks
Ziming LIU

News

88 Reference Citation Format