情 報 学 基 礎 41－3
（1996．4．18）

学術情報センターにおけるオンラインDBとIRシステムの連携
—目録系データベースの構成—

大山敬三　高須淳宏　鶴岡弘

学術情報センター研究開発部

本論文では目録所在情報システムであるNACSIS-CATと情報検索システムであるNACSIS-IRとの連携方式について述べている。オンラインで更新される書誌・所蔵データベース中のレコードに対して、自動的に情報検索用のインデクスを付与してエンドユーザへ検索サービスを提供するシステムの構成を検討している。データベースサーバへの負荷、検索応答性能、ファイル容量などの面からいくつかの方式を評価して実際のシステムの基本設計を行った結果として、データベースの構造と性質に応じた最適な構成が示されている。

Design of an IR System linked to Online DB at NACSIS
—Configuration of Catalog Databases—

Keizo OYAMA, Atsuhiro TAKASU, Hiroshi TSURUOKA

National Center for Science Information Systems (NACSIS)

3-29-1 Otsuka, Bunkyo-ku, Tokyo 112, Japan

This paper describes a method to link NACSIS-IR, an information retrieval system, with NACSIS-CAT, a catalog and location information system. The authors studied the configuration method of a system which automatically builds index of bibliographic and holding records stored in databases updated on-line and which provides information retrieval services to the end users. Several methods were evaluated from viewpoints of load to the database server, response time of retrieval, file capacity and so on in order of the basic design of an actual system. As the result, the optimal configurations corresponding to structure and characteristics of databases are obtained.

1. はじめに

学術情報センターでは図書館向けのサービスとしてオンライン目録所在情報サービス(NACSIS-CAT、以下CAT)を行う一方、大学研究者向けのサービスとして情報検索サービス(NACSIS-IR、以下IR)を提供している。CATは大学等の図書館が所蔵する図書・雑誌の目録と所在情報をオンラインで検索・更新する総合目録DBシステムである。IRは学術情報センターが構築したり外部から導入した学術情報DBのオンライン情報検索システムである。CATは図書館向けのサービスであるが、ここで構築されたDBは研究者に取っても有用なものであり、従来からIRでも検索サービスを提供してきた。

センターでは1996年1月にこれらサービス用のメインシステムの更新を行ったが、このシステムでは従来の互換性と将来のサービスの展開を考えて、いわゆるメインフレーム（互換系）とUNIXシステム（オープン系）の複合構成をとっている。CATではデータベース機能の向上とインターネットへの柔軟な対応を可能とするため、データベース機能をアプリケーションから切り離しオープン系上で再構築するための開発を進めており、1996年8月には移行を完了する予定である。一方、IRでもより高度な検索機能や、インターネットを通じたユーザフレンドリな検索インタフェースの提供、あるいは全文DBへの対応などを可能とするため、オープン系のシステム上に新しい情報検索システム（新IRシステム）の開発を進めており[1]、1996年度のできるだけ早い時期から一部のDBの試行サービスを始める予定である。

本稿では、これらの開発の中で実現しようとしている、オンラインDBシステムと情報検索システムの連携方式について述べる。

2. NACSIS-CATシステムの構成

2.1 データベース

CATのDBは目録所在情報DB（総合目録DB）と参照目録DBから構成されている。これらのDBの種類と規模を表1に示す。

総合目録DBは和文・欧文、図書・雑誌の書誌・所蔵情報と著者名などの典拠情報を集積したものであり、参加図書館が日常の整理業務としてオンラインで検索・更新を行っている。参照目録DBは書誌データの作成の効率化を図る目的で、外部の目録作成機関から提供を受けたMARC (Machine Readable Catalog) をフォーマット変換して提供しており、定期的にバッチ更新を行っている。

総合目録DBのレコード構成を図1に示す。書誌、所蔵、典拠はそれぞれ別テーブルに格納され、書誌レコード中には典拠レコードID、所蔵レコード中には書誌レコードIDを持つことによ

表1. NACSIS-CATのデータベースの種類と規模

テーブル名称			収納件数 (1996.3.15)	週間更新 (1996.3.8~)
総合目録DB	書誌	和図書	1,142,325	8,014
		洋図書	2,261,161	9,948
		和雑誌	82,389	471
		洋雑誌	129,669	721
		合計	3,615,544	19,154
	所蔵	和図書	14,346,920	107,060
		洋図書	7,176,322	47,831
		和雑誌	1,685,324	29,363
		洋雑誌	1,144,491	2,350
		合計	24,353,057	186,604
	典拠	著者名	852,692	2,733
		統一書名	8,990	87
		変遷マップ	22,389	137
		合計	884,071	2,957
参照目録DB	書誌	和図書(JP)	1,791,122	2,758
		和図書(misc)	462,228	2,866
		和雑誌(JP)	95,846	–
		洋図書(LC)	4,808,830	37,006
		洋図書(misc)	1,929,403	–
		洋雑誌(LC)	698,938	–
		非文字(LC)	268,147	–
		合計	10,054,514	42,630
	典拠	著者名(LC)	2,849,477	–
		著者名(JP)	327,561	–
		統一書名(LC)	164,898	–
		合計	3,341,936	0

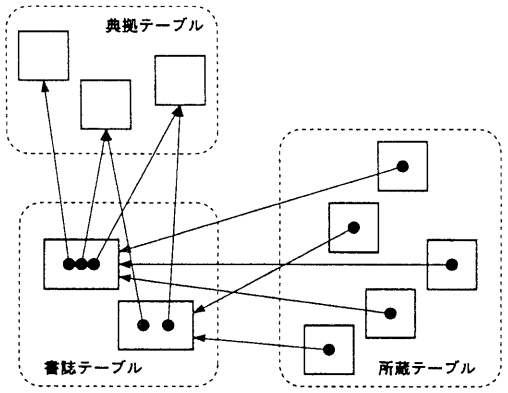


図1. 総合目録DBのレコード構成

りリンクを形成している。書誌ごとに対応して所蔵が作成されるので、レコード間の対応は1:*n*になる。一方、典拠は書誌とは独立なので対応は*m*:*n*になる。

2.2 システム

移行後のCATシステムの構成を図2に示す。現在のCATのシステムはDBサーバ、画面型アプリケーション (SOA)、ユーザインタフェース (UIP) の3つから構成されている。DBサーバはオープン系上に実現され、DBACと呼ぶアクセスインタフェースを通じてDBMS機能を提供する。SOAは互換系上に実現されており、図書館の目録および文献複写業務に対応した機能を提供する。UIPはさまざまなプラットフォーム上

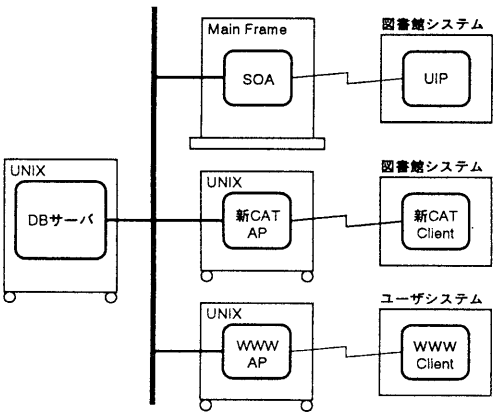


図2. NACSIS-CATのシステム構成

に実装されており、参加図書館のローカルシステム上で稼働し、ローカルなDBとの連携を行っているものも多い。SOAとUIPの間は独自の仮想画面転送プロトコル[2][3]により通信を行っている。

現在、オープン系の上に新しい目録・所在情報アプリケーションサーバ（新CATシステム）を開発中であるが、DBサーバは上記のものを共用し、DBACをAPIとして用いる。また、総合目録に関してはWWWを通じてOPACサービスを提供する予定であり、このアプリケーションもDBACを通じてDBサーバにアクセスする。

3. 目録系IRシステムの構成

3.1 背景

NACSIS-IRでは、学術文献を中心とする各種の索引や抄録などの二次情報や、学術論文などの全文の一次情報などの検索サービスを行っている。これらのDBの多くは外部から導入したりセンターが一括で作成したりしており、更新の頻度は1週間に1度程度から数年に1度程度までさまざまであるが、いずれもIR専用のDBを構築して運用している。

しかし、総合目録DBは参加図書館からオンラインで入力されるDBをサービスしているため、IRとしてもできるだけ最新の情報を提供することが望ましい。現在のIRシステムでは、CATのデータを複製して専用のDBを作り、CATの更新情報をためておいて定期的にバッチ更新するという形態で運用している。しかし、CATシステムでは、サービスシステムの移行でDB機能をDBサーバとして分離するのにともない、オンラインで本体DBを直接即時更新するとともに24時間運転をするようになるため、現状の運用形態を続けることが難しくなる。

また、参照目録DBについても、現在は通常のIRと同じようにCATのデータを複製して専用DBで運用しているが、CAT用とIR用にDBを重複して保守するのは容量的にも時間的にも無駄

が多いため、CAT用のDBを共用できることが望ましい。

目録系DB以外でも、学術雑誌目次速報DBは協力図書館から電子メールなどでデータを随時送ってもらい、情報をDBMSで管理している。IRサービス用には全レコードをダンプして専用のDBを作成しているが、これも管理用のDBをそのまま使えることが望ましい。

そこで、新IRシステムでは、これらのDBをできるだけ統一的な方式で、しかもデータの重複を減らして自動的な運用ができるようにすることを目標にしてシステムの設計を行った。

3.2 設計上の要求事項

CATシステムは図書館の目録業務や文献複写業務に利用することを前提に設計されており、検索もコード類や書名、著者名、出版社名などを主な対象としている。このため、書名中や注記中の自由語検索などへの対応は不十分であり、また、所蔵図書館名の部分検索などもできない。そこで、エンドユーザの利用性を高めるために、インデクスはCATとは別にIR専用のものを構築することを前提とした。従って、目録系IRシステムの設計に当たっては、主にデータ本体（テキスト部分）の構成について検討を行った。検討に当たっては以下のような事項を考慮した。

(1) エンドユーザ用のデータビューの提供

総合目録DBは書誌、所蔵などのレコードが独立した構造となっているが、IRでは書誌・所蔵を統合して検索できるようにする必要があり、IR用のデータビューを提供する必要がある。

IR用のデータビューの概念を図3に示す。所蔵は書誌に包含する形で統合され、任意の組合せで検索可能である。典拠は書誌とは独立しており、検索も独立に行われる。書誌と典拠を組み合わせる場合は、まず典拠を検索してIDを取得し、これと書誌情報を組み合わせて検索を行うことになる。

(2) 適正な応答時間の確保

総合目録DB、参照目録DBともかなりの大きさのDBであり、ユーザが満足できる応答時間を確保できる構成にする必要がある。平均的な負荷での通常の検索応答時間が2秒以内となることを目標としている。

(3) DBサーバへの負荷抑制

CATへの性能上の影響を避けるため、DBサーバへの負荷が過大にならないよう、特に数秒以上にわたってバースト的な負荷が発生しないようにする。

(4) 更新処理の効率化

オンラインでの更新をできるだけ速やかにIRに反映してユーザに最新の情報を提供する。また、データベースの更新時間を短縮することによりシステム運用の負担を減らす。

(5) データベースの重複の抑制

DBを重複して持つことにより発生する記憶容量の増加と保守のコストを抑制する。

特に、(1)と(2)はユーザの使い勝手に直接関係する部分であるので必須の条件である。また、(3)は図書館業務の効率に影響するため重要な条件である。(4)はユーザなどへの直接的な影響はあまりないので(1)、(2)、(3)ほどでないが、運用面から見るとやはり重要な条件である。(5)はディスク資源に関する限りは、他の

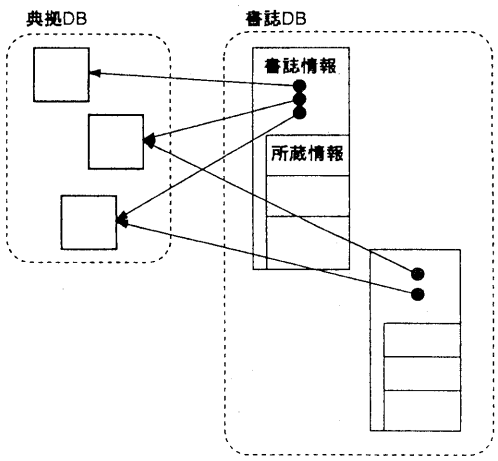


図3. IR用データビュー

条件と比べると追加投資で解決可能であるので、あまり強い条件ではない。

3.3 性能と容量の検討

新IRシステムでは、自由語検索や階層を持つデータの検索への対応や、全文DB検索システムの実現のために、OpenText6をエンジンとして使用する。そこで、ここではOpenText6のメタフィルター機能とオンライン更新機能を用いてシステムを構築することを前提とする。

メタフィルター機能では、ファイルやRDBに格納されたテキストデータに対する検索エンジンからの仮想的なアクセスを提供するため、物理的なデータの記憶システムに依存することなく全文のインデクスを構築し検索を行うことができる。また、データ形式の変換も可能であるため、CATのレコード形式をIRに適した形式（新IRではSGML風のタグ形式）に動的に変換したり、書誌と所蔵のレコードを統合してユーザに適したデータビューを作ったりすることができる。

また、オンライン更新機能では、ファイルやDBのレコードなど、記憶システムにおける論理的な格納単位（オブジェクト）ごとに、インデクスを更新することが可能である。トリガを与えるとそのオブジェクト中のテキストに対するインデクスを作成し、メモリ上のインデクスバッファを即時更新する一方で、これを追いかけながらインデクスファイルの更新を並行して行う。

これらの機能を用いることにより、DBの共用を図りつつ即時に近いIR用DBの更新を実現するためのいくつかの方式が可能となる。

今回検討を行ったデータベース連携方式は以下の3種類である。これらの概念を図4に示す。

(a) 完全共有型

IRのサービスはすべてオンラインのDBを直接使って行う。必要なデータビューの変換はアクセス時に動的に行う。

(b) 複製レコード型

オンラインDBからレコード単位の複製を作り、IRサービスはこの複製を使って行う。必要なデータビューの変換は複製時に行う。つまり、各書誌とそれに付随する所蔵は統合して1レコードとして格納する。

(c) 複製ファイル型

CATのDBのダンプファイルを作成してIR専用のDBを作成する。必要なデータビューの変換は複製時に行う。すべての書誌と所蔵レコードを統合して1ファイルとする。

結論としては、総合目録系DBでは(b)を、参

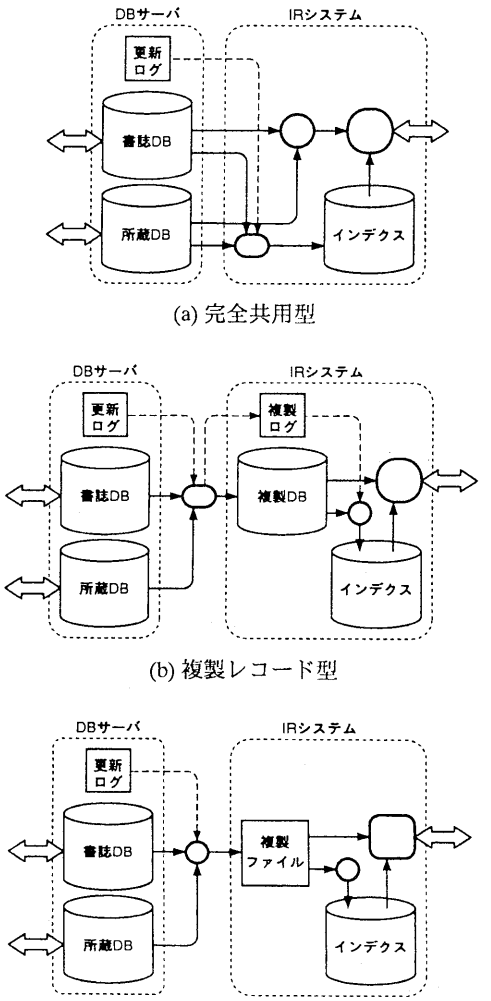


図4. DB連携方式の種類

照目録系DBでは(a)を採用することとした。なお、(b)では複製を格納する記憶システムとしてDBMSとファイルシステムがありうる。当面は保守の容易さを考慮し、DBMSを使うことにしているが、実装が進むに従い変更する可能性もある。

以下ではこの結論にいたる議論として、これらの各方式について、DBサーバへのアクセス回数、更新処理時間、応答時間、ディスク容量に関して行った検討について述べる。

(1) DBサーバアクセス回数

DBサーバへアクセスする局面としては、DB初期作成、インデクス再作成、レコード更新、レコード検索がある。

DB初期作成

書誌、所蔵、典拠の全レコード数をそれぞれN_B, N_H, N_Aとすると、いずれの方式でも全レコードに1回ずつ、すなわち、$(N_B+N_H+N_A)$回のアクセスが発生する。現在の総合目録DBではN_B=3,615,000, N_H=24,353,000, N_A=884,000であるので、これは2,885万回になる。参照目録DBではN_B=10,055,000, N_H=0, N_A=3,342,000であるので1,340万回になる。

インデクス再作成

(a)は初期作成と同じで$(N_B+N_H+N_A)$回、(b)、(c)は複製があるので0回のアクセスが発生する。総合目録DBでは(a)は2,885万回になり、現実的ではない。参照目録DBでは(a)は1,340万回になり、日常的に行うのは不可能であるが、事前に十分計画しておけば不可能な規模ではない。現実的には、DBの定義やインデクスの正規化ルールの変更などを行うときに必要となる程度で、頻繁に行うことではない。

レコード更新

総合目録DBの1日あたりの書誌、所蔵、典拠の更新レコード数の最大値をそれぞれU_B, U_H, U_Aとすると、(a)では書誌、所蔵では1レコードの更新につき、関連する書誌と所蔵のレコードにアクセスし、典拠では更新のあったレコードだけにアクセスする。1書誌に関連する所蔵数の平均は書誌と所蔵の全レコード数の比に等しいのでN_H/N_Bである。従って、この場合は$((1+N_H/N_B)(U_B+U_H)+U_A)$回のアクセスが発生し、過去1年間の統計では$U_B$=6,000, U_H=56,000, U_A=1,000程度であるので、最大51万回程度になる。(b)、(c)では複製のうち、更新のあったレコードの部分のみを入れ換えればよいので$(U_B+U_H+U_A)$回になり、現状では最大63,000回程度になる。

参照目録DBでは定期的にバッチ更新を行っており、最も更新レコード数の多いLC-MARCで1回あたり最大で31,000レコード程度である。所蔵データはないので、(a)、(b)、(c)いずれでも更新には最大で31,000回程度のアクセスが必要になる。

レコード検索

通常の検索対象となるIRのレコードには書誌と所蔵が含まれているため、1レコード表示あたり(a)では$(1+N_H/N_B)$回のアクセスが発生する。(b)、(c)では複製があるのでDBサーバへのアクセスは0回である。ピーク時のレコード表示頻度を10レコード/sと仮定すると、総合目録DBでは(a)は77アクセス/sとなり、CATサービスでのピーク時のアクセス頻度が200~225回/s程度であるのに比べて、無視できない大きさとなる。参照目録DBでは(a)は10アクセス/sとなり、ほぼ無視できる。

(2) 更新処理時間

オンラインDBにレコード更新があったとき、(a)、(b)ではOpenText6のオンライン更新機能を使って逐次更新が可能である。更新時間は主にデータアクセス時間とインデクス作成時間に分けられる。典拠レコードの更新頻度は小さいので無視すると、1更新レコードあたりのデータアクセス時間は、レコードアクセス時間をTとして、(a)では$(1+N_H/N_B)T$、(b)ではTである。インデクス作成時間は更新レコード数に依存する部分と固定部分があるが、総合目録DBでは更新間隔が比較的小さいので更新レコード数も少なく、固定部分の方が大きな割合を占め

ると予想される。原稿執筆時点でまだOpenText6のオンライン更新機能が提供されていないのでインデクス作成時間を正確に評価することができないが、諸般の状況から推定すると10分以内には収まると期待できる。更新間隔を10分と仮定すると、更新レコード数はピークで2,000個程度と見積もられ、T=100ms程度であることから、データアクセス時間は(a)では1,500秒程度、(b)では200秒程度となり、(a)では多重処理をしない限り処理不可能であるが(b)では逐次処理で十分処理可能である。

一方、(c)ではオンライン更新機能は利用できず、インデクスの再作成となり、最大で20時間程度かかると見込まれる。

(3) 検索応答時間

IRでのサーチ自体はインデクス上で完結するため通常のDBと同等であり、問題とはならない。データの表示のときに(a)、(b)ではレコードへのアクセスが発生するため応答時間が問題となるが、(c)は通常のDBと同等であり十分高速である。(a)では1件あたりの表示に$(1+N_H/N_B)$回のDBサーバへのアクセスが発生し、総合目録DBでは770ms程度かかる。通常は10件程度をまとめて表示させるので7.7秒程度かかることになり、許容できない。DBACを改修して複数の所蔵レコードをまとめてフェッチできるようにすれば、1件あたりの時間を300ms程度まで短縮することは可能であると考えられるが、それでも3秒程度かかるので十分とは言いがたい。参照目録DBでは1件あたり100ms程度であり、10件まとめて表示しても1秒であり許容範囲である。(b)では1件あたり1回のアクセスであるので100ms程度となり、10件まとめて表示しても1秒であり許容範囲である。

(4) ディスク容量

インデクスは(a)、(b)、(c)いずれでも同じであるので、本体データの複製に関する容量増加が検討の対象となる。(a)は複製を作らないので容量増加はない。(b)はレコード単位で複製を作成するため、DBMSやファイルシステムのブロックサイズに満たない部分が無効となる。CATのDBの状況から推定すると1レコードあたりの無効容量は300バイト程度となる。これに本体のデータ容量を加えると、総合目録DBでは7.5GB程度、参照目録DBでは11GB程度となる。(c)ではブロック化による無効容量は発生しないので、総合目録DBでは6GB程度、参照目録DBでは7GB程度となる。しかし、サービスと並行して更新を行うためには新旧のデータとインデクスを二重に持つ必要があり、実際にはこの2倍程度の容量がさらに必要になる。(b)、(c)のいずれにしても、容量の増加自体はあまり重大な問題ではない。むしろ、バックアップなどの保守のコストの方が問題になりそうである。

3.4 その他の検討

前節ではおもに性能や記憶容量に関する検討を行ったが、この他にもいくつか検討を要する点がある。

第一に、CATからのオンラインでのデータ更新とIRのインデクス更新の時間差による不整合の問題がある。何らかの理由でインデクス更新が中断すると、(a)ではその間の更新データとインデクスの間に大きな不整合が発生する。(b)ではインデクス更新処理による遅延時間分の更新データに不整合が発生するだけで、比率は許容できる程度に小さい。(c)ではデータとインデクスの不整合は発生しないが、切り替えのタイミングに関する問題が発生する。これらのことから(b)が望ましいことになる。ただし、参照目録DBは定期更新であることからインデクスの更新とタイミングを合わせることが比較的容易であり(a)でも不都合はあまりない。

第二に、CATの保守のためにDBサーバを停止する場合のIRサービスへの影響の問題がある。このような場合、(b)、(c)では複製を用いてIRサービスを行っているので問題ないが、(a)ではデータ本体をDBサーバに依存しているのでIRサービスも停止せざるを得ない。CATの実際の運用がどうなるかはまだ明確ではなく、

DBサーバの停止が頻発するようであれば参照目録DBでも(a)の採用をあきらめざるを得なくなるであろう。

第三に、性能上の問題がない限り、IRからのアクセスインタフェースは総合目録DBと参照目録DBで同一であることが望ましい。つまり、データビュー、タグ形式、記憶システムとアクセスメソッドが同一であることが望ましいことになる。

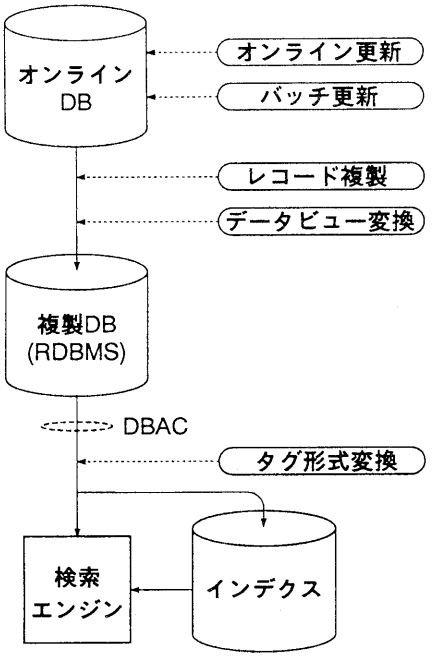
3.5 IRシステムの構成

前節の検討をふまえて決定した目録系DBのIRシステムの構成を図5に示す。

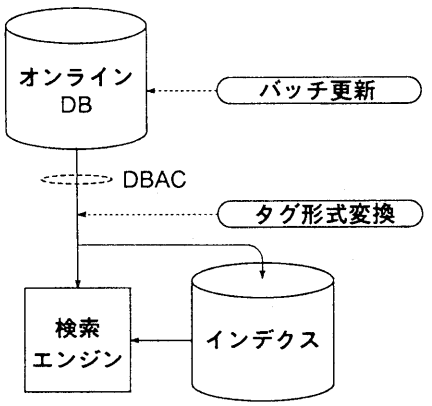
原稿執筆時点ではまだOpenText6のメタフィルターとオンライン更新の機能・性能上の詳細に不明な点があり、この構成で実用システムが実現可能かどうかの最終的判断はまだできないので、いくつかの代替案も念頭におきながら詳細設計とシステム開発を進めている。

4. おわりに

本稿では1996年8月を目標に開発を進めている目録系DBのIRシステムの構成について、オンラインDBシステムであるCATとの運用上の関連、DBサーバへの性能上の影響、IRサービスの利用性などを中心に行った検討について述べた。今後、上に示した構成に基づき、開発と実験による評価を行ってゆく。作業が進むにつれ設計の見直しは避けられないであろうが、ユーザの利用性を最大限に確保しながらサービスシステムを実現したいと考えている。実際のシステムの構成と評価については後日、機会を得てご報告するする予定である。



(a) 総合目録DB



(b) 参照目録DB

図5. 目録系DBのIRシステム構成

参考文献

[1] 神門典子, 木村優, 志津田嘉康, 大山敬三, 越塚美加, 小山照夫: NACSIS-IRの検索機能の高度化, 情報処理学会研究報告（情報学基礎研究会）95-FI-39, Vol.95, No.87, p.57-64(1995).

[2] 安達淳: 画面志向アプリケーション向けネットワークプロトコルの開発, 東京大学文献情報センター紀要, Vol.1, p.86-129(1985).

[3] 安達淳, 橋爪宏達, 大山敬三: TSS接続による仮想画面転送(VTSS)方式, 学術情報センター紀要, Vol.1, p.73-89(1987).