プロジェクト名：　異分野研究資源共有・協働基盤の構築
（サイエンス 3.0 基盤構築）

プロジェクトディレクター：　　新井 紀子　教授（国立情報学研究所）
サブプロジェクトディレクター：　武田 英明、相澤 彰子　教授（国立情報学研究所）

〔1〕 研究計画・研究内容について

(1) 目的・目標

自然科学から人文科学にわたる異分野の「知」と「人」の共有・連携を行い、情報や研究人材の効果的な活用や研究協力・共同研究の促進を行う学術知共有・学術連携促進基盤を構築し、実用に供する。その手段として、まず、インターネット上で様々なところに散在する学術情報および研究支援サービスを結合して利用可能とするプラットフォームを構築する。このように収集された学術データを研究対象として新しい検索技術・機械学習・データマイニング・ユーザインタフェイス技術・可視化技術等の研究開発を通じて、研究者あるいは研究分野・研究プロジェクトごとにパーソナライズされた学術情報・学術サービスの提供を目指す。

具体的には、サブテーマ「研究資源に関する情報推薦基盤の構築」においては、機械学習・データマイニング・オントロジー関する研究を通じて、情報推薦に関して世界をリードする独自技術を開発する。サブテーマ「学術リソースのためのオープン・ソーシャル・セマンティック Web 基盤の構築」において、セマンティックウェブ技術およびデータベース連携の研究開発を通じて、研究者向け次世代ウェブサービスの構造に関する技術開発を行い、散在する学術研究資料が有効活用するための基盤を整える。サブテーマ「多様な知的情報源を結合・融合・再構成する連想情報処理基盤の構築」において、論文情報や書誌情報といった定型的なデータ以外にも、発表資料、コースウェア、研究データなどの異種データをリンケージした上で高速な連想検索を行うための技術の確立を目指す。以上のサブテーマによる、研究開発をサブテーマ「融合研究を加速するための情報共有クラウドサービスの確立」で統合し、世界をリードする次世代研究者サービスを構築し、日本の学術知共有・学術連携を促進することを目指す。

(2) 必要性・重要性（緊急性）

インターネットを通じて様々な学術情報・学術サービスが公開・提供されるようになったが、単に Web に公開しただけは相互運用性がなく、情報を十分に活用することはできない。特に、近年学術分野においても情報爆発が起こっており、これに対応するため、学術情報に関する各種電子アーカイブが整備されつつある。また、多種多様な分野における研究人材データや研究用のデータベースも電子化されてきた。世界的な研究開発の加速・競争の激化の中、整備されつつある研究データ・論文アーカイブ・人材データベース・研究用ミドルウェア等をいかに有機的に連携し、柔軟かつ機動的に共同研究を進めるかということが、日本が科学立国としての地位を維持する上で、鍵となる。しかしながら、現状においては、これらの学術情報・学術サービスを有機的に結合する手段は未成熟であり、人材と研究に関する連携力が充分に発揮されているとはいえない。また、情報技術から遠い学問分野においては、このような潮流の認識が諸外国に比べて進んでおらず、取り残される危険性がある。この問題を解決する手段として、すべての学問分野の研究者にとって使いやすくまた柔軟性のある学術知共有・学術連携促進基盤を構築する必要性がある。

(3) 期待される成果等（学問的効果、社会的効果、改善効果等）

既存の大規模データベースを有機的に結合するための「ハブ」となるシステムを研究者に提供するこ

とにより、学術知共有・学術連携が促進される。特に、異分野での連携促進が期待できる。また、本システムを実運用システムとして全国の研究者に提供することにより、日本最大級の「生きた」学術データベースが構築され自律的に増殖していくことになる。このことにより、主として3つの社会的波及効果がある。第一に、本システムに蓄積されたデータを研究対象として新しい検索技術・データマイニング・情報推薦・ユーザインタフェイス技術・可視化技術等の開発が進むことが期待できる。第二に、本システムをサービスとして利用する研究者は、多様かつ膨大な学術データベースから、自分の研究分野や研究関心にあわせた最適な研究情報が「推薦」され、編集された上でタイムリーに届けられる。また、研究を支援するような各種サービスが、クラウド基盤を通じて提供される。これは、競争が激化している各研究分野において、日本の研究者が国際的優位性を勝ち取る上で、たいへん重要である。第三に、本システムに蓄積された研究情報が国民に随時公開されることにより、多様かつ信頼がおける科学コミュニケーションの場が副次的に実現されることである。

(4) 独創性・新規性等

本プロジェクトでは、多様な異種学術データを大規模に収集した上で、情報および統計の技術を駆使し、各研究者に対して、パーソナライズされた情報およびサービスを提供するという極めて先進的な取り組みを行う。本分野は、1−5でも説明するとおり、世界中の研究機関・研究者向け商用サービスが重要視し、取り組みを本格化させているところでもある。その中で、本プロジェクトは以下の点において、優位性および独創性がある。

まず、国立情報学研究所は国内有数な学術データベースを有しており、また、情報・システム研究機構の融合研究センターはライフサイエンス統合データベースを有している。大学共同利用機関法人として、各種の機関リポジトリやデータベースとの連携関係も深い。これらデータベースと結合することで、他機関では到底実現不可能な大規模な情報流通基盤が実現可能となる。本プロジェクトが具体的に実現される基盤であるNetCommonsは2007年には国際学会IASTED主催第3会国際ソフトウェア競技会で最優秀賞に選ばれたほか、2009年にはIPAより日本OSS奨励賞を受賞するなど国際的評価も高い。その上で、世界最速の連想計算エンジンGETAによるコンテンツ・コンパイル技術を用い、蓄積された情報源の特徴を計算機構として抽出する。さらに情報源同士の相互作用に活用し、研究者の特性をデータマイニング技術によって抽出した上で、パーソナライズされた情報推薦を行うことは、非常に先進的・独創的な取り組みである。また、単に先進的・独創的な研究であるだけでなく、研究開発成果が直ちに、産学官を超えたすべての日本人研究者に提供される。その意味でも、社会貢献の度合い、費用対効果も極めて高い。

(5) これまでの取り組み内容の概要及び実績

本研究に先立って、第一期新領域融合研究「分野横断型融合研究のための情報空間・情報基盤の構築」においては、融合研究を加速するためのバーチャルラボシステムNetCommonsを構築し、オープンソースソフトウェアとして公開している。また、異種情報の結合・分類手法に関しての研究を進め、世界最速の連想計算エンジンGETAによるコンテンツ・コンパイル技術を確立して、異なる情報源同士の相互作用を情報探索に利用する想・IMAGINEシステムを開発した。さらに、大規模リンケージ情報の研究では、国立情報学研究所で公開中の「科学研究費補助金データベース」を情報源として、約13万人の日本人研究者について統一的な研究者ID番号の情報を提供する「研究者情報サーバ」プロトタイプ版システムを拡張し、他のデータベースとの統合のための機能整備を行った。これらの成果を概念レベルだけでなく、具体的に融合させ、平成20年度には、「状況に埋め込まれた人間の相貌をデジタルに表現する技術の研究」において、NetCommonsを基盤として、コンテクスト（状況）の中で、さまざまな

相貌をみせる人間の活動にフィットするポストウェブの技術の開発を目指し、今回提案するサイエンス3.0基盤のプロトタイプとなるResearchmapα版の開発を行った。具体的には、多様な学術情報データベースから、研究者IDをキーとして論文情報・研究者経歴等の学術情報を複数のデータベースから自動取得する方法を開発し、研究者のCVデータとして編集・公開する機能を実装した上で（担当：相澤、大向、新井）、CVデータを軸として、興味関心の近い研究者を分野横断的に検索する技術を開発し（担当：新井・高野・丸川・舛川）、研究者の研究コミュニティの形成および運営を支援するための基盤サービスの提供を試行し、既に1300人を超える研究者が実際に試用している。今後も利用者が増加することが見込まれ、より多くの研究データが蓄積することが確実となっており、次期新領域融合研究を開始する準備が整っている。

(6) 国内外における関連分野の学術研究の動向

海外の学術機関の動向については､フィンランドが健康バイオ分野でセマンティックWeb技術を利用した広範なデータベース連携を実現している。しかし主たるターゲットは公共的機関がもつデータであり、研究データなどはあまり対象となっていない。またEUではEuropeanaプロジェクトが各国の博物館データの統合を進めているが、統合の程度はあまり深くない。

商用サービスを含めた動向としては、研究者が独自のIDを取得できるResearcher IDというサービスをThomson社が開始し、また、研究者の情報発信支援をAcadema.eduが提供するなど、研究者向けに学術情報サービスを提供する試みがまさに始まったばかりであり、世界的関心が非常に高い。しかし、これらのサービスは論文情報販売を目的とした情報収集および顧客囲い込みのためのサービスであり、学術情報を横断的に活用しながら共同研究を推進する基盤を目指しているわけではない。

〔2〕 研究計画

(1) 全体計画

学術情報は、かつてはきわめて狭く固定的な方法で流通していた。流通の範囲は自らの分野の専門家に限定され、方法も学術雑誌における論文といった出版に限られていた。しかし、本来、学術情報はもっと広く柔軟に流通すべきである。学術成果は単に結果を論文として発表するのではなく、利用したデータや結果に関するデータといった情報、研究過程といったものも公開・共有されることが、開かれた科学技術の発展上は望ましい。また学際的な研究も盛んになっている現在、自分の分野だけで利用可能な情報流通は適しているとはいえない。一方で、科学技術における発見や発明が、富の源泉であることは、科学技術の4千年を超える歴史の中で自明のことであり、研究過程を公開することは、研究者にとっても各国の科学技術戦略の上でも、慎重である必要がある。

ここに、研究者最新の学術研究データに1秒でも早くアクセスした上で、自らの研究成果および過程は、適切な共同研究者との間で安全に共有し、それを素早く商用化したり、研究成果として公知としたり、そのサイクルの中で、より大きな競争的資金やより良い共同研究者を獲得する、というニーズが、否が応でも高まる素地があるといえよう。学術研究データに関する多様なデータがデジタル化され、アーカイブされるようになった今、このことは一見、直ちに実現され得るかのように見える。しかしながら、そこにはいくつかの理論的・技術的な困難が存在する。

第一は、多様な学術研究データがウェブ空間上に爆発的に増加した結果、それらのデータにアクセスすることは概念的には可能であるが、現実には不可能に近い。そこで、研究者の知的生産活動にとって効果的で確実な検索技術が不可欠になる。ところが、研究者の在り方や興味関心分野は多種多様であり、必要とするデータも多種多様である。よって、ウェブ上に拡散する学術研究データが多様になればなるほど、個々の研究者に特化した形で、あたかも執事のように情報をリトリーブして的確に提供するため

のプッシュ型の情報検索・情報推薦の技術が望まれる。ここに第二の困難がある。研究者の興味関心に従って、ウェブ上の学術研究データの意味を発見・分類し、統計処理した上で、情報推薦することは、画像処理であればセマンティックギャップ、人工知能であればフレーム問題に相当する、セマンティックとシンタクスをつなぐ非常に困難な問題だからである。そこで、我々は、データマイニングとオントロジーを用いた手法と、ソーシャルメディア的手法を用いてユーザ自身からフィードバックを得る手法と、外部の信頼おけるデータとそれに付与された情報を活用した連想検索の手法を統合することで、この課題の克服を目指す。

<table>
<tr><th rowspan="2">テ　ー　マ</th><th>H22 年度</th><th>H23 年度</th><th>H24 年度</th><th>H25 年度</th><th>H26 年度</th><th>H27 年度</th></tr>
<tr><th>(予備研究)</th><th></th><th></th><th>中間評価</th><th></th><th>事業化</th></tr>
<tr><td>全　　体</td><td colspan="4">実システムへの適用・Web 空間との連携・実証研究・改良</td><td colspan="2">事業化</td></tr>
<tr><td>サブテーマ１</td><td rowspan="3">準備調査研究
プロトシステムの開発</td><td colspan="2">「情報推薦」技術の研究開発</td><td colspan="2">「情報推薦」技術の改良と深化</td><td rowspan="3">他のシステムへの応用</td></tr>
<tr><td>サブテーマ２</td><td colspan="2">セマンティックウェブ技術の研究開発</td><td colspan="2">セマンティックウェブ技術の改良と深化</td></tr>
<tr><td>サブテーマ３</td><td colspan="2">多種データ間の連想検索技術の研究開発</td><td colspan="2">サブテーマ３はサブテーマ４に統合</td></tr>
<tr><td>サブテーマ４</td><td>連携準備</td><td colspan="2">国内学術分野における連携強化</td><td colspan="2">産業界・海外との連携強化</td><td>国際展開</td></tr>
</table>

(2) 各年度の計画

平成 25 年度

サブテーマ１では、XML タグによって表された文書の構造情報を保存しつつ、テキスト部分に構文解析などの言語解析技術を適用して、統合的な構造情報を得るための頑強な解析基盤の枠組みを提案する。平成 25 年度は特に出版社や学会によって決まった形式が与えられている論文の XML 文書を対象に、参考文献や図表の参照、数式、箇条書きなどを含めたままの形で構文解析を適用可能な手法を開発し、実際に大量の論文の解析を行うことで有用性を評価する。

サブテーマ２では、本年度からデータ中心型研究基盤の展開を行う。基盤システムとしての機能強化を図る共に応用的システムを作り、ケーススタディを進める。

まず、異なるデータサイトからくるインスタンス情報（個物に関する情報）は同じインスタンスを指していることがある。このインスタンスマッピングを効率的に行うプログラムを開発する。プロジェクトメンバーが開発したアルゴリズムを発展させ、実問題で適用できるようにし、実際にインスタンスマッピングを行い、データ統合を自動化する。GBIF データといった研究データにおける実データを用いる。次に、構築された統合的データベースを可視化するためにいくつかのアプリケーションを作成する。たとえば、Linked Data アプローチで多様なデータが結ぶつくことを示すため、地図と地名から様々な情報にアクセスできるアプリケーションを作成する。地理・地名情報は多くの分野に共通するので、このアプリケーションを通じて様々な分野の情報、データが横断的に利用できる。地理情報であるので、PC 版だけでなく、携帯できるモバイル版も開発する。以上で開発してきたさまざまなプログラム、システムおよびデータベースを統合的に利用できる環境を構築する。このプラットフォームを用いて、アプリケーションがデータを取得したり投入したりできるようにする。環境プロジェクトや GIS プロジェクトのシステム・データを統合する。

サブテーマ３（連想情報処理基盤の研究）については、中間評価に基づく PD の判断により、 独立のサブテーマとしては研究を中止して、必要に応じてサブテーマ１やサブテーマ４を推進する中で要素技

術の活用を検討する。

サブテーマ４では、これまで蓄積したデータを基に、他のサブテーマ３で研究開発した要素技術を用いて異種データリンケージによる研究情報可視化のプロトタイプの実装を行った。また、サブテーマ１の要素技術を実装するために、研究論文・講演資料等を Researchmap 上で研究者が蓄積するための WEKO ベースの大規模 OpenDepo その上の全文検索を検討・構築した。

昨今、公的機関のウェブサービスに対する悪意ある攻撃が苛烈化している。IPA からはウェブサービス一般に関する注意喚起（http://www.ipa.go.jp/security/ciadr/vul/20140619-oldcms.html）が行われた。このような状況を鑑み、Researchmap およびその基盤となる NetCommons は、いち早くセキュリティを最高基準に揃えるための改修作業を開始した。

平成 26 年度

サブテーマ１では、前年度の検討結果に基づき、XML 文書の言語解析手法を改善するとともに、提案手法を定量的に評価するための評価基盤を構築する。具体的には、文書の構造情報を取り込むことで、従来は扱えなかった箇条書きなどの意味解析を可能にしたり、参照記号や数式を含む文を正しく構文解析したりする効果を実証する。さらに、ウェブページや発表資料など、文書が不定形で構造情報が XML タグで明示されない場合について、頑強性の高い言語解析を行うための手段を探る。

サブテーマ２では、本年度からデータ中心型研究基盤の展開を行う。基盤システムとしての機能強化を図る共に応用的システムを作り、ケーススタディを進める。

基盤開発としては、データの表現力の強化と分野を超えたデータ利用の枠組みを構築する。データの表現力の強化としては、時間的に変遷するデータの記述法の確立を図る。データは作られた時々において正しくても経時的に変化することがある。これを経時的な変化を記述するオントロジーを構築して、経時的にもデータが連続することが可能にする。また、分野を超えたデータ利用の枠組みとしては、DBpedia Japanese を中心とした Linked Data Cloud の構築とその発展的利用を図る。Linked Data Cloud とは LOD のデータセット同士の関係性を保持するデータであり、これを構築することで、新たなデータ利用ニーズにおいても適切なデータセットを見つけやすくなる。

次に、構築された統合的データベースを利用するアプリケーションを作成し、データ利用の可能性を高める。例えば、LOD にあるデータを利用して機械学習を行うことで新たなる発見を行う実験を行う。また以上で開発してきたさまざまなプログラム、システムおよびデータベースを統合的に利用できる環境を構築する。このプラットフォームを用いて、アプリケーションがデータを取得したり投入したりできるようにする。環境プロジェクトや GIS プロジェクトのシステム・データを統合する。

サブテーマ４では、平成 25 年度に構築した OpenDepo を研究者に対してリリースするとともに検索の高速化を図る。また、主要研究大学との Shibboleth 連携、API 連携を深める。これによって集約されたデータを基に、サブテーマ１およびサブテーマ２の実証実験を本格化させる。平成 25 年度は、そのための検討を行いとプロトタイプ（NetCommons3.0α）を開発する。また、Researchmap を研究情報だけでなく他の情報（教育情報等）の循環プラットフォームに応用すべく、学校総覧 edumap を関係機関と連携して研究に着手する。そこにおいて、サブテーマ２の研究成果に基づき API 等の設計を行う。

平成 27 年度

サブテーマ１では、構文解析だけではなく、談話構造解析や照応解析など、より高度な意味解析手法の適用において、XML 形式で埋め込まれた文書構造、外部のリソースへのリンクを含むデータ構造、意味アノテーションなどをシームレスに利用する手法の実現を目指す。これらの基盤技術を用いた意味検索を実現し、研究・教育用コンテンツの情報探索支援における有用性を実証的に評価する。

サブテーマ２では、データに基づく研究基盤として発展させる。まず、論文に含まれるデータを抜き出し、データとして利用できるような発展的ソフトウェアの開発を行う。またデータの由来などの情報も同時に抽出する。この仕組みをつくることで論文とデータが同時に利用可能になり、データ中心型研究の新たな研究成果表現が発展することが期待される。さらに、学術分野ごとに存在する概念体系、専門用語体系を抽出してマッピングを行う。この学術オントロジーと論文、論文抽出データを同時に使うことで分野を超えた研究の理解とデータの利用が可能になる。

サブテーマ４では、平成 25 年度に検討した Researchmap および NetCommons に関するセキュリティ向上のための改修計画に基づき、改修を実施し、NetCommons3.0 をリリースする。また、サブテーマ２の研究成果を、統合プラットフォームのアプリケーションとして OpenDepo 上でデータ操作ができる環境を構築する。リポジトリに投入された情報からデータを取得し統合プラットフォームへデータを送ったり、データを入手できるようにする。リポジトリ自体がデータ中心型研究の環境として機能するようにする。また、これまで構築した、統合プラットフォーム、Researchmap 統合、リポジトリ統合をシームレスにつなぎ、クラウド型データ中心型サービスの構築を構築する。本成果を edumap 他の情報循環基盤にも応用し、サービスとして公開する。

〔3〕 研究推進・実施体制

サブテーマ１：研究資源に関する情報推薦基盤の構築

・研究代表者

〔国立情報学研究所〕　相澤彰子

・共同研究者

〔国立情報学研究所〕　内山清子、高須淳宏、宮尾祐介
〔統計数理研究所〕　持橋大地
〔広島市立大学〕　難波英嗣
〔情報・システム研究機構〕　原　忠義

サブテーマ２：学術リソースのためのオープン・ソーシャル・セマンティック Web 基盤の構築

・研究代表者

〔国立情報学研究所〕　武田英明

・共同研究者

〔国立情報学研究所〕　大向一輝、松村冬子
〔国立遺伝学研究所〕　菅原英明
〔情報・システム研究機構〕　加藤文彦、小出誠二、亀田尭宙
〔東京大学〕　伊藤元己
〔国立科学博物館〕　神保宇嗣
〔人間文化研究機構〕　山田太造
〔東京芸術大学〕　嘉村哲郎
〔慶應義塾大学〕　深見嘉明
〔ATR-Promotions〕　高橋　徹、上田　洋

サブテーマ３：融合研究を加速するための情報共有クラウドサービスの確立

・研究代表者

〔国立情報学研究所〕　新井紀子

・共同研究者

〔国立情報学研究所〕	羽田昭裕、山地一禎
〔国立極地研究所〕	岡田雅樹、野木義史、小林悟志
〔情報・システム研究機構〕	舛川竜治、阿辺川武
〔総合研究大学院大学〕	大田竜也
〔藤田保健衛生大学〕	宮川　剛
〔電気通信大学〕	Neil Rubens

〔4〕 研究の進捗状況

サブテーマ１

サブテーマ１では、多様な観点に基づき論文を推薦する情報推薦手法の研究およびシステム構築を進めている。特に、「論文を単位とする検索」から「論文に書かれている情報の探索」の支援へと研究を展開して、リサーチコモンズ基盤技術の１つとして確立することを目標としている。このために必要な要素技術として、平成 25 年度では特に、(1)論文の XML 文書を対象に、タグによって表された文書の構造情報を保存しつつ、テキスト部分に構文解析などの言語解析技術が適用可能となる、頑強な枠組を考案し、大量の文書に対する有効性の検証を行った。また、(2)可読性の高い情報提示の技術と意味解析に基づく情報探索支援技術の両面から研究者の情報アクセスを支援する論文閲覧のデモシステム SideNoter を構築し、学会にて公開・実演を行った。更に、(3)分野や言語、閲覧者の熟練度などの分け隔てなく論文の推薦を可能にするための文書解析の基盤技術として、(a)「論文内の専門用語検出技術の精度を高めることで、論文内で共参照関係にある用語の抽出精度を向上させる手法」および (b)「日本語テキスト中の専門用語を別言語の辞書項目にリンクする際に、対象言語での候補選択時に起こる曖昧性解消を、テキスト中の語句の一貫性を利用して実現する手法」の開発を行った。

(1) 文書／意味構造を統合した文書解析基盤の構築

論文等文書の内容に踏み込んだ推薦を行うためには、テキスト解析技術を大量の文書群に適用し、深い意味や関係情報を抽出することが必要不可欠である。しかし、従来のテキスト解析技術がその入力に文の連続列を前提とすることが専らであるのに対し、文書は多少なりとも文の構造的配置（構造化）がなされているため、解析のためには、「構造化された文書を入力可能な文の連続列に変換し、解析後の結果を元文書の然るべき箇所に対応づける」という工程が必要となる。更に、文書の表示方法や管理方針によって構造化の形式も多様なため、文書様式毎にこの工程を構築せねばならない。これは、広範な解析、それに基づく推薦情報の取得にとって大きな障壁となる。

我々は、文書が基本的には「表示される文字・実体」と「それを文書上に配置するための情報」の連続列として与えられることに注目し、これを XML フォーマット文書における「文字または XML タグにより導入された文字以外の実体」と「XML タグによる領域指定」によって吸収しうると仮定し、文書解析に関わる特徴により XML のタグが以下の４種に分類可能であることを見出した。

【独　立】 周囲のテキストとは統語的に独立している領域を表すので、別個に解析を行う
【装　飾】 表示上の効果を領域に与えるだけなので、タグのみ除去し解析後にタグを復帰する
【実　体】 自然言語とは異なる原理で構成された実体を文の構成要素として取り込んでいるので、タグを含む領域全体を代替の文字列に置き換えて解析した後、復元する
【非表示】 表示されない付加情報の記述領域を表すので、タグ領域を除去し解析後に復元する

そして、この分類に基づき文書解析を体系的に実現する一般枠組（図 1）を考案、その雛形となるシステムを構築した。

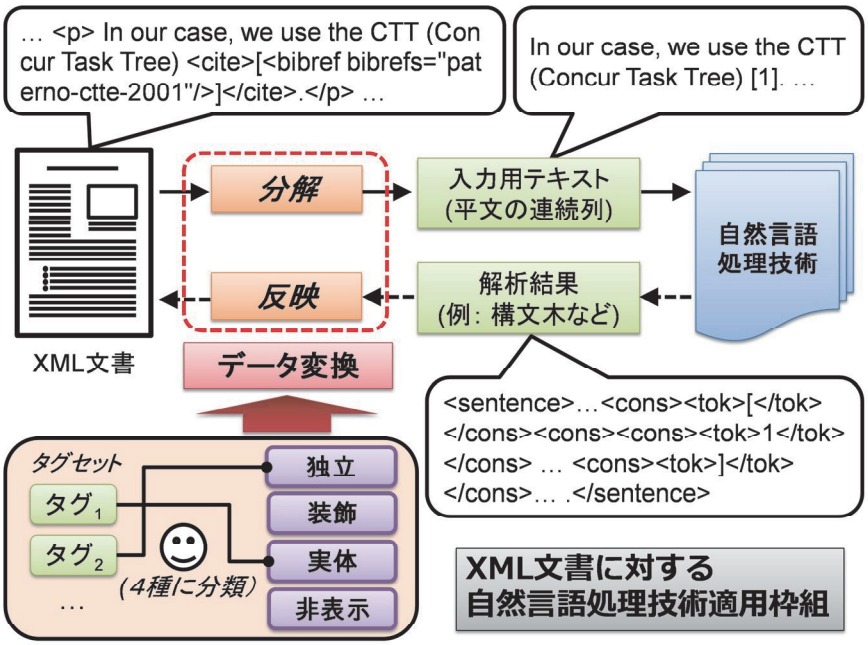

図1：XML文書に対する自然言語処理技術適用枠組

解析対象の XML 文書（群）で用いられるタグを上記 4 種に分類することで、あとは自動的に文書が入力に適した文の連続列へと変換され、解析結果が元の文書に反映される。また、タグ分類作業に関しては、必要なタグ領域のみを適宜検索・展開し分類を促すことにより、最小限に抑えている。このシステムを複数の XML 様式を持つ大量の文書群に対して適用した結果、各文書様式で使用されているタグ名のうち 1/5 程度（ある文書様式で 229 個中 43 個）の分類作業を行うことで、その様式で記述された大量の文書を自動解析できることが示された。また、このシステムを GUI の形で利用しやすく実装し（図 2）、ツールとして公開する方向で調整を進めている。

一方で、今後改良すべき点も観察された。提案枠組では、「実体」タグを単純に代替文字列に置き換えているため、テキスト解析時に以下のような問題が生じる。

- 「実体」タグは、通常の単語にはないような文構成の役割を果たしうるが、代替文字列そのものは全く意味情報を持たないため、従来のテキスト解析技術では正しく扱えないことがある。また、「実体」周辺は句点が抜けやすく、文区切りが曖昧になるという問題もある。
- 「実体」内部にも、部品としてテキストが現れる場合がある（例：箇条書きの各項目など）が、提案枠組では「実体」全体が代替文字列として置き換えられるため、内部の解析は行えない。また、内部のテキストは、「実体」周辺のテキストと併せて解析して初めて意味を成す場合もあり、そのままでは正しく解析できない。

このような構造の解析に関わる問題に関しては、構造の外側と内側、両側で本枠組を拡張する必要がある。今後の研究において、これらの問題を一般的に取り扱うための技術改良を進めて行きたいと考えている。

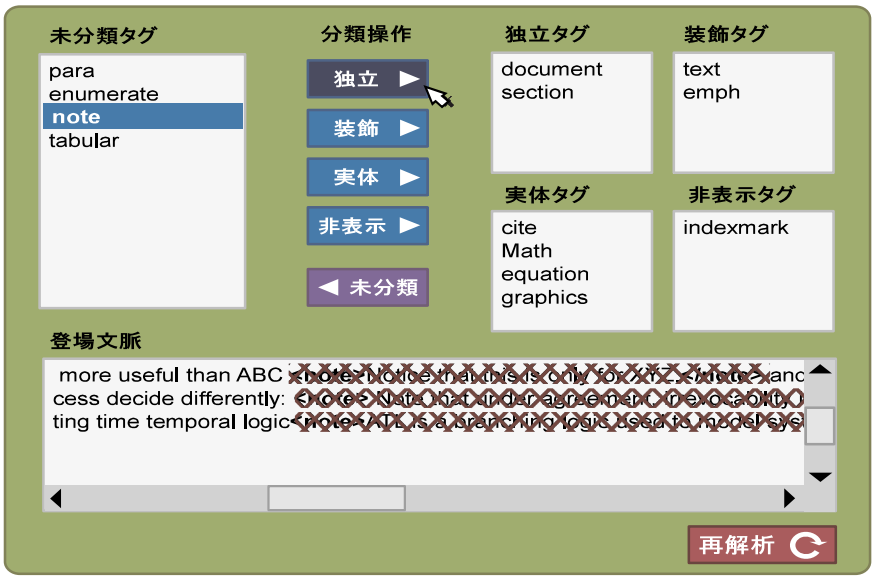

図2：実装ツールにおけるタグ分類のGUI

(2) 脚注表示機能を備えた論文閲覧システム Sidenoter を用いた情報推薦のデモ実施

PDF 文書の閲覧時に、補足的な情報をページレイアウト上にそのまま表示するシステム Sidenoter に対し、論文 PDF 本文の解析から得られる情報をページレイアウト上にそのまま表示するデモシステムの実演を、2014 年 3 月に開催された自然言語処理学会の年次大会で行った。言語処理学会では、20 周年記念事業の一貫として 2013 年 7 月に過去の年次大会の予稿集を学会 Web ページ上で公開しており、本システムは、学術文献の閲覧システムのケーススタディとして、この言語処理学会の年次大会の予稿集を閲覧するものとして作成された。特に年次大会のような学会の会議に参加し、聴講中の予稿集の閲覧を支援するシステムになることをめざしている。

本システムの基本構造は、PDF で配布される論文を画像に変換し、Web ブラウザ上で閲覧する仕組みからなる。図 3 は、そのスクリーンショットである。

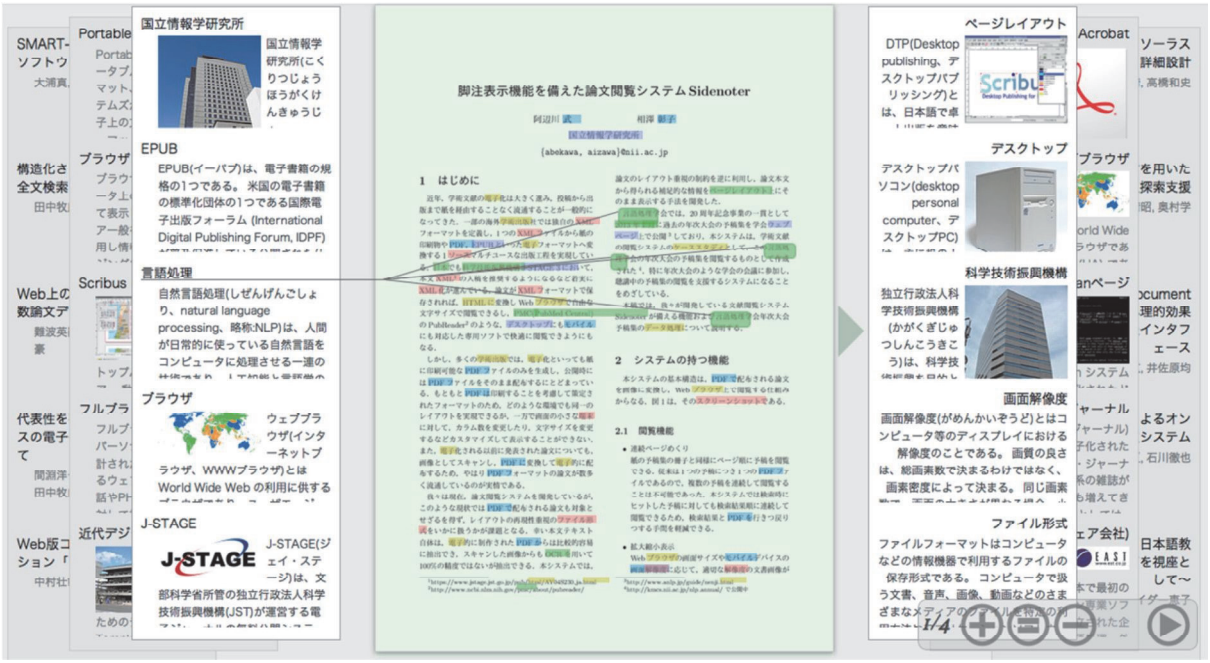

図3：Sidenoter によるデモシステムのスクリーンショット

本システムの実装には以下のような特徴があり、各々がユーザの閲覧支援を実現している。

【閲覧機能】

・連続ページめくり： 紙の予稿集の冊子と同様にページ順に予稿を閲覧できる。従来は 1 つの予稿につき 1 つの PDF ファイルであるので、複数の予稿を連続して閲覧することは不可能であった。本システムでは検索時にヒットした予稿に対しても検索結果順に連続して閲覧できるため、検索結果と PDF を行きつ戻りつする手間を軽減できる。

・拡大縮小表示：Web ブラウザの画面サイズやモバイルデバイスの画面解像度に応じて、適切な解像度の文書画像が表示できる。また画像の拡大縮小表示の他、2〜4 ページまでの割付表示が可能であり、4 ページ表示で次々と文献を移動させていけば、紙の予稿集においてページをパラパラめくる感覚に近い。

・ブックマーク機能：論文を次々と閲覧していく際、後で読みたい論文をブックマークして整理しておくことができる。

・背景色変更：紙に印刷する際は白背景に黒い文字であることが一般的であるが、バックライトが視覚に入るディスプレイで、長時間、白と黒の組み合わせといったコントラストの高い画面を見続けると目が疲れてしまう。そのため背景色を薄い黄や青といった色に変更し、コントラスト値を下げる機能を実装した。

【検索機能】

既存の論文検索サイトにある検索機能と同様で、全文検索で任意のキーワードを含む文献が検索できるほか、タイトル、著者、セッション名などの属性を指定した検索ができる。そして連想検索エンジン GETA を使用した関連文献検索も可能である。

【脚注表示機能】

本システムは、現在表示している論文のページの本文を解析し、ページの補足情報をページの左右の脚注部（Sidenote）に表示する機能を有している。現在表示できる補足情報には次の 2 種類がある。

・本文中のキーワードに関する情報：辞書や百科事典のような見出し語集合とその説明項目が存在するとき、ページ本文中から見出し語を抽出し、説明部分を脚注部に表示する。本システムでは、Wikify や Amazon Kindle の X-Ray5 の技術と同様に Wikipedia をリソースとして用いており、本文中に Wikipedia のタイトル文字列が出現したとき、その説明文と画像を表示している。本文中でマッチしたキーワードは、文書画像の上にオーバーレイでハイライト表示する。現状、キーワードの語義曖昧性解消や表示する説明のランキングなどは力を入れておらず、今後の課題となっている。

・ページの一部と関連する情報：表示しているページの本文の全部もしくは指定した一部に対して、関連する情報を検索し、ヒットした項目を脚注部に列挙する。本システムでは検索アルゴリズムには前述の GETA を、検索対象のリソースとして Wikipedia と言語処理学会の予稿集を用いている。

我々の知る限り、論文に対し外部のリソースを本文に結びつける形で併記して表示するシステムは、いまのところ存在しない。

また、デモシステム以外にも、論文とその発表スライドについて、人手で論文中の段落と発表スライドの各ページを対応付け、それを Sidenoter で表示させるなどの可能性も検討している（図４）。このよ

うに、従来、XML 文書中のアノテーションタグや、テキストに対する簡易な形での付加情報としてしか付与できなかった言語処理の解析結果を、可読性の高いレイアウトで可視化することも可能になる。

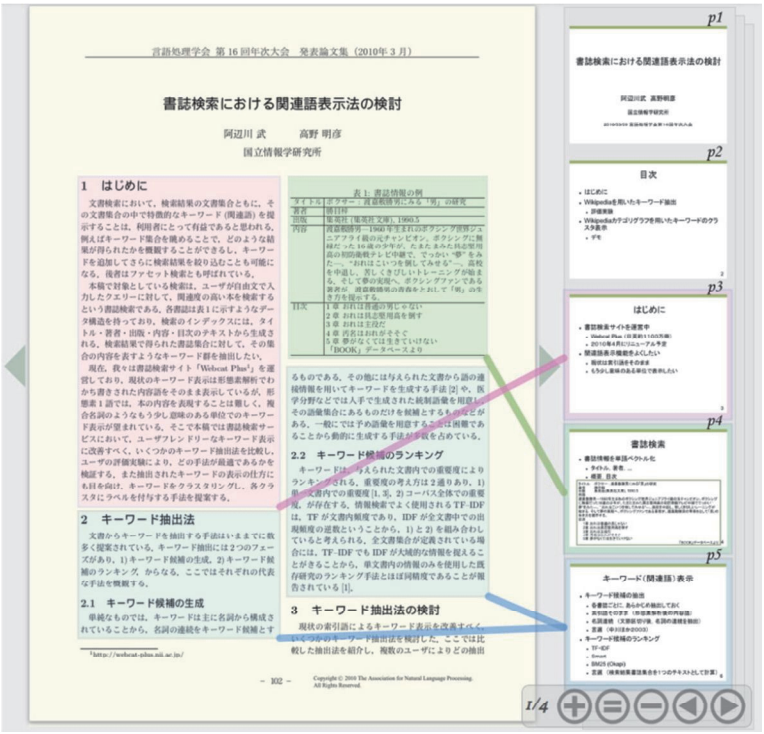

図4：スライドと論文対応の例

今後の開発予定として、言語横断の関連文献（ACL Anthology など）の検索表示、SNS に対応したコメント追記機能などを考えている。電子化にともない XML ファイルとして投稿された論文を、本システムではどのようにして扱うのか、いったん PDF を経由していままで同様に処理するのか、XHTML や EPUB のようなリフロー型のフォーマットで直接ブラウザ上で表現するのか、これも今後の課題である。

(3-a) 多分野の科学論文にまたがる共参照関係コーパスの作成、およびそのコーパスを用いて論文中から用語を抽出し、共参照関係検出に利用する手法の開発

複数の科学分野にまたがる論文に対して、テキスト中の専門用語およびそれらの間の共参照関係を付与することで、科学論文に対する共参照関係のコーパスを作成した。また、このコーパスを用いて、従来の共参照関係検出手法がどの程度有効であるか検証を行ったところ、共参照関係の候補となり得る用語の抽出に失敗することが原因で精度が非常に低くなっていることが判明したため、このコーパス上で学習した技術用語抽出モデルによって自動抽出される用語情報を利用することで、精度の改善に成功した。

コーパス作成の対象文書は、既存の共有タスク SemEval-2010 Task 5 のデータセットを利用した。このデータセットには 4 分野計 284 件の科学論文が含まれており、元は重要語自動検出用のセットである。このデータセットの各論文の抄録に対して、専門用語にアノテーションを行い、その上で、それらの間の共参照関係を付与した。専門用語はそれぞれ考えうる最大スパンのものをとり、冠詞は含めない形でアノテーションを行った。共参照関係に関しては、一般分野でのタスクである MUC-6 task の定義をベースに、命題以外の数式、関係代名詞、結合名詞句に対してもアノテーションを行うよう拡張した。その結果 4,228 の言及間の 1,362 の共参照関係を含むコーパスを獲得した。

このコーパスを訓練用・開発テスト用・評価用に分割し、既存のシステム Decoref による共参照関係