NACSIS-CAT

# 目録システムの多言語対応

第2版

Библиотéка 도서관 Library Βιβλιοθήκη 図書館 Bibliothèque 圖書館 图书馆

2001.2

文部科学省　国立情報学研究所

# はじめに

国立情報学研究所では，1997年からサービスを開始した新CAT/ILLシステムの次の展開として，目録システムの多言語対応を計画しました。現在は多言語対応となり，いままで扱いが保留されていた中国語や韓国･朝鮮語資料の登録が可能になりました。一方，EXC 文字の一部など，文字の扱いが変わりました。

この冊子では，目録システム多言語対応の目的や内容について説明しています。中国語や韓国・朝鮮語資料の目録業務を行っている図書館だけではなく，目録所在情報サービスを利用しているすべての図書館の方々にお読みいただき，各図書館での対応について検討するための資料として御活用くださいますよう，お願いします。

## CONTENTS

# 1 多言語対応とは？

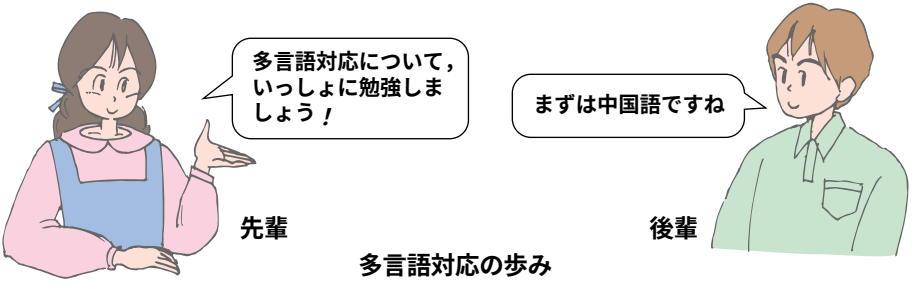
目録システムは，1984年12月にサービスを開始し，1997年には新CATシステムによるサービスを公開しました。目録システムの多言語対応は，この新CATシステムの発展型といえるもので，これにより，ラテン文字（ローマ字。音標符号文字を含む），ギリシャ文字，キリル文字，日本漢字，中国漢字，ハングルなどの入力が可能となります。

特に，従来できなかった中国語資料や韓国・朝鮮語資料について，資料に書かれているとおりの文字で登録することができるようになることがメリットです。

これを実現するため，まず2000年１月にUCS（Universal multiple-octet coded Character Set：国際符号化文字集合）というコード化文字セット（以下，「文字セット」という）に対応したシステムに変更しました。同時に，総合目録データベースに収録されているすべてのデータを，UCSに変換しました。

また，新規の入力作業を効率的に行うため，中国語資料のためにはCHINA-MARC（チャイナ・マーク）（中国MARC），韓国・朝鮮語資料のためにはKOR-MARC（コー・マーク）（韓国MARC）を，参照ファイルとして導入する予定です。

このようなシステム的な対応とは別に，データを入力するための規則の整備も進めています。現在は，総合目録データベースにおける中国語資料（古籍をのぞく）の取扱いを決め，「目録情報の基準」，「コーディングマニュアル」等の改訂を進めています。



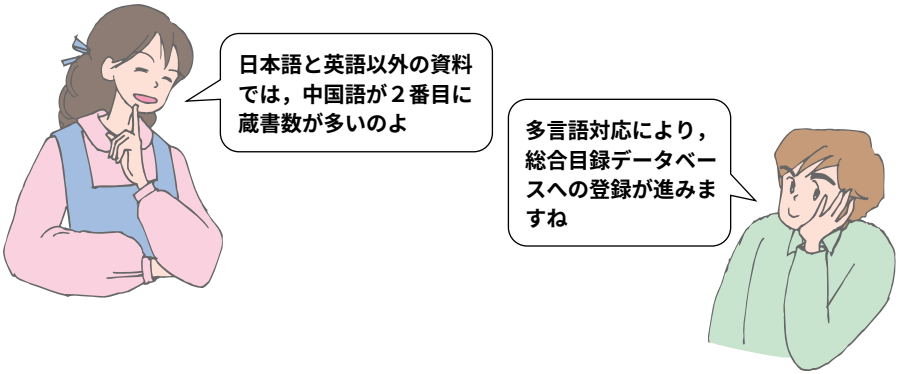

多言語対応の歩み

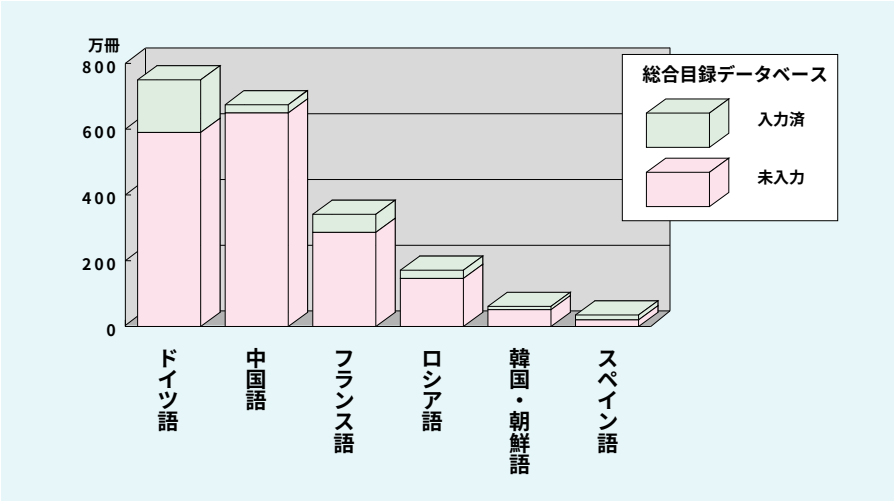
| 年　月 | 内　容 |
|---|---|
| 1995. 11 | 中国語資料データベース化検討WGを設置 |
| 1997. 11 | 新CATシステム業務用・教育用サーバを公開 |
| 1998.　4 | 新ILLシステム業務用・教育用サーバを公開 |
| 1998. 12 | 「中国語資料の取扱い（案）」を発表 |
| 1999. 10 | 多言語対応テストサーバの公開 |
| 2000.　1 | データベースのUCS化，新CATサーバの多言語対応の実施<br>CHINA-MARCの導入<br>中国語対応クライアントの公開 |
| 2000.　4 | 多言語対応クライアントの公開 |
| 2002<br>（目標） | 「韓国・朝鮮語資料の取扱い（案）」を発表<br>KOR-MARCの導入 |

# 2 なぜ多言語対応が必要なのか？

従来の目録システムでは,中国語の簡体字やハングル等の文字を扱うことができないため,国立情報学研究所からは暫定的な入力方法を示すにとどめ,具体的な入力については各参加機関の判断に委ねていました。そのために，総合目録データベースへの中国語，韓国・朝鮮語等の資料の登録は，全体として進んでいませんでした。

全国の図書館では，約 700 万冊の中国語資料，約 70 万冊の韓国・朝鮮語資料を所蔵しているといわれます。目録システムが多言語対応しない限り，これらの言語の目録情報の流通が大きく立遅れ，中国語資料や韓国・朝鮮語資料を必要とする教育・研究に大きな支障をきたすことになります。

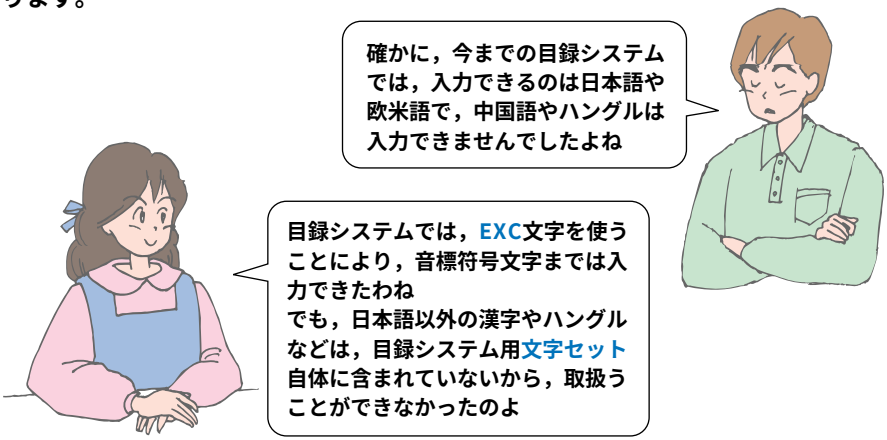



言語別蔵書数の推計（国立情報学研究所調査による）

目録システムの多言語対応により，各図書館で所蔵する中国語資料，韓国・朝鮮語資料の目録作成が可能となります。

総合目録データベースに登録された目録所在情報は，NACSIS-ILL システムやWebcatを通じて活用され，これら資料の共有促進に大きな役割を果たすことになります。




**文字セット**

コンピュータ内部では,文字１個１個をコードとして取扱っています。

コンピュータ同士で情報を交換するためのコード化された文字セットは，日本では JIS で規格化されています。漢字を定義している代表的なものとしてはJIS X0208-1997があり，ここでアルファベットや漢字（第１水準，第２水準）が定義されています。他に JIS X0212-1990 という補助漢字を定義したものもありますが，目録システムでは使えませんでした。

JIS と同様に，中国では GB（簡体字中心の GB2312 が基本セット），韓国では KS，台湾では CNS という規格がありますが，各国独自の漢字はそれぞれにしか定義されていません。

このため従来の各国の文字セットでは,世界の言語を同時に混在させて取扱うことができませんでした。それを解決するために制定されたのが UCS です。UCS では，漢字，ハングル等を含めて約３万６千文字がコード化されています。

# 3 多言語対応の全体像

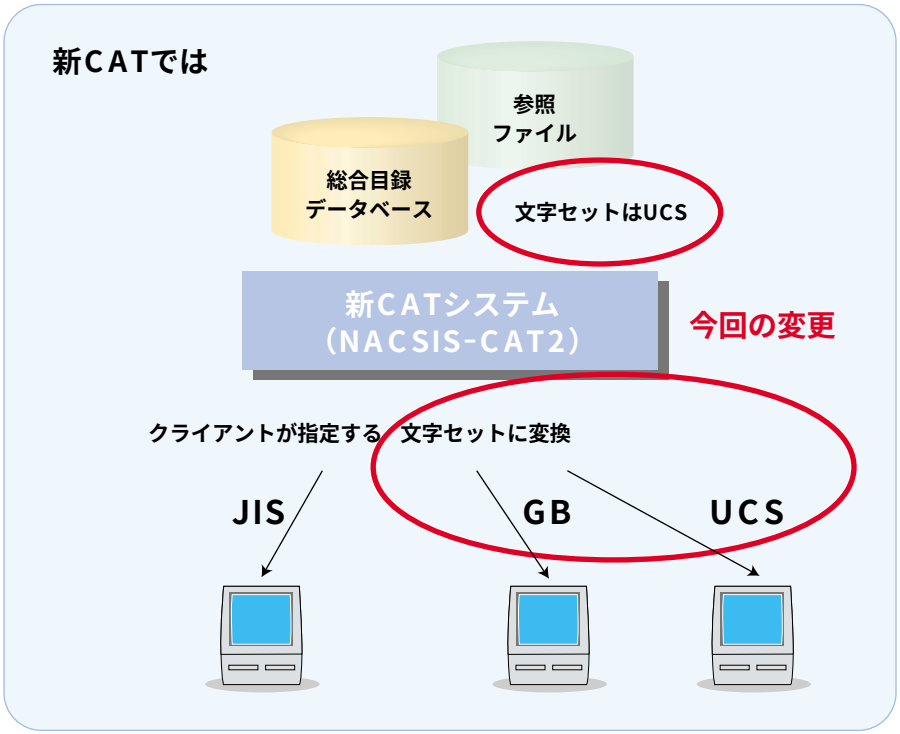
## 3.1 新 CAT システム

多言語対応の新 CAT システムは，次のように，多言語データを保存するデータベースの変更と，多言語データを入出力するサーバの変更により実現されます。

(1) 目録所在情報データベース（総合目録データベース，参照ファイル）の文字セットを，日本語 EUC* から UCS に変更する。

(2) 新 CAT システムのサーバからデータを送る際の文字セットとして，新たに GB と UCS を追加する。従来と同様の文字セット（JIS）もサポートするので，従来のクライアントも継続して使用できる。

(3) 中国語専用としては GB 対応クライアント，中国語以外の文字も扱うには UCS 対応クライアントが必要となる。



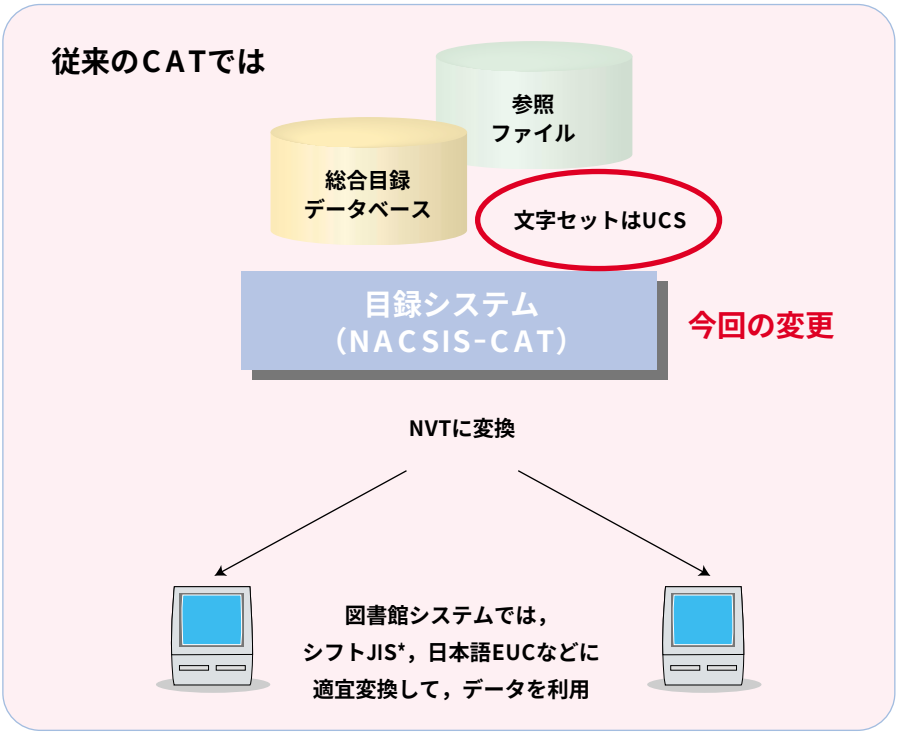

*）日本語 EUC（Extended Unix Code）は，日本語対応 UNIX ワークステーションで使われる内部コード。JIS と同じ文字セットを扱うことができる。目録システムの EUC では，EXC 文字部分を拡張定義している。

## 3.2 従来の CAT システム

従来の目録システムでは，取扱うデータベースが新 CAT システムと同じ多言語データを保存するデータベースとなりますが,データベース内部の文字セットが変わるだけですので，図書館システムに対する影響はありません。

（1） 目録所在情報データベース（総合目録データベース，参照ファイル）の文字セットを，日本語 EUC から UCS に変更する（この部分は，新 CAT システムと同じ）。

（2） 従来の目録システムのサーバからのデータ送信は，従来と同様に NVT という JIS 準拠の文字セットに変換して送る。これにより，図書館システム側の変更は不要である。




*）シフト JIS は，Windows や Mac で使われる内部コード。JIS と同じ文字セットを扱うことができる。EXC 文字を取扱えるかどうかは，図書館システムに依存する。

# 4 多言語対応の仕組み

## 4.1 データベースの多言語対応

総合目録データベースと参照ファイルの文字セットをUCSに変更することにより，収録データは次のようになりました。

| 変　更　前 | 変更後（UCS） |
|---|---|
| 英数字(JIS X 0201),かな･漢字(JIS X 0208)<br>Library<br>図書館 | 対応する英数字，日本語（かな，漢字）<br>Library<br>図書館 |
| 音標符号文字（EXC 文字）<br>bibliothèque | 対応する拡張ラテン文字<br>bibliothèque |
| 漢字，簡体字などの外字<br>◆U56FE◆◆U4E66◆◆U9986◆<br>◆U4E1B◆◆U4E66◆<br>里見◆D9808◆ | 対応する漢字<br>图书馆<br>丛书<br>里見弹 |

従来のEXC文字や外字は，UCSで定義されている拡張ラテン文字や漢字に変換されます。中国語の簡体字，繁体字やハングルなど，ほとんどの文字がUCSに定義されているので，表示が可能となります。

UCSに定義されていない一部のEXC文字（制御文字表示用記号）は，入力の際に省略や置き換えを行い，必要であれば注記することになります。（→「目録情報の基準」第4版 p.62）

漢字などでUCSにも対応する文字がない場合は，従来どおり「◆‥‥◆」という記録方法をとります。

**UCS**

UCS（国際符号化文字集合:Universal multiple-octet coded Character Set）とは，日本語，中国語などの漢字圏を含めた世界的に共通なコード体系の文字セットです。1993年に国際規格ISO/IEC10646-1として制定され，1995年1月にJIS X0221として日本の規格に採用されました。

現在定義されているのは，基本多言語面（BMP:Basic Multilingual Plane）と呼ばれる部分の約3万6千文字です（この部分は，Unicodeと呼ばれる文字セットと同一です）。そのうちCJK統合漢字といわれる中国語，日本語，韓国語で使われる漢字を標準化した部分では，20,902字の漢字が定義されています。

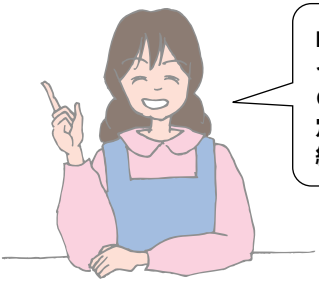
## 4.2 新 CAT サーバの多言語対応

新 CAT システムのサーバは，UCS に対応したデータベースのデータを，クライアントが指定した文字セット（エンコーディング）にあわせて変換して送ります。指定できる文字セットには，JIS7（EXC 文字を含む JIS），ISO 2022JP（EXC 文字を含まない JIS），GB，GBK，UCS があります。指定した文字セットにより，データの表示は次のようになります。

○では，該当する文字そのもので表示されるのに対し，×では，「◆U‥‥◆」のように，１文字ずつ UCS コードを示す番号で表示されます。

| UCS データ ＼ 文字セット | JIS7 | ISO 2022JP | GB | GBK | UCS |
|---|---|---|---|---|---|
| 英数字（JIS X 0201 相当） | ○ | ○ | ○ | ○ | ○ |
| かな・漢字（JIS X 0208 相当） | ○ | ○ | △ * | △ * | ○ |
| 音標符号文字（EXC 文字相当） | ○ | × | △ ** | △ ** | ○ |
| 中国漢字（GB2312 相当） | × | × | ○ | ○ | ○ |
| UCS にしかない漢字，簡体字など | × | × | × | ○ | ○ |
| ハングル | × | × | × | × | ○ |

*）ひらがな，カタカナ及び中国語に対応するものがある漢字だけは○
**）ピンイン表記に使われる音標符号文字は○



ISO2022JPは，EXCをもっていないクライアントのための文字セットです
GBKはGBの拡張版で，漢字21,003字を定義しているから，UCSの漢字部分は網羅しているのよ

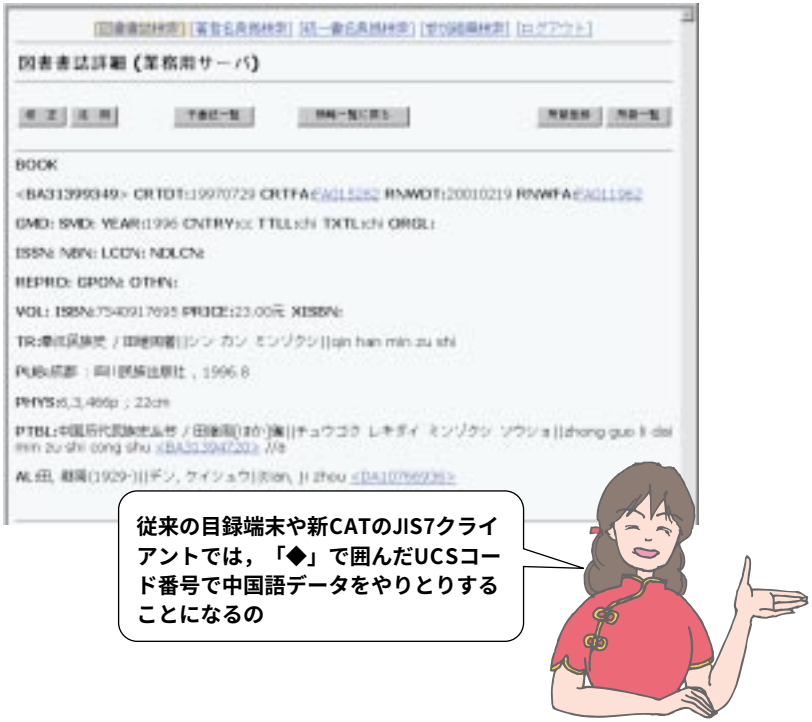
ハングルを扱うには，UCSでないといけませんね

## 4.3 クライアントの多言語対応

UCS によるデータの表示と入力を行うためには，UCS に対応した新 CAT クライアントが必要です。具体的には，Windows2000 等の UCS に対応可能な OS で動作する，UCS 文字セットの入力ツールを備えたクライアントを開発する必要があります。

もちろん，英数字，かな，漢字，EXC 文字の範囲であれば，従来の目録システムや新 CAT クライアント（JIS7 対応）で，従来どおりの入力が可能です。

次の画面例は，NACSIS で開発した多言語対応クライアント（UCS クライアント）です。



# 5 目録システムでの登録と検索

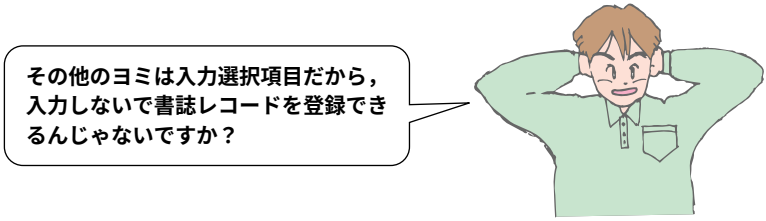
## 5.1 登　録

中国語の入力を行う場合，UCSクライアントを利用するのが便利です。JISクライアントでは，一文字づつ「◆U‥‥◆」という形式でのコード入力*を行わなければ，簡体字などの入力ができないので，実用的ではありません。**

*） UCS コード番号は，「今昔文字鏡」や「ATOK」などの市販ソフトで調べることができる。UCS にない文字については，これまでと同様，大漢和辞典や広漢和辞典の検字番号を使うなどの方法で入力する。
**）中国語以外については，従来と同様の方法で入力が可能。

多言語対応時には，中国語資料のヨミにあたるピンインを記録するために，その他のヨミフィールドを新設します。この新規フィールドの表示・入力ができるように，新 CAT クライアントを対応させる必要があります。

従来の目録システム及びその他のヨミフィールドに対応していない新 CAT クライアントでは，このフィールドにデータが記録されている書誌レコードを確認して，所蔵レコードを登録することはできますが，書誌レコードを修正することはできません。



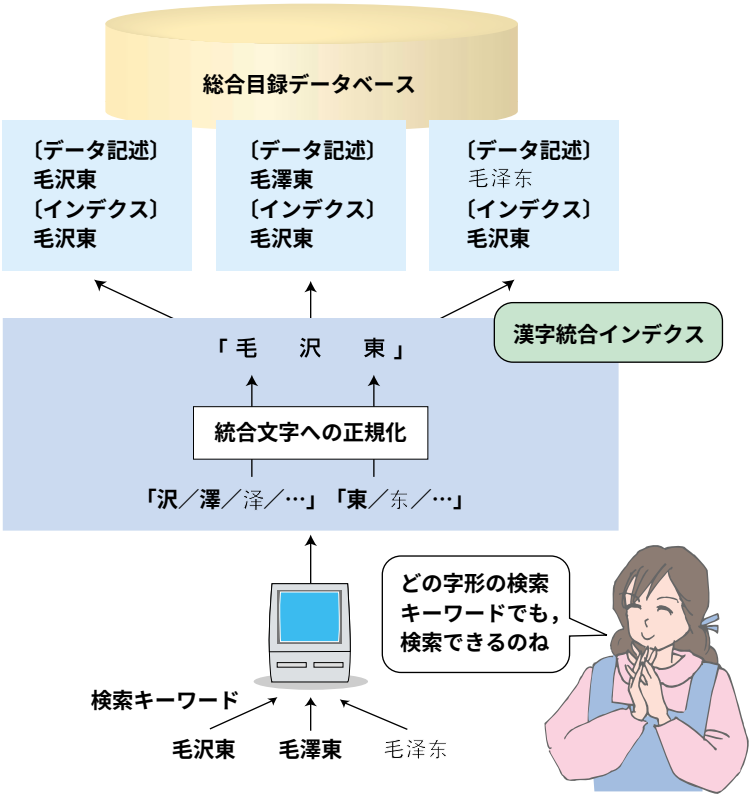

## 5.2 検　索

UCSの統合漢字部分には，約2万字の漢字が含まれており，似た形や同じ意味の漢字が数多くあります。このことによる検索もれを防ぐため，似た形や意味の同じ漢字を含めて検索するように，**漢字統合インデクス**を用意しました。

検索をするときは，検索キーワードに対して漢字統合インデクスによる正規化をしたのち，書誌レコードの検索用インデクスと照合して検索結果をだす仕組みになっています。

書誌レコードのデータ記述は表記そのままの文字ですが，検索用インデクスは漢字統合インデクスにより正規化して登録されていますので，このようなことが可能となるわけです。

なお，漢字統合インデクスは従来の目録システムでも機能しますので，中国語資料の検索は可能となります。



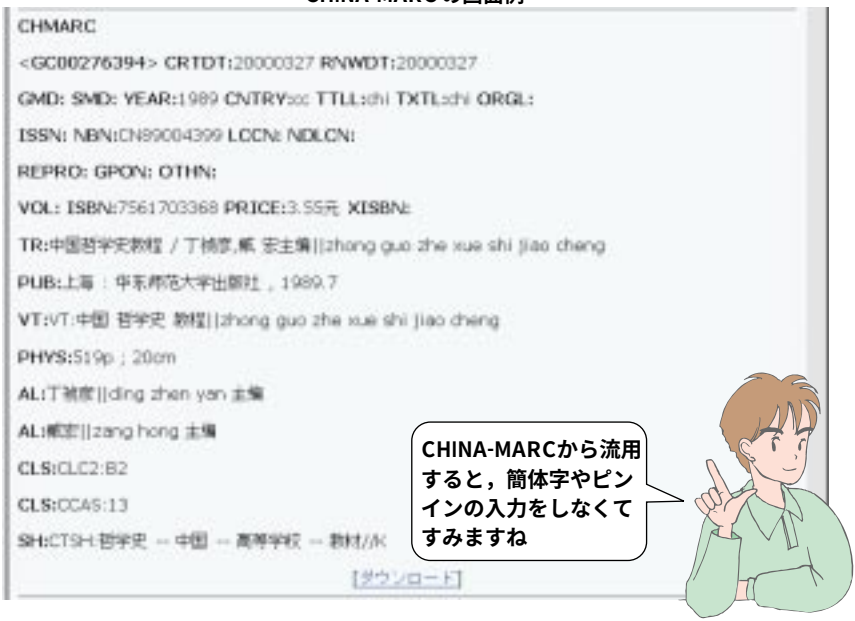
## 5.3 CHINA-MARC

中国語資料の総合目録データベースへの効率的な入力を支援するため,参照ファイルに CHINA-MARC を導入します。CHINA-MARC は，北京図書館（中国国家図書館）で作成する中華人民共和国の図書・雑誌の書誌情報を収録しています。

CHINA-MARC の内容

| | |
|---|---|
| データの収録年 | 1988 年～ |
| データ総件数 | 約 30 万件 |
| 年間増加データ件数 | 約　4 万件 |
| 文字セット | GB2312<br>（将来的には，GBK 対応も計画している） |

CHINA-MARC の画面例



# 6 多言語目録データベースの利用

総合目録データベースに入力された中国語資料は，WebcatやNACSIS-IRなどの検索システムでも公開されます。これらの検索システムについても，将来的には，UCS対応のブラウザを利用すれば中国語などの表示が可能になることが期待されます。

図書館システムとしては，次の３種類の対応が考えられます。

**（1）図書館システム側も多言語対応（UCS対応）を行う**

この場合，図書館側のデータベースも含めて，図書館システム全体をUCS対応にしなくてはなりません。

**（2）多言語対応のシステムを別に用意する**

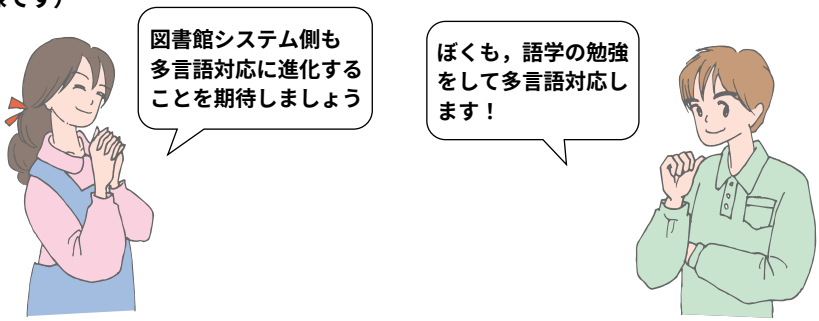
この場合，従来の図書館システムはそのままで，あらたに多言語対応（あるいは中国語対応など個々の言語対応）のシステムを用意し，言語コードなどで目録データを選択して，収録することになります。

**（3）多言語対応はしない**

中国語資料などを登録する必要がない場合は，将来的にも従来のままのシステムで対応を続けても差し支えありません。

（目録システムの利用に関しては，漢字統合インデクスの機能があるため，検索することはできます。この場合，中国語の部分は「◆U‥‥◆」というデータにおきかわり表示されます）

なお，ILLデータベースについても，平成12年１月にUCS対応を行いましたが，ILLシステムのサーバ部分で送信するデータは，従来のままJIS（従来のCATではNVT）となるようにします。これにより，図書館システム側としては，システム的な変更なしに，従来と同様にILLシステムを利用できることになります。（本誌ｐ．４～５のCATと同様です）




# 中国語資料の取扱いの基本方針

## 1 基本方針

### (1) 適用する目録規則

原則として「日本目録規則 1987 年版改訂版」を適用する。ただし，例外的に中国語資料の特性，CHINA-MARCからの流用入力に対応するため，「中国文献编目规则」を適用する場合もある。こうした例外的なケースについては，「コーディングマニュアル」等で明示するものとする。

### (2) 文字の取扱い

記述部分に関しては，転記の原則に従い，書かれたままの字体で記録する。

著者標目に関しては，原則として，最初に典拠レコードを作成する際に用いた資料に表示されている字体のまま記録する。ただし，著名な著者又は清朝までの著者等については，最も良く知られた字体で記録する。

なお，上記のような方針に対応し，異なる字体間での検索を可能とするため，漢字統合インデクスを作成する。

### (3) ヨミの取扱い

漢字の単語単位での検索を可能とするため，日本語ヨミの付与を必須とする。それに伴い，中国語資料について日本語ヨミと分かち書きの規則を新規に作成する。したがって，これは，ヨミの標準化を目的とするものではなく，検索のための便宜的な規則である。

ピンインについては，選択事項として日本語ヨミとは別に記録することができる。

### (4) 古籍の取扱いについて

古籍用として，別ファイルを設定することはないが，入力規則は別に作成する予定である。

### (5) 既存データの取扱い

既存データは，現行の文字コードをUCSに変換し，総合目録データベースに格納する。

NII

お問い合わせ先:
文部科学省　国立情報学研究所
開発・事業部　コンテンツ課
〒101-8430　東京都千代田区一ツ橋2-1-2
TEL.03−4212−2355
FAX.03−4212−2375
E−mail:catadm@nii.ac.jp