

質と量を追求する「情報リンクージ」プロジェクト

急速に増大し、細分化する学術情報。それを有効に活用するには、「参照」など“つながり”を表す「リンクージ情報」によって散在する情報を結びつけ、まとめる仕組みが求められる。それに応えるため新領域プロジェクトの1つとして、異なる分野や組織の研究者の協力で「大規模リンクージ」をテーマにした研究が進められてきた。

あちこちの機関やデータベースに散在する膨大な学術情報を利用しやすくするため、2004年、新領域融合研究センターのプロジェクトの1つである「分野横断型融合」のもとで、NIIコンテンツ科学研究系の相澤彰子教授を代表に「大規模リンクージ」の研究が開始された。それから5年、最終年を迎えて、他のプロジェクトとの連携や今後の方向性を打ち出している。

計算機の同定技術を 統計学で裏付け

学術情報は、学術機関やNIIなどさまざまな場所に散在しており、それらを結び付けるには、まず情報を各場所から1カ所に集めてくるのが基本になる。例えば、検索窓に研究者名を入れ関連した画面が表示されるようにするには、集められた情報を分析して、何と何が同じものを示しているのかを同定し、それに基づいてデータを再整理しなくてはならない。こうした技術は、従来は人手に頼るしかなく、コストもかかるうえ裏方の作業とみなされてきた。そうした領域に対して、計算機を用いた情報処理技術を活用していくのがこのプロジェクトの目的だ。

プロジェクトに参加しているNII情報社会

相関研究系の孫媛准教授は、「科学計量学」といわれる分野で、学術論文のデータベースから、論文数や引用数をカウントし、それを基本的な指標として、研究者の活動を定量的に分析する研究に取り組んでいる。孫媛教授は「こうした統計結果を外に出す際には、高い信頼性が求められるため、いわゆる名寄せなどに基づいてデータを編集し、検証していかなくてはなりません」という。

さまざまなデータベースから収集した情報そのままでは、所属機関や国名の表記が統一されていないことも多く、例えば英語の略称では“Univ. Tokyo”と“Tokyo Univ.”などの表記のぶれもある。また、大学や研究機関の統合・再編などにも気配りして、名寄せをしなくてはならない。これを手作業で行おうとすれば膨大な手間がかかってしまう。

そこで、情報学の専門家がコンピュータを用いた同定技術を駆使して処理すればまず量の問題がクリアされ、さらに統計学の専門家が質の面でお墨付きを与えれば、効率よい仕組みを構築することができることになる。こうして、融合プロジェクトでは、統計数理研究所などからも研究者に参加してもらいコラボレーションによる研究を行っている。

2種のサービスを開発し実用化

参加研究者の協力によりプロジェクトが技術を結集して再構築したデータは、学術コミュニティに不可欠な情報サービスを開発しているNII学術コンテンツ研究開発センターの蔵川圭特任准教授らによってサービスとして組み立てられ、09年4月から公開されている。

その1つが「科学研究費補助金データベー



相澤教授、孫媛教授などを中心に行われるプロジェクトメンバーの研究会。

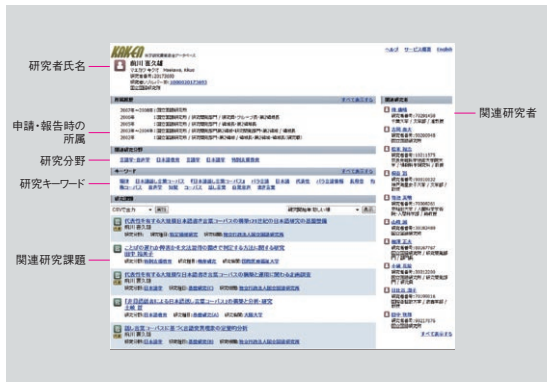


図1 「科学研究費補助金データベース(KAKEN)」の研究者ページ

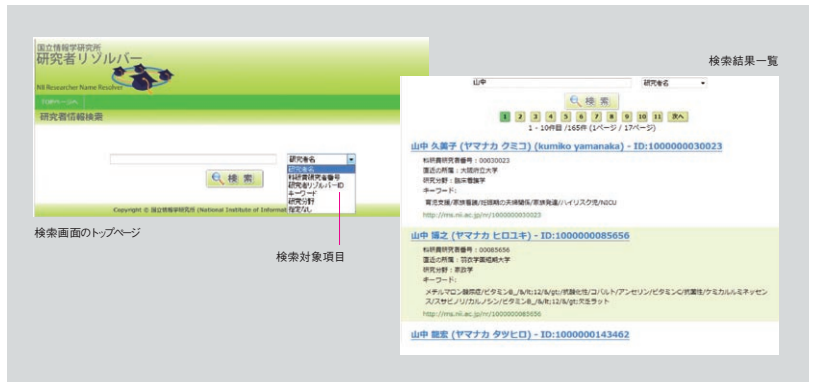


図2 「研究者リゾルバーα2.0版」

ス(KAKEN)」で、論文同定と研究者同定の機能によって、科研費のデータベースから切り出した研究者の個別の成果を集約して見ることができるようになってきている(図1)。成果報告書の文献から、学術文献のデータベースであるCiNii(*)などの文献サイトに飛ぶこともできるため、利便性が非常に高い。

もう1つが、「研究者リゾルバーα2.0版」で、さまざまなサイトにある個別の研究者の情報を集約してあり、容易にアクセスし閲覧できるサービスに仕上がっている(図2)。

これらのサービスの原型となる研究者「名寄せ技術」の基盤を構築したのが、相澤彰子教授と08年度までNIIのプロジェクト研究員だった高久雅生氏だ(08年9月から独立行政法人物質・材料研究機構=NIMS、科学情報室主任エンジニア)。論文情報は、CiNiiや他の国内外の論文データベース、著者の所属機関のサイト、著者のホームページなどから、また、研究者情報は、研究者データベース、文献情報サービスのサイト、所属機関のサイト、個人のホームページなどから収集され、情報同一性を精度よく判断しており、同姓同名の問題にも対応できるようになっている。

高久氏は異動後も、引き続き融合プロジェクトに参加しており、蔵川特任准教授とともに、科研費のデータベースと各研究機関のデータベースとをマッチングさせて、研究者それぞれに“背番号”をつけるような形で、一元的に管理することを目指した研究を続けている。蔵川特任准教授は、「情報入力機関ごとにきちんと管理してもらい、それを連携させることでより精度の高いシステムを作り上げたいですね」という。

共著関係への応用などで研究を活性化

こうした技術を応用すると、共著関係ネットワークについても、ある程度まで自動的に作成することができるようになる。研究者同士のつながりや分野間の広がり、その経時変化などを追うことができるようになり、研究コミュニティの活性化やサイエスマップづくりなどにも利用できる。このほか、NIIの新井紀子教授が開発中の研究者向けサイエンス2.0基盤サービス「Researchmap」(10-11ページ参照)にも再構築したデータを提供したいと考えている。また、この融合プロジェクトのメンバーは、外部の研究者も交えて、年1~2回の研究会を開いており、成果を報告し合っている。

プロジェクトの中心メンバーの1人である統計数理研究所統計科学技術センターの馬場康維特命教授(2008年4月より新領域融合センター特任教授兼務)は、「具体的な成果として、情報を見せる仕組みづくりはNIIが得意で、統計学はそれらのデータを分析する手法を多く持っています。将来は、こうしたコラボレーションの成果を研究の費用対効果の分析などにも応用していきたい」と抱負を語る。

リーダーの相澤教授は、5年目の総括として、「今までは利用価値が低かったデータもリネージュさせることで価値が高められるようになりました。これまで得られた成果は汎用的な技術です。今後はデータベースのウェブ上にある情報を結び付けることをテーマに、メンバーの力もかりて研究を進めていきたいと考えています」と今後の方向を示している。

(取材・構成 塚崎朝子)

(*2)2ページの(*3)を参照。