セマンティック・ウェブをひもとく：学術出版物にとってどのような意味があるのか

このプレゼンテーションでルイーズ・タットン氏は、セマンティック・ウェブの意味と、現代の出版業界内の最終的なメリットを解読します。インターネットは、研究者がコンテンツを発見・消費する方法に大変革を起こしましたが、ウェブというメディアの全潜在能力はまだ実現していません。セマンティック・ウェブはこの進化するプロセスの次のステージを描き、出版者とそのエンドユーザーにわくわくするようなチャンスを提供します。タットン氏は、実生活の例を論じ、セマンティック・ウェブが出版者と学会のために開くことのできる可能性の幾つかを探ります。

ルイーズ・タットン
（パブリッシング・テクノロジー　学術部門　シニア・バイスプレジデント）

ルイーズ・タットン氏はパブリッシング・テクノロジーの学術部門のシニア・バイスプレジデントとして、顧客管理や図書館サービス、商品開発、エンジニアリング、広告、新市場探索など、同部門の管理運営に責任を負っています。タットン氏は電子出版業界で 12 年の経験があり、この期間は顧客関係に主に携わってきましたが、編集およびプロジェクト管理の経験もあります。CatchWord（キャッチワード）と Ingenta（インジェンタ（現在のパブリッシング・テクノロジー））に加え、Taylor & Francis（テイラー＆フランシス）、ABC-Clio（ABC- クリオ）で歴任したポジションによって、幅広い専門知識と実務経験を得ています。タットン氏は世界中のコンファレンスで定期的にスピーチを行い、学術・学会出版を主に扱う幾つかの委員会のメンバーでもあります。

インターネットとセマンティック・ウェブの紹介

ウェブ、言い換えるとインターネットが研究やデータ発見の方法を劇的に変えたと言うことは重要だと考えます。研究技術は大きく変わりましたが、ウェブは今でも非常に伝統的な印刷モデルを踏襲しています。ジャーナルは「ジャーナル・号・論文」のモデルを踏襲し、本は「本・章」のモデルを踏襲しています。ウェブ上には入手可能な情報が大量にあるために、かえってユーザーが探しているものを素早く見つけることが難しくなっています。私たちが直面するもう一つの問題は、コンピューターは構文を理解しますが、意味は理解しないということです。従って、同義語や文脈を常に理解するわけではなく、他の関連資料を提供できるわけでもないということです。検索エンジンは便利ですが、検索語に関連し得るあらゆるコンテンツをやみくもに引っ張ってくることもしばしばです。10 ～ 20 ページの研究を提示され、もともと探していたものを見つけることがむしろ難しくなることもあります。現在、ウェブは実際のところ、個別のウェブサイトのウェブ（網）にすぎません。従って、アプリケーションも個々に独立して機能しがちで、他のウェブサイトと相互に作用することもあまり多くはありません。可能なほどには、互いに統合・接続していないのです。

セマンティック・ウェブとは何か。こうした問題をどのように解決できるのか。――比較対象として、ワールドワイドウェブは、ご存じのとおり文書のウェブ（網）で、一部は孤立して機能しています。一方、セマンティック・ウェブは、データのウェブ（網）であり、接続することやウェブサイトを統合することを伴います。約 5 年前、ティム・バーナーズ＝リー卿は、セマンティック・ウェブを「表示するためだけでなく、さまざまなアプリケーションをまたいでデータを自動化・統合・再利用するために、マシンで使用可能な方法で定義・リンクされたウェブ上のデータ」と表現しました。

セマンティック・ウェブでは、どのようにして上記の問題に取り組めるようになるのでしょうか。セマン

ティック・ウェブは、現在のウェブの強化型として、あるいはもしそうお望みであればその生命の進化における次のステップとして考えられるべきです。データにさらに多くの意味を付加することを可能にします。言い換えると、著者の名前がコンテンツのオブジェクトとして保管されると、マシンも人間も追加情報が理解できるような形で、その著者についてより多くの情報を加えることができるようになります。データを追加することにより、ナビゲーションのルートを改善し、検索をエンドユーザーにより適したものにすることができるようになります。共通のフォーマットとデータのタグ付けにより、以前は孤立してばらばらだった大量のウェブサイトからデータ資源を集めることができるようになります。こうすることにより、コンテンツがいっそう可視化されるのです。また、追加された情報が使えるため、トラフィックをさらに増やし、コンテンツに価値を加えることができるようになります。このようにして、エンドユーザーや購読者、メンバーにとってコンテンツがより使いやすいものになるのです。

若い Google 世代はきっと、インターネット・PC・ブラックベリー以前の生活を覚えていないことでしょう。これらはすべて研究の役に立っています。従って、私たちは適切な信頼できるコンテンツの検索を簡単にし、次世代をも支えられるようにしなければなりません。セマンティック・ウェブは必ず、さまざまな学問分野の研究者が協力して研究し、新たな発見をすることを簡単にします。ティム・バーナーズ＝リー卿の言葉どおり、「セマンティック・ウェブで一番わくわくすることは、これを使ってできると想像できることではなく、できるとは想像の付かないことです」。従って、これは、私たちがまだ知らない多くのことへの第一歩にすぎません。ですから、セマンティック・ウェブがさまざまなデータ資源をいかにして接続・統合するかを視覚化したいと思います。

実生活の例として、DBpedia について考察してみましょう。ウィキペディアはよくご存じかと思いますが、DBpedia は、ウィキペディアのセマンティック・ウェブ版です。DBpedia は、ウィキペディアの 700 万件の記事をよりインテリジェントにリンクしようとする試みです。人や場所、音楽、アルバム、映画についての情報を抽出し、これにタグを付けます。マシンが人や場所などを特定するには、外部のウェブサイトやアプリケーションからデータセットを集めます。このようにして、ウィキペディア上で既に入手可能な情報をさらに充実させるのです。

コンテンツのオブジェクトという観点でその重要性を解明するには、ウィキペディアが 700 万件の記事を持つ一方で、DBpedia は RDF（リソースを記述する枠組み）を用いて 260 万件の「モノ」とそれ以外に 27,400 万件の「事実」を説明しています。これがセマンティック・ウェブのフォーマットです。これが、私たちパブリッシング・テクノロジーが試行している領域です。XML と PDF のフォーマットを取り入れ、これらを RDF に転換して、皆さんのウェブサイトにさらに興味深い展望を打ち出しています。

なぜ DBpedia はこのように機能するのでしょう？コミュニティがユーザー経験を豊かにすることを望んでおり、この追加情報のすべてを使って、さらに高度な自然言語ユーザーの質問にも答えることができるからです。例えば、DBpedia は「人口 1 万人を超えるニュージャージーにあるすべての市を答えなさい」という問題に答えることができます。なぜなら、マシンはニュージャージーが州であることを知っており、その他の人口統計学情報も理解しているからです。もう一つのメリットは、マシンが質問を理解しており、その結果として関連情報を集めることができるからです。

二つ目の例は、著者というコンテクストで聞かれたことがあるかもしれませんが、「FOAF（友達の友達）」プロジェクトです。このプロジェクトは、人とその活動、他の人とのつながりを描くものです。ある人が所属する機関や取り組み中のプロジェクト、誰を知っているのか、誰と共著しているか、どの学会のメンバーかなどの情報が含まれます。言い換えると、そのような情報を蓄えられるようにしているのです。

先に、セマンティック・ウェブのフォーマットの RDF 記録について述べました。このサイトがオープンすると、他のサイトとの統合・接続を始めます。FOAF も Facebook や MySpace、Blogs、LinkedIn、Flickr（写真用）と統合しています。パブリッシング・テクノロジーでは、著者についての追加情報を抽出し、出版者のウェブサイトに入れたり、入手可能な論文情報に入っていない情報を抽出したりしています。しかし、セマンティック・ウェブ技術は、例えば FOAF などの外部サイトから該当情報を引っ張ってくることができます。

三つ目の実生活の例はかなり興味深いものです。日本の東芝の研究センターで今月始まったばかりだからです。これは「クチコミ探索者（Word-Of-Mouth Scouter）」と呼ばれています。「クチコミ探索者」は、ショッピング体験の質の向上を目的としています。現在、書店やネットショップでテスト中です。基本的に何が起こるかと言うと、買い物客が店に行き、バーコードの写真を取り、後でさまざまなブログでレビューを比較するというものです。レビューでは、「良い」「悪い」といった単純な反応が記載され、買い物客がその商品を買うかどうかを決めることができます。

出版者にとってのセマンティック・ウェブ

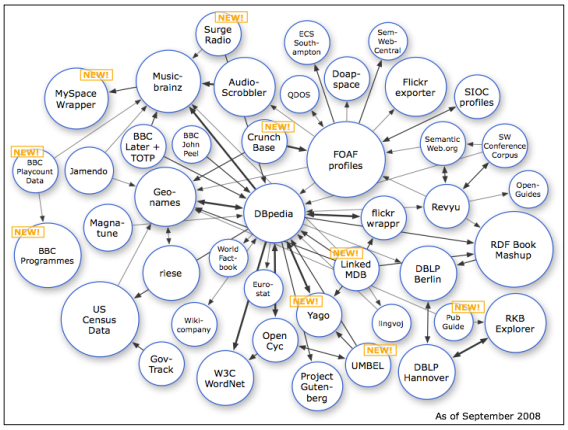
セマンティック・ウェブがどのように出版者に関連しているでしょうか。データをセマンティック・ウェブに準備するために何ができるでしょうか。まずは、ウェブのフローを通して来る新たなコンテンツにできるだけ多くのタグを加えれば加えるほど、良いということになります。これは、分類または類別システムに関係しようとしまいと、テーマに関係する情報については特にそういうことが言えます。PDF や XML で、特に PDF で入手可能なバックファイルのデータについては、数多くの自然言語加工技術とデータマイニング技術を利用することができます。当社のパブリッシング技術は、そのテーマに特有なオープンソースのツールを使い、ジャーナルの論文や本の章などから情報を抽出します。さらにタグが追加されますので、ウェブサイト上で使えるようになります。

データマイニング・ツールの例

コミュニティ内で利用可能なデータマイニング・ツールの幾つかは、欧州バイオインフォマティクス研究所（EBI）によって開発されたものです。例えば、トムソン・ロイターの Calais は、イベントや人、場所を調べるものです。たくさんのツールがありますが、大切なのは、既にあるものの一からのやり直しを試みて、あらゆるテーマの専門家になってしまうことを避けることです。実際には、その分野の研究者がそれを行い、その技術を出版者のデータに適用することができるからです。

ある特定のテーマ領域の中では、利用可能でオープンなデータセットの多くが、セマンティック・ウェブ技術を使っています。例えば、W3C のプロジェクトでは、研究者がこのフォーマットを使うことが奨励されています。現在のところ、既に 170 億のエントリーがあり、それらは 300 万本のリンクでつながれていますが、これは初期段階にすぎません。これによって、可能性が現実になるスピードを想像していただけることでしょう。

9 月時点で、私たちが構築したり、取り組んできたデータセットの幾つかには、前で触れた DBpedia や FOAF、さらにバイオサイエンス領域の Bio2RDF が含まれています。さらに、BBC や US Census のデータなど、さまざまな情報資源があります。エンドユーザーの資源を強化するために数多くの情報が利用可能になっているということです（図 1）。



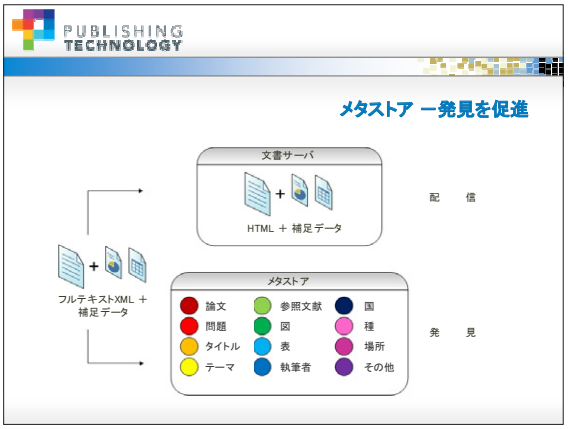
（図 1）

画像クレジット：Chris Bizer
この作品は、クリエイティブ・コモンズ・ライセンスの下でライセンスされています。

パブリッシング技術と セマンティック・ウェブ技術

セマンティック・ウェブ技術を用いたパブリッシング・テクノロジーの仕事の一部を皆さんと共有したいと思います。当社の IngentaConnect および pub2web 製品に隠されたメタデータを管理するためのデータベースを、私たちはこれまで開発してきました。IngentaConnect 製品については、聞いたことがおありになるかもしれません。pub2web は、出版者ブランドのスタンドアローンのウェブ製品です。この二つの製品の両方について、当社は 300 を超える出版者と協力し、かなりの量のコンテンツを管理しています。データベース、すなわち製品の基礎が堅固で、古くならず、非常にすみやかに拡大・縮小できることをお約束することが、私たちにとって重要なことです。データベースにどの技術を使うべきかについて 2 年にわたる研究を重ねた後、RDF とセマンティック・ウェブのフォーマットに決定しました。

メタストア（図 2）はデータベースで、これら二つのサービスの基礎となります。コンテンツを見つける方法を追加することで、エンドユーザーに役立つデータを供給します。製品のクロスプロモーションを提供することによりコンテンツの可視性を高め、入手可能でオープンなデータセットすべてを使って、上で挙げた例のようなデータセットの統合を容易にします。



（図 2）

ここで私たちが取り組んできた試作品を紹介し、こうした技術の幾つかがどのように機能するのかをお見せしようと思います。背景をお示しするために、BioMed Central の XML オープンアクセス・データの幾つかを、極めて基礎的な pub2web サイトに組み込みました。

私たちがこれまでにしたことは、欧州バイオインフォマティクス研究所（EBI）の「What Is It」データマイニング技術を XML データに適用し、種名と遺伝子名を取り出すことです。こうして、そのジャーナルに記載されている種のリストと各種に関係がある論文の数を入手することができます。また、これらの名前がすべて記載された種のマップも手に入れることになります。それだけでなく、さらに重要な名前またはその種名に関連する入手可能なコンテンツがもっと多くあるのはどこか、なども記載されています。従って、テキストから種の情報の抽出を済ませているので、種名をクリックすれば、これを個別のコンテンツのオブジェクトとして保存し、さらに情報を加えることができます。このようにして、出版者から得たデータから、このウェブサイト内で、このデータセット内で、ほかにどれだけの数の論文が入手可能かを解明することができ、更なる情報を集めるために Bio2RDF データセットを使うことができるのです。例えば、ランクや一般的な名称、分類 ID、また画像もあり得ます。こうした追加情報をすべて、外部のデータセットから得た結果、出版者のコンテンツはより豊かに、エンドユーザーがより使いやすいものになるのです。

また、私たちは遺伝子名を抽出し、これを生物過程や細胞成分、分子成分に分けます。その結果、さまざまな方法でコンテンツのナビゲートができることになります。研究者またはエンドユーザーが何か非常に具体的なものを探している場合、検索に関連するすべての情報へのアクセスが非常に簡単になります。同様に、すべての遺伝子名と入手可能な遺伝子名すべてのリストの付いた、遺伝子名の検索マップも得ることができるのです。遺伝子名の一つを選ぶと、その遺伝子名のホームページにナビゲートされます。このシナリオでは、私たちのデータセットからデータを抽出することができ、他の利用可能なコンテンツの中で、どれがこの遺伝子名について述べているかを言うことができます。遺伝子名の説明やその ID、それに関連する同義語を見つけるために、私たちは Bio2RDF データセットを使いました。いったんこれが検索エンジンにキーインされると、結果は、研究者にとって少し関連性が高くなります。これはバイオサイエンス内での例ですが、他の主題領域でもこれがどのように機能するかを想像していただければと思います。

例えば同じ遺伝子名に言及している他の論文へのリンクを追ってみました。もし著者名へのリンクをたどれば、XML 内に既に持っているデータを使って、その著者の E メールアドレスや所属、データセット内にある彼らが書いた他の論文や彼らが共著者である論文も含む、著者ホームページを作成することができます。しかし、ページの潜在能力を高めようと思えば、著者の写真を加えたり、あるいは例えば著者の Flickr に接続したりすることもできます。また、FOAF に接続して、彼らが現在取り組んでいるプロジェクトの情報や、参加あるいは発表しようとしているコンファレンスの情報を求めたりすることができます。このようにさらに多くの情報を加えることが可能です。

さらに一歩進めて、特定の機関が、あるいはその機関に属する著者がこれまでに提出した論文のデータセットから引っ張ってくることで、その機関のホームページも作成することができます。さらに、機関のロゴだけでなく、DBpedia が提供できる事実と数字も加えることができます。私たちが現在取り組んでいる pub2web プロジェクトの幾つかを通して、多くの可能性を発掘中です。

今現在私たちが取り組んでいるプロジェクトの幾つか

についてイメージを持っていただくために例を挙げると（図 3）、私たちは現在、当社の pub2web プラットフォーム上にフランスの統計出版者 OECD の新たなウェブサイトを構築中です。ジャーナルや書籍、統計情報、参考資料などの幅広いコンテンツを持ち、ポッドキャストの試行も始めており、これはとても興味深いプロジェクトになっています。

（図 3）

私たちは世界銀行の新しいウェブサイトにも最近着手しました。このウェブサイトには、主には書籍ですが、ジャーナルや報告書、モノグラムなど、世界銀行の出版物のすべてが含まれます。また、国際通貨基金の出版物すべてをカバーするウェブサイトにも最近乗り出しています。現在のところ、このサイトは海外支店のために作られており、プリントアウトを支店に送付する必要がなくなりました。今後数年のうちに機関市場にも参入したいと希望していることから、これは機関にも販売されることになるでしょう。

一方で、アメリカの学会 “American College of Allergy, Asthma & Immunology（米国アレルギー・喘息・免疫学会）” と共同で、学会メンバーがアクセスできる、ジャーナルに特化した pub2web サイトを構築中です。彼らが特に興味を示している機能の幾つかは、継続的な医学教育と関連があるものです。メンバーのためのオンラインクイズやニュース、イベント領域を提供しています。また、近いうちにイギリスの学会 “Institute of Civil Engineers（土木工学学会）” の出版拠点である “Thomas Telford” のための pub2web プロジェクトへの取り組みも始めます。このプロジェクトは、彼らのジャーナルと書籍、マニュアルを一つにし、機関や学会のメンバー読者を引き付け、印刷物を扱う書店を電子製品に統合するものです。

これらの例すべてからお分かりいただけるとおり、幅広い主題領域とさまざまな種類のコンテンツに、先に述べたセマンティック・ウェブ技術を活用することができます。

技術は進化と変化を続けています。私たちパブリッシング・テクノロジーは、コンテンツができる限り実際的になるような試行と革新を常に目指しています。また、さまざまな層に満足いただくことも目指しています。2 週間前、ロンドン・オンラインで、私たちは、「IngentaConnect Mobile（インジェンタコネクト・モバイル）」（図 4）というモバイル機器を通したコンテンツ配信の試行を開始しました。

（図 4）

これは若い読者をターゲットにしているもので、大学の学部生を対象に、彼らの使用方法について調査しています。ユーザーグループからのフィードバックと、図書館司書や出版者からのフィードバックが、このプロジェクトの次のステップを決める助けとなるでしょう。

この試行では、伝統的な印刷形式から離れるだけでなく、ウェブサイトのみにとらわれないことが重要です。ユーザーは、インターネットとは全く異なるやり方で、モバイル機器からコンテンツにアクセスすると思われるからです。私たちが直面する課題は、モバイル機器でユーザーはどのようにアクセスしたいと思うか、また、そのビジネスモデルはどうあるべきか、というものです。この問いには多くの異なる答えがあることから、これはまさに市場が何を望んでいるかを見出す試行例です。

モバイルのプラットフォームを持つもう一つのメリットは、新たな市場への露出です。実に顕著な数字が幾つかあります。現在、世界中で 26 億人が少なくとも 1 台のモバイル機器を持っています。例えばインドでは毎月、1,000 万人の新しいモバイル機器利用者が生まれている

のです。そして、広く普及したブロードバンドや PC 等へのアクセスのためのインフラを持たないような国々では、ウェブ・モバイルの販売が劇的に増加しています。こうした市場では、今がコンテンツをさらに可視化する時です。国によっては、モバイル機器が優位に立ち、PC の販売が減少しているという傾向についてよく検討しなければなりません。従って、これは今後数年に私たち皆が直面するであろう問題なのです。

結論

最後に、今日伝えたいメッセージは、私たちは、データのウェブ（網）環境へ、複数のマシンで加工が可能な環境へ移行する必要があるということです。これは、ワールドワイドウェブの連続的な進化です。

セマンティック・ウェブ用にデータを準備する方法はたくさんあります。PDF や XML を捨てる必要はありません。

出版者側にも数多くのメリットがあります。これを、今日きちんと立証できていればいいのですが。「私の意見では、今が試行の時であり、試行によってのみ、私たちはウェブ製品がどのようになるかについての次のステップを準備できる」と申し上げて、本日の結びとしたいと思います。