Journal of East Asian Libraries

Volume 2005 | Number 137 Article 7

10-1-2005

Vernacular Searching and Retrieval in OCLC Connexion Client

Daphne Wang

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

BYU ScholarsArchive Citation
Wang, Daphne (2005) "Vernacular Searching and Retrieval in OCLC Connexion Client," *Journal of East Asian Libraries*: Vol. 2005 : No. 137 , Article 7.
Available at: https://scholarsarchive.byu.edu/jeal/vol2005/iss137/7

This Article is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Journal of East Asian Libraries by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Journal of East Asian Libraries, No. 137, October 2005

VERNACULAR SEARCHING AND RETRIEVAL IN OCLC CONNEXION CLIENT

Daphne Wang **University of Oregon**

Introduction

OCLC Connexion client, the new platform for integrated cataloging and metadata services, offers greatly improved bibliographic indexing and retrieval capabilities as well as a wide range of enhancements in cataloging functionality. Notably, the most significant improvement in non-Latin script cataloging is the total access to vernacular data in bibliographical records. For the first time, non-Latin script (vernacular) data in bibliographic records is fully indexed to the same level as Latin data has been treated in bibliographic indexing. As a result, the efficiency and quality of vernacular data retrieval in WorldCat are far superior to the searching capability and retrieval performance rendered by the previous versions of the OCLC CJK software.

There are many unique features to be explored for non-Latin script cataloging in Connexion client, even though much of the cataloging functionality available in the previous version of OCLC CJK software has been integrated and carried over to Connexion client. Apart from the implementation of Microsoft IMEs (Input Method Editors) for CJK data entry, vernacular searching in Connexion client and constructing vernacular search keys represent the areas with the greatest changes in non-Latin script cataloging. The intent of this article is to discuss and share our learning experience in searching and retrieval of non-Latin script (vernacular) records in Connexion client. The examples of bibliographic records used in the discussion are selected from the East Asian languages records in the WorldCat database.

1. New features in Connexion client for non-Latin script searching

In Connexion client, most indexes for Latin script searches also apply to vernacular searches, with one exception: the derived searches are unavailable for vernacular searches. Derived vernacular searches such as VP (personal names), VC (corporate names), VT (titles) and VA (name/title) previously available in OCLC CJK 3.11 are no longer offered in Connexion client. Conventional derived searches for Latin data such as 3,2,2,1 (for titles) and 4,3,1 (for personal names), etc. still work for romanized data (Latin script transliteration of vernacular data) contained in non-Latin script bibliographic record; they just do not work for vernacular data in such records.

Connexion client provides comprehensive indexing of vernacular data, and offers flexibility and a wide range of options in searching methods. With the previous versions of OCLC CJK software, vernacular searching in WorldCat was severely limited. Vernacular data in note fields and in subject headings was not indexed; derived vernacular searches were inadequate to retrieve many relevant records, and the unavailability of vernacular keyword searching further compromised the quality of retrieval results. These problems have been effectively eliminated as a result of extended indexing in Connexion client. Mr. David Whitehair, the Connexion client product manager of OCLC, indicates: "… now the non-Latin indexing is available in almost 100 keyword indexes, and the Latin and non-Latin scripts are included in the same indexes without separate index labels. The same indexes with the same fields and subfields are indexed for all scripts."

The comprehensive indexing in Connexion client provides enhanced access to essential vernacular information contained in the body of non-Latin script records, regardless of where such data is located in a bibliographic record, and whether or not it is in a field traditionally regarded as an access point. Not only is vernacular data in names, titles, and subject headings now searchable, vernacular data in notes, imprints and non-traced titles and series is also accessible with appropriate searching methods.

In the previous version of OCLC CJK 3.11, derived vernacular searches such as VT (titles), VC (corporate names), VP (personal names), etc. only looked for vernacular data in indexed access points, excluding subject headings. Vernacular data present elsewhere in the body of a bibliographic record was not always accessible, unless such data was in indexed fields and could be found by derived vernacular searches. For example, in a bibliographic record for a collection containing four works lacking a collective title, all four individual titles are

transcribed in the title (245) field. It was impossible to search for the fourth vernacular title because this title might not always be traced as an access point based on the relevant provision in AACR2. While in Connexion client, vernacular keyword searching is capable of retrieving the fourth title even though it is not traced in an access point field in the record.

Connexion client also offers increased flexibility and more options in performing vernacular searching such as:

- **Latin and vernacular data may be combined in the same search query.**
- **Browsing (scan indexes) vernacular words, phrases and whole phrases under various indexes.**
- **Boolean operators also apply to vernacular searches to broaden or narrow the scope of searching.**

2. Word searches and phrase searches: differences, strengths and limitations

Searches performed in WorldCat for cataloging may be roughly divided into two broad categories: target searching and reference searching. In target searching, catalogers usually have pieces in hand, and the bibliographic information needed to construct searches is readily available. For most cases simple numeric or derived searches are just as effective as more sophisticated searches in achieving a successful retrieval--finding the matching record in WorldCat if such record indeed exists in the database.
In original and copy cataloging, catalogers often search for reference records in WorldCat for a variety of reasons. This type of search is basically reference searching when catalogers may or may not have sufficient bibliographic information for certain items needed to construct numeric or derived searches. Whether or not one has accurate or adequate bibliographic information for an item definitely affects search strategies, and consequently determines how searches will be constructed.

Since derived searches are not available for vernacular searches, word searches and phrase searches become the commonly used vernacular search methods. Browsing is used less often than other search methods. Word indexes and phrase indexes currently make up the bulk of over 100 indexes available in Connexion client, which facilitate keyword searching, phrase searching and browsing across a wide range of indexes, including titles, personal names, corporate/conference names, series, subject headings, imprints, and notes, etc. To explore the strengths and differences of word and phrase indexes, and examine how the indexes work to support searching and retrieval of non-Latin script records in WorldCat, we focused on four most frequently used searching methods: keyword search (index label kw:), title keyword search (index label ti:), title phrase search (index label ti=), and title whole phrase search (index label tiw=).

Experimenting with these word and phrase searching methods not only allowed us to learn and compare the different features of each index, but also helped us gain a better understanding of how word and phrase indexes work in general in Connexion client. Although the search methods being discussed here are also relevant to Latin script searches, the primary purpose of this study is to evaluate vernacular searching and retrieval supported by various indexes in Connexion client.

■ Title phrase search (ti=)

Title phrase index is essentially the subfield index. Bibliographic information coded in each subfield in a title is independently indexed, and the data in any single subfield in the title can be separately searched for. Title phrase search (ti=) has the search criteria that have to be satisfied for successful retrieval:

- **Multiple words within a single subfield need to be entered in sequence, beginning with the first word in the subfield.**
- **All data belonging to the same subfield needs to be entered completely. If "partial" data is entered, the retrieval could be incomplete or fail, unless phrase truncation is applied in the search key.**

■ Title whole phrase search (tiw=)

Title whole phrase index is in fact the title proper index. Title whole phrase search (tiw=) also has search criteria that have to be satisfied for successful retrieval:

- **Multiple words within the title field need to be entered in sequence, beginning with the first word in the title field, across multiple subfields, and including all data in subfield ‡a, subfield ‡n, or subfield ‡p, as appropriate.**
- **Title remainder information coded in subfield ‡b is not indexed for the title whole phrase search, and thus needs to be excluded in the search key.**

- **Truncation may be applied to title whole phrase search when not all data in the title proper is entered in the search key.**

The differences in indexing between the title phrase index and the title whole phrase index affect searching and retrieval in a number of ways. We have found that the title phrase search is more flexible and often outperforms the title whole phrase search due to the following factors:

- **Data in every single subfield in the title field is indexed under the title phrase index, while title whole phrase index excludes the subfield ‡b in the title indexing. In other words, title remainder information coded in subfield ‡b is searchable under the title phrase index, but not searchable under the title whole phrase index.**
- **Under the title phrase index, bibliographic data coded in each subfield is independently indexed, and thus can be separately searched for. This means one can search for the data in subfield ‡n, or subfield ‡p in titles without having to enter the data in the preceding subfield ‡a. However, this is not the case with the title whole phrase index, in which data across multiple subfields in a title is sequentially indexed from left to right, and one must enter data in the search query beginning with the first word in the subfield ‡a, and include all data in the subsequent subfield ‡n, or subfield ‡p, as appropriate.**

Under the title whole phrase index, directly searching for bibliographic data coded in subfield ‡n or subfield ‡p may not automatically guarantee successful retrieval, unless such data is also traced in other fields such as a 246 field (varying form of title), or 730/740 fields (title added entry). As an index searcher, one does not have control over whether or not bibliographic data coded in subfield ‡n or subfield ‡p is also traced in other fields in a bibliographic record. However, it helps to be aware that the title whole phrase search attempts to match a query (phrase) against the entire title proper being indexed. If a title contains data coded in subfield ‡n or subfield ‡p, such data needs to be entered following the information coded in the preceding subfield ‡a. Otherwise, the retrieval may be compromised unless phrase truncation is applied. Here is one example:

The title whole phrase search, tiw=□□□□□, gets a system message: "No records found for your search." The search failed because the data coded in the subfield ‡p in the title field was not entered in the search key. If we try again with one of the following three search methods,

ti=□□□□□ Title phrase search
tiw=□□□□□* Title whole phrase search with truncation
sca tiw=□□□□□ Browsing (scan title whole phrase index)

three records are retrieved:

□□□□□□□. □**p** □□□□□
□□□□□□□. □**p** □□□□□
□□□□□□□. □**p** □□□□□

The purpose of the title whole phrase index is to collocate main titles (titles proper) of bibliographic entities in WorldCat. The title whole phrase index is very useful for browsing, but less efficient for searching due to its indexing structure. It is hard to imagine anyone would want to use the title whole phrase search (tiw=) to find records with long titles such as the ones in the following examples, as other search methods can retrieve these records more quickly, and with fewer words required in the search keys.

20□□□□□□□□. □**p** □□□, □□□□, □□□□□□□, □□□□□□□, □□□□
□□□□□□□□□□--□□□□□□. □**n** □□□, □**p** □□□□
□□□□□□□□□□ ; □□□□□□□□□□□□ ; □□□□□□□□□□
□□□□□ ; □□□□□ ; □□□□□ ; □□□□□□
□□□□□□ ; □□□□□□ **1** ; □□□□□□ **2** ; □□□□□□ **3**

■ **Phrase or whole phrase searches with truncation**

Title phrase and title whole phrase searches are somewhat rigid searches which require that a searcher have the prior knowledge of the exact title of a bibliographic item and input the complete data from left to right in the search query. If one fails to enter complete bibliographic data from a field or subfield in the search key, the retrieval may be adversely impacted. To reduce this inefficiency or inflexibility, using truncation in vernacular phrase searches is one way to improve the quality of retrieval.

In Latin data search, truncation must be preceded by at least three characters. However, there are no specific rules indicating where truncation should be used in vernacular phrase searches. Applying truncation after fewer words or after more words often produces different retrieval output. Phrase truncation is often useful in cases when a vernacular title is longer, or consists of multiple subfields. For short titles, applying phrase truncation may not be beneficial. The application of truncation in phrase and whole phrase searches is basically a case-by-case judgment based on particular circumstances, and it may take more than one try to get the satisfactory results.

If one prefers not to use truncation with phrase or whole phrase searches, the alternative would be browsing (scan index), as browsing is the only type of search that automatically truncates phrases, and there is no need to enter a truncation symbol (*) in the search query.

Here is one example in which the same phrase search is used with and without truncation. Truncation has obviously broadened the search and rendered a more satisfactory retrieval.

Title phrase search: ti=□□□□ 1 record retrieved
Title phrase search with truncation: ti=□□□□* 20 records retrieved

■ Title keyword search (ti:)

Title keyword index supporting word browsing and searching is a more comprehensive index than the title phrase and title whole phrase indexes; it covers many more fields and subfields in indexing. Although it is generally true that word searches can find a term regardless of its placement in a field or subfield, there are cases when word searches cannot find a term because of its placement in a field. These exceptions are illustrated in Example 7 that we will be discussing.

■ The super index – keyword (kw:)

The keyword index (index label kw:), formerly the subject/title/contents index (index label st), is an expanded index for keyword search offered in Connexion client. The keyword index (kw:) facilitates keyword searches for names, titles, subjects, and notes, etc. in bibliographic records.

The greatest advantage of the keyword index (kw:) is that it is primarily a numeric and word-based index, which goes beyond fields and subfields and indexes the very basic elements of bibliographic data -- words. The majority of the word and phrase indexes in Connexion client, including the title keyword index (ti:), title phrase index (ti=) and title whole phrase index (tiw=), rely on subfield codes in data indexing, which means the placement of subfield codes in a field or subfield determines whether or not the data in the field can be properly indexed. Keyword index (kw:) disregards subfield codes in word indexing, thus eliminating their potentially adverse effect on retrieval in some cases.

The sample cataloging records to be discussed will demonstrate that keyword searches (kw:) consistently outperform other searches such as title keyword (ti:) and phrase searches (ti= and tiw=) in achieving successful retrieval. Often, when all other types of searches fail, keyword search (kw:) is the reliable last resort. Keyword searching (kw:) works equally well for Latin or non-Latin script data, and is especially useful for reference searching when one lacks complete bibliographic data in hand. The merits and strengths of the keyword index (kw:) make it the most comprehensive, flexible and powerful index available in Connexion client. In fact, keyword search (kw:) is the only type of search truly capable of finding a term regardless of its placement in a field or subfield.

Another nice feature about keyword search is that the index label (kw:) is optional, i.e., keywords (Latin or non-Latin data alike) may be directly entered without being preceded by the indexed label (kw:), as shown in the record examples.

While keyword index (kw:) offers flexibility in searching and the benefit of comprehensive retrieval, it may inevitably retrieve some irrelevant records. In such cases, applying appropriate qualifiers or Boolean operators with keyword search (kw:) will significantly improve precision in retrieval.

3. Put the indexes to work: case studies of vernacular searching and retrieval

In the foregoing discussion, we reviewed the different features of four search methods: title phrase search (ti=), title whole phrase search (tiw=), title keyword search (ti:), and keyword search (kw:). To learn how these search methods actually work in retrieving various types of non-Latin script (vernacular) records in WorldCat, a sampling of records was selected to compare the strengths of the search methods and some of their limitations. In each example, we will see how these various search methods work, and why some of them fail. Hopefully, the comparison can offer useful information in helping select the most efficient search method given a particular situation.

Often in reality, one may not have accurate bibliographic information for certain items before entering search queries in WorldCat. Some of the sample records show that even when one has such information available, how the data is actually transcribed and coded in cataloging records may be beyond our control. It is important to be aware of the limitations of some search methods such as title phrase and title whole phrase searches, and know what they do and what they cannot do. Keyword searches prove to be more flexible, and they can overcome the limitations of phrase searches in providing better and more reliable retrieval results. Keyword searches also have more tolerance for information inaccuracy in search queries than phrase searches.

Since the primary interest of this study is to investigate how various indexes in Connexion client work to facilitate searching and retrieval of non-Latin script records in WorldCat, the romanized fields in the sample records have been removed for the simplicity of the discussion. It should also be noted that there are usually many different ways of constructing a search. The search queries demonstrated with each sample record are intended to be illustrative, not to be comprehensive, or prescriptive. Obviously, there are other search alternatives that work equally well but cannot be exhaustively listed in our discussion.

Example 1: OCLC # 24544507

100 1 □□□.
245 10 21□□: □□□□□? : □b □□□□□□□□ / □c □□□□.
250 □1□.
260 □□ : □b □□□□□□□ : □b □□□□□□, □c 1989.
300 6, 419 p. : □b ill. ; □c 19 cm.
440 0 □□□□□□□□□□
504 Includes bibliographical references.
651 0 Pacific Area □x Economic conditions.

Successful search keys	Indexes searched
□□□	kw: pn: pn=
□□□□□/1989	kw: ti:
21□□ □□□□	kw: ti:
□□□□□□□□	kw: ti: ti=
21□□□□□□□	kw: ti: ti= tiw=
□□□□□□□□	kw: ti: tiw= se: se=

Failed searches and the reasons why they failed:

ti=□□□□□	**"□□□□□" is not indexed as part of the subfield ‡b.**
tiw=□□□□□	**"□□□□□" is not at the beginning of the title proper in subfield ‡a.**
tiw=□□□□□□□□	**Title whole phrase index (tiw=) excludes data coded in the subfield ‡b, which is "□□□□□□□□" in this case.**
ti=□□□□□□□□	**Series (440/490/830 fields) are not indexed under the title phrase index (ti=).**

Example 2: OCLC # 46716629

110 2 □□□□□.

245 10 **20□□□□□□□ : □b □□□□□□□□ = 20 century Chinese fine art : collection from China National Museum of Fine Arts.**
246 30 □□□□□□□□
246 31 **20 century Chinese fine art**
250 **□1□.**
260 **□□□ : □b □□□□□□□□□ ; □a □□□ : □b □□□□□□□, □c 1999.**
300 **3 v. : □b ill. (some col.) ; □c 34 cm.**
546 **In Chinese.**
505 0 **Shang juan. 1900-1949 -- zhong juan. 1949-1978 -- xia juan. 1978-1999.**
650 0 **Art, Chinese □y 20th century □v Catalogs.**
650 0 **Art □z China □z Beijing □v Catalogs.**
610 24 **□□□□□ □v Catalogs.**

Successful search keys		**Indexes searched**		
□□□□□/1999	**kw:**	**ti:**	**cn=**	**su=**
20□□□□□□	**kw:**	**ti:**	**ti=**	**tiw=**
□□□□□□□□	**kw:**	**ti:**		
□□□□□□□□*	**kw:**	**ti:**	**ti=**	
20 century Chinese fine art		**kw:**	**ti:**	

Failed searches and the reasons why they failed:

ti=□□□□□□□□
Although "□□□□□□□□" is at the beginning of the subfield ‡b, it is followed by an English parallel title. Just entering ti=□□□□□□□□ in the search key does not satisfy the phrase search criteria that require entering the complete bibliographic data in the entire subfield ‡b. However, title phrase search with truncation (ti=□□□□□□□□*) will find this record.

tiw=□□□□□□□□
"□□□□□□□□" is not the title proper, and is not indexed under the title whole phrase index (tiw=).

Example 3: **OCLC # 42737927**

245 00 **□□□□□□□□□. □p □□□ = Sanwen juan / □c □□□□□ ; □□□, □□□□□□ ; □□□, □□□□.**
246 30 **□□□**
246 31 **Sanwen juan**
250 **□1□.**
260 **□□□ : □b □□□□□□□, □c 1998.**
300 **16, 9, 30, 592, 2 p. : □b ill. ; □c 21 cm.**
504 **Includes bibliographical references.**
650 0 **Chinese essays □y 20th century.**
700 1 **□□□, □d 1939-**
700 1 **□□□.**
700 1 **□□□.**
700 1 **□□□.**
700 1 **□□.**

Successful search keys	**Indexes searched**			
□□□	**kw:**	**pn:**	**pn=**	
□□□/1998		**kw:**	**ti:**	**ti=**
□□□□□□□	**kw:**	**ti:**		

□□□□□□□ **kw: ti:**
□□□□□□□□□ **kw: ti: ti=**
□□□□□□□□□* **kw: ti: ti= tiw=**

Failed searches and the reasons why they failed:

ti=□□□□□□□
In a phrase search, multiple words need to be entered in sequence, beginning with the first word in the subfield, i.e., ti=□□□□□□□□□.

tiw=□□□□□□□□□
Title whole phrase search must include all data in the title proper, including the data coded in subfield ‡p, in this case.

tiw=□□□□□□□□□□□□
Although the title proper "□□□□□□□□□□□□" has been completely entered in the search key, the search still failed due to the parallel title "Sanwen juan" following the section title "□□□". Technically, the parallel title "Sanwen juan" should have been coded in a subfield ‡b. However, when the subfield ‡b is not physically in place, the parallel title "Sanwen juan" is lumped into the subfield ‡p as part of the section title in indexing. For the title whole phrase search to work in this case, the search key needs to be: tiw=□□□□□□□□□□□□ Sanwen juan, or with truncation: tiw=□□□□□□□□□□□□* .

Example 4: OCLC # 39788674

100 1 □□□□, □d 1732-1796.
240 10 □□□□
245 10 □□□□. □b □□□□. □□□□. □□□□ / □c □□□□, □□□□□□.
246 3 □□□□. □□□□. □□□□. □□□□
260 □□ : □b □□□□, □c 1998.
300 v, 488 p. : □b ill. ; □c 22 cm.
440 0 □□□□□□□□□ ; □v 82
505 0 Iso rokujo□ / Mumu Do□jin -- To□shi sho□ / Inoue Randai -- Majirimame hanakuso gundan / Ko□kodo□ Obata-shi -- Kokon niwakasen -- Kuro ururi / Rokitsuan -- Ekyo□dai / Santo□ Kyo□den -- Inaka shibai / Manzo□tei -- Chaban hayagatten / Shikitei Sanba.
650 0 Japanese wit and humor □y Edo period, 1600-1868.
650 0 Short stories, Japanese.
700 1 □□□□, □d 1930-
700 1 □□□□, □d 1935-

This record is a typical case of a collection lacking a collective title. When collections such as this one contain four or more works, individual titles subsequent to the first title are not normally traced in title added entries. With the previous OCLC CJK software 3.11, only the first title could be retrieved via a derived search. There are now more options available in Connexion client to search for other titles following the first title in this type of record. Nevertheless, title phrase and title whole phrase searches will not automatically find all titles in the 245 field in this record. Keyword search (kw:) and title keyword search (ti:) are most effective in retrieving the records of this type, as the keyword and title keyword searches can target any individual title that appears in any position within an indexed field or subfield.

Successful search keys **Indexes searched**

□□□□ **kw: pn: pn=**
□□□□ **kw: ti: ti= tiw= ut=**
□□□□* **kw: ti: ti=**
□□□□ **kw: ti:**
□□□□ **kw: ti:**
□□□□□□□□□ **kw: ti: tiw= se: se=**

Failed searches and the reasons why they failed:

ti=□□□□

In a phrase search, all data belonging to the same subfield needs to be entered completely in the search key. In this case, the subfield ‡b actually includes three titles: □□□□. □□□□. □□□□.

ti=□□□□
ti=□□□□

Since□□□□ and□□□□are not the first titles in the subfield ‡b, they cannot be found by the title phrase searches.

Example 5: OCLC # 31938528

100 0	**□□, □d 1752-1835.**
240 10	**□□□□**
245 00	**□□□□. □□□□. □□□□. □□□□□ □c □□□□□□□□□□□□□□□□□□□□□□□□.**
260	**[Peking] : □b □□□□□□□□□□□□□□□, □c 1992.**
300	**2, 116, 64, 190, 119 p. : □b maps ; □c 27 cm.**
440 0	**□□□□□□□□□□. □p □1□**
500	**Reprint.**
500	**Each page represents 2 leaves of the original texts.**
651 0	**Tibet (China) □x History □v Sources.**
710 2	**□□□□□□□□. □b □□□□□□□□□□□□.**

Successful search keys		**Indexes searched**				
□□/1992		**kw:**	**pn:**	**pn=**		
□□□□	**kw:**	**ti:**	**ti=**	**tiw=**	**ut=**	
□□□□	**kw:**	**ti:**				
□□□□	**kw:**	**ti:**				
□□□□	**kw:**	**ti:**				
□□□□□□□□/1992		**kw:**	**cn:**	**cn=**		
□□□□□□□□*/1992		**kw:**	**ti:**	**tiw=**	**se:**	**se=**
□□ □□□□		**kw:**	**ti:**	**se:**		

Failed searches and the reasons why they failed:

ti=□□□□*

Although the three titles "□□□□. □□□□. □□□□" are considered title remainder information in cataloging, there is not a subfield ‡b physically in place preceding these three titles. As a result, "□□□□. □□□□. □□□□" are not picked up as the title remainder data by the title phrase index. Instead, they are lumped into the same subfield ‡a, and indexed as part of the title proper.

This record for a collection containing four works by different authors is exactly the same type as the one shown in Example 4. However, lacking the subfield code ‡b in the title field has skewed the subfield indexing of bibliographic data in the title (245) field, and consequently affected searching and retrieval. In theory, the three titles (□□□□. □□□□. □□□□) following the first title are treated as the remainder of the title, and should be coded in the subfield ‡b. When this subfield ‡b is not physically present where it is supposed to be, □□□□. □□□□. □□□□ are not indexed as the data belonging to the subfield ‡b; instead, they are inadvertently indexed as part of the title proper in the subfield ‡a.

Example 6: OCLC # 25554923

100 1	**□□□, □d 1866-1910.**
240 10	**Selections. □f 1988**
245 10	**□□ ; □□□ ; □□□□ ; □□ ; □□□□ ; □□□□□ ; □□□□ (□□□)**

/ □c □□□□□

250 □1□□

260 □□□ : □b □□□□□□□□, □c 1988.

300 vii, 613 p., [21] p. of plates : □b ill. ; □c 21 cm.

440 0 □□□□□□□□□

Successful search keys		Indexes searched			
□□□		kw:	pn:	pn=	
□□*		kw:	ti:	ti=	tiw=
□□□		kw:	ti:		
□□□□	kw:	ti:			
□□		kw:	ti:		
□□□□	kw:	ti:			
□□□□□		kw:	ti:		
□□□□□□□□	kw:	ti:	tiw=	se:	

This record has the same indexing problem as the one in Example 5. A subfield □b that should precede the second title (□□□) is not there; therefore, all of the individual titles in the 245 field have been lumped together and indexed in the single subfield □a. As a result, phrase searches for the first title (ti=□□and tiw=□□) will not find this record. Phrase and whole phrase searches with truncation (ti=□□* or tiw=□□*) will retrieve the record.

Attempts to use phrase searches (ti=) for any individual title other than the first title in the 245 field will fail. Keyword search (kw:) and title keyword search (ti:) will both be successful in retrieving the record via any of the titles in the 245 field.

There are more records of this type in WorldCat such as:

OCLC # 26212966 (data coded correctly in the subfield □b)
245 00 □□□. □b □□□□□□□□. □□□□□□. □□□□□□ / □c □□□□□□□□□□□.

OCLC # 39451696 (data coded correctly in the subfield □b)
130 0 □□□□.
245 00 □□□□. □b □□□□□□. □□□□□. □□□□□ / □c □□□□□ **... [et al.]** □□.

OCLC # 24746036
130 0 □□□□□□□.
245 00 □□□□□□□. □□□□. □□□□□□□. □□□□. □□□□□. □□□□□ / □c □□□□□.

OCLC # 24547857
100 1 □□□, □d 1526-1590.
240 10 □□□□□
245 00 □□□□□. □□□□□. □□□□□□. □□□□□ / □c □□□□□.

OCLC # 24746030
100 1 □□□, □d 10th cent.
240 10 □□□□
245 00 □□□□□□. □□□□□. □□□□□□. □□□□. □□□□□□□. □□□□□□□ / □c □□□□□.

OCLC # 34478081
100 1 □□□, □d 17th cent.
240 10 □□□□□
245 10 □□□□□. □□□□. □□□□□□. □□□□□□ / □c □□□□.

OCLC # 27685031
245 00 □□□□□□. □□□□□□. □□□□□□. □□□□□. □□. □□□□ / □c [□□□□□□□□, □□□□□, □□□□□].

Example 7: OCLC # 26975756

130 0 □□.
245 10 □□□□□□ / □c □□□□ ; □□□□□. □□□ / □□□ ; □□□□□□. □□□□ / □□□□ ; □□□. □□ □□□ / □□□□ ; □□□□□□.
250 □1□.
260 □□ : □b □□□□□□□ : □b □□□□□□□□□□□, □c 1991.
300 711 p. ; □c 19 cm.
440 0 □□□□□□
500 Reprint.
500 Each page represents four pages of the original.
650 0 Medicine, Chinese □x Early works to 1800.
650 12 Medicine, Chinese Traditional.
650 22 Acupuncture Therapy.
700 1 □□, □d fl. 762.

Successful search keys	Indexes searched				
□□	kw:	pn:	pn=		
□□	kw:	ti:	ti=	tiw=	ut=
□□□□□□	kw:	ti:	ti=	tiw=	
□□□	kw:				
□□□□	kw:				
□□□□□	kw:				
□□□□□□	kw:	ti:	tiw=	se:	se=

In this case (Example 7), the failed attempts for title keyword search (ti:), title phrase search (ti=) and title whole phrase search (tiw=) to retrieve the record via the second title□□□, third title□□□□and fourth title□□□□□ show the drawbacks of title keyword index (ti:) and title phrase indexes (ti= and tiw=). Because these indexes are strictly subfield-based indexes, the placement of subfield codes determines how data is indexed. In all cataloging records, the subfield ‡c is the last coded subfield in the title (245) field, and there are not any other subfield codes applied following the subfield ‡c.

In this record (OCLC # 26975756) the placement of the last three titles in the 245 field is after the subfield ‡c, which means that these last three titles do not belong to any indexed subfields for indexing purpose. This is why they do not get indexed for the title keyword search (ti:), title phrase search (ti=) and title whole phrase search (tiw=).

This case also indicates that although title keyword search (ti:) is more flexible than phrase searches, it cannot always find a term in a field or subfield. The placement of the term in a field or subfield does make a difference in indexing and retrieval. It is more accurate to state that title keyword index (ti:) can find a term regardless of its placement in a "properly indexed field or subfield." The only search that will find a term regardless of its placement in a field or subfield is the keyword search (kw:). As demonstrated in Example 7, title keyword search (ti:) is unable to find the titles placed beyond the subfield ‡c in the title (245) field; however, keyword search (kw:) succeeded in this regard. In WorldCat, there are more records of this type such as:

OCLC # 34478725
100 1 □□□, □d 1667-1744.
240 10 Poems
245 10 □□□□□□□□c [□□□□□□□]. □□□□□□□[□□□□□□]. □□□□□□□□[□□□□□]. □□□□□□[□□□□□]. □□□□□[□□□□□] ; [□□□ ... et al. □□].

OCLC # 30519210
100 1 □□□, □d 17th cent.
240 10 □□□□

245 10 □□□□□□ / □c [□□□□□□]. □□□□□□ / [□□□□□□]. □□□□□□ / [□□□□]. □□□□□□ / [□□□□].

OCLC # 30519211
100 1 □□□, □d 16th cent.
240 10 □□□
245 10 □□□□□ / □c [□□□□ ; □□□□]. □□□□□□ / [□□□□□]. □□□□□□ / [□□□, □□□□]. □□□□ / [□□□□ ; □□□□]. □□□□□□ / [□□□□ ; □□□].

4. Spacing in vernacular word search keys

Normally in word and phrase searches, vernacular words and phrases are entered in sequence without space between words. Vernacular words and phrases being entered in search keys are adjacent words that are grammatically meaningful. On a number of occasions, Ms. Hisako Kotaka, the Senior Product Manager, Cataloging Products & Services Division of OCLC, advised the OCLC CJK member libraries about the issue of spacing in vernacular keyword searching:

"In the new WorldCat CJK data indexing structure, each single CJK character is treated as a word. (In the internal system data setting, each CJK character in the CJK text is actually separated by a space but such treatment is invisible to the client users externally). The multiple CJK characters are programmed to hold the auto-adjacency to the left for client CJK users. Because of the CJK text 'continuous' writing as one of distinctive characteristics, the use of visible spaces in-between 'words' is inapplicable and improper for client CJK users. Under such CJK-unique data handling condition -- the compounded CJK words is making a phrase in a real sense but are indexed as keyword (ti:, cn:, se:, pb:, etc.) -- client CJK keyword search with non-adjacent word(s) requires a space, that is 'a forced space,' to break the auto-adjacency in CJK 'words.'"
— Hisako Kotaka

This unique feature of spacing between vernacular characters is very useful in keyword searching. It gives one the flexibility to select his/her own "non-adjacent keywords." Basically, if one wants to search for non-adjacent vernacular words in records, he/she can just insert one space between the vernacular words in the search key. It needs to be clarified that spacing between vernacular words applies only to vernacular keyword searches; spacing cannot be applied to vernacular phrase searches and browsing. Here are two examples to demonstrate how this feature works in vernacular keyword searches.

OCLC # 46716629 (the same record as shown in Example 2)

245 10 **20□□□□□□□ : □b □□□□□□□□□ = 20 century Chinese fine art : collection from China National Museum of Fine Arts.**

The search key: ti:20 □□ fine art

In this title keyword search, a space is inserted between 20 and□□, and another space is inserted between □□ and fine art. This is also a numeric/roman/vernacular combined search that retrieves the record in a direct hit.

OCLC # 31126093

245 00 **□□□□□□□□□, 1949-1989. □p □□□□□□□□□□ = □b A Contending series of social science, 1949-1989 / □c □□□□□□□□□□□□□□ □ ; □□□□□□□**
250 **□1□.**
260 **□□: □b □□□□□□□□: □b □□□□□□□□□□□□□, □c 1993.**
300 **6, 7, 927 p. ; □c 21 cm**
504 **Includes bibliographical references.**
650 0 **Arts, Chinese.**
650 0 **Criticism.**
700 1 □□□.
710 2 □□□□□□□□□□□□□□.
740 0 **Contending series of social science, 1949-1989.**

It may not be efficient to use phrase searches to try to find this record, because the title contains multiple subfields, a range of years, and an English parallel title. Word search with spacing works very well in this case (ti:□□□□ □ □ □).

5. What we learned from searching the sample records

The sample records used in this study helped us better understand the features and differences of frequently used indexes in Connexion client, including their strengths and limitations, and how each index works.

Title phrase and title whole phrase searches (ti= and tiw=) are rather inflexible and "demanding" searches that generally require more data in a search query, but the retrieval results are not necessarily great. Phrase searches are often less efficient and have a greater chance for failure than keyword searches. Almost anything that phrase searches are capable of doing can be accomplished as well by keyword searches. However, phrase searches are incapable of doing everything that keyword searches can.

Four out of the seven sample records (Examples 3 and 5-7) that we have discussed show that how bibliographic data is coded in a field or subfield in records could affect searching and retrieval in some cases. Because the subfield code ‡b is not physically present in the title (245) field in Example 5, the same title phrase search (ti=) that is successful in retrieving the record in Example 4 failed to work for the exact same type of record in Example 5. Using keyword search (kw:) is the best way to minimize the adverse impact of such uncontrollable coding factors in indexing, searching and retrieval.

Various subfield codes in headings or fields play a crucial role in title phrase indexing (ti=), title whole phrase indexing (tiw=), and title keyword indexing (ti:). Computer programs do not have the ability to intelligently distinguish a main title, section title, subtitle, or a parallel title in the title (245) field. It is the subfield codes in headings or fields that serve as the markers to identify data belonging to each subfield so that the data can be properly indexed. Because subfield codes guide phrase and word indexing, whether or not appropriate subfield codes are physically in the right place in headings or fields directly affects how the data is indexed, and consequently has the effect to make a search succeed or fail.

Data coding in bibliographic records is sometimes a factor over which an index searcher has no control. However, one does have the option to select the most efficient search strategies and methods that work the best for a given situation. Although the indexes in Connexion client will work the way they are designed to work, human factors in shaping the process of searching and retrieval should not be underestimated. Connexion client does provide an index searcher with options to improve the quality of retrieval. Applications of qualifiers, use of Boolean operators, combining Latin and vernacular data in the same search query, spacing in word searches, and vernacular browsing are some of the strategies to achieve more satisfactory retrieval results.

6. Other issues to be aware of in vernacular searching

■ Center dot (•) in titles and corporate names affect browsing and phrase searches

Vernacular punctuations that are common in CJK languages records generally do not affect search and retrieval. There is no need to enter such vernacular punctuations in search keys. However, there is an exception with the vernacular center dot (•), its presence in titles and corporate names interferes with browsing and phrase searches, causing problems in retrieval.

If a vernacular title or a corporate name heading in a bibliographic record contains a center dot (•) that was entered in the record as a vernacular punctuation, browsing (scan indexes) will not find the targeted record regardless of whether or not the center dot (•) is entered in the scan search keys. In other words, browsing fails entirely when a title, or a corporate name (in a bibliographic record) being searched for contains one or more center dots. We hope that this problem will be eventually corrected in Connexion client, as the center dot is a frequently used vernacular punctuation in CJK records. Here are some of the records from WorldCat:

□□ • □□□□□

□□ • □□ • □□□
□□ • □□□□□□□
□□□ • □□□□□□□
□□□ • □□□ = □b **Fight in northern Shaanix [sic]**
□□ • □□ • □□ : □b □□□□□□□□□□□□□ = **Chuantong xianzhuang weilai**
□□□□ • □□□□
□□ • □□ • □□□□□
□□□□ • □□□ • □□□□
□□□□□ • □□ • □□
□□ • □□□□□□□□□ : □b □□□□□□□
□□ • □□□□□□ : □b □□□ • □□□ • □□□□□□□
□□□100□ : □b □□ • □ □ • □□ • □□ • □□□□ • □□□□
□□□□□□□□□□ • □□□ : □b □□ • □□□□□□□□□□□□

Phrase searches perform better in finding records with titles or headings containing a center dot (•). For instance, if a vernacular title in a bibliographic record contains a center dot (•), and it is faithfully entered in the search key, the record will be found. However, if the center dot (•) is left out in the search key, the targeted record will not be found.

Keyword search is the best and highly reliable method for finding records with access points containing center dots. There is no need to enter the center dot in keyword search keys for successful retrieval, and the center dot has no effect on word searches.

In the following two examples, different searching methods have been experimented with to find targeted records with access points containing center dots, and the search results are recorded for comparison.

OCLC # 47265727
100 1 □□, □**d 1935-**
245 10 □□□□ • □□.
250 □□.
260 □□□□□□ : □**b** □□□□□, □**c 2000.**
300 **40, 179 p. :** □**b ill. ;** □**c 22 cm.**
Title scan failed to retrieve this record because the vernacular title contains a center dot (•). Phrase searches will find the record if the center dot is entered in the search key. Also, when a title or heading contains a center dot, applying truncation (*) in phrase searches will get an error message. Fortunately, both keyword search (kw:) and title keyword search (ti:) are successful in finding the record, disregarding the center dot.

Successful searches to retrieve the record OCLC # 47265727:

□□□□ **(omitting the index label kw:)**
□□□□□□ **(omitting the index label kw:)**
□□ □□**/2000** **(a space between** □□**and** □□**)**
Failed browse and phrase searches:

sca tiw=□□□□ • □□ **no records found for browsing**
sca tiw=□□□□□□ **no records found for browsing**
sca tiw=□□□□ **Retrieved a different record OCLC # 53378495**
ti=□□□□ **Retrieved a different record OCLC # 53378495**
ti=□□□□□□ **no records found when the center dot (•) is left out**

OCLC # 37198153
110 2 □□□□□□□□□□□□□.
245 10 □□□□□□□□□□□□□□□□□, □□□□□□□□□□.
250 □1□.
260 □□ : □**b** □□□□□□□□□□□□□ : □**b** □□□□□□□, □**c 1995.**
300 **1, 1, 392 p. ;** □**c 27 cm.**
500 **Includes indexes.**

610 24 □□□□□□□□□□□□□ □v Catalogs.
650 0 Catalogs, Publishers' □z China.
651 0 China □v Imprints □v Catalogs.

□□ • □□ • □□□□□□is one of the leading publishers in China, and the corporate name appears in hundreds of records in WorldCat. In this particular record, the publisher is both an author and a subject, and its name appears in multiple indexed fields and headings, including 110, 245, 260 and 610 fields. Because the corporate name□□ • □□ • □□□□□□contains center dots, browsing failed to retrieve the targeted record under all phrase and whole phrase indexes (ti=, tiw=, cn=, pb=, ncw=, suw=, etc.). Browse searches can find the records in which the corporate name does not contain the vernacular center dot. Keyword searches successfully find the targeted record.

Some of the searches retrieving the record OCLC # 37198153 in a direct hit:

□□□□□□□□	**(omitting the index label kw:)**
□□□□□□/1995	**(omitting the index label kw:)**
ti:□□□□/1995	
su:□□□□/1995	

Failed browse and phrase searches:

sca tiw=□□ • □□ • □□□□□□	**getting an error message**
sca su=□□□□□□□□□□	**no records found for browsing**
sca cn=□□ • □□ • □□□□□□	**getting an error message**
sca cn=□□□□□□□□□□	**Found more than 20 other records in which the corporate name does not contain the vernacular center dot.**
ti=□□ • □□ • □□□□□□*	**getting an error message**
cn=□□□□□□□□□□/1995	**no records found**
cn=□□□□□□□□□□*	**found 26 other records in which the corporate name does not contain the vernacular center dot.**

■ Word sequence matters and affects retrieval

When vernacular words are entered in search keys, Connexion client compares the search query word by word in sequence against the index being specified in the search, looking for the exact match. The sequence of words entered in the search key determines what will be returned as search results. For example, the searches ti: □□□□and ti:□□□□ produce very different retrieval results. Likewise, ti:□□□□ and ti:□□□□ are not equivalent searches. In these situations, a combined search would produce a more comprehensive retrieval:

ti:□□□□ or ti:□□□□
ti:□□□□ or ti:□□□□

■ Non-Latin script records may not contain transliteration in parallel romanized fields
In Connexion client, non-Latin script records are allowed to have vernacular data only, or romanized data only. This implementation departs from the convention that non-Latin script records must contain bibliographic data in paired vernacular and romanized fields. Although the practice of vernacular data only in records is not encouraged, non-Latin script records without Latin script transliteration in parallel romanized fields can technically enter the WorldCat database. Such records literally do not have roman access points, thus cannot be found by Latin searches. When a roman or a vernacular search fails to find a record, it may not necessarily mean that the record does not exist in WorldCat; additional searches may be necessary.

Concluding remarks

The Connexion client is a superior cataloging platform for non-Latin script users, offering a wide range of choices and options that were not ever available before. The system's remarkable capabilities in handling non-

Latin scripts data and providing greatly enhanced access to vernacular records in WorldCat have brought the processing of non-Latin script materials and resource sharing to a new level.

Vernacular searching is no longer a severely limited access mechanism compared to romanized search methods. In fact, vernacular searching offers unique advantages to remedy some of the intrinsic drawbacks in romanized search methods. For example, word division is constantly an issue in romanized searches, and it can often impact the search results. With vernacular searching, word division is no longer an issue. For CJK languages, homophones (i.e., many different characters pronounce and romanize exactly the same way) can often affect the precision of retrieval. Vernacular searching can sufficiently minimize the adverse effect of homophones on retrieval.

Roman and vernacular search methods are complementary in providing access to materials in the WorldCat database. Since it is the vernacular data that conveys primary information for bibliographic identification and retrieval, the access to such data and materials is important to both communities of resource users and resource access providers connected via the bibliographic utility. With the successful implementation of Connexion client that continues to evolve, OCLC has made great strides in bridging the gap between multi-lingual, multi-script resources and their access in WorldCat.

References

Technical Bulletin 251: Connexion WorldCat Searching
http://www.oclc.org/support/documentation/worldcat/tb/251/

Connexion Client System Guide: Search WorldCat
http://www.oclc.org/support/documentation/connexion/client/cataloging/searchworldcat/

Connexion client non-Latin script cataloging tutorial
http://www5.oclc.org/downloads/tutorials/connexion/client/nonlatin.html

Gabel, Linda (quoting Robert Bremer). Title phrase and Title whole phrase scanning. E-mail to OCLC-CAT list, March 29, 2005.

Kotaka, Hisako. Connexion CJK data indexing issues. E-mail to OCLC-CJK list, May 24, 2005.

Whitehair, David. Multiple scripts in OCLC WorldCat.
http://www.ifla-stockholm2005.se/pdf/Whitehair%20Multi%C9LC%20WorldCat.pdf