書誌データベースの異体字処理
― 谷と穀は同じ字か ―
Processing Japanese/Traditional/Simplified kanji on Japanese Bibliographic Databases

長　坂　和　茂*

抄　　録

本研究では、CiNii Books、NDL-OPAC、国立国会図書館サーチの3つの書誌データベースについて、それぞれのデータベースが異体字を検索する際にどのような範囲で同じ文字として処理しているのか、調査を行った。結果、それぞれのデータベースにより、挙動が大きく異なることが確認できた。異体字を含む可能性がある文字を含むキーワードで検索する際には、データベースの特性を理解し、検索漏れに注意する必要がある。

目　　次

1. はじめに
　1.1　異体字と繁体字・簡体字の歴史
　1.2　問題意識
　1.3　文字コード
　1.4　本稿における異体字の定義
2. 先行研究と本研究の意義
　2.1　先行研究
　2.2　本研究の意義
3. 調査方法
　3.1　対象とするデータベース
　3.2　文字サンプル
　3.3　調査手法
4. 仮説
5. 調査結果
6. 考察
7. NDLサーチ再調査
　7.1　方法
　7.2　再調査　結果
　7.3　再調査　考察
8. まとめ
　8.1　CiNii Books
　8.2　NDL-OPAC
　8.3　NDLサーチ

○1. はじめに○

1.1　異体字と繁体字・簡体字の歴史

一般によく知られている通り、漢字には同じ意味を持つ文字でも、字体の異なる「異体字」が存在する。特に、現代台湾及び戦前の日本や中国で用いられた繁体字（日本では「旧字体」、台湾では「正體字」と呼ばれる）や、現在中国大陸で用いられている簡体字（簡化字）は、各国で過去や現在日常的に用いられていることもあり、日本の大学図書館に所蔵されている資料の書名などにも、もちろん広く用いられている。

戦前までは、日本でも中国大陸でも、ほぼ同じ文字を使用していた。しかし、戦後になって、日本と中国すなわち中華人民共和国はそれぞれ個別に漢字を簡略化する政策を取った。これにより、日本においては「新字体」、中国においては「簡化字（簡化字）」という簡略化された文字が生まれ、それまでの文字をそれぞれ「旧字体」「繁体字」と呼ぶようになった。一方、台湾には、中国大陸で共産党との内戦に敗れた国民党（中華民国）が逃れていた。台湾は日本や中国のような文字の簡略化を行わなかったため、以前の文字がそのまま現代まで使われている。これが台湾でいう「正體字（正体字）」であり、故に日本から見た旧字体、中国からみた繁体字とはほぼ同じ字体を指す。

*ながさか　かずしげ　京都大学　工学研究科　桂化学系図書室　2014年4月1日受理

本稿では特に区別するべき場合を除いて、便宜的に、現代日本で新字体等と呼ばれる文字を「日本漢字」、現代中国で用いられる簡化字を「簡体字」。過去の日本及び中国大陸や現代台湾で用いられる正體字＝旧字体を「繁体字」と呼称することとする。

1.2　問題意識

このことは図書館においても検索上の問題を発生させる。例えば、戦前に書かれた図書がその一例である。『圖書館雜誌』を検索するにあたって「図書館雑誌」というキーワードで検索できなければ不都合であるし、かといって検索キーワードとして「圖書館雜誌」と入力するのは、現在日本で一般的に使われているIMEでは非常に困難である。

また、同様に中国由来の古典籍やその現代版の図書を検索する場合にも問題となる。例えば内容は全く同じ「春秋穀梁伝」であっても、現代の中国大陸で発行された図書であれば「春秋谷梁传」というタイトルとなる可能性が高いし、台湾で発行された図書であれば「春秋穀梁傳」というタイトルとなると考えられる。これらは内容が同じで書名も同じであり、もし「春秋穀梁伝」で検索できないとしたら、図書館員はこの図書を検索する際には「春秋穀梁伝」「春秋谷梁传」「春秋穀梁傳」の3つのキーワードをわざわざOR検索しなければならない。

1.3　文字コード

漢字に限らず、文字をコンピュータ上で扱うためにはコードに置き換える必要がある。コンピュータは原理的に0か1しか扱うことができないためである。この文字コードの設定は基本的に各国ごとに行われる。代表的な各国の文字コードとして、日本ではJIS X 0208など、中国ではGB 2312など、台湾ではBig5などが挙げられる。

その後、文字コードの世界的な統合が図られ、Unicodeが登場した。2014年現在では各種書誌データベースも多くがUnicodeやそれに準拠する文字コードを採用している。

たとえば、CiNii Booksのデータの基になるNACSIS-CATでは、2000年1月にUCS（国際符号化文字集合）にシステムを対応させ、UCSに合わせて既存のレコードの文字コードを変換した[1)]。

一方NDL-OPACがUnicodeに対応したのは2012年1月のリプレイスの際である。しかし、国立国会図書館では1986年から2011年まで、アジア言語OPACを運用しており、こちらはアジアの諸言語に対応する必要があり、Unicodeを採用していた。現在のNDL-OPACは使用する文字コードをUnicodeにした上で、旧NDL-OPACのデータとアジア言語OPACのデータを統合している。そのため、旧NDL-OPACに登録していたデータは、アジア言語での原綴が入力されておらず、翻字タイトルでしか検索できないレコードが存在する。しかし、中国語資料に限れば1986年以降受入の図書・雑誌は全てアジア言語OPACに入力されていたため、Unicodeに対応し、簡体字・繁体字を区別したレコードが作成されている[2)]。

国立国会図書館サーチ（以下、NDLサーチと呼ぶ）はもとになるデータが多岐にわたり、それぞれのデータ元となるデータベースが使用している文字コードを一概に示すことができないが、NDLサーチのシステムそのものとしては、Unicodeを採用しており、日本漢字・簡体字・繁体字の区別をしている。

1.4　本稿における異体字の定義

そこで、各種データベースにおいて、Unicodeを採用していることから、本研究の前提となる「異体字」については本稿においては、Unicodeを基準として定義する。

すなわち、Unicode上で別のコードが与えられているのであれば別の文字、同じコードが与えられているのであれば同じ文字として扱う。データベース内部では、たとえフォントや字形が違っていたとしても文字コードが同じであれば、それを同じものとして扱うし、たとえ人間が見たときにどれだけ字形が似ていたとしても、文字コードが

違えばデータベースは別の文字として扱うためである。このことから、データベースの挙動を確認する本研究では、Unicodeを基準とすることが自然である。

以上の定義により、単にフォントの違いに由来する字形の違い、例えばしめすへんの形が「ネ」か「示」かという違いや、しんにょうの点が1つか2つかといった違いは、Unicodeで同じコードが与えられている限り、異体字としては扱わないこととなる。

○2. 先行研究と本研究の意義○

2.1　先行研究

書誌データベースにおける文字コードや異体字などの問題や処理方法については、過去に多数の発表・報告・研究がなされている。

宮澤[3]は、書誌データベースの中でもNACSIS-CATを例に挙げて、使用する文字セット、包摂規準の適用などについて詳説し、運用にあたって、あいまい性、二重性、不完全性などの問題点があることを指摘する。あいまい性、二重性については、「たとえば宮澤の「澤」と、宮沢の「沢」の違いは同じ字の字体の異なりか、異なる字かという問題にはっきり答えを出してくれる典拠はなく」[4]と、ある2つの文字が、同じ文字か違う文字かという問題に結論を出すことの難しさを指摘している。

一方で、書誌データベースにおける技術的対応について、Miyazawa [5]は原規格分離原則（source separation rule）によって、Unicodeで別の文字として扱われている文字の扱いや、簡体字との関連も含めて説明している。また、米澤[6]はNACSIS-CATにおける漢字の包摂、漢字統合インデクスの仕様について簡潔に解説している。

国立国会図書館関西館アジア情報課[7]や水流添[8]は、NDL-OPACのリプレイスに伴う、旧NDL-OPACとアジア言語OPACの統合に伴う書誌レコードの記述方法の違いという、NDL-OPACに特有の状況について踏まえたうえで、NDL-OPACやNDLサーチにおける簡体字・繁体字の扱いについて、解説している。

2.2　本研究の意義

各種書誌データベースが異体字をどのように扱うか、という点について論説や解説した研究や発表は、以上の通り、当該データベースを提供する機関やその所属者が、そのデータベースについて説明するものが多い。そのため、データベースの検索の仕組みについては詳しいが、実際に検索をした時に発生する検索漏れやノイズなどについて、検索者の立場から見たものは少ない。データベースを使って検索する図書館員の立場から、これらの「公式見解」を検証し、各データベースを比較することでそれぞれの特性を指摘することが、本研究の目的である。

○3. 調査方法○

3.1　対象とするデータベース

調査対象としては、CiNii Books、NDL-OPAC、NDLサーチの3種類のデータベースを選択した。これらのデータベースを選んだ理由は以下の3つである。

1つ目はそれぞれ大学図書館界、国立国会図書館、国立国会図書館及び各種公共図書館など、非常に広い範囲の図書を収録する書誌データベースであることである。その一方で利用者は主として日本人（日本漢字を使用する人たち）であり、検索に使用する言語も日本語である可能性が高いと思われる。ゆえに、日本漢字、繁体字、簡体字などの多様な文字で書かれた図書を、日本漢字をキーワードとして検索することが多いと考えられる。

2つ目は、デザインがGoogle風であると言われるためである[9]。デザインがGoogle風であれば、利用者は当然検索システムもGoogle風であると判断するであろう。Google登場以後のwebの世界では、検索キーワードは適当なものを入力すれば、検索システム内部で適当に処理をして、適当な結

果が返ってくると期待してしまう傾向にある。このため、逆に検索をして資料が見つからなかった場合には、検索キーワードが悪かったのではなく、そもそも探しているものが存在しなかったのだと判断されてしまう危険性が高い。検索漏れを引き起こさないようするために、これら漢字の扱いについて知ることが助けになると考えられる。

3つ目は、これらが非常に新しい書誌データベースであり、今後学術情報を検索するうえで必須となる、息の長いデータベースになる可能性が高いと考えられるためである。ゆえに、これらのデータベースを今の段階で調査しておくことは、今後書誌データベースを利用していく状況において、有用であると考えられる。

3.2　文字サンプル

今回の調査に用いる文字群は表1の通りである。これらの漢字は今回の調査のために適当に選んだ。漢字は大きくA、B、C、D、Eの5パターンに分けることができる。パターンAは日本漢字、簡体字、繁体字の全てが別の字形である文字群で、図・**图**・圖と呉・吴・㠯を用意した。パターンBは日本漢字と繁体字が同じだが、簡体字が異なる文字群で書・**书**と館・**馆**。パターンCは簡体字と繁体字が同じだが､日本漢字が異なる文字群で壱･壹。パターンDは日本漢字と簡体字が同じで、繁体字が異なる文字群で、内・內と学・學。内と內の違いは見た目には分かりにくいが、内部が「人」か「入」の違いであり、JIS X 0208では包摂されて同じコードが与えられているが、Unicode上ではコードが異なる[10)]。パターンEは日本漢字と繁体字が同じ文字という意味ではパターンBと同じだが、簡体字に当てられている文字が別の文字として日本漢字や繁体字で使われている場合である。例えば「後」を簡体字で書くと「后」となるが、后は日本では全く別の文字である。このような現象を「別字衝突」というが、簡体字が別の日本漢字に対して別字衝突を起こしている例がパターンEである。

表1　使用する文字サンプル

パターン	日本漢字	簡体字	繁体字
A	図	**图**	圖
A	呉	吴	㠯
B	書	**书**	書
B	館	**馆**	館
C	壱	壹	壹
D	内	内	內
D	学	学	學
E	後	后	後
E	穀	谷	穀
E	機	机	機
E	葉	叶	葉

3.3　調査手法

調査方法としては、Aという文字とBという文字がある場合、AとBが同じ文字として処理されているかどうか判断するためにNOT検索を用いた。A NOT Bでの検索結果とB NOT Aでの検索結果の両方が0である場合に、AとBは同じ文字として扱われている、と判断することとした。もちろん、たまたまそれらに該当するレコードが存在しない可能性も全く無いわけではなく、必ずしも厳密な調査方法とは言えないが、実際には漢字一文字で検索すれば万単位のレコードが検索できる。そのすべてが「たまたま」AでもBでも検索できるようなレコードだったとは考え難く、実際には大きな問題にはならないと考えられる。

具体的には例えば「図」と「**图**」が同じ文字として扱われているか調査する場合であれば検索キーワードは「図 NOT **图**」「**图** NOT 図」の2種で検索する。もし両方の検索結果が0件であれば、同じ文字として扱われていると考え、どちらか一方でも0でなければ異なる文字として扱われる、と判断する。

ただし、NDLサーチはNOT検索に対応していなかったため、検索結果の件数を比較することで、同じ件数のレコードがヒットしていれば、同じ文字として扱っていると判断することとした。

実験はいずれも2014年3月31日に行った。

○4. 仮説○

調査を行う前に、各データベースの挙動について、仮説を立てた。

CiNii Booksについては、NACSIS-Webcatの後継であることや、NACSIS-CATがシステム上NII漢字統合インデクス[11]を使用していることから、これに従うものと予想される。漢字統合インデクスは非常に広い範囲の漢字を同じ文字とみなす特徴があるため、今回提示した漢字群は、A、B、C、D、E全てのパターンにおいて、同じ文字として扱うだろうと考えられる。

NDL-OPACについては、国立国会図書館のウェブサイトにて「漢字は日本漢字、簡体字、繁体字のいずれでも検索可能です」「ただし、複数の正体字が同じ簡体字に置き換わる場合は文字が正規化されておらず、別の字として認識されます」[12]とあることから、A、B、C、Dでは同じ文字として扱われるが、Eでは別の文字として扱うと予想される。

NDLサーチでは、「日本漢字、簡体字、繁体字は別字として認識されます」[13]とあり、ABCDE全てにおいて別の文字は別の文字として扱われるだろうと予想される。

○5. 調査結果○

CiNii Booksについては、仮説通りの結果が得られた。漢字統合インデクスの範囲の広さを裏付ける結果となった。

NDL-OPACでは、「穀（谷）」の事例で仮説からの乖離が見られた。パターンEは「穀・谷」以外は全て異なる文字として判断するにもかかわらず、「穀・谷」だけが、同じ文字として扱われている。

NDLサーチでは、パターンAの「図・**㊀**・圖」とパターンD「学・學」の例で、日本漢字と繁体字が同じ文字として扱われており、予想と乖離した。一方、同じパターンA・Dでも「呉・吳・吳」と「内・內」 では仮説通り、呉と吳、内と內は別の文字として扱っていた。NDLサーチの場合、検索結果の件数が同じであっても適合度順で並べた際の順番は異なっており、全く同じ文字として扱っているわけではないと考えられる。しかし、今回はあくまで件数のみを判断基準としたため、適合度順が異なる場合も、件数が同じであれば、同じ文字として扱っていると判断した。

以上の結果をまとめたのが表2である。比較した2つの文字が、同じ文字として扱われていれば「同」、違う文字として扱われていれば「異」と表記している。

表2　3種のデータベースにおける異体字の異同

			CiNii Books		NDL-OPAC		NDLサーチ	
パターン	漢字	漢字	仮説	結果	仮説	結果	仮説	結果
A	図	**㊀**	同	同	同	同	異	異
	㊀	圖	同	同	同	同	異	異
	圖	図	同	同	同	同	異	同
A	呉	吳	同	同	同	同	異	異
	吳	呉	同	同	同	同	異	異
	呉	呉	同	同	同	同	異	異
B	書	**书**	同	同	同	同	異	異
B	館	**馆**	同	同	同	同	異	異
C	壱	壹	同	同	同	同	異	異
D	内	內	同	同	同	同	異	異
D	学	學	同	同	同	同	異	同
E	後	后	同	同	異	異	異	異
E	穀	谷	同	同	異	同	異	異
E	機	机	同	同	異	異	異	異
E	葉	叶	同	同	異	異	異	異

○6. 考察○

CiNii Booksでは、ほぼすべての異体字を同じ文字として扱っていることが分かった。これは、漢字統合インデクスに由来するものと考えられる。漢字統合インデクスはこのような異体字のほかに字形が似ている文字（たとえば書と晝（昼の旧字体））を同じ文字として扱っていることから、今回用意した程度の異体字では全て吸収されることは仮説の時点で予想済みである。逆に言えば、CiNii

Booksではこれら異体字を別の文字として検索することはできないので、例えば日本語で書かれたもののみを検索したい場合にはタイトル言語コードや本文言語コードなどを併用して検索結果を絞り込むことが必要になることもあると考えられる。

NDL-OPACにおいては、ほぼ仮説の通りであったが、穀と谷の例だけNDLによるアナウンスと異なる結果となった。この事例について調査するため、鬱（郁）、醜（丑）、雲（云）、乾・幹（干）などグループEと同種の文字を検索してみたが、全て別の文字として扱うことが分かり、「穀・谷」の結果の特異性が際立つ結果となった。ともあれ、結果がほぼ仮説の通りであったことから、基本的にはCiNii Booksと同じく、異体字による検索結果の違いは吸収されている。簡体字が別字衝突を起こしている事例に限って注意が必要であるといえよう。

NDLサーチの事例は今回こちらで用意したパターンの規則性とは全く異なる挙動を示した。パターンAでは、図と圖を同じ文字として扱うのに対して、呉と吳は異なる文字として扱われる。パターンCにおいては、学と學が同じ文字として扱われるのに対して、内と內は異なる文字として扱われる。つまり、今回用意した、「日本漢字と簡体字と繁体字とがそれぞれ同じ字か否か」という枠組みではない部分で区別されている可能性が高いと考えられる。では、圖と學というグループと、吳と內というグループを分ける基準は何であろうか。

この問いに対して直接答えた論文ではないが、富田[14] は、2011年当時の近代デジタルライブラリーの検索システムについて「検索時の異体字の関連づけは第1 ～第4水準内に限られている」と指摘している。近代デジタルライブラリーは国立国会図書館が運営しているサイトであり、検索システムも同様のものを使用していると考えれば、NDLサーチにおいても同様の基準を使用している可能性がある。そこで、JIS X 0208において、それぞれの漢字がどのようなコードを振られているか調べた。JIS X 0208は一般的にJIS第1水準、第2水準と呼ばれている文字の集合である。

その結果、「圖」は5426、「學」は555C、というコードが振られているのに対して、吳と內はどちらも範囲外であり、それぞれ呉と内に包摂されている文字であることが分かった。JIS X 0208を拡張した文字セットであるJIS X 0213でも同様であり、第3、第4水準からも外れる。

このことから、「NDLサーチにおいては、JIS第4水準までの繁体字は日本漢字と同じ文字として扱っている」という仮説を立てることができる。以下、この仮説を検証する。

○7. NDLサーチ再調査○

7.1　方法

この仮説を検証するため、まず、JIS第1水準から第4水準までの文字の中から、繁体字と日本漢字の関係にある文字を選択した。具体的には、国（第1水準）と圀（第2水準）、朗（第1水準）と朗（第2水準）、銭（第1水準）と錢（第3水準）、虚（第1水準）と虚（第3水準）、猫（第1水準）と貓（第4水準）、隲（第2水準）と騭（第3水準）の合計6セットの文字の組み合わせを用いた。

手法としては、先ほどと同様これらの文字をNDLサーチで検索し、そのヒット数を比較した。ヒット数が同じであれば、同じ文字として扱っていると推測している。また、先の調査で分かったことだが、検索キーワードの文字が同じとみなされる場合、ヒット件数が同じであっても、検索の適合度順は異なる。この点については先ほど同様、適合度順が異なっていたとしても、件数が同じであれば同じ文字として扱っていると考えることとした。

7.2　再調査　結果

調査の結果をまとめたのが表3である、NDLサー

チでは、第1水準と第2水準の範囲にあるものだけが同じ文字として扱われ、第3水準、第4水準のものは全て別の文字として扱われていた。具体的には、第1水準と第2水準の組み合わせである「国と圀」及び「朗と朗」の組み合わせのみが同じ文字扱いで、それ以外は全て、別の文字として扱っていた。

表3　NDLサーチにおける文字の異同

パターン	漢字	漢字	結果
第1水準と第2水準	国	圀	同
	朗	朗	同
第1水準と第3水準	銭	錢	異
	虚	虚	異
第1水準と第4水準	猫	貓	異
第2水準と第3水準	隲	騭	異

7.3　再調査　考察

調査結果から、先に立てた「NDLサーチにおいては、JIS第4水準までの繁体字は日本漢字と同じ文字として扱っている」という仮説が誤っていることが確認できた。

NDLサーチにおいて、異なる文字が検索上同じ文字として扱われるのは、JISの第1水準、第2水準の範囲に限られる。第1水準、第2水準を定義する文字コードはJIS X 0208であるので、NDLサーチは、JIS X 0208の範囲内で、繁体字すなわち旧字体を同じ文字として扱っていると言える。

NDLサーチは異体字に対して、「日本漢字、簡体字、繁体字は別字として認識される」と公式にアナウンスをしている。このアナウンスと本調査結果を矛盾なく解釈するには、ここでいう「日本漢字」は、繁体字として認識可能な文字であっても、JIS X 0208に含まれる文字であれば「日本漢字」に含むと考えることができる。

○8. まとめ○

書誌データベースとして似たような扱いを受ける3種のデータベースだが、どのような異体字を同じ文字として扱うかは、大きく異なることが確認できた。

8.1　CiNii Books

これら3つのデータベース中で最も広い異体字を「同じ文字」として扱うデータベースは、CiNii Booksである。CiNii Booksは、NII漢字統合インデクスを使用して、検索の際には書誌レコードの文字と検索語の文字の両方を変換し、内部的に同じ文字として扱っている。そのため、漢字統合インデクスが同じ文字として扱っている文字は異体字でなくても、同じ文字として検索されることがある。具体的には、書と晝（昼）が同じ文字として扱われているという事態も発生している[15)]。ただし、図書館と検索したくて「図昼館」というキーワードで検索する人は、検証目的の人を除いていないであろうから、ノイズはほとんど発生せず、問題はあまりないとも考えられる。書と晝（昼）が同じ文字として扱われている一方で、同様に文字の形が似ている畫（画）が両者と別の文字扱いになっているのは、おそらく、書と画を同じものとして扱ってしまえばノイズが発生すると考えたためであろうと推測できる。例えば「図書」と検索して「図画」という文字列を持つレコードがヒットすれば、大規模なノイズとなることは容易に予想できる。

8.2　NDL-OPAC

続いて、NDL-OPACも、CiNii Booksに次いで、異体字を同じ文字として扱う範囲が広い。NDL-OPACは、簡体字と繁体字が別字衝突を起こしている際に限って、日本漢字・繁体字と簡体字を異なる文字として扱っているが、それ以外の場合は異体字も同じ文字として扱っている。これは、日本の国立図書館であるNDLが、まず日本語におい

て同じ文字であるかどうかを判断しているためではないかと推測される。つまり、书は日本にはない文字であるから書の簡体字であることは明白であり、書と同じ文字として扱ってもいいと判断できる。対して、机は日本語では、機の簡体字である前に、「つくえ」という機とは異なる文字である。そのため日本語で異なる文字は別の文字として扱うべきと判断しているのではなかろうか。ただし、その判断基準では「谷」と「穀」だけが同じ文字であることの説明がつかない。他に同様の事例が見受けられないことから、NDL-OPACにおいて、「谷」と「穀」が同じ文字として扱われていることは、極めて例外的な現象であると考えられる。

8.3　NDLサーチ

NDLサーチについてはまず、現場の図書館員はNDLサーチが「日本漢字、簡体字、繁体字は別字として認識」していることを知り、検索漏れを起こさない様に注意する必要がある。「春秋谷梁传」という図書は、「春秋穀梁伝」では検索できないのである。具体的には、戦前の図書や、中国・台湾の図書を検索する際に問題になるであろう。その上で、すべての場合において別の文字を別の文字として扱っているのではなく、実際にはJIS X 0208の範囲内であれば日本漢字と繁体字を同じ文字として扱っていることも事実である。しかし、ある漢字がJIS X 0208の範囲内であるか否かは、図書館員といえどもカウンターや検索端末の前でとっさに判別できるようなものではない。故に図書館員が日常の業務で使うのであれば、「NDLサーチでは一部の旧字体は新字体でも検索可能である」程度に認識しておき、基本的にはNDLのアナウンス通り別字として扱うと考えていても、さほど影響は出ない。ただし、「適合度順」は異なるので、たとえば旧字体で書かれた書誌をできるだけ上位に出したいのであれば、検索キーワードを事前に旧字体に変換するなど、ひと手間かけると良い。NDLサーチには検索語を中国語、韓国語に翻訳する機能がある[16)]。しかし、これはあくまで翻訳で、文字の変換ではない。本研究で示したような変換に使えるとは限らない

以上のことから、検索漏れが発生しやすい順に並べれば、NDLサーチ＞NDL-OPAC＞CiNii Booksであるし、ノイズが発生しやすいのはその逆であるということができる。図書館員は検索をする際には、これらの特性を理解し、使っている検索キーワードが検索漏れを起こしやすいものでないか、気を付ける必要がある。

謝辞

本稿は、第151回ku-librarians勉強会で発表した内容、およびそれを修正して大学図書館問題研究会第44回全国大会で発表した内容を基に構成しており、どちらの会でも多数の質問、コメントをいただき、本稿に役立てることができました。ku-librarians勉強会及び大学図書館問題研究会の皆様に感謝いたします。

注・参考文献

1）米澤誠「目録システム（NACSIS-CAT）の多言語対応」『全国共同利用大型計算機センター研究開発論文集』23（2001）<https://www.nii.ac.jp/CAT-ILL/about/infocat/pdf/daikei.pdf>（accessed 2014-03-31）
2）国立国会図書館関西館アジア情報課『アジア情報の調べ方: 国立国会図書館サーチ、新NDL-OPACを中心に』<https://rnavi.ndl.go.jp/asia/tmp/H23_asiakensyu_1_opac_2.pdf>（accessed 2014-03-31）
3）宮澤彰「学術情報と漢字」前田富祺、野村雅昭編『漢字と社会（朝倉漢字講座）』朝倉書店　2005, p.143-161
4）前掲3）p.154
5）Akira Miyazawa「Ideograph variant forms and usage control in NACSIS-CAT」『Journal of East Asian libraries』138, 5-8（2006.02）
6）前掲 1）
7）前掲 2）
8）水流添真紀「国立国会図書館サーチ、NDL-OPACにおけるアジア言語資料の検索について」『アジア情報室通報』10（1）（2012）<http://rnavi.

ndl.go.jp/asia/entry/bulletin10-1-3.php>（accessed 2014-03-31）
9）岡田有花「Google風検索ページで「論文を普通の人にも」 NIIの「CiNii」、APIコンテストも」『ITmediaニュース』（2009.08）<http://www.itmedia.co.jp/news/articles/0908/17/news006.html>（accessed 2014-06-25）はCiNii（現在のCiNii Articles）のトップページについて「Google風のシンプルなページ」と評するが、本稿で検証するCiNii Books、NDL-OPAC、NDLサーチも同様にデフォルトでは一つの検索窓のみを表示する形式であり、Google風であると言っていいだろう。
10）「内」はu5185、「內」はu5167。
11）国立情報学研究所『目録システム利用マニュアル　2.7.3 漢字統合インデクス』<https://www.nii.ac.jp/CAT-ILL/MAN2/MAN5/CAT5/2_7_3.html>（accessed 2014-03-31）
12）前掲 8）
13）前掲 9）
14）富田倫生「冊子からテキストへ 文字コードがもたらすものと強いるもの （特集　図書館における全文テキストデータの可能性について）」『現代の図書館』49（2）, 95-103（2011）（引用箇所はp. 101）
15）小島浩之「NACSIS-CATの多言語化と中国書目録--『中国語資料の取扱い（案）』の検討」『大図研論文集』24, 23-42（2002）
16）「ヘルプ - 国立国会図書館サーチ」<http://iss.ndl.go.jp/help/help_ja/help.html>（accessed 2014-03-31） 翻訳検索の項目に「入力されたキーワードを中国語・韓国語・英語に翻訳してから検索します。例えば、キーワードに「図書館」と入力すると、「Library」で検索します」とあり。