文部科学教育通信
記事掲載

○機構長インタビュー

データサイエンスの横串で新分野の創出を支援する　藤井　良一　機構長
・・・9月11日

○連載・データサイエンスによる大学との連携・協働、そして発展へ

①オープンサイエンスと協働が支える生命科学研究の新展開・・・・9月25日
②オープンサイエンスと協働が支える社会・人文学研究の新展開・・・10月23日
③統計学の活用によるビッグデータの解析と複雑現象の予測・・・・11月27日
④ビッグデータ時代に対応した高度人材育成を牽引・・・・・・・・12月25日

大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

機構長インタビュー

データサイエンスの横串で新分野の創出を支援する

大学共同利用機関法人
情報・システム研究機構
藤井 良一（ふじい りょういち） 機構長

情報・システム研究機構は平成十六年、国立大学の法人化とともに発足した四つの大学共同利用機関法人（情報・システム研究機構、高エネルギー加速器研究機構、自然科学研究機構、人間文化研究機構）の一つとして、国立極地研究所（極地研）、国立情報学研究所（情報研）、統計数理研究所（統数研）、国立遺伝学研究所（遺伝研）の四研究所を結集して設立されました。全国の大学等の研究者コミュニティと連携して、極域科学、情報学、統計数理、遺伝学についての国際水準の総合研究を推進する中核的研究機関の役割を担い、生命、地球、人間社会など、複雑な現象に関する問題を情報とシステムという視点からとらえ直すことによって、文理の枠を越えた融合的な研究を行っています。

平成二十八年度から四研究所に横串を貫く組織改革のもと「データサイエンス共同利用基盤施設」を設置して、データ共有支援、データ解析支援、データサイエンティスト育成の取り組みを強化しています。今回のインタビューでは情報・システム研究機構の改革の方向性を中心に、特色ある取り組みについてお話を伺いました。

横糸と縦糸が揃った研究機構

――情報・システム研究機構の使命と改革の方向性について伺います。

機構長　情報・システム研究機構は、わが国の四つの大学共同利用機関法人の一つであり、国立極地研究所、統計数理研究所、国立情報学研究所、国立遺伝学研究所という四つの研究所から構成されています。

それぞれの研究所は、日本学術会議等の学術コミュニティの要望により、当機構ができるずっと前から存続しているのですが、他の三つの大学共同利用機関法人と比べて、当機構の研究所の構成は、一見したところ、まとまりがないように見えるかもしれません。

――情報と統計数理は結びつきますけど、極地と遺伝はどう結びつくのか。

機構長　はい。そういう印象を持たれる方が多いのですが、まず一つは、今の大きなトレンドはビッグデータ、それを広く開放するオープンデータ、それから、それを皆さんが解析をしたりして利用していくオープンサイエンスがＡＩも含めて社会の主流となりつつあります。つまり、データは、セキュリティも含めて、社会のインフラとしては最も基本的なものになっていますし、統計数理については、データをいかに解析するのか――ビッグデータは膨大すぎて、われわれはすべてのデータを見ることはできないので、可視化し、見えるようにするしかないという時代になっています。

つまり、情報と統計数理というのは、すべての学問の基盤として、今後、最も中核をなす部分だと思います。

一方で、例えば、遺伝学はＤＮＡも含めて情報そのものです。しかも大量ですから、まさにビッグデータです。それから極地という

のは、私の専門の宇宙科学もあれば、海洋や雪氷など、地球環境についての専門的で高度な大量のデータを扱っています。それらの膨大なデータを皆さんに使っていただけるように処理をするのは、情報と統計数理の基盤があってはじめてできることなのです。

もちろん、遺伝学や極域科学だけなく、さまざまな分野で、こうした処理が求められているわけですが、最初からすべてに応用することは難しいですから、まずはパイロット的に、しっかりと使いこなせるシステムをつくり、その知見や経験を他の学問、例えば社会科学などに応用していこうというわけです。

つまり、情報・システム研究機構というのは、情報と統計数理という基盤的な横糸と、それを試すための高度な専門分野という縦糸が揃った研究機構であり、そういう意味で非常に良いコンビネーションだと思っています。

「データサイエンス共同利用基盤施設」で新たな展開を目指す

――昨年度（二〇一六）から設けられた「データサイエンス共同利用基盤施設」について伺います。

機構長　情報と統計数理による横串を通した支援を行うためのパイロット的な取り組みを進め、その成果を学問分野全体に展開していくために、四つの研究所にプラスして、昨年度から「データサイエンス共同利用基盤施設」を設けて、新たな仕組みをつくりました。

今、よく言われているように、各コミュニティで必要に応じて集めているデータというのは、その目的のためだけに価値があるわけではなくて、そのデータを他のものと組み合わせることによって、新しい意味や価値が出てくる――いわゆるオープンデータの考え方ですが――新たな展開や新しい方向性というのは、その分野を深化するだけでなく、拡大して異なる分野と接するときに出てくることが多いわけです。例えば遺伝学、極域科学、社会科学、人文学などのさまざまな分野のデータを組み合わせ、違う観点、新たな視点から見ることによって、新しい展開が出てくることが期待されます。

データを集めて統合しツールも用意して提供

機構長　当機構では各コミュニティにあるさまざまなデータを「データサイエンス共同利用基盤施設」に集めて統合し、さらにそれを利用するためのツールも用意して提供することによって、新たな展開を目指しています。それを昨年度から始めました。

データ共有支援と呼んでいますが、いろいろなデータを一緒にまとめて使えるようにするのは、言うのは簡単ですが、けっこう大変なのです。様式も項目もいろいろ違ってきますし、それを解析するためのツールについては、主に統計数理研究所が支援するわけですが、そういうものも用意しながら、全国の方々にデータを使っていただけるようにしていきます。

藤井　良一　機構長

昭和25年７月11日生
昭和52年８月　国立極地研究所助手
平成４年４月　名古屋大学助教授
　７年８月　同　教授
　17年４月　国立大学法人名古屋大学
　　太陽地球環境研究所所長
　21年４月　同　理事・副総長
　27年４月　同　教授
　28年１月　情報・システム研究機構理事（非常勤）
　　４月　同　理事
　29年４月　同　機構長

当機構が持っている高い専門性やアドバンテージを生かし、全国の大学に向けて、分野を超えた横串を通した支援を実施することにより、各大学の機能強化に少しでも貢献していきたいと考えています。

データはみんなに使ってもらう

――さらに国際的な貢献としてはいかがでしょうか。

機構長　分野にもよるのですけれど、多くのデータというのは、すでに国際的に活用されていて、共同利用といったときに、ユーザーは必ずしも日本人だけではありません。外国にも利用可能にしていくことについては、例えばアメリカのＮＡＳＡとかＮＯＡＡの人工衛星のデータなどはかなり良いシステムをつくっていて、誰でも使えるようになっています。日本はまだそこまでは行かない部分があるのですけれど、国際化はかなり進んでいて、国際的な共同利用研究も進んでいます。

データをとることもすごく重要なのですが、今は、そこからどのような成果が出るかということの方が、より重要になっていますので、昔のように自分がとってきたデータを自分で活用するだけでなく、むしろみんなに使ってもらって、より良い成果をなるべく早く出す。国際協力と国際競争――連携と競争が同時に行われているので、長い時間をかけて単独で一つの結果を出すというよりも、連携しながらできるだけ早く結果を出すというのが今のトレンドなのです。そのためには、データはなるべくみなさんに使っていただいてこそ意味がありますから、そこに乗り遅れないようにすることが重要だと思っています。

――データは〝使ってもらってなんぼ〟ということですが、知財という観点からはいかがでしょうか。

機構長　一般論として、データを利用する場合は、どこで誰にとられたデータなのかをきちんと明記します。特に論文には、例えば当機構のデータを利用した場合、このデータは情報・システム研究機構のここのデータですというクレジットがきちんと明記されます。それを証拠にして、私たちのデータはこれだけ使われていて、これだけ重要な成果が出ているということを示すことができるんですね。それによって十分なリターンがあると思っています。

効率化を図り機能強化と負担軽減を実現

――予算が削減される中で「データサイエンス共同利用基盤施設」を設けたねらいについて伺います。

機構長　一つは、別々にやるよりも、より効率的にできるということです。

それから、ご存じのように運営費交付金の一部については、私たちのところは毎年一・六％減ということで、毎年どんどん減ってきています。一方で、本部については、例えば機構長の裁量経費などは増えてきています。研究所の予算は減る一方なので、本部の増えた予算を使って機能強化を行うことが求められているわけです。そういう中で、各研究所が今まで行ってきたデータベースの共同利用について、それを一本化するというのは、機構の機能強化であるとともに、研究所の負担を減らす方向にもなっていると思います。

――ハードも日進月歩の技術の世界ですが、その辺りは予算的にはいかがですか。

機構長　それは大きな問題です。例えば、遺伝研には二八〇〇名ぐらいの共同利用の方がいるのですけど、ゲノムの配列等を見ていくためには非常に大型の計算をしなくてはいけなくて、かつデータもすごく大きいのですね。計算能力もだんだん足りなくなってきていますし、データのストレージも足りなくなってきています。新たに設備投資を考えているのですが、今後のデータの伸びなどを考慮すると、予算的には苦しいのです。そこをどうしていくのかは大きな課題です。

例えば、シーケンサという機械があるのですけれど、本当に日進月歩で性能が上がっています。ということは、何年かに一度は買い換えないといけないのですが、なかなか難しい。特に大学共同利用機関というのは、大学で持てないような高度な設備があるからこそ

意味があるので、それをみなさんに使えるようにしなければ、役目を果たせなくなってしまいます。コンピュータのファシリティも含めて、そこは最優先だと思うのですが、予算的には非常に苦しくなってきているというのが現実ですね。

人材育成を外からも見えるようにしていく

――人材育成の取り組みについて伺います。

機構長　総合研究大学院大学の基盤機関として、人材育成については、各研究所でさまざまな取り組みを進めています。

また、特にデータサイエンス人材の育成については、国でも議論をしており、大学院の博士後期課程を修了するぐらいのトップレベルの人材、いわゆる〝棟梁レベル〟の人材が毎年五〇〇人は必要だと言われています。上場企業が約四千社あって、そういうところに最低一人、二人は統計数理ができる人材が必要だということになると、毎年数名というレベルでは、とても足りないのですね。もちろん、学士課程から準備しなければなりませんし、われわれだけで養成できるわけではなくて、さまざまな機関が連携しながら養成していくわけですが、機構として最低五〇人、少なくとも一〇%は養成しようということで、いわゆる質だけでなく、量も含めて、きちんと養成する体制をつくろうと取り組んでいます。

統計数理に限らず、各研究所で専門家人材を、必ずしも学生に限らずに養成していくということは当機構の一つの大きなミッションです。他の大学院とも連携しながらさまざまな取り組みを進めています。分野によっては留学生や社会人が数多く在籍していますし、研究所ごとに海外の機関と協定を結んでいて、海外インターンを実施するなど国際交流も進めているのですが、あまり外から見えない部分もあるかもしれません。それを今年度から機構長裁量経費を使って外からも見えるようにしていこうと考えています。

サイバーセキュリティを担う人材の育成も課題

――平成二十九年度の重点的な取り組みについて伺います。

機構長　例えば、「データサイエンス共同利用基盤施設」の中に、平成二十九年度から極域環境データサイエンスセンターと人文学オープンデータ共同利用センターという二つのセンターが新たに加わりました。まずそれをきちんと充実させ、大規模データを統合して、共有できるようにすることが一つです。

それから、情報研では、ネットワークを経由したサイバー攻撃に対し、国立大学法人が迅速にインシデント対応できる体制構築の支援を行うため、平成二十八年度から「大学間連携に基づくサイバーセキュリティ体制の基盤構築」事業を実施しています。さまざまなサイバー攻撃を感知して、それを大学等にお知らせして対応するということと、サイバーセキュリティを担う人材の育成も行っていく予定です。アメリカなどと比べて予算の配分に格段の差がありますので、まだ十分に対応できているとは言えないのですけれども、重点的に取り組んでいきたいと考えています。

マスタープランの1／4に関与

――研究活動等の状況について伺います。

機構長　日本学術会議が、わが国の大型研究計画の重点課題を選定して、そのあり方などの指針を「マスタープラン2017」として策定しているのですが、選定された二八件のうち、当機構は七件に関与しています。

このうち代表機関になっているのが機構本部の「アカデミック・ビッグデータ活用研究拠点の形成」と情報研の「電子ジャーナル・バックファイル等へのアクセス基盤の整備」の二件です。このほかに、極地研一件、統数研一件、遺伝研三件と五件に参加機関として関わっていますので、当機構は重点課題の四分の一に関与していることになります。

これは、わが国の学術研究の中核的研究拠点として全国のコミュニティの要望を汲んだ研究活動を行っているということを意味し、大学共同利用機関としての役割を果たしていることの証左であり、われわれの励みになり

ます。

5年間の伸び率は76・5％

機構長　それから、今年（平成二十九年）三月に公表された「Nature Index 2017 Japan」というのがあるのですが、私たちはＴｏｐ１００研究所の中で三五位なのです。順位の高低は別として、二〇一二年から二〇一六年の五年間での伸び率を見ると、七六・八％、七位です。日本の科学力が失速していると言われている中で、非常に伸びていることがわかります。これは非常に良いことですので、さらに伸ばしていきたいと思っています。

また、予算減の中で、研究者に行くお金はどんどん減ってきていますので、今は自分の研究は競争的資金を獲得して進めるというのが基本になっています。

科研費が獲得できているということは、ある程度、研究できるだけの資金があるということを表しますので、研究活動の状況を示す重要な指標だと思っています。昨年（平成二十八年）十月に公表された科研費の獲得状況をみると、統計科学、ソフトウェア、遺伝・染色体動態の三細目で細目別採択件数が一位になっています。そのほか超高層物理学など、二〇細目で二位から一〇位に入っていますので、まずまずではないかと思っています。科研費というのは研究者のピアレビューによる、最も重要な競争的資金の一つですから、これからもさらに研究力を強化していきたいと思います。

コーディネーターが支援ニーズを掘り起こす

――共同利用の推進方策について伺います。

機構長　広報、すなわち認知度の向上がすごく重要だと思っています。私たちは共同利用機関と言っても、公募をして、応募された取り組みを採択して支援するので、どちらかというと待っているところがあるのですが、むしろもっと積極的に外に出ていこうということで、平成二十八年度から、研究コーディネーターという職の者がいろいろな学会に出向いて、当機構の活動を紹介して、支援ニーズを掘り起こすという取り組みをはじめています。

昨年は、生命科学系の学会等に行って、「自分たちはこういうことをできます」「ああいうこともできます」と……

――ピーアールしたわけですね。

機構長　はい。それによって共同利用がより深まって、広がるということで、非常に効果あることがわかりましたので重点的に取り組んでいきたいと考えています。具体的には、いろいろな分野に専門のコーディネーターを入れて共同利用を深めたり広げたりしていきたいです。

例えば、各大学でＵＲＡとして活躍している方々がいますけれども、そういった方々と同じように、基本的には研究のバックグラウンドを持った方を採用できればと思っています。将来的にはコーディネーターセンターのようなものを設けて進めていきたいと考えています。

ＩＲ資料持参でトップセールス

機構長　さらに、私や担当理事が各大学を訪問して、学長や研究担当の理事にお会いして説明するといったことも昨年からはじめました。

その際に、大学の機能強化や研究力の強化に役立てていただくために、その大学と当機構との共同利用についてＩＲ的に調べたデータを示しながら、さらにご活用いただこうということで、各大学を回っています。

昨年は三大学しか回れなかったのですが、今年は既に五大学にお邪魔しており、さらに現時点で一〇大学の予定があります。国大協の会議などで学長の先生方にお会いしてお願いしたところ、みなさん気持ちよく「ぜひ来てください」と言っていただいていますので、今後も継続していきたいと思っています。もちろん国公私を含めて全国の大学を回りたいと思っています。

――大学共同利用機関法人四機構の連携はいかがですか。

異分野融合により新分野創出を支援

機構長　さまざまな側面で連携が求められています。一つは、事務局の連携です。四つがバラバラにやるよりも、一つにできるものがあれば進めていこうということで、例えば調達、契約事務や知財対応など、いろいろ検討を進めているところです。

もう一つは、今年度から四機構が予算を出し合って、「異分野融合・新分野創出支援事業」を始めました。また、当機構においても、例えば「未来投資型プロジェクト」は、特に若手の研究者を中心に、新しい研究領域を生み出すような挑戦的な案を出していただくもので、昨年度（平成二十八年度）は一三件採択して、そのうち六件を次年度以降も継続することにしています。今年度（平成二十九年度）も新たに公募します。それから、「文理融合研究プロジェクト」は他機構の研究者との共同研究であることを要件としており、昨年度フィージビリティ・スタディーとして三件採択し、一件を本格研究として採択しました。「日本列島人の深化とその言語文化の起源」という遺伝研、人間文化研究機構、科学博物館、九州大学の連携による研究で、日本人のゲノム史と日本語の歴史を解明しようというものです。このように新たな分野を模索できる可能性があるものを積極的に進めていきたいと思っています。

よく話し合えるシステムをつくる

――機構長のリーダーシップと機構のガバナンスについてお考えを伺います。

機構長　やはり各研究所をいかに活性化させるかが最も重要だと考えています。そして、それをいかにうまく統合していくか――もちろん各研究所の自主性、自律性を尊重しつつ、大学共同利用機関としてのミッションをしっかり果たせるように担保していくことが重要だと思っています。特に研究所個別ではなかなか対応できないことが出てくる時代になっているので、それを俯瞰して、きちんと発展させるのが大きな仕事だと思っています。

各研究所はそれぞれの歴史があり、別々の個性を持っていますから、その中で意思を共有していく難しさはあります。一つの試みとしては、今までは月一回程度だった役員と所長の懇談会を毎週開くようにしました。毎週のことなので、テレビ会議によるランチミーティングというかたちで、各研究所の意見や希望を聞き、ざっくばらんに話し合いをしています。最終的には機構本部が決めるとしても、しっかりと意見を聞いて、希望を十分に採り入れないと改革の意味がありませんので、スピード感を持って重要な案件について、よく話し合えるようなシステムをつくりたいと思っています。

それから、よりガバナンスを強化すべく専任理事を新たに二名置きました。理事職というのは専任でないととてもできませんので、国立大学で部局長経験のある方に専任で来ていただき、研究戦略等の担当としました。さらに事務局長を理事に任命し、より強い権限で事務業務をまとめてもらっています。

戦略企画本部を設置

機構長　基本的な戦略については、戦略企画本部を設けて、各研究所と意見交換しながら、機構全体として一つの案をつくって、決定するようにしています。戦略企画本部には、各研究所の副所長級の方に入っていただいて、研究所としての意見や問題点を出していただくと同時に、各研究所の所員に戦略企画本部の案を説明するという二つの役目を担ってもらっています。

昨年からはじめたばかりなので、まだ道半ばというところですが、お互いに、取り組んでいることやどういう進捗になりそうかということも基本的に理解しながら議論できるようになってきています。研究所と一体感を持ちながら運営していくためには双方向の意見交換が大事です。本部と研究所の横串を通すかたちで企画立案体制を整えたので、今後より質の高い戦略が企画できるのではないかと期待しています。

新連載　情報・システム研究機構の支援事業

データサイエンスによる大学との連携・協働、そして発展へ①

オープンサイエンスと協働が支える生命科学研究の新展開

藤山秋佐夫　情報・システム研究機構　データサイエンス共同利用基盤施設長
小原　雄治　同施設ライフサイエンス統合データベースセンター長
箕輪　真理　同センター特任准教授
野口　英樹　同施設ゲノムデータ解析支援センター長
馬場　知哉　同施設データサイエンスコーディネータ

はじめに

大学共同利用機関法人情報・システム研究機構では、データサイエンスを合い言葉に大学との連携・協働を強化する目的で平成二十八年四月にデータサイエンス共同利用基盤施設（DS施設）を設置しました。DS施設では、生命科学、環境、データ融合計算から人文・社会科学に至る多様なデータサイエンスを推進するため、本機構から選りすぐった研究者を大きく六つのセンターに組織しました。このシリーズでは当初の三回でDS施設の全体像と各センターの活動内容を概説し、最終回では本機構が重要視しているデータサイエンス人材育成についての取り組みを紹介します。

さて、一口にデータといっても様々で、私たちが研究活動の中で扱うものだけでも、ゲノムや遺伝子に関する生命情報データから、国勢調査のような社会データまで広い範囲に及びます。ここでは、学術研究に限らず広い範囲のデータを科学的に取り扱い、知識として使えるようにする学問分野を広くデータサイエンスと呼ぶことにします。

今回は、ライフサイエンス統合データベースセンターとゲノムデータ解析支援センターを紹介します。

ライフサイエンス統合データベースセンター

ライフサイエンス統合データベースセンター（DBCLS）は、平成十九年四月に、内閣府総合科学技術会議（当時）の議論を受けて始まった文部科学省「統合データベースプロジェクト」の中核組織として設立され、所在や整理の仕方がばらばらで使いにくかった生命科学のデータベース（DB）全体の把握と整理を行い、DBの利用を容易にしてきました。平成二十三年度以降は、DBをより高度かつ柔軟に活用するために、科学技術振興機構（JST）のバイオサイエンスデータベースセンター（NBDC）と連携した活動を進めています。当センターは基盤技術開発を担当しており、ウェブ上に分散するDBをあたかも一つの巨大なDBの様に使いこなすための技術や、解析技術の進展に伴って日々大量に生産される分子データを活用する技術の開発を進めています。

当センターの組織体制は、センター長と専任教授各一名のほか、様々なバックグラウンドを持つプロジェクト研究員一六名および学術支援技術専門員一名が、モデル生物ゲノム

解析、トランスクリプトーム解析、ＲＮＡ生物学、プロテオーム解析、糖鎖生物学、自然言語処理、テキストマイニング、計算機科学、知識工学、ゲノムアノテーション等の分野に従事しています。

ライフサイエンスは扱う生物種がさまざまで、解析の対象も遺伝子（ゲノム）やタンパク質等の分子から細胞や個体まで、多岐にわたっています。それらのデータや情報を記述する方法も分野や対象ごとに異なるため、一つの大きなＤＢを構築するということは困難で、研究者が情報解析をする際には目的に応じたデータ形式の整理やデータの関連づけに多くの時間を費やさざるを得ませんでした。そこでＤＢＣＬＳでは、セマンティック・ウェブ技術を応用し、分野や対象の異なるデータを一括して扱うことができる情報環境の構築を目指しています。具体的には、Resource Description Framework（ＲＤＦ）というデータ記述方式を採用し、すべてのデータを「主語（Ｓ）―述語（Ｖ）―目的語（Ｏ）」という簡単な形で記述することによって、ＳとＯの関係性（データの意味）を表現するようにしています。ＳＶＯのそれぞれに整理された用語を使用することで、別々のデータが持つ共通項の重ね合わせが可能となり、複数のＤＢのデータを合わせた巨大なネットワークを構築することができるのです。これを実現するために、ＤＢＣＬＳでは国際バイオハッカソンを開催し、分野ごとの用語の整理、データ記述のためのガイドライン作成などの国際標準化を進めています。また、これに準じてデータの変換を行うためには、データ作成機関の協力が必須です。このため、ワークショップや講習会を通じて技術移転や協力関係の構築に努めています。一方、ＲＤＦデータの活用法も同時に開発しており、一般ユーザーがデータの構造を意識しなくても、多種多様な情報を活用できるツールやサービスの提供を行っています。

ゲノムデータ解析支援センター

ゲノムデータ解析支援センターはＤＳ施設発足と同時に設立され、ＤＳ施設が推進するデータサイエンス（データ駆動型サイエンス）の支援事業の一環として、ゲノム関連データの解析支援を担当しています。最新のバイオインフォマティクス技術を駆使し、大学等の研究機関の研究者が解読したゲノムデータから生物学的に重要な情報を抽出する解析支援です。

近年の次世代シーケンシング（ＮＧＳ）技術の急速な発展によりＤＮＡシーケンサーのデータ生産能力は飛躍的に向上しました。生命科学の幅広い分野（生物学、医学、薬学、工学、農学、環境学など）で、多くの研究者があらゆる生物種を対象に塩基配列レベルでの多様な解析に取り組めるようになっています。例えば、次のような解析です。

- ゲノム配列が決定されていない生物種のゲノムを決定する「新規ゲノムシーケンス解析」
- ゲノム配列が決定された生物種における個体（系統）間でのゲノム変異を調べる「ゲノム再シーケンス解析」
- ゲノム上で発現している遺伝子を調べる「トランスクリプトーム解析」
- 環境中からゲノム情報を抽出する「メタゲノム解析」

しかし、ＤＮＡシーケンサーが出力するデータは断片的な塩基配列データであり、しかもそのデータ量は一個体分でもゲノムサイズ（例えばヒトでは三〇億塩基）の何十倍、何百倍に達するほど膨大なものになります。このような配列データを効率的に解析し、目的に応じた結果を正しく得るためには、生物学の知識に加えて専門的なバイオインフォマティクスの知識と技術が必要不可欠です。

当センターでは、平成二十八年度に哺乳類、魚類、昆虫、植物、菌類から原核生物まで幅広い生物種について一三課題の解析支援を受け付けました。また、五つの学会（日本分子生物学会、日本ゲノム微生物学会、日本農芸化学会、日本薬学会、日本応用動物昆虫学会）の大会で当センターを中心としたＤＳ施設の活動を紹介するとともに、共同研究に至る前の相談対応を行い、研究者の方々が求めるような解析を行うためにはどのようなデータが必要で、どういった解析手法を取るのかといった技術的なアドバイスも行っています。また、解析パイプラインの構築や高速・省メモリの解析アルゴリズムの開発による解析支援の効率化や、解析講習会の開催等による人材育成にも取り組んでいく方針です。

当センターはＤＳ施設のライフサイエンス統合データベースセンター（ＤＢＣＬＳ）をはじめとした各センターや国立遺伝学研究所の先端ゲノミクス推進センター及びＤＤＢＪ（DNA Data Bank of Japan）とも連携し、わが国の生命科学とデータサイエンスの推進に貢献します。

連載　情報・システム研究機構の支援事業

データサイエンスによる大学との連携・協働、そして発展へ②

オープンサイエンスと協働が支える社会・人文学研究の新展開

藤山秋佐夫　情報・システム研究機構　データサイエンス共同利用基盤施設長
吉野　諒三　同施設社会データ構造センター長
山下　智志　同センター教授（兼）
越前　功　同センター教授（兼）
北本　朝展　同施設人文学オープンデータ共同利用センター長

大学共同利用機関法人情報・システム研究機構では、データサイエンスを合い言葉に大学との連携・協働を強化する目的で平成二十八年四月にデータサイエンス共同利用基盤施設（DS施設）を設置しました。DS施設の組織構成については**図**をご覧下さい。シリーズ一回目の前回（本誌№420）はライフサイエンス統合データベースセンターとゲノムデータ解析支援センターを紹介しましたが、今回は社会データ構造化センターと人文学オープンデータ共同利用センターを紹介します。

社会データ構造化センター

本センターの主要な任務は、社会の状況を反映するデータの有効利用を進め、実証的なデータにもとづく学術研究の発展、政策立案、産業振興等を支援することです。そのため当センターでは、(1)社会調査関連事業、(2)公的ミクロデータベース事業、および(3)ソーシャル・ビッグデータ事業が一体となって活動しています。

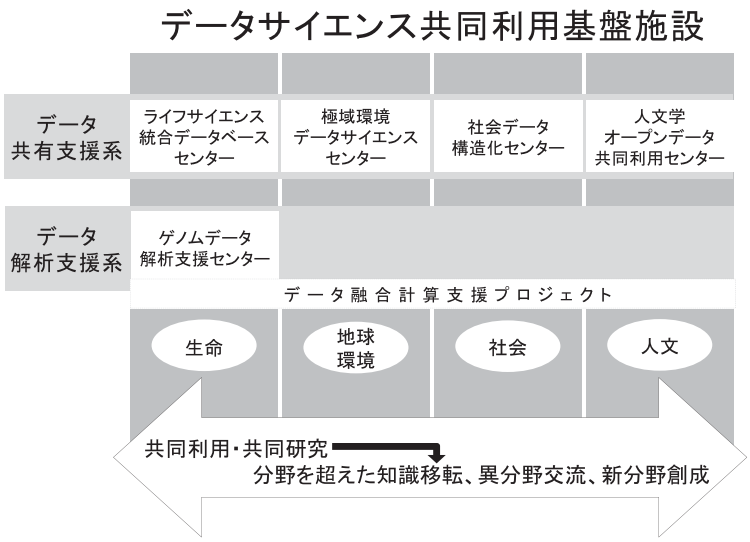
(1)社会調査関連事業では、既存の社会調査データを公開するために必要な整備の実施、全国の大学等を組織した社会調査共同ネットワークの形成、新たな調査データや関連情報の収集と調査報告書の整備等を進めています。これまでに公開したデータには、統計数理研究所が六〇年以上にわたり実施してきた「日本人の国民性」調査や「意識の国際比較」調査が含まれています。調査個票レベルのデータについては個人情報や所有権等に関する倫理的・法律的・社会的な諸問題の解決が必要なため、これらを検討しながらDS施設の共同研究等を通じて公開を進めています。

(2)公的ミクロデータベース事業では、政府や自治体が実施した統計調査データの二次的利用を推進しており、それに必要な基盤整備とデータ構造化の方法論を構築するとともに、学会・シンポジウム等の機会を通じて活動の成果を発信しています。また、本センターは総務省によるオンサイト・ネットワークの接続実験に参加しており、立川地区に新設されたデータサイエンス棟にオンサイト拠点を開設しました。（公財）統計情報研究開発センターや、アジア各国の政府統計関係者とも連携協力を進めており、国際政府ミクロ統計データベースの拡充やワーショップを通じてデータベース提供国の教育や統計技術への理解の増進を図っています。このほか、アジア人材育成事業としてセミナー等も開催し、国際研究ネットワークの形成と人材育成を進めてい

ます。

(3)ソーシャル・ビッグデータ事業では、公共サービスや公益性を有するデータをもとに「地（知）の拠点大学とのネットワーク型データ連携基盤」を構築し、さらに地域経済の活性化や雇用機会の創出という社会問題の解決支援を目指しています。そのため大学、地方公共団体、企業等と協働し、地域の施設稼働率、交通システム利用率、イベント集客率などのデータを集積したソーシャル・ビッグデータの構造化事業を進め、人々の交通移動データから観光、防災政策を立案するための情報基盤整備を行っています。

さらに上記三事業の連携による「人間・社会データ・コンプライアンス管理プラットフォーム」の整備を進めています。急速に発展しているＡＩやビッグデータの産業活用が期待されている時代ですが、他方で個人情報の保護、情報の所有権等の諸問題に関して法的整備が追い付いていない危惧があるためです。われわれはこの問題に対して慎重に検討を進め、社会に建設的な提言を行うことを目指しています。



人文学オープンデータ共同利用センター（ＣＯＤＨ）

本センターでは、人文学分野におけるデータのオープン化と共同利用の推進を目的とした研究・支援活動を進めています。しかし、人文学研究コミュニティでは、大規模でオープンなデータを基盤とした研究（データサイエンス）は未だに発展途上であり、単にデータをオープン化するだけでは共同利用が拡大する状況にはありません。そこで本センターでは世界的に発展しつつあるデジタル・ヒューマニティーズ（人文情報学）の方法論を取り入れてデータベースやツールを開発・公開するとともに、こうした研究資源の利活用を推進するためのセミナーやチュートリアルを開催し、この分野へのデータサイエンスの普及に努めています。

本センターの活動は、人文学者だけでなく情報学者（＋機械）や市民との協働を視野に入れている点が特徴です。近年の機械学習（人工知能）技術の発展に伴い、従来は人間が行ってきたが今後は機械に任せたほうがよい作業が増えつつあります。機械に仕事を任せるためには、まず機械が仕事の内容を学習するためのデータが必要ですが、これに使えるオープンデータが大幅に不足している状況が人文学分野に人工知能を導入する際の障害となっています。研究者と市民が互いに学び協創できる情報基盤と大規模な学習データの構築が、本センターの重要な課題です。

次に、データ公開の先行事例として、国文学研究資料館が中心となって推進する「歴史的典籍ＮＷ事業」と本センターとが協力して公開したオープンデータを紹介します。

まず「日本古典籍データセット」では、古典籍七〇一点の画像データをダウンロード可能な形式で提供しています。この古典籍データの各々にＤＯＩ（デジタルオブジェクト識別子）を付与することにより、同一タイトルの古典籍が複数個存在しても画像データを特定できるようになりました。次に「日本古典籍字形データセット」では、くずし字を対象とした文字のデータセットとして三九九九文字種、四〇万三二四二文字のデータを公開しています。これは機械学習を用いた文字認識やテキスト化に向けたアルゴリズム開発にも利用可能な形式となっており、「人工知能はくずし字を読めるか？」を課題とした「くずし字チャレンジ！」コンテストを開催中です。最後に江戸時代の料理本『卵百珍』から現代でも調理可能なレシピを作成し「江戸料理レシピデータセット」として公開しました。このレシピを日本最大のレシピサービスである「クックパッド」から公開したところ、市民から大きな反響が得ることができました。

本センターは、人文学分野を対象とした多様なデータセットの公開とそれを利用するためのツール開発を進め、さらにＣＯＤＨセミナーなどの開催を通じて人文学データの普及啓発活動にも取り組みます。

連載　情報・システム研究機構の支援事業

データサイエンスによる大学との連携・協働、そして発展へ③

統計学の活用によるビッグデータの解析と複雑現象の予測

情報・システム研究機構　データサイエンス共同利用基盤施設長　藤山秋佐夫
同施設　極域環境データサイエンスセンター長　門倉　昭
同施設　データ融合計算プロジェクト長　中野　慎也

はじめに

大学共同利用機関法人情報・システム研究機構では、データサイエンスを合い言葉に大学との連携・協働を強化する目的で平成二十八年四月にデータサイエンス共同利用基盤施設（DS施設）を設置しました。本シリーズ二回目で紹介したDS施設の組織のうち、今回は極域環境データサイエンスセンターとデータ融合計算プロジェクトの活動を説明します。

極域環境データサイエンスセンター

本センターは、極域科学のデータサイエンス活動の中心となることを目指しており、国立極地研究所（極地研）が所有する、南極域・北極域から得られた貴重なデータの公開と共同利用を進めることで、地球環境研究の推進に貢献します。

極地研では四つの研究グループ（宙空圏、気水圏、地圏、生物圏）が国内外の大学や諸機関の研究者と共同研究を行っています。それぞれ、オーロラや超高層大気、大気や海洋・雪氷、地質・地形や地震・重力、陸上生物や海洋生物、といった分野で行われている多種多様な研究・観測活動から得られたデータが様々な形で蓄積されています。それらを大きく二つに分類すると、一定時間ごとの観測から得られる連続観測データ（時系列データ）と、ある特定の場所・時期に採取された試料（試料系データ）に分けられます。後者には、岩石、隕石、アイスコア（氷床を掘削して得られる筒状の氷の柱のこと。南極では三〇〇〇メートル程度の深度まで掘削が行われています。）、海水、大気などが含まれており、標本そのものに関するデータの他に、様々な解析・分析を施した結果も高次処理データとして蓄えられています。

こうした研究データは分野別、種類別に個別のデータベースが作られており、それぞれが共同研究に供されていますが、データごとにデータベース化の程度やデータ公開の進み方にばらつきのあるのが現状です。また、分野を横断した検索や利用を行うためには、極域科学データ全体を俯瞰出来るような総合的な仕組み（統合データベース化）も必要です。

こうした問題に対処するため、本センターでは、極域環境データに関する統合データベースを構築し個別データのデータベース化やオープン化を進めています。極域科学分野におけるデータ中心科学（データサイエンス）の推進と、国内外の研究コミュニティとの連携をさらに発展させることが本センターの主な目標です。

平成二十九年度における本センターの体制

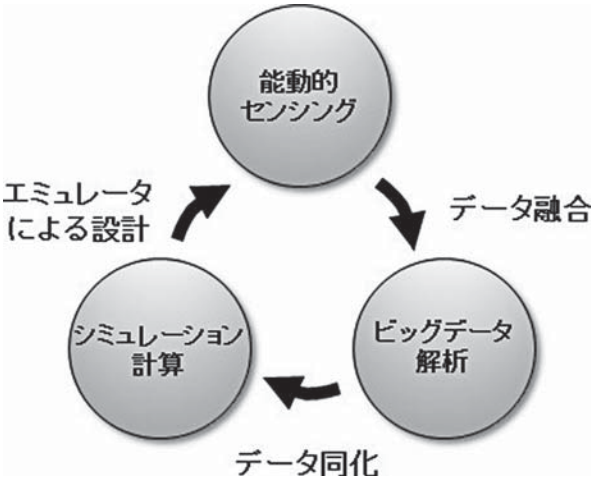
は、センター長の他に専任の准教授一名、特任准教授三名という構成になっており、極域科学学術メタデータベースシステム、北極データアーカイブシステム（ＡＤＳ）、大学間連携超高層大気データシステム（ＩＵＧＯＮＥＴ）、昭和基地大型大気レーダー（ＰＡＮＳＹ）データアーカイブシステムなどの運用を担っています。これらのデータベースシステムの充実化や拡張、相互運用化も、本センターの事業に含まれています。また、極域研究に関連するデータベースの利用法やデータ解析方法についての講習会などを通じ、大学や諸機関の学生や研究者に対する教育・研究面での支援を行うことも、本センターの大事な任務の一つです。

データ融合計算支援プロジェクト

本プロジェクトでは、数値シミュレーションを効果的に利用する統計的手法の研究・開発と、統計学を活用した数値シミュレーションを実用的な課題に応用するための支援を行います。数値シミュレーションとは、複雑な現象をコンピュータ上で再現する研究手法です。コンピュータの性能の向上とともにシミュレーションの精度も向上しており、最近では現実と区別がつかないような高精度シミュレーションも珍しくはなくなりました。しかし、シミュレーションは基本的に与えられた条件設定に対して結果を予測するものです。シミュレーションを現実の問題に適用するためには、条件設定を適切に与えることが重要です。この作業を効率的、効果的に行うために、統計的な手法が活用できるのです。

シミュレーションの応用例のひとつが天気予報です。シミュレーションで明日の天気を精度よく予測するには、今日の気象の状態を詳細に設定する必要があります。各地の気象データから、ある程度は気象の状態を把握できるのですが、当たる天気予報をするためには、もっと詳しい情報をコンピュータに与える必要があるのです。そこで実際の天気予報では、その日のデータだけでなく、過去数日間のデータにも合うようにシミュレーションの条件を調整する「データ同化」という作業が行われます。この作業には大規模な統計処理が用いられており、それによって「これからの気象の変化」を精度よく予測できるようになります。

一方、機械や構造物などのシステム設計においては、「これからの気象の変化」のようなシナリオ予測よりも、様々な条件設定に対してどのような結果が得られるかを知りたいという場合があります。例えば、地震による建築物への影響や自動車の設計などです。シミュレーションを繰り返せば様々な条件に対する計算結果は得られますが、実用的なシミュレーションには最新のスーパーコンピュータを使っても数時間、あるいは数日かかることがあり、試行できる回数にも限度があります。そこで、シミュレーションの挙動を模倣するモデルを統計的に作成し、様々な条件に対する結果を高速に予測する方法が考えられるようになりました。このようなシミュレーションを模倣するモデルを統計的エミュレータ（あるいは単にエミュレータ）と呼んでいます。エミュレータは統計的手法によって構築されるため、予測の信頼性を評価する目的にも使うことができ、システムの設計にも役立ちます。



データ融合計算支援の概念図

本プロジェクトでは、データ同化やエミュレータなどの手法を活用する研究相談を受け付けており、平成二十八年度のプロジェクト開始以来、すでに一〇件程度の相談を受けています。相談の中には、大学等の研究者だけでなく、民間企業からの相談もあり、共同研究に発展して現在進行中のものもあります。また、手法の普及のために、講習会や体験学習（ハンズオン）を随時開催しており、実習を通じて、データ同化などの基本的な考え方や現実の問題への活用方法を勉強する機会を提供しています。

連載　情報・システム研究機構の支援事業

データサイエンスによる大学との連携・協働、そして発展へ④

ビッグデータ時代に対応した高度人材育成を牽引

情報・システム研究機構　統計数理研究所　所長　樋口　知之
同所　統計思考院　特任准教授　神谷　直樹

はじめに

大学共同利用機関法人情報・システム研究機構では、機構内の各研究所が総合研究大学院大学の基盤機関として人材育成に取り組んできました。一方で、データサイエンス人材の育成については、大学院博士後期課程修了レベルの人材、いわゆる「棟梁レベル」人材育成が喫緊の課題として、国レベルでも議論されています。当機構の「ビッグデータの利活用に係る専門人材育成に向けた産学官懇談会」報告書を経て、文部科学省「数理及びデータサイエンス教育の強化に関する懇談会」による報告書『大学の数理・データサイエンス教育強化方策について』がまとめられ、数理及びデータサイエンスに係る教育強化事業の拠点校が選定されました。この事業では、拠点校の特色に応じて、数理・データサイエンス分野の専門能力を向上させつつ、他分野へ応用展開が可能な教育を目指しています。

今回は、情報・システム研究機構が取り組んでいる棟梁レベル人材育成プログラムを中心に紹介します。当機構では、データサイエンスを合い言葉に大学との連携・協働を強化する目的でデータサイエンス共同利用基盤施設を設置し（本シリーズ一～三回目で紹介）、データ共有支援、データ解析支援、データサイエンティスト育成に取り組んでいますが、特に棟梁レベル人材の育成体制をきちんと整えるために「データサイエンス高度人材育成プログラム」を開始しました。

データサイエンス高度人材育成プログラム

情報・システム研究機構には、統計数理と情報という分野横断的な横糸と、それが適用される高度な専門分野である極域科学と遺伝学という縦糸が揃っています。本プログラムではこの特徴を生かして棟梁レベル人材を育成します。各プログラムは、Aタイプ（講義などを活用した教材作成）とBタイプ（演習・コンペティション）の二種類に大きく分かれます。

⑴ライフサイエンスデータベース統合およびデータサイエンス応用を担う人材養成ハッカソンシリーズ（Bタイプ）

ライフサイエンス統合データベースセンターは、国内外で開発されてきた数万件以上あるといわれる生命科学・医科学の公共データベースを統合的に利用するための技術開発を進めてきました。さらに技術開発を進めるとともに、この統合データベースを実問題へ応用できる人材を育成します。科学技術振興機構バイオサイエンスデータベースセンターと連携

し、これまでに行ってきた五つのハッカソン（国際版バイオハッカソン、国内版バイオハッカソン、SPARQLthon、RDFサミット、Biomedical Linked Annotation Hackathon）を発展させて、ハンズオン形式のチュートリアルを付随させることにより、ライフサイエンス分野における人材育成を加速します。

⑵Global and Polar Data Science Schoolによる研究力強化と大学への貢献（Bタイプ）

国立極地研究所は、極域観測データを中心に地球環境データを多分野にわたって蓄積してきました。これらは、長期的な地球環境変動の要因を分析し、地球の未来を予測する上で貴重なデータです。そこで同所と極域環境データサイエンスセンター（本シリーズ三回目、十一月二十七日号で紹介）は、九機関（名古屋大学、京都大学、九州大学、東北大学、宇宙航空研究開発機構、国立天文台、豪州南極局、韓国極地研究所、インド国立南極海洋研究センター）と連携して、国内外の研究者を対象とした人材育成に取り組みます。極域科学に関する専門家が、データ利用の個別指導スクールや、特定課題に対して集中的にデータスキルを向上させる講習会で講師を務めます。また国外研究者の育成のため、国際インターンシップ制度と連携して国外の研究者を二週間から九〇日間招聘します。

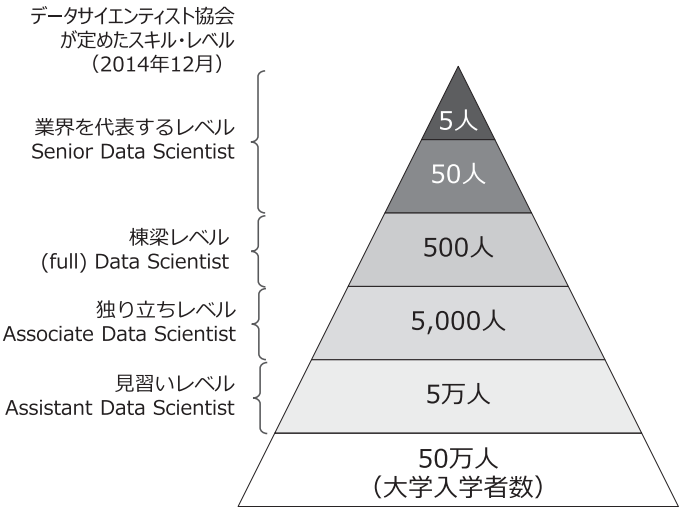


図1　人材育成レベルと毎年の育成目標人数

⑶大学研究者のための研究データ管理に関するリテラシ教育環境の構築（Aタイプ）

国立情報学研究所は大学図書館等と協力して、図書館職員向けの基礎的な研究データ管理教育用コンテンツ開発を行ってきました。今年度からは研究者を含む幅広い大学教職員向けの教育コンテンツ開発を行い、研究データ管理リテラシ教育環境（大規模公開オンライン講座（Massive Open Online Courses：MOOCs））を構築・提供していきます。そして利用者コミュニティの育成を通して、文部科学省科学技術・学術審議会からその不足が指摘されている、研究データ管理の適切なリテラシを有する研究者や研究支援職員の育成を促進します。

⑷リーディングDAT（Data Analytics Talents）プログラムによる棟梁レベルデータサイエンティスト育成（Aタイプ）

統計数理研究所は、約半世紀に渡って、統計数理を学びたい一般の方々を対象とした公開講座を実施してきました。当機構の「ビッグデータ利活用に係る専門人材育成に向けた産学官懇談会」で指摘されている人材輩出エコシステムを実現するために、従来の公開講座をベースにした、棟梁レベル人材輩出のためのリーディングDATを開始します。このプログラムは「リーディングDAT講座」と「リーディングDAT養成コース」から構成されます。リーディングDAT講座では、棟梁レベルを目指すデータサイエンティストに必須の統計数理の知識を効率的に習得することができます。リーディングDAT養成コースでは、リーディングDAT講座に加えて実践的な問題の演習や特別講演などを含めた集中的なトレーニングが行われます。リーディングDAT養成コースを完了すると修了認定証が交付されます。これらの系統的学習プログラムは、順次オンライン教材化していくことが予定されています。

⑸大量ゲノムのためのバイオインフォマティクス講習プログラム（Bタイプ）

国立遺伝学研究所は、次世代ゲノムシーケンシング技術の急速な発展に伴うデータ解析需要に対応できる人材を育成します。同所内で運営されているDDBJ（DNA Data Bank of Japan）センター内に生命データアナリスト育成ユニットを設け、ゲノムデー

タ解析支援センター（本シリーズ一回目、九月二十五日号で紹介）と連携して大量ゲノム情報を扱える技術者を育成します。現在、データ解析講習会等を実施して実践的な人材育成に取り組んでいます。講習会の資料と内容はライフサイエンス統合データベースセンターと連携してインターネット上でも公開し、全国規模でゲノムシーケンスを扱える人材の育成を目指しています。

プログラムタイプ	Aタイプ 講義などを活用した教材作成	Bタイプ 演習・コンペティション
取組名称	• 大学研究者のための研究データ管理に関するリテラシ教育環境の構築 • リーディングDAT（Data Analytics Talents）プログラムによる棟梁レベルデータサイエンティスト育成	• ライフサイエンスデータベース統合およびデータサイエンス応用を担う人材養成ハッカソンシリーズ • Global and Polar Data Science Schoolによる研究力強化と大学への貢献 • 大量ゲノムのためのバイオインフォマティクス講習プログラム

情報・システム研究機構の特徴を生かした棟梁レベル人材育成体制の構築

図2　データサイエンス高度人材育成プログラム

その他の特徴的な人材育成プログラム

⑴トップエスイープロジェクト「サイエンスによる知的ものづくり教育プログラム」

国立情報学研究所は、「自ら課題を発見し、新しい技術の意義や限界を理解した上で、課題解決のためにその技術を活用できる」スーパーアーキテクト育成を目指しています。このプロジェクトでは社会人を対象とした二コース（トップエスイーコースとアドバンス・トップエスイーコース）が設置されており、修了するとトップエスイー、あるいはアドバンス・トップエスイーとして認定されます。トップエスイーコースでは演習を重視した講義と実課題に取り組むソフトウェア開発実践演習を通じて、ソフトウェア工学の基礎技術を学習することができます。アドバンス・トップエスイーコースでは、最先端の技術を駆使し、難度の高い先端課題を解決する人材を育成します。アドバンス・トップエスイーコースには受講者個人で取り組む「プロフェッショナルスタディ」と、複数の受講者で取り組む「最先端ソフトウェア工学ゼミ」があります。

⑵大学間連携に基づくサイバーセキュリティ体制の基盤構築

国立情報学研究所は、ネットワークを経由したサイバー攻撃に対し国立大学法人等が迅速に対応できる体制構築の支援を通して、情報セキュリティを担当する技術職員を育成しています。実環境および実習環境を用いた研修を通じて実践的な人材育成を行っています。

⑶北極域研究推進プロジェクト（ArCS）

国立極地研究所は、北極の自然科学研究と人文・社会科学研究の融合を目指し、海洋研究開発機構と北海道大学とともにこのプロジェクトを実施しています。このプロジェクトでは、若手研究者を北極に関する研究を行う海外の研究機関に派遣し、技術・知識の習得や人的ネットワークの構築を踏まえた人材育成をしています。人材育成以外にも、北極関連会合への専門家派遣、国際連携拠点の整備、国際共同研究推進に取り組んでいます。

⑷データサイエンティスト育成事業

データサイエンティストに求められるスキルセットは多岐にわたり、現実的な人材として国際的にはT型データサイエンティスト（幅広い知識と任意のドメインの専門知識を持つ人材）が提案されています。統計数理研究所は、「共同研究スタートアップ」、「公募型人材育成事業」、「特別共同利用研究員制度」、「データサイエンス・リサーチプラザ」などの取り組みを通して、データサイエンスを横糸とするT型人材の育成に取り組んできました。そして、データを活用したモデリングや研究コーディネーションなど、ビッグデータ時代に求められている統計数理の知識とスキルを持った人材を育成できるよう研究・教育活動を行ってきました。この他にも棟梁レベル育成を目的としたデータサイエンティスト育成コースや、大学や企業との「組織連携に基づくDS講座企画」などの新たな取り組みを通して、人材輩出エコシステムを構築しています。