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）

〔1〕 研究プロジェクト

(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 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 26 年度

サブテーマ１では、論文に書かれた知識の探索を支援する基盤技術の確立を目指して、引き続き情報推薦・閲覧システムの開発に取り組む。具体的にはまず、実際に流通する論文の書式に対して言語の意味解析を適用するための構造解析手法の研究に取り組み、特に XML 文書の言語解析の性能を改善するための解析ツールを開発・公開する。また、前年度に開発した PDF 論文の閲覧支援システム SideNoter について、言語を横断して関連論文を推薦する機能を改善するとともに、抽出した参考文献に対する自動同定処理を適用したデモシステムを構築して、有効性を実証的に検証する。さらに、推薦の単位を論文からセクション、さらには「文」まで拡張し、論文の閲覧・執筆時に文脈にあわせて関連情報を提示する手法の検討に新たに取り組む。

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、科学技術文書の構造・意味解析に基づく高度な情報推薦基盤の実現を目指して、引き続き要素技術の開発に取り組む。特に、国内の学術コンテンツを国際的な学術コンテンツに結びつけるための言語横断的な機能の強化を目指す。また、科学技術文書中の専門的な概念とその関係を抽出して既存のオントロジーに対応づけるためのリンキングサーバを立ち上げて、大量文献の内容解析に基づく既存オントロジーの補完・拡充の方策を探る。人材育成では、関連研究者との会合を定期的に開催し、学術研究基盤の強化に取り組むコミュニティの活性化に取り組む。

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

〔情報・システム研究機構〕　原　忠義、Hubert Soyer
〔国立情報学研究所〕　高須淳宏、宮尾祐介
〔湘南工科大学〕　内山清子
〔広島市立大学〕　難波英嗣
〔九州大学〕　冨浦洋一、石田栄美

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
〔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
〔情報・システム研究機構〕　舛川竜治
〔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では、論文に書かれた知識の探索を支援する基盤技術の確立を目指して、H25 年度に引き続き情報推薦・閲覧システムの研究およびシステム構築を進めた。特に本年度は、現在の電子図書館が提供する「論文を単位とする検索」機能を「論文に書かれている情報の推薦や探索」機能へと進化させるための解析・ナビゲーション手法を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、ツール構築や実証的なデモシステム開発にも取り組んだ。具体的には以下の研究課題を実施した。

①　実文書解析

実際に流通する論文の書式に対して言語の意味解析を適用するための構造解析手法の研究に取り組み、平成 25 年度に提案した XML 文書の解析技術をツールとして実装して公開した。

②　専門用語リンキング

科学技術文書中の専門用語を抽出して、外部知識ベース等の既存のオントロジーに対応づけるための専門用語リンキング手法の研究に取り組み、特に言語横断的なリンキングについて評価用データを構築して誤り分析を行った。分析結果に基づき H27 年度に継続して性能改善に取り組む予定である。

③　科学技術文書の閲覧支援

前年度に開発した PDF 論文の閲覧支援システム SideNoter について、言語を横断して関連論文を推薦する機能を改善するとともに、抽出した参考文献に対する自動同定処理を適用したデモシステムを構築・公開した。

④　文脈に応じた詳細な情報推薦のための類似文検索

本年度の新たな取り組みとして、日本語（英語）の句や文を入力として、意味的に類似する英語（日本語）の例文を提示するための手法を検討し、深層学習に基づく新たな方式を提案した。その成果は国際会議等に採択されており、H27 年度では他の研究成果とあわせて論文理解・執筆支援のための API を実装する予定である。

上記 4 つの研究課題の詳細について以下にまとめる。

サブテーマ 1　研究課題①　実文書解析のための文書解析手法の開発とツール公開（学術論文 5、6）

構造化されたテキストを含む実世界の文書を自然言語処理（NLP）ツールで解析することは、情報検索や情報推薦の高度化、専門知識オントロジーの構築等に必須の情報処理技術である。しかしながら現実の文書には、印刷のためのレイアウト情報やメタ情報などが混然一体となって埋め込まれており、言語テキスト抽出の難しさが、大量文書を扱う際の効率や解析性能を損ねる結果となっている。そこで H26 年度ではこれまでの研究成果を踏まえて PlaneText と呼ぶテキスト抽出ツールを開発して公開した（図 1 文書解析・変換ツール PlaneText、http://kmcs.nii.ac.jp/planetext/）。

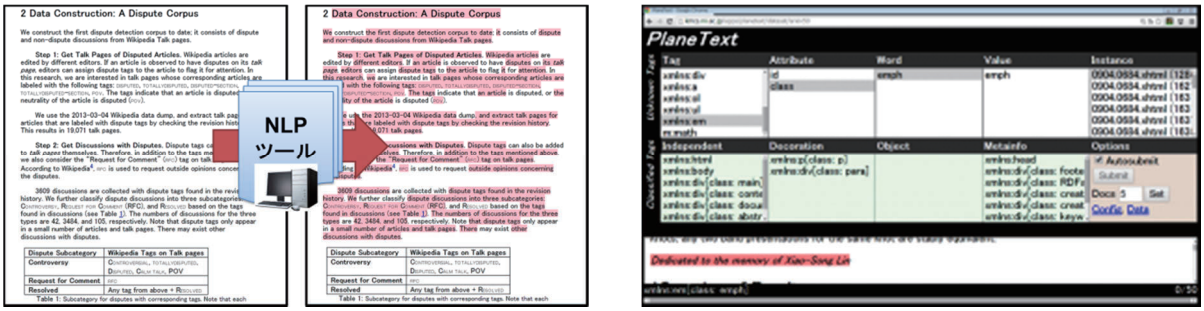

図 1　文書解析・変換ツール PlaneText

PlaneText の基本的なコンセプトはタグ分類に基づく XML 文書の平文テキスト化である。すなわち、PlaneText を用いることで、XML タグで囲まれたテキストを NLP ツールに直接入力可能な文に変換することができる。その処理の流れを以下に示す（図 2 PlaneText における処理の流れ、http://kmcs.nii.ac.jp/planetext/）。

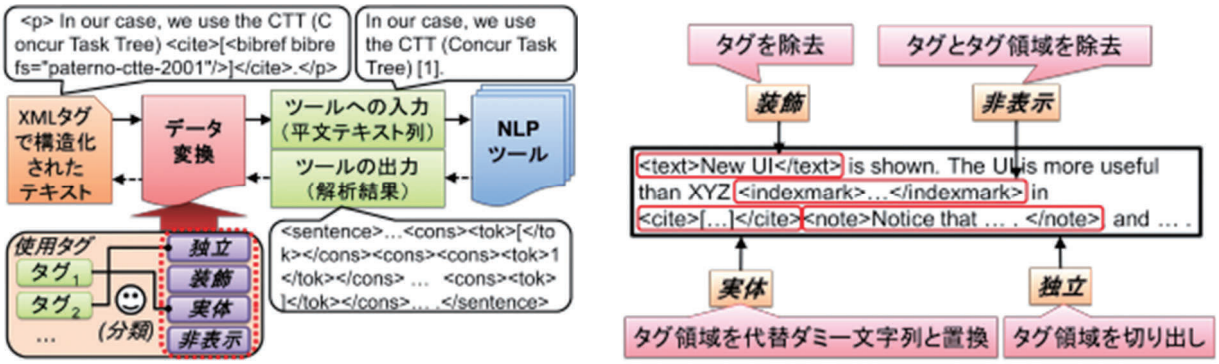

図 2　PlaneText における処理の流れ

現在のバージョンでは、XML のタグをその機能に基づいて「独立」「装飾」「実体」「非表示」の 4 種類に分類し、これに基づき文書の自動変換を行う仕組みを実装している。

・ 独立（Independent）： タグで囲まれた部分を、周囲とは区切られた文（文章）として見なす
・ 装飾（Decoration）： タグで囲まれた部分の表示スタイルを変化させる
・ 実体（Object）： 自然言語ではない何らかの構造を文章中に導入する
・ 非表示（Meta-info）： 実際にはディスプレイには表示されない、テキストに関する情報を挿入する

PlaneText では、上記の 4 種の分類にしたがって、「独立タグで囲まれたテキスト部分毎に切り出し」、「装飾タグを除去し」、「実体タグで囲まれた領域を代替のダミー文字列で置換し」、「非表示タグで囲まれた領域を除去する」の操作を行うことで、構造化された文書を、NLP のツールで処理せずに解析できるような平文のテキスト列へと変換する。ここで、ユーザが効率的にタグ分類を行えるよう、コマンドラインおよび Web ブラウザを介した GUI ツールインタフェースを提供している。たとえば実験に用いた複数の文書群では、文書中の使用タグの 5 分の 1 以下のタグを分類するのみで、NLP ツールで解析できるテキスト列への変換ができることを確認している。文書を発行・管理している組織や出版社ごとにタグ仕様が決まっている場合には、保存した設定ファイルを読み込むことで、簡単に一括処理が実現できる。

PlaneText は単なる前処理ツールではなく、構文解析などの重要な言語解析の性能改善に大きく寄与するものである。その有効性を評価するため、PubMedCentral、arXiv.org、英語 Wikipedia 記事、言語処理学会論文誌（和英）の 5 つの文書集合（表 1 実験で用いた文書集合とタグ分類数）に対して PlaneText を適用し、構文解析の誤り数を比較する実験を行った。各文書集合の概要を表 1 に示す。英語文書に対しては、深い統語・意味解析を行うパーザである Enju パーザ*1（文区切りのため GeniaSS*2 を適用）、および句構造／依存関係解析を行う Stanford パーザ*3 の 2 つを用いた。日本語文書に対しては、代表的な形態素解析ツールである JUMAN*4 および MeCab*5 の 2 種類を用いた。

*1 Enju (Ninomiya et al., 2007)　*2 http://www.nactem.ac.uk/y-matsu/geniass/　*3 de Marneffe et al., 2006
*4 http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN　*5 http://mecab.sourceforge.jp

表 1 実験で用いた文書集合とタグ分類数

文書群	ジャンル	言語	スタイル	使用文書数
PubMed Central (PMC)*1	科学論文	英	XML	1,000
arXiv.org*2	科学論文	英	XHTML	300
Wikipedia*3 エントリ	ウェブ	英	HTML†	300
言語処理学会論文誌「自然言語処理」*4	科学論文	日	XHTML††	384(全)
	科学論文	英	XHTML††	68(全)

(† XML ファイルから生成されたもの)
(†† LaTeXML*5 を用い XML から変換したものが公開中)

*1 http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/　*2 http://arxiv.org/
*3 http://www.wikipedia.org/　*4 http://nlp20.nii.ac.jp/resources/
*5 http://dlmf.nist.gov/LaTeXML/

文書(数)	全タグ数(異なり数)	分類されたタグの数(異なり数)					獲得列数
		独立	装飾	実体	非表示	合計	
PMC (1,000)	**1,357 k (421)**	32k (12)	62k (9)	48k (9)	34k (56)	**177k (85)**	**26 k**
arXiv (300)	**1,969 k (210)**	6k (15*)	47k (12*)	60k (8*)	8k (17*)	**121k (52*)**	**4 k**
Wiki. (300)	**224 k (60)**	3k (12*)	11 k (8*)	1k (28*)	11k (67*)	**28k (115*)**	**2 k**
JNL-E (68)	**142 k (57)**	8k (25*)	12 k (16*)	6k (9)	23k (19)	**29k (69*)**	**6 k**
JNL-J (384)	**699 k (58)**	50k (23*)	56 k (18*)	32k (10*)	14k (21*)	**153k (72*)**	**38 k**

比較手法としては、(1)タグを単純除去した場合（「単純除去」）、(2)実体・非表示タグを提案枠組で処理し、装飾・独立タグについては単純除去を行った場合（「実/非のみ」）、(3)提案枠組で全タグを処理した場合（「提案枠組み」）の 3 通りを用いた。各方法について、検出される文の数、解析時間、解析失敗文の数(%)を比較したものを以下に示す（表 3 実験結果（英語構文解析の性能に対する影響）、表 2 実験結果（日本語形態素解析の性能に対する影響））。このように、対象文書中の 20%以下のタグを分類することで平文テキスト列を獲得することが可能で、従来のタグの単純除去に比べ、NLP ツールの大幅な解析性能改善（高被覆・効率化）が実現されることを確認した。このことは、実文書を NLP と適切に繋ぐ技術の重要性を示すものである。

表 2　実験結果（日本語形態素解析の性能に対する影響）

タグの扱い方が JUMAN / MeCab の性能に与える影響

JUMAN

文書（数）	手法	検出文数	解析時間（秒）	失敗文数（%）
JNL-J (384)	単純除去	96,668	122	10 (0.01)
	実/非のみ	76,277	86	8 (0.01)
	提案枠組	114,250	59	2 (0.00)

MeCab

文書（数）	手法	検出文数	解析時間（秒）	失敗文数（%）
JNL-J (384)	単純除去	97,312	7	10 (0.01)
	実/非のみ	78,461	6	8 (0.01)
	提案枠組	116,424	6	2 (0.00)

表 3　実験結果（英語構文解析の性能に対する影響）

タグの扱い方が Enju パーザ の性能に与える影響

文書（数）	手法	検出文数	解析時間（秒）	失敗文数（%）
PMC (1,000)	単純除去	159,327	209,783	4,721 (2.96)
	実/非のみ	112,285	135,752	810 (0.72)
	提案枠組	126,215	132,250	699 (0.55)
arXiv (300)	単純除去	74,762	108,831	2,047 (2.74)
	実/非のみ	41,265	89,200	411 (1.00)
	提案枠組	43,208	87,952	348 (0.81)
Wiki. (300)	単純除去	10,561	14,704	1,161(10.99)
	実/非のみ	5,026	6,743	67 (1.33)
	提案枠組	6,893	6,058	61 (0.88)
JNL-E (68)	単純除去	23,196	24,881	271 (1.17)
	実/非のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

タグの扱い方が Standord パーザ の性能に与える影響

文書（数）	手法	検出文数	解析時間（秒）	失敗文数（%）
PMC (1,000)	単純除去	170,999	58,865	18,621(10.89)
	実/非のみ	126,176	50,741	11,881 (9.42)
	提案枠組	139,805	63,295	11,338 (8.11)
arXiv (300)	単純除去	75,672	27,970	10,590(13.99)
	実/非のみ	48,666	24,630	5,457(11.21)
	提案枠組	50,504	26,360	5,345(10.58)
Wiki. (300)	単純除去	14,883	3,114	1,651(11.09)
	実/非のみ	6,173	2,248	282 (4.57)
	提案枠組	8,049	2,451	258 (3.21)
JNL-E (68)	単純除去	24,942	9,069	1,577 (6.32)
	実/非のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10,154	892 (4.48)

サブテーマ１　研究課題②　専門用語リンキング（学術論文 1、2、3）

知識獲得では、テキストの構成要素を外部知識ベースに対応づける「エンティティ・リンキング」操作が必須の技術である。ところが既存のエンティティ・リンキングは、組織名や人名などの固有表現を対象とするものが大半で、専門用語のような抽象的な概念を対象とする研究はほとんど行われていなかった。そこで本研究では、テキストから①「専門用語」を抽出して、②機械翻訳手法で「翻訳」し、③知識ベースに対応づけた上で、④関連情報を推薦する、一連の機能を提供する辞書サーバを試験的に実装し（図 3 用語リンキングサーバによる用語抽出と Wikipedia 記事への対応付け）、後述する PDF 文書閲覧支援システムに適用して、言語横断論文推薦における有効性を示した。

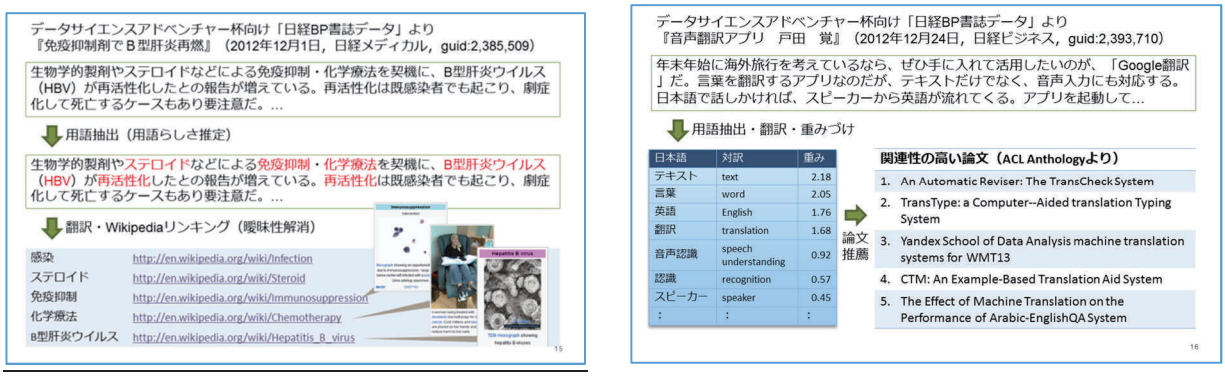

図 3　用語リンキングサーバによる用語抽出と Wikipedia 記事への対応付け

サブテーマ１　研究課題③　科学技術文書の閲覧支援（学術論文 4、7）

近年、学術文献の電子化は大きく進み、投稿から出版まで紙を経由することなく流通することが一般的になってきた。多くの学術出版では、PDF ファイルを用いて電子出版を行っている。そこで本研究では、阿辺川らによる PDF 文書閲覧システムに注目し、論文閲覧時に、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 Sidenoter の開発および実証に取り組んでいる。SideNoter は論文 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である。H25 年度ではその成果として、2014 年 3 月に開催された自然言語処理学会の年次大会において、過去 20 年分の全国大会の予稿集を SideNoter 上で閲覧可能にして参加者に対して提供した。この経験に基づき本年度はさらに SideNoter に改良を加え、2015 年 3 月に開催された言語処理学会年次大会において再度デモを行った。SideNoter は現在、ウェブ上で一般公開をしており、http://kmcs.nii.ac.jp/nlp_annual/でアクセス可能である。

SideNoter 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。現在、本システムで対応する論文の形式は、紙に印刷された論文および PDF で流通する論文である。紙の論文に対しては、スキャナーを使用してデジタル画像データに変換後、OCR ソフトウェアを用いてテキストを認識し、透明テキスト付き PDF に変換する。一方、LATEX や Microsoft Word から作成された PDF 形式の論文は大半が PDF 内にテキスト情報が保存されているが、一部の PDF ではフォントのグリフ情報と既存の文字コードの対応表が存在しないものもある。そのような PDF では正しくテキスト情報が抽出できないため、一度画像データに変換した後、OCR を施しテキスト情報を得る。まず画像変換のための処理フローを以下の図に示す（図 4 SideNoter におけるデータ処理フロー）。

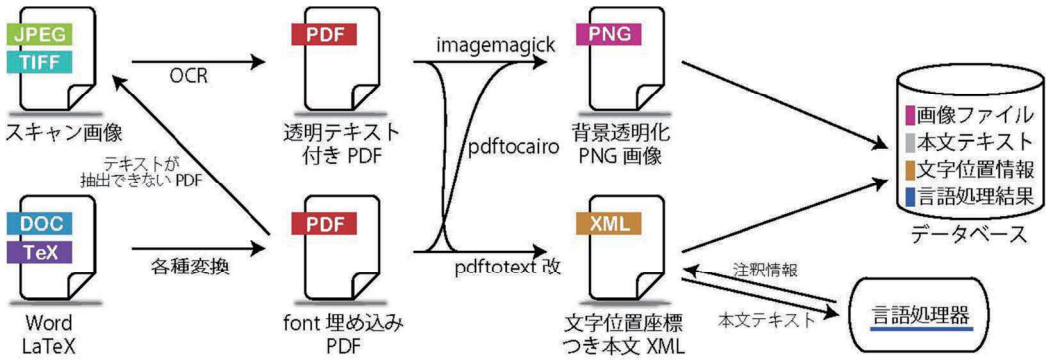

図 4　SideNoter におけるデータ処理フロー

PDF からテキストを抽出するにあたり、本システムで要求する機能を実現するためには、日本語のような分かち書きのない言語の文字は文字単位で、分かち書きのある言語では単語単位でページ内座標を得る必要がある。しかし、オープンソースソフトウェアライセンスで使用できるプログラムを各種検討したが、条件を満たす使い勝手のよいツールが見つからなかったため、Poppler パッケージ（http://poppler.freedesktop.org/）に含まれる pdftotext に独自にパッチをあて対応している。本システムの本文表示で使用している画像形式は、本文の背景色を変更する機能に対応するため、PNG 形式を採用している。PDF から PNG への画像変換には、フォント情報を持つ PDF からダイレクトに透過 PNG ファイルを作成できる Poppler パッケージ中の pdftocairo を使用している。スキャン画像から作成された PDF については、背景色（白色）を透明色に指定した PNG に変換する。通信トラフィックを減少させるため、Web ブラウザの表示ウィンドウの大きさに合わせて解像度を変えた画像を用意し、白黒のページについては 6 色に、カラーのページでは 64 色に減色した画像を表示に用いている。

SideNoter のスクリーンショットを以下に示す（図 5 SideNoter によるデモシステムのスクリーンショット）。システムは PC もしくはモバイルデバイスの Web ブラウザ上で動作し、マウス、キーボード、タッチ機能で操作する。論文本文は、画像変換から得られた画像そのものを画面中央に表示する。ハイライトや特定の領域を指し示す線分を、画像の上にオーバーレイする形で配置できる。

図 5　SideNoter によるデモシステムのスクリーンショット

SideNoter における本文画像の両サイドには、論文読解を支援する各種リソースを掲載するスペースがある。利用者の閲覧を支援するため、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）に自動表示している。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。具体的には、日本語 Wikipedia や YouTube 上の情報が自動的に取り込まれ、画像や要約文とともに提示されるようになっている。
- 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。検索アルゴリズムには GETA を用いている。具体的には、言語処理学会年次大会・論文集、および ACL Anthology に登録された国際会議論文の内容等が提示されるようになっている。

SideNoter による各論文のエントランス画面を以下に示す（図 6 SiceNoter のエントランスページと関連論文推薦機能）。SideNoter に特徴的な機能として、PDF 文書から論文のセクション構造を自動抽出してサムネイル画像とともに提示し、さらにセクションごとに関連論文の推薦が行えるようになっている。日本語論文については、推薦対象文書もセクション単位で分割しているため、指定した論文に対して「実験セクションが類似する論文」など、きめの細かな論文推薦が行えるようになっている。

図 6　SiceNoter のエントランスページと関連論文推薦機能

サブテーマ１　研究課題④　文脈に応じた詳細な情報検索のための類似文検索（学術論文 8、9）

大量の論文を計算機で処理して、効率の良い知識探索・知識獲得を実現するためには、をさらに細かく分析し、知識を獲得するためには、文単位での文書を横断する情報推薦が必要になる。そこで本年度の新たな取り組みとして、日本語（英語）の句や文を入力として、意味的に類似する英語（日本語）の例文を提示するための手法を検討し、深層学習に基づく新たな手法を提案した。具体的には、Inclusion criteria と呼ぶ評価基準を用いた新しい意味獲得法を提案し（図 7 Inclusion Criteria に基づく文の意味ベクトルの獲得）、従来手法に対して言語横断文書分類タスクの誤り率が 30%改善できることを示した。

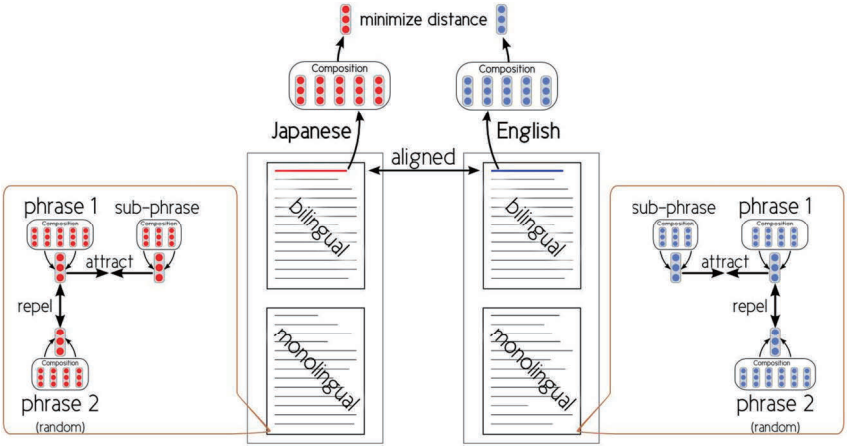

図 7　Inclusion Criteria に基づく文の意味ベクトルの獲得（学術論文 8）

さらに、獲得した意味ベクトルを利用した言語横断の類似文推薦機能を実現して、インタラクティブに類似文を検索したり、ウェブベースのエディタから呼び出したりするユーザインタフェイスを実装した（図 8 深層学習の適用による言語横断類似文検索の実現）。これらの機能は、知識獲得やマイニングだけではなく、たとえば研究者が英語論文を執筆する際の例文検索などにも役立つことが期待される。

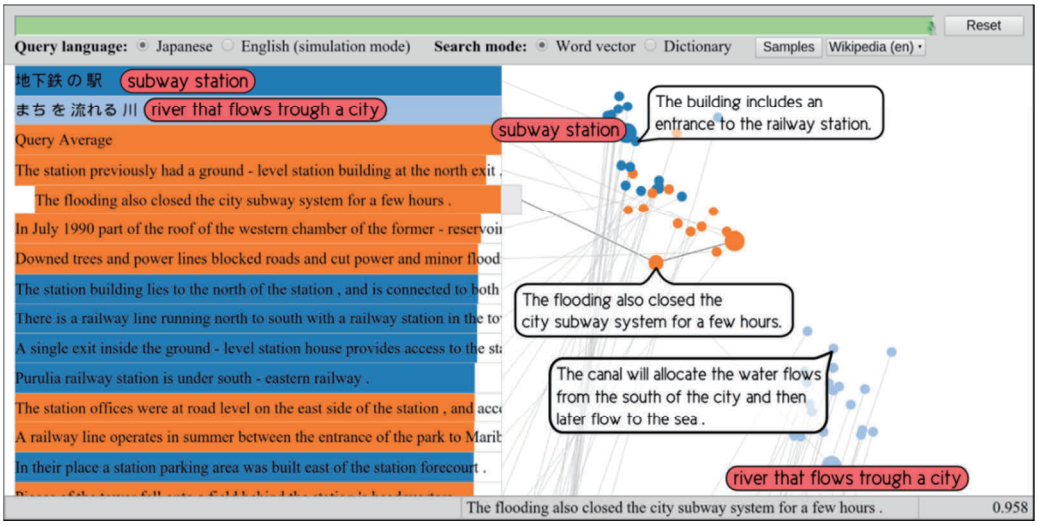

図 8　深層学習の適用による言語横断類似文検索の実現（学術論文 9）

サブテーマ 2 では平成 26 年度はこれまでの継続で、DBpedia Japanese の運営と利用促進と生物種データの LOD 化とその利用などを行った。DBpedia 関連では、国内 LOD データセット 19 件が利用するなど、着実に利用が増えている。またこれを利用した論文を国内ジャーナルに投稿、採録されている。また、これまで行ってきた博物館 LOD では、スミソニアン博物館の LOD 化プロジェクトで先行研究として言及、じんもんこんでのパネル登壇など、着実に浸透している。関連して、H–GIS（Historical GIS）研究会と合同研究会を開催するなどの活動も行った。生物種データの LOD 化では、このデータを利用した絶滅危惧種データとの接続をおこなったり、生物の相互作用の機械学習の研究などを行った。また、これまでの研究をまとめた論文を日本語ジャーナル、英語ジャーナルに 1 編づつ投稿している。

(1) DBpedai Japanese の運営と利用促進

Linked Data [Berners-Lee 06] についての説明でよく使われる図として、The Linking Open Data cloud diagram[Cyganiak 11]（以下、本家図）がある。本家図は 2007 年 5 月に最初の版が発表されて以来、2011 年 9 月まで更新されており、Linked Data の発展や現状を紹介するための端的な図として広く用いられている。2011 年 9 月の段階では本家図内に 295 のデータセットが描かれている。

本家図における 1 つの課題は、日本のデータ公開者による日本語のデータセットが ndlna と NDL subjects のみであることである。これらはどちらも国立国会図書館が公開しているデータであり、実際は Web NDL Authorities として統合されているため、1 つのデータセットのみが 2011 年 9 月の本家図に存在すると言える。

一方で、DBpedia Japanese [Kato 13] や日本語 Wikipedia オントロジー [玉川 13] のように、ここ数年で日本においても様々なデータセットが Linked Data として公開されるようになってきており、人工知能学会セマンティック Web とオントロジー研究会や、2011 年から 3 回開催されている Linked Open Data チャレンジ Japan 等においても数多く報告されている。そこでここでは、日本で公開されている日本語のラベルを含んでいる Linked Data についてのリンク関係等を調査することで、日本語 Linked Data Cloud 図（以下 JLDC 図）としてまとめることを試みた。

2014 年 3 月 10 日現在での JLDC 図は図 1 の通りである。現在対象となっているデータセットの数は 27 個である。調査方法は手動であり、過去のセマンティック Web とオントロジー研究会や Linked OpenData チャレンジ Japan にて報告されているデータセットを主な調査対象としている。SPARQL エンドポイントがある場合は SPARQL による問い合わせを行い、RDF ファイルが取得可能な場合はダウンロードしてローカルで計測している。

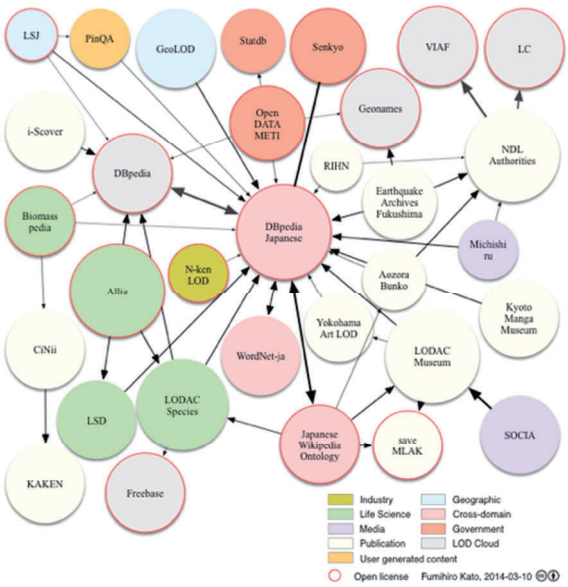

図 1　日本語 Linked Data Cloud 図

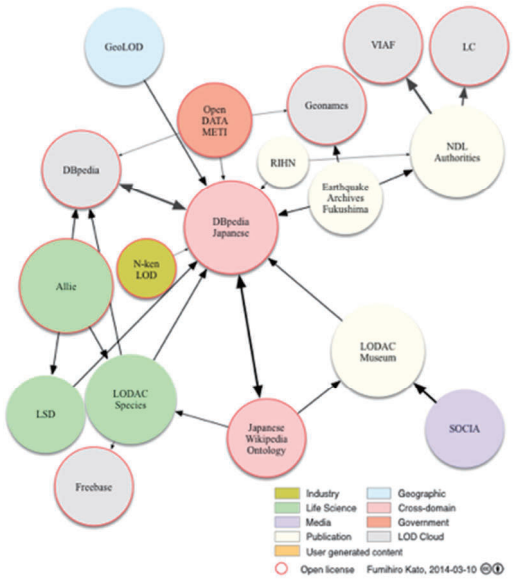

図 3　本家基準による Linked Data Cloud 図

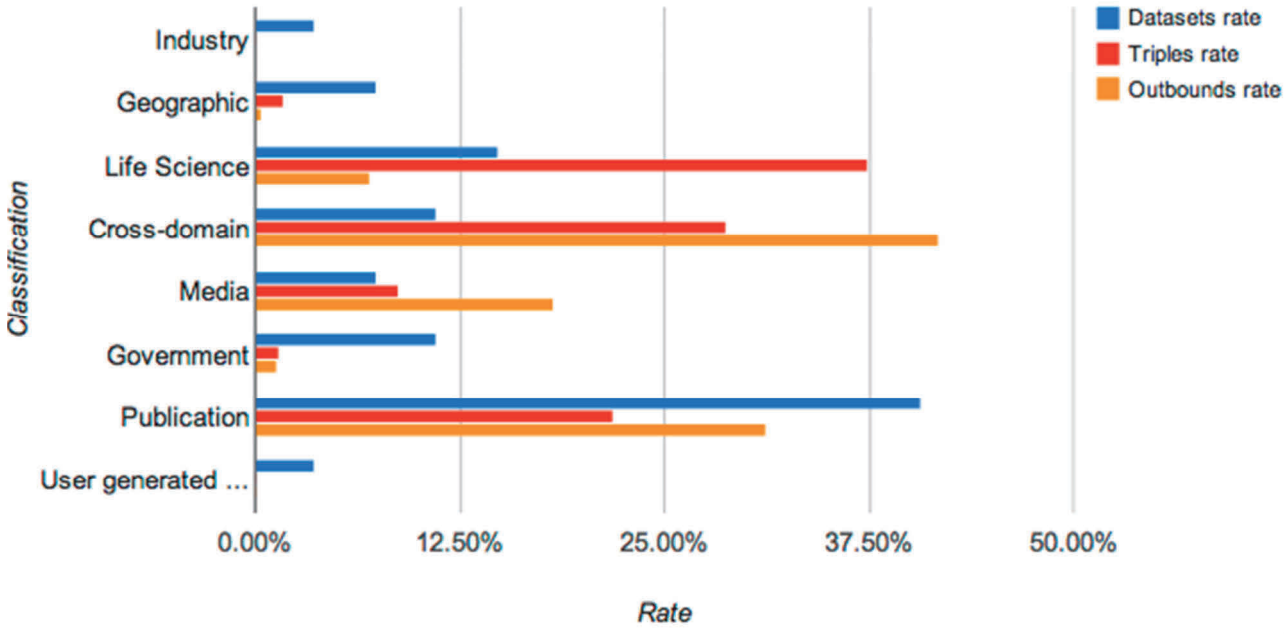

図 2　分類別の割合

JDLC 図に採用するデータセットの基準は以下の通りである。

- データ公開者が日本にいる人・組織等である
- 日本語ラベルを含んでいる

・ 1000 トリプル以上含んでいる
・ 本家図か JLDC 図の既存のデータセットとの RDF リンクが 10 以上ある
・ 参照解決可能な状態、データダンプ、あるいは SPARQL エンドポイントのいずれかによってデータセットを公開している

本家図に入っているデータセットについても、本基準に合致するものは JLDC 図のデータセットとして扱うことにした。その理由は、将来 JLDC 図にしかなかったデータセットが本家図に含まれるようになったときにも変わらずに JLDC 図内でも描けるようにするためである。現在該当するのは Web NDL Authorities のみとなる。

JLDC 図において、Open Definition に適合するオープンライセンスを採用しているデータセットは赤丸で明示している。また、データセットの分類は本家図を参考に独自に行った（表 1、図 2）。唯一、ねじ LOD [Fujii 13] のみは本家図のカテゴリでは分類できないと考えたため、Industry という分類を新設した。

JLDC 図の基準に近いが採用しなかったものとして、RDF リンクの作法が間違っているデータセットが存在した。RDF リンクが間違っているケースとして良くあるのが、Cool URIs [Ayers 08] に関するものである。Cool URIs においてはリソース識別子としての URI とそのリソースに関する文書の URI を明示的に分けることが求められている。しかし、そのように実装されたシステムに対するリンクを生成する際に、文書、特に HTML 文書の URI を使用してしまう例が見受けられた。これは恐らくブラウザでリソースにアクセスした際に、アドレスバーにある URI をそのままリンクに使用したことが考えられる。

JLDC 図におけるデータセットの採用基準は、本家図の採用基準とは異なる。そうした理由は、本家図の基準をそのまま適用すると基準に満たないデータセットが多かったからである。本家図の採用基準をそのまま用いるよりも、まず RDF としてのリンク関係があるデータセットを拾えるようにすることで、個々に何が不足しているのかを示すことが重要だと考えた。

本家図の採用基準 [Cyganiak 11] は Linked Data の 4 原則 [Berners-Lee 06] を解釈したものであり、以下の通りである。

1．解決可能な http://（または https://）URIs でなければならない
2．content-negotiation か何かで良く使われる RDF 形式（RDFa, RDF/XML, Turtle, N-Triples）のいずれかで RDF データを解決できなければならない
3．1000 トリプル以上含んでいる
4．本家図上の既存データセットとの RDF リンクが 50 以上必要である。
5．RDF クローリングまたは RDF ダンプ、あるいは SPARQL エンドポイントによってデータセット全体にアクセスできる。
6．認証なしかつ無料でアクセスできる。

表 1　除外理由の件数

	LOD cloud基準	データセット数（重複あり）
1	解決可能なhttp URIs	8
2	RDFデータの解決	9
3	1000トリプル以上	0
4	50以上のRDFリンク	4
5	データセット全体へのアクセス	2
6	認証なしかつ無料のアクセス	1

最後の項目は基準として記述されているのではなく、オープンについての説明で前提として記述されていることであるが、JLDC 図で本項目に満たないものがあるため基準として明示した。本家図では、オープンは Open Definition に適合するオープンライセンスを採用しているという意味ではなく、認証なしかつ無料でアクセス可能であることとしている。これは、現実にはライセンスを明示的しているデータセットというのが少なかったという事情を考慮したものである。

本家図の基準に合致するデータセットだけを残して図を描くと図 3 のようになる。採用データセット数は 13 個と、JLDC 図の半分以下である。

基準を満たしていない 14 個のデータセットについて、満たしていない基準毎に表 2 としてまとめた。ここでは複数の理由を許している。該当数が多い理由が 1 と 2 であるが、これらは Linked Data の 4 原則における 2 と 3 に相当する。つまり、これらの基準が満たされていないのであれば、本来は Linked Data とは呼べない。そのため、データセットの提供方法が改善されることが望ましい。基準 4 は 4 原則の 4 に相当するが、50 というリンク数には根拠がなく、外部リンクが極端に少ないデータセットを足切りするための数値以上の意味があるとは考えにくい。例えばヨコハマ・アート・LOD は現在 DBpedia Japanese へのリンクが 42 個のため、本家図では基準外になる。しかし、外部リンク数が多いから良いデータセットであるとは必ずしも言えない。なお、本家図の基準に合致するだけで、データセットが本家図に採用されるわけではない。本家図に実際に採用されるためには、基準をクリアしていることを表明するための手続きが別に必要である。具体的には datahub に所定の方式でデータセットのカタログを記述した後に、Data Hub LOD Datasets で個々の要件を確認する必要がある。日本語のデータセットでこの手続きを既に行っているのはわずかしかないが、今後本家図が更新されるときに採用されるためにも手続きを行なっておくことが望ましい。また、SPARQL エンドポイントがある場合は、この手続きによって SPARQL Endpoints Status にも掲載されるようになる。

ここでは日本語 Linked Data Cloud 図を通して、DBpedia Japanese を中心とする日本における Linked Data の現状と課題を紹介した。今後の課題としては、継続的に日本語 Linked Data Cloud 図を維持していくためには、手動で全ての調査を毎回行うには限界があるため、本家図のようなシステマチックな手段を検討する必要がある。

参考文献

[Ayers 08] Ayers, D, V¨olkel, M: Cool URIs for the Semantic Web, W3C Interest Group Note, http://www.w3.org/TR/cooluris/ (2008)

[Berners-Lee 06] Berners-Lee, T: Linked Data, http://www.w3.org/DesignIssues/LinkedData.html (2006)

[Cyganiak 11] Cyganiak, R, Jentzsch, A: The Linking Open Data cloud diagram, http://lod-cloud.net (2011)

[Fujii 13] Fujii, A, Egami, S, Shimizu, H: EDI support with LOD, 2013 Linked Data in Practice Workshop (2013)

[Kato 13] Kato, F, Takeda, H, Koide, S, Ohmukai, I: Building DBpedia Japanese and Linked Data Cloud in Japanese, 2013 Linked Data in Practice Workshop (2013)

[玉川 13] 玉川奨, 香川宏介, 森田武史, 山口高平: 日本語 Wikipedia オントロジーの Linked Open Data への取り組み, 第 27 回 人工知能学会全国大会論文集 (2013)

(2) 生物種データの LOD 化とその利用

我々はこれまで学術情報に関する LOD 基盤の構築を LODAC（Linked Open Data for ACademia）プロジェクトとして行ってきた。対象は博物館情報に始まり［嘉村 10］、生物多様性情報に拡張され［武田 12］、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の閲覧が可能になった。さらにデータの追加とデータモデルの整理を経て［南 13］、今回、絶滅危惧種情報とのデータ統合を試みた。

絶滅危惧種に関するデータとしてはレッドリストとレッドデータブックが有名である。それぞれ、国際機関、国、都道府県など異なるレベルで作られ、大抵の場合、名前と保全状況についての最低限の情報を速やかにレッドリストとして編纂し、後に詳細な情報を加えてレッドデータブックとして発行するという手順を踏んでいる。我々はまず、国レベル、都道府県レベルでのレッドリストのデータ化を試みた。具体的には環境省の生物多様性センターが 2012 年から 2013 年にかけて編纂した第 4 次レッドリスト（以下、環境省レッドリストと呼ぶ）と、2013 年に編纂された京都府レッドリスト（以下、京都府レッドリストと呼ぶ）をデータ化した。

環境省は学名、和名、保全状況、その種のカテゴリーといった情報を公開している。そこで学名「Oceanodroma castro」を持つ種が、和名として「クロコシジロウミツバメ」を持ち、「絶滅危惧 IA 類（CR）」の保全状況で、「鳥類」にカテゴライズされていることは Turtle 形式の RDF で表 2 のように記述できる。

表 2　レッドリスト 1 項目の LOD 化例

```
<http://lod.ac/species/Oceanodroma_castro> a speciesOnto:ScientificName;
speciesOnto:hasCommonName <http://lod.ac/species/クロコシジロウミツバ
メ>;
    speciesOnto:hasSuperTaxon <http://lod.ac/species/鳥類>;
rdfs:label "Oceanodroma castro";
cnsvOnto:hasRedListEntry redlist:jibis-redList2012_tyorui-17.

redlist:jibis-redList2012_tyorui-17 a cnsvOnto:RedListEntry;
    rdfs:comment "クロコシジロウミツバメ"@ja;
    cnsvOnto:ofSpecies <http://lod.ac/species/Oceanodroma_castro>;
    cnsvOnto:currentStatus cnsv:CR;
    cnsvOnto:ofArea "日本"@ja.
```

前半は学名を主語としたトリプルであり、species オントロジの語彙を用いて和名とカテゴリを記述している。カテゴリに関してはデータ源の記載を尊重してそのまま記載している。例えば「汽水・淡水魚類」といったカテゴリが環境省レッドリストに存在するが、そもそも魚類というカテゴリ自体が、系統分類学的ではない用語であり、体系の異なる用語をマッピングすることは困難であるし、また LOD 化の段階でするべきではないと考えている。それを speciesOnto:hasSuperTaxon というプロパティで単に上位の分類群として登録することで、異なる分類体系の情報を共存させることを可能にしている。実際、この種については LODAC Species 内の既存の情報と統合され、NCBI や DBpedia といった他のデータベースへのリンクや、ミズナギドリ目ウミツバメ科に属するといった他の上位分類群の情報、［EOL］等から取ってきた関連する写真が閲覧できるようになっている（図 4）。また学名を主語としたトリプルの一つとして、保全情報オントロジ語彙の cnsvOnto:hasRedListEntry プロパティを用いてレッドリストのエントリを参照しており、その項目の情報を後半で記述している。このように情報を分けているのは、複数のレッドリスト情報が一つの種について存在し得るからである。そして、レッドリストのエントリを主語としたトリプルで具体的な保全状況や指定されている地域などの情報を記述してい

る。京都府レッドリストについても同様であるが、そちらは学名が記載されていなかったため、和名を主語として情報を記述した。

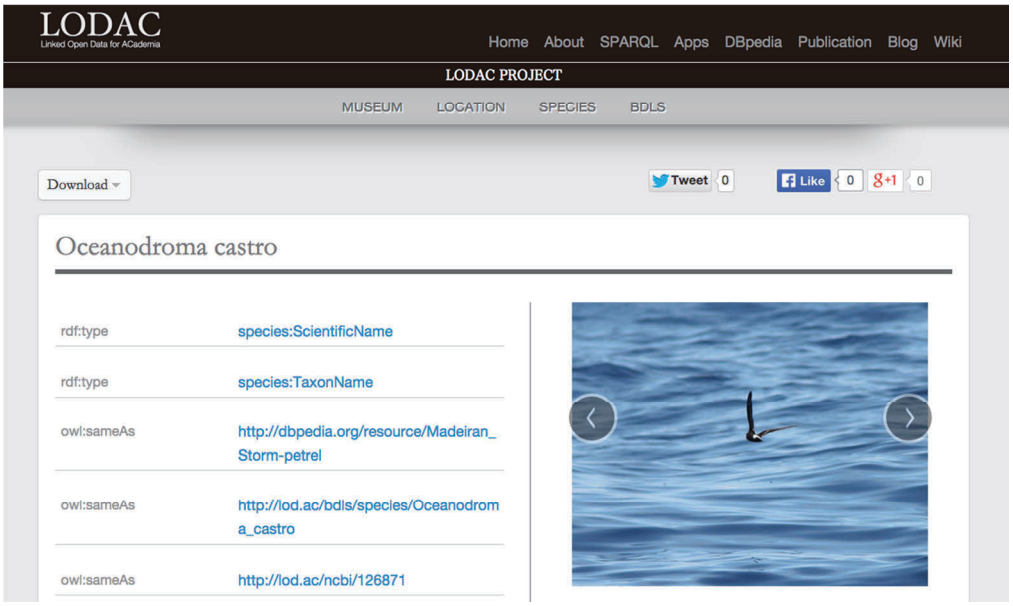

図 4　LODAC における生物種情報の提示の例

統合に際しては、LODAC Species が名前ベースのアーキテクチャを採用し、各生物種に対応する URI が http://lod.ac/species/種名のような形式になっているので、種名が一致すれば自動的に統合されるようになっている。

環境省レッドリスト 5690 件には学名と和名が存在するが、それぞれを用いて LODAC Species との統合を試みたところ、学名を介して 3294 件（57.9%）が既存のデータと統合された。一方、和名を介しては 4145 件（72.8%）の統合に成功した。その和集合は 4711 件（82.8%）となっている。京都府レッドリスト 1871 件については和名のみを用いたが、1598 件（85.4%）という比較的高い割合で統合に成功した。全体的に高い割合で統合に成功したのは、生物の種名が比較的ゆれがなく使わていること、LODAC Species のデータが広い範囲をカバーしていることを示している。和名の方が効率的に統合できた一因は、和名の方が表記ゆれが少ないことではないかと考えている。

元のデータと統合できなかったデータについて、統合すべきデータがそもそも入っていなかったのか、データがあるにもかかわらず何らかの理由で統合に失敗しているのか、を正確に知る術は無い。しかし、統合に失敗した名前について調査することで、より統合率を向上させるのに役立つ知見が得られたので、以下に報告する。

1．統合すべきデータが入っていなかったと思われるもの

例えば、京都府版レッドリストにある「ヨドゼゼラ」に関しては、2010 年に新種「Biwia yodoensis」の発見の論文が出ており、この論文の著者である細谷が琵琶湖生物多様性画像データベースに和名として「ヨドゼゼラ」を記載していることから、比較的新しい種であるために、元のデータの中に対応する種が含まれていなかったと考えられる。また、同じく京都版レッドリストにある「ルイスムネボソヨツメハネカクシ」については、「Boreaphilus lewisianus」という学名が付けられている種の発見自体は 1874 年と古いが、和名がつけられたのが柴田らによって 2013 年に編纂された『日本産ハネカクシ科総目録』においてであるため、元のデータの中に対応する種が含まれていなかったと考えられる。これらの例については、データベースや図鑑、目録といった形で発行される新し

い情報を積極的に入力する他、そのフローを効率化するために各分野の分類学者と協力していく必要がある。

2．複数の和名を持つ種

京都府版レッドリストにある「イモリ」や「モモンガ」は族や科の名前である一方、種として「モモンガ」といった場合には「ニホンモモンガ」（「ホンドモモンガ」ともいう）のことを指し、種として「イモリ」といった場合には「アカハライモリ」を指す。例えば、Wikipedia の生物分類表テンプレートには「和名」という項目があり、それぞれの種について「アカハライモリ、ニホンイモリ、イモリ」「モモンガ、ニホンモモンガ、ホンドモモンガ」という形で名前が列挙されている。また、イモリについては京都府がウェブ上に公開している 2002 年版レッドデータブックにおいて学名「Cynopus pyrrhogaster （Boie, 1826）」が付されているため、「Cynops pyrrhogaster」という名前を持つアカハライモリであると推測され、それは環境省のレッドリストにおいても「準絶滅危惧（NT）」指定されていることがわかる。これらの例については、DBpedia や他のレッドリストデータから同じ種を指す複数の和名を抽出し、それらを関連付けることで解決できると考えられる。

3．ミススペルと表記ゆれ

前述の例の前半「Cynopus pyrrhogaster」と同様の学名を用いた論文、文書はウェブ上に他にも存在したが、タンパク質に関するデータベースにおいて、Cynopus は Cynops のミススペルであるとされている。また、発見者名や年号を後ろにつけるのは学名において多くみられる表記法であり、1700 万以上の学名についての情報を提供している Global Names Index では、「Cynops pyrrhogaster」「Cynops pyrrhogaster (Boie,1826)」「Cynops pyrrhogaster Boie」の３つを表記グループ（Lexical groups）としてまとめている。一方、和名についてはこういった形の表記ゆれはないことが、前述のように効率的に統合できた一因だと考えられる。また環境省レッドリストにある「Aerobryum speciosum(Dozy & Molk.) Dozy & Molk. var. nipponicum Nog.」については、非常に近い表記の「Aerobryum speciosum (Dozy et Molk.) Dozy et Molk. var. nipponicum Nog.」が米倉らの作成した YList に記載されており、それが LODAC Species にも含まれているが、記号表記「&」とラテン語表記「et」の差異やスペースの数の差異といった細かな差異によって統合に失敗している。これらの例については前節と同様、ミススペルや表記ゆれについて扱っているデータベースの情報を追加する他に、すでにあるエントリとの表記上での類似度を計算し、統合先として推薦するという手法が考えられる。我々はすでにオープンソースの検索エンジンである Apache Solr*16 を用いて類似の文字列を検索するシステムを実装しており、今後異なる情報源からのデータの統合の際にこのエンジンを活用したいと考えている。

4．同名異種

統合できたものについては統合成功としているが、その統合が適切だったかどうかについては十分に検討できていない。異なる種に同じ学名が付けられることはほぼ無いと思われるが、和名については昆虫類のカマキリと淡水魚類のアユカケの別名であるカマキリが同名になってしまっているといった例が実際に存在し、それらを区別する仕組みが必要と考えられる。

本研究では、生物情報を共有する LOD 基盤として構築を進めてきた LODAC Species へ、生物多様性に関する重要なデータである絶滅危惧種情報の統合を行った。学名や和名を手がかりとして多くの絶滅危惧種情報を既存の情報と統合できた一方、失敗例の分析を通して、より効率的な統合のためにデータの追加や統合手法の改善の必要性が示された。

参考文献

[Heath 09] Bizer, Christian; Heath, Tom; Berners-Lee, Tim: Linked Data—The Story So Far,

International Journal on Semantic Web and Information Systems 5 (3), pp. 122, Solving Semantic Interoperability Conflicts in CrossBorder EGovernment Services (2009).

[Heath 13] Tom Heath, Christian Bizer: Linked Data: Evolving the Web into a Global Data Space, (邦訳 :Linked Data: Web をグローバルなデータ空間にする仕組み, (2011) 武田英明 監訳, 大向一輝, 加藤文彦, 嘉村哲郎, 亀田尭宙, 小出誠二, 深見嘉明, 松村冬子, 南佳孝 訳 (2013)).

[UNEP 92] UNEP CBD, Convention on Biological Diversity, (1992).

[EOL] Encyclopedia of Life. http://www.eol.org

[GBIF] Global Biodiversity Information Facility. http://www.gbif.org/

[嘉村 10] 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: LOD.AC: Linked Open Data によるミュージアム情報の結合, 第 3 回知識共有コミュニティワークショッ

プ, 情報社会学会, (2010).

[武田 12] 武田英明, 南佳孝, 加藤文彦, 大向一輝, 新井紀子, 神保宇嗣, 伊藤元己, 小林悟志, 川本祥子 : 生物情報基盤構築のための生物種データの Linked Open Data 化の試み, 人工知能学会全国大会 (第 26 回) 論文集, No. 3C2-OS-13b-3, 山口 (2012).

[南 13] Y. Minami, H. Takeda, F. Kato, I. Ohmukai, N. Arai, U. Jinbo, M. Ito, S. Kobayashi and S. Kawamoto: Towards a Data Hub for Biodiversity with LOD, in H. Takeda, Y. Qu, R. Mizoguchi and Y. Kitamura eds., Semantic Technology - Second Joint International Conference, JIST 2012, Nara, Japan, December 2-4, 2012. Proceedings, Vol. 7774 of LNCS, pp. 356361, Springer (2013).

サブテーマ 4 では、平成 26 年度も引き続き研究者情報のエコシステムとしての Researchmap の研究開発を続けた。従来の Researchmap では、業績データと資料公開が個別の内部システム構成により管理されていた。Researchmap に登録されたコンテンツの機械によるハンドリング性能を向上するためには、それらを一括して管理する仕組みの導入が不可欠である。コンテンツ統合管理機能には、メタデータと添付ファイル管理に加え、メタデータのフリーワード検索や詳細検索、全文検索などの機能が必要とされる。コンテンツ統合管理機能が整備されれば、コンテンツの検索機能に加え、研究成果に対する統計解析機能やコンテンツ情報に関連する他のシステムとのマッシュアップなど、Researchmap に登録されたコンテンツを活用した付加価値機能の実現が容易となる。そうした拡張性を実現するためには、コアシステムである Researchmap 本体とは独立にコンテンツ管理のためのリポジトリシステムを用意し、両者のマッシュアップにより Researchmap システム全体を構成するのが妥当である。平成 25 年度には、図 4.1 に示すように、NetCommons2 上で動作するリポジトリシステム WEKO を用いた OpenDepo をコンテンツ統合管理機能として用い、SWORD（Simple Web-service Offering Repository Deposit）プロトコルによるコンテンツ登録、削除、更新 API と、OpenSearch プロトコルによる検索 API を実装した。

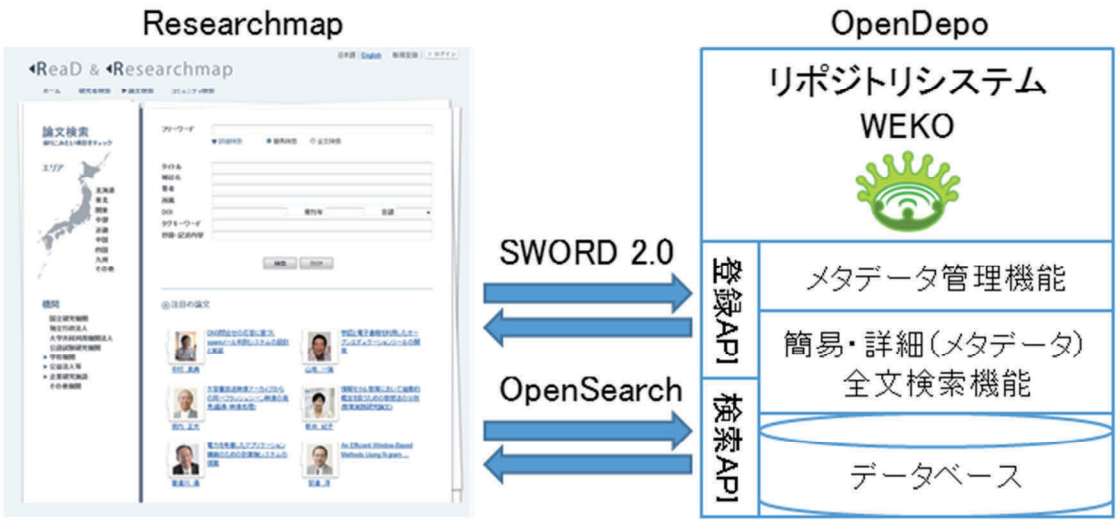

図 4.1　Researchmap と OpenDepo の関係図

この実装により、Researchmap に登録されたコンテンツを統合的に検索できるインターフェースを提供できることになるが、これだけでは CiNii や Google Scholar が既に提供する論文検索機能のサブセット的なものにしかなり得ない。Researchmap の特徴である研究者としてのエンティティやその属性を活用した研究成果に対する統計解析機能などが加われば、単なる論文検索エンジンにはない付加価値機能として活用できる。また、現在の Researchmap に登録されるコンテンツはメタデータ（書誌情報）が主であり、本文ファイルが登録されているものは、論文で約 0.5%、講演口頭発表等で約 0.1%程度である。Researchmap は、業績データベースとしてのメタデータの登録は進んでいることから、大学図書館等がオープンアクセスを推進する目的で運用する機関リポジトリとの連携が実現すれば、メタデータだけではなく Researchmap からの本文ファイルへの到達性を高めることができる。こうした可能性を実現するために、平成 26 年度の研究開発では、Researchmap における検索機能の機関別統計機能と機関リポジトリとのハンドシェイク機能を実現した。

Researchmap における検索機能の機関別統計機能

従来の検索 API では、フリーワードに加えてタイトル、誌名・会議名、著者・講演者、出版年・開催年、記述言語による詳細検索機能を提供した。また、メタデータの著者・講演者名から Researchmap 内の研究者マイページにリンクすることにより、業績データベースとしての Researchmap の特徴を生かし、その研究者に関する研究内容の詳細を閲覧できるようにした。これらは主として研究者や学生を対象とした検索機能である。一方で、機関を横断した研究者の業績データベースは、研究者のみならず、大学経営者・大学評価者・リサーチ・アドミニストレーター（URA）などが、それぞれの機関の特徴を他大学と比較・評価するための情報源としても有用である。

Researchmap 上では、ID 化された所属機関情報が研究者情報とともに管理されている。この研究者の属性情報を各コンテンツのメタデータの一部として管理し、詳細検索の AND 条件として追加検索できるように機関別統計機能を実装した。このクエリに従って、詳細検索の結果を OpenDepo から Researchmap に図 4.2 に示したような JSON 形式で戻す API を用意した。

```
{
   "opensearch:totalResults":28,   # 検索結果数
   "items":
   {
        "1234567890":12,          # key=>10 桁の数値(機関 ID), value=>検索結果数
        "9876543212":3,
        ...
   }
```

図 4.2　機関別統計機能における OpenDepo から Researchmap への応答例

Researchmap 上では、図 4.3 に示すように論文検索タブに加えて新たに機関統計タブを用意する。例えば、フリーワード入力後に論文検索が実行され、次に機関統計タブがクリックされた場合には、そのフリーワードに対する機関統計の結果が表示される（図左）。この状態において、ある大学の統計グラフがクリックされた場合には、フリーワードと共にその機関名が入力された状態での論文検索画面に遷移する（図右）。

図 4.3　Researchmap 上での機関別統計機能提供例

Researchmap 上で管理される所属機関属性情報は、その研究者が現在所属する機関となる。したがって、本統計機能で得られる情報には、過去に他機関で得た業績情報が含まれる。研究者の所属機関の編纂情報も加味した統計情報の提供方法については、今後の課題となる。また、フリーワードを起点とした単純な統計情報では、大規模大学が常にランキングの上位を占めることになる。こうした単純なランキングでは発見することが困難な、各機関の特徴をいかに利用者に提示するかについても、更なる研究開発が必要である。

Researchmap と機関リポジトリとのハンドシェイク機能

欧米では、研究情報システム（CRIS：Current Research Information System）と機関リポジトリの連携が積極的に進められている。しかしながら、それらの多くは双方のメタデータの共有に留まってお

り、機関リポジトリの運用ワークフローにまで踏み込んだものは多くない。それに対し、本研究開発で実装する機能は、機関リポジトリにおける本文収集のワークフローも加味した連携機能となる。

Researchmap と機関リポジトリとのハンドシェイク機能の全体構成図を、図 4.4 に示す。OpenDepo で管理されるメタデータは、研究者とその所属機関情報と連結して保存される。IR 連携エンジンは、OAI-PMH による差分更新で最新のメタデータ情報を OpenDepo から取得する。本機能を提供する機関リポジトリ群は、JAIRO Cloud とした。JAIRO Cloud では、利用機関の情報を UMS（User Management System）と呼ばれるシステムで管理している。IR 連携エンジンは、UMS とマッシュアップして利用されるシステムとして構築した。機関リポジトリ運用担当者は、UMS のユーザインタフェイスから、ハンドシェイク機能の利用の可否と情報の送信方法が選択できる。IR 連携エンジンで抽出された、各機関の研究者業績データの差分情報は、2 種類の方法で機関リポジトリ側に通知される。

1 つは、UMS で管理されている機関リポジトリ運用担当者に電子メイルアドレスへ通知する方法である（図中①）。メイル本文には、メタデータ情報とともに UMS 上のコンテンツ登録 URL が記載される。機関リポジトリ運用担当者は、その URL にアクセスし、自機関のリポジトリに登録したい研究者業績データの差分情報を選択する。選択されたコンテンツメタデータは、SWORD プロトコルにより UMS から機関リポジトリに送信される。その後、各機関の運用フローに従い、機関リポジトリ運用担当者は当該コンテンツの本文ファイルを研究者から取得後、本文とともにコンテンツを公開することになる。もう一つの方法は、機関リポジトリ運用担当者へのメイル通知を介すことなく、UMS から研究者業績データの差分情報を全て機関リポジトリに直接登録する方法である（図中②）。この場合にも、コンテンツの登録には SWORD プロトコルを利用する。

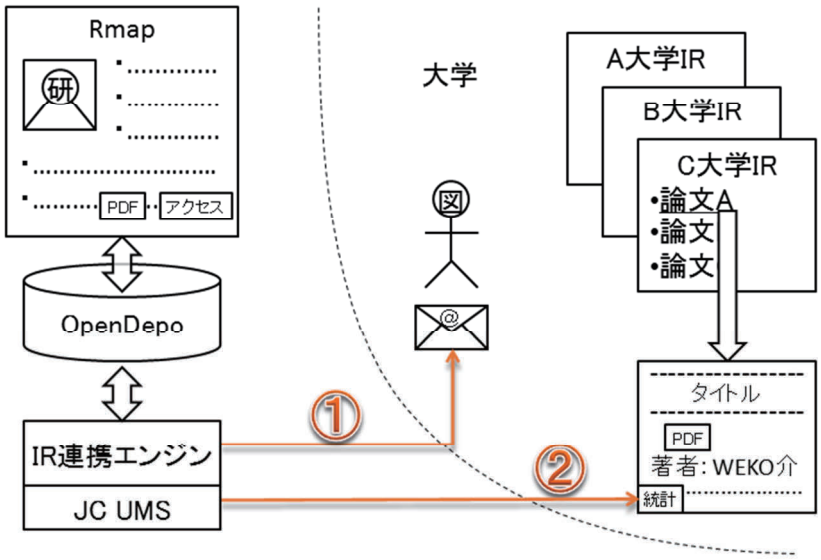

図 4.4　ハンドシェイク機能の全体構成図

実装した機能の評価は、平成 27 年度に実施する。Researchmap と機関リポジトリの更なる連携機能として、機関リポジトリに登録されたコンテンツへのリンクを Researchmap 上に表示することが挙げられる。これについては、今後、本連携機能を介して登録されたコンテンツ以外に、既に機関リポジトリに登録されているコンテンツの本文情報についても同様に Researchmap 上で表示されることが望まれる。その機能を実現するためには、Researchmap 上に登録された業績情報と機関リポジトリ上の論文情報を突合させる必要がある。Researchmap に登録されたメタデータ情報は、研究者自身が登録することから、メタデータ情報としての精度に欠くことが事前の調査から明らかになっている。この問題に対

応するためには、適切な文字バイグラムの一致率で両者の情報を突合するなどの方法が考えられる。本サブテーマで開発する連携機能の強化により、Researchmap からの本文情報へのリンク率を高めることができれば、論文検索機能の有用性が増すばかりでなく、Researchmap 上での研究者ネットワークの俯瞰機能や、より利便性の高いリコメンドやサジェスト機能への発展にも繋がるものと考えられる。

サブテーマ 4 においては平成 26 年度にもうひとつ大きな研究の展開を行った。それは Researchmap というソフトウェアを研究者情報エコシステムに留まらず、初等中等教育を新たなコンテンツとして edumap のプロトタイプを設計・構築したことである。

学校は、日本の最も重要な公的な施設のひとつであるにも関わらず、学校の基本的な情報（学校名・所在地・電話番号・メールアドレス・校種・校長名・ホームページアドレス等）は、これまで、それぞれの学校を所管する教育委員会が把握をするだけで、文部科学省では整備をしてこなかった。そこで、Researchmap を用いて、日本の約 4 万に上る学校の基本情報をオープンデータ化し API で公開する edumap というプロジェクトを開始した。

図 4.5 edumap のトップページ

図 4.6　edumap の学校ポータルページ

図 4.5 はプロトタイプのトップページ、図 4.6 は各学校のポータルページである。

図 4.7　位置情報との連動

サブテーマ２においてこれまで研究を進めてきた Linked Open Data の考え方に基づき、各学校の各項目情報を RDF で記述し、可用性を高めている。また、第一期新領域融合研究で設計構築した公共機関向けコンテンツマネージメントシステム NetCommons を用いて構築されている学校ホームページの新着情報を RSS で読み込み、それを県別・校種別にまとめて配信する機能を備えている。これによって、各県のサイト更新率を自動的に把握する他、各県内での更新率上位の学校のリストが自動的に計算される。

本サイトのコンテンツの基盤となる各学校の基本情報は wikipedia から自動取得・整理した上で、リンクが切れている場合など情報が古い場合には人手で修正を加え整備した。Google Map の API を用い、当該の学校の所在地を地図上で表現するほか、近隣の学校を併せて表示できるようにした。これにより将来的には、ある学校でインフルエンザが発生し学級閉鎖になったという情報等を、自動収集し、自然言語解析によってそれが「流行性の疾病の発生」であることを把握した上で、それが一週間でどれくらい伝染する可能性があるか等を位置情報から予測し、近隣の学校に通知すること等が可能となる。これにより、今までは学校内・教育委員会内で閉じていた情報が、校種・教育委員会を超えて共有されるようになるメリットがある。

さらに、校舎の耐震化工事実施状況や災害時の避難所指定の有無などを項目として持つことによって、震災時にどの学校は被害が大きいかなどを震源地と耐震化の有無、学校規模等から予測し、ターゲットを絞って迅速に対応したり、授業開始の柔軟な判断などをしたりすることにつながることが期待できる。

〔5〕 研究成果物

① 知見・成果物・知的財産権等

1. 文書変換ツール PlaneText を公開した（http://kmcs.nii.ac.jp/planetext）
2. 言語処理学会年次大会で論文推薦・閲覧支援のデモシステムを公開した（http://kmcs.nii.ac.jp/nlp_annual）

② 成果発表等

＜論文発表＞

〔学術論文〕

1. 相良毅，古川竜也，相澤彰子：「LDA を用いた学術用語の対訳選択手法」、情報知識学会 第 22 回（2014 年度）年次大会　2014 年 5 月（査読なし国内会議）
2. 古川竜也，相良毅，相澤彰子：「言語横断エンティティリンキングのための語義曖昧性解消」、情報知識学会 第 22 回（2014 年度）年次大会（学生奨励賞受賞）　2014 年 5 月（査読なし国内会議）
3. Panot Chaimongkol, Akiko Aizawa: "Corpus for Coreference Resolution on Scientific Papers, " The 9th Language Resources and Evaluation Conference (LREC 2014)　2014 年 5 月（査読付き国際会議）
4. 阿辺川武，相澤彰子：「内部構造解析機能と脚注表示機能を備えた論文閲覧システム」、インタラクティブ情報アクセスと可視化マイニング研究会 (SIG-AM) 第 7 回研究会　2014 年 6 月（査読なし国内会議）
5. 原忠義，トピチ ゴラン，宮尾祐介，相澤彰子:「実文書を自然言語処理技術と適切に繋ぐ技術の重要性」第 217 回自然言語処理研究会（SIG-NL）　2014 年 7 月（査読なし国内会議）
6. Tadayoshi Hara, Goran Topic, Yusuke Miyao, Akiko Aizawa: "Significance of Bridging Real-world Documents and NLP Technologies, " Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT), Coling 2014 Workshop　2014 年 8 月（査読付き国際会議）
7. Kiyoko Uchiyama, Takeshi Abekawa, Akiko Aizawa: "A Study of Analyzing Comprehensive Reading Behavior for Research Activities based on a Web-based Experiment, " the 2015 Conference of the

Pacific Association for Computational Linguistics (PACLING 2015)　2015年5月(査読付き国際会議、採録済)

8. Hubert Soyer, Pontus Stenetorp, Akiko Aizawa："Leveraging Monolingual Data for Crosslingual Compositional Word Representations,"　International Conference on Learning Representations (ICLR 2015)　2015 年 5 月　(査読付き国際会議、採録済)
9. Hubert Soyer, Goran Topić, Pontus Stenetorp, Akiko Aizawa: "CroVeWA: Crosslingual Vector-Based Writing Assistance," Demonstration Track, in the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)　2015 年 5 月 (査読付き国際会議、採録済)
10. 込山悠介, 番野雅城, 鑓水優行, 加藤文彦, 大向一輝, 武田英明, 清水謙多郎： 生命科学の複数 LOD の統合による目的別タンパク質分子間相互作用 Linked Open Data の構築, 人工知能学会論文誌, Vol. 29, No. 4, pp. 356-363 (2014)
11. The impact of A.I. – Can a robot get into The University of Tokyo?,Noriko H. Arai, National Science Review,,, 査読有り
12. The impact of A.I. on education - Can a robot get into the University of Tokyo?,Noriko H. Arai, Takuya Matsuzaki, The Proceedings of The 22nd International Conference on Computers in Education,, 1034-1042, 2014 年 12 月 査読有り
13. CMS を合理的に選択するためのソフトウェア特性指標の策定, Fumihiro Kumeno, Yuuta Kohama and Noriko Arai, ソフトウェアエンジニアリングシンポジウム 2014 (SES2014),,, 2014 年 9 月 査読有り
14. Development of a database module for information literacy education through the construction of collective knowledge,Shingo Sugawara, Ryuji Masukawa, Kazuki Hyodo, Noriko H Arai, Proceedings of the 16th IASTED International Conference Computer and Advanced Technology in Education (CATE2014),, 15-22, 2014 年 7 月 査読有り
15. Automated Natural Language Geometry Math Problem Solving by Real Quantier Elimination, Hidenao Iwane, Takuya Matsuzaki, Noriko Arai and Hirokazu Anai,, Proceedings of the 10th International Workshop on Automated Deduction (ADG2014),, 75-84, 2014 年 7 月 査読有り
16. Mathematics by Machine,Noriko H. Arai, Takuya Matsuzaki, Hidenao Iwane, Hirokazu Anai, Proceedings of 39th International Symposium on Symbolic and Algebraic Computation (ISSAC 2014),, 1-8, 2014 年 7 月 査読有り
17. Cognitive model of generic skill: cognitive processes in search and editing,Akira Fujita, Masayuki Suzuki, Noriko H. Arai, Proceedings of 36th Annual Conference of the Cognitive Science Society (CogSci 2014),, 2234-2239, 2014 年 7 月 査読有り
18. The most uncreative examinee: a first step toward wide coverage natural language math problem solving,Takuya Matsuzaki, Hidenao Iwane, Hirokazu Anai, Noriko H. Arai, Proceedings of 28th Conference on Artificial Intellegence (AAAI 2014),, 1098-1104, 2014 年 7 月 査読有り
19. 機械による数学, 新井紀子, 数学文化,(21),, 2014 年 4 月 査読無し

〔著書等〕

1. 武田英明: Linked Open Data (LOD), インターネット白書 2015, pp. 235-240, インプレス R&D (2015)
2. ロボットは東大に入れるか (よりみちパン! セ) (よりみちパン!セ), 新井紀子, 2014 年 8 月, イースト・プレス
3. 人間とは何か―先端科学でヒトを読み解く (科学のとびら), 榊佳之, 新井紀子, 唐津治夢, 山極寿一,

2014 年 10 月, 東京化学同人
4. 日本未来図 2030,, 2014 年 12 月, 日経 BP 社

〔解説・総説〕
1. 相澤彰子：「デジタル化された学術文献の言語解析について」、情報の科学と技術　特集：デジタル時代の日本語 64(11) 2014 年 11 月

〔その他〕

＜会議発表等＞
〔招待講演〕
(国際)
1. H. Takeda: Knowledge is Power (now again), in 4th Joint International Semantic Technology conference (JIST2014), Chiang Mai, Thailand (2014), (KeyNote Speaker)
2. H. Takeda: Dimensions of Open Data Activities in Japan: Policy, Technology and Community, in Open Data Towards Open Governmen conference, Bangkok, Thailand (2014), Electronic Government Agency (EGA), (Invited Talk)
3. ロボットは東大に入れるか, 新井紀子, IEEE Japan Council Women in Engineering 2014, 2014 年 7 月 5 日
4. Mathematics by Machine, Noriko H. Arai, The International Symposium on Symbolic and Algebraic Computation 2014, 2014 年 7 月 23 日
5. Todai Robot Project, Can an AI get into the University of Tokyo?, Noriko H. Arai, IEEE 3rd Global Conference on Consumer Electronics (GCCE 2014), 2014 年 10 月 8 日
6. Creating the artificial intelligence program able to qualify for university entry, 新井紀子, 10th Global Meeting of Women's Forum, 2014 年 10 月 16 日, Women's Forum for the Economy and Society
7. The impact of A.I. on education - Can a robot get into the University of Tokyo?, Noriko H. Arai, The 22nd International Conference on Computers in Education, 2014 年 12 月 2 日,
8. When and how an A.I. will be smart enough to design?, 新井紀子, ASP-DAC 2015 (twentieth annual international conference on VLSI design automation in Asia and South Pacific region), 2015 年 1 月 22 日

〔一般講演〕
(国際)
1. R. Chawuthai, H. Takeda and T. Hosoya: Link Prediction in Linked Data of Interspecies Interactions using Hybrid Recommendation Approach, in 4th Joint International Semantic Technology conference (JIST2014), Chiang Mai, Thailand (2014)
2. M. Thompson, S. Battle, J. Padget and H. Takeda: ArtFinder: A Faceted Browser for Cross-Cultural Art Discovery, Workshop on Human-Semantic Web Interaction (HSWI'14) together with the 11th Extended Semantic Web Conference (ESWC'14) (2014)
3. Yamaji, K., Kato, H., Aoyama, T., Yamada, T., "Handshake ecosystem for Educational Contents between Institutional Repository and OER based Repository", 9th International Conference on Open Repositories, 2014.

4. Aoyama, T., Suzuki, Y., Yamaji, K., “Tag Cloud of Search Queries for Repository System”, 9th International Conference on Open Repositories, 2014.
5. Yamaji, K., Aoyama, T., “Shared Supplemental Data Repository for Japanese Academic Societies in Information Sciences”10th International Digital Curation Conference, 2015.

（国内）

1. 亀田尭宙, 加藤文彦, 神保宇嗣, 大向一輝, 武田英明： Linked Open Data による絶滅 危惧種情報共有の試み, 人工知能学会全国大会(第 28 回), No. 1G4-OS-19a-3, 松山市 (2014).
2. 加藤文彦, 武田英明, 小出誠二, 大向一輝： 日本語 Linked Data Cloud の現状, 人工知能学会全国大会(第 28 回), No. 1G5-OS-19b-7, 松山市 (2014).
3. 小出誠二, 加藤文彦, 小林巌生, 大向一輝, 武田英明: 企業コードと XBRL データの LOD 化, セマンティックウェブとオントロジー研究会, No. SIG-SWO-035-09 人工知能学会 (2015)
4. ロボットは東大に入れるか 2014, 新井紀子, 人工知能学会全国大会特別セッション, 2014 年 5 月 12 日, 人工知能学会

＜受　賞＞

1. 古川竜也, 相良毅, 相澤彰子 :「言語横断エンティティリンキングのための語義曖昧性解消」、情報知識学会 第 22 回（2014 年度）年次大会　 2014 年 5 月（査読なし国内会議）(学生奨励賞)
2. 独立行政法人科学技術振興機構 データサイエンス・アドベンチャー杯「言語部門優秀賞」TermLink : 言語横断論文推薦のための専門用語処理（受賞者: 相澤彰子, 相良毅　チーム名 : T-linkage）
3. 情報知識学会 第 11 回（2014）論文賞 機関リポジトリコンテンツの多面的な学内利用フレームワークの提案と実装, 受賞者: 青山俊弘, 山地一禎, 池田大輔, 行木孝夫

③ その他の成果発表

1. 人工頭脳プロジェクト「ロボットは東大に入れるか」, 新井紀子, QCon2014, International Software Development Conference, 2014 年 4 月 30 日
2. ロボットは東大に入れるか, 新井紀子, 財務省職員セミナー, 2014 年 5 月 14 日, 財務省
3. イノベーション創出を支える博士人材の育成, 新井紀子, 政策のための科学シンポジウム, 2014年6月2日,
4. NetCommons3、こうなります！, 新井紀子, NetCommons ユーザカンファレンス 2014, 2014 年 8 月 6 日,
5. edumap - オープンデータが拓く教育の未来, 新井紀子, NetCommons ユーザカンファレンス 2014, 2014 年 8 月 6 日,
6. コンピュータが仕事を奪う, 新井紀子, ガートナージャパン　特別講演, 2014 年 9 月 12 日,
7. 人工知能がもたらす人間と社会の未来, 新井紀子, NTT-GLOCOM 研究会, 2014 年 9 月 17 日
8. 情報力（知）の共進化, 新井紀子, 第 55 回自律分散システム部会研究会, 2014 年 11 月 7 日,計測自動制御学会システム・情報部門 自律分散システム部会
9. 統合タスク「ロボットは東大に入れるか」の意味と意義, 新井紀子, Cloud Days 2015, 2015 年 3 月 12 日,
10. わが国の経済成長に向けてロボット技術が果たす役割, 新井紀子, 第四回経済好循環実現委員会, 2015 年 3 月 26 日, 自由民主党
11. ロボットと拓く明日, 新井紀子, 教育情報 ,14,, 3-3, 2014 年 4 月

12. ホワイトカラーの職場はロボットに奪われる，新井紀子，文芸春秋 ,92,(10), 156-163, 2014 年 7 月
13. 財務省「職員セミナー」ロボットは東大に入れるか，新井紀子，ファイナンス ,50,(5), 67-73, 2014 年 8 月
14. 進化をつづける人工知能，新井紀子, Newton 別冊　注目のハイテク 35 ,,, 150-155, 2014 年 11 月
15. ロボットビジネス主役交代，朝日新聞社，朝日新聞, (朝刊 2 面（ザ・テクノロジー）), 2014 年 5 月 1 日
16. ロボットは東大に入れるか，埼玉県立川越高等学校，スーパーサイエンスハイスクール全校講演会,, 2014 年 5 月 19 日
17. ロボットは東大に入れるか，高知工科大学環境理工学群，理工学のフロンティア,, 2014 年 5 月 30 日
18. 人工知能　米追う中国，朝日新聞社，朝日新聞, (朝刊 2 面（ザ・テクノロジー）), 2014 年 6 月 7 日
19. 東大合格を目指すロボットにホワイトカラーは勝てるか，ダイヤモンド社，週刊ダイヤモンド，(特集ロボット・AI 革命), 2014 年 6 月 14 日
20. 「東大ロボット」で未来を問う，朝日新聞社，朝日新聞, (土曜版 be　フロントランナー), 2014 年 7 月 26 日
21. 「機械学習」革命　～的中したビル・ゲイツの予言，日経 BP 社，ITPro, (「機械学習」革命　5), 2014 年 8 月 8 日
22. 「機械学習」革命　～的中したビル・ゲイツの予言，日経 BP 社，ITPro, (「機械学習」革命　5), 2014 年 8 月 8 日
23. 人工知能、どこまで進歩，日本経済新聞社，日本経済新聞, (創論), 2014 年 8 月 24 日
24. 焦点，日本電気協会新聞部，電気新聞, (朝刊 1 面コラム), 2014 年 9 月 11 日
25. 書評（ロボットは東大に入れるか），信濃毎日新聞社，信濃毎日,, 2014 年 9 月 25 日
26. 書評（ロボットは東大に入れるか），東洋経済，週刊東洋経済,, 2014 年 10 月 4 日
27. 東ロボくん猛勉強！！国立情報学研の人工知能，毎日新聞社,毎日新聞, (夕刊一面（チェック）),2014 年 10 月 23 日
28. 偏差値 47…476 大学で「A 判定」　人工知能「東ロボくん」センター模試に挑戦，産経新聞社，産経新聞, (朝刊 24 面), 2014 年 11 月 3 日
29. 人工知能「普通の高 3」？，静岡新聞社，静岡新聞，(朝刊 26 面), 2014 年 11 月 3 日
30. 東ロボくん偏差値 47　東大目指す人工知能　模試成績，朝日新聞社，朝日新聞，(朝刊 3 面), 2014 年 11 月 3 日
31. ロボットが国立大合格!?，電経新聞社，電経新聞, (朝刊 4 面), 2014 年 11 月 3 日
32. 人工知能　東大への道遠く　センター模試 470 大学で合格 A 判定，中日新聞社，東京新聞, (朝刊 25 面), 2014 年 11 月 3 日
33. センター模試受験　国立 4 大学　AI が A 判定，日刊工業新聞社，日刊工業新聞, (朝刊 13 面), 2014 年 11 月 3 日
34. 人工知能、偏差値アップ　センター模試「東大は無理」，日本経済新聞社，日本経済新聞，(朝刊 34 面), 2014 年 11 月 3 日
35. 産経抄，産経新聞社，産経新聞, (朝刊 1 面), 2014 年 11 月 4 日
36. 人工知能と歩む未来はバラ色か？，東京都立国立高等学校，進路指導講演会,, 2014 年 11 月 5 日
37. ロボットは人間になれるのか？，集英社，メンズノンノ, (おしえてわかる人。), 2014 年 11 月 10 日
38. 人工知能が拓く未来，毎日新聞社，週刊エコノミスト,, 2014 年 11 月 18 日
39. NEWS Web,NHK,NEWS Web,, 2014 年 11 月 28 日
40. 人工知能の未来を注視し研究を怠るな，日本経済新聞社，日本経済新聞, (社説・春秋), 2014 年 12 月 21 日

41. 勝負の先に見えてくるもの，文芸春秋社，文芸春秋，(資生堂トークセッション　美しき挑戦者たち vol.5), 2015 年 1 月
42. 人工知能のスペシャリストが観る「人類×機械」の創造的未来,プレジデント社, Forbes Japan, (未来を創る日本の女性 10 人), 2015 年 1 月
43. 必要なのはロボットバリアフリー社会，経済産業新報社，経済産業新報, (6 面), 2015 年 1 月 1 日
44. ロボットとくらせる日はくるの？，朝日新聞社,朝日新聞, (Re:お答えします　朝刊 37 面), 2015 年 1 月 1 日
45. コンピューターが未来をどう変えるのか〜人工知能の最前線から〜，ラジオアクセスフォーラム，ラジオフォーラム，(105 回)，2015 年 1 月 5 日
46. 働き方 Next，日本経済新聞社，日本経済新聞，(1 面、12 面)，2015 年 1 月 7 日
47. 1995-2015 私のこの 20 年，TBS ラジオ，久米宏　ラジオなんですけど,, 2015 年 1 月 17 日
48. 競争相手は「ロボット」？，毎日新聞社，毎日新聞，(朝刊 11 面), 2015 年 1 月 19 日
49. 超人的な能力は吉か狂か　ロボットと共存する未来，朝日新聞社，AERA, (vision その③
50. ホワイトカラーの仕事半分を奪うだろう), 2015 年 1 月 26 日
51. 自動運転・AI・ロボット，毎日新聞社，週刊エコノミスト,, 2015 年 1 月 27 日
52. 知的労働　半分奪う，東京新聞社，東京新聞, (朝刊 28 面), 2015 年 1 月 27 日
53. かぞく百景 I　未来教室に向けて　情報通信技術と学校（下），西日本新聞社，西日本新聞, (朝刊 20 面), 2015 年 1 月 27 日
54. 橋下市長が進める“未来の教室”，朝日放送，キャスト特集,, 2015 年 2 月 25 日
55. 人工知能が東大に合格する日，東洋経済新報社，週刊東洋経済, (三人三談　第 7 回　機械 vs.人間), 2015 年 2 月 28 日
56. 人気職種はどう変わる　わが子の未来予想図，プレジデント社,プレジデントファミリー,, 2015 年 3 月 5 日
57. ロボットと拓く明日は何色？，都立小石川中等学校，進路講演会,, 2015 年 3 月 23 日
58. ロボットと歩む未来はバラ色か？，福島県教育委員会，オールふくしまリーダー育成プロジェクト,, 2015 年 3 月 25 日