「XMLに基づくコンピュータ処理可能な構造化デジタルオブジェクト」の概念による Web リソースの検索システムの研究と開発

Research and Development of a Web Retrieval System on the Concept of “Computer-Processable Structured Digital Object based on XML”

李 穎* 石塚英弘*

Ying LI and Hidehiro ISHIZUKA

膨大な学術情報を如何に表現し、提供し、また如何なる検索機能を提供すれば良いか。これは古くかつ新しい課題である。これまで種々の手法が開発されてきたが、それぞれに欠点があり、研究者、すなわち、特定の情報源に限定されることなく、広く情報を調査し、よりノイズの少ない状態で有用な情報を得て、得られた結果を自らの観点から再構成しようとする利用者にとって、満足できる状況とは言いがたい。これを解決するためには、新たな概念に基づく方法論が必要であると考え。本研究では、「XMLに基づくコンピュータ処理可能な構造化デジタルオブジェクト」、すなわち、情報の最小単位を構成要素とし、XMLを用いて表現する構造化情報表現を考え、その概念に基づいて、Web リソースの検索システムのモデルを考案した。そして、典型的なファクトデータベースである Merck Index のデータを例として、プロトタイプを開発した。そのデータ構造は XML schema を用いて表現し、ASP.NET 上に実現した。

How is huge scientific information described and serviced? What retrieval functions should be provided? These are old and new problems to be solved. So far various methods have been developed, however each of them includes some defects. The situation is not satisfying for the researcher's needs who wants to investigate information widely not restricted to the particular resource, to get valuable information with less noise, and to reorganize obtained information according to his/her view. A new methodology on new concept is necessary to resolve the situation. We proposed concept of “Computer-Processable Structured Digital Object based on XML (Extensible Markup Language): (CPSDO/XML)” that is composed of elementary units of information, and designed an information retrieval system based on CPSDO/XML for Web resources. We also developed a prototype for Merck Index data, that is a typical fact database, on ASP.NET. The data structure is described with XML schema.

キーワード: Web 情報検索システムモデル, 検索対象, デジタルオブジェクト, XML, XML schema, XSLT, XPath, DOM, ASP.NET, Web information retrieval system model; retrieval target; digital object

1. はじめに

膨大な学術情報を如何に表現し、提供し、また如何なる検索機能を提供すれば良いか。これは古くかつ新しい課題である。これまで

種々の手法が開発されてきたが、それぞれに欠点があり、現時点では全ての問題点を解決できるシステムは見当たらない。そのため、さらに良い方法を求めて研究開発がなされている。我々もこのテーマを研究しており、本論文でその成果を報告する。

* 図書館情報大学

University of Library and Information Science

まず、これまで行われてきた方法、および最近提案された方法のうち代表的なものを以下に挙げ、それらの長所短所を述べる。

- (A) 文献情報データベースとその情報検索システム
- (B) ファクト・データベースとその検索システム
- (C) 引用関係を利用した文献検索システム
- (D) 人工知能のフレームによる情報知識表現
- (E) 多数のファクト・データベースと文献情報データベースを搭載した検索システム
- (F) Web-HTML による文書群と Web 検索エンジン
- (G) オブジェクトデータベースによる表現と検索
- (H) オンライン・ジャーナル
- (I) XML(Extensible Markup Language)
- (J) Topic Map^[1]
- (K) Semantic Web^{[2][3]}

(A) の典型例は、化学分野では CAS (Chemical Abstract Service)^[4] による抄録、事項索引、化学物質索引、書誌情報のデータベースとその情報検索システムであり、医学分野では米国国立医学図書館 (National Library of Medicine) による MEDLINE, Elsevier Science^[5] による EMBASE (Excerpta Medica database) などである。これらは有効で広く使用されたが、これらの方法には限界があった。すなわち、使用料金の高さ、原文献の出版とデータベース公開とのタイムラグ、検索結果と文献が電子的に接続していなかったために検索結果を基に文献を書棚から探し出す必要があること等である。最後の点は全文データベースあるいはオンライン・ジャーナルとの連携がその後実現して解消し、タイムラグは出版段階からデータを文献情報データベース作製機関に送信することによる改善が見られるものの、事項索引など高度な索引の場合にタイムラグを無くすことは困難である。また、料金の高さの問題は現在も未解決である。

(B) は整理され、信頼できる情報を得ようとする利用者のために作られた便利なデータ

ベースであるが、使用料金の高さとタイムラグが欠点である。また、ファクト・データベースはそれぞれの対象分野・主題があり、収録しているデータ項目、内容は微妙に異なる。そのため、利用者は自分が得たい情報が収録されているファクト・データベースはどれかを知る必要があり、また複数のファクト・データベースから得られる結果を利用者が手作業で整理・組み合わせなければならない。

(C) の典型例は米国 ISI^[6] による Science Citation Index であり、引用関係で検索できる点が特徴である。或る文献に引用された文献だけでなく、引用した文献も検索することができ、それを繰り返すことによって関連する文献群を得ることができる。ただし、引用索引データベースの編集はコストが掛かり、タイムラグも発生し、また利用料金もかなり高い。

(D) 人工知能の分野では情報知識の表現にフレームが用いられてきた。Minsky^[7] が提唱したフレームは、その中にスロットを持つことによって、知識を適度な大きさのまとまりとして表現し、また、スロットに他のフレームへのリンクを付加することによって知識を構造化する。スロットには付加手続きを付けることもできる。フレームの実装としては、KRL^[8], FRL^[9] 等々様々なシステムが作られたが、そのことは逆に夫々のシステムで記述された知識／情報データの交換利用が難しいことを示した。また、百科事典の知識を記述しようとした Lenat の CYC プロジェクト^[10] は大量の知識／情報を 1 箇所で作ることの難しさを浮彫にした。大量の場合は、分散して作製された知識／情報を互いに交換利用する方が実践的であるが、フレームではそれが困難なことが欠点である。

(E) の典型例は STN International^[11] である。ファクト・データベースだけでなく、文献情報データベースも在る点が特徴である。個々のデータベースを検索でき、また統合検索機能を用いれば、複数のデータベースを同時に検索することができるが、得られる結果は人間が見て分かる形式であり、利用者側の

ソフトウェアによってデータ項目ごとに、あるいはそれらを組み合わせて二次処理することは容易でない。また、使用料金の高さは依然として解消されていない。

(F) は、全文が得られる点、関連ページに跳ぶリンク機能がある点が有用である。しかし、HTML で記述されているため、(A)、(B)、(E) のようにデータ項目が区別されていない。また、Web 検索エンジンは HTML を対象としているため、データ項目を区別することができない。その結果、検索ノイズが頻発する。これは明らかな欠点であり、解決されていない。

(G) オブジェクトデータベースは、オブジェクト指向データベースあるいは object-based database のことであり、プログラミング言語の観点とデータベースの観点の 2 つの面がある。後者では、オブジェクト間の関係によってデータ構造をモデル化する。知識を構造化して表現する点では (D) のフレームの影響を受けている。オブジェクトデータベースの例としては、O2^[12] や ODMG 標準^[13] などがある。複雑な構造を持つ情報の場合は複合オブジェクトを用いることになるが、学術情報のように複雑な場合は複合オブジェクトを用いても容易ではない。オブジェクトデータベースの場合も処理系の間でのデータの互換性の問題が存在する。

(H) は全文が得られる点、リンク機能を用いて参照文献のページに跳ぶことができる点が有用である。この 2 点は (F) と同じである。さらに、(A) の文献情報データベースと連携すれば、データベースの検索結果から当該ジャーナル論文のページに跳ぶことができ、この点は (F) よりも便利である。もっとも、実際にそのページに跳ぶためには、そのページの著作権を自社で持つか、あるいは、著作権を持つ出版社あるいは学会との協定を結ぶ必要がある。それは「寡占状況における情報の囲い込み」という別の大問題を引き起こした。たとえば、Elsevier Science 社は自社のみで大部分の情報を提供しようとする戦略を持ち、ヨーロッパの老舗の出版社を次々買収して傘下に

収めた。その結果、同社が提供するサービス：Science Direct^[14] は圧倒的多数のオンライン・ジャーナルを以って様々な学問領域をカバーし、また、EMBASE も以前から持ち、前述の機能を実現している。これは、他のオンライン・ジャーナルを提供する出版社・学会にとっては脅威である。そのため、米国の大学会などは互いのページの参照を認め合う協定の締結を進めて対抗しているが、状況は厳しいようである。また、学術雑誌の価格の高騰が問題になっているが、その一因は寡占状況にあると推測されている。

(I) XML (Extensible Markup Language)^[15] は、情報の階層構造、同じ階層の上の要素間の順序関係、関連情報へのリンクなどが表現できることが特長である。XML は単純なデータだけでなく、ドキュメント、テキストにも適用できる。XML による表現は処理系に依存しないため、データ交換に適している。また、Web を想定して設計された仕様であるため、URI を介して、インターネット上に分散して存在する情報を取得することができる。また、スキーマを自分で設計でき、適用分野や適用対象に制限は無いが、このことは逆に分野や対象に適合したスキーマを設計する必要があることを示している。

(J) は、トピックマップの作成者が認識している概念体系に則って、情報リソースを分類、整理、管理、検索、ナビゲートする一つのパラダイムである。情報リソースとは独立したメタレイア（上位層）として、個々のトピックを互いに関連付けた一種のトピック間関連図を構成し、個々のトピックから関連する情報リソースに参照のリンクを張る。トピックには、情報リソースが持つ主題を選んでもよいし、概念を選んでもよい。トピック間関連図により、主題と概念の間の関係と概念間の関係を示す。情報リソースの種類は問わず、ドキュメントであってもよいし、データベースであってもよい。また、情報リソースには個々のトピックから単にその該当箇所へのリンクを張るのみであって、情報リソースを加工す

ることではない点が特徴である。しかし、このことは逆に得られる情報を加工・再構成する機能は無いことを示している。

(K) は、次世代の Web とされ、データの記述言語に HTML ではなく、データ項目を区別できる XML を用いること、データ項目間の関係を RDF (Resource Description Framework)^[16] で記述すること、特定の Web サイトに集中するのではなく、分散型であり、各 Web サイトで異なる用語を用いることを許し、Ontology を用いることによって用語の問題を解決しようと考えていること、検索には Web 検索エンジンではなく、エージェントを用いることなど興味深いパラダイムである。しかし、現時点では構想段階であって、多くの未解決の先進技術に依存するため、実装までは至っていない。

(A) から (K) に述べた現状は、研究者、すなわち、特定の情報源に限定されることなく、広く情報を調査し、よりノイズの少ない状態で有用な情報を得て、得られた結果を自らの観点から再構成しようとする利用者、にとって満足できる状況とは言いがたい。そのためには、新たな概念に基づく方法論が必要であると考える。

そこで 2002 年 7 月、我々は情報の構成単位として CPDO (Computer Processable Digital Object) の概念を考案し、利用者が検索によって得たい情報の対象、すなわち「検索のターゲット」は CPDO によって表現できることを述べ、また、様々な分野の情報が CPDO を基に再構成できる可能性を示し、CPDO の概念に基づくデジタル図書館システムのモデルを提案した^[17]。その後、我々は、CPDO の概念規定を改良して、CPSDO/XML (Computer-Processable Structured Digital Object based on XML) 「XML に基づくコンピュータ処理可能な構造化デジタルオブジェクト」を考案した。また、その概念に基づいて、Web リソースの、様々な情報リソースの典型例を採り上げて、その中のデータ項目の CPSDO/XML への変換、CPSDO/XML に依る検索システム

のプロトタイプの XML 処理系: ASP.NET^[18] 上での作成などを行い、前述の課題に対して一定の成果を得たので、本論文でそれを報告する。

2. CPDO および CPSDO/XML の概念またそれに基づく検索システムのモデル

本論文では、利用者が扱う学術情報の最小単位を CPDO (Computer Processable Digital Object) という。また、CPDO を構成要素とする構造化情報表現を CPSDO (Computer Processable Structured Digital Object) という。たとえば、「検索のターゲット」すなわち「利用者が検索によって得たい情報対象」が情報の最小単位であれば、それは CPDO であり、もしも、CPDO を組み合わせたものであれば、それは CPSDO の一部分である。

CPDO と CPSDO は XML によって記述されるデジタルオブジェクトであり、XML のタグ名称は記述内容を示す。XML で記述された CPSDO を CPSDO/XML (CPSDO based on XML) という。そのため、CPSDO/XML はソフトウェアによる処理が可能である。また、もしも検索対象の Web ページあるいはデータベースが CPSDO/XML に基づいて構築されるようになれば、CPSDO/XML の該当部分のみを検索するだけでよくなるため、ノイズの少ない効率的な情報検索が可能になる。

なお、CPDO および CPSDO で言うオブジェクトとは XML で言うオブジェクト (物あるいは対象とする構成要素) であって、メソッドは持っておらず、その点でオブジェクト指向で言うオブジェクトとは異なる。

我々の先の論文^[17]では、CPDO と CPSDO を区別せず、CPDO の中に CPSDO を含めていた。また、CPDO の粒度は小さくても良いし、大きくても良いとしていた。利用者が検索して得たい情報の粒度は様々だからである。検索対象として論文の場合を検討し、大きい方から小さくなる順に、ある主題 (たとえば、

炭疽菌に効く薬)に関する論文の集合, 同じく1つの論文, 論文の一部(たとえば, 薬の薬効を記述した部分)の集合, 同じく論文の一部, 語(たとえば, 薬の名称)の集合, 語などから成る階層構造を考えたが, 対象となる文献は論文以外に, 単行本, 辞書, データベース, Web ページなどがあり, そのため, 様々な粒度がありえることになる. たとえば, 炭疽菌に効く薬の名前を検索したい, その薬の製法を検索したい, 製法が書いてある文献の書誌情報が欲しい, あるいは, 製法が書いてある記述そのものを得たい, あるいは製法のエッセンスである化学反応式を得たい, あるいはそれらが記述された文献を得たい, 等々の要望がありえる. これらに対応する CPDO の粒度は異なる. しかし, 論文^[17]の考え方には分かり難い点があるため, 本論文では CPDO と CPSDO を区別することにした.

また, CPSDO を XML で記述することによって, CPSDO と CPDO の間に階層関係が存在する場合は, それを XML の階層関係を用いて表現できる. また, CPDO と CPSDO の間にリンク関係があれば, それを XML のリンク関係として表現することができる.

このように本論文では前論文の考え方をさらに進めて, 既存の情報リソースのデータ項目を XML 要素で記述した CPSDO/XML あるいはその構成要素である CPDO へ変換し, これにアクセスすることによって検索対象となった CPDO を取り出すことを可能にした. その結果, ①単一の CPDO による検索のみならず, ②複数の CPDO を組み合わせた検索, ③他のリソースの該当箇所への連鎖検索などが可能な検索システムのプロトタイプを作成した. 図1にそのシステムのモデルと機能を示す.

図1: CPDO および CPSDO/XML に基づく検索システムとその機能 (概念図)

図1のⅠFact DBは化学物質・医薬品の辞書またはファクトDBとして定評のあるMerck Index^{[19][20]}の各データ項目を、対応するCPDOに変換し、XMLで表現したものである。これはWebサーバ上のXML処理系を用いることにより、検索することができる。実例は本論文の中で後に詳しく述べる。また、Ⅱ一次文献は特許のDBのCPSDO/XML版であり、Ⅲ一次文献は雑誌文献のDBのCPSDO/XML版である。なお、ⅡとⅢは説明用の仮のものであって、実際に変換作成したものではないが、変換作成は原理的に可能であり、検索も可能である。

①1つのCPDOによる検索

例えば、抗生物質名“Ciprofloxacin”(炭疽菌に効く薬)の検索である。図1のⅠFact DBの該当するCPDOのXMLタグ: MonographTitleを指定して、その値が“Ciprofloxacin”であるものを検索すればよい。こうすることにより簡単にノイズレスで検索目標を達成できる。

②CPDOの組合せによる意味的な高度検索:

例えば、「Ciprofloxacinの製法」を検索したいときは、図1のⅠFact DBのCPDOのXMLタグ: MonographTitleの値が“Ciprofloxacin”であるものについて、その下のレベルのCPDOがSubject, その下のCPDOがPrepn.(preparation:製法)の下のCPDOがRef(参考文献)のものをアクセスすることによって、製法が書かれた文献の書誌情報を知ることができる。このようにCPDOを組み合わせることによって「Ciprofloxacinの製法」のような意味検索を実現することができる。

③CPDOの連鎖検索:

従来の情報検索システムでは、二次文献DBにより一次文献のIDを入手しても、両者は接続していないことが少なくなく、一次文献を得るには手間が掛かった。しかし、CPSDO/XMLに基づいた図1に示す検索システムが成立すれば、「Ciprofloxacinの製法」を記述した文献のXMLベースのDB、

すなわち、Ⅱ一次文献(特許)あるいはⅢ一次文献(雑誌)の該当文献へリンク機能を用いて直接行くことができる。これは連鎖検索と言えよう。また、一次文献中の参考文献リストのCPDOから他の一次文献の全文DBのCPSDO/XMLのCPDOにリンクすれば、検索の連鎖はさらにつながることになる。図1の◎と←→あるいは■と←→は、それを示している。

3. CPDO および CPSDO/XML に基づく検索システムの設計

図2の上半分に示すように、CPDOおよびCPSDO/XMLに基づく検索システムにおける「検索」とは、図2の④「CPSDO/XMLのXMLスキーマ(XML Schema^[21])」を用いて、XMLで記述された情報リソース、たとえば、Merck Indexから、⑤「検索ターゲットとなったCPDOを取り出すこと」である。この検索システムによって、ニーズの異なる様々な検索者にノイズの少ない検索結果を与えることが可能になる。

また、この検索システムを構築するには、まず、CPDOから構成される情報の構造(スキーマ)を設計し、次いで、各CPDOの値を既存の情報リソースの対応するデータ項目から得て、XML記述形式に変換すること(図2の下半分参照)が必要である。既存の様々な情報リソースのデータ項目は、CPDOのそれと一致するとは限らない。なぜなら、既存のリソースのデータ項目は人間(検索者)が見て理解できればよいから、ある程度まとまった単位の情報となっており、必ずしも情報の最小単位ではないからである。CPDOの考え方から言えば、2節で述べたように、CPDOは情報の最小単位であり、もしも、それより大きな単位が必要であれば、CPDOを組み合わせて大きな単位のCPDOを作ることになる。この考え方に依り、CPDOから構成される情報のスキーマを設計する。

各CPDOの値を既存の情報リソースの対応

図 2: CPDO および CPSDO/XML に基づく検索システムとその構築の概念図

するデータ項目から得て、XML 記述形式に変換するには、次に示す処理を行う。なお、具体例は次節に述べる。

①既存のリソースから、そのスキーマに従ってデータ項目を取り出す。

②取り出したデータ項目が CPDO と一致するか否かを判定し、YES ならば、そのスキーマの構成要素単位を流用して対応する CPDO の XML データに変換し、XML ベースのリ

ソースに格納する。

③判定結果が NO ならば、データを加工して対応すべき CPDO に一致させ、XML データに変換して該当する XML ベースのリソースに格納する。

4. CPDO および CPSDO/XML に基づく Web 検索システムの実装

我々は、既存のリソースの典型例として、薬学および化学の分野で最も広く利用されているリソースの一つである Merck Index^[19]を採り上げ、前節で述べた設計方針による実装を行った。以下に述べる実装方法は勿論、他の種類の既存リソースにも適用可能である。

4.1. 実装の概要

Merck Index のデータ項目を冊子体の記述を参考にしつつ、情報検索システム：DIALOG の Merck Index データベースのスキーマに依って分析した結果、DIALOG システムにおけるデータ項目の一部は、利用者が検索ターゲットとしたい情報の単位：CPDO に比べて大きいことが明らかになった。そこで、CPDO に比べて大きい単位のデータ項目は CPDO の単位に細分化し、CPDO と同じ単位のデータ項目はそのまま CPDO に対応させて、CPSDO/XML の XML スキーマを設計した。なお、本論文で設計した XML スキーマについては後で述べる。

次いで、既存のリソースのデータ項目の値を対応する CPDO の値に変換し、CPSDO/XML の XML DB を作成した。そして、この XML DB を持つ Web サーバを ASP.NET を用いて実装した。この検索システムは Web ブラウザ：Microsoft I.E. を用いて検索でき、検索者が必要とする CPDO を XML DB から取り出し、検索結果として表示する。

図 3 に前述のプロセスを示した。①は既存の Merck Index データベースを示す。②は既存の Merck Index のデータ項目を CPDO にマッピングした結果、必要に応じて細分化し、対応する CPDO に変換することを示す。③は CPDO を XML で記述して CPDO に基づく XML DB を作成し、それを持つ検索システムを ASP.NET を用いて構築したことを示す。④は、CPDO に基づく XML DB から検索者が必要とする CPDO を ASP.NET の機能を用いて取り出し、Web ブラウザにデータを送信し、表示することを示す。XML データは Web サーバ上で XSLT を利用して HTML に変換して Web ブラウザに送る。また、一部分の内容だけを表示するなどの処理が必要な場合は、その処理を VB Script を用いて記述した。

図 3: CPDO および CPSDO/XML に基づく Web 検索システムの実装のプロセス

4.2. 既存リソース／DBのデータ項目とCPDOの対応関係

以下、具体例を挙げて述べる。図4に、炭疽菌に効く抗生物質：CiprofloxacinのMerck Indexの冊子体の記述^[20]を示す。なお、冊子体ではデータ項目名は明示されていないため、凡例に則してデータ内容を記述したものを○で囲んで示してある。Monograph numberとはMerck Indexの中の物質を識別する番号であり、その右のTitleとは物質の名称、この場合はCiprofloxacinである。これらは情報の最小単位であり、CPDOに一致する。しかし、Literature referencesはChemical InformationとBiological, pharmacological, etc. informationの2つに分かれ、さらに前者は、「Prepn(製法): 参照文献の書誌事項; 参照文献の書誌事項.」, 「Antibacterial spectrum in vitro: 参照文献の書誌事項.」等々のように主題ごとに分かれるなど内部構造を持っており、情報の最小単位とは言えない。Merck Indexは情報検索システム：DIALOG^[19]にも収録されているが、そこでもLiterature referencesのデータ項目は大きな単位のままである。大きな単位のままでCPDOに対応させることもできるが、より小さい単位に分けてCPDOに対応させる方がCPDOおよびCPSDO/XMLの本来の趣旨に合っている。

図5はDIALOGに収録されているMerck Indexのデータ項目をそのままCPDOに対応させ、XMLで記述して作成したWeb検索システムの検索結果表示である。図5の中ほどの矢印で示した“LITERATURE REFERENCES”項目を見ると、全てのレファレンス文献を主題ごとには分けて、一つの項目として扱っていることが分かるであろう。Merck IndexではLiterature referencesはChemical informationとBiological, pharmacological, etc. informationの2つに分かれているが、DIALOGのThe Merck Index OnlineではLITERATURE REFERENCESのみであって分かれていない。この設計では、例えば、検索者は「炭疽菌に効く薬 Ciprofloxacin」の製法だ

けを調べたい場合でも、LITERATURE REFERENCESの項目に目を通して、その中からPreparationに関連するレファレンスを抽出しなければならない。これではCPDOおよびCPSDO/XMLの利点を生かしているとは言いがたい。

そこで、より小さな情報の単位、この場合は主題ごとに分けてCPDOとする設計が考えられる。図6はその設計によるシステムの結果表示である。図6の1から7までの矢印に示すように、DIALOGのMerck IndexのLITERATURE REFERENCESのデータをChemical InformationとOther Information (Biological, pharmacological, etc. informationのこと)の2つに分け、その中をさらにSubjectごとに分け、CPDOの概念に従って表示している。そのため、検索者は必要なデータ、例えば、製法(Preparation)が欲しい場合、矢印1に示す項目だけを取り出すことができる。粒度の異なるCPDOもCPSDO/XMLで記述した階層構造と対応しているため、XMLの手法：DOM^[22]やXPathなどを利用して抽出することができる。

4.3. CPSDO/XMLのXMLスキーマ

次に、XML DBでCPDOを記述する際に使用するXMLスキーマ^[21]を検討した。Merck Indexの分野に対応する既存のXMLスキーマとしては、Peter Murray-Rustによって開発されたCML (Chemical Markup Language)^{[23][24]}および、それを科学技術及び医学分野に拡張したSTMML (Science, Technology and Medicine Markup Language)^[25]がある。これらはオブジェクト指向プログラミングの観点に立ってデータ型を豊富に用意しており、また、STMMLは出版物の構造を意識した結果、大きなスキーマになっている。しかし、CPDOの視点から見ると、これらはMerck Indexの検索で必要な要素をカバーしておらず、CPDOの記述には適していない。

図 4: Merck Index における Ciprofloxacin の記述.

The Merck Index., 2001, p.403. Merck & Co., Inc. より引用

アドレス http://telipe4/paper/OnLineMerckindex.aspx 移動	
Merck Indexデータベースにおける項目をxmlの要素に変換した結果(炭疽菌に効くCiprofloxacinを例とした)	
Merck Index	
THE MERCK INDEX Monograph Number:2377	
Chemical Properties	
Chemical Properties, THE MERCK INDEX Monograph Number:Ciprofloxacin	
Chemical Properties, CAS REGISTRY NUMBER:85721-33-1	
Chemical Properties, CHEMICAL NAME:1-Cyclopropyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid;	
SYNONYMS(Alternate names and/or trademarks (capitalized) of title compound):	
MOLECULAR FORMULA: C ₁₇ H ₁₈ FN ₃ O ₃	
MOLECULAR WEIGHT: 331.34	
MOLECULAR COMPOSITION:C61.62%,H 5.48%,F 5.73%,N 12.68%,O 14.49%.	
Sources/References	
LITERATURE REFERENCES:Fluorinated quinolone antibacterial. Prepn: K. Grohe et al., DE 3142854;idem. US4670444(1983, 1987 both to Bayer AG); K. Grohe, H. Heitzer, Ann.1987,29. Antibacterial spectrum in vitro: B. Watt, F.V.Brown, J.Antimicrob. Chemother.17,605(1986); C.M.Bassey et al.,ibid.623. HPLC determin in biological fluids: W.Gau et al.,Liq. Chromatog.8,485(1985). Pharmacokinetics: G.Hoffen et al., Antimicrob. Ag. Chemother. 27,375 (1985). Clinical trials: C.A. Ramirez et al.,ibid. 28,128(1985); B.E. Scully et al., Lancet 1,819(1986). Symposia on antibacterial spectrum and clinical use: Am. J. Med. 82, Suppl. 4A, 1-404 (1987); J. Antimicrob. Chemother.26, Suppl. F.3-193(1990). Review of clinical safety and efficacy in children: R. Kubin, Infection 21,413-421(1993)	
PATENT INFORMATION:DE 3142854;US4670444	
PATENT INFORMATION, Patent Number:DE 3142854;US4670444	
Physical Property Information:Dec 255-257°C	
Therapeutic category (In humans):THERAP CAT: Antibacterial	
DERIVATIVE INFORMATION	
SUBSTANCE :Monohydrochloride monohydrate	
DERIVATIVE CAS RN :86393-32-0	
DERIVATIVE NAME:Bay o9867;Baycip;Ciflox;Ciloxan;Ciprobay;Ciproxan;Ciproxin;Flociprian;Septicide;Uniflox;Velmonit	
DERIVATIVE MOL FORMULA :C ₁₇ H ₁₈ FN ₃ O ₃ .HCL.H ₂ O;	
DERIVATIVE MOL WEIGHT :mol wt 385.82.	
DERIVATIVE PHYSICAL DATA :Light yellow crystalline powder. Mp.318-3200.	

図 5: DIALOG における Merck Index の各項目をそのまま CPDO に対応させ、XML で記述した結果の表示

そのため我々は独自に Merck Index 用の XML スキーマ (図 7 参照) を設計した。

図 7 の 3 行目に示すように、Merck Index の XML のルート要素 (element) の名は Merck Index である。それは下位要素: Monograph の繰り返しで構成される。また、Monograph の下位の要素として、順に AccessionNumber (DIALOG の Merck Index Online ファイルのアクセス番号で、Merck Index の Monograph number に当る)、ChemicalProperties (その内容は後述)、SourcesANDReferences (内容後述) 等を設定した。ChemicalProperties の下位要素は、図 7 の 20 行目からの数行に示すように、順に MonographTitle (化学物質名で、上述の例で言えば、Ciprofloxacin)、RegistryNumber (CAS Registry Number, CAS 化学物質登録番号) 等々、MolecularComposition までが並ぶ。SourcesANDReferences の下位要素は、31 行目からの数行に示すように、ChemicalInformation, OtherInformation, PATENTINFORMATION である。最

後の要素は DIALOG で設定されているデータ項目に対応する。40 行目からの記述は ChemicalInformation の下位要素に Subject を、48 行目からの記述は OtherInformation の下位要素に Subject を、繰り返し可として設定したことを示す。55 行目からの記述は Subject の下位に繰り返し可で Ref(参考文献の書誌データ) を置けることを示す。これら個々の要素が CPDO に対応する。また、ChemicalInformation のように、要素によっては下位の要素を持たせることによって、CPDO の階層構造を実現した。

このように、CPDO に基づく XML スキーマが設計でき、これを用いることによって様々な階層に属する CPDO に対する処理が可能になる。

アドレス http://felipe4/paper/Paper/2DOCPMerckindex.aspx	
移動 リンク	
CAS REGISTRY NUMBER:85721-33-1	
CHEMICAL NAME:1-Cyclopropyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid;	
SYNONYMS(Alternate names and/or trademarks (capitalized) of title compound): Bay q 3939.	
MOLECULAR FORMULA: C17H18FN3O3	
MOLECULAR WEIGHT: 331.34	
MOLECULAR COMPOSITION: C61.62%, H 5.48%, F 5.73%, N 12.68%, O 14.49%.	
Sources AND References :	
Chemical Information :	
Generic Name:Fluorinated quinolone antibacterial.	
1	Subject,Prepn:
	REFERENCE:K. Grohe et al, DE 3142854;
	REFERENCE:eidem. US4670444(1983, 1987 both to Bayer AG);
	REFERENCE:K. Grohe, H. Heitzer,Ann.1987,29.
OtherInformation :	
2	Subject,Antibacterial spectrum in vitro:
	REFERENCE:B. Watt, F.V.Brown, J.Antimicrob. Chemother.17,605(1986);
	REFERENCE:C.M.Bassey et al.,ibid.623.
3	Subject,HPLC determ in biological fluids:
	REFERENCE:W. Gau et al, Liq. Chromatog.8,485(1985).
4	Subject,Pharmacokinetics:
	REFERENCE:G.Hoffken et al, Antimicrob. Ag. Chemother. 27,375(1985).
5	Subject,Clinical trials:
	REFERENCE:C.A. Ramirez et al,ibid. 28,128(1985);
	REFERENCE:B.E. Scully et al., Lancet 1,819(1986).
6	Subject,Symposia on antibacterial spectrum and clinical use:
	REFERENCE:Am. J. Med. 82, Suppl. 4A, 1-404 (1987);
	REFERENCE:J. Antimicrob. Chemother.26, Suppl. F.3-193(1990).
7	Subject,Review of clinical safety and efficacy in children:
	REFERENCE:R. Kubin, infection 21,413-421(1993)
PATENTINFORMATION :	
Patent Number.DE 3142854;US4670444	

図 6: 主題ごとに分けた CPDO に基づく Merck Index の XML データの表示

4.4. CPDO および CPSDO/XML に基づく XML DB からの検索と表示

Merck Index の XML DB から検索ターゲットである CPDO を取り出した例を図 8 に示す。ここでは「炭疽菌に効く薬 Ciprofloxacin」の製法に関する特許と雑誌文献の書誌データの CPDO を、ASP.NET 上で XML スキーマと DOM を用いて取り出し、表示している。さらに、特許や雑誌の原文献リソースが Web 上に存在すれば、リンク機能を用いる連鎖検索も可能となる。

CPDO を XML データの要素として記述したため、基本的に全ての XML 関連技術が適

用できる。たとえば、必要な項目をメニュー選択により特定し、該当する CPDO を検索し、その結果を表示した例を図 9 に示す。

5. 考察

既存の情報検索システム、たとえば DIA-LOG では、ピンポイントの情報を与えるとは限らず、関連の情報も含めた形で検索結果を表示する場合が多く見られる。これには次に示す 2つの理由があると考えられる。第 1 に、既存の情報検索システムが採用しているデータベースの方式では情報に階層構造を持たせることが難しいため、たとえば、Literature

```

<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="MerckIndex">
    (中略)

  </xs:element>
  <xs:element name="Monograph" />
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="AccessionNumber" />
        <xs:element ref="ChemicalProperties" />
        <xs:element ref="SourcesANDReferences" />
        <xs:element ref="Structure" />
        <xs:element ref="PhysicalProperty" />
        <xs:element ref="TherapeuticCategory" />
        <xs:element ref="DerivativeInformation" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="AccessionNumber" type="xs:short" />
  <xs:element name="ChemicalProperties">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="MonographTitle" />
        <xs:element ref="RegistryNumber" />
        <xs:element ref="ChemicalName" />
        (中略)
        <xs:element ref="MolecularComposition" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="SourcesANDReferences">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="ChemicalInformation" />
        <xs:element ref="OtherInformation" />
        <xs:element ref="PATENTINFORMATION" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="ChemicalInformation">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="GenericName" />
        <xs:element ref="Subject" maxOccurs="unbounded" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="OtherInformation">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Subject" maxOccurs="unbounded" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="Subject">
    <xs:complexType mixed="true">
      <xs:choice minOccurs="0" maxOccurs="unbounded">
        <xs:element ref="Ref" />
      </xs:choice>
    </xs:complexType>
  </xs:element>
  <xs:element name="Ref" type="xs:string" />
    (中略)

</xs:schema>

```

図 7: CPDO および CPSDO/XML に基づく Merck Index のための XML スキーマ

図 8: 検索対象となった CPDO(参考文献の書誌データ)のみを取り出した例

図 9: 必要な CPDO を特定した検索と表示

references, Chemical information, Subject, Reference といった階層構造を扱うよりも, Literature references の単位でまとめる方が処理が容易だからである. 第2に, 検索者は検索結果を見て判断するのであるから, ピンポイントに絞って情報を提供して検索者の意図に外れるよりも, 関連の情報を含めて提供した方が得策だからであろう.

しかし, これからの検索者は単に結果を読んで理解するだけでなく, 得た結果のデータを再構成, 再処理することもあり得るし, これにより調査・研究の効率が上がることになる. これを可能とするのが, 我々が提唱する CPDO および CPSDO であり, その XML による記述と処理システムは以上述べたように可能である.

CPSDO/XML の考え方と関連する他の手法を比較する. まずフレームとの比較では, 情報の構造化, すなわち, 階層構造とリンク関係の導入はフレームと同じであるが, CPSDO/XML は表現に XML を用いるため, 同一階層の上の要素間の順序関係も表現でき

ること, インターネット上の異なる処理系の間でデータを交換し利用できることが長所である. オブジェクトデータベースとの比較では, XML を用いることにより, 複合オブジェクトよりも容易に表現できる. また, 単に XML を用いるだけでは, オブジェクト指向の長所が活かせないが, スキーマにオブジェクト指向の考え方を取り入れた XML スキーマを用いることにより, 処理手続きを記述する際にオブジェクト指向の長所を取り入れることができた. 最後に XML と比較する. XML による構造設計の際には情報をどのように見るかによって記述形式が異なり, また, 設計の良し悪しが分かれるが, CPDO および CPSDO の考え方を取り入れることによって初めて有用な設計ができたと思われる.

CPDO および CPSDO/XML に基づくシステムの実現に問題があるとすれば, 実際に CPDO をどのようにして得るかであろう. 考えられる方法は次の2つである. (A) 既存の情報検索用データベースあるいは情報リソースの中のデータ項目から変換プログラムによっ

て自動的に得る方法と、(B) データベースを作る段階から CPDO を意識して作るという方法である。

(A) の方法の場合、既存のデータベースあるいは情報リソースの中で、CPDO に 1 対 1 に対応するデータ項目については、プログラムによる自動変換が可能であり、問題はない。しかし、Merck Index の DIALOG 版に見られるように、現状ではデータベースとして提供されるデータ項目の粒度は必ずしも細くないため、CPDO に対応するより粒度の小さな単位に分ける必要が出てくる。Merck Index の場合は、「:」、「;」、「.」といった記号によってデータが区切られているため、かなりの程度、自動変換が可能であるが、全データが 100% 人手無しで自動変換できるとは言いがたい。

(B) の方法の場合、データベースの作成コストが上がる可能性があることが問題である。ただ、たとえば、Merck Index のデータが、「:」、「;」、「.」といった記号によって明確に区切られていることをみると、製作段階からデータの構成要素が意識され、区別されていると推測できる。また、実際に、たとえば、CAS では論文の書誌項目のうち、雑誌タイトル、巻、号、ページなどは別のデータ項目として作られている。さらに、作成コストが上昇するとしても、そのコストが何らかの方法で回収できるようになれば、その問題は解決できる。

そこで我々は、CPDO の考え方をデータベース製作段階から導入すること、情報検索などで情報を提供する際には、CPDO 単位での提供も行い、その場合は CPDO の単位での課金を行うことを提案する。これが実現すれば、CPDO の利用が進み、異なるデータベースあるいは情報リソースから集めた CPDO を組み合わせることによって、目的別のデータベースあるいは情報リソースを、従来よりも低いコストで作ることが可能となり、それがまた新しい情報活用を進めるといった好循環が現出されると期待できる。

6. 結論

我々は、CPSDO/XML (Computer Processable Structured Digital Object based on XML) の概念とそれに基づく検索システムを提唱した。CPSDO は、利用者が扱う情報の最小単位を構成要素とする構造化情報表現である。たとえば、もしも「検索のターゲット」すなわち「利用者が検索によって得たい情報対象」が情報の最小単位であれば、それは CPSDO の構成要素であり、もしも情報の最小単位を組み合わせたものであれば、それは CPSDO の一部分である。

CPSDO/XML は XML によって記述されるデジタルオブジェクトであり、XML のタグ名称は記述内容を示す。そのため、CPSDO/XML はソフトウェアによって処理可能である。また、検索対象の Web ページあるいはデータベースが CPSDO/XML に基づいて構築されれば、該当する情報の構成要素あるいはその組み合わせのみを検索でき、ノイズの少ない効率的な情報検索が可能になり、また、検索者が得たいピンポイントの情報が得られることを、典型的なファクト・データベースである Merck Index を例とするプロトタイプを構築して実証した。

本論文では、CPDO を、既存のデータベースあるいは情報リソースのデータ項目からプログラムによる変換処理によって、ほぼ自動的に得られることを示したが、この方法が全てではなく、CPDO の考え方をデータベース製作段階から導入すること、情報検索などで情報を提供する際には、CPDO 単位での提供も行い、その場合は CPDO の単位での課金も行うことも提案した。これが実現すれば、CPDO の利用が進み、目的別のデータベースあるいは情報リソースが従来よりも低いコストで製作でき、これがまた新たな情報活用を促進するといった好循環が期待できると主張した。

参考文献

- [1] International Standard Organization: "ISO/IEC 13250 Information Technology - SGML Applications - Topic Maps", ISO/IEC 13250, 2000.
- [2] The Semantic Web: <URL:http://www.w3.org/2001/sw/>
- [3] T. Berners-Lee et al.: "The Semantic Web", Scientific American, May 2001. (日本語訳が「自分で推論する未来型ウェブ」のタイトルで日経サイエンス, 2001年8月号に掲載されている)
- [4] Chemical Abstract Service: <URL:http://www.cas.org/>
- [5] Elsevier Science: <URL:http://www.elsevier.com/>
- [6] ISI: <URL:http://www.isinet.com/isi/index.html>
- [7] M. Minsky: "A framework for representing knowledge", in P. Winston (ed.) "The psychology of computer vision", pp.211-277, McGraw-Hill (1975).
- [8] D.G. Bobrow, T. Winograd: "An overview of KRL, a knowledge representation language", Cognitive Science, Vol.1, pp.3-46 (1977).
- [9] R.B. Roberts, I.P. Goldstein: "The FRL primer", MIT AI Lab. Memo 408 (1977).
- [10] D.B. Lenat; R.V. Guha: "Building large knowledge-based systems : representation and inference in the CYC project", Addison-Wesley Pub. Co., 1989, 360p.
- [11] STN International: <URL:http://www.cas.org/istn.html>; <URL:http://pr.jst.go.jp/db/STN/index.html>.
- [12] F. Bancilhon et al. eds. "Building on Object-Oriented Database System", Morgan Kaufmann, 1992.
- [13] D. Jordan: "C++ Object Databases Programming with the ODMG Standard", Lucent Technologies, 1998; 玉井浩 訳: 「C++オブジェクトデータベース ODMG 標準によるプログラミング」アジソン・ウエスレイ, 1998年.
- [14] Science Direct: <URL:http://www.info.sicencedirect.com/>
- [15] Extensible Markup Language (XML): <URL:http://www.w3.org/xml/>
- [16] Resource Description Framework (RDF): <URL:http://www.w3.org/RDF/>
- [17] H. Ishizuka; Y. Li: "Digital Library System Based on Concept of Computer-processable Digital Objects", Proceedings of Digital Library - IT Opportunities and Challenges in the New Millennium, p.315-326, 2002.7, Beijing.
- [18] Microsoft ASP.NET: <URL:http://www.asp.net/>
- [19] Thomson, DIALOG: The Merck Index Online: <URL:http://library.dialog.com/bluesheets/html/b10304.html>
- [20] The Merck Index., 2001, p.403. Merck & Co., Inc.
- [21] XML Schema: <URL:http://www.w3.org/XML/Schema>
- [22] Document Object Model (DOM): <URL:http://www.w3.org/DOM/>
- [23] Peter Murray-Rust, Henry S. Rzepa: "Chemical Markup, XML, and the Worldwide Web", J. Chem. Inf. Comput. Sci., 1999, Vol.39, p.928-945.
- [24] CML: <URL:http://www.xml-cml.org/information/dictionaries/index.html>
- [25] STMML: <URL:http://www.xml-cml.org/stmml/stmml.xsd>

(2002年11月19日受付)

(2002年12月25日採録)