プロジェクト名：　異分野研究資源共有・協働基盤の構築
（略称：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）

〔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 Researchmap α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><td rowspan="2">テーマ</td><td>H22 年度</td><td>H23 年度</td><td>H24 年度</td><td>H25 年度</td><td>H26 年度</td><td>H27 年度</td></tr>
<tr><td>（予備研究）</td><td></td><td></td><td>中間評価</td><td></td><td>事業化</td></tr>
<tr><td>全体</td><td colspan="5">実システムへの適用・Web 空間との連携・実証研究・改良</td><td>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>「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>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>多種データ間の連想検索技術の研究開発</td><td colspan="2">多種データ間の連想検索技術の改良と深化</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>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 23 年度

サブテーマ１では、実験的に立ち上げたプロトタイプシステム上で論文推薦システムの課題や解決法を検討した。また、著者同定と引用文献同定に基づく個人プロファイルの自動獲得、引用および共著者・共同研究者ネットワーク情報の補間手法について検討した。前年度で設計したデータベースに、実際に運用に使われている論文データベースおよび著者情報を入力して、効率や性能の予備調査を行い、自動データ更新に向けて準備を進めた。また、コンテンツベースの推薦性能を高めるため、論文主題スキーマの設計や論文主題の抽出などに向けた課題を整理した。

サブテーマ２では学術研究はどの分野でもデジタル化され、多種多様大量のデータが生産され利用されている。本サブプロジェクトの目標は、このようなデジタル化された学術情報を統一的に扱い相互の関係をつけることのできる Linked Open Data の方法論を展開して、学術活動を支援することである。この実現に向けて以下の 5 つの活動を行う。

1. 基盤ソフトウエア環境構築
2. 基盤データベース構築
3. 研究コミュニティ支援サービス構築
4. 発展ソフトウエア開発
5. プラットフォーム構築

22 年度は第 1 項の基礎的なソフトウエア環境構築と美術情報を例に第 2 項の基盤データベース構築を行った。23 年度も基本的に第 1 項と第 2 項を中心に研究開発を行った。

第 1 項としてはまず汎用的半自動スクレイピング技術の開発を行った。これは Linked Open Data (LOD)ではなく一般の Web ページとして公開されている情報を半自動で効率的に収集することができる仕組みである。Web ページの定型的な構造を抽出して明示化すると同時に簡単なマッピング言語を使って既存のスキーマと結びつきを定義できる。この方法はデータベース型の Web ページからの RDF 生成として効率的である。また、LOD における変遷を記述する変遷のオントロジーの開発を行った。後述の生物種情報などにおいてはある時点で正しかったデータが書き換えられることがある。このような変

遷の種類や役割を列挙して適切なオントロジーを定義し、この定義を使って変遷を記述するシステムを構築した。

第 2 項としてはバイオ分野の基盤データベースとして生物種メタデータデータベースに着手した。これは極地研、遺伝研を初めとする GBIF 活動グループ、DBLCS との協力の下で行っている。生物種の情報は生物を研究するにあたって基盤的な情報であるが、現在は分野や注目するレベルなどによって多種多様なデータベースが散在する状況である。ここでは生物種の情報を一元的にアクセス可能でかつ多様なデータベースへガイドすることができるメタデータデータベースを構築する。また国内情報と国際情報を適切に関係づけられるために和名にも取り組む。今年度は実験的に 3 つのデータを LOD 化した。1 つは蝶類のデータベースで和名図鑑のデータに基づき国際的なデータベースとリンクを行った。2 つめは苔類のデータベースで、これは上位タクソンを含めてデータベース化した。3 つは DBCLS が開発した生物学辞書（BDLS: Building Dictionary for Life Science)の LOD 化である。

また第 4 項の発展的ソフトウエア開発として、GIS を含む多様な情報源から LOD 情報を組み合わせるソフトウエアの試作を行った。

サブテーマ 3 では、連想情報処理基盤の研究については、文化遺産オンライン由来の文化財情報と、Webcat Plus 由来の人物情報（著者情報）を起点として、高信頼な公開情報を収集・整理した。文化財情報と人物情報の 2 つを軸として情報を集約整理する方式について検討した。また、ウェブ上の学術関連のニュースと Researchmap に登録された研究者情報を紐付けた上で、分野ごとに編集して公開する研究ニュース集約システムの開発を行った。

サブテーマ 4 では、ユーザインタフェイス等を検討した上で、以上の 3 つのサブテーマの研究成果を Researchmap に反映させ、研究者コミュニティに提供、融合研究を支援した。他機関・他プロジェクトのデータベースとの連携を開始した。具体的には科学技術振興機構（JST）が提供している ReaD の基盤ソフトウェアとして Researchmap を提供し、世界有数のデータ量を誇る研究融合基盤を構築した。また、集約された研究情報を大学等の機関にフィードバックするための API を設計し、これを表示するためのウェブシステムを開発した。これにより、大学等が研究情報を網羅的に入力するインセンティブが高まり、日本の研究者情報が一箇所に集約され、効果的な機械学習を可能にすると考えられる。集約された研究情報をもとに、サブテーマ１、２、３のそれぞれの要素技術と連携し、実システムに昇華させていくための技術的ハードルを洗い出し、研究課題としてフィードバックを行った。

平成 24 年度（中間評価）

サブテーマ 1 では、研究者の論文検索における嗜好を調査し、利用要求を分析する。研究者個人プロファイルと論文との関連度指標について検討し、特に研究者の多様な利用要求に対応するための推薦手法の開発・実証を行う。専門用語辞書やウェブ情報などの外部情報源の活用方法を検討するとともに、言語的な手法に基づく論文記述の解析や記述どうしの相互の参照関係抽出の手法について課題を整理する。また、論文記述の深い解析を支える基盤として、構文解析技術の解析対象を、抄録などの平坦で整えられた文から、構造・表現に柔軟性のある論文全体へと拡張するための基本枠組を構築する。

サブテーマ 2 では、本年度までにデータ中心型研究の基盤のプロトタイプを完成させる。具体的には基盤的ソフトウエア開発および基盤的データベースを完成させる。また応用的ソフトウエアとして GIS を含む多様な LOD データを連係される。

1. 基盤ソフトウエア環境構築

1.1　データ統合ソフトウエア開発

複数のデータサイトからデータを収集すると、それらの関係の管理が重要になる。本プロジェクトではデータの多様性を維持したままデータを統合するために、共通で核となるデータとそれに結びつけられた個別のデータという構造でデータを管理する。このようなデータ管理を可能とする仕組みを考案して、実装を行う。

2.　基盤データベース構築

2.1　生物種情報 LOD

前年度から廃止した生物種情報 LOD のスケールを拡大して、一通りの生物種を格納し、和名を含む名称検索やタクソン構造が辿れるようにする。また EOL や GBIF など国際的に重要なデータサイトとの連携を実現する。

2.2　シソーラス・百科事典・辞典情報

分野横断的情報としてはシソーラス、百科事典や辞書の情報はハブとして機能する。百科事典としては日本語 Wikipedia を対象として、その Linked Data 化（日本語 DBpedia）を行う。辞書情報としては日本語 WordNet などのオンライン辞典を統合して Linked Data 化する。また関係する学術分野のシソーラスの Linked Data 化も行う。

3.　研究コミュニティ支援サービス構築

3.1　環境プロジェクト連携

前年度に引き続き、国立遺伝学研究所の GBIF(Global Biodiversity Information Facility)活動と連携する。上記の生物種メタデータデータベースを基盤して使い、GBIF データの取り込み支援システム、データの連係支援システム、可視化システムなどを実装する（共同研究者：神保宇嗣（国立科学博物館））。

3.2　GIS プロジェクト連携

前年度に引き続き、国立極地研究所と連携して、GIS とのデータ統合の仕組みを研究する。（共同研究者：小林悟志（極地研））。

3.3　データベース検索支援

上記での構築した生物種メタデータベースを用いて、DBCLS が収集したデータベース検索およびデータ検索における検索支援を行う。ユーザが入力した語を生物種メタデータベースに問い合わせて関連するタクソンの追加やその和名学名変換を行うことでユーザが必ずしも適切な語を入力しなくても検索できるようにする。（共同研究者：川本祥子（DBLCS））

サブテーマ３では、震災アーカイブ内の写真や国立極地研究所の観測データなど、情報が取得された日時と場所だけが記録されたコンテンツが大規模に生成されている。これらのコンテンツについて日時や場所を指定しての検索は可能であるが、基本的には写真データ内、場所データ内のように同一種類のコンテンツ内で閉じており、今まで培われている大量のテキストコンテンツと同じ枠組みで検索することができない。そのため時空間コンテンツを格納したデータベースとテキスト情報からなるデータベースとを横断的に検索する仕組みが求められている。

連想検索は、単語ベクトル化された文書集合に対し、自然文や複数文書からの検索機能を提供するものである。複数のデータベースに対して横断検索するときには、どのデータカラムを検索対象として指定するかなど、それぞれのデータ構造を把握しお互いに連携する必要があるが、連想検索ではおのおののデータベースが自身で文書･単語行列を構築しておけば容易に横断検索できる仕組みを持つ。

画像や映像などのテキストベースでないコンテンツでも、タイトル、作者、説明文などのテキストから構成されるメタ情報が付与されていれば、それらの情報を利用して連想検索が可能である。

本研究では、テキストベースのコンテンツに時間･空間情報を自動的に付与し、メタ情報にテキスト情報を持たないコンテンツと横断的に連想検索ができる仕組みを提案する。本研究は主に 2 つの要素技術から構成される。

1. テキスト情報から時間、空間の情報を抽出する

   この技術は自然言語処理分野の固有表現抽出に該当する。ただし、固有表現抽出はテキスト中に明示的に出現している時間、空間を表現する文字列だけを抽出するのに対し、提案手法では、時空間情報を表さない固有表現に対して Wikipedia などのデータリソースの情報を用いて、時空間情報を付与する仕組みを開発する。

2. 連想検索に時空間の概念を取り込む

   従来の連想検索では、単語から構成される語空間を検索対象としていたが、ここに時間、空間を取り入れる。複数のコンテンツを検索クエリーとした場合に対応するため、時間および空間の情報を確率分布の和で表現し、この値を離散化したベクトルを連想検索に用いる。連想検索時には、どの空間を重視して検索をするかをユーザが指定できるようにし、それぞれの空間の寄与度合いを調節可能とする。最終的には国立極地研究所と連携して GIS 情報を核とするデータと国立情報学研究所の持つ論文や書籍の情報とを横断的に検索できる仕組みを開発する。

サブテーマ 4 では、前年度設計した API を他のデータベースへの提供を開始する。特に、府省共通研究開発管理システム（e-Rad）および、大学研究者総覧への提供を行う。これにより、研究者情報の循環が促され、異分野の研究者の融合が促進されることが期待できる。これまで研究開発された要素技術を改良した上で、Researchmap 上に統合し、ユーザからの評価の他、アクセシビリティ等に関して、外部の評価を受ける。

平成 25 年度

サブテーマ 1 では、論文抄録や全文データの言語的な解析による拡張について検討する。論文の専門度や「手法」、「応用分野」などに関する情報を抽出して、推薦対象を論文から研究資源に拡大した推薦システムの実現を目指す。また、前年度に構築した論文全文に対する構文解析技術を各種推薦手法と組み合わせることにより、より深い論文内容理解から実現される、推薦システムの更なる拡充を目指す。

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では、連想情報処理基盤の研究については、引き続き文化遺産オンライン由来の文化財情報と、Webcat Plus由来の人物情報を起点として、高信頼な公開情報を収集・整理する。個々の文化財の高解像度写真、年表、歴史地図などを軸にして情報を集約整理して、それらに対する直感的な対話環境について検討する。

サブテーマ４では、前年に受けた中間評価を元に、要素技術を改良し、準備が整い次第、Researchmap上に再統合する。この年度までに、各関連機関と連携して、各種学術研究データのリサイクルの仕組みを完成させるとともに、実サービスとしての精度を高める。また、学術コミュニティだけでなく、産業界、また海外との連携を開始する。

平成26年度

サブテーマ１では、論文から抽出した情報を他のデータベースやウェブ上の情報に結びつけるとともに、APIを経由して、外部の学術コンテンツサービス関連サーバと連携して、推薦を通して研究者どうしの協働を促進するための基盤システムを開発する。また、より広範囲な情報リソースとの結びつけを実現するべく、多様な記述スタイルを持ったテキストの解析を実現可能とする基盤技術の実現を目指すとともに、論文から抽出した情報とその他リソースから得られる情報との質的な差異を摺り合せる手法について検討する。

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では、連想情報処理基盤の研究については、公開コースウェアを起点として、高信頼な公開情報を収集・整理する。収集された関連情報をコースウェアの参考情報として適切にひも付けるための技術的課題について明らかにする。コースウェアにおけるトッピクの変遷を自動的に捕捉することにより、関連性の評価を調整する方式を検討する。

平成 27 年度

サブテーマ１では、これまでに構築した基盤システムを活用するための実証システムを構築し、論文推薦手法および推薦の基本的アルゴリズムの評価を行うことによりその有効性を確認する。

サブテーマ２では、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。

以上を、統合プラットフォームのアプリケーションとしてリポジトリシステム WEKO 上でデータ操作ができる環境を駆逐する。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。

サブテーマ３では、連想情報処理基盤の研究については、これまで研究開発した要素技術を、文化遺産オンライン、Webcat Plus、想・IMAGINE などの公開サービスにフィードバックして、それらの新しいサービス構築に活用する。

平成 28 年度以降の展開

これまでの研究成果の下、Researchmap は、異分野研究資源共有・協働基盤システムとして完成し、産業界や海外の共同研究者も含めて幅広い層の研究者に提供されるサイエンス 3.0 基盤として成果を上げていることが予定される。本システムを、事業として成立させることも可能であるが、ResearchersID 等の商用サービスが海外では成立していることから、民間への移転も可能であろう。どちらがより研究者のメリットになるかを検討した上で、自律的なサービスとしてテイクオフさせる方策を検討する。

また、大学機関リポジトリ、ライフサイエンス統合データベース、包括脳支援データベースなど各種データベースから CiNii や KAKEN に至る、研究情報サイクルの輪を完成させ、一度の入力によって、すべての情報が再利用可能な状態にし、利便性を向上させる。

各サブテーマの研究成果である要素技術については、オープンソースとして、あるいは、WebcatPlus、文化遺産オンライン等の学術情報サービスに反映させ、広く社会に還元する。

特に、サブテーマ１では、これまで構築した基盤システム、実証システムの評価を行うとともに、安定したサービス運用という形で研究成果を発信するための方策を検討する。

また、今後の研究はデータ中心型研究の方向に向かっていくことが予想される。サブテーマ２では、データを保管場所、保管方法などにとらわれずシームレスに利用可能し、データ操作が自由に行え、データを用いた研究がその場で実施し公開できる、データに基づくバーチャル研究環境が必要なると思われる。これまでの本研究の成果を生かしてこのような分野を超えてデータを自由に操作できる環境を構築する。

サブテーマ３では、世界最速の連想計算エンジン GETA によるコンテンツ・コンパイル技術が、我が国を代表する博物館・美術館・図書館・文書館で蓄積されている高信頼な情報源に適用され、それらが我々ユーザの日常的な意思決定に活用できる情報サービスとして提供される。本研究の技術をさらに深

めて異種複数コンテンツの融合・再構成のための連想情報処理基盤が確立されれば、コンテンツ間に“化学反応”を起こして“化合物”を作り出す情報環境を実現できると考えられる。

〔3〕 研究推進・実施体制

本研究の推進にあたっては、(1)既存学術データベースと連携するためのセマンティックウェブ技術(2)異種データベースをつなぐリンケージ技術および検索技術(3)大規模データマイニングおよびオントロジー技術が不可欠となる。そこで、本研究プロジェクトを 4 つのサブテーマに分け、3 つのサブテーマにおいて、(1) (2) (3)に関する研究開発を行った上で、4 つ目のサブテーマにおいて、他のサブテーマでの研究成果を、研究者にサービスするための基盤の研究開発を行い、実際に大学・国内主要研究グループ・学会等に提供しながら、実証的に研究を推進していく。

本研究の実施に先立って行った、第一期新領域融合研究センターおよび次期「新領域融合研究センター」プロジェクト立案のための調査研究において、日本全国の研究者を対象とした研究者向けサイエンス 2.0 基盤サービス「Researchmap」の試行版を公開、運用を始めた。平成 22 年度中に 300 を超える組織から 1300 人を超える研究者が登録した。現在（平成 24 年 7 月）は、22 万人を超える研究者が参加、800 万件以上の研究情報データが研究者本人によって分類・登録されており、本研究が目指す異分野研究資源共有・協働基盤の構築を開始するための下地としては理想的な状態が整っている。また、本サービスの上には、既に 188 の研究コミュニティが構築され、「包括脳支援データベース」「共生社会に向けた人間調和型情報技術の構築（CREST 領域）」など重要な研究コミュニティの共同研究ツールとして実際に活用されている。筑波大学大学院生命環境科学研究科、総合研究大学院大学先導科学研究など、組織としての利用も増加しており、今後、日本の研究サービスの一翼を担っていることが期待されている。

(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

〔広島市立大学〕 難波英嗣

〔鳥取大学〕 村田真樹

(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、加藤文彦、原　忠義、小出誠二

〔国立遺伝学研究所〕 菅原英明

〔人間文化研究機構〕 山田太造

〔東京芸術大学〕 嘉村哲郎

〔ATR-Promotions〕 高橋　徹、上田　洋

〔慶應義塾大学〕 深見嘉明

(3) 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

研究代表者

〔国立情報学研究所〕 高野明彦

共同研究者

〔国立情報学研究所〕 西岡真吾、丸川雄三、阿部川武、萱島礼香、荒井紀子

〔国立遺伝学研究所〕 大久保公策

(4) 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〕 山地一禎、羽田昭裕、舛川竜治、南　佳孝

〔総合研究大学院大学〕 大田竜也

〔国立極地研究所〕 岡田雅樹

〔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

CiNii のデータベースを利用した論文推薦システム「オススメ論文検索システム」のプロトタイプを平成 22 年度に完成させ、平成 23 年度は、以下の 3 点を中心に進めた。

1. ユーザインタフェースの改良
2. 機能拡張
3. 推薦アルゴリズムの基礎となる研究

1. ユーザインタフェース(UI)の改良

既存の UI を改良するにあたって、問題点を洗い出した。昨年度開発した UI では、ログイン方法と推薦論文の表示方法について検討し、改良を行った。まずログイン方法については、キーワード検索と著者検索の二通りについて昨年度実装した。キーワード検索は、自由にキーワードを入力し、そのキーワードに関連する論文をリストにして表示する方法である。著者検索では、自分の名前を入力してログインすると、自分の過去に執筆した論文をベースに類似論文や類似研究者を表示し、シードとなる論文を選択して、推薦論文の表示を行う方法を実装した。この二つに加えて、Shibboleth 認証による Researchmap 連携として、Researchmap の ID でログインすると Researchmap ユーザが「公開」に設定した情報をベースにして論文を推薦する連携部分を構築した。

次に推薦論文の表示方法について、改良を行った。昨年度は、ユーザが選択したシード論文に基づいて 8 つの推薦手法（レコメンダ）により論文を表示する UI を実装した。推薦論文の表示方法は、2 列 4 行に配置して各レコメンダ上位 3 論文を結果として表示していた。各レコメンダの推薦手法が類似しているものもあったため、推薦結果が同じような内容を表示し、各レコメンダ間での推薦論文にあまり差異がないなどの問題点があった。そこで、UI を意識して表示結果の棲み分けを行った。論文のタイトルをリストにして表示する部分と、論文の中身に着目した部分とに分けた。まず、論文のタイトルをリストにして表示する方法として、速報性、類似度、人気度、異分野、入門性を横に並べた。コサイン類似度で計算し類似論文を 100 論文抽出した後、各レコメンダの数値に従って論文をソートできる機能を加え

た。以下の図 1 が改良した表示画面である。

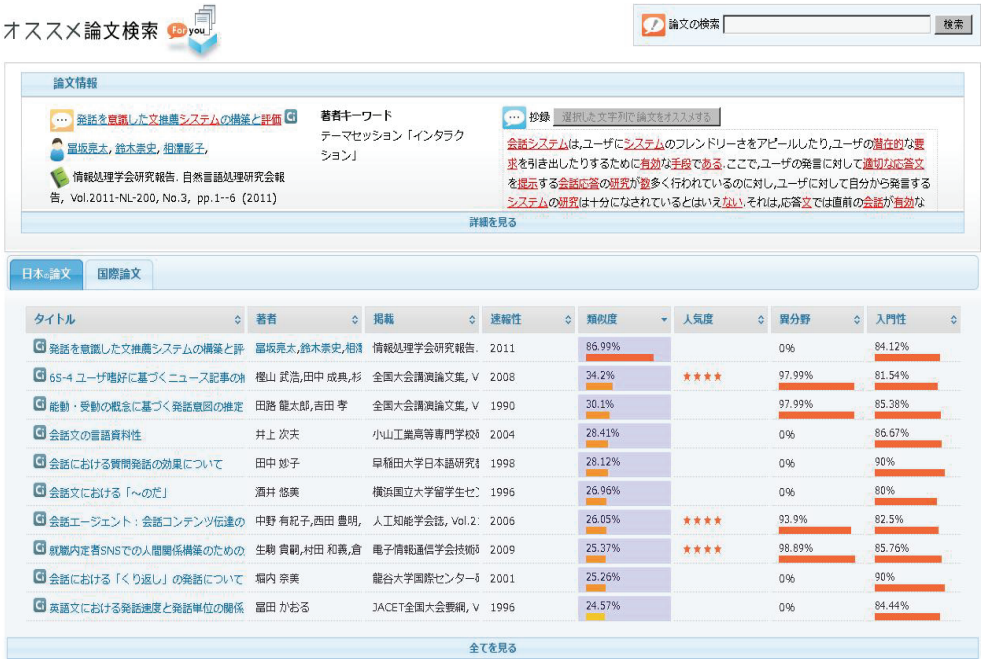

図 1: オススメ論文検索システム推薦結果表示画面

二つ目は、論文の内容を解析して、詳細な情報を提示する機能を加えた。これまで対象レコメンダ、手法レコメンダによって、論文抄録に含まれる同じ手法や対象を扱っている論文タイトルを提示していたが、特定の用語に対して同じ手法や対象を扱う論文を表示できるようにした。その際、用語に注目するため、用語に対する注釈として Wikipedia に記述されている情報や市販されている用語辞書の定義文を用語に付与して参照できるようにした。推薦論文の表示方法について、国際論文はユーザから日本語を入力するだけで、英語の論文が推薦される機能は非常に便利であるとの評価を受け、更に詳細に表示できるように別タグとして、日本語論文と分離した。これにより国際論文についても類似度や年度でソートできるようになった。

2. 機能拡張

今年度は、センテンス単位検索機能と管理者用機能の二つを追加した。まず、センテンス単位検索機能として、これまで論文を推薦する際、一つの論文の抄録を単位としていたが、より狭い範囲で論文を検索したい場合に対応できるよう、センテンス単位で論文を推薦できるような機能を追加した (図 2)。論文の抄録の中に含まれるセンテンスをドラッグして、「選択した文字列で論文をオススメする」のボタンをクリックすると、選択した文字列を含む類似した論文を推薦することができる。

次に管理用機能では、辞書管理と研究用テキストデータ管理を操作する機能を追加した。辞書管理機能では、管理者が簡単に辞書を追加、削除、編集できる機能を実装した。管理者が自分で用意した、あるいは購入した辞書を随時追加することによって、用語の注釈を網羅的に表示することが可能になる。また、用語編集では、適切ではない見出し語などの削除をすることにより、見出し語のフィルタリングを行うことができる。研究用テキストデータ管理では、研究者がオススメ論文検索システムのデータベースとなっている言語リソースに簡単に、必要なだけダウンロードできるインタフェースを追加した。論文推薦のために、さまざまなリソースを収集しているため、そのデータを研究に利用できるような基盤を構築した。

図 2: センテンス単位検索機能

3. 推薦アルゴリズムの基礎となる研究

推薦アルゴリズムのうち、国際レコメンダと入門レコメンダについて基礎となる研究を行った。まず、国際論文レコメンダは、日本語の論文をキーにして、多言語の論文を推薦するためのレコメンダである。多言語文書の処理には、機械翻訳や対訳辞書などを使って言語間の変換をした後に処理する方法と複数言語で記述された文書を同一の特徴空間にマップして処理する方法が考えられる。学術文献は新しい概念や技術を扱うことが多く、新しい用語が使われる頻度も通常の文書より多いことが予想される。このような文書に対しては作成コストの高い対訳辞書等を用いる手法よりも統計的な特性を用いて同一特徴空間にマップするほうが適していると考え、後者のアプローチによる日英 2 言語で書かれた論文を推薦するための論文の特徴の抽出法を提案した。提案手法では、Latent Dirichlet Allocation (LDA) と同様に生成モデルに基づいた特徴を抽出した。

単一の言語で記述された論文が与えられた場合、単語の生成確率を固定してギブスサンプリングを行い、トピック生成確率を求める。この分布を当該論文の特徴ベクトルとして用いる。トピックと単語の関係は、単語生成確率として求められているため、この確率分布より論文に含まれる単語から対応するトピックの分布を求める。論文間の類似度には、トピック分布の KL-divergence に加え、分布を特徴ベクトルと見なしてコサイン尺度やユークリッド距離をもちいることも考えられる。このモデルを論文に付与されるキーワードを推定する問題に適用し、日本語の概要から英語のキーワードを付与したり、逆に英語の概要から日本語のキーワードを付与する性能を調べたところ、同一言語間での推定とほぼ同じ精度で 2 言語間での推定が行えることが実験的に確かめられた。

次に入門レコメンダについての研究について報告する。入門論文レコメンダは、これからその分野を学ぼうとする初学者（学部学生から他の分野の研究者など）に、分野の概要が分かるような解説論文や、理解しておくべき基礎的技術を説明している論文などを推薦することを目指したレコメンダである。入門論文を推薦するために、初学者が理解しなければならない基礎的な用語を抽出することが重要となる。その用語の理解がなければ論文を読み進めることができない重要で基礎的な用語のことを分野基礎性が高い用語と定義する。分野基礎性の高い用語は優先度によってレベル分けをする必要がある。用語基礎性が高い順にレベル分けされていれば、効率的に学習を進められる。

入門レコメンダに必要な要素技術として、(1) 分野基礎性が高い語の抽出方法、(2) 分野基礎性の度合いを識別する方法、の二種類が必要となる。分野基礎性が高い語の抽出は従来行われてきた専門用語抽出の手法が有効であるかを確認する。度合いを識別する方法は、先行研究において対象となる学習者を設定し、それに対応する用語のレベル分けを提案したが、ここでは(1) の抽出方法において、上位に順位づけられている語が分野基礎性の度合いが高いものであるかを調べた。論文において分野基礎性が高い語は、経年推移性、網羅性、語構成性から考えることができる。

まず、経年推移性の観点において、分野基礎性が高い語は、論文中の出現パターンが比較的安定して、長い期間出現しつづけると考えられる。たとえば、自然言語処理分野で基礎性が高い用語の「形態素解

析」は、その分野の論文が出版された時から出現し、初期の段階では手法に関する研究が多かったが、現在はすでにツールとして広く用いられているため、継続して論文中に記述される。網羅性の観点においては、特定の論文に多く出現するような用語ではなく、平均的に論文中で用いられる用語や、事典の場合には、複数の異なる章にまたがって出現するような用語が分野基礎性が高い用語と考えられる。

語構成性とは、多くの複合語を生成する性質のことで、新しい複合語の基になる用語は基礎性が高いと考えられる。前述の「形態素解析」という用語では「日本語形態素解析」「形態素解析システム」など前後に多くの用語が接続して複合語を生成したり、文章中に「形態素解析中」や「形態素解析失敗」など接尾辞や文章を短縮したような表現の臨時一語を生成する傾向にある。専門用語の前後に連接する頻度や異なり語数に基づいた用語抽出手法も提案されており、分野基礎性を測る上で語構成性は重要な観点であるため、語構成性に絞って用語の抽出を行う。

用語抽出を行う際、論理構造における出現パターンを利用した方法 が有効であったことから、論理構造を考慮した。論理構造とは「タイトル」「著者」「抄録」などの書誌情報や、本文中の「関連研究」「実験」「考察」「まとめ」など論文の詳細な構造のことを指す。実験に使用したデータは、情報処理学会自然言語処理研究会 14 年分 1993 年から 2006 年までの 1421 論文の書誌情報（タイトル、抄録、キーワード）を対象として分野基礎性が高い語の抽出を行った。デジタル言語処理学事典の索引語と論文の著者キーワードの中から、自然言語処理の専門家が特に分野基礎性が高いと判断した 500 語を 4 段階に分類したものを正解データとして利用した。対象となる複合名詞は、MeCab で形態素解析をし、名詞が連続している単語列を抽出し、各属性の頻度を計算した。評価方法として、分野基礎性が高い語の抽出精度については、各属性の単独頻度と、これらの属性値を利用した用語スコア付けの FLR 法、C-Value 法、MC-Value 法を用いて上位 100 語(@100)、300 語(@300)、500 語(@500) の順位を求め、それぞれ F 値を算出して評価を行った。全体的に精度が低かったが、その理由として従来の専門用語抽出と異なり、正解セットを基礎性が高い用語に絞ったため、用語として適切であっても正解と見なさなかったことが考えられる。また、C-Value と MC-Value は語構成の特徴を利用したスコア付けであったが、二字漢語の場合、多くの専門用語や複合語の語基となっているため頻度が高く、語構成要素数を考慮してもスコアが高くなってしまう傾向にある。

今後は分野基礎性が高い用語に対しては、各属性値に文字数や語構成数を総合的に組み合わせて拡張する必要がある。次に、論理構造別の精度では、特に目立った差がなかった。著者キーワードが有効であることは確認できたが、全ての論文にキーワードが付与されているわけではないため、書誌情報全体の頻度を扱うことが妥当だと考える。将来的には、書誌情報以外の「はじめに」「関連研究」「実験」「考察」「おわりに」等の論理構造毎の出現傾向も分析する予定である。

サブテーマ２

1. はじめに

Web が普及し、多くの情報がインターネットを通じて入手可能になった。しかし、その情報は人間が読むことを前提に作られており、コンピュータを通じて利用しているも関わらず、コンピュータがその内容を処理することは容易ではない。Web の発明者である Tim Berners-Lee は、Web の前提として、人間だけでなく、コンピュータもその内容を処理可能であることが必要だと考えており、その仕組みとしてセマンティック Web を提唱している[Berners-Lee01]。しかし、人間が読む情報の共有という点で Web は大変強力であり、Web が急速に普及したことに比べて、セマンティック Web は必ずしも発展・普及したとはいえない。

ところが、Web が社会にあまねく普及し、膨大な情報が Web にのるようになって、再びセマンティック Web の考え方、すなわちコンピュータが処理可能な形式による情報の公開の重要性が認識され

るようになった。特に大量のデータにおいてはこの必要性が強く認識されるようになった。そこで、概念的な定義ではなくて個別の情報をコンピュータが処理できる仕組みとして Linked Data という方法が提唱された。Linked Data はセマンティック Web の分野で開発されてきた言語(RDF, RDFS, OWL)を利用するが、主に個別の情報、データを記述する手段としてそれを用いる。

Linked Data におけるデータは RDF を用いて記述される。RDF はシンプルで柔軟性があり、多様なデータの記述が可能である。このような理由から、近年、Linked Data が情報流通の仕組みとして普及しつつある。ヨーロッパや米国では、すでに新しい情報公開・共有の仕組みとして認知されつつあり、我が国でも様々な研究や活動が行われている[武田 11]。

本研究では、生物学の中でも生物多様性の分野に焦点をあてた。この分野は、現在、生物多様性の損失や保全など、地球環境問題の 1 つとして社会問題にもなっている[UNEP92][環境省 10]。これらの問題を解決するには、対象生物のみではなく、地球規模の観測から人間活動まで様々な情報を横断的に利用できる基盤が必要である。しかし、生物の種名や分布、各種の特徴や保全状況といった生物的な情報でさえ、現状では形式や公開場所が分散しており関連が弱い。

そこで、本研究では、Linked Data の技術を用いて、分散的に公開されている生物多様性の情報を統合的に利用できるようにすることを考えた。

2. 生物多様性情報基盤整備の現状

生物には、分子レベルから生態系レベルまで多層のレイヤーが存在し、生物多様性もこうした多層レイヤーから構成されている。

本研究は、中でもその中核をなす種の多様性に着目する。このレイヤーでは、主に個体や種の名称・特徴といった情報が扱われ、大きく分けて(1) 生物名の目録の情報（種名情報）、(2) 標本や観察記録などの情報（分布情報）、(3) それぞれの生物種の特徴を示す情報（種情報）からなる。このようなデータを情報技術により保存・解析・活用することを目的とした横断的分野は、生物多様性情報学（biodiversity informatics）とよばれる[Bisby00]。

生物多様性情報は、生物分類学の研究成果として、18 世紀より紙媒体に蓄積されてきたが、情報技術が発達した現在では、膨大な情報を扱うデータベースに重要な情報ストレージとして蓄積されている。その例としては、グローバルなものとして地球規模生物多様性情報機構（The Global Biodiversity Information Facility：GBIF、種名・分布情報）、Encyclopedia of Life（EoL、種情報）、Catalogue of Life（CoL、種名情報）、Barcode of Life Data Systems（BOLD、DNA・標本情報）などが、国内では国立科学博物館が運営するサイエンスミュージアムネット（S-Net、標本情報、GBIF と連携）が挙げられる。

生物多様性情報は、様々なデータベースを通じて、データの種別や目的に特化した形式で公開されている。そのため、ある課題の解決のためには、複数の異なる Web サイトにわたる検索やデータの統合が必要であり、その統合には多かれ少なかれ摺り合わせが必要である。実際、同一種の情報であっても、異なる Web サイトに掲載されている場合、同一の検索キーでは、それぞれから適切な検索結果を得られないことがある。このため、相互運用性の確保は生物多様性情報学分野で最重要な課題の 1
海外では、Linked Data を提供しているプロジェクトとして GeoSpecies があるものの、まだ発展途上の研究プロジェクトであり、実用化には至っていない。

3. Linked Data 化のプロセス

生物多様性情報の Linked Data 化では、(1)基本となるデータの選定、(2)データ公開のための構成の決定、(3)関連データとのリンクと公開、(4)対象データの拡張の順で作業を行った。

3.1　データの選定

本研究で対象とする生物多様性のデータは、様々な組織から複数の Web サイトで公開されており、分類群、データの種類が多岐にわたっている。

本研究では、まず、対象分類群として蝶類を選定した。その理由として、パイロット研究に適当な種数であること、一般によく知られており科学的データをはじめとした様々な情報がリッチであり、Linked Data の利用が様々な場で期待できること、基本情報がデータベースの形で公開されていること、分類学者とも連携可能なことがあげられる。

様々な多様性情報をリンクする最も重要な要素は生物の名前、すなわち種名である。そこで、本研究ではまず種名情報の整備を目標とし、ソースとして日本産蝶類和名学名便覧を選択した［猪俣 11］。日本産蝶類全種にあたる 327 種・亜種について、所属分類群・学名・和名などが記されたもので、専門家が最新の知見に基づいて編纂しているため、基本となるデータとして適当であると判断した。

3.2　データの構成

蝶類の種名データは、分類体系を構成する要素、すなわち分類群に関する項目（界名・門名・綱名・目名・科名・亜科名・族名・亜族名・属名・亜属名・種小名）と種を表現する要素、すなわち種名に関する項目（学名・著者・出版年・和名・和名の別名）に大別できる。そこで、データを公開するために、分類体系を表現することと種名に関する情報を表現することを考えた。

まず、分類体系を表現するために、各分類群名に対して URI を定義し、さらに分類体系の階層性を表現するため、木構造の根にあたる界名以外の分類群名に項目については、上位階層の分類群を指し示す URI にリンクした。分類に関するデータについて、図 3 に Lepidoptera 鱗翅目（チョウ目）の例を示す。

次に、種に関する情報を表現するために、学名に URI を定義し、和名、著者、出版年に加え、所属する分類群を指し示す URI にリンクした。種に関するデータについて、図 4 に Papilio xuthus アゲハ（ナミアゲハ）の例を示す。

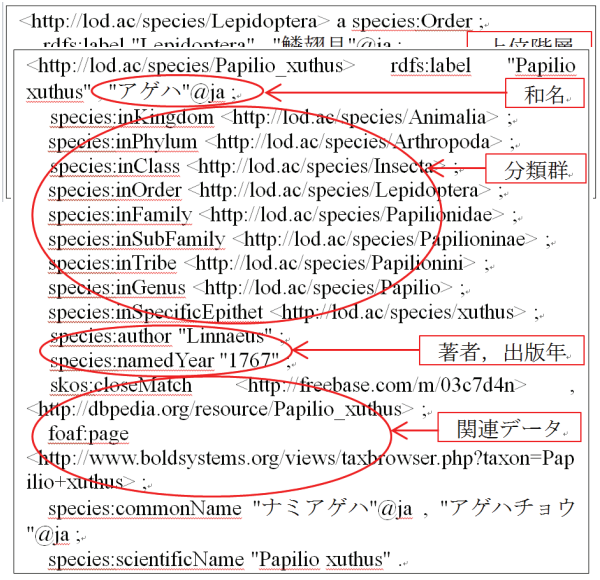

図 3:　分類に関するデータ（Lepidoptera の例）
図 4:　種に関するデータ（Papilio xuthus の例）

3.3　関連するデータとのリンク

蝶類に関する多様性情報リソースは数多く、前述した GBIF，EoL，CoL，BOLD，S-Net，GeoSpecies などに保存されているほか、DBpedia や Freebase など既存の Linked Data リソースにも蝶に関する情報が存在する。そこで、これらの Web サイトに対して、分類名と学名をキーに検索し、該当した Web ページへのリンクを作成した。図 3、図 4 で付与されているのは、GoogleRefine を用いて変換した RDF（Resource Description Framework）であり、BOLD，DBpedia，Freebase のデータに対しては GoogleRefine 上でリンクを生成できた。その他の Web サイトに対しては、Web サイト毎にスクリプトを記述してスクレイピングを行い、必要な情報を取得した。このうち、S-Net のデータは、各博物館の所蔵標本情報である。そのため、S-Net のデータには、標本を所蔵する博物館の情報が含まれていたため、その情報を活用して LODAC の博物館情報にリンクした。また、標本は 1 種に対して複数存在するので、各標本データに URI を定義し、種の情報へリンクした。Linked Data 化した S-Net の標本情報について、図 5 に Papilio xuthus の例を示す。

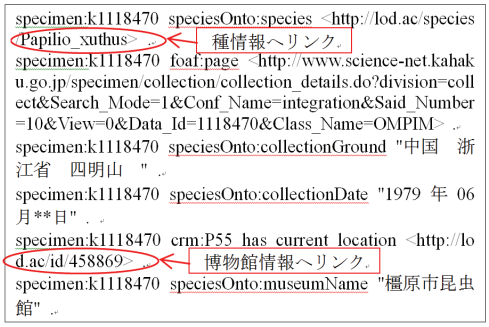

図 5：S-Net のデータ（Papilio xuthus の例）

3.4　データの拡張

データを拡張する対象として、国立極地研究所に収蔵されている蘚苔類の標本データについて、同様に Linked Data 化することを考えた。標本データには、蝶類データの項目以外に、亜綱名、亜種名、ID、採集日、採集者、緯度・経度、地域番号、地域名、収蔵場所、分布していた状況、標高、採集者、（掲載標本集：CBM，HIRO，HIRU，HYO，KOCH，NICH，NIPR，TNS）といった情報が記されている。そこで、前節までと同様のプロセスで分類に関するデータ、種に関するデータ、標本に関するデータについて、Linked Data 化を行った。

4.　Linked Data 化の結果

本研究で、生物多様性の情報について Linked Data 化を行った結果、図 6 に示すように、関連する情報を一覧で提供できるようになった。また、リンクした他の生物多様性情報の Web ページをインラインフレーム（iframe）によって表示することで、関連する情報の俯瞰が可能になった。さらに、種情報の Web ページに DBpedia の画像を表示することで、掲載情報が利用者の探している蝶の種類かどうかを特定できるようになった。

また、本研究では、Linked Data 自体の外部インタフェースとして SPARQL Endpoint を公開した。これにより、複合的、横断的な検索が可能になった。

図 6：結果表示例

5.　BDLS の Linked Data 化

生物種に関わる異なる情報源として BDLS の Linked Data 化も行った。BDLS（Building Dictionary for Life Science)とは、情報・システム研究機構ライフサイエンス統合データベースセンターが 100 近くの多様な生物に関わる辞書を統合して一つの辞書として構築したものである [1]。主に生物種を中心とするタクソン情報（学名、和名）と用語（日本語と英語）が含まれている。

[1] http://lifesciencedb.jp/bdls/

5.1 Linked Data 化の方針

BDLS は個別の情報に必ず出典が付されている。このため、この出典ごとに Named Graph として表現した。語彙については、タクソン間の関係などは geospecies[2] で定義されている語彙を参考にして構築した（図 7 参照）

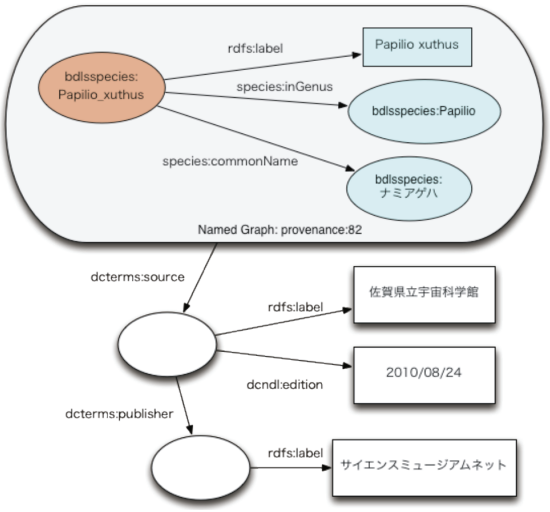

図 7: BDLS のデータモデル

5.2 結果と応用

上記の方法で RDF Store に登録を行った。現在、6,366,545 トリプルがある。HTML によるインタフェースと SPARQL Endpoint を用意している。

このデータを利用した応用例として、文献検索システムとの連携を作った。ここでは、種名で検索するとその和名や同属の別種も含めた検索結果を得られる。（図 8 参照）。

図 8: BDLS データを検索に利用した例

6. おわりに

本研究では、生物情報基盤構築に向けて、生物関連データの Linked Data 化を行った。これにより、分類体系・種名・種の特徴・標本に関する情報を柔軟に組み合わせて閲覧できるようになった。

本システムが生物多様性分野に与えるインパクトとして、以下の 2 点があげられる。1 点目は、既存のシステムで困難だった複雑な検索が可能となったこと、2 点目は、分野横断型の情報システム作成を可能にする基盤が構築できたことである。これらは、生物多様性情報学の課題である相互運用性の向上ともつながっている。

[2] http://lod.geospecies.org/

今後は、更なるデータの拡充を考え、データを容易に追加できる仕組みや、目的に応じて利用しやすいインタフェースを備えたアプリケーションの開発を目指す。情報とシステムを整備することで、本研究の成果が、社会問題として生物多様性保全問題を扱う際に不可欠な情報基盤になると期待される。

参考文献

[Berners-Lee01] T. Berners-Lee, J. Hendler, James and O. Lassila: The Semantic Web, Scientific American, May 2001, p. 29-37.

[Bisby00] Bisby, F.A.: The quiet revolution: biodiversity informatics and the Internet, Science, Vol. 289, No. 5488, pp. 2309-2312, (2000)

[Edwards00] Edwards, J.L., Meredith A.L., and Nielsen E.S.: Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science, Vol. 289, No. 5488, pp. 2312-2314, (2000)

[UNEP92] UNEP CBD, Convention on Biological Diversity, (1992)

[猪俣 11] 猪又敏男，植村好延，矢後勝也，神保宇嗣，上田恭一郎：日本産蝶類和名学名便覧. http://binran.lepimages.jp (2010)

[環境省 10] 環境省，生物多様性国家戦略 2010，(2010)

[武田 11] 武田英明，嘉村哲郎，加藤文彦，大向一輝，武田英明，高橋徹，上田洋：日本における Linked Data の普及にむけて, 2011 年度人工知能学会全国大会，人工知能学会, 2011.6.

サブテーマ３

サブテーマ３では、研究者情報を含む異種データベースをつなぐリンケージ技術および検索技術のひとつの実装として「研究成果発表に特化したデイリーニュースサイトの開発」を行った。

日本の大学･研究機関に所属する研究者の研究成果をより広く伝えることを目指し､科学技術や社会調査に関連するニュースを収集後、整理された形で提示するサイトを構築した。本サイトは、ニュース内で言及されている研究者に着目することで、各研究者の業績を１つにまとめて提示するとともに、研究者の詳細が掲載されている Researchmap へ誘導する役割も有する。主な機能は、①毎日更新されるニュースを自動的に収集、②ユーザによるニュース記事の投稿と削除依頼を可能にする機能、③記事中で言及されている研究者の同定、④各ニュース記事を専門分野ごとにカテゴライズ、⑤web サイト構築とニュースレイアウト機能、⑥管理機能がある。

システム設計

ニュース記事のクローリング機能およびスクリーンショットの保存機能

研究に関連するニュースを特定のニュースサイトよりクローリングし、ニュース記事をニュース DB へ保存する。クローリング対象のニュースサイトは google ニュースおよび、RSS を用いている任意のサイトとする。これらのサイトには、大学や研究所が広報として配信している RSS を利用することを考えているが、当面は RSS を用いているサイトとしてサイエンスポータルを利用する。ニュース記事がニュースサイト上から消去される可能性を考慮し、掲載記事の存在を保証するために、ニュース記事のスクリーンショットを画像ファイルとして、掲載媒体や後述する抽出した研究者名と関連づけて内部サーバに保存する。

ユーザによるニュース記事の投稿と削除依頼

google ニュースに掲載されなかった研究に関するニュース記事の存在を考慮し、一般ユーザ自らが関

連する研究についてのニュースの URL を投稿する機能を持つ。研究成果と関係のない記事の投稿を防ぐために現在は、Researchmap の登録ユーザに限定してニュースを投稿できるようにしている。また、自分が関連すると判断されたニュースに間違いなどが存在した場合に削除の依頼を行う機能も有する。これらの投稿は、システム運用側が内容の真偽について確認し、その投稿を採用するかどうかを決定する。

記事中で言及されている研究者の同定

Researchmap に登録されている研究者がニュース記事中で言及されている場合、システムが自動的に研究者を識別し、記事から Researchmap の該当研究者へのリンクを生成する。これによりニュース記事と研究者を紐づけることができるため、研究者ごとの言及されたニュース記事一覧ページを生成することができる。

現状では、Researchmap 上に同姓同名の研究者が複数いる場合や、研究内容とは関係ない箇所で言及されているなどの問題には厳密な対応を行っていないため、運用者が記事を公開する前に正しく研究者が同定されているかを確認する必要がある。

ニュース記事のカテゴリ分類

クロールしたニュース記事をあらかじめ定義した専門分野のカテゴリに分類する。分類には機械学習アルゴリズムを使用し、訓練データから分類器をモデル化し、任意のニュース記事に対し、いずれかのカテゴリを付与する仕組みを開発した。分類するカテゴリは、科学研究費補助金「系・分野・分科・細目表」や既存の科学系ニュースサイトの分類、さらに実際のニュース記事の配信頻度を鑑みた結果、以下の 7 カテゴリとした。

表 1

医歯薬学	Medical
工学	Engineering
生命科学	Life science
地球惑星科学	Geoscience
社会人文学	Humanities
理学	Science
その他･不明	Other

次に Web からニュース記事を収集し、人手で上記カテゴリを付与し、「その他･不明」以外のカテゴリについて訓練データとして各 100 記事ずつ収集した。これらの訓練データをもとにナイーブベイズ分類手法を用いてカテゴリ分類を行った。訓練データを用いたカテゴリ分類実験を行った結果、全体の精度は 5 分割交差検定で約 69%であった。

web サイト構築とニュースレイアウト機能

ニュース記事の表示は新聞の紙面をイメージし、大きさの異なる矩形で記事を囲い平面的に配置した(システムのスクリーンショットを参考)。ニュース記事内で記事に関連した画像が用いられている場合には、その画像を大きく表示することでインパクトを与えるとともに、ニュース内容を視覚的に理解できるようにしている。読者が閲覧した当日に収集された記事を最初に表示し、ボタンをクリックし時間をさかのぼって過去の記事を表示する。また、分類したカテゴリごとに記事を一覧できる機能を有する。

運用者管理機能

すべての記事は一般読者に公開される前に、運用者のチェックを受ける。このときにカテゴリ分類、写真の有無、同定された研究者の確認を行い、間違った情報は人手で修正し、最終的に公開可能であるかを判断する。収集したニュース記事には犯罪、資金流用などの不祥事についての内容を伴うものもあり本サイトで扱う趣旨とは異なるが、「窃盗」「汚職」といったあらかじめ登録しておいたキーワードが記事の本文中に含まれている際、自動的に非公開とする機能を有している。

今後の課題

本システムは、カテゴリ分類や研究者同定を自動的におこなっているが、訓練データの不足や同定アルゴリズムの不備により、誤りが発生する。カテゴリ分類の精度向上のためには訓練データの量を増やすことがもっとも有効であり、これは日々の運用で徐々に蓄積していくことが可能である。研究者同定については、研究者の所属情報を用いることが効果的であると考えられるが、Rsearchmap に登録されている研究者情報が古い場合、ニュース記事中の所属情報と異なる場合があり、必ずしも信頼性の高い情報源とは言えない。異なる研究分野に属する同姓同名の研究者の同定には、ニュース記事のカテゴリ情報が有効であり、一方で記事のカテゴリ分類には、同定された研究者の情報が有用であると考えられ、それぞれの解析は相互に関連しているためブートストラップ的に精度を向上させていく手法が有効である。

図 9： デイリーニュースサイトのスクリーンショット

また、サブテーマ３では、研究者情報を含む異種データベースをつなぐリンケージ技術および検索技術のもうひとつの実装として「外部情報源を効果的に提示するコンテンツ表示システムの開発」を行った。

学術論文や書籍といったテキストを主体として構成されるコンテンツは、近年、コンテンツ表示デバイスや電子化技術の発展とともに、電子的に利用される機会が格段に増加している。しかし電子的に利用するといっても、パソコンや携帯端末の画面で紙面と同じものを閲覧することや、電子テキストに対し参照したい場所を検索するくらいの利用方法が現実である。本研究では、電子化されたコンテンツを最大限利用すべく、コンテンツに対し外部情報源から関連する情報を結びつけ表示するコンテンツ表示システムを開発する。

コンテンツビューア設計理念

テキストを主体とするコンテンツの利用目的としては、もちろん、テキストを読み込み、テキストで表現される情報を理解することである。このとき本研究で開発するコンテンツビューアでは次の 2 つの要求を満たすことを目標とする。

より簡単に

より深く

一般にこの 2 つの要求は相反する概念であるが、本研究では両要求を実現すべく開発を行う。テキストを主体とするコンテンツとしてここでは学術論文と一般書籍を対象とする。それぞれのコンテンツに対し「より簡単に」読むことが要求される状況とは、学術論文では大量の関連論文をサーベイすることに、書籍では購入の際の試し読みや大量の蔵書からキーワード検索では探し出せないページを探し出すことに当てはまる。「より深く」読むこととは、学術論文では、本文で言及される専門用語や専門知識の説明、参考文献や関連文献を提示することに相当し、書籍では、言及される固有名詞の詳細な説明、その固有名詞からうかがえる社会背景を提示することに相当すると考えている。

学術論文と書籍を比較したとき大きく異なる概念として、学術論文はそれ自体では完結せず過去からの研究成果の積み上げとしてお互いにつながりがある一方、書籍では基本的に一冊で閉じた世界を構成している点がある。開発するコンテンツ表示システムではこの点を重視し、学術論文ビューワでは、本文内で言及されている専門用語や数式の説明を提示する基本機能の他に、その専門用語を研究課題とした代表的な論文の提示、同一著者の関連文献など、外の世界とのつながりを読者に意識づけられるような機能を搭載する。書籍ビューワの方では、テキストでは補えない固有概念の画像による補足説明を提示する基本機能に加え、内部構造の把握が容易になるような章別の情報の提示、本文から認識したキーワードをもとにしたカテゴリ別索引ページの自動構成など、1 冊の書籍全体を多様な手段で把握できる機能を搭載することを目指す。

システムのプロトタイプ

平成 23 年度は、以上の設計理念をもとにコンテンツ表示システムのプロトタイプを開発した。図のスクリーンショットは書籍ビューワで、本文中のテキストに対して Wikipedia および世界大百科事典で出現する見出し項目を抽出し、ページ左右にサイドノートとして図画とともに説明を提示したものである。この書籍はテキストのみから構成されているため文章を読解しないと中身が理解できないが、サイドノートで提示されている情報を一目見るだけでページの内容がおおまかに把握でき、また、画像が表示されることで文字でしかなかった情報に視覚的な情報が加わり、より一層理解が進むと考えられる。このような機能を実現することが、「より簡単に」「より深く」理解するといった要求を満たすことにつながると考えている。

図 10：コンテンツ表示システムのスクリーンショット

サブテーマ 4

サブテーマ 1 から 3 においては、テキストを中心とした大規模データの格納・検索・共有・分析・可視化に関する要素技術を研究しているが、サブテーマ 4 では、Researchmap という 22 万人の日本の研究者情報という具体的な大規模データを基盤として、それらの要素技術を検証し、その具体的な課題を

フィードバックする研究サイクルの確立を目指している。同時に、本テーマで構築した研究情報のデータベースと共同研究基盤がそれ自体として、日本の共同研究、特に学際的な融合領域の研究が促進することを目指す。

コンピュータ群に研究情報をより高精度でより効率よく処理させるためには、さまざまな観点から機械が可読であるような情報アーキテクチャを十分に検討する必要がある。たとえば、印刷された論文をスキャンして画像として保存された文書と、タイトルや著者氏名、キーワード等にアノテーションを施し、実験データにはローデータの所在などがリンクとして埋め込まれた形式の文書では、人間の見た目には同じように映る「デジタル化された論文」であっても、その機械可読性には雲泥の差がある。特に、深い検索、深い分析を可能にするには、粒度がそろった正しい情報が大量に蓄積されていることと、機械による処理を可能にするデータベースの設計と厳密な情報の入力が求められる。しかし、その一方で、機械可読性を求めるあまり厳密な情報の入力をユーザに求めれば、入力コストが効用を上回る。生命科学の文献情報を収集したオンラインデータベース Medline では、正確な検索を保証するためにデータ整備に毎年膨大な人件費を支出している。一方、完全な自動化を目指した情報科学分野の文献情報システム CiteSeer では、検索精度が極めて低い。このことからも精度とコスト削減の両立がいかに困難な課題かがわかる。

どのようなプラットフォーム（制度）を導入すれば、機械可読性やユーザの効用の向上と、コストの縮減を同時に実現しうるだろうか。本サブテーマでは、研究資源が発生時点からデジタルであるようなボーンデジタル時代の学術研究情報のエコシステム（循環型情報活用基盤）を今後いかに確立すべきかについて検討を行い、平成 22 年度に Researchmap という基盤ソフトウェアとして実装した。

Researchmap の機能を要約すると以下のようになる。

1 PubMed，Amazon，CiNii, KAKEN など標準的な規約(RSS, ATOM 等）に基づきデータをオープンに公開している学術データプロバイダーから、研究者の研究業績・経歴・競争的資金の獲得状況などをフィードできる。
2 研究者リゾルバー[3] を用いて研究者の名寄せを行い、高い精度で研究業績等のフィードが行える。
3 1 および 2 の機能を用いて、研究者が半自動的に本文所在情報の URL を埋め込んだ研究業績リストおよび Curriculum Vitae(CV)を備えたウェブページを作成し、このページに対して不変の URL を付して、研究者にホームページサービスとして提供する。
4 登録研究者が「自分の業績」として認めた項目情報を CiNii や J-GLOBAL 等のデータプロバイダーにフィードバックすることにより、機械だけでは困難であった研究者情報の名寄せを補完する。
5 登録研究者は CV ページ以外にも、研究ブログを公開したり、研究・教育資料のリポジトリ機能を利用したりすることができる。
6 CV に登録したデータはテキストまたは CSV 形式でダウンロードでき、各種の申請や報告書(大学評価・年報・競争的資金の応募書類・報告書)、調査等に流用できる。
7 Researchmap の CV データは ATOM1.0 に準拠した形式で大学等の研究機関に提供する。
8 登録研究者は学術・研究イベントを登録し、広報することができる。登録されたイベントは、RSS 形式で配信される他、twitter 上でロボットが情報を拡散する。
9 研究プロジェクトを推進するためのバーチャルラボ（コミュニティ）機能を提供する。コミュニティ機能には以下のツールが予め搭載され、利用することができる。
9.1 メーリングリストと連動した掲示板機能

[3] NII が提供している研究者の情報を集約してアクセスを可能にするためのサービス。

9.2　データを共有するためのキャビネット機能および汎用データベース機能
9.3　ToDo を管理するための ToDo モジュール
9.4　予定を調整・管理するためのスケジューラー機能およびカレンダー機能
9.5　オンライン会議のためのチャット機能
9.6　プロジェクト内でアンケートを取るためのアンケート機能および投票機能
9.7　リンクを共有するためのリンクリスト
9.8　写真・画像を整理するためのフォトアルバム機能
9.9　登録研究者間で個人的なメッセージをやりとりするためのプライベートメッセージ機能

10　Researchmap 上で起きている情報をパーソナライズした上で整理・分類して可視化する「新着情報」機能を利用できる。

11　以上の機能のうち CV 以外の機能を携帯電話で閲覧・編集できる。

12　登録研究者を名前・所属・研究分野・研究キーワード・所属学会・地域などから多角的に検索できる。

13　登録研究者の研究内容上の距離を定義し、関連研究者（おとなりの研究者）を計算し、可視化する。

9、10 および 11 の機能は、Researchmap に先だって開発した NetCommons というオープンソース CMS に搭載されている機能である。

以上の機能を研究者に提供することで、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ボーンデジタル時代の学術研究情報のエコシステム（循環型情報活用基盤）を構築する（図 11）。

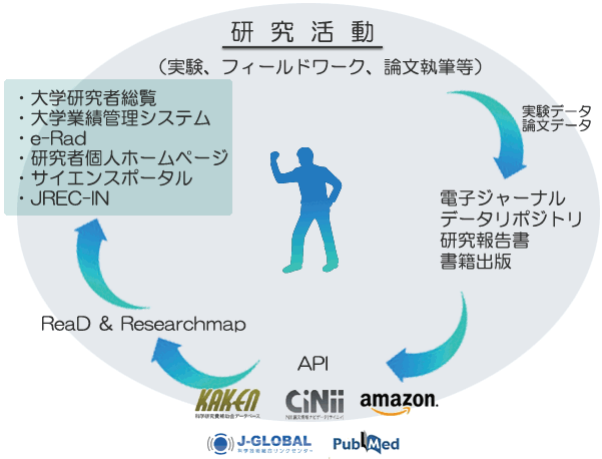

図 11：学術研究情報のエコシステムのイメージ図

4、5 は「機械可読性やユーザの効用の向上と、コストの縮減を同時に実現」するための Researchmap のひとつの具体的提案である。大学が提供する研究者総覧ではデータはひとつずつ手入力しなければならず、入力コストが膨大である。また、手入力による入力ミスやデータの欠損も避けられない。しかし、Pubmed や CiNii 等が提案する相当程度の精度で名寄せされた論文リストの中から自分の論文を選別するだけならば、研究者の入力コストは激減する。ユーザは小さなコストで自らの網羅的な業績リストを手に入れることができる。一方、この「選別」は機械にとって非常に困難な“last one mile”である。この部分を人間しかも著者の支援を受けることで、元のデータベースの名寄せ精度は機械だけでは到達

できない域にまで向上する。つまり、機械と著者はここで win-win の関係を結ぶことができ、精度とコスト削減の両立をはかることができる。

大学が提供するホームページサービスは任期がついている研究者や非常勤講師に提供されることは稀である。仮に提供されていたとしても、各大学の研究者総覧は独自仕様で構築されていであるため、異動の度に初めから入力をし直さなければならない。一方、Researcmap は生涯わたって利用できるサービスなので、異動のたびに書き直す必要がない。研究者の流動化が高まり、多くの研究者が調査疲れ・評価疲れを感じている今、Researchmap はまさに研究者にとってなくてはならないサービスになり得る。

日本の研究者データベースとしては、JST の ReaD がもっともデータ量が多いが、ReaD はデータを登録する研究者にも大学機関にもメリットが少なく、依頼ベースでの登録に過ぎないため、これまでデータの精度に大きな課題があった。また、データの再利用性が極めて低く、研究情報のエコシステムの基盤とすることができなかった。平成 23 年度、JST は既存の ReaD システムに代わり、Researchmap をその基盤として採用した。これにより、Researchmap のデータ量は表 2 のように飛躍的に増大した。

カテゴリ	10 月 8 日 （統合前）	12 月 13 日 （統合後）
学歴	597	489,198
書籍	4,433	715,278
経歴	19,598	530,559
委員歴	420	256,374
競争的資金	10,481	664,798
講演口頭発表等	13,848	1,134,597
研究キーワード	62,045	700,624
MISC	65,429	6,725,707
その他	1,174	5,990
論文	53,297	82,604
特許	69	84,842
受賞	2,779	127,320
研究分野	8,302	263,942
所属学協会	8,119	739,567
担当経験のある科目	3,502	6,204
Works	331	524,268
登録人数	5,464	218,758

ReaD のシステムとして採用されることにより、Researchmap において、外国誌掲載の書誌情報や特許情報を含む J-GLOBAL の情報もフィード可能となり、J-STAGE で無償公開されている日本の学術雑誌の本文データへのリンクも可能となった。また、事業として永続的に運用されることが保証され、データの集積がさらに加速すると予想される。

平成 23 年度の ReaD と Researchmap の統合により、本プロジェクトの要素技術の精度検証に必要な研究情報が集約される見込みが立ったといえる。サブテーマ 3 で平成 23 年度に構築したプロトタイプシステムの実証実験によれば、Researchmap に蓄積されたデータに基づいて行った「学術ニュースの自動分野推定」の精度は約 69%（3 月末時点）であり、まだ大きな課題が残る。精度の向上のためには、機械学習精度の向上のほか、より信頼性の高い情報の集約が必要となる。ReaD に記載されているデータには過去の経緯もあり信頼性にばらつきが大きいため、これを向上させることに取り組むひとつの方

法として、Researchmap データを大学等研究者の所属機関が再利用できるような API を設計し、これをウェブアプリケーションとして実装した。これにより、大学等ではより正確で信頼性の高い情報を入力するインセンティブが高まり、データの信頼性向上につながると考えている。

〔5〕 研究成果物

① 知見・成果物・知的財産権等

1. Researchmap と連携した推薦エンジンの構築
   ユーザ認証機能（Shibboleth 認証）による Researchmap 連携として、Researchmap に登録されている研究者が Researchmap のユーザ ID とパスワードでログインすることによって、その研究者のプロフィールを使った論文推薦を可能にする機能を追加したシステムを構築。
2. 奈良国立博物館で開催された特別展「天竺へ 三蔵法師 3 万キロの旅」にあわせて、国宝「玄奘三蔵絵」（藤田美術館所蔵）の高精細画像を鑑賞できるサービス「国宝 玄奘三蔵絵の世界」を奈良国立博物館と共同で制作し、一般に公開しました。
   会期：2011 年 7 月 16 日～8 月 28 日
   場所：奈良国立博物館

② 成果発表等

＜論文発表＞

〔学術論文〕

1. 梁成基、阿辺川武、「強化学習によるテキスト自動要約手法の提案」、言語処理学会第 18 回年次大会発表論文集、p.1067-1070
2. 研究資源・研究情報のエコサイクルの確立を目指して ReaD と Researchmap の統合がもたらすもの、新井紀子、坂内悟、情報管理、54(9), 533-544, 2011 年 12 月
3. デジタル教科書の諸問題、新井紀子、数学文化、(17), 35-49, 2012 年 3 月

〔データベース〕

1. LODAC 本体（美術館）
   http://lod.ac/
2. LODAC 生物種
   http://lod.ac/species/
3. LODAC 位置情報
   http://lod.ac/location/
4. LODAC 生物学辞書
   http://lod.ac/bdls/
5. TogoDB（コケ）
   http://semantic.togodb.dbcls.jp/togodb/view/antmossdb

〔解説・総説〕

1. 武田英明：Web 時代の識別子と典拠を考える、情報の科学と技術、Vol. 61, No. 11, pp. 441−446 (2011).
2. 武田英明：Linked Data の動向、カレントアウェアネス、No. 308, pp. 8−11 (2011)
3. 武田英明：Linked Data とアイデンティティ、人工知能学会誌、Vol.27, No.2 (印刷中)
4. 高野明彦、「情報の信頼性を支えるもの」、ことばパティオ（Web 掲載）、第 45 回

5. 阿辺川武、「全文テキスト検索技術－サービスの最新動向」、現代の図書館、vol.49、No.2、p.117-124
6. 阿辺川武、「未来の読書が体験できる『e 読書ラボ』」、専門図書館、No.252、p.42-45
7. 畑林一太郎、新井紀子、研究機関における ReaD&Researchmap を利用した研究者総覧の構築について、情報の科学と技術 61(12) 511-515 2011 年 12 月
8. 新井紀子、数学的思考力を高める「珠玉の４問」、日経ビジネス Associe 10(10) 24-29 2011 年 6 月
9. 新井紀子、言語活動の充実と数学的活動（特集 言語活動の充実と指導の改善(1)国語，社会，地理歴史，公民，数学)、中等教育資料 60(6) 34-39 2011 年 6 月
10. 新井紀子、言語としての数学（特集 初学者を悩ます数理の概念－理解のためのヒントを探る)、数理科学 49(5) 11-16 2011 年 5 月
11. 新井紀子、学校の危機管理としての ICT（下)、内外教育（6081) 6-9 2011 年 5 月
12. 新井紀子、学校の危機管理としての ICT（中)、内外教育（6080) 6-9 2011 年 5 月
13. 新井紀子、学校の危機管理としての ICT（上)、内外教育（6079) 6-8 2011 年 5 月

＜会議発表等＞

〔招待講演・国内〕

1. 新井紀子、コンピュータが仕事を奪う　～IT 化の知られざるインパクト、人事部長クラブ　2011 年 4 月　公益財団法人 日本生産性本部
2. 新井紀子、大震災を踏まえて―災害に強い学校 IT 環境づくり、ＩＴ活用セミナー　2011 年 5 月　家庭教育新聞
3. 新井紀子、次世代へのメッセージ　科学者に問われるもの、ノーベル賞フォーラム　2011 年 10 月　読売新聞社
4. 新井紀子、「超チャレンジング研究をどう進めるか」～サイエンス 2.0 基盤の構築～
「生命をはかる」研究会　第 37 回公開研究会　2011 年 10 月
5. 新井紀子、ICT の活用による地域と連携した安全・安心な学校づくり、全国生涯学習ネットワークフォーラム　2011 年 11 月　全国生涯学習ネットワークフォーラム 2011 実行委員会

〔一般講演・国際〕

1. Takafumi Suzuki, Shin Hasegawa, Takayuki Hamamoto and Akiko Aizawa, Document recommendation using data compression, Procedia - Social and Behavioral Sciences, Volume 27, Pages 150–159 , 2011 年
2. Yoshimasa Tsuruoka, Yusuke Miyao, Jun'ichi Kazama, Learning with Lookahead: Can History-Based Models Rival Globally Optimized Models?, Proceedings of CoNLL 2011, 2011 年 6 月
3. Shun'ya Iwasawa, Hiroki Hanaoka, Takuya Matsuzaki, Yusuke Miyao, Jun'ichi Tsujii, A Collaborative Annotation between Human Annotators and a Statistical Parser, Proceedings of the Linguistic Annotation Workshop, 2011 年 6 月
4. Tadayoshi Hara, Yuka Tateisi, Jin-Dong Kim and Yusuke Miyao, Parsing Natural Language Queries for Life Science Knowledge, The 10th Workshop on Biomedical Natural Language Processing (BioNLP2011), 2011 年 6 月
5. Tadayoshi Hara, Takuya Matsuzaki, Yusuke Miyao and Jun'ichi Tsujii, Exploring Difficulties in Parsing Imperatives and Questions, The 5th International Joint Conference on Natural Language Processing (IJCNLP2011), 2011 年 11 月
6. Kiyoko Uchiyama, A Study for Identifying Domain-specific Introductory Terms in Research

Papers,The 9th International conference on Terminology and Artificial Intelligence (TIA2011), pp.147-150, 2011 年 11 月

7. Atsuhiro Takasu, A Multicriteria Recommendation Method from Data with Missing Rating Scores, International Conference on Data and Knowledge Engineering (ICDKE 2011) , pp.60-67, 2011 年 9 月
8. Atsuhiro Takasu, Saranya Maneeroj, A Recommendation Algorithm Using Positive and Negative Latent Models, IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011) , pp.72-79, 2011 年 12 月
9. 武田英明：共有・再利用のための知のデジタル・アーカイブ、国際シンポジウム：デジタル化時代における知識基盤の構築と人文学の役割―デジタル・ヒューマニティーズを手がかりとして―東京大学文学部次世代人文学開発センター/ 東京大学大学院情報学環メディア・コンテンツ総合研究機構/ 科研基盤 A「国際連携による仏教学術知識基盤の形成―次世代人文学のモデル構築」(2011)
10. 加藤文彦：LODAC Museum -A platform for linking museum data-, Culture & Creative Industry EXPO, Taipei, Taiwan

〔一般講演・国内〕

1. 森信介、Daniel Flannery、宮尾祐介、Graham Neubig、部分的アノテーションから学習可能な係り受け解析器、情報処理学会自然言語処理研究会、2011 年 5 月
2. 内山清子、高須淳宏、相澤彰子、難波英嗣、宮尾祐介、オススメ論文検索システム：OSUSUME、人工知能学会全国大会、pp.1-4 、2011 年 6 月
3. 亀田尭宙、内山清子、宮尾祐介、武田英明、相澤彰子、論文中の引用文における構文パターンを用いた論文・概念間の関係抽出、pp.25-31、第 94 回人工知能学会知識ベースシステム研究会、2011 年 12 月.
4. 亀田尭宙、武田英明、相澤彰子、関連研究に関する記述の分析による論文間の意味的関係の抽出、人工知能学会全国大会、2011 年 6 月
5. 原忠義、建石由佳、Jin-Dong Kim、宮尾祐介、生命科学知識を得るために入力される自然言語クエリの構文解析、NLP 若手の会(YANS) 第 6 回シンポジウム、2011 年 9 月
6. 亀田尭宙、相澤彰子、論文のスクリーンショットを用いた文献紹介スライド作成補助の試み、第 12 回 AI 若手の集い（MYCOM2011）、2011 年
7. 福田悟志、難波英嗣、竹澤寿幸、武田英明、相澤彰子、大向一輝、宮尾祐介、内山清子、CiNii データベースを用いた研究動向分析システムの構築、言語処理学会第 18 回年次大会予稿集、2012 年 3 月
8. 松林優一郎、宮尾祐介、相澤彰子、語彙概念構造を用いた日本語述語項構造コーパスの設計、言語処理学会第 18 回年次大会予稿集、2012 年 3 月
9. T. Kamura, H. Takeda, I. Ohmukai, F. Kato, T. Takahashi and H. Ueda: Study Support and Integration of Cultural Information Resources with Linked Data, in Proceedings of the Second International Conference on Culture and Computing, pp. 177–178, Kyoto, Japan (2011). 2011 年 10 月 20-22 日
10. 武田英明、嘉村哲郎、加藤文彦、大向一輝、高橋徹、上田洋：日本における Linked Data の普及にむけて、人工知能学会全国大会(第 25 回)論文集、No. 3E3-OS20-9、盛岡 (2011)、人工知能学会. 2011 年 6 月 1-3 日
11. 南佳孝、加藤文彦、大向一輝、武田英明、新井紀子、神保宇嗣、伊藤元己、小林悟志：生物情報基盤構築に向けた生物関連データの Linked Data 化の取り組み、第 25 回セマンティックウェブとオントロジー研究会、No. SIG-SWO-A1103-02 人工知能学会 (2011). 2011 年 10 月 11 日

12. 小出誠二、武田英明、大向一輝：WordNet 日本語化への LOD アプローチ、第 26 回セマンティックウェブとオントロジー研究会、No. SIG-SWO-A1103-05 人工知能学会（2011). 2011 年 12 月 14 日
13. 松村冬子、小林巌生、嘉村哲郎、加藤文彦、高橋徹、上田洋、大向一輝、武田英明：Linked Open Data による博物館情報および地域情報の連携活用、じんもんこん 2011 論文集、第 2011 巻、pp. 403−408 情報処理学会（2011). 2011 年 12 月 10-11 日
14. 松村冬子、小林巌生、嘉村哲郎、加藤文彦、高橋徹、上田洋、大向一輝、武田英明：Linked Open Data による博物館情報および地域情報の連携活用、じんもんこん 2011 論文集、第 2011 巻、pp. 403−408 情報処理学会（2011). 2011 年 12 月 10-11 日
15. 加藤文彦：LODAC Project, セマンティック Web コンファレンス 2012、慶応大学、2012 年 3 月 8 日
16. 加藤文彦、南佳孝、神保宇嗣、川本祥子、武田英明：生物学辞書の Linked Open Data 化とその応用、第 48 回人工知能学会分子生物情報研究会（SIG-MBI)、2012 年 3 月 23-24 日
17. 丸川雄三、「研究情報統合サービスについて」、ネアンデルタールとサピエンス交替劇の真相・2011 年度春期大会（第 3 回全体会議)、学術総合センター、2011 年 4 月 23 日
18. 丸川雄三、「文化遺産オンラインにおける作品情報の連携基盤についてー奈良国立博物館との連携事例を中心にー」、アート・ドキュメンテーション学会 2011 年度年次大会、東京国立博物館平成館大講堂、2011 年 6 月 11 日
19. 高野明彦、「文化情報の整備と活用」、シンポジウム「文化情報の整備と活用～デジダル文化財が果たす役割と未来像」、丸の内・コンフェレンススクエアエムプラス、2011 年 7 月 22 日
20. 中村佳史、「複合情報拠点としての図書館をめざして」、長野県図書館協会大学専門図書館部会図書館研究会、清泉女学院大学清泉女学院短期大学、2011 年 8 月 30 日
21. 中村佳史、「本と街の案内所」、専門図書館協議会関東地区協議会第 17 回情報サービス研究会、国立情報学研究所、2011 年 11 月 1 日
22. 間下亜紀子、「読書中における着目表現の調査」、第 59 回日本図書館情報学会研究大会、日本大学文理学部、2011 年 11 月 13 日
23. 高野明彦、「文化遺産オンラインとデジタル・アーカイブについて（基調講演)」、「文化遺産オンライン構想」成果報告フォーラム、一橋記念講堂、2011 年 12 月 2 日
24. 丸川雄三、「文化遺産オンラインの新機能」、「文化遺産オンライン構想」成果報告フォーラム、一橋記念講堂、2011 年 12 月 2 日
25. 梁成基、「強化学習によるテキスト自動要約手法の提案」、言語処理学会第 18 回年次大会、広島市立大学、2012 年 3 月 16 日
26. 高野明彦、特別講演「いま改めて書店について考えるー本屋の機能を問い直す」、東京国際ブックフェア 2011、東京ビッグサイト、2011 年 7 月 9 日
27. 新井紀子、「震災から学ぶ～学校の危機管理としての ICT」、Netcommons ユーザカンファレンス、一橋記念講堂、8 月 18 日

＜受　賞＞

1. 松村冬子、Linked Open Data チャレンジ Japan 2011 goo 賞（2012 年 3 月 8 日）
2. 高野明彦、丸川雄三平成 23 年度　文部科学大臣表彰科学技術賞（理解増進部門）（文部科学省）「連想情報技術による自発的学びのための情報理解増進」
3. 灘本明代、荒牧英治、阿辺川武、村上陽平、Emerald Literati Network 2011 Awards for Excellence 選考　最優秀論文
   「Extracting content holes by comparing community-type content with Wikipedia」

4. 国立情報学研究所、第 5 回（2011）JEPA 電子出版アワード　ベンチャー・マインド賞（一般社団法人日本電子出版協会）「e 読書ラボ」

③　その他の成果発表

＜新聞報道など＞

1. 高野明彦、「キーマンインタビュー　連想検索エンジンの開発者が提案する電子出版の新たな地平」、On Deck、2011.5.12
2. 「高野明彦、知的体験をさらに豊かにする文化施設空間をめざして」、AMeeT - Art Meets Technology 京都から世界へ、2011.5.25
3. 高野明彦、「東京国際ブックフェア…活字は安らぎ・希望　『本の力』見直された震災後」、YOMIURI ONLINE、2011.7.12
4. 高野明彦、「街の書店　個性が生命線　国際ブックフェア　生き残り策議論」、朝日新聞夕刊 3 面、2011.7.12
5. 高野明彦、「東京大学大学院情報理工学系研究科・国立情報学研究所　高野明彦教授インタビュー」、東新進学情報　vol.150、2011.7.15
6. 高野明彦、「高精細画像で国宝絵巻再現　奈良国立博物館の特別展」、科学新聞　2 面、2011.7.29
7. 高野明彦、「いま改めて書店について考える　本屋の機能を問い直す　東京国際ブックフェア 2011 特別講演から」、文化通信 BB 増刊　1-2 面、2011.8.1
8. 高野明彦、「奈良博、NII　『国宝　玄奘三蔵絵の世界』を一般公開　高精細画像で絵巻を鑑賞」、文部科学教育通信　no.273、2011.8.8
9. 高野明彦、「国立情報学研究所（NII）、『e 読書ラボ』を正式オープン」、NDL Current Awareness Portal、2011.9.30
10. 高野明彦、「未来の読書体験スペースが神保町に開館、国立情報学研究所らが企画・運営」、INTERNET Watch、2011.9.30
11. 高野明彦、「IT・サイエンスの R25 コラム　噂のネット事件簿　10 種類の電子書籍端末や紙の本が読み比べられるラボ」、web R25、2011.10.5
12. 高野明彦、「ユニバーサロンリポート　《電子書籍》　誰でもじっくり触って選べる『e 読書ラボ』　東京神田神保町にオープン」、毎日 jp、2011.10.14
13. 高野明彦、「NII　未来の読書を考える　e 読書ラボで電子と紙を読み比べ」、電経新聞　1 面、2011.10.17
14. 高野明彦、「今月の本トピ　『読書ラボ』が本格始動　電子書籍の端末あれこれを実体験してみよう」、散歩の達人　vol.16, no.11 (no.188)、2011.10.21
15. 高野明彦、「【연재기획】 책의 거리를 탐방하다(하)」、KYODO NEWS、2011.10.27
16. 高野明彦、「新しい読書体験を模索する『e 読書ラボ』」、マガジン航[kɔ：]、2011.10.31
17. 高野明彦、「おやこで読書の秋　AKB メンバー、神田を探検　本好きには天国のような街」、msn 産経ニュース、2011.11.1
18. 高野明彦、「これが未来の読書だ　体験施設オープン　東京　電子書籍端末 11 種　紙の本との比較も」、朝日小学生新聞　3 面、2011.11.2
19. 高野明彦、「らいふプラス　読書スタイル新たな“1 ページ”　使い方指南　紙と見比べも」、日本経済新聞夕刊　4 面、2011.11.12
20. 高野明彦、「Book Review　永江朗の出版業界事情　神保町に電子書籍端末の展示場がお目見え」、週刊エコノミスト　vol.89, no.52 (no.4201)、2011.11.15
21. 高野明彦、「be report 『「出会い』を大事にする古本屋　案内所で敷居を低く」、朝日新聞　be on

Saturday　be4 面、2011.12.3
22. 高野明彦、「文化遺産オンライン　10 万件目前」、読売新聞朝刊　15 面、2011.12.7
23. 高野明彦、「ユーザー視点からみた“電子書籍のいま”　神保町『読書ラボ』訪問記」、プリバリ印　vol.34、2011.12.12
24. 高野明彦、「キッズ通信　電子書籍と紙 読み比べよう『e 読書ラボ』東京・神田神保町の『本と街の案内所』」、読売新聞夕刊　8 面、2011.12.24
25. 高野明彦、「想・記・伝 3　頼れる記録の器求めて つきまとう『はかなさ』」、朝日新聞朝刊　17 面、2012.1.4
26. 高野明彦、「文化の扉　はじめての青空文庫　タブレット広まり利用者急増」、朝日新聞朝刊　31 面、2012.1.23
27. 新井紀子、「[新刊 GUIDE]コンピュータが仕事を奪う」 日経サイエンス、4 月号 P.107、2011/4/1
28. 新井紀子、「筑波大ベンチャーエデュケーションデザインラボ全国の教育機関にクラウド無償提供災害への強さ注目」、茨城新聞、2011/5/13
29. 新井紀子、「災害時に有効な CMS の活用」 内外教育時事通信社、P6～P8、2011/5/17
30. 新井紀子、災害時に機能する学校 IT 環境 7 面、教育家庭新聞、2011/7/4
31. 新井紀子、インターネット、ウェブホスティング、「Joe's ウェブホスティング、NetCommons が無償ですぐに使える「NetCommons 標準サーバー」を提供開始」、2011/8/17
32. 新井紀子、「埼玉版災害時に強い Netcommons」読売新聞、29 面、2011/10/5、新井紀子、学校コンピュータ、2011/10/1
33. 新井紀子、日経リナックス Netcommons で情報共有、P38～P45、2011/10/1
34. 新井紀子、「国立情報学研究所『ドラえもん』計画人工知能東大合格目指せ、読売新聞夕刊 13 面、2011/11/5
35. 新井紀子、「ノーベル賞受賞者を囲むフォーラム『次世代へのメッセージ　科学者に問われるもの』」、読売新聞朝刊 21－21 面、2011/11/12
36. 新井紀子、「NII キックオフシンポ『人工頭脳プロジェクト』」、文教ニュース第 2163 号 P.5、2011/11/14
37. 新井紀子、「東大合格めざす人工頭脳プロジェクト」、科学新聞、2011/11/18
38. 新井紀子、「科学朝日」放送、テレビ出演、朝日ニュースター、2011/11/24
39. 新井紀子、「情報研『人工頭脳プロジェクト』キックオフシンポジウム」、文教速報第 7659 号 p.23、2011/11/25
40. 新井紀子、TBS ラジオ“GAKUSHOCK”、ロザン、2011/12/4
41. 新井紀子、「みんなのライバル『東ロボくん』」、朝日小学生新聞 1 面、2011/12/6
42. 新井紀子、「情報研『人工頭脳プロジェクト』」 12 月 5 日文教ニュース「キックオフシンポジウム、第 2166 号、2011/12/14
43. 新井紀子、「ロボットは東大に入れるか」、WEB ニコニコニュース、2011/12/15
44. 新井紀子、「ロボットは東大に入れるか」、カラパイヤニュース、2011/12/15
45. 新井紀子、「ニッキィの大疑問＝コンピューター、脳に迫る？　主観・ひらめき…まだ人間の領域」日本経済新聞、夕刊 7 面、2012/2/20