社会基盤を支える第四の科学

# 「データサイエンス」という未来

## ～データサイエンス共同利用基盤施設の紹介～

大学共同利用機関法人　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
Joint Support-Center for Data Science Research

# 情報・システム研究機構について

国立極地研究所
大学共同利用機関法人 情報・システム研究機構

■南極観測事業による共同研究（昭和基地）
■北極域における共同利用（ニーオルスン基地）
■試料解析装置や試料の提供
（アイスコア、隕石など）

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

■超高速学術情報ネットワーク（SINET5）
■情報セキュリティ体制支援
■オープンサイエンス推進基盤構築

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

大学共同利用機関法人 情報・システム研究機構 統計数理研究所 The Institute of Statistical Mathematics

■NOE（Network Of Excellence）形成事業
■スーパーコンピュータ“AIC”
■統計思考力育成（統計思考院）

大学共同利用機関法人 情報・システム研究機構 国立遺伝学研究所

■遺伝子改変＆モデル生物資源
（マウス、線虫、イネなど）
■遺伝情報の解読・利用支援（DDBJ センター）
■遺伝研スーパーコンピュータ

大学共同利用機関法人 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
Joint Support-Center for Data Science Research (DS)

■データ共有支援
■データ解析支援
■データサイエンスのための人材育成

**施設長** P.3

**運営会議**

**データサイエンス推進室**（プロジェクト事業推進のマネージメントを担当）

### データ共有支援

**ライフサイエンス統合データベースセンター** P.4
生命科学分野のオープンサイエンス推進、ライフサイエンス・データベース統合化のための研究開発を推進
（JST/NBDC との共同研究事業）

**極域環境データサイエンスセンター** P.5
過去から現在に至る長大な時間軸を持った極域環境変動・地球システム変動に関する貴重なデータと、その分析・解析支援を提供する共同利用を推進

**社会データ構造化センター** P.6
大学研究者のための社会調査データ、公的統計ミクロデータ、ソーシャルビッグデータに関するデータベースを整備。また、データ利用コミュニティを形成し、環境、治安、経済を含む各種の社会的課題の解決のための実証的研究を促進

**人文学オープンデータ共同利用センター** P.7
データサイエンスに基づく人文学（人文情報学）という新たな学問分野を創生するとともに、データを中心としたオープン化を推進することで、組織の枠を超えた研究拠点を形成・強化

### データ解析支援

**ゲノムデータ解析支援センター** P.8
最先端のバイオインフォマティクス技術を駆使して大量のゲノム・トランスクリプトームデータから生物学的に重要な情報を得るためのデータ解析支援

**データ融合計算支援プロジェクト** P.9
データ融合計算技術による諸科学・産業界での課題解決

以降のページ（P.3～）は、平成 29 年 2 月 20 日（月）に開催された機構シンポジウムの講演内容の要約です。

# 「データサイエンス共同利用基盤施設の取組み」

データサイエンス共同利用基盤施設／施設長
藤山 秋佐夫

## データサイエンス共同利用基盤施設設置の目的

情報・システム研究機構には、国立情報学研究所があり、国立遺伝学研究所があり、国立極地研究所があり、統計数理研究所があって、世の中の通常の常識で考えると、一緒になっても何も起こらないと思われる研究所の集まりなんですね。それらを集めて混ぜて、そこからさらに新しいものを作ろうとする努力が法人第一期、第二期に渡って行われ、いろいろと新しい成果も出てきたと思います。

現在の第三期では、データサイエンス推進という形で次のステップを目指している状況でございます。

じゃあ、こういう活動をして結局何ができたかというと、大学共同利用機関としての成熟度がかなり増大したということです。もともと普通においておくとなかなか混ざりにくい4つの研究所を強引に混ぜて、いろいろなことをやった結果として、共同研究体制がかなりできあがってきたということです。

特に第二期の後半あたりから、実際にわたくしが関係しているゲノム科学の世界ですと、データの生産量というのは飛躍的に増大しましたので、データ駆動型サイエンス、データサイエンスというのを生物学の世界でも実現できました。もちろん地球、物理、大気のデータ観測などでは、それ以前からビッグデータ時代に突入した科学が進んできたわけでございます。

## 大量・大規模データ共有のための仕組みを作る

データサイエンスを推進するためには、大規模なデータを生産するということが必要です。一言に大規模データといっても、ただデカければいいというわけではなく、データのどこかを見れば必ず全部がわかるという網羅性が大事です。そういう網羅的なデータは各研究所で作っていただくとして、この施設では、そのデータをどうやって大学の先生方に還元するかということを考えていきます。つまりデータ共有を進めるための仕組みを作りましょうということをまず掲げています。

それからもう一つ大事なことは、大規模データは通常のパソコンのレベルは超えてしまいますので、どうやって解析しようかということになります。そういう解析を支援するというユニットで、一つはこの機構の強みを生かしてゲノムデータのデータ解析支援ユニット、それからデータ融合計算支援ユニットの、二つのユニットを作ってございます。

データサイエンス共同利用基盤施設は、データ駆動型サイエンス（データサイエンス）の観点から、大学などの多様な分野の研究者に対し、大規模データの共有およびデータ解析の支援事業と人材育成を行い、我が国の大学などの機能強化に貢献する。これが本施設の基本目標でございます。事業内容は、データ共有支援、データ解析支援、T型・Π型人材育成となっております。

それからもちろん大学の先生方からいろいろなアイデアをいただきたいし、内部から上がってくるいろいろなアイデアも伸ばしたいと思いますので、戦略プログラムという形でいろいろなプログラムを走らせていこうと、さまざまな計画を進めているところでございます。

## データサイエンス推進に向けて果たす役割

設立からちょうど一年たったところなんですが、当施設が狙っていることはこういうことになるかと思います。

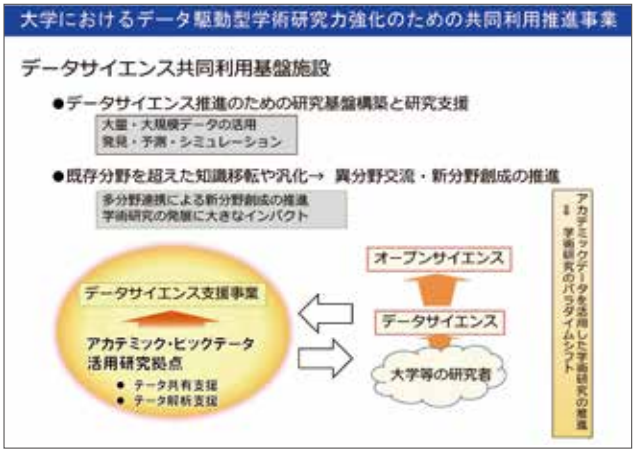
一つは大量・大規模データを活用し、発見・予測・シミュレーションといったデータサイエンス推進のための研究基盤構築と研究支援を行うことです。

それからもう一つは、データということに落とし込みますと、かなりいろいろな横串を通せるようになると思っておりまして、そこからかなり新しい芽も出てくるのではないかと思っております。既存分野を超えた知識移転や汎化から、異分野交流・新分野創成を推進したいと考えております。

ひとまずここは大学共同利用機関なので、アカデミック・ビッグデータの活用研究拠点として、データサイエンスの支援事業を進めていくということでございます。

我々がミッションとしてやらなければいけないことは、データサイエンスを自由自在に使えるような環境を大学などの先生方へ提供し、実際に解析の支援などを行っていくことでございます。

やがて行きつく先はオープンサイエンスですが、その意味では機構そのものに国立情報学研究所もございますので、オープンサイエンスを進めるための基盤を提供する仕組みはもうできあがっております。そういったところにまで持ち込んでいって、機構全体としてデータドリブンのサイエンスを進めていきたいと思っております。



# 「ライフサイエンス統合データベースセンターの取組み」

ライフサイエンス統合データベースセンター／センター長
小原 雄治

## データベース戦略、中核センター確立に向けて

そもそもライフサイエンスのデータベース（DB）は、たくさんの大学・病院でそれぞれの研究所が作ったので、バラバラでどこに何があるのかわからない、信頼性の高い注釈が付いていないから使いにくい、大型プロジェクトはあるがなかなかDBが公開されない、バラバラに構築・管理されていて検索・解析・応用が困難、さらにこれをまとめる戦略や中核センターがない、という数々の問題がありました。

そのためにライフサイエンス統合データベースセンターを作って、統合DBプロジェクトが始まったわけです。DBの所在情報を明らかにするために、Integbioという1543件（うち国内1105件）のDBカタログができていますし、簡単な横断検索（キーワード検索）もできています。DBを作ったのはいいが、研究終了後のメンテナンスが非常に難しいという点に関しても受け入れ体制ができています。権利関係に関してはクリアなもののみを受け入れていますので、自由に再利用できるという形になっています。

こういうことを実現するために平成23年度以降、JSTのNBDC（バイオサイエンスデータベースセンター）と共同で統合DB事業を進めて参りました。DBの利用に必要な、整理する・探す・つなげる技術を開発し、データそのものにアクセスして、意味をちゃんとわかった上で利活用できるようにしましょうということです。

## RDFによるデータ統合とSPARQLによる検索

データにはどうやって取ったのかという説明が必要ですし、再利用するためにはいわゆるメタデータが必要です。使っている用語に関しても取り決めや辞書が必要です。次世代のデータドリブンサイエンスのためのDB、統合DBを作るために、RDFというセマンティックウェブの技術を使うことを考えてきました。データそのものはポータルもできていますし、どこに何があるかという情報もどんどん貯まっていますが、それを統合的に利用する技術を開発しているということです。

RDFとは、Resource Description Framework の略です。データを主語＝モノのID、述語＝オントロジーで定義された属性、目的語＝別のモノのIDまたは値、この3つの組で表現するわけです。それぞれに、URIというユニークな番地を付けて管理するというものです。

たとえば遺伝子には名前がありますし、塩基配列がありますし、たとえば薬だったら作用・副作用がありますし、構造もあります。遺伝子が何かの薬剤に対して阻害されるとか、色々な関係も出てきます。そういうものを重ね合わせていくと、どこかでつながってくることがあるわけですね。それをつなげて行きますと、2つのものの関係が見えてくる。これまでわからなかった関係が大量のデータの中から見えてくるということです。こういう情報を提供したい、これが統合DBの概念です。

それを探すときに使うのが、SPARQLという検索用の言語です。SPARQLの文を書きますと、適切な関係のものがデータとして出てくるということです。例えばLODQAというシステムでは、自然文で「アルツハイマー病に関連する遺伝子は何ですか？」と質問すると、SPARQL検索を行い、関係する遺伝子のリストが出てくるというものです。これをアカデミックなデータに全部適応できるようにしようということで頑張っております。

## RDFデータの利用促進と国際的標準化

これまでJSTと一緒に、微生物ゲノム、タンパク質立体構造、遺伝子発現などさまざまな研究グループと協力をして、DBをRDF化していくということを進めてきました。私たちのセンターはそれをサポートするということです。すでに10数件のDBがRDF化できましたので、これらの利用促進を進めたいと思います。

DBの標準化は非常に重要でして、国際的なBioHackathon会議を通じて、ヨーロッパ、アメリカの方々と一緒になってRDFに関係する標準化を進めてきた次第です。セマンティックウェブは平成13年に提唱され、その後だんだん広がってきています。ヨーロッパのUniProtというのが一番早かったですけれども、私たちも結構早く始めていまして、阪大のPDBjも割と早く始めています。最近ではアメリカ、ヨーロッパのDBがどんどんRDF化を進めているという状況です。データのメンテナンスが非常に楽になりますので、コストも下がるというメリットもあります。こうやってBioHackathonをやって標準化を進めてまいりますと、日本、アメリカ、ヨーロッパ、たくさんのDBがつながっていくということになり、分野や国境を越えたDB統合が実現してくるのです。

# 「極域環境データサイエンスセンターの取組み」

極域環境データサイエンスセンター／センター長
門倉 昭

## 4つのグループから多種多様なデータが集結

極域環境データサイエンスセンターの目的とするところは、国立極地研究所の所有するデータの公開と共同利用、有効利用を促進し、極域科学のデータ活動の中心を担い、地球環境研究に貢献するということです。

「宙空圏」「気水圏」「地圏」「生物圏」という、4つのグループがさまざまなデータを取っており、時間的に連続した時系列データと、大気や海のサンプルやアイスコアなどといった試料系データの2種類があります。

「宙空圏」では、IGY（1957-1958年）という昭和基地開設以来のオーロラデータがあり、昔の400ftのフィルムで記録されたデータもあります。南極、北極、多点の観測機を使ったオーロラデータも取得され、100Hzで取った最先端のデータなども蓄積されています。最近ではPANSYレーダーという最先端の大型大気レーダーが昭和基地に設置され、ここから大気の流れを観測するデータが生み出されています。また、北欧のEISCATという大型のレーダー組織の国際プロジェクトに日本の代表として国立極地研究所が参加しており、電離圏の観測データといった最先端のデータもあります。

「気水圏」では長年、温室効果ガスを観測しており、二酸化炭素の増加、地球温暖化に関係するエアロゾルや雲の観測、人工衛星データからの南極域の雲や海氷の観測データもあります。ドームふじ基地では、地下3000mのアイスコアの分析から72万年前の歴史の解析が行われています。時系列データはせいぜい30年程度の蓄積ですが、試料系データは地球46億年という長いスパンの歴史を分析・解析できるデータになっております。

「地圏」では、岩石、隕石の試料が蓄積されています。時系列データで、地震、重力を観測しており、地球の重力の変化、地震波形を使った地球内部のモニタリング観測なども行われています。東日本大震災の影響は昭和基地まで及び、地震波、重力の大きな変化が観測されました。昭和基地とホバートの間の距離が年々離れているという測地データも蓄積されております。

「生物圏」では、ペンギン個体数の長期モニタリングや、最近ではペンギンあるいは大型動物の行動を解析する非常に先端的な研究がされています。陸上生物については、コケ類、種子類、さまざまな試料データが蓄積されデータベースが構築されています。

また、海洋生物から遺伝子解析をする設備やアイスコアラボラトリー、隕石ラボラトリーといった非常に大きなラボラトリーもあり、各種試料の高次処理・解析データも蓄えられ、データベースも作られています。

## 多彩な公開用（汎用）データベースシステム

さまざまなデータを外側に公開する仕組みとして、学術データベースがあります。これはメタデータベースで、いわばデータのカタログです。4つの研究分野のほぼすべての観測メタデータを見ることができるシステムも整備されてきています。

SCARという南極関係のコミュニティのデータ管理委員会と深く連携した形で開発していて、このメタデータをアメリカのGCMDに提供するために、GCMDのメタデータ形式に変換されて送られる仕組みもできています。南極観測や北極観測は常に国際的な共同研究や連携で進みますので、データについても国際的な連携を意識した仕組みを作っております。

最近は北極域の観測が集中的に行われておりまして、GRENEあるいはArCSといった北極プロジェクトの関係のデータ・情報を扱うシステムとして、ADS（Arctic Data archive System）というものが開発されています。これは、メタデータを検索するのみではなく、実際極域で取れるいろいろな実データを検索・表示・オンライン可視化・解析もできるアプリケーションも備えた総合的なデータベースシステムになっています。

IUGONETというデータシステムもあり、大学間連携プロジェクトで作られた超高層大気関係のデータのメタデータシステムです。特に各機関が所有する超高層大気関係のデータを横断的に検索できるというものです。

ソフトウェアも開発され、アメリカの同様のシステムであるSPEDASとも連携しています。

南極GISというシステムもあり、これは主に南極域の地形図、地質図、航空写真、衛星写真などの地理情報を検索し、地図上に表示するための地理情報システムで、南極観測隊活動支援としても利用されています。

## 極域環境データサイエンスセンターの目標

個別のデータベースや分野限定のデータベースはあるが、極域科学全体を横断的に俯瞰出来る総合的な仕組みがないというのが現状です。データベース化や公開の進み方にもばらつきがあります。また、公開用データベースシステムは作られてきた目的や背景が異なり、それぞれに特化したシステムが並列する状態にあります。

当センターでは、極域データ全体を抱合し、検索・可視化、解析までできる総合的なシステムの設計や構築を一つの目標としております。

また、データを有効利用できるという意味では、国立極地研究所が今年の1月19日に創刊した「Polar Data Journal」というデータジャーナルを有効利用したデータ出版の積極的な促進も図っていきたいと思います。

国内・国際コミュニティとの積極的な連携やデータサイエンス・共同研究の推進を図り、極域科学のデータ活動の中心を担っていきたいと考えております。

# 「社会データ構造化センターの取組み」

社会データ構造化センター／センター長
**吉野 諒三**

## 社会データ構造化センターのミッションと目的

社会データ構造化センターのミッションは、大学共同利用機関として、社会調査データ、公的ミクロデータ、ソーシャルビッグデータの収集や公開のための「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」の構築、その利用性向上のためのデータ構造化を進める。これを展開するなかで、全国のオンサイト拠点などを構築し、それを通じて、官民学のデータ利用コミュニティを形成し、環境、治安、経済を含む各種の社会的課題の解決のための実証的研究を促進させるということになっております。

目的としては、一つ目に社会調査データ、公的ミクロデータ、ソーシャルビッグデータを連携・統合したアーカイブを整備するということがあります。二つ目はこれらを展開するなかで、新たな調査データ収集システムと解析システムの開発をしていこうということです。三つ目は結果として実証的データに基づく人文社会科学の発展と政策立案の実現のための研究基盤の発展です。

三つのタイプのデータに対応しそれぞれの事業がありますが、それらは相互に連絡を取り合いながら連携したシステムを作っていくという共通課題もございます。

## 社会調査データから見える日本人の国民性

社会調査データに関して、ご説明いたします。先日のアメリカ大統領選挙で世論調査に基づく選挙予測が大外れしました。またイギリスではEU離脱の国民投票の予測が外れたという出来事もありました。それと比較すると、日本では住民基本台帳や選挙人名簿が整っていて、理想的な統計的標本抽出調査が遂行でき、それに基づく調査は比較的信頼できて、大間違いには至っていません。

そのような統計的標本抽出調査の基盤として、統計数理研究所が1953年から5年ごとに実施している「日本人の国民性調査」というものがあります。一例をあげますと、50年以上前に一度聞いた質問に「あの世を信じますか」という項目がありました。数年前にまた聞いてみると、50年以上前も今も、女性の方があの世の存在を信じる人が多い。それは一貫しています。けれども年齢層で比べてみると、50年以上前、年寄は信じているが若い人は合理的でそんなことは信じていなかった。ところが今はそれが逆になっている。若い人の方がむしろ信じていると。それはいったいなぜか。老若の感性が変わったのか。50年以上前、今の年寄は若者だったわけです。昭和10年代に生まれた人たちは、それよりも年上の人たちと比べても、あるいはそれよりも年下の人たちと比べても特別の意識を持っているコホート（特異な集団）をなす人たちが多い。つまり戦争が終わって価値観が急変した時代を経験した世代で、いろいろなことを簡単には信じない。そういう意味での「合理派」の世代が今は年寄りになったためなのか、それとも時代の効果で年寄と若者が変わったのか、それは断定できませんが、今後研究してみる価値はありそうです。

さて、「データ・アーカイブ構想」というものに関しては、戦後の比較的早い時代から各国で社会調査データアーカイブが構築されてきて、日本もそういうのを造らなければいけないという議論が長い間続いていました。近年では日本学術会議２０１４年５月社会統計アーカイブ分科会の提言「社会調査基盤のリノベーションに向けた官民学連携研究拠点の構築」の中でもまた提言が出されています。

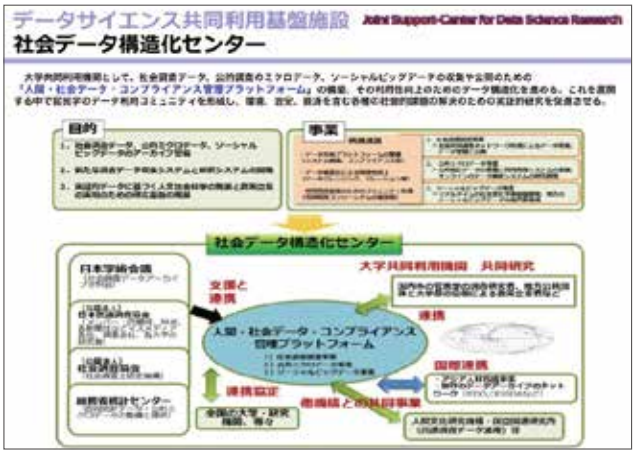
最近の状況を確認しますと、海外では立派なものがすでにあり、日本でも頑張って造り始めているものもある。しかし、ごく近年の様相としては、IT技術の発展で、一機関にデータを大量に集めてしまうというよりも、各データを収集したところがそれぞれのところでデータ公開も実施する「分散型アーカイブ」のシステムへ移行しているようです。データ活用の需要と供給のバランスを考えながら、現実にふさわしいあり方を求めている印象です。

## 人間・社会データ・コンプライアンスの重要性

公的ミクロデータというのは、政府が集めているデータで、多くの場合、内閣府の大臣広報室で管理し、各省庁の調査を取り仕切ってることが多いのですが、そういうデータが総務庁系の独立行政法人の統計情報研究開発センターを中心として二次収集されています。それは順次公開されていて、特に日本全国にオンサイト拠点と称するものを置き、一般公開はできないようなレベルのデータでも扱えるようにしていくということになっております。我々のセンターは、これをサポートするとともに、「アジア人材育成事業」などを含めて海外との連携も進めようとしています。

もう一つ、ソーシャルビッグデータの観点です。駅の改札での乗客の移動や観光客の移動、コンビニで何歳の男性がどんなものを買うとか、ビッグデータの活用についていろいろ報道されています。我々は、そういう情報を二次収集して、各地方自治体などと連携協力して観光や産業を支援することなどを考えております。

ここまでいろいろなタイプのデータについて申しましたが、一つ一つは公開しても問題ないデータでも、さらなる活用のために各データを連携・統合しようとなると、個人情報の問題を含めていろいろな問題が上がってくることに注意しなければなりません。我々のセンターでは、「人間・社会データ・コンプライアンス」---個人情報保護の問題を含め、コンプライアンスを管理するプラットフォーム---を構築して、将来的にはそのシステム自体をいろいろな大学や自治体に提供するというようなことを考えています。



# 「人文学オープンデータ共同利用センターの取組み」

人文学オープンデータ共同利用センター／センター長
**北本 朝展**

## オープンデータの3つの価値と3つの関係者

オープンデータの価値にはさまざまな側面があり、「利用」が促進できる、「透明性」が確保できる、みんなが「参加」できるという基準があります。我々のセンターは、この「参加」という側面を重視したいと思っております。

「本は図書館に行って読めばいい」とも言われますが、実際そうはいかないという側面もあります。例えば、図書館が1000㎞離れた場所にあったらどうするのか。人文学の本は一点物が多く、1000㎞移動しないと読めない場合も多々あります。また、例えば米国の研究者が東アジアの本を研究したいと思ったら海を渡って来ないと読めません。そのとき、例えば日本と中国を比較して中国のデータの方がオープン化されていたら、そちらの方が研究しやすいので研究対象に選ばれやすいという状況も生じています。いろいろな人が参入できないと、研究分野自体が盛り上がらないという状況があり、そういう点からもオープンデータ化を進めていくことが重要な課題となっています。

オープン化にあたって、3つの関係者について考えています。一つは「研究者」です。ここでは、データをより深く読み、分析して新しい知見を得ていくという方向性が重要になります。

一方、大量の情報を高速に処理するためには機械の力が不可欠ですから、いかに「機械」が使いやすいデータに整えるかということも重要な課題になってきます。

また人文学データは非常に多様ですので、定型化されたデータを自動的に処理するだけでは済まない場合も結構あります。そこには「市民」の協力が必要ですし、また市民自体が学習して賢くなっていくということも、オープンサイエンスの重要な側面だと考えております。

## 情報学と人文学の協働で歴史的典籍活用を推進

人間文化研究機構の国文学研究資料館との共同研究についてご紹介します。これは我々の人文学オープンデータ共同利用センターと統計数理研究所、国立情報学研究所の研究者が一緒になって取り組んでいるものです。国文学研究資料館では、日本の歴史的典籍30万点をデジタル化し、国際共同研究を推進する大型プロジェクトを進めています。そこでまず、研究者のためのオープンデータとして、「日本古典籍データセット」を公開しました。ここで公開した古典籍は、くずし字で書かれていまして、江戸時代のくずし字を読めればこのデータも活用できます。最終的に30万点をデジタル化する計画ですが、現在はまだ700点の公開に留まっていますので、今後どんどん増やしていくことが重要です。なお、翻刻テキストは一部の古典籍のみで、基本的には画像ファイルに書誌のメタデータなどを同梱したオープンデータとなっています。こうした画像公開によって、本を見ることは可能となりましたが、データを活用できるのはくずし字が読める研究者だけです。くずし字をスラスラ読める人は日本に数千人程度しかいないと言われており、これだけだと日本人でもほとんど活用できないことになります。

そこで、機械が使えるデータにしないといけません。では機械が使えるデータとは何か？ それはどこに何の文字が書かれているかということを機械に教えるための、いわゆる教師付き学習データというものです。「日本古典籍字形データセット」はそのためのもので、これを使えば機械が学習して文字を認識するということも可能になります。

たとえば、「あ」という文字には、我々が読める「あ」と別の形の「あ」があります。これは変体仮名と言いまして、元となる漢字が異なる「あ」なんですね。江戸時代は「あ」が何種類もあったのですが、今は一つの「あ」しか残っていません。

機械が学習するデータとして、現在86,176文字のデータがあり、今年度末には40万文字以上になる見込みです。座標情報が入っていますので、この画像のどこに何の文字があるかという認識もできます。このようなデータを使って学習すると、文字認識のプログラムが作れます。ただ、この40万文字というのは実は全然十分ではないと考えていまして、もう2桁ぐらい多い文字数がないとなかなか全部は読めないのかなとも思っています。

## 江戸料理レシピデータセットに未来を感じる!?

江戸時代の料理の本に冷やし卵羊羹というレシピが載っていますが、これを読んで作れる人はほとんどいません。しかし、それを「江戸料理レシピデータセット」として、現代の人が料理できるような形にすれば、使えるようになるわけです。

デジタル化しただけでは十分とはいえません。くずし字を翻刻すればいわゆる古文の状態になりますが、古文も読めない。そこで現代語訳し、さらにレシピ化すると、そのまま使えるようになります。さらにこれをクックパッドで公開します。センターのウェブサイトで公開するだけではなく、クックパッドのような一般の人が使うプラットフォームに乗せることによって、データをファインダブル（可視化）にすることに意義があります。これには、非常に大きな反応がありまして、国立情報学研究所のTwitterアカウント史上最大の反応がありましたし、「江戸料理＋クックパッド」に未来を感じるというツイートもありました。関心が非常に高いデータ公開だったと言えます。

近代の本もOCRが難しく自動解読できていません。統計数理研究所と国立国語研究所との共同プロジェクトとして、今後取り組んでいきたいと思っております。

このような課題を、我々はディープアクセス技術と呼びたいと考えております。メタデータにアクセスするというのではなく、データの中身、コンテンツの中身にまでアクセスできる技術が必要です。

# 「ゲノムデータ解析支援センターの取組み」

ゲノムデータ解析支援センター／センター長
**野口 英樹**

## 次世代シーケンサーがゲノム解析を変えた

我々のセンターは、大学などいろいろな研究機関の研究者がお持ちのゲノムデータをバイオインフォマティクスの技術を用いて解析し、そのための解析技術を提供するセンターになります。

ではゲノムデータというのはどういうデータなのかということですが、動物、植物、微生物までどんな生物種であっても、細胞の中にゲノムDNAを遺伝情報として持っています。DNA以外にもそこから転写された転写産物（RNA）なども存在していまして、これを計測したデータがゲノムデータということになります。

計測機器にはいろいろなものがあり、例えばDNAマイクロアレイとか、最近ですとアイリスとかいろいろ装置がありますが、現在メインで使われているのはDNAシーケンサーです。DNAの情報、ACGTの配列を直接読み取るという装置を使って出てきた塩基配列データが、当センターで主に取り扱うデータということです。

ただ問題はこれらのデータというのは基本的にはどれもこれもACGTの配列なんですけれども、生物種が違ったり、実験条件が違ったりすると、その意味することというのは大きく変わってくるわけで、当然解析方法も変わってきます。そういう意味で多様であるということと、もう一つの大きな問題はデータが大量に出てくるというところです。

これによって、なかなか解析が困難だという状況がございます。そういう大量のデータが出てくるようになった背景に、この次世代シーケンサーと呼ばれる新しいシーケンサーの登場があります。

国立遺伝学研究所の先端ゲノミクス推進センターに置いてあるシーケンサーには、Illumina HiSeq2500とPacBioRSⅡというものがあります。ゲノムのDNAというのは、例えば人の場合、3G、30億塩基対あるわけですけれども、このシーケンサーはその30億塩基対を端から端まですらっと連続で読み取れるというものではありません。ゲノムの非常に短い断片を読むことしかできないのです。その一個一個の断片を配列リードと呼んでいますが、そのリードの長さは、Illumina HiSeq2500で最大250 bp、PacBioRSⅡの場合、若干長いのですけれども、それでも平均15 Kbp塩基対しか読めない。非常に短い配列しか出て来ないということです。

ただ数の方は大量で、Illumina HiSeq2500で一連あたり、最小単位で約2.5億本くらい出てくる。つまり掛け算して約60 Gbのデータというものが一度に得られるわけです。PacBioRSⅡの場合、若干数として少ないのですが、それでも前世代のシーケンサーが96本とか384本でしたので、それに比べると、100倍以上のデータが一度に得られるようになってきています。

シーケンシングに係るコストも、次世代シーケンサーの登場した2007年以降、非常に速いスピードで低下してきています。コストが下がってくると、皆さん使いたくなってくるわけです。ゲノム情報というのは生物の基盤の情報になりますので、使えるのなら使いたいということで、いろいろな分野の先生方がゲノムデータを取得するようになってきました。ただ、これまで配列を扱ってこなかった先生が、いきなり1ファイル50 Gbとか60 Gbになるようなデータを数十億本とか扱わなければいけないということになってもなかなか対応ができないわけです。ですから、解析されずに眠っているデータもたくさん出てくるという状況になっているわけです。

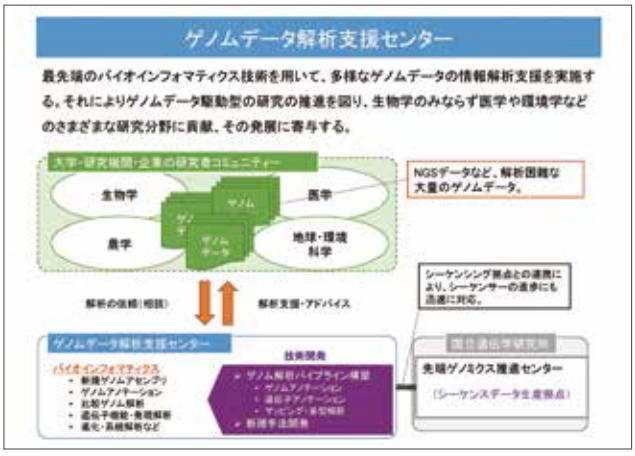
## 多様なゲノムデータの情報解析で研究を支援

我々のセンターでは、最先端のバイオインフォマティクスの解析技術を用いてこれらのデータを解析し、共同研究という形で研究を回したり、あるいはその研究者が求めるような結果の解析を行うためにはどういうデータを用意して、どういう解析手法をとるのかという技術的なアドバイスなどをしていくことで、ゲノムデータ駆動型の研究の推進に貢献することを目的としております。

実際にどういうデータ解析を行っているかということですが、新規にゲノムを決定する「De novoゲノムシーケンス」、まだゲノム配列が決まっていない生物種のゲノムを決定する、あるいはもうすでに決まっている別の個体とか変異系のものや変異仮想を調べる「ゲノムリシーケンス」、あるいは発現している遺伝子の解析をする「トランスクリプトーム解析」、さらに「メタゲノム解析」と幅広いデータ解析を行っております。

今年度は、11課題で支援依頼がありまして、生物種としてもかなり幅広く哺乳類、昆虫、魚類、植物、あるいは原核生物と、非常にさまざまな種類の生物種の解析依頼が来ております。

また、この解析支援を続けていく上で、効率化のためのパイプラインの構築、高速・省メモリな解析アルゴリズムの開発、解析講習会などの人材育成などに取り組んでいきたいと思います。



# 「データ融合計算支援プロジェクトの取組み」

データ融合計算支援プロジェクト
**中野 慎也**

## シミュレーションと観測を融合した研究手法

データ融合計算というのは一般的な用語ではないかもしれませんが、シミュレーションとデータサイエンスを融合した、この研究手法のアプローチをいろいろな分野に紹介していくことが、このプロジェクトの概要です。

すでに研究相談の受付を開始しておりまして、共同研究の実施を通して他分野に展開しており、さらに随時共同研究を受付中といったところです。

データ融合計算に関するノウハウや手法、ソフトウエア化したものを整備し、今後の公開に向けての準備も進めております。

また来年度以降、セミナーやハンズオンの活動を通じて、我々の研究手法をいろいろなところに紹介していこうと考えております。

具体的には、シミュレーションと観測を統合した研究手法というものに取り組もうとしております。

シミュレーションというのは、第3の科学と言われて、ひとつ前の世代の研究手法かもしれませんけれども、今でもスパコンなどいろいろな分野で高精度なシミュレーションが展開されています。

ただ、スパコンの性能が上がるとともに計算量がどんどん増大したこと、インプットを与えないと結果が出て来ないという問題がシミュレーションにはあります。

一方、観測データですが、最近はいろいろな種類のデータが取られていますけれども、そのデータから意味のある情報や知識をどう抽出するかが問題で、本質的な情報が見えにくくなっているという現状もあります。

そこで我々は、シミュレーションと観測を融合させ、第3の科学のアプローチと第4の科学のアプローチを融合させたような手法を提案させていただき、いま展開しているところであります。

シミュレーションというのは、入力変数と入力パラメータを与えて出力を何か返すというものです。大体出力というのは部分的には観測できるものになっています。シミュレーションを実行するときの計算量がすごく重いという状況の下で、何を入力したのか、どういうパラメータ設定で実行したのかということを推定するという方法として「データ同化」と呼ばれる手法があります。

もう一つは、データサイエンスのアプローチを使って、シミュレーションを模倣するモデルを作るという手法があります。シミュレーションは計算時間がすごくかかるものですから、これを簡単なモデルで、ある程度ローコストで予測を行う「エミュレータ」というものを作るという手法です。この2つがシミュレーションと観測の融合の主な手法です。

## システム設計と検証の作業を一体化

観測データからいろいろなデータを組み合わせて解析し、それをシミュレーションのデータ同化と呼ばれる手法を使って、シミュレーションに食べさせてやります。さらにそのシミュレーションの振る舞いを模倣するようなモデルを作れば、どういったところのデータを取ればシステムの改善ができるかといった判断ができますので、そこからまたデータを取り、解析するというサイクルになります。

これによって、システム設計やアウトプットの定量的予測、不確実性の評価などの実現を進めていけるような方法論を呈示していきたいと考えているところです。

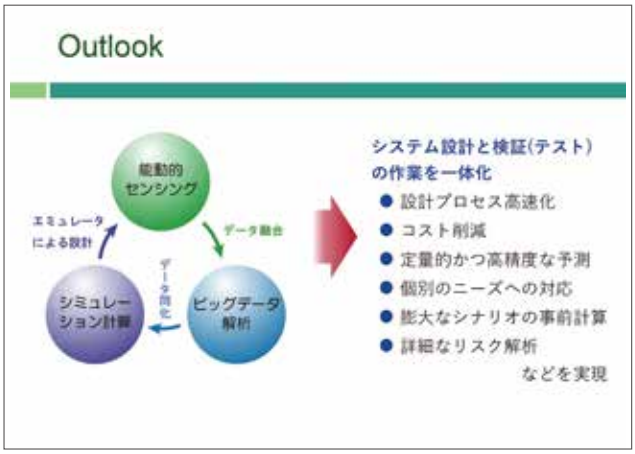
現在、我々は方法を研究しているところがございまして、具体的なテーマは外からいただきたいという立場にありますので、相談の窓口を開きましてメールなどで受付をしております。我々の持っているノウハウの提供で解決できる問題もあると思いますが、新しい方法論を開発しないといけないようなテーマの場合は、共同研究として実施させていただくといったことを考えています。

もちろん、ノウハウの提供で解決しそうな問題の場合でも、統計的な考えに詳しくない方もいらっしゃると思いますので、そういった場合には共同研究として実施させていただいて、より密な情報交換をしながら進めていくといった出口も用意しております。

## データ融合計算プロジェクトの実績

これまでの実績としましては、細胞質シミュレーションのパラメータ推定(国立遺伝学研究所との共同研究)、南極氷床コア年代推定手法の開発(国立極地研究所との共同研究)、マルチモデルによる将来気候変動予測(防災科学技術研究所との共同研究)、人の動きとデータのシミュレーションから人の動きを推定するといった人流シミュレーションの状態推定(東京大学生産技術研究所との共同研究)など、いろいろな応用分野で研究を進めております。

また、民間企業との共同研究も受け付けておりますし、現在進行中のものもございます。



# 「情報・システム研究機構の新時代に向けて」

大学共同利用機関法人 情報・システム研究機構／機構長
藤井 良一

## 研究所のミッション遂行と機構の果たす役割

情報・システム研究機構のミッションは、生命、地球、環境、社会などの複雑な問題を、物質とエネルギーの観点に替って「情報とシステム」という立場から捉えるための方法の研究、研究基盤の整備および融合研究による新分野の開拓を行なうということでございます。

機構を構成する４研究所は、国立極地研究所と国立遺伝学研究所という分野型の研究所と、国立情報学研究所と統計数理研究所という分野によらない基盤的な科学を行う研究所から成っております。各研究所の歴史は、2004年の機構の設立より十分に早くて、各々のコミュニティからの付託を受けて、先端的な学術研究、共同利用、人材育成を行うことをミッションとしてきております。

これらの研究所は各学術領域でCOEであることが求められております。これは大学共同利用機関として、最も優先度が高い大学共同利用、共同研究事業を行うための前提条件になるものと考えております。世界をリードする技術レベルを持って初めて、全国の大学などの研究機関や企業の方々が共同研究をしたいと思っていただけ、かつ大学などの機能強化に貢献できると考えるからでございます。

機構は研究所による最先端研究と科学の発展、それを支える基盤を支援することが役割と考えております。研究所と機構の共同により、社会や大学などと連携して基盤的な環境を提供し、基礎的そして応用的な成果を社会に還元することが求められていると自覚しております。

2016年に法人第三期に入りまして、それまでの事業の成果を踏まえ、研究所の力を結集して、今後訪れるオープンサイエンス時代に向けて、機構本部の機能をより高めようとしております。

一つは、本部に戦略企画本部を立ち上げ、各研究所の執行部から参加をいただき、機構全体の研究戦略や将来計画、そしてガバナンス施策などの提案を行う組織を立ち上げました。

また、データサイエンス共同利用基盤施設を立ち上げまして、研究所が作り出すデータのデータベース統合を行い、オープンデータ、オープンサイエンス推進の加速を目指しております。これにより大学の共同利用の拡大とともに、統合された異分野データから新分野創成につながることを期待しております。

## 社会と学術のデータサイエンスへの要請

情報やそれに関連するテクノロジーは現在非常に急激に発展を見せております。急激なICT化と多種多様なビッグデータの出現、それに呼応しようとする計算能力の成長は、社会を変容させ、研究環境を大きく変化させております。機械学習は発達し、社会での応用は拡大しつつあるということです。人間の能力は、例えば2500年前の孔子の時代からあまり変わっていない、進歩していないように見えますけれども、テクノロジーの進展は凄まじくて、素人目にはビッグデータやAIは人間が制御できる範囲を既に超えているのではないかという不安に陥るほどでございます。

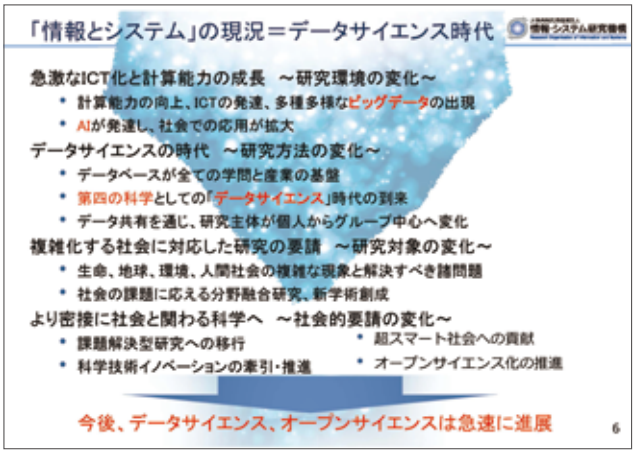
国立極地研究所が共同運営しております、欧州にある超高層大気観測所の大型レーダーEISCAT（アイスキャット）では、数年後の2020年過ぎに新しいレーダーを建設して観測を開始する予定ですが、最速サンプリングでのデータ量は1.6 Tbit/秒(生データ）でございます。リアルタイム処理で1/30に減らしますが、それでも54 Gbit/秒、3.2 Tbit/分、4.7 Pbit/日、1.7 Ebit/年という巨大な量になります。もちろん連続運転は不可能でございますけれども、いかにインテリジェントに有用なデータを引き出すか、数年後には情報・システム研究機構として対応が迫られるチャレンジングな課題となっております。

今後データベースがすべての学問と産業の基本となることが予想され、それに呼応して第4の科学としての「データサイエンス」時代が到来して、データ共有を通じて研究主体が個人からグループ中心へ変化すると言えます。

また研究対象も生命、地球、環境、人間社会の複雑な現象と解決すべき諸問題が重要課題となって、さらに社会の課題に応える分野融合研究、新学術創成が求められているということであります。

すなわち社会的要請の変化によりまして、課題解決型研究への移行、科学技術イノベーションの牽引・推進、超スマート社会への貢献が求められ、これらの基礎となるオープンサイエンス化の推進が強く求められているということであります。

ビッグデータには膨大な知識や価値が埋もれておりますけれども、現在の方法・技術では必ずしも十分な有効活用はなされていないと言えます。新たな科学的な手法によりまして、知識発見や価値創造を行うことが必要だということで、それが正に第４の科学と呼ばれているデータサイエンスに他なりません。



ビッグデータの利用には、レーダーの例で申しましたけれども、大量の散在するデータをリアルタイム処理する技術、大規模なデータ処理技術、それから膨大な高次元データや計算結果を人間が把握可能にするデータ可視化、そしてビッグデータから意味ある知識獲得のための方法、データ解析手法が必須となります。

## 分野融合・新分野創成のための機構の施策

情報・システム研究機構では、中期的対応としまして、データサイエンス、オープンサイエンスを推進して、その結果として分野融合研究や新分野創成を行うために、データサイエンス共同利用基盤施設をフラッグシップとして、データ共有支援、データ解析支援を行い、オープンサイエンスを推進していきたいと考えております。

また、学術動向や社会の要請の変化に対応するためには、従来の縦型の教員配置に加えまして、横断型の教員配置を作りまして、スピーディーに対応していくことが重要で、これは各研究所で実施されつつあります。

このデータサイエンス共同利用基盤施設は、機構の研究所の所有する、または他機構との融合で得られる、今までは個々に利用されてきた生命、遺伝子、地球環境、人間社会、人文学に関するデータベースにメタデータやRDF化などを行いまして、データ共有の支援をし、統合データベース化します。さらにデータ同化などのデータ解析支援を行うことにより、データベースを相互に利用することを可能にして、データ利活用を格段に高めようとするものでございます。その結果、データサイエンスが推進されて、異分野融合が進むことで、新分野創成が可能になると期待しております。

統合されたデータは、大学などのすべての研究者が利用可能になるというものであります。また、国立情報学研究所は全国の大学のために、オープンサイエンス研究のデータ基盤の準備を進めております。さまざまなデータベースを管理・公開するための基盤で、データサイエンス共同利用基盤施設のデータベース、これも将来、この基盤の上に乗ることになると考えられます。

一方、大学や研究機関などの学術界や、国や自治体、それから企業と産業界など社会にはさまざまなデータがあります。データサイエンス共同利用基盤施設での統合データベース構築のためのデータ共有支援やデータ解析支援を横展開することによりまして、学術界、社会全体のデータベースの統合と相互利活用への道が拓かれて、オープンサイエンス実現への貢献ができると考えております。

これによって、全体のデータが統合されるということと、大学の機能強化や異分野融合、新分野創成の促進にもつながると考えております。共同利用はいくつかのパターンがございますけれども、全研究者のための基盤の提供は大学共同利用機関の大きな役割であると考えております。全国に張り巡らされたSINET5、セキュリティ、クラウドそれからさまざまなCiNiiとかJAIROなどの学術情報の提供などがこの基盤にあたると考えます。

## 時代が求める研究者・高度技術者を社会・大学などに輩出

また、データサイエンス共同利用基盤施設も全国に提供される基盤施設の一つでございます。共同利用機関の大きな役割の一つは人材育成であります。情報・システム研究機構の各研究所はデータや情報に関するさまざまな人材育成を行ってきております。統計数理科学人材の育成、それから情報科学人材の育成、そしてバイオインフォマティクスの育成と、大変精力的に行われてきておりまして、ほかにも多くのプログラムが計画されているとお聞きしております。機構としては、こういうものの見える化を図り、支援していきたいと思っております。

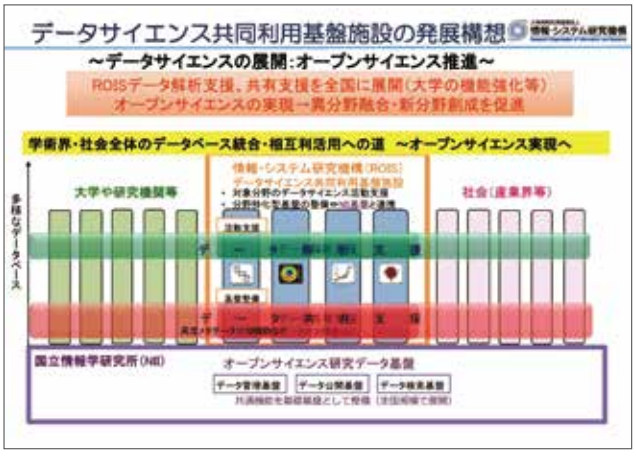
一つの課題として挙げたいのは、社会が必要としている人材数と研究所が育成できる人材数との関係です。統計数理研究所では棟梁レベルの人が毎年500人は必要と言われております。統計数理研究所はその主要な部分を育成することを期待されていると思いますが、これは大変なことだと思います。人材育成制度の整備と拡大が急務であると機構でも考えております。社会や大学などのニーズに、質だけでなく量的にも応えることができるように、機構として支援をしていきたいと考えております。

もう一つ重要なのは、人の交流、循環です。共同研究のレベルを超えて、研究者の相互の循環は両者の活力を高めて、大学の機能強化とともに、我々研究所・機構の発展、双方に貢献すると信じます。機構では、研究者交流プログラムがありますが、今後さらにクロスアポイントメント制度などを用いた双方向の人材循環を支援していきたいと考えております。

一方、大学院生の教育は、機構の最も重要なミッションの一つでございます。機構は、総合研究大学院大学の一員として、大学院生を受け入れております。研究所で最先端の研究に接して、また野外のフィールドで生の自然の神秘に接して感動するということは、学生にとって得難い経験でありまして、学術研究に進む人材だけでなく、社会において指導的な立場に立つ総合的能力を持つ人材を輩出することを可能にすると思います。

できるだけ多くの学生たちにこの教育研究環境を経験してもらうために、総合研究大学院大学を中心に、連携大学院や特別共同利用研究員制度、インターンシップなどを十分活用できるように機構として支援していきたいと思います。

各研究所の学理をもとに学術基盤の充実を図って、データサイエンス、オープンサイエンスを推進して、今まで以上に大学などの機能強化や社会のイノベーションに機構一丸となって貢献していきたいと思っております。



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


大学共同利用機関法人
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

〒190-0014
東京都立川市緑町 10-3
TEL:042-512-9254
https://ds.rois.ac.jp/