# 目録システム（NACSIS-CAT）の多言語対応

米澤 誠
国立情報学研究所開発・事業部コンテンツ課

## 1. はじめに

国立情報学研究所で運用する目録システムは，1984 年 12 月にサービスを開始し，1997 年には新しいプロトコル：CATP（Cataloging information Access & Transfer Protocol：目録情報を検索・転送するための専用プロトコル）を使った新システムを稼動した。目録システムの多言語対応は，この新システムの発展形といえるもので，これにより，従来のラテン文字（ローマ字。音標符号付き文字を含む），ギリシャ文字，キリル文字，日本漢字に加えて，中国漢字，ハングルなどを扱うことが可能となった。

これを実現するため，まず 2000 年 1 月の計算機リプレースにあわせて，UCS（Universal multiple-octet coded Character Set：国際符号化文字集合）に対応したサーバシステムに変更した。同時に，データベースに収録されているすべてのデータを，UCS に変換した。

本報告では，多言語対応の概要とシステム内容について報告することとしたい。

## 2. 多言語対応の必要性

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，目録データの入力規則としては，日本漢字を代用するという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た。そのために，目録システムで作成する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なかった。

全国の大学図書館では，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る。目録システムが多言語対応しない限り，これら言語の目録情報のデータベース化と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る。

目録システムが多言語対応することにより，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った。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る。

## 3. 多言語対応の実際

多言語対応の目録システムは，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している。

(1) 目録システムのデータベース（総合目録データベース，参照ファイル＝外部機関作成 MARC）の文字セットを，日本語 EUC から UCS に変更する。

(2) 新システムのサーバ（以後，「CATP サーバ」という）からデータを送る際の文字セットとして，新たに GB（中国語用文字セット）と UCS 等を追加する。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。

### 3.1 データベースの UCS 対応

データベース及び CATP サーバ内部で使用する文字コードを，EUC から UCS に変更した。データベースでは，DBMS として日立製作所製のリレーショナルデータベース HiRDB を使用している。OS は，HP-UX11 である。使用する UCS は，以下のような仕様としている。

(1) 目録システムで用いてきた，従来の EXC 文字（音標符号付き文字等）を含む。

(2) UCS の実装水準は 3 とする。これは，EXC 文字の全てを表現するのに，合成文字を使用するためである。

(3) JIS X0208 の漢字包摂規準を採用する。

データベースの文字セットを UCS に変更することにより，収録データは次のような文字となっている。

表 1. データベースの収録データ

| 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 |
| 音標符号付き文字（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>里見弼 |

従来の EXC 文字や外字は，UCS で定義されている拡張アルファベットや漢字に変換された。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，表示が可能となった。

UCS に定義されていない一部の EXC 文字（制御文字）及び X0208 に定義されていない記号は，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とした。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っている。

### 3.2 CATP サーバの多言語対応

従来のサーバでは，JIS7 エンコーディングだけに対応していたが，新たに GB，GBK，UTF8，ISO2022JP の 4 種類のエンコーディングに対応可能となった。

(1) ISO2022JP： EXC 文字を含まない JIS7 文字セットを扱うものであり，EXC 文字は UCS の番号（UCS 外字）で表示される。

(2) GB： 簡体字中心の中国語用文字セットで，日本語の表示はできない。

(3) GBK： GB の拡張版で，UCS と同等の CJK（中日韓）統合漢字を含んでいる。

(4) UTF8： UCS の文字セットを扱うエンコーディング。

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る。○では，該当する文字で表示されるのに対し，×では，「◆U…◆」のように，1 文字づつ UCS コードを示す番号で表示される。

表 2. CATP サーバのエンコーディング

| 文字セット / UCS データ | JIS7 | ISO 2022JP | GB | GBK | UTF8 |
|---|---|---|---|---|---|
| 英数字（JIS X 0201 相当） | ○ | ○ | ○ | ○ | ○ |
| かな，漢字（JIS X 0208 相当） | ○ | ○ | △* | △* | ○ |
| 音標符号付き文字（EXC 文字相当） | ○ | × | △** | △** | ○ |
| 中国語（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付き文字は ○
***) 日本語に対応するものがある漢字だけは ○

クライアントが使用する文字コードとサーバ内部の文字コード（UCS）との変換は，CATP エンコーディングの指定により多言語対応サーバで行う。この UCS 変換は双方向の変換であり，CATP リクエスト時と CATP レスポンス時に実行される。

クライアント文字コードに変換できない文字を，UCS 外字という。これは，◆で UCS コード値をはさんだ形に変換する。

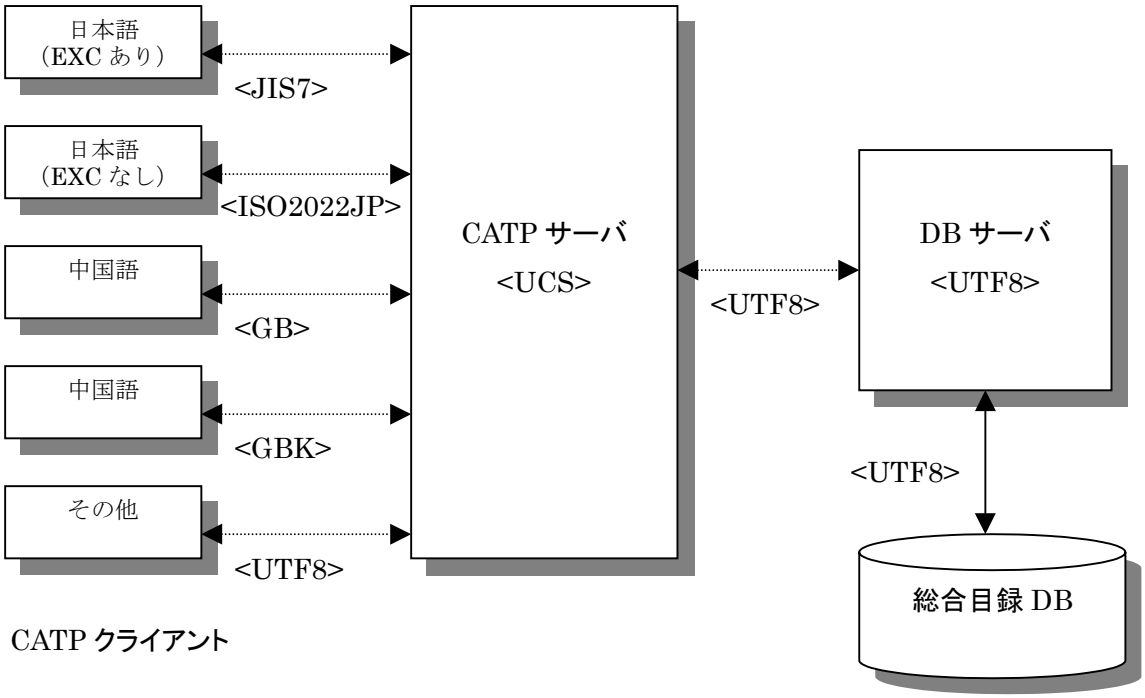



図 1. 新 CAT サーバの多言語対応

### 3.3 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 CATP クライアントが必要である。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る。CATP クライアントを開発している各図書館システムメーカでは，UCS 対応のクライアントの開発を進めている。

下の画面例は，国立情報学研究所で開発した UCS 対応クライアント：WebUIP で，2001 年 1 月から公開している。WebUIP は，Web に対応した新 CAT/ILL ゲートウェイシステムで，標準的な Web ブラウザを CATP クライアントとして使い，CATP サーバを利用することができる。

WebUIP 自体は CATP サーバの外部に位置し，ブラウザからのリクエストを受けて，CATP サーバを経由して，データベースを利用するという構成をとっている。Internet Explorer 5 等の，UCS（UTF8）の入出力が可能なブラウザ環境で，中国語資料等の入力・表示が可能となっている。


画面 1. UCS 対応クライアント（WebUIP）

## 4. CATP サーバの正規化処理

CATP サーバでは，クライアントから送られてきたデータに対して，以下の処理を行った上で，データベースに格納している。項番 1 の文字コード変換については前述した。項番 2，3 の処理について説明する。

表 3. CATP サーバの正規化等処理

| 項番 | 処理 | 内容 |
|---|---|---|
| 1 | 文字コード変換 | サーバ内部文字コードと外部文字コードの相互変換を行う |
| 2 | UCS 包摂 | CATP クライアントから送られた外部文字データを正規化する |
| 3 | 検索キー正規化 | 検索キーに対して正規化処理を行う |

### 4.1 UCS 包摂処理（文字の正規化処理）

UCS 包摂とは，データベースに格納するデータ，および CATP サーバの内部処理のために行う正規化である。

サーバでの CATP メッセージの受信時に CATP オブジェクトボディのデータに対して，UCS コードへの変換，UCS 包摂の順に行う。UCS 包摂で行う処理の一覧を次に示す。

表 4. UCS 包摂処理

| 処理順 | 項目 | 対象 | 概要 |
|---|---|---|---|
| 1 | ラテン文字包摂 | ラテン文字 | バイト数に関わらず，同一のコードポイントとする |
| 2 | 数字包摂 | 数字 | 同上 |
| 3 | 記号包摂 | 記号 | 同上 |
| 4 | 漢字包摂 | 漢字 | JIS X0208 の漢字包摂規準に従って漢字を包摂する |
| 5 | カタカナ包摂 | カタカナ | バイト数に関わらず，同一のコードポイントとする |
| 6 | 合成文字包摂 | 合成文字 | 合成文字の正規化（カノニカルオーダリングによる）を行う* |

*) UCS では，EXC のような合成文字は，1 文字で表しても，複数文字の組み合わせで表してもよい。それらを統一するために，異なる組み合わせパターンの合成文字（基底文字＋結合文字）を，一つの合成文字に正規化する。 例）「Á」←「A」＋「´」

### 4.2 検索キー正規化

検索キー正規化とは，更新レコード及び検索条件から検索キーを生成する際に行う正規化である。検索キー正規化で行う処理の一覧を次に示す。

表 5. 検索キー正規化処理

| 項番 | 項目 | 対象 | 概要 | 備考 |
|---|---|---|---|---|
| 1 | 大文字化 | ラテン文字 | ラテン文字の小文字を大文字へ変換する。 | |
| 2 | | ギリシャ文字 | ギリシャ文字の小文字を大文字へ変換する。 | |
| 3 | | キリル文字 | キリル文字の小文字を大文字へ変換する。 | |
| 4 | | カタカナ | 拗音，促音を大文字へ変換する。 | |
| 5 | 小文字化 | ラテン文字 | ラテン文字の大文字を小文字へ変換する。 | コード類のみ |
| 6 | カタカナ化 | ひらがな | ひらがなをカタカナへ変換する。 | |
| 7 | ラテン文字化 | EXC | EXC をラテン文字へ変換する。 | |
| 8 | 漢字統合 | 漢字 | 似た形や同じ意味の漢字を，そのグループを代表する漢字へ変換する。 | 漢字統合インデクスによる |

### 4.3 漢字統合インデクス

UCS の漢字統合部分には，約 2 万字の漢字が含まれており（拡張領域を除く），似た形や同じ意味の漢字が数多くある。このことによる検索漏れを防ぐため，似た形や同じ意味の漢字を含めて統合検索を可能とするのが，漢字統合インデクスである。

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，書誌レコード等の検索用インデクスと照合して検索結果を出す仕組みになっている。

書誌レコードのデータ記述は表記そのままの文字であるが，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るため，このようなことが可能となる。

なお，漢字統合インデクスは従来の目録システムや Webcat でも機能するので，それらシステムでも中国語資料の検索は可能となっている。

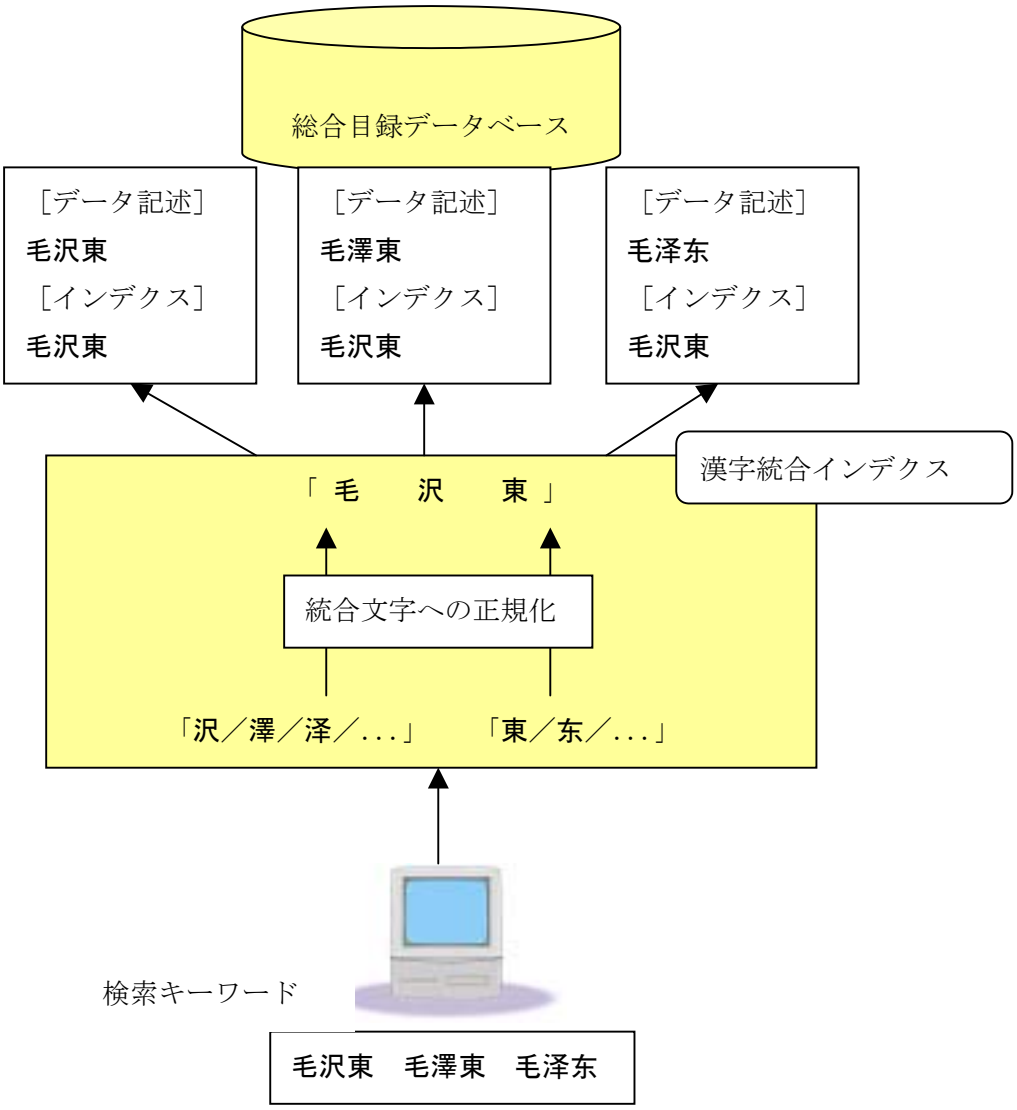
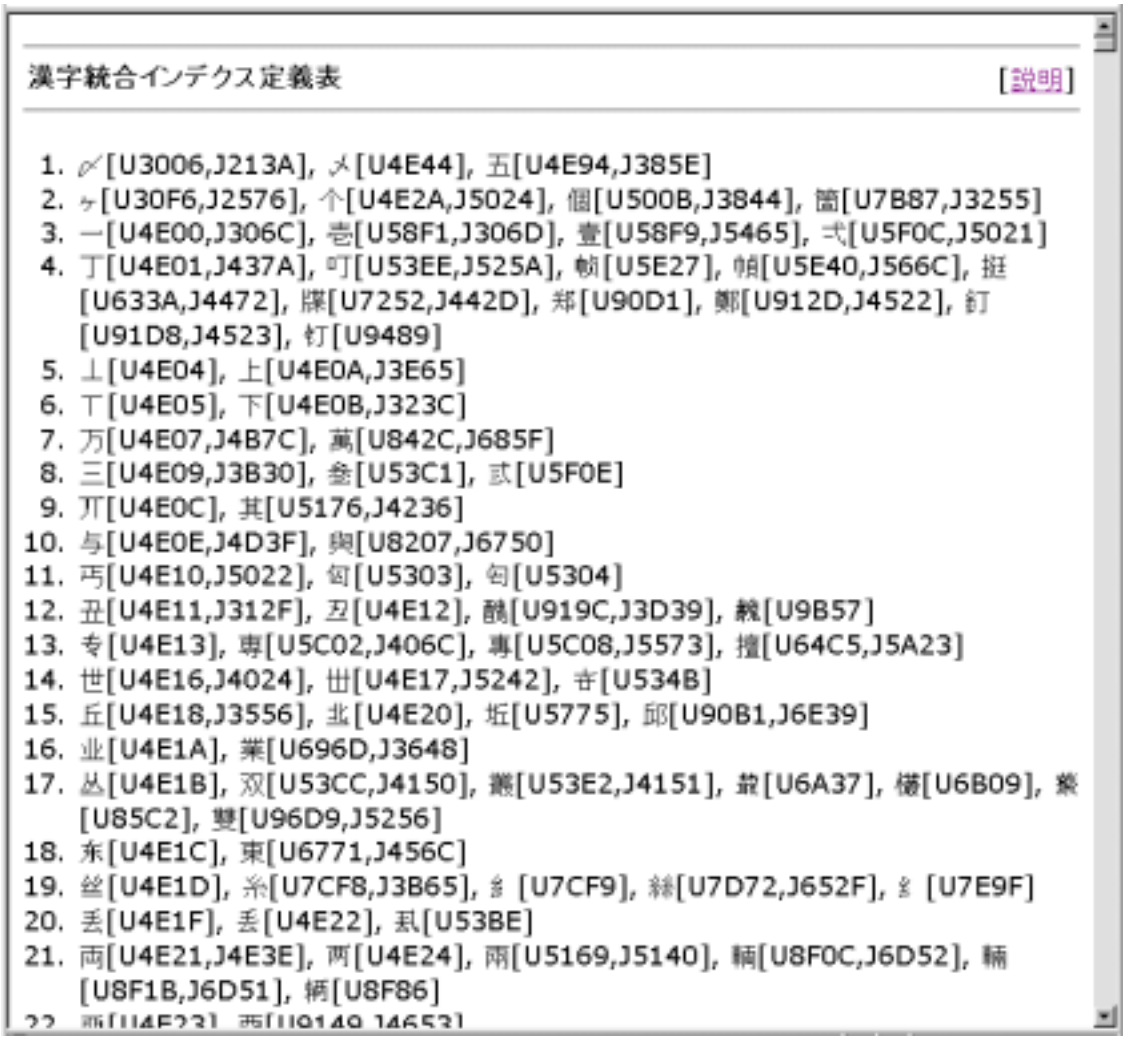



図 2. 漢字統合インデクスの仕組み

漢字統合インデクスは，次のような表として定義されている。



漢字統合インデクス定義表 [説明]

1. 𠂇[U3006,J213A], 乂[U4E44], 五[U4E94,J385E]
2. 𠂉[U30F6,J2576], 个[U4E2A,J5024], 個[U500B,J3844], 箇[U7B87,J3255]
3. 一[U4E00,J306C], 壱[U58F1,J306D], 壹[U58F9,J5465], 弌[U5F0C,J5021]
4. 丁[U4E01,J437A], 叮[U53EE,J525A], 帄[U5E27], 帥[U5E40,J566C], 挺[U633A,J4472], 牒[U7252,J442D], 郑[U90D1], 鄭[U912D,J4522], 釘[U91D8,J4523], 钉[U9489]
5. 丄[U4E04], 上[U4E0A,J3E65]
6. 丅[U4E05], 下[U4E0B,J323C]
7. 万[U4E07,J4B7C], 萬[U842C,J685F]
8. 三[U4E09,J3B30], 叁[U53C1], 弐[U5F0E]
9. 丌[U4E0C], 其[U5176,J4236]
10. 与[U4E0E,J4D3F], 與[U8207,J6750]
11. 丐[U4E10,J5022], 匄[U5303], 匂[U5304]
12. 丑[U4E11,J312F], 刁[U4E12], 鵰[U919C,J3D39], 貔[U9B57]
13. 专[U4E13], 專[U5C02,J406C], 專[U5C08,J5573], 擅[U64C5,J5A23]
14. 世[U4E16,J4024], 丗[U4E17,J5242], 吉[U534B]
15. 丘[U4E18,J3556], 北[U4E20], 坵[U5775], 邱[U90B1,J6E39]
16. 业[U4E1A], 業[U696D,J3648]
17. 丛[U4E1B], 双[U53CC,J4150], 雙[U53E2,J4151], 叢[U6A37], 樂[U6B09], 叢[U85C2], 雙[U96D9,J5256]
18. 东[U4E1C], 東[U6771,J456C]
19. 丝[U4E1D], 糸[U7CF8,J3B65], 糹[U7CF9], 絲[U7D72,J652F], 糸[U7E9F]
20. 丢[U4E1F], 丟[U4E22], 乱[U53BE]
21. 两[U4E21,J4E3E], 两[U4E24], 兩[U5169,J5140], 輌[U8F0C,J6D52], 輛[U8F1B,J6D51], 裲[U8F86]
22. 严[U4E23], 严[U9149,J4653]

図面 2. 漢字統合インデクス定義表

定義表は，漢字統合グループごとに並んでおり，先頭の数字は，漢字統合グループの連番である。漢字統合グループは，インデクス作成時に同じ漢字に統合される漢字のグループである。漢字のあとの角括弧([ ])には，その漢字の UCS コードとそれに対応する JIS コード(JIS X 0208)を記入している。ただし，対応する JIS コードがない場合，JIS コードは記入されていない。UCS コードにはその先頭に「U」が，JIS コードには「J」が付加されている。

統合先漢字は，原則として次のように決定される。

(1) 同一グループに属する漢字のうち，JIS コードの最も小さい漢字を統合先とする。

(2) ただし，そのグループ内に対応する JIS コードのある漢字が一つもない場合，UCS コードの最も小さい漢字を統合先とする。

このインデクスは，本研究所ホームページで公開しており，目録所在情報サービス参加機関の図書館システムを支援する目的に限って利用できる（URL：http://www.nii.ac.jp/CAT-ILL/INFO/newcat/kanji/kui_about.html）。

## 5. Webcat の多言語対応

目録システムで構築された総合目録データベースを，簡便に検索するためのシステムが，Webcat である。標準的なブラウザから，インターネットを通じて自由にアクセスできるために多くの利用があり，一日平均 4 万件の検索がなされている（2001 年度平均）。

この Webcat では，2001 年 1 月に UCS（UTF8）対応を実施し，中国語が表示できるようになった。UCS（UTF8）に対応したブラウザであれば，どこからでも利用できる（URL：http://webcat.nii.ac.jp/）。


画面 3. Webcat の多言語対応

## 6. おわりに

以上，国立情報学研究所における目録システムの多言語対応についての，特にシステム面での実装状況について報告した。

各図書館システム側の対応状況，データ入力規則の整備，総合目録データベースへの多言語資料データの入力状況，多言語対応に関する今後の課題等，報告すべき点は数多くあるが，それらについては，稿をあらためて報告したい。