学内のデジタルコンテンツの公開と「知の編纂（へんさん）と再編」のチャレンジ

慶應義塾大学デジタルメディア・コンテンツ統合研究機構では、学内で運用されているデータベースやWebサーバーに蓄積されたさまざまなデジタルコンテンツを、学外の一般利用者に広く公開するDMCシステムのテスト運用を2006年8月にスタート。本番運用に向けて、引き続き研究開発を進めています。

DMCシステムの検索エンジンには、IBMが提唱し、現在はオープンソースとなっている自然言語処理機能の統合技術UIMAを採用したIBM OmniFind™ Enterprise Editionを導入し、検索語と関連の深いキーワードも同時に検索できるなど、高度な検索能力を実現しています。

DMCシステム構築プロジェクトのプロジェクトリーダーである慶應義塾大学 デジタルメディア・コンテンツ統合研究機構 助教授 嶋津 恵子氏に、プロジェクトの概要を説明していただくと同時に、嶋津氏がスピーカーとして参加されたIBM Information On Demand 2006の印象についても語っていただきました。

Interview ❷

Publication of university digital content and the challenges of “compiling and reorganizing knowledge”

In August 2006, the Research Institute for Digital Media and Content (DMC) at Keio University launched a test operation of its DMC system, which makes available to non-university users at large, a broad range of digital content stored on the databases and web server operated at the university. They are currently proceeding with their ongoing research and development in preparation for full-scale operation of the system.

IBM OmniFind™ Enterprise Edition has been adopted for the DMC system search engine. It is strongly recommended by IBM, and it utilizes the UIMA integration technology of the natural language processing function which is currently open source. Such a search engine allows for an advanced search capability, including the ability to simultaneously search for keywords that are of great relevance to the search term.

We spoke with Ms. Keiko Shimazu, an associate professor with the Research Institute for Digital Media and Content (DMC) at Keio University, and who is the leader of the project building the DMC system. We asked her to provide an overview of the project, and spoke with her on her impression of IBM Information On Demand 2006 at which she participated as a speaker.

DMC機構の設立と DMCシステム構築プロジェクト

慶應義塾大学デジタルメディア・コンテンツ統合研究機構（以下、DMC機構）は、文部科学省科学技術振興調整費の「戦略的研究拠点育成プログラム」いわゆるスーパーCOE（Center of Excellence）の採択を受けて、慶應義塾大学により2004年7月に設立されました。

デジタルコンテンツ素材を利用目的に沿って生成・編集・加工・統合することによりデザインされ、かつシナリオを持ったデジタルコンテキストの統合的構築および人材育成を行うことが、その目的です。

DMC機構では、毎年、研究テーマや研究員を募集し、同時に幾つかの研究プロジェクトを進めています。

その中で「特命プロジェクト」は、DMC機構が行うべき課題を解決し、特定の業務を遂行するためのプロジェクトとして位置付けられ（**図1**）、わたしがプロジェクトリーダーを務めている「DMCシステム構築プロジェクト」もその一つとして2005年度にスタートしました。

慶應義塾大学
デジタルメディア・コンテンツ
統合研究機構
助教授
博士
嶋津 恵子氏

Ms. Keiko Shimazu, Ph.D.
Associate Professor
Research Institute for
Digital Media and Content (DMC)
Keio University

デジタルコンテンツの価値を共有するために

DMCシステム構築プロジェクトの狙いは、慶應義塾大学のネットワーク上にデジタルコンテンツとして蓄積されている研究成果や学術成果を、学外にも広く開放し、専門家だけでなく一般の方々も自由に利用できる仕組みを構築することです。そのために、コンテンツを共有し、発信するために必要な機能を実現するシステムの設計・開発に取り組んでいます。

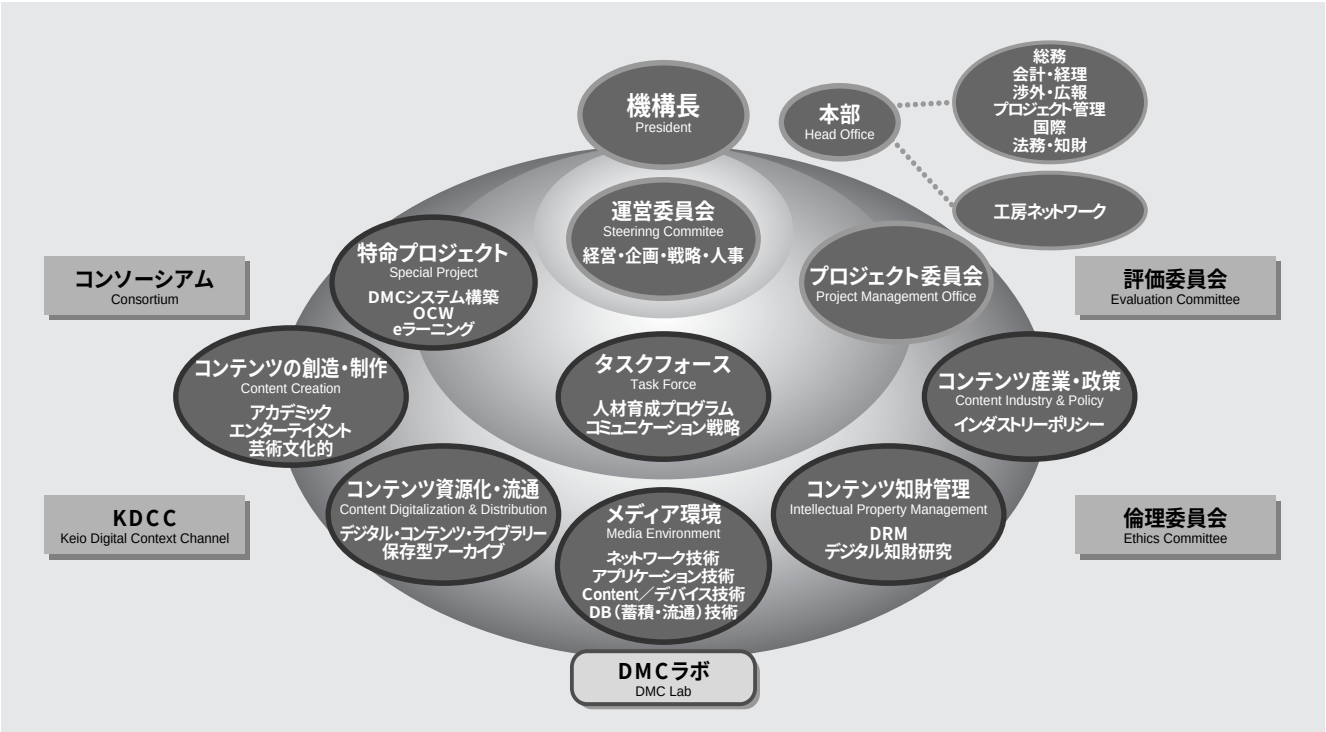

図1. DMC機構の組織図

インターネット上では既にさまざまな検索エンジンが利用され、キーワードの入力によりコンテンツをリストアップできるようになっています。しかしながら、慶應義塾大学の研究者・研究室が保存しているさまざまなデジタルコンテンツは、多種多様なアプリケーションで作成され、保存形式も異なります。学外の利用者にも広く公開し、効果的に利用してもらうには、単なる検索エンジンの能力では力不足です。

例えば、専門領域ごとに使われている用語が異なることも多いため、専門外の利用者がダイレクトにキーワード検索を行っても、必要なコンテンツを取り出せるとは限りません。検索結果を効率的・効果的に絞り込むためのキーワードの候補を提示する仕組みが必要です。

もう一つ、単なる検索エンジンは、ターゲットコンテンツが存在することを前提とし、それを探し当てることしかできません。わたしたちの目標は、ターゲットコンテンツを高速かつ正確に探し出すのはもちろん、利用者が新たな研究テーマに取り組んだり、新しく何か物事を考え出そうとしたときに、その発想を支援する材料を提供することです。利用者が設定したシナリオに従って、すなわちコンテキストに従って既存のコンテンツを自由に取り出し、並び替え、見方を変えることによって、新しく何かをクリエイトするときの助けになる仕組みをつくろうということです。

民間企業のシステムエンジニアから転身

実はわたしは、ドキュメント管理ソリューションの開発メーカーに長年在籍し、システムエンジニアとしてネットワークシステムやエキスパートシステムなどの設計に携わりました。エキスパートシステムとは、専門分野に特化した知識データベースを基に情報を解析し、利用者の問題解決を手助けするシステムです。そして最後の2年間は、研究部門で帰納論理プログラミングの応用技術の研究開発に携わりました。

民間企業での研究開発を辞めて、DMC機構に移った理由は、今回のような大規模システムの全体設計にかかわることができるのはとても魅力的だったからです。企業では、大規模システムの構築に部分的に携わることはできても、かなりの権限を持った立場にならないと、システム全体のデザインを任せられることは極めてまれだからです。

それともう一つ。プロジェクトそのもののスケールも特徴的だと感じました。理工学系の研究者によるコミュニティーの研究プロジェクトは幾つか経験しましたが、DMC機構では、専門分野を横断し、さまざまな研究分野の知識や経験を統合活用することを目指しています。また完成するシステムの想定利用者が特定の業務システムの枠に特定されていないことなど、プロジェクトそのもののスケールも大きな魅力でした。

ナレッジマネジメントの研究成果利用に着目

専門分野を横断し、さまざまな研究分野の知識や経験の活用は、システムのグランドデザインをする際にも試みられています。

具体的には、1990年代後半に新しい経営手法の研究としてブームとなったナレッジマネジメント（Knowledge Management）で提案されている手法の利用です。これは、背景が異なる専門家同士が、それぞれが持つナレッジを共有する方法であり、具体的には5W1Hで特定できる情報を確実に引き渡すというものです。

同様の主張はリチャード・S・ワーマンが「それは『情報』ではない」という著書の中で「インフォメーションデザイン」の観点から、LATCH（Location：位置、Alphabet：アルファベット、Time：時間、Category：分野、Hierarchy：階層）が示された段階でその本質が伝わると述べています。

そこでわたしたちは、「人」「場所」「時」の三つの要素が重要であると考え、これらを使って、コンテキストを表現し、コンテンツ検索に応用することにしました。

検索エンジンにOmniFindの導入を決定

この考え方に基づき、デジタルコンテンツを効果的に検索する仕組みを実装していくことになりました。データベースやWebサーバーに蓄積されているデジ

タルコンテンツを、システム横断的に検索して取り出す仕組みは、IBM OmniFind Enterprise Edition（以下、OmniFind）をはじめ、企業の情報共有向けに開発された検索エンジンが存在していることから、その機能をDMCシステムのミドルウェアとして利用することにしました。

そこで各ベンダーが企業向けに開発した検索エンジンを比較するために、次の三つの評価基準を定めました。

・日本語処理能力

高品質な日本語処理能力を備え、特定の固有名詞を取り出して、それが人名なのか、地名なのか、どんな属性なのかを自動的に判断して振り分けられる機能を備えていること。

・スケーラビリティー

大量のコンテンツを高速で扱えるだけではなく、学内の研究者や研究室がそれぞれの目的に応じて運用しているデータベースシステムやリポジトリー、Webサーバーなどから、システムの違いにかかわらずに情報を収集できる透過性を備えていること。すなわち、学内のネットワーク上に散在している各種リソースを、保管場所やシステムの違いなどを意識しないで利用できること。

・モジュールの組み込み可能性

研究のベースとなるシステムとして、プロジェクトの研究成果を反映させながらモジュールを組み込むことで、スパイラルアップによりらせん階段を上っていくように機能を拡張できること。

その結果、OmniFindが要件を最も満足していると判断し、2005年6月に導入を決定しました。

広範囲なデータソース、ファイルフォーマットをサポートし（**図2**）、学内に蓄積されている多種多様なデジタルコンテンツを大量かつ高速に検索できることが決め手になったのです。

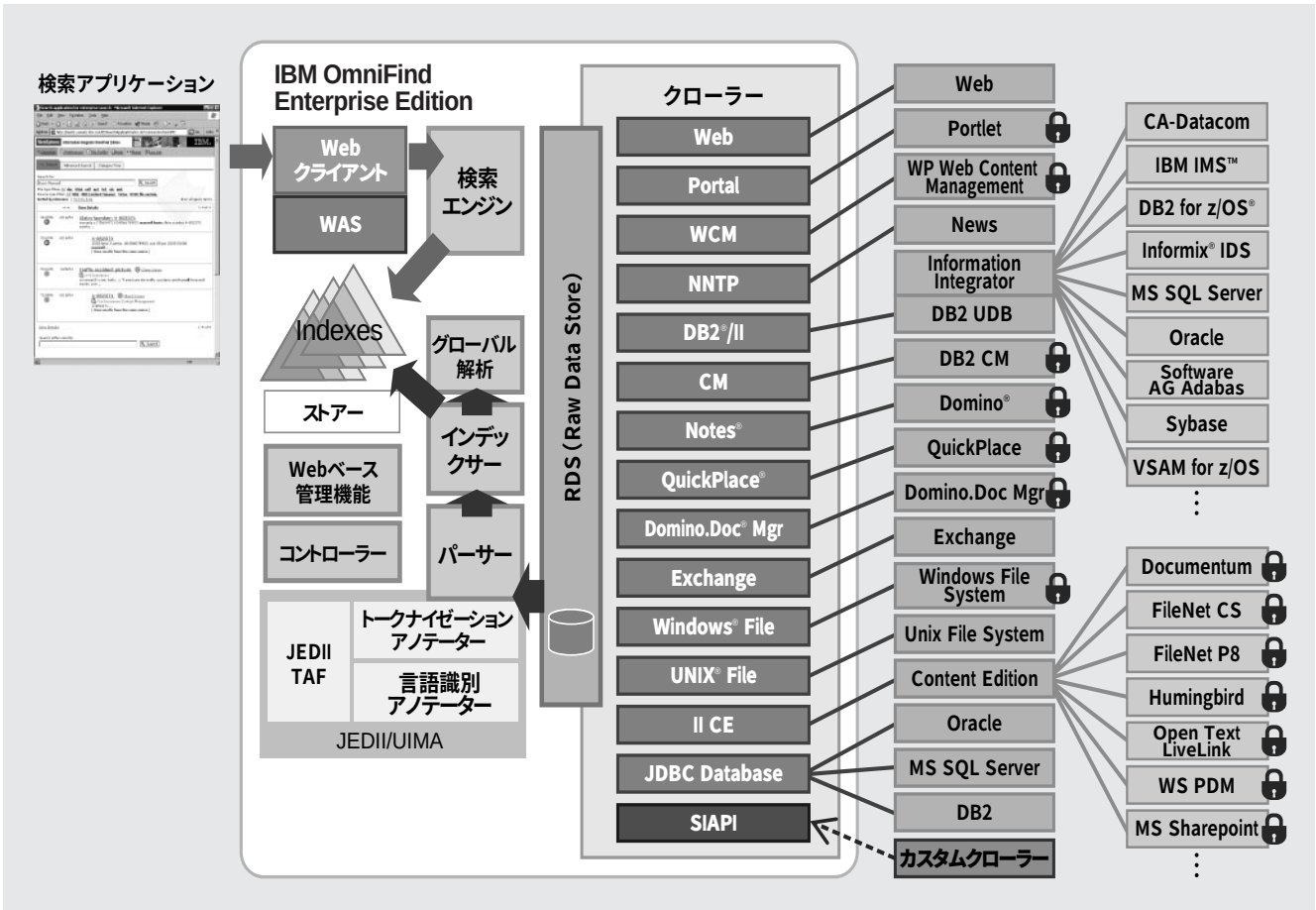

図2.　OmniFindのコンポーネント

UIMAに基づく自然言語処理技術を活用

わたしたちはOmniFindが採用しているUIMA（Unstructured Information Management Architecture）という技術にも注目しました。

UIMAは、構造化情報だけでなく非構造化情報を処理するソフトウェア同士を連携させるためのアーキテクチャーで、自然言語処理技術を簡単に組み合わせることができるため、キーワードだけでなく、さまざまな関連性や意味を発見するのに役に立つ技術です。

非構造化情報とは、電子メールやWordファイルなどの一般的な文章に含まれる情報であり、人間なら文章の内容を判断して、さまざまな関連性や意味を発見できます。一方、コンピューターは、例えばExcelやCSV（Comma Separated Values）形式などに代表される構造化情報の処理は得意ですが、非構造化情報の処理は苦手としていました（**図3**）。

UIMAは、さまざまな解析技術を使って非構造化情報を処理することができ、その機能をOmniFindに組み込むことで、文章が持つさまざまな関連性や意味を発見することを可能にしました（**図4**）。

構造化情報：

名前	職業	会社名	年齢	オフィス・エリア
山田太郎	エンジニア	ABC株式会社	29	大和
田中花子	デザイナー	XYZ株式会社	32	東京
斉藤健太	営業	XYZ株式会社	42	大阪

非構造化情報：

山田太郎はABC株式会社に勤めており、エンジニアの仕事をしています。彼の年齢は29歳で、事業所は神奈川県の大和にあります。…

図3.　構造化情報と非構造化情報

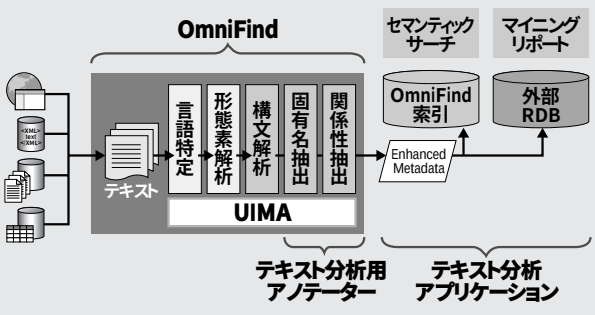

図4.　UIMAに基づく自然言語処理機能とOmniFindの統合

日本アイ・ビー・エム株式会社 東京基礎研究所で開発された固有表現を抽出する機能と、わたしたちが提案する5W1Hの中の重要要素である「人」（who）、「場所」（where）、それに「時」（when）の変換機能をUIMAを通じて統合することで、コンテンツをコンテキストに従って分類できると考えました。つまりわたしたちの研究テーマは、コンテキストを表現するための属性の特定とそれらの組み合わせ方であり、その要素技術としてUIMAを採用したということです。コンテキストの表現に関しては、人工知能の研究分野で著名なJohn MaCarthyの様相論理による数学表現、すなわち「～でなければならない」「～であり得る」「～べきである」といった可能性や必然性にかかわる命題を扱う論理学をベースにしています。

DMC機構が開発したコンテンツシステムは、デジタルコンテンツの検索機能を優先して2005年8月にリリースしました。先に述べたように、この機能はOmniFindを利用したものです。

そして現在は、コンテンツが持つ、もしくは関係する「人」「場所」「時」を使って、コンテキストによる分類を行い、コンテンツをダイナミックに配列・配置し直す試みも始めました。コンテンツが学術成果である場合、これはまさしく「知の編纂（へんさん）」の狙いに一歩近づいたことになります。

IBM Information On Demand 2006に参加

2006年10月、6日間にわたって米国・アナハイムでIBM Information On Demand 2006（62ページ参照）が開催され、わたしはスピーカーの一人として、今回のDMCシステム構築プロジェクトについて発表する機会を得ることができました。

このコンファレンスは、800以上のセッションを数え、来場者は5,000人を超える大規模なものでした。

外国での論文発表は、今までにも何回か経験がありますが、今回のセッションはまったく印象の異なるものになりました。アカデミックな場では、コンピューターサイエンスや基礎技術、要素技術についての発表が中心であり、最近は実際のシステムにどう応用・

展開されたかということも重要視されるようになってきたものの、質疑応答などではやはり数学的な表現への展開や、科学的な実験に基づく視点での質問がほとんどです。

ところが、今回のコンファレンスでは「UIMAをどう使ったのか」とか「OmniFindのどこを拡張したのか」「ユーザーが何を求めているのか」「どうすれば使いやすくなるのか」といった質問を多く受けました。中には「そのユーザーインターフェースのチェックボックスはどういう仕組みになっているのか」とか、「ソースが欲しい」というリクエストまであったほどです（笑）。

IODの広がりと深みを実感

コンファレンスは6日間にわたって開催されましたが、特に印象に残っているのは、IBMの圧倒的な地力というかパワーを感じたことです。IOD（Information on Demand）というテーマを論じ合うために5,000人ものエンジニアが1カ所に集まるというのは注目に値すると思いました。

IODというキーワードでこれだけ多くの人が世界中から集まって、熱心に討論するということは、単なる興味だけでなく、問題意識・課題意識を持っている方が世界中にそれだけたくさんいるということでしょう。

また、IODという考え方が、横に広がりを持つと同時に、深みが出てきたということも強く感じました。

IODは、エンドユーザーに情報を渡す出口のところが注目されがちですが、実際には、どのようにデータを持たせるか、どんな処理をさせるのか、いつ処理させるのか、といったことがすべて出口にかかわってきます。それにもかかわらず、今まで出口の辺りばかり見ていて、その内側に対する議論が少ないと感じていました。それが今回のコンファレンスでは、出口だけではなく、例えばコンテンツの設定や、メタデータの持たせ方についても、多くの技術者や技術にかかわる方が興味を持っていました。

熱心な討論が繰り広げられたBOF

コンファレンスでは、わたし自身の発表の時間以外には、BOF（Birds of a Feather）と呼ばれる討論会に参加しました。BOFとは「同じ羽毛の鳥」という意味合いで、あるテーマに興味を持つ方たちが集まって、形式にとらわれずに技術的な課題を議論し、親睦を深めるために行われる討論会・座談会のことです。

今回のコンファレンスでは、テーマごとに明確な問題意識を持つ人たちが集まり、真剣に議論を交わす場となっていました。特に、顧客向けのシステム開発や導入に際し、プロダクトのインテグレーションの仕方や、システムのグランドデザインにおける重要ポイントなど、要素技術の開発ではなく、実利用に焦点を当てた利用方法や応用方法に議論が集中していたことが印象的です。

次の機会には、さらに高度な応用技術の発表を

今回はわたしにとって、企業が開催するコンファレンスでの初めての発表経験でした。そこで、わたしたちの研究テーマから実際に動いているシステムのイメージ、またOmniFindやUIMAのミドルウェアとしての利用方法など、幅広く話題を準備しました。一方、発表に参加された方々の期待は、プロダクトの応用方法に集中していたようです。これはいただいた質問やコメントから容易に理解できました。例えば、GUI（Graphical User Interface）との連携部分など、ソースコードを出しながら説明するぐらいでも良かったのではないかと思っています。

もし、仮に来年も自分のセッションを持つことができるのなら、慶應義塾大学におけるDMC機構の狙いとわたしのプロジェクトの目標をより詳しく説明し、それに対してどういうアプローチで目標に到達しようとしているのかを具体的に説明したいと考えます。そうすることにより、より正確にOmnifindやUIMAを選択する理由が明らかになると思うからです。特に今回触れることのできなかった「知の再編」について、参加者の方々からご意見をいただきたいと思います。

「知の再編」とは、今までアナログで処理されていた知的アウトプットがデジタル化されることで、知りたいことをすぐ知ることができるようになるということです。つまり、今まで蓄積されてきた知の資産を、それが作られた順番ではなく、自由に並べ直すことで、新しい可能性が広がるということです。「ビジュアライゼーション」（可視化）という言い方もできますが、見え方をコンテキストに従って並べ替えていくことで、人間が直接「見る」ことのできない情報を「見る」ことができるようになります。それがデジタルコンテンツの良さなのですから。

例えば、福澤諭吉に関するコンテンツについて、それがどこで保管されているのか世界地図上にマッピングすると、ヨーロッパと米国西海岸、東京、福岡に集中しているということが一目で分かります。コンテンツのリストを一覧表示して都市名が表示されるだけで終わるのか、それとも世界地図上にどのように分布しているのかが分かるのでは、そこに大きな違いがあります。まさにビジュアライゼーション上の大きな改革ということもできるでしょう。

まだテスト稼働の段階ですが、DMCシステムは既にURL（Uniform Resource Locator）が公開されていますから（http://context.dmc.keio.ac.jp/）、その違いをぜひ体験していただきたいと思っています（**図5〜9**）。

図5.　デジタルメディア・コンテンツ統合システムのホームページ

図6.　全文検索の例
（「福澤諭吉」で全文検索を行った結果）

図7.　メタデータ検索の例-1
（「福澤諭吉」で以下のメタデータを入力した検索前画面
メタデータ：人＝福澤諭吉、場所＝三田、時＝2008年）

図8.　メタデータ検索の例-2
（「福澤諭吉」で以下のメタデータを入力した検索結果画面
メタデータ：人＝福澤諭吉、場所＝三田、時＝2008年）

図9.　メタデータ自動抽出の例

「知の再編」に向けて

わたしは民間企業の出身ですから、その経験を踏まえて今回のプロジェクトを検証してみると、実用システムを設計開発するための体制が必要なのではないかということを感じました。

DMC機構の成果創出に向けてプロジェクト活動を進めていく中で、いつも悩んでしまうのは、大学の理工系の研究者が実用システムを開発する機会があまり持てないことによる、経験知の不足です。大学内のプロジェクト活動も、企業におけるそれと同様にトップダウンで動いていくことのメリットをもう少し取り入れるべきではないかと感じていますし、わたし自身としても、今後はそういった提案活動も行っていく必要があるかなと思っています。

IBMに対する要望としては、自然言語処理技術、つまり人間が日常的に使っている言語をコンピューターに処理させる技術については、特に固有名詞抽出技術にさらに磨きをかけることで、自然言語から人名や地名、組織名などを抽出する精度をさらに上げて、技術的な先進性と実用性を両立してもらえたらと思います。日本語処理の精度が高まることで、わたしたちの研究のベースとしてさらに役に立つはずです。

学内のデータベースやサーバーからコンテンツ情報を集めるクローラーについては、安定性と操作の簡便性を高め運用の手間をさらに省けるようになると助かります。クローラーとは、周期的にWeb上を巡回して文書や画像などを取得し、自動的にデータベース化するプログラムのことです。例えば、クローラーが巡回先でコンテンツを探し出してこられなかったときに、クロールし直す方法は何段階かあり、クローリングしてきた結果を書き直す方法も何種類かあります。しかし、どんなときにどの方法を採れば最も効果的なのか、実際にやってみなければ分かりません。そのため情報を全部消して最初からやり直すことも多いのですが、それを繰り返している限り、運用手順書に落として標準化することができません。こうした作業は可能な限り研究者の手から離して、システムの運用スタッフに任せられるようにしたいと思っています。わたしたちも工夫していくつもりですが、IBMにも支援をお願いしたいですね。

先ほども述べたように、わたしたちのプロジェクトの狙いは、既存のコンテンツを高速かつ正確に見つけ出すだけではなく、「知の再編」という知的活動を支援するシステムの構築です。ある意味では、IBMが取り組んでいるIODと目指すところは近いのではないでしょうか。その意味でも、お互いに協力できる部分で、より良いパートナーになっていければと思っています。