国際シンポジウム「ビックデータ時代の図書館の挑戦－研究データの保存と共有」議事録

日時：平成 26 年 2 月 5 日 13:30～17:00

場所：国立国会図書館東京本館新館講堂

主催：国立国会図書館

プログラム

※以下、〔 〕内は、各講演者の講演スライドのページ番号を示しています。

開会挨拶

中山正樹（国立国会図書館電子情報部長）

本日は多数の皆様にお集まりいただき、誠にありがとうございます。国際シンポジウムの開催に当たりまして、一言挨拶申し上げます。

昨今、ビッグデータの利活用に注目が集まり、データのオープン化に向けた動きが加速しています。日本政府におけるオープンデータの動きとしては、2012 年 7 月に IT 戦略本部において「電子行政オープンデータ戦略」が決定され、世界的な動きとしては、2013 年 6 月の G8 首脳会合で、各国首脳が「オープンデータ憲章」に合意しました。オープンデータ推進に向けた具体的な議論が始まっています。

オープンデータとしては、コンピュータ、計測装置、観測装置から生み出される、膨大な量の研究データがあり、それを保存し、共有することによる期待が高まってきています。また、自然科学に限らず人文・社会科学分野も含め、異分野の研究データの共有・統合により、新たな知見が生み出されることが期待されていると考えます。

国立国会図書館は、2011 年度に策定しました科学技術情報の整備に関する基本計画において、「新たな知識の創造と還流」を目指した「知識インフラ」の構築への積極的な関与を掲げています。「知識インフラ」は、科学技術研究活動の過程で生じる研究データも含めた多様な学術情報について、収集・保存・組織化・公開等の機能を実現し、知識の循環を促すものです。国立国会図書館は、2011 年の大震災直後から取り組んでいる東日本大震災アーカイブの構築を通じて、関係機関と分担して、大災害のあらゆる記録を集め伝える活動に取り組んでいます。この東日本大震災アーカイブでは、文献以外の音声や動画、また研究データも収集対象としており、「知識インフラ」構築の先行事業として位置付けています。

また、従来からの事業として、当館が収集した資料の書誌情報、7,300 万件をはじめ、ウェブサイト情報、電子書籍・電子雑誌の本文テキスト情報、画像情報などアーカイブしたデジタルコンテンツはビッグデータと言えるものです。例えば、当館がインターネット資料収集保存事業(WARP)で収集したウェブサイトのデータは、昨年末現在で、ファイル数にして、実に 20 億に達しています。これらについて、全文テキストを活用することにより、文献の書誌情報、引用情報レベルでの検索でなく、本文のテキストマイニング等による意味的内容により、資料と資料、資料と研究データとを有機的に関連付けて、必要となる情報を的確に入手することが可能となるサービスを実現することも想定されます。また、画像情報を認識した検索や、アクセス統計を活用した文献候補のサジェスションなど、情報の利用スキームも大きく変革する可能性を秘めています。

本日のシンポジウムの御報告及び議論については、当館及び関係者の皆さんの進むべき方向性への御示唆を頂けるものと思います。

以上、簡単ではございますが、開会の挨拶とさせていただきます。

「知識インフラ」構築に向けて（講演者紹介）

川鍋道子（国立国会図書館科学技術・経済課長）

国立国会図書館の科学技術情報整備に関する基本計画を担当しております。

このスライド〔川鍋-2〕は、国立国会図書館の使命と目標を紹介しています。図書館、特に国立国会図書館は、出版物を中心に国内外の資料・情報を広く収集し、保存して、知識・文化の基盤となるべく各事業を展開してきました。近年は印刷出版物にとどまらず、電子的に流通する情報も文化的資産として収集し、保存することを使命・目標に掲げ、ウェブ上の出版物については実行してきているところです。

先ほど、開会の挨拶でも触れました、国立国会図書館が策定した科学技術情報整備に関する基本計画の概要が、こちらのスライド〔川鍋-3〕になります。この基本計画では、八つの柱を掲げており、そのうちの一つとして、当館が「知識インフラ」の構築に積極的に関与することとしています。「知識インフラ」の構築は、研究データも含め多種多様な資料・情報について保存・共有・利用の循環を目指すものです。研究データについては、何をどこまでどのように保存・共有するのか、国立国会図書館も含め図書館という機関がどこまで関与できるのか、または、できないのか、まだ見えていない状況と捉えています。本日は、「知識インフラ」の構築に向けて、研究データの保存・共有に図書館界がどうアプローチしていくべきか議論の場にしたいと考えています。

続いて〔川鍋-4〕、本日の講演・報告について講師の先生方を紹介します。

最初に「研究データをめぐる国際動向」と題しまして、情報通信研究機構統合データシステム研究開発室長・京都大学生存圏研究所客員教授　村山泰啓先生に御報告いただきます。村山先生は、科学者の国連ともいわれます国際科学会議（ICSU : International Council for Science）に設置された国際組織である世界科学データシステム（WDS : World Data System）に深く関わっておられます。また、科学データに関する国際的なプログラムであります研究データ同盟（RDA : Research Data Alliance）に関する動向についても御紹介いただき、研究データの保存・共有について研究者のお立場からの概説をお願いしております。

続いて、基調講演として、ドイツ国立科学技術図書館の戦略：研究データの保存と共有」と題しまして、ドイツ国立科学技術図書館・ハノーファー大学図書館 研究開発部門長 ペーター・レーヴェ先生にお話しいただきます。レーヴェ先生には、ドイツ国立科学技術図書館（TIB）における研究データリポジトリのプロジェクトである RADAR（Research Data Repositorium）を中心に御紹介いただき、研究データの保存・共有をマネジメントするうえで、直面している課題、研究者・技術者・図書館員との連携の在り方、図書館員に必要なスキルについてもお話しいただきます。レーヴェ先生は、ヴュルツブルク大学で地理学、コンピュータ科学を専攻され、ドイツ地球科学研究センター研究員、プロジェクトマネージャーを務められ、昨年 2013 年の 10 月からドイツ国立科学技術図書館の研究開発部門長及びドイツ地球科学研究センター客員研究員を務められています。

次に、国内の研究データの保存・共有の先駆的取組について、自然科学、社会科学それ

ぞれから事例報告をお願いしております。農業研究におけるデータ共有の実態と題しまして、農業・食品産業技術総合研究機構 中央農業総合研究センター情報利用研究領域 上席研究員 木浦卓治先生、SSJ(Social Science Japan)データアーカイブにおけるデータの保存と普及と題しまして東京大学大学院情報学環教授〔社会科学研究所兼務〕の佐藤博樹先生に、それぞれ運営の現状と課題等について御紹介いただきます。

講演・報告のあとは、「研究データ・マネージメントの将来像：図書館ができること」と題しまして、国立情報学研究所長・東京大学生産技術研究所教授でありデータベース工学が専門の喜連川優先生、レーヴェ先生、村山先生による鼎談を行います。モデレータは村山先生にお願いしています。

最後に、質疑応答を予定しています。以上、本日のシンポジウムの趣旨について説明しました。

【講演】研究データをめぐる国際動向

村山泰啓氏（情報通信研究機構 統合データシステム研究開発室長・京都大学生存圏研究所客員教授）

私は情報通信研究機構（NICT）という研究組織にいますが、実は専門は地球科学で、ITの専門家というわけではありません。科学の分野から、データの保存や研究についてどのように関われるか、という視点でやっています。京都大学でも一緒に研究をさせていただいており、また、先ほどお話のあった WDS でも国際的な事業をしています。〔村山-2〕講演の内容としては、オープンデータの現状の議論、政府の動き、アカデミーの意識、動きといったものです。私が関わっているのは、WDS の学術データ事業、また、先ほどお話のあったコンソーシアムの RDA などにも関わっています。現在進めていることとして、データパブリケーション、データサイテーションなども御紹介できたらと思います。

〔村山-3〕ビッグデータをめぐる政治の大きな動きとして、オバマ大統領が「ビッグデータイニシアティブ」を 2012 年に発表しています。以前から研究コミュニティではデータについての関心が随分高まっていたわけですが、それがこうして一般社会に見える形で出てきました。〔村山-4〕G8 サミットが昨年 6 月に行われ、そこで「Open Data Charter（オープンデータ憲章)」が合意されました。これは一般的なデータのオープン化についての議論だったわけですが、〔村山-5〕G8 サミットと同じ月に、科学技術大臣及び各国アカデミー首長が集まる会合が行われて、「Science Ministers Statement」という合意に達しました。その中で「Open Scientific Research Data」が明記されており、一般的なデータだけでなく、研究データについても国際社会が十分な関心を持つべきだというメッセージが発信されました。この会議には、日本からは日本学術会議の会長と副会長、内閣府から総合科学技術会議の原山議員が閣僚級として参加され、オープンデータについて大いに議論されたそうです。その関係で「国際動向はこうだそうだが、日本ではどうか」あるいは「国際的な議論はどのようになっているのか」といったことについて、私が呼ばれて色々御相談させていただいている次第です。

〔村山-6〕国際的な議論の背景の例としてここに挙げたのは、「Science as an open enterprise」です。これは英国王立協会（Royal Society）が発表した報告書ですが、「公開事業としての科学」と仮訳してみました。これは、王立協会に設置された有識者部会による学術データ・情報のオープン化についてまとめたものです。そもそもヨーロッパでは、近代科学の成立時から、「科学とは何か」「科学情報の共有とは何か」という非常に根本的なところの認識がよりしっかりできているように見受けられます。〔村山-7〕近代科学において、原著論文という形で情報が固定化され出版される中で、実験方法が記録されていたり、それが第三者による再現実験が可能な記述をすることによって、科学的真実が担保され保障されることを、中高生向けの科学入門書などで科学的方法論として紹介されています。加えて、現代において非常に重要なのはコミュニティです。専門家コミュニティ、学会による自由な相互批判と検証が必要になってきます。論文に書かれた内容の再現がどうなっているか、それを色々な研究者・科学者の目で見て大丈夫かをチェックし続ける体制が成立します。

したがって、二つ目に「「科学」という「制度」」という項目を挙げたように、「科学」というものはそれだけでは社会における立場を議論するのは容易でない。自然や研究対象を見つめてそのメカニズムを理解して、理解したことを相互に共有することで、初めて「科学」として他の人々に問いかけることができるわけです。例えば、オゾン層にオゾンホールができていることを発見するだけではなく、それを皆が確かだと信じるに足るだけの根拠を提出する、また北極や南極に近い国、遠い国の科学者も含めて、オゾンホールの議論をして必要なら相互批判して結果を改善する、それがこれまで科学がやってきたことだと思います。これは、共有されたデータや情報に基づいて行われます。これが先ほど申し上げた検証や再現性の担保ということです。

現代では社会が何かアクションを起こすための判断基準は何かというと、科学が非常に重要な国内外の認識基盤の１つを築いていると思います。それだけ科学が社会的に信頼されているわけです。「気候変動が問題であれば CO2 を削減すべき」という国際社会の合意形成は科学的研究成果に基づいて行われています。そういった信頼を国内外において我々はどのように維持していくことができるのか。例えば、原著論文を固定して、評価して、公表して、保存して、引用する、更に再利用できる、といったメカニズムが、科学者、研究機関、図書館業界、出版社などなどの皆様を含めて成り立っています。そういったことが科学にとってどんなに重要かということです。「「科学」という「制度」」と書いていますが、研究活動による新たな科学的知識という情報が生成されたのち、その情報を共有し、皆で認識を共有することも含めるならば、出版や文献の保存、維持、検索も含めて、科学者コミュニティやひいては一般社会が合理的と認める共有された知識や認識をつくりあげる。それを「科学」というシステムと考えてよいのではないでしょうか。システムとしての「科学」が存在することで、研究者が研究した結果が、確実に人類の共有する知となっていくのではないか、と考えています。そういう意味では、新たな科学技術成果やイノベーションのためにも、知識や情報の保存管理がいかに重要かということです。

〔村山-8〕例えば、情報を正しく保存することの重要性について、これは『Nature』に発表された論文についての、先ほどの英国王立協会のワーキンググループの方の報告です。簡単にいうと、癌の研究論文を対象にしたもので、インパクトファクターの非常に大きなジャーナルに素晴らしい論文が掲載されて、それを参照して、さらに研究が発展していく過程で多くの論文が出版されるわけですが、調査対象となった論文のうち研究成果を再検証できない論文が多数あり、またそれを引用した論文が二百数十もあると言う結果が出たそうです。つまり、癌の研究論文の結果を基に病院で更に治療方法の研究がなされる場合に、それが途中でおかしかったら、「何がおかしかったのだろう」と研究結果の原典、すなわち過去の論文に立ち戻ることが必要になることがあるでしょうが、その論文をいくら読んでもその実験・研究結果を再現・検証できない。そういう問題を抱えたままで研究を進めていいのだろうか、という問題提起をしています。これは、たとえば巨大地震でも、気候変動でも同じだと思います。社会が科学を信頼して、その結果を皆で「科学知」として共有するならば、その再確認を保障できることが科学の制度の信頼を維持するうえでたいへん重要です。それができないものが現にあるという、非常に大きな問題が指摘された、ということになります。

〔村山-9〕現代においては、「論文から得られる情報は、果たして十分だろうか」ということが国際社会の中で問いかけられています。今挙げたのは記述が不十分、ということもあるでしょう。そのことは今後、我々研究者自身も考えていかなければならない。同時に、研究基盤としての情報が不十分なのではないかということもあります。つまり、先ほどの気候変動の問題でいうと、数式や論考を積み重ねることで新しい科学的理解が得られる論文であるならば、それらは論文にきちんと記述をするべきです。しかし、気候変動の長期変動データを調べた結果を分析・報告する場合には、論文中のどこを探してもその根拠は出てきません。なぜならばその根拠はデータそのものにあるからです。〔村山-10〕したがって、科学的な発見・理解において、特に社会に大きな影響を及ぼす分野では、近年は「データがなければ科学的な真実を共有できない時代」になってきているのではないでしょうか。つまり、科学や研究活動は、ライプニッツやニュートンといった近代科学確立の時代から同じ科学というものとして考えがちですが、実は科学というものが既に色々と変わってきている、という議論もあります。その中で、新しい科学的情報の扱い方を考えるべきじゃないか、ということです。

先ほどお話しした通り、巨大な地震や気候変動というものは、一度起って過ぎ去ってしまうと二度と誰も測れません。気候変動も、今この瞬間も刻々と進行しており、今この時間のデータというものも後からでは手に入らない、そういう意味で、過去を定量的に記録したものとしてデータの重要性は大変大きなものがあります。そういった一過性の現象も、実験室で再現できる種類の検証も、科学的なデータは科学のシステム上、重要・不可欠なものであって、それを科学知の基礎としてどのように共有していけばよいのか、を考えていく必要が今後あるでしょう。論文が固定されたり、評価されたり、公表されるのと同様のプロセスでデータを共有するという概念が成立するのか、それは「データパブリケーシ

ョン」として成立しうるのか、それが今、国際会議等で議論されている問題と認識しています。それは、データもパブリッシュすればよい、と簡単には言い切れない様々な問題を抱えています。電子的手段の課題、情報発信の課題が様々にあると思います。

〔村山-11〕現在の図書館、文献の文化というものは、グーテンベルグが活版印刷による印刷文化を創り出してから既に 300 年以上経っています。その中で我々人類は、文献を参照したり、検索したり、引用したりするシステムを作り上げてきました。ところが、今は PDF などの電子的方法で論文が検索でき、データもオンラインで入手できるものがあります。しかし、それらの手段は未だ生まれてから一世紀も経っていません。私はハードディスクの発明前後当たりに生まれていますが、考えてみれば人の一生の時間をまだ経ていないわけです。すなわち、科学的文献や科学的データが電子的手段でのみ提供される社会が来るとすれば、300 年以上の歴史を持つ図書館文化、出版文化と共に受け継いできた科学的遺産、人類の知的資産をどのように電子に移し替えて使っていけるか、というのが非常に大きな問題であることを改めて認識しているところです。

では個々の事業についてお話をします。〔村山-12〕まず、ICSU（イクス、と読みます）の WDS についてです。〔村山-13〕アカデミーの側から科学データの重要性、その保存や利用について問いかけ、活動している事業です。「ICSU」というのは、「International Council for Science」の略で、1931 年創設ですが、起源は 19 世紀末の各国アカデミーの連合体にあるとも言われています。これが 1950 年代から、各国の科学的データを保存する事業「WDC (World Data Center) 」を実施してきました。並行して「FAGS (Federation of Astronomical and Geophysical data analysis Services)」といった事業もありました。これらのデータ保存事業では、終戦後は紙やフィルムを主体として記録されたデータを、データ保管をミッションにしたある国の機関へ送り、本棚や倉庫にしまって無くならないようにする、という活動が行われました。ところが、あれよあれよという間にインターネットが普及し、ウェブ上で情報がオープンになっていき、「現代的なデータの扱い方は違うのではないか」ということで、制度から作り直すことになり、2008 年に「World Data System」が設立されました。これは今のところ組織であり事業であって、具体的なコンピュータシステムの名前ではありません。この WDS の前身であった「WDC（World Data Center)」の例を見ていただくと、〔村山-14〕WDC では 1950 年代から、このように世界で 50 の国際センターを認定して、データ保全事業が続けられてきました。その 50 のうち七つを日本の研究機関が担当・維持してきたので、日本は非常に大きな寄与をしてきたと言えます。例えば、国立極地研究所のオーロラデータセンターなどは国際的にも重要な地位を占めています。このような組織が世界各国にあります。見ていただくと分かるのですが、地球科学が中心になっている感じがあります。〔村山-15〕一方、WDS は、それだけではなく社会科学や災害、人文科学の研究なども含めて幅広く学術そのものを捉えようということで国際的に進められています。そして一般公開ポリシーのもと、長期的データ保存を共通形式で進め、お互いに利用できる、アクセシブルな形でデータを提供するという理念を掲げて進めています。〔村山-16〕WDS には国際プログラムオフィスという実行部隊があり、これのホストを私

の所属する NICT が引き受けているわけですが、実は、このような国際アカデミーの国際オフィスを日本で引き受けるのはこれが初めてのことです。私は事務局のお世話担当室として特別に作られた研究室の室長をやっていますが、ICSU の体制の下、WDS 科学委員会の委員に私と東京大学の柴崎亮介先生の二名が任命されており、その下に国際プログラムオフィスがあります。WDS に加盟してこのポリシーで動いている機関・団体は NASA や中国科学院の組織など現在、76 機関あり、年々増えています。〔村山-17〕プログラムオフィスのオープニングセレモニーは 2012 年に行われて、日本学術会議の会長、総務大臣、文部科学大臣政務官などもいらっしゃいました。〔村山-18〕各国研究機関がデータを持ち寄って、お互いに使いあうというのがそのコンセプトで、〔村山-19〕データパブリケーション、オープンなメタデータカタログを作る事業などを今進めているところです。

〔村山-21〕次に、RDA についてです。研究データの共有・利用を加速するためのコンソーシアムとして 2013 年に発足しました。G8 の議論があって始まったと聞いています。こちらも結局のところは WDS と同じように、国際的な科学データの共有を目標にしており、WDS と RDA はお互いに連携しながら進めています。〔村山-22,23〕RDA は原則、自由参加ですので、現在、数百名程度の参加者が総会に来て、理事会など様々な運営機構が出来ているところです。〔村山-24〕草の根的に色々な研究テーマを立ち上げてワーキンググループを作る、といった形で進められており、〔村山-25〕中でも Library Science（図書館情報学）の研究者・図書館員が国際的に非常に重要な立場で重要な働きをしている、というのが私にとっては印象的でした。科学データ共有に関して日本国内ではあまり会うことがない方ともお会いできました。〔村山-26〕研究データのオープンデータ化については、「オープンガバメント」とは違うカテゴリとして EU の組織内では動いているということで、ガバメント系とは全く違う議論をしているというのが RDA の特徴です。〔村山-27〕この丸い図を書いたのは、マイクロソフトの人と英国の大学の人なのですが、データについての研究を進める上で、技術的基盤のほかにも法的枠組や研究文化、ビジネスモデルなど様々なことを一緒に解決しなければならない、というのが RDA のポスターセッションでの議論で紹介されました。〔村山-30〕国連をはじめ様々な機関がデータの共有事業を進めていますが、データの共有、引用、再利用のための仕組みがどのようにできるか、今、具体化が進められている最中です。データパブリケーションについてはエルセビア社やワイリー社、トムソンロイター社等が実際に入って WDS の委員とともに動き出していまして、非常に現実味を増しています。〔村山-31〕オーストラリアでも具体的に動いています。〔村山-33〕政府機関もこのようにデータの共有について動き出していますので、「我々は科学データについて、日本として何ができるのか」ということも含めて、今後様々なコミュニティの皆様と議論をしていかなければならない、と思っております。

【基調講演】ドイツ国立科学技術図書館の戦略：研究データの保存と共有

ペーター・レーヴェ氏（ドイツ国立科学技術図書館・ハノーファー大学図書館研究開発部長・ドイツ地球科学研究センター客員研究員）

私は昨年図書館に入ったばかりの地理学者でして、図書館についての専門家ではありません。図書館員のように話したいと思いますが、少し外れてしまうかもしれません。〔レーヴェ-1〕本日のトピックは、TIB の戦略、TIB における研究データの保存と共有についてです。〔レーヴェ-2〕まず、私が属する TIB の紹介をしてから、次に研究データ・マネージメントの状況、コンソーシアムである DataCite について話します。また、TIB の多くの部門から構成されている「GOPORTIS」ネットワーク、更に「RADAR」プロジェクトについても話したいと思います。RADAR は独立した研究イニシアティブとして、科学及び産業界向けに GOPORTIS とは独立した形で行われているリポジトリです。最後に、ビッグデータ時代における図書館の戦略について、将来にわたって図書館はどのように適応していかなければならないか、ヨーロッパ及びドイツの視点からお話しします。将来に向けて図書館がどのように適応していけばよいかについても提言します。

〔レーヴェ-3〕TIB は、ドイツの国立科学技術図書館です。TIB が中心としているのは、工学、建築学、科学、化学、情報技術、数学、そして物理です。創立は 1959 年で、日本と違って中央集中型の国立図書館ではなく、幾つかの国立図書館が分野ごとに設立されています。連邦政府及び州から資金拠出を受けて設立されています。〔レーヴェ-4〕こちらが TIB の本館になります。〔レーヴェ-5〕閲覧室の内部の様子です。TIB は国立図書館としての役割とハノーヴァー大学図書館としての役割と二つの役割を持っているので、多くの学生も使っています。〔レーヴェ-6〕また、昔ながらのいわゆる「マスタール塔」という建物もあります。ザクセン州はもともと王国でしたので、王様のための城があったわけですが、その中の馬小屋を図書館として改装したわけです。〔レーヴェ-8〕その城は現在ハノーヴァーのライプニッツ大学として馬小屋、すなわち TIB 図書館の隣にあります。

〔レーヴェ-9〕これが TIB の購入予算です。出版物購入予算は 1470 万ユーロで、約 5 万 2700 点のジャーナルを購読しており、うち 3 分の 2 がデジタル化されたもので、残り 3 分の 1 が紙媒体のものです。ほかに 900 万点の蔵書があり、職員数は 400 人ほどです。〔レーヴェ-10〕TIB は GOPORTIS ネットワークにも参加していますし、RADAR プロジェクトのパートナーとしても、様々な図書館・研究機関のネットワークに参加しています。〔レーヴェ-11〕様々な国際ネットワークにも所属しています。Open Planets Foundation、WorldWide Science Alliance、TechLib、中国科学技術院、もちろん日本の国立情報学研究所のパートナーにもなっています。〔レーヴェ-12〕利用者の 71%はドイツ国内です。10%ほどが EU 圏内、14%がアメリカ、残りの 5%がその他の国という構成になっています。〔レーヴェ-13〕ビジョンと戦略についてです。TIB は 1959 年の設立当時の中心は紙の印刷物でした。デジタル時代になり、研究データ、科学映画、ソフトウェア、3D オブジェクト、シミュレーションデータ等が主流となっており、それらに対応しなければなりません。〔レーヴェ-14〕2013 年、TIB では新たな研究開発部門を設立しました。館長を中心にしたかな

りフラットな階層になっており、研究開発部門は、その重要性から、館長に近いところに位置しています。

〔レーヴェ-15〕研究データは、簡単なトピックではありません。我々は繰り返し、「研究データの問題とは何か」「既に解決済みなのではないか」「なぜ解決できないのか」と自問しています。〔レーヴェ-16〕氷河のクレバスの上で足を広げている男の人がいますが、現時点ではこの隙間がどんどん広くなっています。出版された論文や文献と、それを支える基礎データとのギャップが広がっているわけです。そのギャップをなくそうとしてもなくならない。つまり、基礎データの発見とアクセスが難しいわけです。これが研究者にとっても、図書館司書にとっても難しい課題になっていると思います。〔レーヴェ-17〕ここで、簡単にまとめてみましょう。研究の経路としては、まず「データ」があります。次にそのデータを分析・解析することで追跡可能な「情報」に変わります。この情報が発表・公表されることで「知識」になり、アクセスが可能になります。これが研究の一連のサイクルですが、〔レーヴェ-18〕それを機能させるためには、新規の又は既存のデータセンターの強化が必要です。持続的な識別子を用いることによって、既存のカタログを通じて世界中からデータセンターにアクセスできるようにしなければなりません。また、新たな技術のトレンドを科学一般の中からモニタリングしていくことも必要と思われます。

〔レーヴェ-19〕現在、科学データの識別子として使われているのが DOI（Digital Object Identifier）です。これは世界中で使われているもので、デジタルネットワーク上での知的財産の相互運用を可能にするものです。DOI の仕組みは、この例にあるように、非常に単純な「Prefix」と「Suffix」という二つの要素から構成されています。例えば、ISBN 番号を識別子として用いることで、アマゾンで本を買ってもそれが何か区別できるように、DOI も研究データのための識別子となっています。〔レーヴェ-20〕科学者にとっての DOI の価値を説明すると、今日、科学者はデータを FTP サーバでアップして、URL によるリンクをつけて公開することができます。しかし、データを保存又は移動しなければならなくなった場合、URL による識別は古くなったら無効になることがよく起こります。例えば、文献の引用元データ、基礎データを見つけたいと思った場合、URL がヒットしてもそれが 20 年くらい前の古いものであった場合、エラーメッセージが出てくる可能性が高いわけです。一方、同じデータが DOI を使ってリンクされていれば、20 年経っても 100 年経っても、さらに遠い未来でもデータに有効なレファレンスを与えることができます。〔レーヴェ-21〕TIB のカタログから最近の例を示しますと、これは本日のスライドのために作ったわけではなく、私が以前勤務していたポツダムにあるドイツ地球科学研究センター（GFZ）のデータですが、たまたま DOI につながっていたために検索できました。これは非常に重要な例だと思います。〔レーヴェ-22,23〕もう一つ、これは Scientific Drilling Database の例です。DOI を使うことで特別なデータセットを参照することができます。このデータセットは Google マップでトレースすることができますし、反対に Google マップを使ってデータセットを検索することもできます。特別な関心領域についてのデータセットを検索することができるわけです。このように、検索効率を高めることで、科学者の仕事を楽にするこ

とができます。

〔レーヴェ-24〕理想的な研究サイクルを考えると、「実験」をすると「データ」が得られ、それを「情報」に加工して「出版物」になります。この四つのステップ、つまりデータアーカイブからパブリッシャーまで、それぞれに DOI を通じて紐づけることができます。黄色い線で示した通りです。我々は、この矢印部分で示していることを今やろうとしているわけです。〔レーヴェ-25〕DOI はそれほど新しい概念ではありませんが、TIB は早くから取組を行っていました。1999 年に出版社らが DOI のための機関「CrossRef」を創立し、その 6 年後に TIB が最初の登録機関として一次データ（primary data）の登録機関となりました。最初の 4 年間はうまくいっていましたが、やがて、一つの登録機関だけでは全世界をカバーできなくなりました。そこで、この DOI 登録業務は TIB から世界中に拠点を置く「DataCite」に移管されました。今でも TIB はその一部として機能を続けています。来年の 2015 年は、TIB が DOI 登録機関として 10 周年を迎えることになるので、ハノーファーでシンポジウムを開催する予定です。

〔レーヴェ-26〕次に、世界中の DOI の登録を手掛けている DataCite についてお話しします。研究者が研究データを特定し、信頼感を持って検索、引用することができるようにしています。また、データセンターには、ワークフロー又はデータ出版のための基準を提供しています。出版社に対しては、文献から基礎データへの紐付けを可能にすることでサポートしています。〔レーヴェ-27〕DataCite は、多くのローカル機関が参加する世界規模のコンソーシアムです。データセットや非テキスト情報に関する学術的基盤を改善することに注力しています。データセンター及び様々なコンテンツを持っている機関との協力を進めており、基準やワークフロー、成功事例などの提供をしています。〔レーヴェ-28〕DataCite は 2009 年に創立して以降、そのネットワークを拡大してきました。いまやヨーロッパ、アジア、北米などにまたがる 19 の本会員組織、10 の準会員組織（日本の準会員として WDS の国際プログラムオフィス）があります。〔レーヴェ-29〕5 年が経過した今年 2014 年、データサイトの状況は、その間、272 のデータセンターから、250 万もの DOI ネームが登録されました。2013 年には 1 年間で 800 万ものコールアップ又はリンク、アクセスが、URL ではなく、DOI でアクセスされるようになりました。メタデータスキーマが発表され、メタデータストアも出されています。〔レーヴェ-30〕次に、こちらのピラミッド図をご覧ください。データの扱われ方を示したものです。一番下のレベルには「科学者」がいます。彼らがデータの生成、ハーベストをしています。つまり、他者のデータを使って新しい情報を再生産しています。次のレベルは「データセンター」で、保存や品質管理のための活動及びメタデータなどを作っているところです。さらに一つ上がると「図書館」があります。検索結果の提供や登録などを行っています。次に GOPORTIS という図書館ネットワークと RADAR プロジェクトの二つについてお話ししますが、RADAR は図書館とデータセンターの間くらいに入るという位置付けです。

〔レーヴェ-31〕まず「GOPORTIS」について紹介します。このライプニッツ図書館ネットワークは、研究情報のためのネットワークです。ライプニッツアソシエーションが有名

な数学者の名前を付けたもので、ドイツにある三つの国立図書館のネットワークによって成立しています。〔レーヴェ-32〕GOPORTIS は、研究の卓越性を保障し、情報科学におけるアプリケーション指向の研究に対して情報インフラを提供し、それを継続的に発展させていくためのネットワークです。また、科学の利益を保護し、政治的な意思決定をサポートし、戦略的に国内外のパートナーとの協力を拡大・維持していく、そして学術的な作業の手法を変革し、ドイツを科学の拠点としていくことをミッションに掲げています。〔レーヴェ-33〕ドイツの国立図書館は一か所だけではありません。国立、公立の機関それぞれに、連邦政府、州政府から資金が拠出されています。それぞれが関連分野の情報、文献、その他の媒体などを収集し、資料へのアクセス提供やアーカイビング、保管をするなどを担当しています。また、特定の関心領域についての文献情報を提供しています。いわゆる灰色文献と呼ばれるものも含めて、ほぼ完全なコレクションを擁しています。ここでいう灰色文献とは、科学的研究であるがジャーナルのような公式の出版物になっていないもので、図書館においてレファレンス用に保管することが合意されている資料を指します。例えば、スライドショーを含むプレゼン用資料などです。このようなコンテンツを 50 年後もみることができるかは、アーカイビングなど別の問題も関わってきます。〔レーヴェ-34〕TIB のパートナーの一つは、ドイツ国立医学図書館という、医学、栄養学、環境学、農学分野を担当している、ヨーロッパで 2 番目に大きな図書館です。ドイツ中部のケルンとボン、2 か所にあります。〔レーヴェ-35〕もう一つのパートナーは国立経済学図書館です。経済学の分野では世界最大で、北部沿岸地域のキールとハンブルグにあります。〔レーヴェ-36〕次に協力・連携についてです。GOPORTIS は、科学的コンテンツの提供、研究とイノベーション、政治的活動の三つの領域において活動しています。協力は運用レベルで行われています。その分野に関心があること、戦略的な意義をもっていることが協力できるかどうかの基準となります。〔レーヴェ-37〕具体的な GOPORTIS の領域は、農学、建築学、経済学、化学、コンピュータ科学、環境科学、数学、医学、栄養学、物理学、技術といったように、かなり広範囲になります。全てのパートナーに関連・影響があるトピックが多くなっています。〔レーヴェ-38〕これが組織図です。科学的コンテンツの提供、研究とイノベーション、政治的活動の三つの分野に別れており、このネットワークにとって科学情報が価値や妥当性、時宜を得ているということを示していると思います。

〔レーヴェ-39〕次に、もう一つ TIB が行っている連携プロジェクト「RADAR」についてお話します。RADAR とは「Research Data Repository（研究データリポジトリ）」の頭文字です。まだオンラインでつながっていませんが、2 か月後につながる予定です。〔レーヴェ-40〕RADAR プロジェクトの対象者は、研究プロジェクト、科学学術団体、科学機関、そして図書館です。その機能は、有名な又は特定の分野に限った既存のデータリポジトリを拡張して提供することであり、それらの横断的なプラットフォームを、API を通じて提供することです。まだ始まったばかりで、昨年の 9 月に立ち上がったところです。2015 年の第三四半期まで開発が行われる予定で、順調に進めば、開発期間は 1 年間の延長が認められています。RADAR は、いわゆるスモールサイエンス及び産業界を対象に情報提供して

おり、二つのステップで取り組んでいます。まず「スターターパッケージ」ですが、分野を問わない汎用的な分野を対象に研究データを保存します。もう一つ、高度な「スーペリアパッケージ」というものもあります。これはスターターパッケージと同じようにデータを保存しますが、データの公開内容も統合するという機能を持っています。こちらはビットストリーム保存、そしてコストモデルを網羅しています。これについては後ほど説明をします。〔レーヴェ-41〕プロジェクトパートナーは、ライプニッツ研究所（FIZ）、カールスルーエ工科大学、ライプニッツ植物生化学研究所（IPB）、ルートヴィッヒ・マクシミリ大学ミュンヘン、そして私たち TIB です。

〔レーヴェ-42〕この表は、研究データのランドスケープで、データ公表が科学者にとってどういう意味をもつのかを説明しています。まず一次データのセットが一番下にあります。これはハードディスクに保存してよいのですが、それが適切な保存場所でない場合にはデータ公表の仕組みは失敗します。その上にデータ収集や構造型のデータベースがあります。これは既にデータセンターに保存されています。ということは、データ公開を阻害する要素が除かれているわけです。その上は、研究データの編集、つまり出版です。これは単独のデータの出版物でもあり得ます。一番上にあるのは本格的な科学データを含んでいる出版物です。そしてそれが研究データに紐づけられ、位置付けられる、というのが理想的イメージです。〔レーヴェ-43〕一方、現在、直面している現実がこちらです。先ほどのピラミッドとは、バランスや構成要素が変わっていることが分かると思います。単一のハードディスクにデータが保存されているということはデータ公表を阻害するリスクを伴っていることを示しますし、その中の 75%のデータはそこに保存されているだけで、一切公開されることのないデータになっています。つまり、存在しないのと同じことになっているのです。これではどこを頼ればよいのか分かりません。また、保管情報はデータのダンピングによってあふれてしまう状況に直面しています。また、データとリンクした記事も非常に不足している状況です。こうした状況を変えるためのデータへのアプローチを採用しなければなりません。これに対し、〔レーヴェ-44〕研究データのデータセンターとリポジトリを強化し、その規模を大きくする、また、一次データの保管は薄い層になり、記事に含まれている研究データが大きくなる、といった状況が求められています。データセンターとリポジトリは、出版物においてデータを補完しあう位置付けとされています。出版社が研究データのアーカイブを必要とし、科学者もそれを重視することが求められます。そのため、データセンターは一般的なデータへのアプローチに加え、科学の特定分野に特化したデータへのアプローチの確立が必要であって、それによって接続性が構築できるということになります。研究データと記事はリンクでテキストデータに紐づけされます。すなわち、非常に改良された出版物になりうるということです。こうした手法により RADAR プロジェクトが、アーカイブや出版、インターフェイスの確立に役立つようになるわけです。

〔レーヴェ-45〕ご存知の通り、デジタルデータ生成は地球科学だけでなく、様々な科学・生物学的領域において急速に拡大しており、その膨大なデータを保管していくためのメカ

ニズムが必要になっています。そこに RADAR プロジェクトの意義があります。データ基盤の管理を促すという、現時点で欠如している機能を提供することで、RADAR はより重要な貢献をすることができます。つまり、情報を公開し、永続的に保存し、出版するシステムの確立です。そのための仕組みは次の通りです。〔レーヴェ-46〕まず RADAR アーカイブにアクセスする際には登録、サインインをしてもらいます。次に、幅広いデータの中から、どういったサービスタイプが望まれるのかを選びます。スターターパッケージでは、特定の保管期間に応じてデータが保存されるという仕組みです。これは非公開のデータが保管されます。スモールサイエンスにとって優位な仕組みかもしれません。例えば、出版社が最初にデータを使う場合は、非公開のデータであることが非常に重要です。ただし、もう一つのスーペリアパッケージでは、同様にデータ保存の機能をもっていますが、こちらは統合されたデータ出版の機能を持っています。DOI を通じて対外的にアクセスすることができ、登録されたユーザーでなくてもアクセスすることができます。それを可能にするためのデータの取込み、ライセンシー、データの変換・検証、データの転送・保存、更には永続的識別子の付与やデータ提供者に対するフィードバック機能を持っています。このメリットは、永続的識別子を使って研究データを活用できること、複数の冗長性をもつ機能を通じて分散型データを持続的に蓄積できること、データの一貫性を確実にし、研究データを任意の期間保存できることです。例えば、ドイツでは特定のリサーチ期間が終わっても、その後 10 年間はデータを保存しなければならないという規定があります。その際、ずっと使われなければ保管されるだけですが、5 年目に使われればその健全性は保たれます。そのため、永続的に該当データを残すための仕組みがあります。また、コストモデルもカバーしています。このコストモデルは、科学的研究モデルに重要なものであり、一括払いができます。一括払いであればプロジェクトの計画とその予算編成に組み込めるからです。いわゆるコストの「見える化」ができます。

〔レーヴェ 47〕では、研究データドメインにおいて RADAR の役割とは何かを考えてみましょう。まず「個人の領域」、つまり科学者の職場領域があります。ここで新しいデータが創造、生成されます。次に「協業の領域」があります。これは研究機関のインフラで、データが処理されるところです。この二つの領域で、良好なデータ、有用なデータ、無効なデータがどれであるのかの選別ができ、また、データの文書化が行われます。三つの目は「公開情報の領域」です。RADAR はここで役立てられます。例えば、データを DataCite や出版社とのリンクによって公開することができます。ここで図書館の役割が発揮されるのですが、ここで保管されているデータは、図書館の目録などを通じての活用、日本や海外においての活用が可能なのです。〔レーヴェ-48〕こちらは研究データ・マネージメントについてですが、一番左に「科学者」がいて、ここでデータ生成がされます。次に「職場」があり、そこでデータが役立てられるのですが、そのデータのリポジトリとして RADAR がデータを保存するという機能を果たします。その次の段階にウェブ・ポータルがあります。API や DOI が活用される領域で、そこから世界のコミュニティに公開されます。将来的には科学的出版物にも紐づけて活用されることが可能であるという考え方です。

以上が RADAR プロジェクトで、ここまでが TIB のこれまでの取組についてです。次に、ヨーロッパ全体の今後のイメージを考えてみたいと思います。

〔レーヴェ-50〕これは、宮本武蔵『五輪の書』の言葉です。「がらりと大局を見る大きな心になって、大が小に変わることであり、これは兵法一つの心構えである」。この英語を直訳すると、「山頂まではいくつもの道があるということを理解しなければいけない」となります。つまり、我々の仕事においても「頂上にいくまでの幾つかの道」を追及しなければならないわけですが、〔レーヴェ-51〕ヨーロッパにおいては「Riding the wave」という EC のレポートがあります。ドイツには「小さな大根（Radieschen）」という名称のリサーチプロジェクトが展開されています。〔レーヴェ-52〕これらは、将来像をイメージするために、どういうチャンスや脅威がこれからあるのかを説明し、先を読むための仕組みをつくる取組です。〔レーヴェ-53〕このブルーの丸のところが今現在、右側が時系列、将来像です。我々の取り組み方が変革されなければ、この薄いブルーの部分、つまり起こりそうな領域にとどまります。もし視野を広げることができれば、薄紫やグリーンの領域まで広がっていきます。また、不確実な要素も網羅できれば、オレンジの領域にも広げることができます。しかし、これはあくまでも先見性を持って提起しなければここの領域まで広がることはできません。〔レーヴェ-54〕「Riding the wave」では「知識は力なり」という構想のもと、これからのヨーロッパでは、デジタル資産の管理が必要だと謳っています。〔レーヴェ-55〕これについてヨーロッパのシナリオは五つあります。まず「科学とデータ・マネージメント」のシナリオです。これは多様な国が関与するヨーロッパでの研究において、プロジェクトやコンソーシアムにおける課題を明らかにし、国際的な場でどのように取り組んでいくかを考えなければならない、というものです。次に「科学と市民」のシナリオです。私のもっているこのスマートフォンは、10 年前の PC よりもずっと機能が向上しています。これがまさに将来像を示しています。非常に膨大な量のデータを素人が作ることができ、科学者がそれを収穫することができる世の中になっています。そうした世の中の変化に対応しなければなりません。三つめは「科学者とデータセット」のシナリオです。これは、ビッグデータの時代に、個々の科学者のレベルにおいてどういったビッグデータが活用され、科学者が研究に使い得るのかどうかを考えなければならない、というものです。四つめのシナリオは「科学と学生」です。これは明日の科学を考えることです。今日の学生が将来的にどのようなネットワークを確立し、どのようなことを学んで、進化していく社会の中でその存在意義を訴えていけるかを考えてもらうためのシナリオです。一方的に大学から教わり、自分の中の知識として蓄えるだけでなく、その知識を活用してネットワークをグローバル的に広げていくという可能性を含めて対応していかなければなりません。五つめのシナリオは「科学とデータ共有インセンティブ」です。研究データの 75%は失われてしまう、というのは先ほど言及した通りですが、ドイツにおいては科学的データが“工具箱”の中に収められていてそこから出ていくことのないという時代になっていますが、そういった古いデータ、例えば地図や旅の記録、写真といったものを取り出して今後データとして活用することの重要性を改めて感じなければなりません。

〔レーヴェ-58〕次に「Radieschen」というプロジェクトについてです。これは、ドイツ国内の今後の科学構想の観点から〔レーヴェ-59〕「今後の図書館の在り方はどうあるべきか」を考えるプロジェクトです。〔レーヴェ-60〕現在ステークホルダーの分析によってその検証がなされており、幾つかのシナリオがあります。〔レーヴェ-61〕まず、「科学にとっての新しいパフォーマン指標」、すなわち学術的パフォーマンスを評価するに際して、ソフトウェアや研究データに基づいて研究を進めていくという考え方です。また、国際的評価システムが確立されることで、データや記事、ソフトウェアを生成する際に得たデータを大きな情報源として科学者が使える仕組みが必要となります。〔レーヴェ-62〕二つ目のシナリオは、「図書館は未来の一部」であるということです。つまり、図書館が進化して、革新的な仕組みをもち、連結された情報のセンターにならなければならない、というものです。また、データの品質管理やキュレーション、アーカイブなどの機能を持たなければなりません。データサイエンティストがそれらを活用する仕組みは非常に重要です。〔レーヴェ-63〕三つめのシナリオは、「データサイエンティストが学問の世界を確立する」というものです。データサイエンティストは、従来の科学図書館からさらに進化した形で、現代的情報プロバイダーを学術界に対して提供していくという考え方です。また、そのデータサイエンティストの職責のデータを取り込みながら、アーカイブし、分析します。〔レーヴェ-64〕四つ目のシナリオは、「新しい役割を担っているデータセンター」です。研究者がデータ・マネージメント、ソフトウェアサービス及びあらゆる種類の出版物にアクセスできる仕組みを構築する、というシナリオです。これは計算センターが進化してこうしたデータセンターになる、コミュニティが使えるようになる、というものです。〔レーヴェ-65〕五つめのシナリオは「ゆるぎない国」というものです。現状は、順調に進んでいるものの、私はあまり望ましいとは思いません。なんらかの理由でイノベーションが阻害されている状況です。ドイツの科学者は国際コミュニティから断絶している状況なので、その阻害要因を取り除かなければならないと思います。

最後のトピックに移りたいと思います。〔レーヴェ-66〕これはドイツのシナリオからの教訓としてですが、今後の取組を変革・進化させていかなければならないと思います。〔レーヴェ-68〕このヒストグラムは Google トレンドのものですが、グリッドコンピューティング（青）、クラウドコンピューティング（赤）、それからビッグデータ（黄）を比較検証しています。こういった領域機能は重要なトピックでして、グリッドコンピューティングはもはやホットなトピックではありません。クラウドコンピューティングも衰退時期にあり、代わりにビッグデータが注目されるようになっています。トレンドには、浮上し、やがて衰退していくサイクルがあることを示しています。〔レーヴェ-68~70〕破壊的（disruptive）テクノロジーというものもあります。これは「予測できないテクノロジー」です。一つの実例はマウンテンバイクです。マウンテンバイクはかつて市場には存在しませんでしたが、従来の自転車を山でこぎ始めた時に需要が発生し、今では破壊的テクノロジー、スキマ市場として確立されています。印刷物の書籍についても、グーテンベルグが高く評価されていない時期、インターネットが予測されていない時期がかつてありましたが、それらの技

術の持つ意味が認識されるようになったのもここ数年のことです。

〔レーヴェ-72〕続いて「Gartner Hype Cycle」について説明します。最初に、技術的トリガーがあってそれが市場に導入され、期待を膨らませる時期があります。次に、その期待が大きく拡大し、驚きの要素を提供している時期があります。コンピュータ、パソコンの時代がそうです。昔のコンピュータは 64kB でしたが、それが様々な機能をもって市場に出ていました。それが衰退時期に入り、64kb のコンピュータでは求められている機能が実現できない、という時期に入りました。そして再び、いろいろなところでパソコン機能を活用できるトレンドが浮上していますが、現在期待が膨らんでいるのがこの 3D プリンタの技術です。〔レーヴェ-73〕つまり、図書館員にとって何が重要で何が重要でないのか、ということを自己類似型のハイプ・サイクル（The hype cycle's self-similar hype cycle）で検証することが必要なのです。これはヨーロッパで普及している考え方ですが、現状は衰退しており、単なるグラフィックに過ぎなくなってしまっているというのが現状です。

〔レーヴェ-74〕まとめに入りましょう。〔レーヴェ-77〕今後の図書館の在り方、利用者ニーズへの対応の仕方はどうあるべきか、という点について言えば、研究のインフラ提供を成功させるためには、モジュール化されたサービス、共通プラットフォームに基づいたサービスでなければなりません。科学の変容するニーズに応えるためには、共通プラットフォームに基づくモジュール化されたサービスに加えて、その安定性を確立することが重要です。長期にわたる科学の進化を可能にするためには、こういった基礎プラットフォームの構築が重要になるでしょう。また、こうした基盤の安定性とアプリケーションの柔軟性のギャップを埋めることも必要になるでしょう。それが、我々の期待する、「頂点に到達するための道筋」であるということなのです。

【事例報告 1】農業研究におけるデータ共有の実態

木浦卓治氏（農業・食品産業技術総合研究機構　中央農業総合研究センター情報利用研究領域　上席研究員）

〔木浦-2〕まず、農業研究は単独の学問でないということにご注意ください。複数の学問が関わって農業に対して研究しています。違う学問でも、観測の対象、研究の対象は農業だということです。農業の場合、多くは環境をコントロールすることができませんので、一度採ったデータは二度と採ることができません。その意味で、ちゃんとデータは残しておかなければならないと常々思っていますが、これがなかなか困難なことになっています。

〔木浦-3〕農業研究に関係するデータベースには、国際連合食料農業機関（FAO）が運営している「AGRIS」があります。農業関係の文献データベースとしては最大のものです。今度新しくなりまして、これについては、カレントアウェアネス-E252,2014（http://current.ndl.go.jp/e1523）をご覧ください。そのほかに言語資源として「AGROVOC」、このシソーラスとコンセプトサーバ、言語間の関係を定義したオントロジーを FAO が作っています。あと、全体的な統計データとして「FAOSTAT」があります。

ただ、FAO が世界中のデータの記述をしてくれるのはよいのですが、日本独自のことを

記述しようとすると、足りません。これに関して私どもの方でも頑張ってやっているところです。〔木浦-4〕農林水産研究情報総合センター（AFFRIT）というところがあり、そこは「AGROPEDIA」を通じて幾つかのデータ、衛星や基礎数値に関するデータなどを提供しています。〔木浦-5〕研究者の方に限って提供しているものに「Satellite Image Data Base（SIDaB）」があります。この中の衛星観測のデータベースとしては三つありまして（WeSIDaB、MoSIDaB、Landsat）、基本的には NOAA からもらったデータが入っています。MODIS のデータベースだけ自社でデータをとって加工していましたが、今はもらっています。衛星データを使う人は限られているということで、今後の運用をどうするか検討しているところです。現在、DIAS（地球環境情報統融合プログラム）と SIDaB とは関係がありませんが、どうにかしてデータを共有したいと考えています。〔木浦-6〕また、農業にとっては気象データが非常に重要です。この農林水産基礎数値データベース（NDB）には、農林水産省が作っている統計データや国土交通省が作っている数値データ等があります。NDB の利用には認証が必要で、誰でも使えるというわけではありません。ただし、ユーザー登録は受け付けてくれるという状況です。

このほか、農林水産省関係の独立行政法人がいろいろなデータを提供しています。〔木浦-7〕これは農業環境技術研究所が提供しているデータベース・画像情報のページです。書いてある通り、システム更新のため、一部のデータベースのサービスを休止しています。これは農業環境技術研究所自体がシステムの運用をしており、一生懸命システムの更新をしている関係で継続的にデータを提供できる環境に今のところなっていないということです。「モデル結合型作物気象データベース」というのがありますが、ここの一部のデータが、先ほど紹介した DIAS に入っていっています。〔木浦-8〕独立行政法人の農業生物資源研究所も、同様の大きな二つのデータベースを持っています。一つは「農業生物資源ジーンバンク」です。これは種子の管理をしているときに種子に関するデータがすぐに出せるようになっています。種子があったとしてもそれがどういうものか判らなければ使い物にならないので、こういうデータベースがあります。もう一つは「DNA Bank」で、こちらは DNA の研究をする方々のサポートをしています。ご存知の通り、DNA の研究者はデータの共有が最初からできていて、論文を載せるためにはデータを共有しなければなりません。そのため、こうしたシステムがないと論文すらかけない状況です。

〔木浦-9〕私が所属している中央農業総合研究センターでも幾つかデータベースを運営しています。「メッシュ農業気象データ」、「カバークロップ DB・雑草 DB」といったものや、「Field Server Data Archive」というものがあります。Field Server とは、私たちが作ったワイヤレスのセンサーネットワークの観測ノードでして、これで圃場を観測しています。特徴は画像があるということで、データは基本的に全部公開しています。データは DIAS にミラーしており、一生懸命メタデータを登録させてもらっています。「MetXML with METBroker」というものもあります。随分前から気象データは公開されているものがたくさんあるのですが、どこにあるのかわからない、データのある場所によってデータの取り方が違っている、個々の研究者ではとてもじゃないが堪らないので、それを解決しましょ

う、というものです。簡単な URL を叩けば欲しい場所の気象データを手に入れられるようになっています。また、これらの他にも様々なデータを提供していますが、ここでは割愛します。

〔木浦-10〕最近の売りは「メッシュ農業気象データ」です。1km メッシュの日別気象データと、気象庁からもらっている通知予報データをやはり 1km メッシュ化して、観測データと予報データをシームレスに提供しています。最大 1 週間先までの予報が行えるというシステムを提供しています

〔木浦-11〕農業機械分野では、国際標準化が進んでいます。農業機械の中でどのようにデータを提供して更新するかについて ISO 標準として定められています。ISO 標準に則った機械に端末をつなげると、勝手に通信してデータがどんどん溜まっていくという仕組みになっています。この表示端末は同時にデータロガーであることも多いのですが、これらのデータは基本的に農業機械を提供している会社、表示端末を提供している会社にストアされます。残念ながら、そのデータをどのようにストアするかは規定がありません。このため、実際に出たデータの交換方法については考えられていないのが現状です。

〔木浦-12〕そのほか、農業関連の独立行政法人や大学でも農業関連のデータはかなり出ています。全てを網羅できないので一つだけ紹介します。「つくば知的資源サイバーモール」というもので、つくば地区で公開されているデータや自由に使える公開実験施設をリストアップして検索できるようにしています。データは人手で一生懸命にかき集めていますが、現在、このデータ自身をオープン化できないかということで、つくば WAN 情報資源共有研究会と協議を進めています。

〔木浦-13〕先ほどクラウド関係がだんだん落ちてきているという話がレーヴェ先生からありましたが、農業関係の方では、クラウドコンソーシアムという組織があって活動しています。私の知人もやっていますが、昨年からページが更新されていないのでどうなっているのだろう、と心配していたところです。「データをどうするか」、いろいろ考えているようです。〔木浦-14〕その他、農業クラウドサービスも幾つか立ち上がっています。富士通では「Akisai」というのを提供していて、これは海外でも結構有名です。日本語ページしか見当たらないので海外で聞かれても困っているのですが。また、NEC も提供していますし、ソリマチ、農業関係のソフトウェアでは有名な会社ですが、そこも出しています。ですが、基本的にこれらのシステムは「Software as a service」になっていまして、全て自分たちのサービスの中で閉じています。データの互換性はありません。どうやってデータを交換するか別途協議しなければならない状況です。〔木浦-15〕そのために CLoud Open Platform（CLOP）協議会というものがあります。ここで API と XML スキーマを収集していて、API 層で統合できないかということを模索しているようです。富士通のほか、中央農業総合研究センターも係わっていますが、私は少し距離を置いて係わっている状況です。

〔木浦-16〕遺伝子源のデータベースについては先ほど DNA Bank を紹介しましたが、他にもどんどん出てきています。しかし、表現型の方のデータは、従来の方法で集めていることが多くなかなか手に入らない状況にあります。もちろん、大規模な研究施設でインハ

ウスやグリーンハウス、植物工場のようなところで自動計測をしている事例はいっぱいありますが、それは実際に農業でものを育てる場所とは違う環境です。そうした状況で、表現型のデータの共有というのは非常に重要です。イネについては 国際稲研究所（IRRI）が主導して、「Global Rice Phenotyping Network」を作っています。重要なのは、表現型のデータの収集にはとても時間がかかるので、ICT を利用した効率的なデータ収集が必要といわれている点です。

〔木浦-17〕まとめますと、農業研究では多様な情報を利用します。FAO は世界的な統計データや文献データ、言語資源など世界の農業技術のためのデータを提供しています。AFFRIT は農業研究の独立法人や大学の雑多なデータを出そうとしています。ですが見ていただいた通り、これらのデータは同様に分野が限られています。ほとんどのデータは研究者の手元にあるままです。農業機械の方では ISO 標準のデータのやり取りはあっても、それは機械の中だけで外には広がっていない。農業クラウドサービスも立ち上がってはいるが、共有されていない。農業関係でも情報そのものはかなり持っているが、有用な情報を発見して分析できるような状況には残念ながらなっていません。私としては「ミニ DIAS」の農業版のようなデータサービスができたらいいな、と思っています。

【事例報告 2】SSJ（Social Science Japan）データアーカイブにおけるデータの保存と普及

佐藤博樹氏（東京大学大学院情報学環教授〔社会科学研究所兼務〕）

私は学内出向で 3 年間だけ情報学環に所属ということで、この 4 月には社会科学研究所に戻ります。本日は、社会科学研究所で 20 年近くやってきた授業についてお話します。

〔佐藤-2〕SSJ（Social Science Japan）データアーカイブは、社会科学分野、特にマイクロデータの収集・整理・保存・提供を行うデータアーカイブです。マイクロデータについて説明すると、世の中にはいろいろな社会調査、世論調査又は統計データがあります。例えば選挙のとき、新聞社が投票所の前で出口調査をやって、開票よりも早めに速報が出たりします。翌日「なぜ今回は自民党が勝ったのか」といった分析が出ますが、新聞に載るのは、年齢別の支持政党、男女別の支持政党といった集計（aggregate）されたデータです。我々は集計前のデータを収集し、整理・保存・提供するという事業を行っています。その社会的機能は「実証的研究における再現性の担保」です。例えば、投票行動を研究している政治学者が大規模な科研費をとって、選挙前後の投票行動を研究し、「こういったことが今回の政権交代に大きな影響を及ぼした」といった論文を書いたとします。別の人が同じような手続を踏めば同じ結果が出るかどうか、それが「再現」できるかどうかということです。もちろん「もう一度調査すればよい」という話もありますが、同じ選挙は二度ないわけです。ある研究者が集めたデータを別の研究者が同じ手続でやれば同じ結果が出る、ということが大事です。そういう意味では、誰かがやった研究を自分だけが見られます、ではなくて他の研究者もアクセスできるようにする、ということが非常に大事です。

研究者一人ひとりがデータを保存し提供する、リクエストがあれば提供するということでもよいのですが、これだと手間暇がすごく大変です。海外では、ドイツだと「ZUMA（Zentrum für Umfragen, Methoden und Analysen）」、イギリスでは「UKDA（UK Data Archive）」、アメリカには「ISPS（Institution for Social and Policy Studies）Data Archive」といった大きなデータアーカイブがありますが、日本では残念ながら我々が作り始めるまでそういうデータアーカイブはありませんでした。極端な言い方をすれば、社会学の研究分野で査読雑誌に載っている論文でも、今検証することはできません。データがありません。そういう研究がたくさんあります。そういう意味で、少なくとも我々にデータを寄託していただければ、そのデータについては検証できるようになっています。

実は、検証できることも大事ですが、もっとプラスになることがあります。例えば同じデータに基づいて、ある人がやった仮説をもう一度検証できるというだけじゃなく、同じデータに基づいて異なる仮説を立てて議論ができるようになるわけです。そうでないと「データが違うのではないか」、「僕のやったデータでは違う結果になっている」という議論が出てくる。同じデータを使って仮説を競わせる、理論を競わせるということが可能になるわけです。そういう意味で、研究の深化にも大きく貢献します。これは実はとても大事なことです。さらには、最初に調査を設計して実施した人は考えていなかったが、別の研究者が同じデータセットを使って別の研究ができること、例えば「選挙の時に誰の意見を聞きましたか」という回答をソーシャルネットワークのような別の観点から使うなど、という既存データを活用した研究、これを「Secondary analysis」といいますが、そういったことが可能になってきます。その意味では、一つの大きな研究資源、例えば、全国調査をすると何千万とかかるが、そういうものは一回きりではなく、他の方も使いたい。特に若手の研究者は実証研究をするための十分な研究費をなかなか取れません。従来は、大きな研究費を取れる研究室にいる若手研究者はデータにアクセスできても、そういうところにいない若手研究者はデータにアクセスできない、論文が書けないということもありました。あるいは、海外の研究者が日本に来たときになかなかそういったデータにアクセスできない。アメリカだったらすぐにデータが手に入るのに、日本に来たら入らない、がっかりして帰る、ということがあります。また、「日本でデータが手に入らないのだったら、アメリカのデータでアメリカの研究をしよう」ということが起きてしまうこともあるわけです。そういう意味で、データの共有化は、単に再現性の担保というだけでなく、データの有効活用や若手が実証的研究をできるためにも大事なことです。

それからもう一つ大事なのは、特に社会調査の場合は、企業の調査ですと毎日調査票が送られてくる、個人でも様々なところから調査票が送られてきます。すると、「もう答えない」、「協力しない」ということもあったわけです。つまり、皆がデータを公開しないとどういうことが起きるかというと、皆が研究費を取って皆が同じような調査をやる、ということになってしまう。研究費の無駄というだけでなく、調査対象になる人の調査負荷が非常に高まります。これは「調査環境の悪化」と言われています。要するに、データアーカイブがないことによっていろいろなところでマイナスの状況が出ていた、ということにな

ります。

〔佐藤-3〕そうしたところで、我々は1998年にデータアーカイブを立ち上げたわけです。もちろん、先進的に行っていたところはありますが、我々の場合は社会科学分野の領域を問わず、ソーシャルサイエンス全体のマイクロデータを集めるということをやっています。そういう意味では、おそらくアジア最大だと思います。最初にそういう形ではじめたのは立教大学です。まだ20、30のデータセットだと思いますが。あとはレヴァイアサン・データバンクですが、これは選挙意識や投票行動についての調査データで、選挙データのアーカイブでは我々の先輩になります。神戸大学の三宅一郎先生はじめ、いろいろな先生たちが係わっていた先進的なデータです。その後、調査研究機関が自らの機関で行った調査データを提供するという事例が出てきました。国のデータについては、統計法にカバーされていて我々が扱えません。ただし、国のデータもパブリックデータとして、研究者が分析したときに回答者を特定できないようデータを加工した形での提供が始まっています。今年は国勢調査、センサスのサンプル、例えば10%抽出のデータが公開されるようになるはずです。国際的にみれば遅れているわけですが、政府統計についても、匿名データの提供がはじまったということになります。〔佐藤-4〕我々はこうしたことを早めに、1998年に始めたということです。

〔佐藤-5〕次に、我々はどういうことをやっているかについて話します。四つの活動からなります。一つはデータアーカイブの事業です。二つ目は自ら調査データを作る、つまり、公開を前提としたデータの作成です。なかなか研究者はデータを提供してくれないので、我々がデータを公開するモデルを作る、ということで始めたものです。日本の場合、残念ながら科研費を使って大規模な調査をしてもデータを寄託することが義務になっていません。アメリカであれば、「データをいつ公開するのか」ということを書かないといけない。できるだけ早く公開する、というところにお金が付くわけです。ただ、少しずつ変わってきたのは「公開する」ことを書かないと科研費の審査で落ちるということになってきました。三つ目の活動は、「二次分析の普及」です。これは、データアーカイブをつくり始めたが、なかなか利用者が出てこないためです。既存調査を使う「Secondary Analysis」は、結構難しい面もあります。どういうことかというと、自分が仮説を考えてその仮説を実証するような調査を設計する場合は、自分が必要な変数を設問に入れられるわけです。ところが、Secondary Analysisの場合は自分の仮説にぴったり合う変数になるデータがないので、第2変数、その変数に近いような変数を工夫して作っていくか、仮説を変えていくといった工夫が必要になってきます。また、日本の場合は特に、分野にもよりますが、社会学などですと既存データに基づく論文が研究として評価されず、自分で調査しなければいけないという風潮があったりします。海外を見ると、アメリカ社会学研究のジャーナルでは、かなりの部分がSecondary Analysisです。特に、修士論文の1章や2章はSecondary Analysisを書くわけですが、日本はまだまだそうなっていなかったわけです。そういう意味では、既存データを使って分析するというやり方が、特に若手でわからない研究者が多かったので、二次分析の啓発もやってきました。四つ目の活動は、海外のデータアーカイ

ブ、例えば、ドイツの「GESIS -Leibniz-Institut für Sozialwissenschaften」、「ZUMA」、アメリカの「ISPS」、「IFDO（International Federation of Data Organizations）」などと連携し、海外の先進事例を学びながらデータを作ってきました。

〔佐藤-6〕どのようなデータを収集しているかというと、海外のデータアーカイブと同様に、研究分野を限定せず、社会科学全般のデータセットを収集しています。また、個人調査だけでなく企業調査も集めています。我々はデータを集めるだけではなく、マスキング、匿名化などデータの整備も行っています。特に個人調査の場合、研究者が色々分析した際に、誰が回答したのかわかってしまわないようにすることは大事です。「年齢が 120 歳、北海道」というようなケースですと、トップコーディングをします。あるいは地域について、例えば愛知県で何万人もの製造業というと「豊田市かな」とわかってしまうことがあるので、地域をブロックでつくるなど、いろいろと調査データに応じて、マスキング作業をします。日本の場合はいろいろな調査機関が調査をしていますが、全部集めると大変ですので、研究機関の行っている大規模調査や大規模な科経費によるもの、継続的なもの、例えば NHK の「日本人の意識調査」、これは 5 年ごとに同じ調査票でやってもらっている調査ですが、あとはパネル調査など、同じ個人を追いかけているというようなもの、官庁の委託調査といったものを基本的に収集範囲としています。〔佐藤-7〕ここにはありませんが、化粧品のポーラで行っている美容に関する調査などもあり、いろいろな調査を寄託してもらっています。

〔佐藤-8〕実際にどんな運営状況かということですが、現在 1500 データセットを提供していて、毎年 80-100 データセットは増えている状況です。提供数は毎年 2,300 データセット、利用者数は 2,500 人くらいです。実際にどのくらいの論文ができているかというと、社会調査のデータ分析の授業に使う場合もあるため、論文数は 140～150 と利用者数よりも少ないですが、これらは二次分析によって作成されたものです。つまり、自分たちで調査しなくても、既存データセットを使ってジャーナルに投稿してアクセスされる、あるいは博士論文の一部になっている論文が書かれているということです。そういう意味では、いろいろな方に研究できる機会を提供できていると思います。

〔佐藤-11〕次に、運営上の課題についてです。一つは、データアーカイブは、データを寄託してくださる方がいてできる事業、ということです。ですので、利用者の方には「寄託者に報いる研究をしてください」と書いています。当たり前のことですが、データを利用して論文を書いたら、どこからデータを入手し最初の調査者は誰かということを参照させる、ということはやってもらえますが、書いた論文を我々の方に 2 部送付する、そのうちの一部を寄託者に渡す、ということがなかなか守られません。データを使うときは熱心なのですが。「はやくデータを提供してください」といっても、その後のアクションが遅い。寄託者からしてみれば、「我々が寄託したデータを使ってどんな研究がでたのか」ということに大変関心があります。そういう意味で、良い研究をして、その結果を寄託者に返す、というのは非常に大事ですが、なかなかそれが守られない。海外のデータアーカイブでもそういったことが起きるそうです。また、我々は、海外も含めた研究者の方に提供してい

ますが、国内で非常に著名な先生で、利用はしても自分が所蔵するデータの寄託には無関心、ということがあります。この辺りの意識を変えなければなりません。若い研究者はデータアーカイブのデータを使って論文を書く、ということを大学院のときにやってきていますので、研究者になると積極的にデータを寄託してくれる、というように変わってきてはいます。〔佐藤-12〕運営上の課題で一番大きいのは、単にデータを預かって提供するだけでなく、その間にマスキング作業やメタデータを作るなど、様々なデータの整理することです。データは、エクセルやアスキーなど、いろいろな形で入ってきます。そういうデータにラベルを付けて SPSS や SAS といった汎用的な統計パッケージで処理できる形にします。利用申請の際に「SPSS で使いたいです」と言われると「SPSS」のファイルで渡します。既にラベルがついていてすぐ動かせます。また、我々が公開する前にデータのエディティング （検票）をやりますと、正直、ひどい調査もあります。そういうのはあまり言えませんが、これは正しいデータ、ということではなくて、バージョン１、バージョン２、バージョン３という形で、最初その研究者が書かれたデータを、このようにクリーニングし直しました、というものをつけて出したりもします。これは間違っているというわけではなく、いろいろな論理矛盾とか出てくるのに対応するためです。利用者から「ここを直した方がよい」といったことを連絡していただき、新しいデータセットのバージョンを提供する、ということもやっています。そういったことで、ここでの作業には一定のスキルを持った人が必要ですが、こうした「データライブラリアン」をどう育てるかが大事です。海外に行きますと、各データアーカイブにデータライブラリアンがいますし、大学図書館でも、大学院生に本のレファレンス「こういう研究をするが、どんなものを読んだらいいのか」以外に、データのレファレンス「こういうデータセットがあるか」を相談してくれる、データライブラリアンがいます。我々のデータアーカイブにもデータライブラリアンは必要ですし、やはりこれからは大学図書館で、紙だけではなく電子データの利用に関するライブラリアンも必要と思います。あと、DDI（Data Documentation Initiative）、これはデータセットの標準フォーマットですが、その書式を使うとメタデータとコアのデータが同じ書式で管理できます。ヨーロッパは既にそうなっています。例えば、台湾、韓国にそれぞれあるデータアーカイブで同じ形式でデータを管理すると、一括でデータを検索できるようになります。あとはデータを提供しなくとも、ネット上でデータの集計ができるようになります。我々は既に始めていますが、マスキング前のデータを使いたい場合、ネット上でのデータ処理も有効と思っています。「データアーカイブ」といいつつ、システムやコンピュータに詳しい人材も必要で、これもなかなか大学で確保するのは難しいと思います。そういう意味で、この分野の専門家をどう育てて、評価していくか、ということも大事であると思っています。

【鼎談】研究データ・マネージメントの将来像：図書館ができること

喜連川優氏（国立情報学研究所・東京大学生産技術研究所教授）
ペーター・レーヴェ氏
村山泰啓氏：モデレータ

【鼎談冒頭の喜連川優氏によるショートスピーチ】

昨年 4 月から国立情報学研究所（NII）の所長をしつつ東大に出向している形をとっています。NII の最も大きな役割は、サイエンスの基盤を支えるネットワーク、SINET を運用しているということです。いま約 800 の大学・研究機関が参加していて、国立大学は当然ですが、私立大学も半分くらいは加盟しています。SINET の利用例としては、先日ヒッグス粒子というのがノーベル賞をとったのですが、あのデータは、実は欧州原子核研究機構（CERN）の大型ハドロン衝突型加速器（LHC）から日本に全部来ています。また、ALMA（Atacama Large Millimeter/submillimeter Array）の場合は、5000 メートルのところから引いていまして、この回線も SINET がプロキュアしているところです。

面白いのは、今のサイエンスは、実験をしている人よりもデータを前にして考えている人の方がはるかに多い、という時代になってきたということです。これは、特に「ビックサイエンス」と呼ばれているような、非常に大きな装置を使う場合に、その装置から生まれるデータをたくさんの研究者が使うようになってきたことがあります。その時、データ基盤とネットワーク基盤が極めて重要になってきます。事例として先ほど話題になった「DIAS」があります。地球環境というものは、観測対象として極めて自然な研究対象と思います。地球の健康をどう維持するかは、人類にとって非常に大きな課題です。センサーネットは山のようにありますが、それらを横串でデータ解析する基盤がないということで我々がそれを作ってきました。いわば大きなデータのインフラですが、30 年ほどかけてコツコツとこの中に様々なデータが投入され、計算機のシミュレーションの出力も、あるいはリモートセンシングのスペースエージェンシーのデータも、地上観測点のデータも全てが 20PB 程度の大規模なデータに入っています。このデータは東京大学だけでなく、特に 3.11 以降は電力不足の関係もあり、いろいろなところに置いています。

重要なメッセージは、「データを集める」ということが主体になっているのではなく、それをどう使うかということが国民目線としては映ってくると思います。その一つの例ですが、最近日本は地震もですが洪水でも随分被害が出ていて、その洪水にどう対処していくかという視点で地球環境のデータを見るということがあります。グローバルな全球モデルのデータがありまして、毎日気象庁が観測状況を提供しているわけですが、そういうものでデータを同化し、近似解をどんどん作っていきます。それから予測モデルを作り、最終的に川の上に雨がどう降るか、ということを考えます。川には高低差がありますから、どこからどこへ水が流れるかは計算できます。川が水を取り込むことができれば洪水は起きないのですが、反対に取り込めないと大きな洪水になってしまいます。日本は、戦後開発が進む中で膨大なダムを作りました。自然に対して人間が作用することはなかなか困難で

すが、このダムを利用して調整することができます。雨を予測してダムの水を放流し、そこに雨が降って川の流量がこのくらい上がる、ということが予測できるのです。そんな絵に描いたようなことができるのかと思われるかもしれませんが、平成 23 年の台風 12 号と 15 号をある時点で計ってみたところ、実測と予測とがかなり一致することが分かりました。レーダーからのリアルタイムのデータ、河川のデータなど、ありとあらゆるデータをフュージョンしていくと、過去のストックを元にして、ダムをどのくらい開けたらよいかの予測ができます。これは非常に複雑なプログラムで動くのですが、このユーザーインターフェイスをいま国土交通省に見せているところです。

ですが、失敗すると大変です。ダム（の水位の違いが）1 ㎝で、ざっくりと 1 億です。失敗すると、農家の方々がお持ちになっている水がスッと無くなって、東京大学が怒られることになるわけです。ただ、そういうことをしないとやっていけないくらい、気候の変化が激しく、極端事象の多い温暖化の中で、我々は次の手を打とうということになっているわけです。これは日本だけでなく、ベトナムのハノイにもダムがあり、またはフィリピンのマニラでも同じようなことがあります。こういった地域は、国際協力機構（JAICA）やアジア開発銀行（ADB）から随分支援いただいており、我々のデータはひと月に 100TB くらい、いろいろな国からダウンロードされています。つまり、データを蓄えると同時に、開発途上国においてはこういうデータを利用して、20 年後、30 年後を見据えたデザインをする。そういう時代に入っているということです。

また、我々は IT メディアを収集していまして、オープンになっているメディアに、ウェブ、ブログ、twitter の三つくらいがありますが、それらの膨大なデータを過去 15 年ほど集めてきています。15 年前に twitter はありませんでしたが。こういうものもデータも集めるだけでは不十分で、社会学研究者、若い研究者の方々にも活用いただければありがたいと思っています。丁寧に集めたデータの方が価値は高いのかもしれませんが、網羅性という意味では、IT メディアから集めたデータから補足できる部分も少なくはないだろう、というように考えています。世界で起こっていることのほぼ全てがサイバー空間に入っている中で、IT をうまく利活用できるのではないか、ということです。例えば、東日本大震災後にあった「ヤシマ作戦」（非公式の節電キャンペーン）の動きをみてみたところ、大震災 3.11 の直後にブログは 3 月 12 日から始まり、Twitter は 3 月 11 日から動いています。これらのディフュージョン（拡散）のパターンが図示化できます。情報が伝播するパターンというのは、明らかに最近のマイクロブロギングといったものの方にサブスタンスがあるということで、こういうものが利用できる時代になった、ということです。こういったデータについては、社会学の先生方にもご利用いただける部分があるのではないかと思っています。もう一つ、「情報がどう伝わるか」ということですが、震災のとき、日本人の情報の伝播というのは極めて草の根的でした。今までマーケティングであちこち騒がれていたのが、いわゆる「シングル・インフルエンサー・モデル」でした。声の大きい人がワッと発言すると皆が同調する。ところが、震災の時の「節電をしよう」という草の根の声は、全然そういうパターンではなく広まっていったということは、日本の「つながり感」を非

常に典型的に示したパターンになっているように思います。こんな風にデータはいろいろ利用できます。実は文法解析もできまして、「何々が不足する」ということを解析しますと、電力やガソリンが不足している中で、実は「情報」が不足しているのが分かります。何をしているんだ、IT 関係者は、といったお叱りもあります。これは 3.11 以降、福島第一原発で何が起きているかわからないということが一番の不安要素であった、ということを如実に表わしているデータだと思います。また、形容詞も入れることもできまして、震災のときに、「音が怖い」というデータがあります。津波の時に家がミシミシっと壊れる音が怖いのかな、と思いましたら、そうではなく、「昼夜を分かたずに鳴る緊急地震速報の音がどうにも怖い」という PTSD が生まれている、といったこともこの中から感じ取れるように思います。

このように、社会学上のいろいろなデータ、それから先ほどお話にあったサイエンスのデータは、種類は違うかもしれないが、様々な役割を果たすようになってきています。このためには非常に多くの基礎学問が必要になってきます。話すときりがありませんが、この他、レセプトといいまして、全国合わせると 370 億レコードという、「皆さんがお医者さんに行かれたときにどんなお薬を飲んだか」のデータがあります。これを分析しますと、非常に多くの日本の病態、医療経済の情報がわかります。このポイントは何かというと、ビッグデータ時代において、サイエンスデータや人のデータについて、保存のスタンスで議論することも重要かもしれませんが、いかにそれを活用するか、について議論することも重要ということです。データを活用するときには IT が必要である、ということは、皆さんおそらく同意されているでしょう。レーヴェ先生の発表の中にもあったと思うのですが、「free portability」というものを考えたときに、データだけで十分かというと、サイエンスの場合はコードがついてないとほとんど意味がありません。従いまして、データサイテーションだけでなく、コードサイテーションというものもしっかりと作る必要があると思っています。最後に、誰がそのインフラを作るのか、誰が費用を負担するのかといったことも議論することが必要と思います。簡単ではありますが、議論のためとしてお話しました。

【鼎 談】

村山：喜連川先生の今のお話も大変参考になり、面白い講演を聞かせていただいたと思います。興味深かったのは、最後の方でコードサイテーションのアイデアについてお話されていたことで、研究をされる上でいろいろなアイデアや関連情報が結びついてきちんと使われることは、プリザベーションにもそうですし、次のノイノベーションにも非常に重要だというご提案だと受け止めたのですが、レーヴェ先生はどうお考えですか。

レーヴェ：先ほど喜連川先生のおっしゃったポイント、つまり誰かが資金を出さなければならないということは非常に重要で、資金を拠出するに当たってはそれを計算しないとダメだし、国が出してくれるなら事前に計画を立てなければならないということがあると思います。

村山：喜連川先生は、そのあたりはやはりご自分のビジョンの中にお持ちですか。

喜連川：これは非常に難しい問題です。例えば、データというものを蓄えなければいけないという皆さんの気持ちは、蓄積するに値するデータという前提で議論が進んでいます。ところが、多くの実験をしたときのデータというのは、本当に残しておかなければならないデータよりもはるかに多くの、残しても仕方がないということもないのですが、希少性が判断不能な“柔らかな”データがいっぱいあるわけです。

同時に、先ほど私はコードが重要だという言い方をしましたが、コードというのは、自分の興味のあるところだけをたまたま動かしています。他の人がそのコードを使ってその人と違うものを動かしたら動く保障なんて 100%ありえない、というくらいコードというものはボロボロです。こんなことを言うと、「僕のプログラムはそんなじゃない。もっと上等だ」と反論される方もおられるかもしれませんが、一般的にはそうです。

この二つはかなり共通していまして、論文を残すというのは極めて大きな努力をして、クリスタライズされた最終知を残す。そういう意味で非常に純度の高い、クオリティの高い情報を残している。一方で、これはどなたも賛成してくださると思いますが、データもコードもそこまで質が高くないものもある。そうなってくると経済原則が出ざるを得なくなって、「じゃあ、どこまで残すのか」という話を我々は真摯に議論せざるを得なくなってきていると思います。何もかも、データも蓄えるしコードも蓄える、ということでは成り立たないということも、一方で正しく伝えていく必要があると思います。

村山：全くおっしゃるとおりです。私はオープンデータの理念や国際動向の話の中で、データを出しましょう、使いましょうという直接的なメッセージだけをお伝えしました。しかし、国際会議や国際社会の場では、「信頼できるデータ」をどのように共有するか、という話が必ずついて回ります。「信頼できるリポジトリ」を誰が認定するのか。論文のレフェリーと同じようにデータ出版のレフェリーを設定するのか、それはどのようにレビューするのか、といったいろいろな問題がある。そうした「信用できるもの」をなんらかの形でセレクトする動きが今後非常に重要であろうという議論です。科学データの蓄積を以前から始めていたドイツですと、RADAR プロジェクトやリポジトリの動きがありますが、そういった「どのようなデータを蓄えるか」についてご意見をお持ちでしょうか。

レーヴェ：もちろん、今後進めていくにあたって幾つかの選択肢があると思います。いま重要なポイントが挙げられました。コードも重要ですし、ソフトウェアも重要です。科学者の位置付けも考えなければいけません。つまり、科学者としてなんらかの利益を得られるような仕組みがなければなりません。価値のあるデータを創造するためのソフトウェアを開発し、そのソフトウェアを用いて作ったデータを論文の執筆に使えるようにする、データサイテーションに使えるようにする、ということが重要なのです。したがって、科学

者そのものの権限を移譲し、上質なデータやコードを開発してもらわなければなりません。そのためには、こうしたコンテンツを公開することです。データを出して、斬新なやり方でやっていかなければなりません。積み木のような考え方で、オープンジャーナルの構想を立てていくことです。本文や抄録と併せて、データ及びコードがオープンにされて、同僚の科学者の間で議論の対象になるような仕組みが作られなければならないと思います。そして、そのコード、データの質に関する議論がなされなければならない、そういう場を持たなければいけないと思っています。

村山：非常に総合的な取組を考えておられるということですね。そういった意味では、科学の再現性の確保や、それに基づいた次のイノベーションのために必要な情報という意味でも、データだけでは足りない、というメッセージも同時に非常に重要だという、この問題をきちんと捉えれば捉えるほどそういう視点になっていくのだろうと思います。ソフトウェアもそうだし、ソフトウェアコードを走らせるときの OS、例えば MS-DOS で作ったシミュレーションコードを走らせるときには MS-DOS 環境を保存する、といったように何もかもがついて回ることになる。どれを残してどれを残さないかというセレクションプロセスは、考えるだけでも非常に頭が痛いのですが、次のイノベーションと人類のためになんらかの手を打たなければならない、と思うわけです。

こういったことについて、例えば、喜連川先生は NII 所長というお立場で、日本という国のレベルでどうだ、ということを仰るのは難しいかもしれないですが、何か展望やお感じになるようなこと、例えばドイツだとナショナルライブラリ・ネットワークがありますが、今後日本ではこう、ということがありますでしょうか。

喜連川：私はコードが重要であるということを敢えて申し上げたのですが、これは分野によって随分違うといいますか、サイエンスのいろいろな領域の中で、どんぶり勘定で一般的な器が作れるかというと、そうではないと思います。一つ例を挙げますと、シュルンベルジェ という石油探査をされている会社がありますが、ここは図書館がなくてもデータは一生捨てない、データが命ですという会社です。どういうことかと言うと、何十年前に測ったデータの上にどんどん都市開発が進んできますと、同じような実験で探査をすることすらできなくなってくるわけです。一方で、プログラムはいくらでも改変することができる。データがあるところで、プログラムが進化することによって新しい支援を見出す可能性が出てきます。そういったように、純粋に観測データが非常に貴重だという分野もあると思います。ただ、一方で、そうではない分野も多々ありまして、これらを十把一絡げに議論することはかなり厳しい問題だと思います。先ほどの佐藤先生の、人文社会系の出口調査のような情報はそんな大きなボリュームになりませんので、置いといてもだれも文句を言わない、というと失礼になるかもしれませんが、我々の持っているボリューム感とはだいぶ違います。先ほど SAS の統計解析パッケージの話が出ていましたが、そういうデータを置いておくところと解析をするところが必要になってくる。レーヴェ先生は「コンピ

ュータのセンターからデータのセンターにシフトしていくだろう」という言い方をされたと思うのですが、エリア・バイ・エリアかもしれませんが、そういうところを日本中に作っていくことになるのではないかと思います。この時に、NDL をはじめ図書館の方々から、従来の本を探すのと同じようにデータの在りか、コードの在りかを探す支援をしていただけるのではないかと思っています。ただ、インフラに関しては IT プロパーに頼らざるを得ないところがあるので、そこの部分が協調していくような形が自然かな、と個人的には思っています。

村山：いろいろな面で非常に同意します。データ一つとっても、イギリスで「データ・キュレーション・センター」というものができて、データの維持作業だけの専門家がいて、そのための専門機関がある。例えばですが、社会学のデータは社会学のことがわかった人がお世話しないといけない。地球環境なら地球環境がわかった人でないといけない。昔聞いたところですと、アポロ計画で採った月の地震のデータが、地震の専門じゃないロケットセンターの壁際に押しやられていたということがあって、それを日本人研究者が「捨てるのならもらう」といってもらってきた、という噂を聞いたことがあります。専門家が見ないと宝の持ち腐れになる。コードにとって IT システムはその専門家が必要ですし、あるデータにとってはそのデータの専門家がやはり必要です。そのデータやコードの整理・体系化をしたり、レファレンスサービスをしたりするのは図書館の方がエキスパートであり、持分といいますか、ネットワーク化した総体としてのサービスのイメージがこうあってもいいのかなと思っています。ドイツ国内ですと、そういった専門家、あるいは機関間での役割分担はどうなっているのでしょうか。

レーヴェ：ドイツでは、状況評価をはじめています。少なくとも、その時点ではどのような資金供与を受けているのか、ということを考えなくてはいけません。科学者に対しては政府・EU からの資金援助が多いわけです。研究はだいたい 3 年間のサイクルで終結しなければいけません。いわゆる「オーファン（孤児）データ」というものがあります。リサーチプロジェクトを行ったが、データはそのままでどうしたらよいかわからない、「孤児」の状態になってしまうわけです。そして、どこかに入ってしまっている。そこで非常に重要なのは、データ・キュレーションをする人たちがいるということ、最適なのはそれぞれの専門施設においてそういった専門家がいる、ということです。ドイツ国内にはだいたい 200 の研究機関があります。それらに少なくとも一人ずつデータライブラリアンを置くことができれば大きな改善になると思いますが、今はいません。それぞれの機関でデータライブラリアンの配置にかかるコスト計算もできていない状況です。そしてプロジェクトを申請するときに、どれだけのお金が必要なのか、長期的成果として出てきたデータの保存のためにどのくらいの期間、コストをかけるのかという計画が必要だと思います。

村山：まずはデータでしょうけれども、その保存・維持にそれぞれの国でそれぞれの工夫

をするといったことだと思いますが、それを日本ではどうするか、ということが今後の議論になると思います。図書館が科学情報の取り締まりをしてきたこと、IT の発展とともに科学情報の多様化が進んでいること、どのように今後の将来像に組み込んでいくかという意味では、もちろん図書館が主要な、あるノードを成します。同時に、多様な機能をどのように実現していくかということを、今後ともぜひ皆さんと議談させていただけたら、と感じます。

まとめ・質疑応答（講演者全員登壇）

【まとめ】

川鍋：それでは最後のまとめ、質疑応答に入りたいと思います。最初に木浦先生と佐藤先生に、これまでのお話を聞かれて一言ずつコメントをいただければと思います。

木浦：農林水産省のつくば事務所は、NDL の分館でもありまして、そういう意味でも皆さんと協力してデータセンターみたいなものを作っていけたらといいなと思っています。本日のお話を聞いて、その期待がどんどん高まっているところです。

佐藤：先ほど鼎談で、「どういうデータを保存するのか」という議論があったのですが、我々のところでも課題になっていまして、社会調査でいうと代表制に問題があるとか、あるいはサンプルフレームやサンプリングの仕方についてきちんと書いてないデータをどうするか、といったことがあります。その際、やはりメタデータ、「どういうふうに調査が行われたか」という情報をどういう基準で用意しているかが大事です。利用者がデータセットの質を評価できる情報をつけておくことがすごく大事だと思います。質が低ければ保存しないのか、というのはなかなか難しくて、これはもちろん我々の能力的なキャパシティもあるのですが、どういう情報なのか、調査の偏りがあるのかといった情報が載っていれば、それは有用ではないかと思います。つまり、調査に偏りがあっても、事後的に分析できるツールが出てきたりする可能性はありますので、メタデータは残すようにする、ただし、どのようなものを残すのかというところを我々で作っていかなければいけない。もう一つは、やはり質のいいデータは利用者が多いです。ちょっとどうかな、と思うものは利用者が少ないです。そういう意味では事後的に評価されていくので、あまり利用者の少ないような調査をやるのはだんだん評価されなくなりますので、公開の仕組みを作ることでデータの質が高まっていくという効果もあると思っています。

川鍋：今佐藤先生のお話しにあった、データの質の判断も含めて、図書館の立場から「どういうものを残すか」について、レーヴェ先生からコメントをいただけますか。

レーヴェ：まず、データを生成している科学者自身に聞くのがよいと思います。科学的なプロセスの中で、まだファイナル・データセットになっていないものを収集するのが合理的なこともあるかもしれませんが、他の分野ではそうではないかもしれません。最初にエキスパートに相談した方がよいかもしれません。その後、一歩下がって図書館員の

視点から、一次データをどのようにアーカイブすればよいかを考えるわけです。例えば、DOI システムにおいては、データに対して DOI の識別子を自動で毎分配布していくという仕組みがありますが、レファレンスデータとしてはハンドル・システムにした方がやりやすいということがあります。しかし、そのような場合、データ生成者と対話しなければなりません。もう一つの検討事項は、データを持っている人のクオリティについてどのように考えるか、ということです。データの所有者、データを生成した人の視点から評価すること、そしてクオリティの評価はピアレビューが必要で、またコミュニティの視点から評価してもらうことが重要だと思います。うまくキュレーションをすればリサイクルが可能になってくるからです。データセットによって、用途によっては付加価値を、これまで予見できなかったような価値を追加することができるかもしれません。また、他の利用者のコミュニティでは、非常に重要な、不可欠なデータかもしれません。どのように保存するか、どのくらいの期間保存するか、ということについては、他の分野の人に聞くと全く違う結論が出るかもしれません。

川鍋：次に、村山先生、他の先生への質問やコメントがありましたらお願いします。

村山：佐藤先生のお話にありました「淘汰されていく」というお話は興味深くて、私のスライドにも DOI でデータを引用するという話があったのですが、あのシステムの一つの視点は、論文中で使われたデータの DOI を必ず論文に書くということです。そうすると今までの論文の参照と同じように、データの参照、サイテーションやインデックスが出せる、データがどれくらい使われたかという統計情報が出せます。データを出す方も、たくさん引用されたデータを出した人はよいデータプロデューサー、クリエーターであることとなり、よいデータができたらぜひ DOI をつけて出版したい、というよいサイクルが回り始めることで、使う方、出す方のインセンティブが上がるというのが望ましいと思います。

川鍋：では喜連川先生、ここまでのところでご質問などありますでしょうか。

喜連川：なかなかエンドレスな議論だとは思うのですが、（データを公開して）淘汰されていくということもありますが、「他人にお使いいただける」ようにきちんとデータを整理して発信することは、かなり多大な労力が必要になってくるわけです。ということは、科学者の研究スタイルを啓蒙していかなくてはいけません。論文を書いて終わり、その論文がいくら引用されたからこの先生はこう、といった大学の評価機構や研究者の評価機構を変えていかなくてはなりません。多くの方々が利用に供するデータを作った人を高く評価する仕組みを作っていかないといけないと思います。そうでないと、一般にデータをきちんとキュレートして、そこにメタデータをつけて、というのはかなり膨大なコストを要求するので、そういった全体的な科学者のマインドの変化を起こしていくことも必要だと思います。

佐藤：社会科学分野でいうと、調査を行ってメタデータを作る、第三者が使えるようなコ

ードブックを作るわけですが、例えば、科研費にはそれを作る費用は入っていません。研究者がボランティアでやっているわけです。データを寄託すればデータアーカイブ側がそういう作業をすることが必要です。こういうことに研究者が研究時間を取られるようなことになっていると、なかなかうまく回っていかないので、そういった点が大事と思います。それから、ささやかですが、我々も好循環を作るため、寄託されたデータで利用者が多い、論文が多く書かれているものについては寄託者を表彰するという取組を行っています。「おたくのデータを使ってこんなによい研究ができました」という。そういった好循環をどう作るかというのは、お金の面も大事ですし、村山先生が仰ったように研究者として評価されることも大事だと思います。

喜連川：私も発表の中で、最後に結局誰がサポートするか、という話をしましたが、アメリカでうまくいっているのは米国国立衛生研究所（NIH：National Institutes of Health）しかない。NIH は米国科学財団（NSF：National Science Foundation）の 10 倍以上の潤沢な予算があるので、ファンドをとって何かを公表する場合には、全てのデータをここに入れなさい、入れない限りは発表すらさせません、ということになっています。そのデータの面倒は、NIH が見ているわけです。NSF からお金をもらうのは、科研費をとるのと変わらないが、NIH から予算をもらうというのは、それとは違っています。NIH は米国国防総省高等研究計画局（DARPA：Defense Advanced Research Projects Agency）以上の予算規模でやっているのですが、アメリカでも基礎研究のようなものに対してきちんとデータを残そうという動きが出てきており、その部分がリサーチファンドの中に入っているのは、実状はそれなりにつらいことかもしれません。

川鍋：他に先生方で質問はございますか。

レーヴェ：高等教育について教えてください。これまでの話では、現在膨大なデータに直面していますが、それに対して折り合いをつけなければいけないということでした。そこには技術なソリューション、科学的、政治的なソリューションもあるかもしれません。また、次世代、将来の科学者になる生徒に対しても支援することが必要かもしれません。今日のサイエンスにおいては、メタデータのパターンを考えていくといったことは、まだまだ一般的には行われていません。しかし、ヨーロッパの視点からいうと、次世代あるいはそれ以降の科学者ではそのようなことを当たり前に考えるようになればよいと思います。現在、我々はそのようなことを当然のこととして考えていないわけですが、なんとかしてデータサイエンティストとしての行動パターンを他の分野にも統合していこう、という取組があるのですが、日本ではどうでしょうか？

川鍋：データサイエンティストと科学者への教育について、日本での事例をということですが、喜連川先生、いかがでしょうか。

喜連川：これは高等教育局の所掌で、我々は研究振興局というところなのであまりうかつな発言はできないのですが、教育というのはなかなか動かすのが大変で、新しい専攻や学科を作ることは、憲法を変えるくらい大変なことです。例えば、センター入試でも、

どの科目の試験をするかというのが大変な議論になっています。そういう意味で、日本の次を支える人材に対して「基礎素養は何か」という考え方を軽率に変えるべきではないという意見もあります。一方で、「これだけ IT メディアが発展しているのに、どうして学校の先生たちはもっと iPad を使わないのか」という話もある。シンガポール等では教育でもっと IT メディアを利用しています。私が感じるのは、「全ての大学がこういったデータに対する素養を持つべきである」ということを上から命令することは難しいですが、それぞれの大学の中で重要だと考える点について、先導的にカリキュラムを作っていくようなことが起これればそれは素晴らしいということだと思います。ただ大学のカリキュラムというのは基礎を教えるところですので、むしろ大学院の話なのかもしれないですが。

レーヴェ：データサイエンティストのカリキュラムは学部生向けというよりは、院生向けだと思いましたが、やはりドイツにおいても専攻を簡単に変えるということはなく、教科書通りに行くということ、簡単にルールを変えられないのが実情です。ただ、小刻みに進化させていくことは非常に重要だと思います。私はドイツを代表して、また地球科学の領域を代表してお話することしかできませんが、努力を重ねてトップダウンを取り入れていこうとしてはいますが、非常に大変です。ただ、既にムーヴメントとして、ボトムアップの動きが出てきています。今は地理学や科学など自分の領域だけの取組ですが、そういったコミュニティの間で対話を起こすことによって、それが大きな取組につながっていくような活動をハノーヴァーで行っています。こういった動きははじまったばかりで、まだ歩みは非常に遅いと思います。

【質疑応答】

川鍋：では、会場からの質問を紹介します。

　まず「木浦先生と佐藤先生の事例報告についてもう少し詳しく」ということですが、木浦先生、スライドにあった「表現型のデータ」について教えていただけますか。

木浦：「表現型のデータ」というのは、例えば「私の身長が何センチあります」とか、稲でいうと「どのくらいの穂の数が出ていてどのくらいの籾がつきますよ」といった、環境とのインタラクションの結果として出てきている形がどうなっているのか、というものです。

川鍋：佐藤先生の報告事例についての質問です。SSJDA に登録されたデータセットを再利用する場合、その利用条件は明示されているのか、という質問です。

佐藤：基本的に研究目的で、営利利用は不可ということにしています。「研究目的」とは何かというとなかなか難しいのですが、一つは、研究した結果を皆で共有するということです。ですから自分だけで調べて「面白かったな」ではなく、データを利用して学会で報告する、論文にする、研究成果をアカデミックコミュニティで共有する、ということが「研究目的」だと思います。もう一つは、マイクロデータを使うには一定の訓練が必要なので、原則として大学院生以上、研究機関に属していること、という形式審査をし

ています。もちろんそれ以外で使えないわけではないのですが、一応、統計研究での訓練を受けているということを条件にしています。さらに、マスキングはもちろんしていますが、個人を特定する分析はしないことを誓約させています。また、発表するときはデータ寄託者の名前を、「誰々がやった調査である」とリファーすることにしています。

川鍋：村山先生後半の、RDA のお話についてもう少し詳しく、という質問をもらっています。「研究データ同盟（RDA）の動きに対して、なじむ分野となじまない分野があるのではないか」、「分野によって温度差があるのではないか」、例えば「地球科学などはなじむだろうけど、他の分野はなじまないのではないか」、「地球科学の他にもなじむ分野があるのではないか」といった分野ごとの差や、ネガティブな分野がある場合、それに対して「共有してください」といったような動きは RDA の方であるのでしょうか、という質問です。

村山：RDA は、私の理解ですと、そもそも G8 の下で「Global Research Infrastructure」という議論があって、これは言葉のとおり「地球規模での研究基盤」をどう共有するかということです。私はそのワーキンググループには参加していませんので詳細は知らないのですが、ハッブル望遠鏡やカミオカンデなどの大規模な研究施設がサイエンスにとって不可欠な時代で、国際社会としては、各国がそれに個別に投資していられないので、各国で共有しませんかというニュアンスのものと聞いています。

そのアクティビティの下に、「Data Infrastructure Working Group」というグループができました。これは CERN で出たデータをデータ基盤として国際社会がどのように使えるようにするか、といったものと思っていたのですが、そこを契機として、RDA のような連盟が設立されたように聞いています。その背後には「RDA のようなものをほしい」と言って動いていたアメリカやオーストラリアの関係者があったようなのです。そういった研究施設の延長線上の研究データという別の形のリソース、リサーチ・リソースをどう扱っていくか、という視点で、フィットするものとしないものがあるかといえば、おそらくあるだろうと思います。先ほどの議論にもありました通り、出しづらいデータですとか、「それを外に出すとうちの価値がなくなる」という種類のものを出すということはないと思います。例えば、国連の下で海洋データを国際的に共有する機構が動いています。これは海洋データを各国が共有して、海のことですから地球規模でないといけないので「各国はみな参加してください」ということになります。ただ海のことは経済や国土など議論がいろいろありますので、出したくない国は当然あります。そこで出せるものは出しましょう、ということになります。天文学は「Virtual Observatory」によって共有が進んでいます。私が知らなかった分野ですと、「toxicogenomics」、遺伝毒性学とでも訳すのでしょうか、とかがあります。毒性学の分野では、毒性検査をする EU や北米が共有してデータベースとして持ちたい、という話を聞きました。EU で一度調査したものをアメリカで投資したくない、だから共有データベースにするということです。業界がその目的に沿って有益だというところは、こういう感じでどんどんその活動を高

めていくようです。それこそ、喜連川先生がおっしゃったような、コミュニティやその分野でよいと思ったところは伸びていく恰好だと思います。ネガティブな分野まではさすがに見えてこないのですが、今のところ、地球科学は比較的フィットする分野に思える、医学・遺伝関係は進行中というのが当面のお答えかなと思います。

川鍋：レーヴェ先生に RADAR プロジェクトについての質問です。ご報告の中で、「75%のデータが埋もれている」というお話がありましたが、その 75%のデータをリポジトリに登録する場合、その登録コストは誰が負担するのでしょうか、という質問です。

レーヴェ：大変、重要な指摘だと思います。75%というのは驚くべき数字で、研究データの大部分が失われているということを想定するならば、いま継続されている研究は、また新たに幾度も繰り返さなければならないということです。一度失われたということは、二度、三度と失われる可能性があるわけです。RADAR プロジェクトのアプローチとはそういった状況を回避することです。全ての分野において改善することは範囲が大きすぎて難しいかもしれませんが。RADAR を採用すれば、例えば、バイオ・ジェネティクスという領域、あるいは生物学の領域に特定するのであれば、その特定のリポジトリを確立することができます。もちろんその資金を誰かが捻出しなければなりませんが、RADAR プロジェクトは二つの方法を用いてこのリポジトリの維持費用を賄っています。一つは研究プロジェクトに費用を充てることです。まず一括払いでリポジトリの確立資金を支払うという仕組みです。このようなファイナンスは、リサーチプロジェクトの費用の一環として計上できるというメリットがあります。つまり、前もって資金提供機関から予算を編成してもらうことができます。もう一つは、どこの団体が融資してもよいという考え方です。この場合、長い期間研究が続けられれば、年々費用が減少していくというメリットがあります。とはいえ、リポジトリにおいてデータを保全し、管理していくための費用を持たなければならないというのは事実であります。

川鍋：最後に、レーヴェ先生、今回参加されてのご感想をいただけますか。

レーヴェ：楽しかったです。参加できてうれしく思っています。本当にたくさんのことを学ぶことができました。ヨーロッパでも本当に苦しんでいる課題が、日本でもこのように活発に議論されていることを知って安心しました。このようなディスカッションはお互いにとって有益ですし、先ほどもお話ししたとおり、75%のデータが 1 回だけではなく 2 回、3 回と失われるという状況にあります。そういった経験からお互いに教訓を学ぶことができれば、持続可能性のあるデータ基盤の構築を、一国だけで行うよりも早く達成することができると思います。

川鍋：以上で終了したいと思います。先生方、どうもありがとうございました。

（会場拍手）

以上