6 オントロジーに基づく知識の構造化と活用

橋田 浩一　（独）産業技術総合研究所

和泉 憲明　（独）産業技術総合研究所

異種のコンテンツやサービスを利用者の意味のレベルで相互連携させて知識の循環と拡大再生産を支援するため，データ形式とオントロジーの標準化に基づいてサービスやコンテンツを収集・構成・蓄積する技術について述べる．オントロジーを用いて意味構造を明示することによりコンテンツの作成コストを低減し品質を高める方法（セマンティックオーサリング）や，Web ブラウザ上でのコンテンツの操作をドラッグ＆ドロップで行うための JavaScript ライブラリ（WebSLIT），ならびに，Java ネイティブの Web アプリケーションサーバのソフトウェアフレームワーク（YOSEE）を紹介する．

情報爆発から知識循環へ

Web の登場によって利用可能な情報の量は日々爆発的に増大し，検索技術の進展によって情報アクセスの精度も向上を続けている．このような情報量の増大は，価値観を多様化させ，多品種少量生産，サービス化，個人化等のトレンドを加速している．

しかしこれまでのところ，情報は爆発しているが知識は爆発しているとは言い難い．見かけの情報量の爆発は，データベースからのコンテンツの自動生成，コンテンツのコピー，さまざまなセンシング等の結果であり，これらは必ずしも真に新たな知識の創造にはつながっていない．しかも，特定の検索エンジンや文書フォーマットによる技術の画一化がもたらす情報リテラシーの低下や寡占的な情報操作も危惧される．

知識が社会のさまざまな場や人々の間で共有され絶え間なく拡大再生産される知識循環型社会（knowledge-spiral society，または知識社会 knowledge society，知識に基づく社会 knowledge-based society など）を実現するには，知識の社会的共有と高度利用を支援する情報技術が必要であろう．それは，人間とコンピュータとが共有する意味に基づいてコンテンツやサービスを作成・流通・運用する技術（セマンティックコンピューティング）だと我々は考える．そこでは，さまざまな利用者がオントロジーを協調的に構築する環境，オントロジーおよびそれに基づくコンテンツを可視化する手法，それに即したユーザインタフェース，コンテンツの翻訳や検索，要約などの技術が重要な役割を果たす．

以上の観点から，本稿では，知識循環を実現するための基盤技術として，オントロジーの共有に基づいてさまざまなコンテンツやサービスを作成・収集・連結・蓄積する技術の枠組みについて述べる．

知識循環型社会の基盤としてのオントロジー

オントロジーは，概念の定義と概念間の関係をコアとした基本概念や構成ルールの仕様としてしばしば定義される．そして，機械（情報機器やソフトウェア等の人工物）同士，人間と機械，人間同士など，さまざまなエージェント間での共有情報の意味構造に関する合意を明示的に表示するためにオントロジーが有用と考えられる．記述方法が異なっていても，意味の同一性や差違に関する合意がオントロジーによって明確になるため，人間同士の知識共有を通じたコラボレーションの精度と生産性が高まるだろう．たとえば，大量の情報を厳密な解釈の下に共有する必要がある共同作業として大規模な情報システムの構築[5),6)]があるが，そのうち少なくともシステムの仕様設計の効率がオントロジーによって飛躍的に向上することが分かっている[1)]．

オントロジーに基づいて情報コンテンツを明示的に意味構造化することにより，他にもさまざまな仕方で知識循環を活性化することができる．たとえば，そのような構造を用いることによって情報検索や視覚化や翻訳や要約の精度が向上し，それが知識循環の活性化につながることは説明を要しないだろう．本稿ではその詳細に立ち

入らない．

以下では，オントロジーによる知識循環の活性化をさらに他の 2 つの観点から論ずる．第 1 に，次の章において，そのような意味構造を初めからコンテンツに含めておくことにより，コンテンツの作成コストを低減し品質を向上させることができることを述べる．第 2 に，その次の章において，オントロジーを介してコンテンツ間，サービス間の相互連携ができることにより，知識の再利用と拡大再生産が生じやすくなることを論ずる．

セマンティックオーサリング

セマンティックオーサリング（semantic authoring）[2] とは，オントロジーに基づいてコンテンツを作る作業である．典型的には，**図 -1** に示すようなグラフの形のコンテンツを人手によって作成することである．我々は，セマンティックオーサリングをサポートするコンテンツ作成支援ソフトウェアツールとして**セマンティックエディタ**（semantic editor）を開発中である．セマンティックエディタは Java Web Start で起動する Java アプリケーションであり，ローカルキャッシュを H2 ライブラリに基づくデータベースに格納し，サーバを介して多数の利用者の間でコンテンツを共有・共著できるグループウェアとしての機能を備える．

セマンティックエディタによって作成・編集できるコンテンツは何らかのオントロジーのインスタンスとしての実体ー関係グラフ（entity-relationship graph）である．その各ノードはそのオントロジーで定義される概念のインスタンス，各リンクは同じくそのオントロジーで定義される属性（property）のインスタンスを表す．ノードの内容はテキストや映像のショットや音声データであり，自然言語の 1 つの単文程度のまとまった意味内容を持つものとする．また，リンクはノード間の談話関係（因果関係や目的ー手段関係）などを表す（リンクを端点とするリンクもあり得る）．このようなやや大きな粒度のグラフ型コンテンツを**粗粒度知的コンテンツ**（coarse-grain intelligent content）と呼ぶ．

粗粒度知的コンテンツは，談話構造等を明示することにより，文字や音素の 1 次元の列で表示される通常の文章よりも，人間の伝えたい内容を明確に表現し伝達することができる．これに対し，たとえば各ノードが自然言語の単語程度の内容を持ちリンクの多くが文内の意味関係を表す，いわゆるセマンティックネットワークのような細粒度のコンテンツも考えられるが，そうした過度に詳細な構造化は，人間にとってのコンテンツの可読性を低下させ，作成コストを高めてしまう．通常の文章に相当する意味内容を表現する際には，各ノードが単文程度

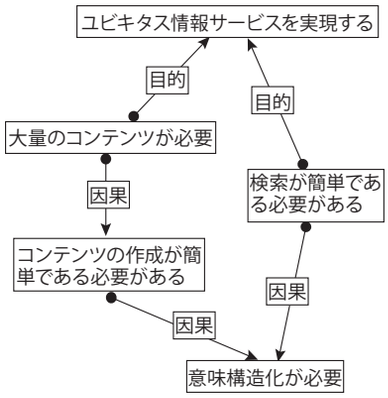

図 -1　粗粒度知的コンテンツ

の内容を持つ粗粒度知的コンテンツが人間にとって最も扱いやすい（理解しやすく作りやすい）というのが我々の予想である．そのような適切な粒度の意味構造を規定するオントロジーを共有することが社会的な知識の共有と循環のために肝要である．

次に，粗粒度知的コンテンツのセマンティックオーサリングによってコンテンツの作成がいかにして支援されるかを考えよう．ある文章（読売新聞 2005 年 5 月 14 日朝刊の原田泰氏の記事）の内容をセマンティックエディタによって人手で事後的に構造化したものを**図 -2** に示す．図の上の方の破線の内側はこの文章の骨子に当たる．つまり，この文章の要約はたとえば次のようになる．

> 70 年代の経済成長率は 10% から 3% に低下した．競争をやめて仕事を分け合おうとしたためである．それは生産性の高い製造業と生産性の低い非製造業が並存する二重構造経済をもたらした．低生産部門では飛躍が可能だから，非製造業を活性化すべきである．

図 -2 を見ると，この文章の前半（図の左の方に相当する）は事例や関連事項の説明を多く含むが，後半にはそれがほとんどないことが分かる．もしもこの文章の内容が最初からセマンティックオーサリングによって粗粒度知的コンテンツとして構造化されていたとすれば，文章の前半と後半のバランスをとるような改良がなされただろう．また，実は図の右上の「非製造業を活性化すべし」というノードに対応する内容は原文にはなかったのだが，これがないとグラフが非連結になってしまい，話がつながらない．実際，原文を読んでみると最後の方で腑に落ちない感じを受ける．逆にいうと，もしも最初からセマンティックオーサリングを使っていれば，当然ながらグラフは連結になり，筋の通った文章ができたはずである．このように，セマンティックオーサリングはコンテンツの品質を向上させる．

いわゆる発想支援ツールにも同様の効能がある．しか

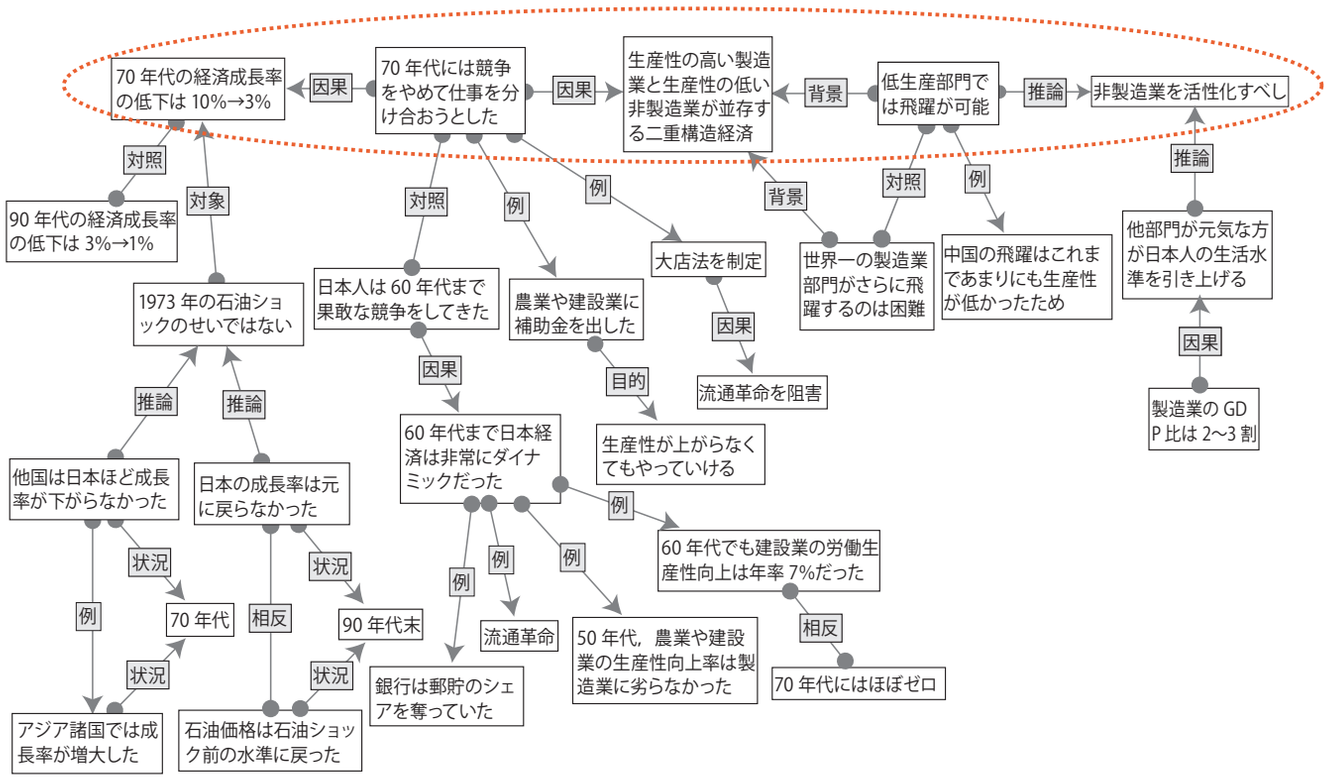

図 -2　文章の意味構造

し，従来の発想支援ツールでは，談話関係等の関係が規格化されていなかったため，リンクが標準的な関係でラベル付けされず，グラフ全体の意味が作成者以外には理解し難かった．したがってそれらのグラフは，通常の文章と違って，まとまった意味内容を伝達する手段たり得なかった．これに対し，セマンティックオーサリングにおいては，ISO/TC 37/SC 4 で策定中の標準的なオントロジーによって意味的な関係を規格化することにより，文章と同じ内容を文章よりも明確に粗粒度知的コンテンツで表現し伝達することができる．

セマンティックオーサリングによって検索や翻訳の精度が高まることは容易に想像できるだろう．そのほかにも，セマンティックオーサリングは，さまざまな観点からのコンテンツの分析を高度化する．たとえば，セマンティックオーサリングで構造化された議論においてはさまざまな発言の貢献度の高さなどを自動的に高い精度で判定することができる[3]．これにより，さらなる貢献の期待が高い参加者の発言を促したりそのような参加者を支援したりすることによって議論の品質を高めることが可能と考えられる．

コンテンツとサービスの循環と共創

セマンティックオーサリング等によりオントロジーに基づいて構造化された情報コンテンツにおいては，コンテンツの多くの部分の間での意味的な関連付けが体系的にできる．したがって，Web 上の異種のサービスの間でのコンテンツの相互運用が容易になる．これに基づき，利用者の意味のレベルにおいて Web 上のサービスを簡便に相互連携できるようにすることにより，利用者主導で新たなサービスが創出されるようになる．こうして，サービスとして具現化された知識が不特定多数の利用者の間で循環し拡大再生産される環境が実現できる．以下では，そこで用いるソフトウェアツールに関して述べる．

構造化コンテンツ運用環境 WebSLIT

セマンティックオーサリングはオントロジーに基づく明示的な意味構造を含めて新たなコンテンツを作ることだが，既存のコンテンツを事後的に構造化する必要が生ずることも多い．WebSLIT は，既存の Web コンテンツの意味的な構造化を支援する JavaScript のライブラリである．利用者の意図に応じて Web ページの中の構成素に意味的な構造化を施しながら取り出し，さまざまなサービスに仮想的にドラッグ & ドロップする仕組みを提供する．

これにより，Web 上のデータを保存するツール，および，保存したデータを選択し，指定の Web フォームにドラッグ & ドロップで入力するツールが実現できる．専用のクライアントと連携させ，**図 -3** のように，コンテンツの部分抽出と，その蓄積ならびに構造化を，Web ブラウザ上でのドラッグ & ドロップ操作で実現している．表示と内部構造を分離して扱っているので，フォームの

図 -3　WebSLIT によるクリッピング

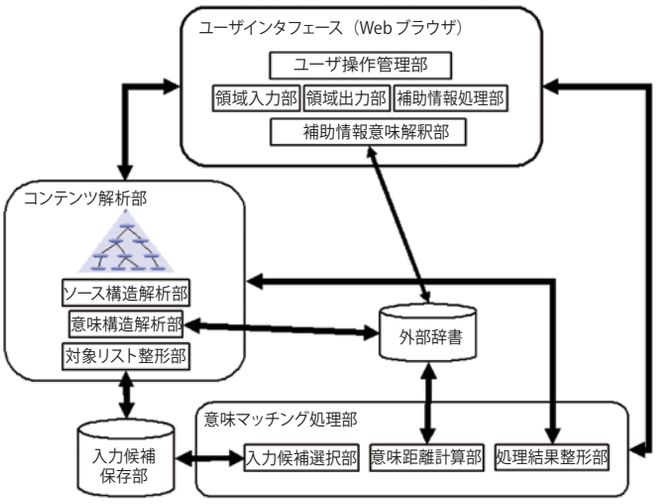

図 -4　WebSLIT のアーキテクチャ

ようなサービス（動的なコンテンツ）のクリッピングも可能である．図 -3 では，Yahoo! の検索サービスの部分（右側の赤い破線の囲み）をクリッピングして蓄積し，それを作成中のコンテンツの中にドラッグ & ドロップ（左側の赤い破線の囲み）している．

ここで重要なのは，Web ページの中の統語的な構成素（HTML のエレメントなど）が単にクリッピングされて貯められるのでなく，クリッピングの際にその構成素が意味的に構造化されるということである．その構造化とは，たとえば「この Web フォームへの入力は ISO 8601 の構文に従う文字列であって日付を意味する」のように，クリッピングされた構成素の各部分の統語論と意味論に関する一種のアノテーションである．このような仕方で構造化されたサービスは後述の YOSEE のような技術によって意味的に相互連携可能になる．したがって，そうしたアノテーションのメタデータを社会的に共有することにより，多様なサービスに関する知識が不特定多数の利用者の間で共有されながら協調的に構造化され，利用者主導でサービスが連携して新たなサービスが絶え間なく創出されるようになるだろう．

図 -4 に示すように，WebSLIT は，Web コンテンツを提示するためのブラウザ機能をフレームワークとして，利用者操作管理部や領域入力部，領域出力部，補助情報処理部，補助情報意味解釈部のサブモジュールがフレームワークにプラグインされるかたちで動作する．

コンテンツ解析部は，図 -3 に示したようなクリッピングの際に Web コンテンツの統語的な構造を解析するソース構造解析部，外部辞書を参照してその構造の中の各部分に意味情報を付加する意味構造解析部，意味情報と操作情報を付加したソースとして整形する対象リスト整形部から構成される．対象リストとして整形された Web コンテンツは，入力候補保存部に保存され，適宜，意味マッチング処理部から読み出される．

意味マッチング処理部は，同じくクリッピングの際に，入力候補保存部に保存されたマッチング処理の候補を選択する入力候補選択部，外部情報を参照しつつソースを構成するプリミティブ間の意味距離を計算する意味距離計算部，処理結果を整形する処理結果整形部からなる．

外部辞書は，前記の構造解析において抽出された文字列に意味情報を付加するための概念辞書や同義語辞書，概念間の意味距離を定義するための概念階層などから構成される．

入力候補保存部は，意味情報と操作情報を付加して整形された構造を保存する部分で，単一ホスト内での動作の場合は OS のカットバッファを利用するが，一般には，Web アクセス可能な SQL サーバ上に構成される．

コンテンツ駆動型アーキテクチャ YOSEE

YOSEE（Yarn Of Semantically Enhanced Entities）は，インターネット上で流通可能なさまざまな情報コンテンツを作成，蓄積，公開するための CMS（コンテンツマネジメントシステム）である．YOSEE の目的は，主に次の 3 つの機能を提供することにある．

「貯める」機能：Wiki，ブログ，スケジュール，デジカメの画像や動画，Web ページのクリッピング情報，ファイルなど電子的データを蓄積する．

「つなぐ」機能：蓄積したコンテンツを組み合わせることによって新たなコンテンツの作成を支援する．

「公開する」機能：蓄積したコンテンツを見つけやすくするための検索の仕組みを提供する．

これらの関係を**図 -5** に示す．

YOSEE では，共同文書管理の仕組みである Wiki 機能をインタフェースとして以上を実現する．Wiki を拡張することにより，YOSEE では通常の文書のみならず，画

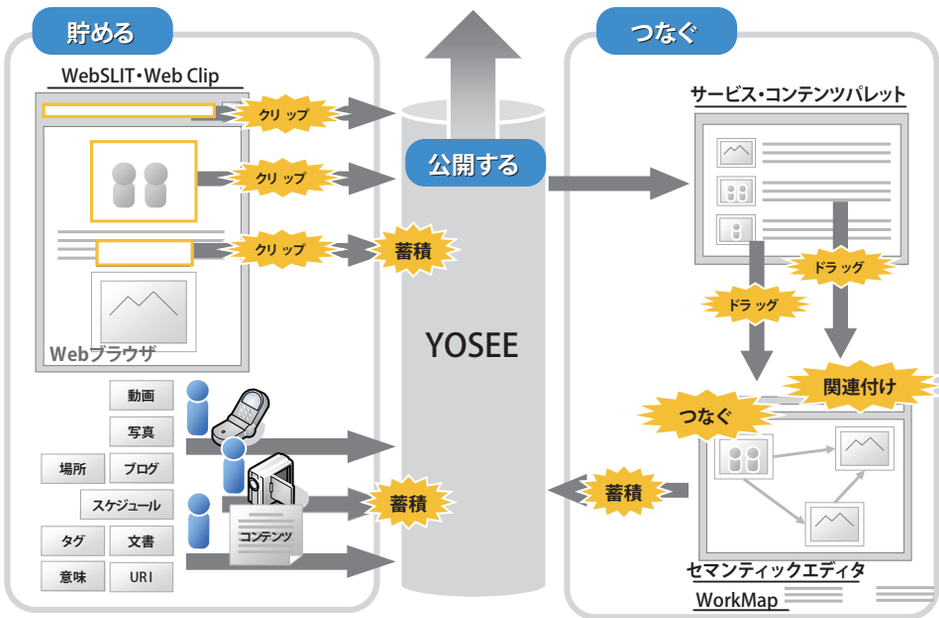

図 -5　YOSEE と WebSLIT に基づくコンテンツ駆動型アーキテクチャ

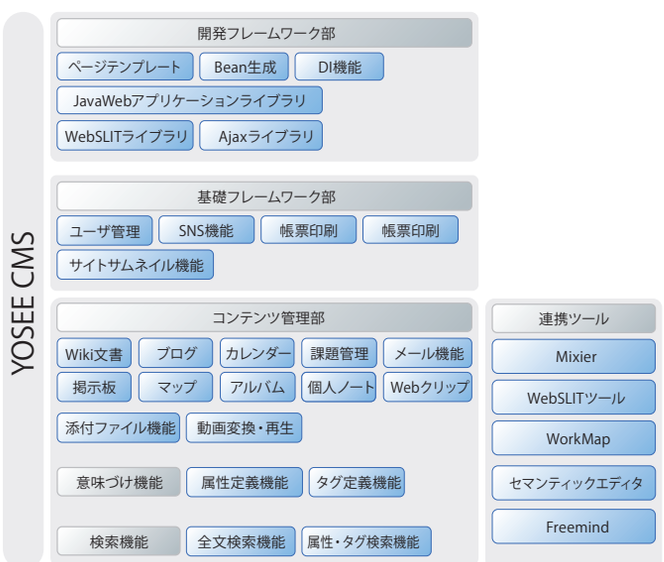

図 -6　YOSEE の機能構成

図 -7　蓄積したコンテンツへの属性の付与

像，動画のほかにも多くのファイルを複数の利用者が協調しつつ蓄積管理することができる．

YOSEE はまた，コンテンツを蓄積する際に施される意味的な構造化により，意味に基づくコンテンツ同士の関連付けを支援する．この意味的な関係構造を使えば，全文検索とは異なる意味に基づく知的なコンテンツ検索が実現できるだろう．

YOSEE を構成する主な機能モジュールは**図 -6** の通りである．以下では，これらのうちで前記の「貯める」機能と「つなぐ」機能を担う連携ツールに関して具体的に述べる．

(1) 拡張 WebSLIT

WebSLIT によって前述のように蓄積したコンテンツは，**図 -7** のようにリストアップしてタグや属性を付けたり，またグループ分けをして後で検索したりすることができる．

(2) Mixier

Mixier は，YOSEE 内に蓄積されている粗粒度の単位コンテンツを組み合わせることによって新たなコンテンツやサービスに相当する Web ページを構築するためのソフトウェアツールである．蓄積されたコンテンツから Mixier によって要素を取り出して他のコンテンツと組み合わせる様子を**図 -8** に示す．これは，赤い破線で示すように，蓄積してあった Goo 乗り換え案内を取り出して複合的なサービスの部品として組み込んでいるところである．

(3) WorkMap

Mixier によって組み合わされたサービスの間を意味的

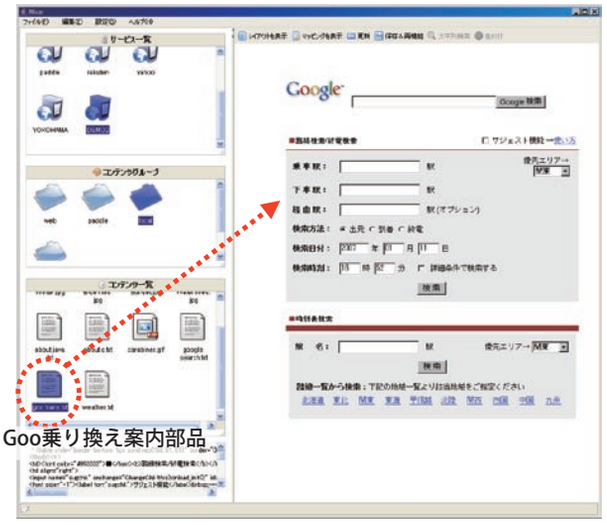

図 -8　蓄積コンテンツの組合せ

図 -9　コンテンツ間連携の編集

に関連付けることによってそれらのサービスを連携させ，複合的なサービスを作成するためのグラフィカルな編集ツール WorkMap を実装しつつある．YOSEE のコンテンツを取り込んで関連付けや編集ができる機能を実装する予定[4]である．**図 -9** に WorkMap のユーザインタフェースを示す．これは，Yahoo! カレンダーや goo 乗換案付等のサービスの間での入出力の受け渡しによる連携を入出力の意味的な整合性を担保しつつ設定している様子である．このようなことが簡単にできるのは，前述の通り，サービスをクリッピングして蓄積した際に各サービスの入出力の統語的・意味的構造に関するアノテーションが施されているからである．

知識循環型社会の展望

オントロジーに基づいて（サービスを含む）コンテンツの意味構造を明示し，それによってコンテンツの作成や利用を不特定多数の利用者が主導するかたちで高度化する技術について述べた．これらの技術は一般に，明示的に共有された意味内容に基づいて個人の行為や社会的なインタラクションがなされるようにする効果を持つ．

たとえば，セマンティックオーサリングのような仕方でさまざまな議論が構造化されることにより，各主張がその根拠や論拠と明示的に結び付けられる（または根拠や論拠を欠くことが明示される）．こうして多数決や人気投票によらずに意味内容に即してそれらの主張の妥当性を判断することが容易になる．また，WebSLIT と YOSEE のような仕方による利用者主導でのサービス連携が普及するにつれて，連携によって生まれる多数の複合サービスの検索は Page ランク等の意味での人気やキーワードの出現ではなく意味構造に基づいてなされるようになるだろう．

こうして，本稿で述べたような技術が普及することにより，ブランドや人気や権威や多数決や Page ランクではなくその意味内容や品質や科学的信頼性に応じてさまざまなコンテンツが流通し利用される，「知識の完全競争市場」が現出すると考えられる．それが知識循環型社会の究極の姿であろう．

参考文献

1) Kondo, K., Hoshii, S., Morita, T., Yamaguchi, T., Izumi, N. and Hasida, K. : Semantics Driven Development of Software Systems Based on Business Ontologies, Knowledge-Based Software Engineering, Frontiers in Artificial Intelligence and Applications, Vol.140, pp.176-185, IOS press (2006).
2) 橋田浩一 : オントロジーと制約に基づくセマンティックプラットフォーム，人工知能学会誌，Vol.21, No.6, pp.712-717 (2006).
3) Kamimaeda, N., Izumi, N. and Hasida, K. : Discovery of Key Persons in Knowledge Creation Based on Semantic Authoring, The Learning Organization : An International Journal, Emerald (2007, to appear).
4) 産業技術総合研究所プレスリリース : 産業変革を先導する戦略的な産学官連携プロジェクトを開始，http://www.aist.go.jp/aist_j/press_release/pr2005/pr20050713/pr20050713.html (2005).
5) 産業技術総合研究所プレスリリース : 開発企業のしばりから解放された大規模情報システムの開発に着手，http://www.aist.go.jp/aist_j/press_release/pr2006/pr20061219/pr20061219.html (2006).
6) 横浜市役所記者発表 : http://www.city.yokohama.jp/me/gyousei/it/news/news061214.pdf (2006-12-14).

（平成 19 年 7 月 9 日受付）

橋田　浩一（正会員） hasida.k@aist.go.jp

産業技術総合研究所情報技術研究部門長．専門は自然言語処理，認知科学，言語学，知的コンテンツ．最近の興味は，セマンティックコンピューティングおよびその応用としての文脈依存型情報サービス，知の社会的共創など．

和泉　憲明（正会員） n.izumi@aist.go.jp

1969 年生．産業技術総合研究所主任研究員．1996 年大阪府立大学大学院博士後期課程（3 年次）を中途退学し，1996 年静岡大学情報学部助手，2002 年より，現所属．博士（工学）（慶應義塾大学）．知識モデリングの観点から知識管理の研究に従事．最近は，セマンティック Web と大規模情報システムの融合を試みている．