解　説

知の空間を構成する大規模知識ベース

——知識インフラの構築と インテリジェント化の技術†——

横　井　俊　夫††

1. はじめに

知識としての知の構造の解明，それに基づく大規模な知識としての知の構築とその（再）利用のメカニズム，これが今後の情報処理分野の最重要のテーマになる．知能としての知（知の内部メカニズム）解明の困難さが切実に実感される中で，現象として明確に把握しうるものである知識としての知（外在化された知の現象）に焦点をしぼろうというのである．知識を社会的に共有化されるものとみると，その構造の解明，そして構築と利用のメカニズムは，21世紀の地球規模の高度情報化社会実現の礎となる知識インフラ☆ の構造の解明，そして構築と利用・共有のメカニズムにマクロに結びつく．

この大課題を字義どおり体現するものとして大規模知識ベース☆☆を位置付ける．重要性を反映してか大規模知識ベースを標榜したり，それを目指したりする多くの試みが行われるようになってきた．しかし，この課題の大きさを的確に自覚した試みは，まだこれからのものである．そこで，本稿も既存の事例を細かに紹介するよりこれからの方向付け，本来あるべき姿の解説のほうに重きを置く．そのほうが読者諸氏の関心を喚起し，議論を活性化するには良いとの判断からである．ただし，そのために全体として概論的にすぎるきらいとなるが，ご容赦願いたい．

2. では，大規模知識ベースの枠組を定めることになる知識としての知の基本的な構造をマクロに反映する知識インフラとその構築の基本ステップを解説する．**3.** では，大規模知識ベースへのアプローチの現状を二つの側面から概観する．大規模知識ベースを暗に目標とする個別技術分野からの側面と，大規模知識ベースを陽に目標とする事例の二つの側面からである．**4.** では，アプローチの現状を**2.** で述べられる枠組にそって整理することによって浮かび上がってくる大規模知識ベースのあるべき姿を解説する．基本的な要件，知識を規定する知識表現メディア，システムとしての基本機能の3点から説明する．**5.** ではより踏み込んだ議論の糧となるように，あるべき姿をもう一段具体化する．これは筆者らが計画・立案中の研究開発計画の一部を題材としたものである．**6.** では大規模知識ベースという課題のポイントをトピックス的に再整理して本稿のまとめとする．

2. 知識インフラの構築

大規模知識ベースという課題を知識インフラの整備と利用・共有化のための新しい技術の研究開発を目標にするテーマとしよう．したがって，知識インフラの基本的な構築のステップが課題の輪郭を描き出してくれることになる．

グーテンベルグ以来の印刷文字文化から電子文字文化への壮大な移行が始まったといわれる．電子化された知の空間（情報空間，知識空間）の構築である．巨大な知の空間をコンピュータによって縦横に駆けめぐることによって人間の知能活動は強化され，その知的生産性は飛躍的に向上する．コンピュータの人工知能化ではなく知の人工空間の建設である．この人工空間は物理的世界に

† A Very Large-Scale Knowledge Base Embodying an Intelligence Space—Technologies for Building and Evolving Knowledge Infrastructures—by Toshio YOKOI (Japan Electronic Dictionary Research Institute, Ltd.).

†† 日本電子化辞書研究所

☆ 情報インフラという言葉が使われるようになったが，ネットワークのようなハードウェア環境の整備を指すのに使われるのがほとんどである．欧米で information infrastructure という場合は，ネットワークだけではなく，その上に乗せる社会資本としての情報そのものをも指す．ここでは新社会資本としての情報・知識そのものを強調する意味で知識インフラという言葉を用いる．

☆☆ 以降，この言葉は多義に用いられるので注意されたい．大規模に集積された知識そのものを指す場合，大規模知識ベースにかかわる技術を指す場合，大規模知識ベースシステムというソフトウェアシステムを指す場合などがある．

対峙する情報の世界を形作り，建設作業には21世紀の大半を要することになるであろう．そして，人工空間の多くは社会的に共同利用される施設として建設される．ここに知識インフラ，社会資本としての知識という発想が生まれる．

注意されたいのは，ここで取り上げる大規模知識ベースという課題は知識インフラそのものの構築を始めようというものではないということである．知識インフラの構築を高能率化し，極力自動化するための技術，利用度や共有化度を高いものにし，高レベルの利用を可能にするための技術などの研究開発が目標である．もちろん，そのためには共通となる基本的な大規模な知識そのものの構築も手掛けることになる．

知識インフラ（知の空間）は，機械可読化（電子化），ハイパ化☆，インテリジェント化のステップを大きく踏みながら高度な知の構造物へと成長していく（**図-1**）．

機械可読化：すべての情報（知識）をコンピュータ処理可能な状態にするステップである．ワープロや印刷工程でのコンピュータ利用の普及によって機械可読化は急速に進みつつある．情報を表現するいろいろなメディアに対し，データとしてのデジタル化，メディアの表現要素（文字など）のコード化，情報の提示構造（印刷仕様など）のコード化，情報の論理構造（文書構造など）のコード化とさらに細かなステップが踏まれる．そのステップにそって技術開発が行われ，コード体系やコード化のための言語の標準化が進められている．まだまだ，解決すべき問題も多い．文字コードでさえも，世界の全文字に対する統一コードの標準化となるとまだ落ち着いたものになってはいない．文書の論理的な構造の表現に対しては，SGML などの利用が始まったところである．これに関しては文書化習慣の希薄さから日本の遅れが目立つ．図形・画像・音響情報については，マルチメディアの本格的利用の開始を前にして問題が山積した状態である．

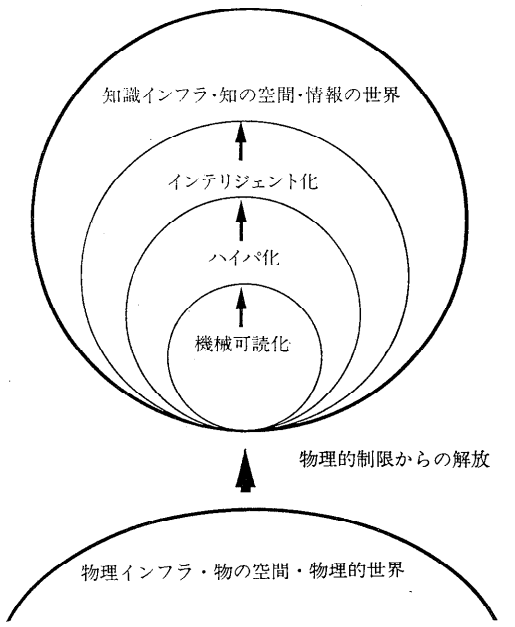

図-1　知識インフラ形成の基本ステップ

☆ハイパメディアのハイパとほぼ同一趣旨であることから借用した言葉である．

ハイパ化：情報（知識）をその本来の構造へと再構成するステップである．それぞれのハード的媒体の制約のままに表現されていた情報をそのままコンピュータに移すのが機械可読化のステップである．ハイパ化は媒体の制約という夾雑物を取り除き情報間の本来のつながり（関係）で構造化しようというものである．ハイパメディアやデータベースの情報などが対応する．

インテリジェント化：情報を知識としての体系に整えるステップである．ルール（規則）やケース（事例）として整理された知識と適切な推論のメカニズムによって蓄積された知識から多くの新しい知識を導出できるようになる．ファクト（事実），ルール，メタルール，……という軸にそった知識の階層化も行われる．

知識インフラは，大きく以上の3ステップを踏みながら知の空間としての拡がりを増していくことになる．

現在，知識インフラを本格的な議論の対象にできる諸条件がようやく整ったのであるが，その中で特に重要なものが，知の空間が物の空間（物理的世界）の制限から解放されてきたことである．

コンピュータ，ネットワーク，マルチメディアなどの技術進歩と製品の普及によって，物理的制限の多くが急速に取りはらわれ情報，知識の本来の性質のみから知の空間を構築できる状況が整えられてきている．コンピュータの高性能化と低廉化は，処理技術や記憶容量という制限を緩めてくれつつある．ネットワークの通信容量の増大と普及は離れた場所にある知の空間を物理的距離にわ

ずらわされることなく一体化できるようにしてくれつつある．これによって知の空間における距離を，情報の内容や論理的属性に基づいて定義できるようになる．マルチメディアの進歩と普及は，コンピュータ上のメディアと人間が日常使用しているメディアが大きく異なるという制限を取り払い，情報表現を一体化しつつある．

もちろん，物理的制限がまったくなくなったわけではない．知識インフラが巨大になるにつれ，また新たな制限の解消に向けた努力が求められるようになるであろう．しかし，予想されるハードウェアの進歩は十分な余地を約束してくれているようである．

以上の知識インフラ構築の基本ステップに重ね合わせると，大規模知識ベースという課題の解決はハイパ化を含むインテリジェント化の部分に大きな役割をはたすことになる．しかし，機械可読化にも重要な役割がある．通常，インテリジェント化された情報（知識）からみれば，ハイパ化された情報はファクトデータ，機械可読化された情報は素材データとみなすというところであろう．しかし，知識処理技術の現状からも明らかなように，人間のもつ知識を広範囲にわたり高精度にインテリジェント化する技術はまだ遠い将来のものである．したがって，人間の知識を最も正確に表現しているのは機械可読化された情報であり，ハイパ化，インテリジェント化によって知識は近似され，ある側面，ある部分のみが取り出されるにすぎないとみるほうが妥当である．

3. 大規模知識ベースへのアプローチ

知識インフラとの対応からも明らかなように，大規模知識ベースは多くの技術分野にかかわる．その現状をみるとき，表立って大規模知識を標榜しているもののみを取り上げたのでは今後の動向を正確に把握することはできない．そこで，個別技術分野と事例の両面からアプローチの分析を試みることにする．

3.1 個別技術分野の動向

5つの技術分野における動向を説明する．これらは大規模知識ベースへアプローチするときに各分野がはたすべき役割とみなしうるものである．

文書処理：自然言語で表現された機械可読化情報の作成・編集，蓄積・検索，変換，伝達を目標とした分野であるが，頑健な技術に育ちつつある自然言語処理[1]によって新しい展開が望めるようになった．特に，機械翻訳，フルテキスト検索，テキストからの知識抽出に目立つ動きがある．機械翻訳は商品化が一段落し，新しい努力が始まっている．たとえば対訳用例データを利用して翻訳処理をする事例ベース機械翻訳[2]である．フルテキスト検索は，専用ハードウェアを含め実用化の活発なテーマである[3]．制限キーワードから自由キーワードへという動きは注目すべきものである．テキストからの知識抽出は，文書要約やキーワード抽出を含むものでこれからの重要なテーマである．現在では米国 ARPA の MUC[4] から TIPSTER プロジェクトへの動向がまとまりのある例である．これからはさらに新しい発想で取り組むべき課題である．

知識工学：知識ベースという考え方を打ち出し，インテリジェント化の先頭を切った分野である．しかし，知識獲得ボトルネックが切実になるにつれ，その解決のための地道な努力が始まっている．知識の再利用をいかに高めるかという健全な発想を中心に解決策が議論されている．米国における Knowledge Sharing Effort[5] がまとまりのある代表的な取組みである．さまざまな知識表現言語に対する中間言語，共通の知識表現言語，知識ベースシステム間の共通のプロトコル，共通のオントロジという4テーマに取り組んでいる．オントロジとは，存在論 (ontology) という哲学用語からのものであるが，記述対象をどうみるのかという世界観と記述するための語彙セットを指す用語である．これから多用される用語になる．もう一つの重要な動向はルールからケースへというものである．ケースベース，事例ベース，メモリベースなどと称される一連の動きである[6]．知識要素の相互のインタラクションをできるだけ簡明なものにするのが，再利用度を上げる大きな手立てになる．

データベース：大規模な情報を扱い，すでに大量の蓄積が行われている分野である．いろいろなインテリジェント化に向けて努力が行われ始めている．まず，データベースの知識ベース化ともいうべきもので，演繹機能やオブジェクト指向機能を付加し入れ物としての高機能化をはかろうというものである[7]．次は，データの知識化ともいうべ

きもので，大規模データベースからの知識獲得の試みである[8]．そして，データベースの共通オントロジともいうべきものもある．各応用領域に対し，関係 (relation) 名や実体 (entity) 名の共通語彙を定めようというものである．

マルチメディア (ハイパメディア)：多様なメディアによる情報を統合的に扱い，永い歴史の中でメディアの共通化をはかり再利用度の高い表現形式を工夫してきた人間の努力を最大限に利用しようという分野である．マルチメディアの目新しさに関心が集められ，とりあえずその範囲で各種のハイパメディア・ソフトが開発され，蓄積され始めている．しかし，大規模なハイパメディア・ソフト，すなわち，大量マニュアルのハイパ化や電子図書館などが指向されるにつれ，リンク付けの自動化やリンクのインテリジェント化への要求から他の分野との連携が始まっている[9]．

ソフトウェア工学：プログラムや仕様という形式の情報をいかに効率良く生産できるようにするかをテーマとする分野である．生産性向上の健全な手立てはやはり再利用である．CASE 環境においても大量のプログラムや仕様を蓄積し検索要求に適切に応える機能が重視される[10]．その機能もプロダクトライブラリからリポジトリへと機械可読化からよりインテリジェントなものへと高度化の努力が続けられている．この努力は，プログラムや仕様や作成過程をいかに明確に表現できるようにするかという工夫と表裏をなして進められている．

以上の動向をふまえ大規模知識ベースの位置付けを図示すると図-2 のようになる．大規模知識ベースというテーマはソフトウェアに関する技術分野を統合的に含む大きなテーマであり，また，そのようにアプローチされねばならない．

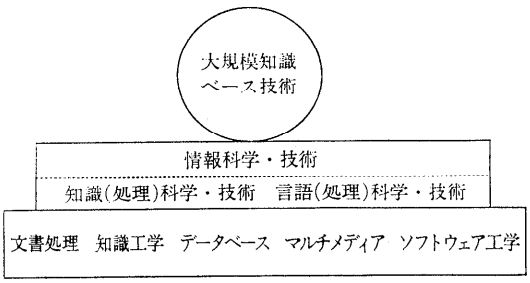

図-2　大規模知識ベース技術の位置付け

3.2 事　例

大規模知識ベースを標榜する事例を少し広めにとって現状のアプローチを整理する．大規模というからには，知識ベースとして高い汎用性，一般性，共通性をもつものでなければならない．知識ベースとして汎用性などを議論する部分は大きく次の三つである．

①記述対象：対象とする知識分野や知識のレベル

②記述方式：コンピュータが理解可能か否かより，データとしての記述力の高さや記述の容易さ

③問題解決（推論）方式：コンピュータの理解力の高さや範囲の広さ

この3点すべてにわたって高い汎用性，一般性，共通性を達成しようというのはかなりの将来にわたって不可能である．どこかに重点を置くことになる．置き方によりアプローチのバラエティが生まれる．知識工学分野からのものではどちらかといえば③に，文書処理，ハイパメディア分野からのものでは②に重点が置かれる．

ＣＹＣ[11]：大規模知識ベースのさきがけとなった Lenat らの CYC は，当初は上記の3点すべてにわたって高い汎用性を実現しようということを表看板に始められた．机上タイプの百科事典に含まれている全知識を記述対象に選び，記述方式，問題解決方式ともに高い汎用性をもたせるように努力した CycL という知識表現言語で 1,000 万項目にわたる知識を集積しようというものである．この知識量は有効な自動学習が可能になる量として設定された．これは明らかに過大すぎる目標である．現在は試行錯誤の結果，かなり妥当なところに落ち着いてきているようである．すなわち，CycL も記述方式の一般性に重点を置いたものとの位置付けになり，さらに百科事典を対象にすることによりまとめられた時間，空間，因果関係，信念などに対するオントロジで強化されたものにまとめられている．そして，この CycL によって計算機，アパレル，法律，化学などの範囲を限定した記述を対象にすることに重点を移しているようである．しかし，なにぶんにも当初かかげた看板が過大すぎたこと，また研究開発の行われている MCC という組織の性格から現状が具体的に示

されにくいことから批判も多い☆.

EDR 電子化辞書[13]：記述対象のある種の一般性に重点を置く．記述方式や問題解決方式は意味ネットワーク程度の簡素なものである．自然言語（日本語，英語）の基本語を主とする語彙周辺の知識を記述対象に選んだものである．言語，特に，汎用性の高い言語に関する知識というのは現時点で安定した一般性を得る重要な方法である．ただし，これは言語処理のためにとりあえず有用である知識に限定したものである．大規模知識ベースとの関連は文献 14) により詳しく解説されている．

工学分野の共通知識ベースとして：まだ大規模とはいえないが，知的 CAD などを目的にした共通知識ベースの試みがいくつか始まっている．記述対象を限定することにより，問題解決方式や記述方式にある程度の汎用性を確保しようというものである．機械設計を対象にフィジカルフィーチャと呼ぶ単位を知識の基本単位に表現，収集を行う試み[15]や物理デバイスをモデリングするために必要な工学的知識などを対象に工学分野の共用知識ベースの構築を目標にしたもの[16]などがある．

分散協調システムとして：今まで述べてきたものとは別の方式で大規模さを達成しようというアプローチである．ネットワーク上で各所に分散するさまざまな知識ベースを協調して働くようにしようというものである．そのためにインタフェースやプロトコルの標準化やシステムアーキテクチャのオープン化などを図ろうというものである．今までの議論にそえば，知識ベースのシステム仕様を記述対象にした共通知識ベース作りとみることもできる．前節で述べた Knowledge Sharing Effort 内のグループによる Collaborative Testbed という試みがある．分散協調するものの単位をさらに小さくとろうという試みもある．知識コミュニティ[17]などの提案である．これは，大規模知識ベースを大規模ソフトウェアと見立てるアプローチである．再利用性の高い柔軟な構造をもつ大規模ソフトウェアを実現する手立てとして自立して動作し，互いに協調し合うエージェントの集合という構成法を採用する．エージェントの粒度をどのくらいにするかが大きな論点となるが，究極は Minsky の“心の社会”である．このあたりになると賛否さまざまであろう．

☆ このあたりは，Artificial Intelligence (Vol. 61, No. 1, 1993) 誌の文献 12) に対する多くの書評と著者らの回答という長文の記事からもうかがい知ることができる．

4. 大規模知識ベースのあるべき姿

知識インフラの構築ステップに合わせ，現状のアプローチを整理し，これからの大規模知識ベースの研究開発がどのような枠組に納まるべきか，あるべき姿を説明する．満たすべき基本的な要件，骨格となる構造，システムとしての基本構造の3点から解説する．なお，骨格となる構造は知識を表現する言語やメディアによって定まるものと考える．

4.1 基本的要件

目標とすべき大規模知識ベース，すなわちこれからの研究開発テーマとして魅力あるものとなるべき大規模知識ベースが満たすべき要件は以下のようなものになろう．

①大規模化が技術的ブレークスルーにつながること：まず，大規模知識ベースの構築技術の研究開発自体が，ブレークスルーを必要とする魅力的な多くのテーマを含み，成果が広範囲にわたる波及効果を生み出すことである．次に，構築技術によって作り出された大規模知識を利用する側となる技術やシステムに新しい機能や次世代への展開が得られることである．

②大規模に知識を獲得，集積することが可能であること：大規模知識ベース構築の仕組が知識インフラ構築のステップ(**2.**)を素直に反映したものであることである．知識素材の多量入手が可能であり，知識獲得の自動化，あるいは効率の良い支援機能の実現が可能であるところから着手することである．そして，ある種，あるレベルでの網羅性が十分に達成されることも重要である．

③発展性を保証できること：大規模に集積されたものは自ら強い慣性をもつだけに，以後のさまざまな展開の障害になる要素を含むようなものであってはならない．大規模といえども，人間のもつ全知識に比べればごく初段階のものである．後に続く，さらなる知識の解明や知識ベースの高度化を促進，加速するものでなければならない．そのためにも，**3.**で述べた種々のアプローチのそれぞれの長所を統合化しうるものであることが重要である．無理なく大規模化しうること，そして，大

規模化が無理を生むことのないことである．
④それ自体としても十分な存在価値をもつこと：まず最初に作り上げられる大規模知識ベース自体が知識ベースシステムとして十分に有用な機能を実現するものであることである．高い機能をもつ非常に一般的なエキスパートシステム，ユニバーサル・エキスパートシステムになることである．エキスパートシステムとしてのユニバーサルさは，より広くコンピュータシステム（ソフトウェアシステム）の新しいアーキテクチャの提案にもつながる．

4.2 知識と知識表現メディア

知識をどのような形式のものにするのかによって大規模知識ベースの性質や構造が定まる．対象にしうる知識は客観的に観察しうる形式をとるものだけとしよう．言語，広くは各種メディアによって表現されることにより，知識は分析したり，処理したりできるものになる．知識を表現するメディアを何にするかによって大規模知識ベースの性質や構造が定まることになる．したがって，大規模知識ベースのあるべき姿を見通すために，それぞれの知識表現メディアの性質，役割，互いの関係などを的確に見極めねばならない．

知識表現メディアは大きく二つに大別される．とりあえずそれぞれをヒューマンメディア（人間が認識・理解するメディア），コンピュータメディア（コンピュータに処理可能なメディア）と呼ぶことにする．以下の説明には，**図-3** を参照されたい．

ヒューマンメディアは，人類が知識の表現，蓄積，利用，伝達のためにその永い歴史の中で育んできたもので，人間を処理系，推論系，理解系とする知識表現メディアである．記号化能力と汎用性の両面から最も高い能力をもつ自然言語（日本語，諸外国語）を汎用知識表現言語として，形式言語（代数式，論理式，コンピュータメディアなど），図形言語（建築設計図法，電子回路設計図法，楽譜など）という応用向知識表現言語がある．そして，情報，知識の直截的な表現能力に富む画像（静止画，動画，アニメーションなど），音響（音声，音楽，一般音）が加わる．それぞれのメディアは他で置き換えることのできない固有の役割をもち，適切に組み合わされることによって高い表現能力を発揮する．

コンピュータメディアはコンピュータを処理系，推論系，（理解系），とする知識表現メディアである．人工知能分野での（狭義の）知識表現言語，プログラム言語，データベース言語など多くのものが含まれるが，表現の対象や能力，そして処理効率などによってそれぞれの役割が割り振られる．また，ニューロコンピューティングやファジィロジックなど，まだ表現能力に十分な広がりを得るまでにはいたっていないが，新しい能力を求めての工夫が続けられている．

重要なことは大規模知識ベースとしてこの二つのメディアにいかなる役割を割り振るかということである．知識工学からのアプローチでは，大規模な知識をコンピュータメディアによって表現し

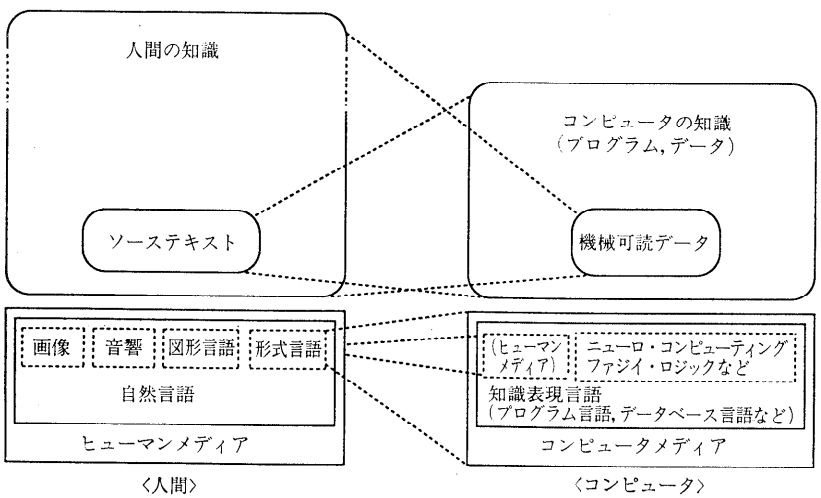

図-3 知識と表現メディア

つくそうとする．一方，ハイパメディアのようなアプローチはほとんどヒューマンメディアに頼って進められる．しかしながら，**3.** で述べたアプローチの現状から次のことは明らかになったといえるであろう．

- たとえ分野を限ったとしても人間のもつ知識をコンピュータメディアで表現しつくすにはまだ程遠いところにある．コンピュータメディアにおける最近の新しい工夫も，その効果のほどはごく限られたものである．
- コンピュータメディアでさえもコンピュータが確実に処理（理解）できるのはオブジェクトとして実行することだけである．メタな処理，たとえば正しさを判定する，等価性を判定する，表現を作り出すなどに関してはコンピュータが理解できるのはごくわずかであり，やはりほとんどは人間の役割である．この事実を反映して，図-3 ではコンピュータの知識がソーステキストとして人間の知識に含まれている．また，コンピュータメディア全体が形式言語の一部としてヒューマンメディアに含まれている．

ただし次の2点の例にみるようにメディアの技術に新しい展開がある．

- マルチメディアの技術によって，ヒューマンメディアで表現された知識の相当広範囲のものをとりあえず機械可読データとして扱うことが可能になった．そのレベルに限るならヒューマンメディアもコンピュータメディアとなりつつある．さらに，コンピュータ上で融合することによりヒューマンメディアは新しい機能をもつにいたっている．この事実を反映して，図-3 では，人間の知識が機械可読データとしてコンピュータの知識に含まれている．また，ヒューマンメディアがその表層部分のみということで“(ヒューマンメディア)”としてコンピュータメディアに含まれている．
- ヒューマンメディアをマルチメディア的なレベルを越えてコンピュータに理解可能なものにしようという努力も少しずつ実を結びはじめた．特に，自然言語に関しては確実な進歩が得られようとしている．コンピュータの処理能力はマルチメディア的な文字列レベルから相当向上し，形態素や構文レベルに至っては，かなり広範囲に利用できるほぼ安定したものとなりつつある．意味処理に関しては，ごく浅い理解に限ればもう少しの努力の段階になっている．文脈処理に関しても限られた範囲内であるが同様の試みが行われている．このようにして，図-3 の“(ヒューマンメディア)”は表層から少しずつ深層への機能をもつようになる．“機械可読データ”は機械認識可能データ，機械理解可能データへと高度化していく．

以上のことから，大規模知識ベースでのメディアの役割付けを考える上での指針としては，知識は人間とコンピュータの複合系に対して表現されるものということになろう．すなわち，理解の主役は人間であり人間の理解を適切に支援する機能を十分に発揮しうる程度にはコンピュータも理解できるようにするというのが出発点としては妥当である．

4.3 システムの機能

大規模知識ベースが基本的にもつべきであると期待されるシステムとしての機能の要点は以下のようなものになろう．

知識ベース機能：大量の知識を体系的に蓄積するための機能である．当然，適切なレベルの学習能力，自己組織化能力をもつことになる．また，この機能を規定する（狭義の）知識表現言語には，あるレベルの高い汎用性と効率の良い処理系の存在が求められる．

知識獲得支援機能：文書などの知識素材からと専門家からの知識の獲得，収集を高度に支援する機能である．知識素材からの知識獲得は極力自動化するのが望ましいが，良質の知識を得るには獲得過程への人間の介入や人手による事前編集が必要である．専門家からの知識獲得では，獲得作業を行うインタビュアを介入させるときもあるが，専門家自身やグループによる作成を支援する環境の整備を原則とすべきである．

知識利用支援機能：いろいろな利用に特化された知識ベースの作成を高度に支援する機能である．

大規模知識ベースはあくまでもマスタ知識ベースである．ある部分，あるレベルのみを取り出したり，適切に組み換えたり，知識コンパイルしたり，バラエティに富んだ広範囲の応用向きの効率の良い知識ベースが生成できねばならない．もちろん大規模知識ベースシステムはそのままでも一

つのエキスパートシステムとして機能する．知識獲得支援機能の実現にはこのシステム機能も利用される．

5. 研究開発へのステップ

この分野に取り組みたいと思っておられる方々のために前章で述べた大規模知識ベースのあるべき姿にわずかではあるが肉付けをする．これは，筆者らが計画・立案中のプロジェクト計画[18]を参考にしたもので一例として解説する．

知識分野（記述対象）：情報処理に関する知識を対象にする．この分野に関する科学技術的・学術的知識，政策・産業・経済・社会問題にかかわる知識，歴史的・科学技術論的知識，特許や知的所有権や法律にかかわる知識，教育や資格試験にかかわる知識，製品・システム仕様・言語仕様・製造技術にかかわる知識，機関や人物にかかわる知識などを対象にする．

知識表現メディア（それぞれのヒューマンメディアの位置付け）：日本語を中心にし，次のような段階を追ってコンピュータの理解能力を拡大していく．①日本語のみを用いる．他のメディアによる知識は日本語による近似表現で置き換える．また，知識抽出が容易になるように制限を加えた日本語の仕様も考案する．②日本語と形式言語を用いる．形式言語としては（狭義の）知識表現言語，プログラム言語，代数式，論理式などを用いる．知識表現言語としては大規模知識ベースシステム用のものをはじめとして代表的なものを取り上げる．③日本語，諸外国語，形式言語，図形言語，画像，音響などを総合的に用いる．図形言語，画像，音響に関するコンピュータの理解能力はごく限られたものに限定するのが無難である．

知識表現言語（中核となるコンピュータメディア）：大規模知識ベースの理解能力を表し，記述方式や問題解決方式を代表する．問題解決能力の高さより，記述能力の一般性に重点を置く．どの知識のどのような構成要素になっているのかという構造的関係，同意や上位—下位などの意味的関係，時間，空間，因果関係などに対する推論・理解能力をもつ．これらにかかわる詳細なオントロジが大規模知識の骨格を形作る．

知識素材：対象とする知識分野の機械可読化レベルの情報であり，知識の源である．情報処理関連の科学技術文献，教科書，ハンドブック，用語辞書，新聞記事，特許文書，法律文・判例文，マニュアル，仕様書，プログラム，データベースなどである．

6. むすび

今まで述べてきたことの補足としてこれからの大規模知識ベース研究のポイントをトピックス的に4点にまとめ，むすびとする．

①日本語コンピューティング：日本語をコンピュータシステムの情報表現・処理の中核言語に設定する．ソフトウェアからハードウェアまでを含む一般的なコンピュータシステムの中核に日本語を位置づけてみようという試みである．日本語が情報側からシステムアーキテクチャを規定する．日本語で表現された情報の作成，変換，蓄積，検索，伝達がシステムの基本機能となる．ただし，コンピュータ側からシステムアーキテクチャを規定するのはプログラム言語である．プログラムはプログラム言語で書かれる．日本語プログラミングと混同しないでいただきたい．

②コーパスベース言語処理：大量の言語データを収集し，現実の言語現象に対応付けながら（自然）言語処理ソフトウェアを頑健なものに育て上げたり，新しい言語理論の研究を展開する．言語データに関しては，EDR 電子化辞書によって単語レベルのものの整備は一応なされた．次は，句，文，文章レベル，いうなればコーパスレベルの電子化辞書の研究・開発が必要である．これにより文，文章を単位とした意味処理，さらには文脈処理への本格的な取組みが始まる．

③ケースベース知識処理：知識はほとんどがケース（事例）の形で蓄積される．ケースは知識現象を直截に捕捉する再利用に適した形式である．蓄積された大量のケースを分析することによって，より複雑な知識，ルール化された知識などが取り出される．大規模知識ベースは大規模ケースベースである．

④大規模オントロジ：言語処理と知識処理を結び付けるのがオントロジである．通常，知識工学でいわれるオントロジはどちらかといえば小語彙であるが，ここでは大語彙のものを想定する．言語処理におけるシソーラスが知識処理のオントロジに融合する．概念のゆれの少ない専門用語を対象

に融合化がはかられる．単語だけではなく，句や文も対象になる．オントロジの規模の拡大や精度の向上を学習能力によって手順を追って達成していく．そして，オントロジを刻々入力される知識にダイナミックに適応させたり，観点の違いに合わせダイナミックに骨組みを変化させたりするダイナミックオントロジともいうべきものが目標である．大規模知識ベースの知識構造としても，研究テーマとして大規模オントロジが焦点となる．

参考文献

1) 松本祐治：頑健な自然言語処理へのアプローチ，情報処理，Vol. 33, No. 7, pp. 757-767 (July 1992).
2) 佐藤理史：実例に基づく翻訳，情報処理，Vol. 33, No. 6, pp. 673-681 (June 1992).
3) 小川隆一，菊池芳秀，高橋恒介：フルテキスト・データベースの技術動向，情報処理，Vol. 33, No. 4, pp. 404-412 (Apr. 1992).
4) Lehnert, W. and Sundheim, B.: A Performance Evaluation of Text-Analysis Technologies, AI Magazine, Vol. 12, No. 3, pp. 81-94 (1991).
5) Neches, R. et al.: Enabling Technology for Knowledge Sharing, AI Magazine, Vol. 12, No. 3, pp. 36-56 (1991).
6) 特集「事例ベース推論」，人工知能学会誌，Vol. 7, No. 4, pp. 558-607 (1992).
7) 横田一正：演繹オブジェクト指向データベースについて，コンピュータソフトウェア，Vol. 9, No. 4, pp. 3-18 (1992).
8) 西尾章治郎：大規模データベースにおける知識獲得，情報処理，Vol. 34, No. 3, pp. 343-350 (Mar. 1993).
9) 黒橋禎夫，長尾 眞，佐藤理史，村上雅彦：専門用語辞典の自動的ハイパーテキスト化の方法，人工知能学会誌，Vol. 7, No. 2, pp. 336-345 (Mar. 1992).
10) 鯵坂恒夫：開放型 CASE プラットフォーム，コンピュータソフトウェア，Vol. 10, No. 2, pp. 4-12 (1993).
11) Guha, R. V. and Lenat, D. B.: Cyc: A Midterm Report, AI Magazine, Vol. 11, No. 3, pp. 32-59 (1990).
12) Lenat, D. B. and Guha, R. V.: Building Large Knowledge-Based Systems, p. 372, Addison-Wesley (1990).
13) EDR 電子化辞書仕様説明書，p. 134, 日本電子化辞書研究所 (1993).
14) 横井俊夫：知識処理と自然言語処理の融合としての大規模知識ベース―電子化辞書から知識アーカイブへ―，人工知能学会誌，Vol. 8, No. 3, pp. 286-296 (1993).
15) 松本幹雄他：設計向き大規模知識ベースの研究，人工知能学会全国大会（第5回）論文集，pp. 717-720 (1991).
16) Feigenbaum, E. et al.: Large Knowledge Bases for Engineering: How Things Work Project of the Stanford Knowledge Systems Laboratory, Proc. of the Symposium on VLKB, ASTEM, pp. 9-20 (1990).
17) 西田豊明：大規模知識ベース構築への私案，人工知能学会研究会資料，SIG-KBS-9104-6 (1/17-18), pp. 39-46 (1992).
18) 知識アーカイブ研究開発計画，p. 127, 機械振興協会・経済研究所，日本システム開発研究所 (1992).

（平成5年6月4日受付）

横井 俊夫（正会員）

1941年生．1965年東京大学工学部電子工学科卒業．1966年通商産業省工業技術院電気試験所（現在，電子技術総合研究所）に入所．オペレーティングシステム，計算機アーキテクチャ，人工知能などの研究に従事．1982年(財)新世代コンピュータ技術開発機構へ出向し第5世代コンピュータ・プロジェクトの推進に従事．1987年日本電子化辞書研究所へ出向し電子化辞書プロジェクトの推進・運営に従事．現在，日本電子化辞書研究所所長．電子情報通信学会，人工知能学会，日本ソフトウェア科学会，日本認知科学会各会員．