データベースシステム 112－3
（１９９７．５．１３）

医薬品・化学物質情報を例としたWebとデータベースの連携検索

灘岡陽子、石川恵司*、中野達也**、五十嵐貴子**、神沼二真**

（財）東京都臨床医学総合研究所、*石川電気、**国立衛生試験所

現在、インターネット上には膨大な情報が雑然と公開されている。従って関連した情報も分散して存在しているので、相互に関連づけてサーチするのは大変難しい。そこで、関係するWeb上のコンテンツやデータベース群を検索しその結果を統合して提供するシステムへの必要性が増大している。われわれは、医薬品や一般化学物質など化合物を対象として、インターネット上の関連情報を相互参照できるシステムを開発している。これまでに実験したのは、Web上のコンテンツをHarvestを用いて収集、検索するシステムと、Webとデータベースの連携機能を用いてデータベース群の統合検索を行うシステムである。いずれも化学物質に関する情報を対象とし、データベース群の共通キーとしてはCAS番号を用いている。

Development of An Integrated Web and Database Search System for Drug and Chemical Information

Yoko Nadaoka, Keiji Ishikawa*, Tatsuya Nakano**, Takako Igarashi**, Tsuguchika Kaminuma**

The Tokyo Metropolitan Institute of Medical Science,
*Ishikawa Electronic Company ,
**National Institute of Health Sciences

It becomes increasing important to devise tools that search and link rapidly expanding information contents on the Internet. We have developed two basic tools which enable one to take interesting information from the Web space, and to search distributed databases on the Internet. The former tool was constructed using Harvest, an information collection software developed at the University of Colorado. The latter tool is based on WWW and database connection interface. The system has been implemented for cross-searching databases that have at least one common retrieval key. Presently the key is CAS registry number.

1．　初めに

近年、インターネットを介して提供されるデータは日々増大しており、その巨大な情報空間から効率よく必要な情報を入手することは困難になってきている。ネットワーク上に分散して存在している、お互いに関連した内容を有するさまざまなデータベースを連結して統合的に参照することは、さらに困難である。Webコンテンツの情報を検索する一般的なツールとしては、Yahoo、Infoseek、AltaVistaなどが知られているが、これらは、単純なキーワードでヒットする膨大なURLから目指すデータを自分で選択して探さざるを得ず、非常に効率が悪い。一方、データベースに関しては、これまで各データベースを一ケ所（一つのマシン）に集め、各データベースに共通に含まれているデータ項目を頼りにして、連結検索を可能にする方法が取られてきた。化合物の分野でこの方式を採用していたのは、1970年代にNIH/EPAの共同プロジェクトとして開発されたCIS(Chemical Information System)である。

現在でもこの方式によって各種の化学物質データベースを統合して巨大な複合データベース（メガデータベース）を構築しようという構想があるが、統合する側に大きな負担がかかること、データの更新に対応することが難しいという大きな欠点がある。

これに対して、我々は、普及が著しいインターネットとWWWを利用するシステムを開発した。これは、分散して管理サービスされているデータベースを、そのままの形態で案内機能によって連結、統合するものであり、開発コストが少なくてすむ、更新が楽である、などの長所がある。また、 Webコンテンツの検索効率を上げるために、化学物質・医薬品に特化した分野における独自のサーチエンジンを作成した。以下では、これらのシステムについて報告する。

2．現状

生命科学分野では、いわゆるファクトデータがインターネット上に自由に入手できるように置かれていた。ただし昔は、データベースに収められていたデータや情報、知識が、今ではWWW、データベース、FTPサイトなど、さまざまな場所にさまざまな形で置かれるようになった。これらは、すべてブラウザーで見ることが可能である。けれども、研究の素材として選択的に収集、編集するには、これでは不便である。さらに、それぞれのサイトが工夫をこらしているだけ、必要なデータや知識を探し出し、ローカルなマシンに移すには、かなりの技術が必要である。以下に2、3の例を挙げる。

Chem Finderは、WWWから分子式、沸点、凝固点、CAS番号、部分構造などで検索できる化学物質の物性データベースを持っている。さらに、検索結果が1エントリに絞れた段階で、実際のデータを表示するとともに、追加情報として、関連情報を提供している他のサイトのリストを提示している。しかし、単にサイトのURLでしかなく、そのサイトへ飛んだ後、改めて検索する必要がある。

ドイツのエルランゲン大学のグループが開発しているThe WWW Chemical Structure Databaseは、サイバースペースを探索して、２０００以上の化合物に関する構造情報を集めて編集したもので、ユーザは、全構造、部分構造検索が可能である。構造情報に特化して成功した例である。

国立衛生試験所（国立衛試）では、化学物質の安全性や毒性、医薬品に関するデータをパソコンシステムから、UNIXワークステーションのRDBに移植し、Webから検索に対応できるようなシステムを構築した。これらのデータベース間では、CAS番号や物質名で統一的に利用可能な環境を実現している。さらに、Web上で公開されている関連情報サイトを積極的に網羅したページを作成しており、

それらは、組織別、分野別、アウトプット別に分類されている。しかし、個別の物質（化合物）に関する種々の情報は、それぞれのサイトで別個に検索を行わなければならない。

そこで、衛試内で所有するデータベースに対して統合化を図ったように、インターネット上に分散し、個別に管理されているデータベースも合わせて同時検索できる環境を作成することにした。

3．Webページの検索エンジンの作成

3．1　Harvest

基本となるソフトウェアとして、コロラド大学で開発されたHarvestを用いた。これは、図1に示すような4モジュールから構成されるサーチエンジンで、次のような特徴を持つ。

1）ネットワークトラフィックに負荷をかけないため、Internet上に複数のGathererを分散させたり、Brokerをcascadeさせて他のBrokerから情報を検索することが可能である。

2）defaultのindex/search engineとしてGlimpseを使用しているが、WAISなど他のindexerも使用可能である。

3）HTML, SGMLドキュメントだけでなく、Postscript、LaTeXなどのフラットファイル、tar、sharなどのアーカイブファイル、uuencode、gzipなどの変換プログラムを必要とするファイルなどを自動認識し、タイプに応じてサマライズするが、これらの各ステップをユーザがカスタマイズして、必要とするデータのみをデータベース化することが可能である。

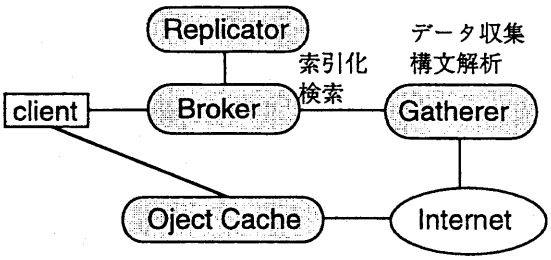

図1　harvestのソフトウェア構成

ただし、現在は、英語ないし、1バイトコード言語にのみ対応している。

3．2　対象サイト

国立衛試で収集分類している化学物質の安全性および医薬品に関する412サイト（URL）を対象として、そこで提供されている情報を元に、化学物質の安全性に関する検索エンジンを作成した。検索条件は、指定したURLと同一ホスト内へのリンクのみを辿ることとし、他のホストへのハイパーリンクで情報収集を停止するよう設定した。

3．3　結果

指定した412サイトから同一ホスト内にリンクを張られているファイル約83,000件のうち、情報のメンテナンスの不備で実体とハイパーリンクが対応しなかったか、なんらかの理由でアクセス不能のファイルが約5,000件あった。全文検索を行うために、一旦すべてのファイルをコピーし、その後インデキシングを行うが、これに約4日半かかった。

このような収集状況で試作したサーチエンジンを用いて、化学物質名、組織名、単語などで検索してみた。図2にquery画面、キーワードchloroformの検索結果を示す。chloroformでヒットした24サイトのうち、クロロフォルム自身のデータは3件だけで、あとはクロロフォルムへの溶解度やクロロフォルム臭といった他の化合物の記述であった。

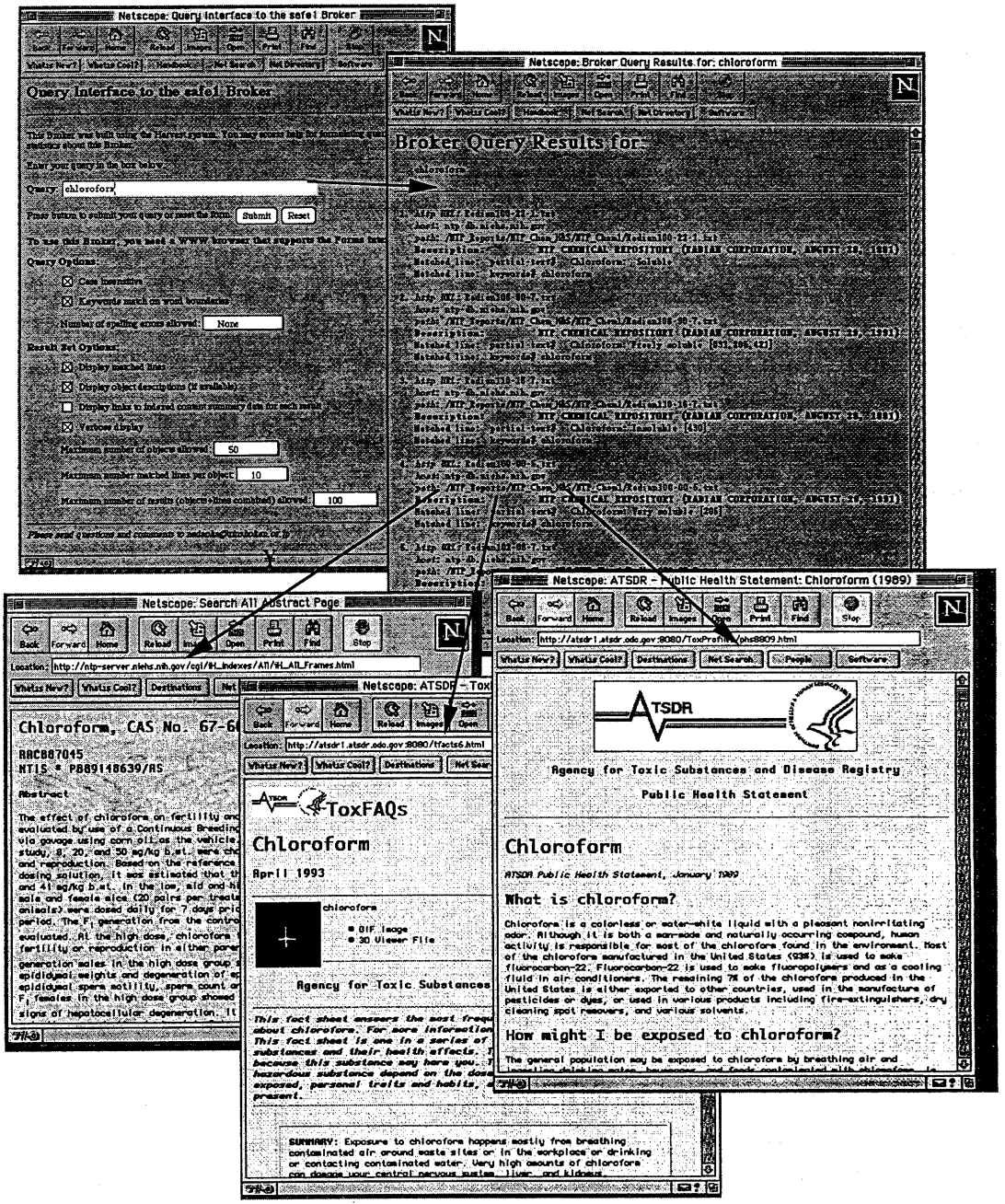

図2　検索画面

4　データベースの統合検索

4．1　サイトの選択

まず、化学物質情報を提供している各データベースの形式を調査するため、これまで収集した関連サイトのうち、重要度の高いサイトとして列挙されている22（URL）ヶ所を辿って約１００件のデー

タベースを調べた。その結果、統合検索の共通キーと採用したCAS番号をデータとして含む２３件のデータベース群を対象とした。それらは以下の４つのグループに分類される。第１は、国立衛試(NIHS)と厚生省（MHW）のデータベースである。これらは、データベース管理者と直接連絡がとれる利点がある。残りは全く独立して管理運営されている我々にとっては外部のデータベース群である。第２のグループは、データがテキストファイルとして提供されているだけでなく、ディレクトリサービスのような形式で、物質名の頭文字から目的の物質情報を選択できるメニューが追加されているサイトである。第３は、テキストファイルとCGIプログラムを介してキーワードによる検索機能も追加されているサイトである。第４としてデータベース管理システム上に登録してあるデータをCGIプログラムを介して検索するサイトから、CAS番号で検索可能で、かつ重要と思われる４件である。以下では、これらをサイトデータベースと呼ぶ。

４．２　統合検索用テーブルの作成

我々が開発した統合検索システムは、検索のための基本情報として、検索対象となるサイトデータベースに関する情報、化合物、および化合物とサイトデータベースとのリンク情報のテーブルを持っている。図３と４に、作成した４個のテーブルと、その関係を示す。これらのテーブルはSybase 10 にインストールした。

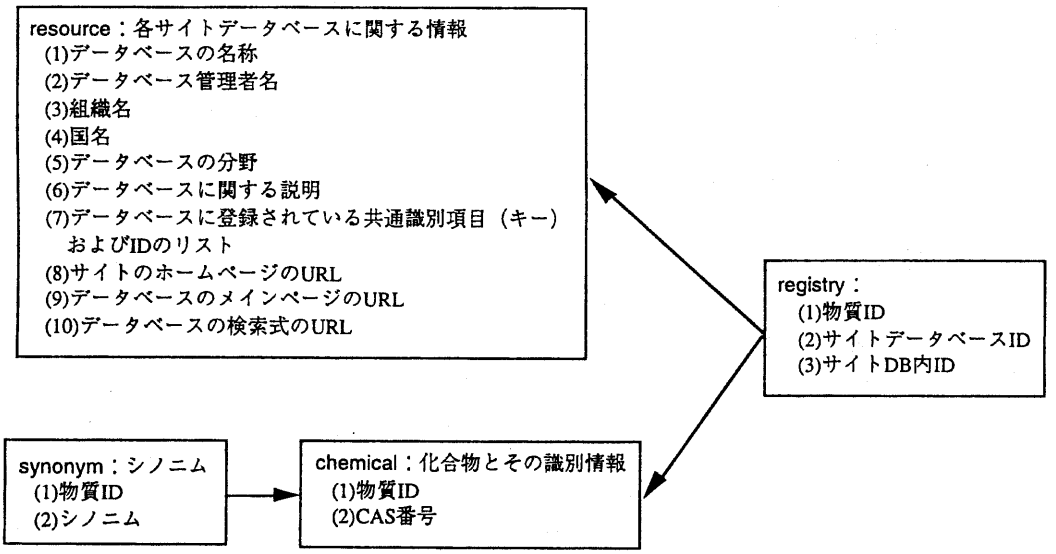

図３　テーブル構造

４．３　CGIプログラム

検索キーは化学物質名かCAS番号のいずれかである。また、入力文字列のマッチングは完全一致、部分文字列一致、前方一致、後方一致のいずれかを選択することができる。ブラウザから入力されたキーワードは、相当するSQL文に変換され、対象となるデータベースへの問い合わせが行われ、その回答はHTML文にしてWebサーバに返される。

この処理の流れを図５に示す。CAS番号が特定できた時点で、registryテーブルから、その物質情報

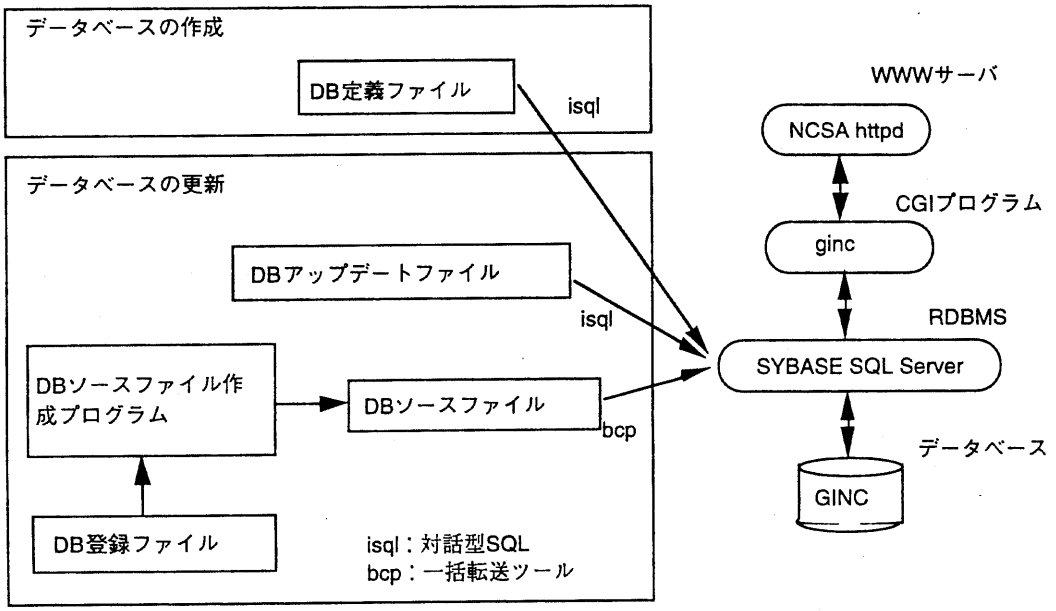

図4　統合検索システムの概念図

を含むサイトデータベースの検索式を作成する。従って、その物質の情報を持たないデータベースは表示されない。ただし、第4のデータベース群に対しては、該当物質情報の有無にかかわらず、すべてのデータベース名を表示することにする。

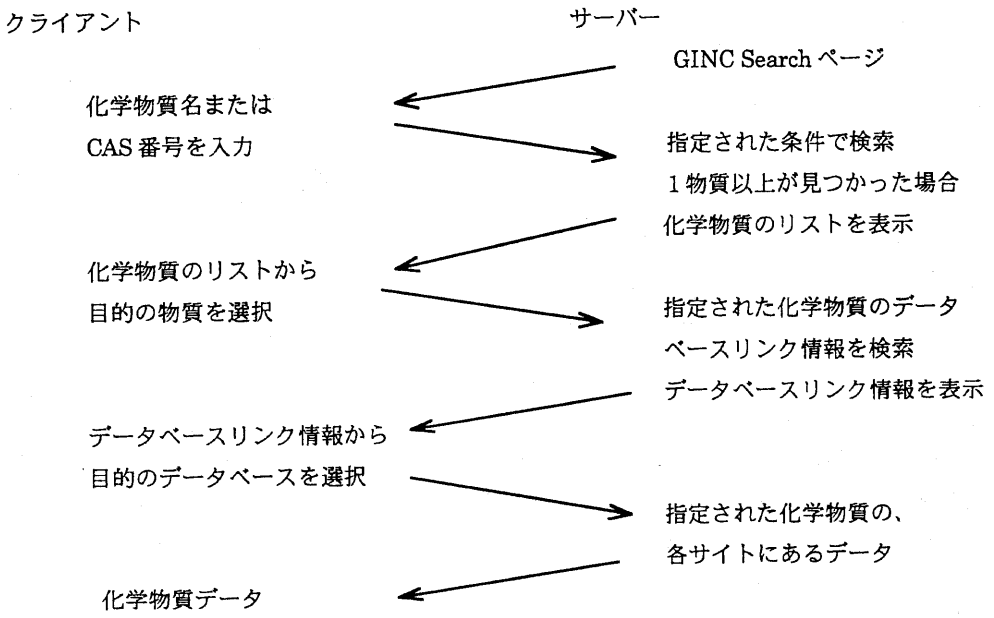

図5　問い合わせフロー

4．4　結果

現在のシステムはデータベースシステムとしてSunSparcStation 20 （OS: Soralis 5.1)、データベースサーバとしてSYBASE SQL Server、Webサーバとしては NCSA httpd（version 1.4.2）を用いている。当初は、Web サーバとしてApacheを利用していたが、CGIプログラムが正常に動作しなかった。実行画面を図6に示す。

本システムは1月末に実験を開始したが、アクセスログを調べると、3月における検索ページへのアクセスは557件、検索実行回数は1060件であった。なお、4月は16日までで12,655件の検索が実行されている。

5．考察

サーチエンジンの設定に不備があり、指定したサイトすべてから情報を収集できてないが、InfoseekやAlta Vistaのような一般的なサーチエンジンと比較し、化合物データに限定された情報だけがヒットするため、短時間で目指す情報が入手可能である。さらに必要な情報を網羅していくためには収集対象の選択が重要である。現行の設定はすべてのサイトに対して同一サーバに限定して情報を収集しているが、ESMGのように関連URLをまとめてあるようなサイトからは、ハイパーリンクされている他のホストも対象とする必要がある。また、Chem Finderのように、Webのページでユーザから関連サイト情報を募る方法や、専門家の協力のもとに、使用Browserのアクセスログから有益なURLを抽出する方法を検討していく予定である。情報収集時には、それぞれのサイトへのアクセスが集中しサーバに負荷をかけるため、CPUの空いている時間帯に収集したり、あるいは収集先に許可を求めることなど検討していかなければならない。

化学物質データベース統合検索システムによって、複数のサイトから提供されている同一物質に関する種々の情報を、サイトを意識せず、容易に入手することが可能となった。また、CAS番号や物質名で複数のデータベースに登録されているデーター覧が一度の検索で得られ、効率よくデータが獲得できるようになった。現在の方式ではキーとしてCAS番号を用いるが、この情報が欠如しているデータベースも扱えるようにすることが次の課題である。また現在の方式では、多くのサイトのデータベースを登録した場合、大量の不必要な情報まで表示されてしまうので、サイトデータベースの分野、国別情報などをユーザが指定できるようなシステムに改良する必要がある。

現在の統合検索システムは、インターネット上に分散して存在している個別データベースとは独立に機能する。ただ、実際問題として統合の対象となる各データベース（管理者）が自らのデータベース項目を自主的に統合検索システム側に登録したり、その内容を更新したりした方が、後者の運用はスムーズである。さらに、それぞれのデータベースのデータ項目や内容、情報の記述方式、用語、単位などに整合性をもたせた方が、使いやすく、自動処理にも向いている。この意味で、統合の対象となる個別のデータベースの提供者、開発者と統合検索システムの提供者、開発者とはなんらかの協力関係があるのが望ましい。

以上、本システムの有用性は実験によって証明されているので、現在、化学物質だけでなく、医薬品や分子生物学の分野における同様なシステムへの拡張を検討している。

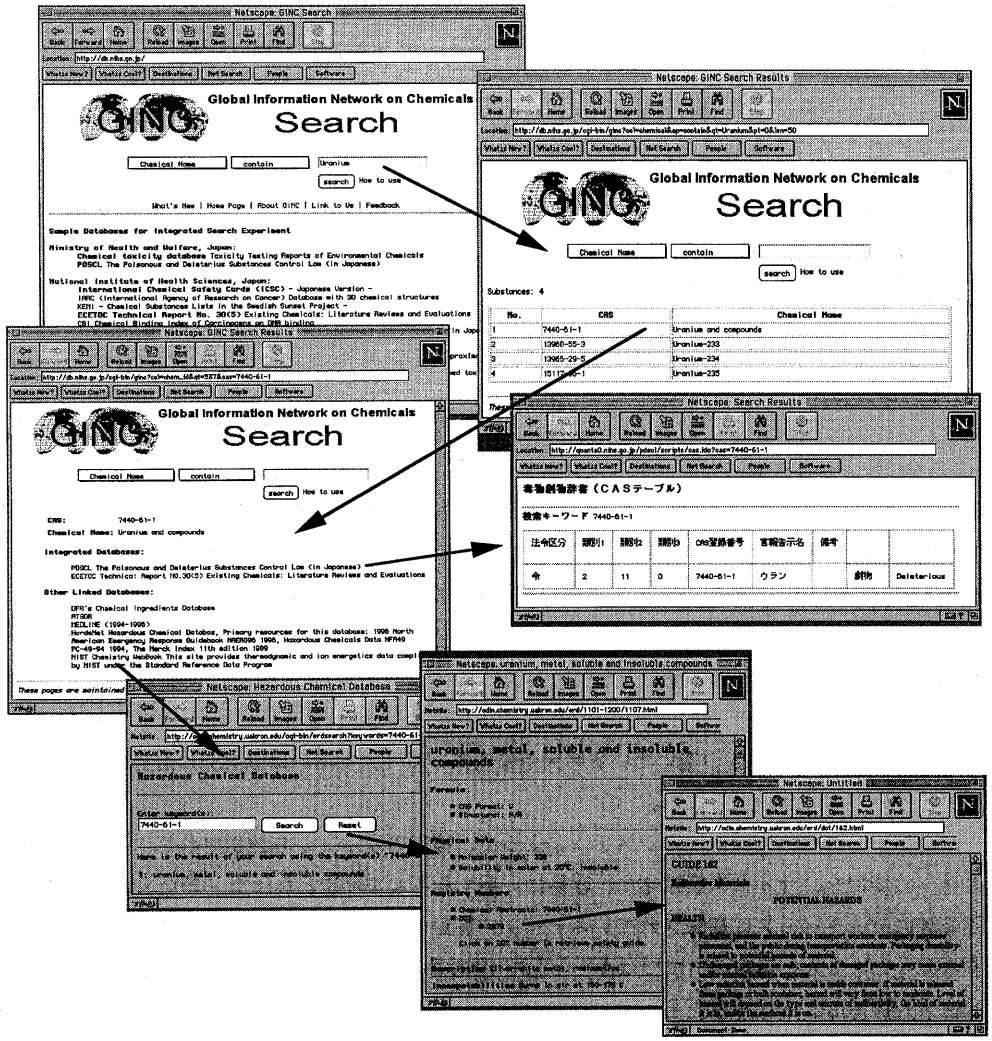

図6　検索画面

参考文献

http://harvest.cs.colorado.edu/

日経データプロ、WWWーデータベース連携システム構築法、日経BP社、1996

L.Buhle,M.Peace,V.Kumar,et al., Webmaster's Professional reference, New Readers Publishing Indianapolis, 1996

J.Rowe., Webmaster's Building Internet Database Servers with CGI, New Readers Publishing Indianapolis, 1996

UNEP Chemicals, Internet on Chemicals, United Nations, 1996

Windows NT World, 1996年10月号 p79

キーワード

化合物データベース、Harvest、サーチエンジン、統合検索システム、WWW、CGI

Chemical substances database、Harvest、Search Engine、Integrated Web and Database Search System、WWW、CGI