AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ（第5回）　議事録

https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu29/006/gijiroku/mext_00005.html

記事ページ本文

現在位置
トップ
>
政策・審議会
>
審議会情報
>
科学技術・学術審議会
>
情報委員会
>
AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ
> AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ（第5回）議事録
AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ（第5回）議事録
1．日時
令和8年5月22日（金曜日）16時00分～18時00分
2．場所
文部科学省東館17階局4会議室 及び オンラインのハイブリッド形式
3．議題
AI for Scienceを支える研究データの管理・利活用及び流通の在り方について
AI for Scienceを支える研究データの管理・利活用及び流通の在り方ワーキンググループとりまとめ素案について
その他
4．出席者
委員
尾上主査、石田委員、江村委員、工藤委員、千葉委員、林委員、宮田委員、吉田委員、若目田委員
文部科学省
山本学術基盤整備室長、麻沼参事官補佐、池田参事官補佐、鈴木科学官、込山学術調査官、國本学術調査官
オブザーバー
国立情報学研究所
副所長/アーキテクチャ科学研究系 教授 合田 憲人
アーキテクチャ科学研究系 教授 佐藤 周行
アーキテクチャ科学研究系 教授 栗本 崇
オープンサイエンス基盤研究センター長 谷藤 幹子
物質・材料研究機構
技術開発・共用部門長 出村 雅彦
理化学研究所
放射光科学研究センターデータ処理系開発チームリーダー 初井 宇記
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 教授 五斗 進
5．議事録

【尾上主査】それでは、定刻になりましたので、科学技術・学術審議会情報委員会AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループの第5回会合を開催いたします。
委員の皆様におかれましては、お忙しいところお集まりいただきまして、ありがとうございます。
本日は、現地出席とオンライン出席のハイブリッドでの開催としております。また、通信状態に不具合が生じるなど続行できなかった場合、委員会を中断する可能性がありますので、あらかじめ御了承ください。
まず、事務局より本日の出欠状況などについて御案内願います。
【麻沼参事官補佐】事務局でございます。本日の出席者につきましては、矢守委員が御欠席ですので、9名の先生方に御出席をいただいております。
また、本ワーキングですが、国立情報学研究所からもオブザーバーで参加をしていただいておりますが、本日は、合田副所長、佐藤先生が現地から、栗本先生、谷藤先生、武田先生がオンラインから出席をいただいておりますので、よろしくお願いいたします。
続きまして、陪席いたします科学官・学術調査官に関しまして、鈴木科学官がオンラインから、國本学術調査官と込山学術調査官は現地から御出席をいただいております。
また本日はゲストスピーカーとして3名の先生方に御参加をいただいております。国立遺伝学研究所から五斗先生、物質・材料研究機構から出村先生、理化学研究所から初井先生に御出席いただいておりまして、五斗先生は現地から参加をしていただいております。また、初井先生は17時頃からオンラインにてご参加でございます。
以上でございます。
【尾上主査】ありがとうございました。次に、配付資料の確認とハイブリッド開催に当たっての注意事項について事務局より御説明お願いいたします。
【麻沼参事官補佐】事務局でございます。本日の配付資料は議事次第に記載のとおりでございまして、資料1から6、参考資料1と2、また非公開の資料として机上配付資料を御出席の委員の先生方のみ、本ワーキングの取りまとめ素案として御用意をしております。
もし資料の欠落等不備がございましたら、議事の途中でも構いませんので、事務局までお知らせをお願いいたします。
続きまして、ハイブリッド開催に当たって注意事項を申し上げます。御発言時を除きまして、マイクは常にミュートとしていただきますようお願いいたします。
ビデオのほうは常時オンにしていただき、通信状況が悪化した場合にはビデオを停止していただくようお願いいたします。
また運営の都合上、現地出席の先生方も含めまして、御発言いただく際は挙手ボタンを押して御連絡をお願いいたします。
尾上主査におかれましては、参加者一覧を常に開いていただきまして、手のアイコンが表示されている委員の御指名をお願いいたします。
また、議事録作成のため、速記の方に入っていただいております。御発言される際は、お名前をおっしゃってから御発言をお願いいたします。
恐れ入りますが、マイクの数が限られておりますので、現地出席の先生方が発言される場合には大きめの声で御発言をお願いいたします。
本日傍聴希望をいただいた方には、YouTube配信により御参加いただいております。
トラブルが発生した場合には、現地出席の先生方は手を挙げていただき、オンライン出席の先生方は電話にて事務局まで御連絡をお願いいたします。
事務局からの御案内は以上でございます。
【尾上主査】ありがとうございました。本日は、1、AI for Scienceを支える研究データの管理・利活用及び流通の在り方について、2、AI for Scienceを支える研究データの管理・利活用及び流通の在り方ワーキンググループ取りまとめ素案について、その他の3件の議題を予定しております。
前回、国立情報学研究所様からは、黒橋所長にお越しいただき、研究データ基盤とAIの融合として、知識基盤の構築に係る構想を伺い、議論を深めました。
本日は、6月の取りまとめに向けての報告書についての審議と、研究データの利活用を進める上で分野ではどのように進めてきているのか、この点がミッシングリンクとなっておりました。中核的な研究データ基盤と位置づけるNIIの基盤とはどのように接続しているのかなど、主要分野の状況を3名の先生方から御紹介いただきます。その上、NIIにおいても、この1か月間で、NII中心の基盤からより俯瞰した基盤として確立する上で、様々な分野との対話を重ね、構想実現に向けて検討を進めていた

だきました。この状況についても御報告いただきます。
それでは、まず、文部科学省から資料1に基づき、取りまとめに向けた整理として、前回御説明いただいたものに前回の議論要素を追加いただきましたので、この点を御説明いただいた後、3名の先生方にお取組を御紹介いただきたいと思います。その後、NIIから資料説明をお願いしたいと思います。
それでは、麻沼補佐、よろしくお願いいたします。
【麻沼参事官補佐】それでは、資料1の2ページ目を御覧ください。こちら、「AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループのとりまとめに向けて」という資料でございますが、前回もお示ししたものでございます。本日のワーキングは全体スケジュールの真ん中の黄色い部分でございまして、主要分野の取組状況等について御議論をいただきます。また、右側の青い四角のところ、「セキュリティ強化」と赤字で書かせていただいておりますけれども、ここは論点が抜けておりましたので、追記をさせていただいたものでございます。
続きまして、3ページ目を御覧ください。取りまとめの骨子となっておりますが、前回は、右側の枠内のとおり、整理すべき内容をまとめておりました。そこから、左側のように骨子案とした形で整理をしております。こちらにつきましては、本日の最後に再度御説明いたしますので、ここでは詳細、割愛をさせていただきます。
4ページ目以降ですけれども、これまでのワーキンググループでの主なポイント等をまとめた資料になっておりまして、10ページ目と11ページ目に第4回目の分を追加してございます。
10ページ目はNIIからの御説明資料の構想についてのまとめになっておりまして、知識基盤は学術研究プラットフォームの最上位に位置し、AIにより研究データを知識へ変換する中核機能として、研究データ基盤の上にAI機能を統合し、データから知識を創出する基盤であることを御説明いただいております。
また、一番最後のポツですけれども、AI for Scienceにおける計算資源の需要は、研究データの整備・解析・学習等を担うバッチ処理と、研究者の日常的な思考支援等に用いる対話型処理に大別される。これらの利用を前提とした場合、学術分野全体において大規模な計算資源が必要とされることが示されており、研究者がAIを活用した研究を円滑に実施するためには、当該資源を共通基盤として整備することが必要といったことを御説明いただいたところでございます。
また、11ページ目が、委員の先生方からの主な御意見をまとめてございます。効果としまして、提示された機能が実現すれば、研究活動において有用な支援が提供される可能性が高く、研究者の利活用が進むことが期待されることや、研究環境の改善に寄与することにも期待が持たれております。
また、AIの導入によって、従来の研究分野の枠を超えた新たな研究機会の創出につながる可能性があるといった期待の御意見とともに、従来型の研究プロセスが前提になっているという点や、各基盤の一体的な連携が重要で、研究全体を支える統合的な基盤として再設計する必要性があるのではないかといった課題も示されたところでございます。
12ページ以降が政策文書における関連記載をまとめた資料になっておりまして、こちら前回から変更はございません。
その後、飛んでいただいて、参考資料が続くのですけれども、追加させていただいた資料は最後になります。「AI for Scienceを支える次世代研究インフラの構築」とした資料になっておりまして、こちら、AI for Science推進委員会第4回目が4月23日に開催されましたが、AI for Scienceの戦略方針の具体化に向けた方策の案の1つとして頭出しされた資料となっております。
本ワーキングにおいてもデータ基盤と計算基盤の整備は一体的に考えたほうがよいことや、実験基盤も含めまして各基盤の一体的な連携が必要といった御意見もいただいておりますので、本ワーキングでの御検討も踏まえまして、このような方策の検討が進んでいる状況でございます。こちら紹介としてお示しをしているものでございます。
事務局からは以上でございます。
【尾上主査】ありがとうございました。ただいまの御説明に対して御質問などもあるかと思いますが、後ほどの議題でまとめて御議論できればと思います。
それでは、先生方からのプレゼンテーションをお願いいたします。まずはライフサイエンス分野の取組として、情報・システム研究機構国立遺伝学研究所、DBCLSセンターより五斗先生から御説明いただきます。15分程度御説明いただいた後、10分程度の質疑応答の時間を設けられればと思います。
それでは、どうぞよろしくお願いいたします。
【五斗先生】よろしくお願いいたします。情報・システム機構国立遺伝学研究所で、ライフサイエ

ンス分野のデータベースの統合関係の研究開発をしています五斗と申します。よろしくお願いいたします。
本日はこういう機会をくださいまして、ありがとうございます。
本日は、我々が取り組んでいるライフサイエンス分野のデータ基盤を中心に紹介させていただきたいと思います。
まず、ちょっと簡単に、本当に初歩的なことになって恐縮ですが、ライフサイエンス研究データとDBの特徴ということで、研究データのほうの特徴としては、大きく、データがすごく多様性が高いということ、それから、データの規模が大きいということが挙げられると思います。
データの多様性としてはいろいろなところがありますが、大きく、研究分野がすごく多様であること。生物学、医学、薬学、農学、その他生物に関することであれば何でも、あと、環境の情報なんかも入ってきます。
それから、研究対象、関連して研究対象も、分子レベルのゲノムとか、遺伝子とか、タンパク、化合物、それから生物の情報、それから環境であったり、疾患、病気の情報、様々な対象があるということが多様性の1つです。
それから、それらの研究対象を研究するために様々なデータを実験によって出してくるわけですが、そのための計測装置も研究対象の多様性と相まって多様になっていると。例えば塩基配列であればDNAシーケンサー、それから、タンパク質とか化合物の同定を行うための質量分析装置、それらの基になるような画像データを解析するような装置。そういう情報がたくさん出てきているという状況です。
それから、大規模、この規模に関しては、ほかの分野と比べて大きいか小さいかというのはいろいろ比較の仕方がありますが、塩基配列に関しましては、現在、欧米も含めて大体100ペタバイトクラスのストレージを確保して、データを置いている。
また、ほかのプロテオームデータベースなんかに関しても、1回の実験で相当大きなデータが出てくると。
あと、文献に関しても、ライフサイエンスの分野は、PubMedというデータベースが既にスタンダードで提供されていまして、4,000万件以上の論文が、アブストラクト、少なくともアブストラクトは見れるようになっていますし、オープンアクセスのものはフルテキストで見れるようになっているという状況になっております。
こういうデータを研究者が扱えるような形にするために、大小様々なデータベースが既に存在しているという状況になっていまして、それが右側の箱に書いてあるものです。現在言われているのは、世界全体では7,000以上、1万ぐらいのデータベースがあると言われていて、いろんなデータベースのカタログも作られています。
それから、それらのデータベースの中の情報を記述するためのオントロジーも、分野が多岐にわたりますので、1,000個以上オントロジーが開発されていまして、右に書かれていますように、いろんなサービスでそのオントロジーが使えるようになっていると。
実験で出てくるデータが、ちょっとこれ図が小さくて申し訳ないんですけど、一番大きな部分で100ペタバイトクラスのデータがリポジトリに登録していると言われていて、そこからのデータを整理して、マニュアルのキュレーションも含めて知識として整備して、いろんなデータベースが開発されてきているという状況になっています。
そういういろんなデータベースが世界的に分散されて、いろんなフォーマット、オントロジー、インターフェースが用いられたデータベースが開発されてきているので、それらを連携・統合して使いやすくするためのプロジェクトが20年ぐらい前に統合データベースプロジェクトとしてライフの分野のプロジェクトとして立ち上がっています。
その後、DBCLSという組織が立ち上がって、JSTにナショナルバイオサイエンスデータベースセンターというのができたんですが、それは協力してプロジェクトを進めてきていました。去年から、これらの一部が文科省のナショナル・ライフサイエンス・データベース・プロジェクト、NLDPとして進められてきています。
その手段とかターゲット層はそこに書いてあるとおりなんですが、オープンサイエンスを推進するというところは当然あるんですが、その方法として、いろんな統合のための技術開発をしていきましょうと。
それから、文科省だけじゃなくて、いろんな省庁も含めて枠組みをつくっていきましょうということを進めてきています。
ターゲット層としては、ライフサイエンス研究者がメインですが、データサイエンス、バイオイン

フォマティシャンを中心に使ってもらうという形で進めてきています。
その中で、ほかの分野でも最近使われていますが、FAIR原則というのをベースとしたデータの利活用促進というのを進めてきています。これはもともとのバイオの分野で提唱されてきたものだと理解しているんですが、Findable、Accessible、Interoperable、Reusableということで、Findableに関しては、データベースのカタログを構築してきていますし、Accessibleに関しては、横断検索とかデータベースのアーカイブをつくってきています。これは世界各地のデータベースにも、リポジトリとか、まとめたようなものが今できている状況です。
それからInteroperable、Reusableに関しては、機械可読な形にして、機械が理解できるような形にもしてデータを統合していくということで、これはDBCLSが基盤技術開発を中心にやってきたところになります。
そのプロジェクトの成果の一部をここに載せていますが、詳細はあまりここでは述べませんけれども、データベースをまとめて、データセットを整備して、そこからそれを使いやすくするためのアプリケーションを開発したり、データベースを整備したり、アプリケーションを開発するためのツールをゼロからつくっていくということをずっとやってきていました。
こういうふうにして整備してきたデータについて、また後でちょっとお話ししますけれども、こういうデータというのは、バイオインフォマティシャンとかデータサイエンティストが使えるような形で整備していくという形にしてきたわけですが、結局、AIでも同じように知識としてまとめられたデータというのは必要不可欠なので、AIのえさというか、基盤になるようなデータ基盤として提供できるようにもなってきています。それぞれの活動についてはここでは省略いたします。
その中でできたものについて少し書いたのがこの資料になりまして、先ほど言ったFindableのカタログについては省庁連携で行っていまして、現在、このカタログに関しては2,500ぐらいのデータベースが登録されています。そこでは、ヨーロッパのグループとデータのやり取りもしていたりして、そこで合わせて5,000ぐらいのデータベースがカバーされていると。そのうち横断検索できるようにしているのが819データベースあって、さらにアーカイブとして研究データを登録して置いておくということもできるようにしていまして、そこにアーカイブ化されたデータも幾つかあります。それ以外に、データを集めるときの、集めたり利用するための仕組みとして、例えばヒトのゲノムの情報なんかだと、個人情報なんかが関係してきますので、倫理審査を通った人だけがちゃんと使えるような仕組みとか、そういう仕組みも導入して、ヒトデータベースというものを構築してきています。
最後のInteroperableとかReusableといった意味では、形式として、知識グラフの形で知識基盤を構築して使えるようにしていきましょうということで、そうやって集めてきたものをRDFポータルという形で提供しています。
現在、こういう形で、幾つかの、最初にお話しした多様なデータセットがあるというか、多様な種類のデータがあるということを言いましたけれども、現在、分子レベルの情報が今はメインですけども、塩基配列から、疾患の情報とか、それ以外の様々な生物学の情報が知識グラフの形で統合された形で今使えるようになっています。これは現在もどんどん増えていまして、黄色で書かれた部分が昨年度増えた分で、現在合わせて、知識グラフというグラフの形にするのでトリプルで表しますが、2,200億トリプルが皆さんに使えるような形で提供されています。
こういうことをやっていくと、例えばヒトのゲノムの日本人のバリアントの情報というのは、日本人の様々なコホート研究からゲノムの情報が集められてきて、バイオバンクなんかに登録されたデータからゲノムが得られてきますが、ゲノムの中から個人情報として公開できるような形で、頻度情報の形で、ゲノムのどの位置にどういう変異がどの程度あるかというのを集めたようなデータベースをつくっていまして、そういうのをつくっていくと、先ほどのRDFポータルで提供している、それ以外の疾患の情報であったり、遺伝子の情報であったり、タンパク質の機能の情報なんかと簡単に結びつけることができるので、バリアントからそういう別の情報が見れるようにするみたいなことができるようになっています。
様々なデータベースを使っていますので、データベースによっては同じ遺伝子でも違うIDで管理しているようなデータベースもありますので、データベース間のIDの対応サービスというのも提供しています。
そういうことを提供していくことによって、一番上に書かれているような、例えば、患者さんの症状と疾患、病気とそれから遺伝子を結びつけるような仕組みをつくることができますので、例えばお医者さんが診断するときに、症状を見て、それがどういう疾患と関係しているかといったことを検索できるような仕組みもつくることができるので、そういうのをつくって提供しています。
こういうのを、我々のところでは様々なデータベースを構築して、使って解析しているんですが、日

本ではそれぞれがいろんな組織でつくられているデータベースです。
一方、海外では、例えばアメリカのNCBIとかヨーロッパのEBIなんかは、大きな組織として、組織の大きさは大分違うんですが、自分たちのところでちゃんとデータを集めて、関連するデータベースを自分たちの中で統合して使えるようにしていると。
一方、日本は、割とその辺がまだ小さいので、我々が提供している統合的な技術だけを使って今あるデータベースを統合的に扱えるようにしているということをやっています。
ただ、その中でも、国際連携でやっている部分というのも結構あって、例えばリポジトリ、例えば塩基配列なんかの生データを登録するようなリポジトリというのは、日・米・欧の三極でそれぞれリポジトリを運営していて、ただ、日本人の研究者は日本のデータベースに登録するんですが、登録されたデータは中で全部共有していて、三極がそれぞれ全部同じデータを常に持っているというような仕組みを構築して、国際連携でデータをちゃんと集めるようにしているという形になっています。
ここまでが現状、我々だけじゃないですけど、ライフのデータベースとしてどういうことが今されているかというのをすごく簡単に説明させていただいたことになるんですが、NIIの知識基盤とどういうふうに関係しているかということについて簡単にまとめたのが最後のスライドです。
当然、共通の基本理念というのがあって、特に大きなところではオープンサイエンスとか、データを集めて研究透明性に結びつけるとか、データを利活用を促進させるという、目的は多分全く同じだと思います。
ただ、分野特異的な背景というのが結構あって、バイオの分野では、一番最初にお話ししたようなデータの多様性というのがすごく高いと。それから、規模も割と大きいと。データの多様性の中でそれぞれのデータに対して専門家によってアノテーションとかキュレーションとかして、それをするためのオントロジーというのも整備されてきているので、専門家による知識というのがいまだに結構重要な要素になってきている。
国際連携に関しては、先ほどお話ししたような生データのリポジトリの運用というのは、塩基配列だけじゃなくて、いろんな分野でも進んでいます。そういうのも集めて、知識を抽出してきて、知識として提供するということをやっているわけです。
大規模データの解析というのも必要になってくるので、それが入っているストレージと解析基盤を一体的に提供するということも、これはまだあまり十分にできているところが多くはないですが、そういうのも必要になってきています。
最後に、そういうものがあった上で、NIIの知識基盤への期待ということを書かせていただいていますが、データベース運用の基盤となるような標準ツールを、分野に限らないようなところで部品を提供していただけると、我々としてはすごく助かるかなと考えています。
例えば、「共通認証機構の提供」と書いていますが、現在、GakuNinで提供していただいているものもありますが、例えばバイオの分野だと、海外と共同研究したりすることもあるので、海外の研究者も使えるようにするとか、それから、倫理審査が必要なデータを扱うときの研究者の要件確認なんかができるような仕組みがあるといいかなと考えています。
それから、CiNiiなんかで連携検索の基盤を提供されていると思いますが、メタデータの共通化に関しては分野間でいろいろ協力してやる必要がありますが、それ以前のポリシー的な問題になるんですが、データがまだ十分にオープンになってこないというようなものが、研究者の手元にとどまっちゃうというものがまだ結構あったりするので、その辺のデータシェアリングのポリシーを国の国策としてどういうふうにしていくかというのをトップダウンで決めるようなところが、そういう司令塔になっていただけるといいのかなという気がしています。
それから、「共通大規模ストレージの提供」と書かせていただきましたが、これは分野ごとにいろんな必要性がありますので、それも反映した環境が提供されているといいのかなと考えています。
以上、簡単ですが、私のほうからの説明とさせていただきます。
【尾上主査】五斗先生ありがとうございました。それでは、ただいまの御説明に対しまして、何か御質問等ございましたら挙手にてお知らせいただければと思います。
いかがでしょうか。
石田先生、どうぞ。
【石田委員】九州大学の石田と申します。御説明いただきまして、ありがとうございました。
もう少し最初のデータの登録というところについて少しお聞きしたいんですけれども、これ多分研究者がそれぞれつくったデータを登録するという形になっていると思うんですけれども、それに関しては、分野の中での慣習として、義務化ではなくても、皆さん積極的に登録するという流れになっているのか、それとも、例えば、根拠データのように査読で必要なので登録するというような形になって

いるのか、その辺の積極性といいますか、そういうのを一つお聞きしたいのと、それからもう一つは、データの登録に際しての、システム的にはきちんとなっていると思うんですけど、人的支援みたいなものが必要なのかどうかというところを聞かせていただければと思います。
【五斗先生】最初の積極的にデータが登録されているかどうかということですけれども、これ分野とか、もちろん研究者によっても大分違うんですけれども、仕組みとして今提供できているのは、例えば塩基配列とか、タンパク質の立体構造という形を決めたら、データとして登録するデータベースがあったりするんですが、そういう研究者が入れるリポジトリというものが幾つかのタイプのデータでありまして、その幾つかのもの、特に塩基配列に関しては特に強力に決まっているんですが、論文を発表するときに、ジャーナルの出版社のほうが、まずデータをリポジトリに登録してくださいと要請します。登録された証拠としてデータのIDを出版社に渡します。それがあって初めて論文として出版できるという仕組みがもうあるので、協力してくれる出版社が、塩基配列なんかも結構多いので、そういうのが多いところは自然と出てくるというところはあります。ただ、論文として発表されないものに関しては、そういう制約がないので、出てくるものもあれば出てこないものもあるというような、それは研究者によるところになります。
塩基配列以外では、タンパク質の立体構造もそうですし、タンパク質の発現情報、プロテオームというデータがあるんですけど、そういうものに関してもそうなってますので、そういうのが幾つかあります。
登録するときのサポートというのは、研究者に対するサポートということですね。
【石田委員】はい。
【五斗先生】それは今、受け付ける側で、キュレーターとかデータベース側の人たちが、こういうデータが来たときに、例えば我々、ヒトデータベースというのを扱っているんですが、それもちょっと厳しくて、例えば倫理審査の書類が必要になるわけなんですが、倫理審査の書類がちゃんとそろっているかどうかというのは受付側でちゃんとチェックして、こういうのが足りないからこういうのを入れてくださいというようなことはやっています。
そこでどういうものが必要になるかというのは、やっぱりなかなか大変なので、リストアップとかはしていますけれども、それは随時対応しているような感じになっています。
あと、もう少し自動的にできるところでいうと、例えば生物種の名前とか、いろんな、どういうタイプのデータがあるかというのは、オントロジーがある程度準備されているので、オントロジーの中のタームを選べるような、そういう仕組みというのは割と機械的なインターフェースとして用意することができるので、そういうものを準備しているリポジトリというのはあります。そういう状況でお答えになっていますでしょうか。
【石田委員】分かりました。ありがとうございました。
【尾上主査】それでは、江村委員、お願いいたします。
【江村委員】ありがとうございます。資料の中でちょっと気になったところが、ストレージの枯渇というところです。一方で、お話伺っていると、データベースというのは個別につくられていっているものを統合しようというような議論でなっていると感じます。要は、これからますますデータ量が増えていったときに、仕組みのつくり方というんですかね、ばらばらにデータベースがあるものを統合していくというアプローチを今後も続けるのか、ストレージ枯渇みたいな問題も含めたときに別のやり方があるのかという辺りがどうなのかなと思って聞いていたのですけど、いかがでしょうか。
【五斗先生】ストレージの枯渇って多分幾つかのパターンがあって、例えばリポジトリごとに今はストレージを持って提供しているんですね。例えばDDBJなんかは割と大きなストレージを持っているんですけど、それでもやっぱり足りないという状況にはなっています。ほかのところもリポジトリとして持っているんですけど、足りない。そういうことをそれぞれのサイトでやっているので、それぞれのサイトが毎年ディスクを請求して、購入して、管理していかないといけないということがあるので、今、NLDPのプロジェクトの枠組みの中でどういうふうにしようと考えているかというと、やはり1つの大きなセンター、ナショナルセンターをつくって、そこに大きなディスクを準備して、そこでデータベースも開発できるようにしていきましょうと。
特に小さいデータベースなんかに関しては、本当にそれぞれが、皆さんアイデアを持って、データをきれいにするようなアイデアを持っているんですけど、そこをセンターのディスク、計算機資源はセンターのものを使ってもらって、自分たちのアイデアはそこで実現して、そのときにほかのデータベースともちゃんとつながるように我々のほうでサポートします。そういう仕組みづくりをちゃんとしていこうかなと今計画というか、話しているところです。
なので、ディスク、それぞれの研究者がディスクを準備するということは、これ多分NIIの知識基盤

もそういう考え方なのかもしれませんけど、そういう個人の研究者がディスクというか、計算機資源を準備するのではなくて、中央的に大きな計算環境を提供して、そこでつくってもらうというような形にすると、そこで最初から統合されたデータベースにもなるというようなことを考えています。
【江村委員】 分かりました。それでNIIという話になっていると思います。最終的には本当に必要なキャパシティーを持ち切れていっているのかという議論もしないといけない感じですよね。
【五斗先生】 そうですね。そのとおりで、最終的に必要なキャパシティーもそうですし、ここでちょっと書かせていただいたのは、分野ごとの事情を反映できるような環境というのは、例えば計算のプログラムとか、データの解析のソフトウエアとかも、分野によってかなり特徴的なものになっていますし、ディスクにどういうふうにアクセスするかとかいうところも多分分野ごとに全然違うので、そこら辺も考慮した大きな仕組みがもしあるといいのかなという気はいたします。
【江村委員】 分かりました。ありがとうございます。
【尾上主査】 ありがとうございます。本日、ちょっとスケジュール、タイトでございますので、宮田委員、若目田委員、千葉委員、全員、御質問をまとめてお受けしたいと思います。
まず、宮田委員、お願いいたします。
【宮田委員】 宮田です。ありがとうございました。さっきの江村先生のにもちょっと関係するんですけど、データの拡充の方向性みたいなのをちょっとお伺いしたくて、今は論文を出すときに証拠として登録するというのが多分主というようなことなのかなという気もするんですけど、知識基盤を形成するって考えると、戦略的にここは今足りてないのでここを拡充していこうかみたいなのが学術としてはあってもいいのかなという気もちょっとして、ただ、最初におっしゃっていたように、分野的にはすごく多様性が非常にあるというような分野ということなので、それを言い出すと、どれもこれも似ているけど違うみたいな感じで、どうやって何を拡充していくべきかみたいなのって、難しいところはもちろんあるとは思うんですけど、その辺りについてみんなで議論するような機会があったりされるんですかとかいう、その辺りをちょっとお伺いしたかったです。よろしくお願いします。
【尾上主査】 若目田委員、お願いいたします。
【若目田委員】 NII基盤との連携のところですけれども、メタデータの共有化とか、あとポリシーですね、これが本当に全体設計ちゃんとしているところはまさに賛同します。ぜひ、その辺、NIIさんと連携をいただいて、特にこれからポリシーの1つとして、安全保障の問題とか、そういう部分あるんじゃないかなと思っていました。FAIR原則として、公開してくれという部分から、むしろ安全保障の観点から、特に関連のデータというのは、データを一切渡さないのか、もしかするとデータクリーンルームのようなものの形で、渡さないんだけども、共有はしていくのか、この辺のポリシーであるとか仕組みというものは、ぜひ具体的に実装をしていただくべきかなと思いました。
ちょっと伺いたいのは、データをどんどん充実させていくという部分はいろんなモチベーションがあると思うんですけども、研究者の方の中、異動とか、退職とか、そういったようなことも含めて、延々とデータを残し続けるのか、廃棄をしていくようなポリシーというのはどうなっているのかみたいなところを教えていただければと思います。
以上です。
【尾上主査】 千葉委員、どうぞ。
【千葉委員】 私からの質問は、データベースというのは具体的にどういうものを指しているのかということを伺いたくて、単なるリポジトリということはないと思うんですが、いろいろあると思うんです。それから、どなたがつくっているかというのに興味があって、研究者が御自身でつくっているものもあるかもしれないんですけれども、研究者がどこかの企業に発注してカスタムでつくっているものなのか、それとも、分野全体に標準的なものがあって、それを使っているのか、あるいは商用のものを使っているのかということを伺いたいです。
以上です。
【尾上主査】 五斗先生、お願いいたします。
【五斗先生】 簡潔に答えられるかどうか分からないですけど、まず何を拡充していくかという決め方は、これ結構難しい問題がありまして、JSTのNBDCというところがずっとデータベースのサポートとかをやっていたんですが、そこは例えば今どういう生物種がカバーされていて、どういう分子レベルの情報がカバーされているのかというのを一応マトリックスみたいなものをつくって、どこが抜けているかというのを調べるといったようなことはやったりしていました。
研究の進展で新しい技術ってどんどん出てきますので、それはそれに対応して、対応していかないといけないなというのはあるので、そういうのをフォローしていくということをやっていっているのが現状かなと思います。

それから、廃棄のポリシーは、それほどちゃんとはしてないです。アーカイブというのがあると言いましたが、例えば科研費なんかでつくられたデータベースというのが、おっしゃったように、研究者が引退すると、そのまま更新もされなくなるし、サポートもされなくなる。ただ、有用なものもあるので、物だけはちゃんと保存しておきましょうということで、一応ダウンロードできる形でアーカイブとしては持っておくというような形でやっています。それを今、廃棄するという形には特にはなってないです。
それから、データベースの定義なんですが。
【千葉委員】定義というか、何を。広いじゃないですか。
【五斗先生】そうなんですね。NIIのデータベースと言っている考え方と我々が言っているデータベースの考え方、多分いろいろ違っていて、例えば研究データのリポジトリって割とファイル置場的な感じになっているものが多いと思うんですけど、ライフの場合も、ファイルを置くところもあるんですけど、さらにそこから、そのデータがどういう意味を持っているかというのをちゃんとアノテーション、メタデータをつけたりというのがありますが、アノテーションして、ほかの人がちゃんと使えるように、もう1回再解析できるようにするといったようなことをやっています。そこは基本的には研究者がやっています。
バックエンドのデータベースに何を使うかというのはデータベースごとによってやっぱり違います。データベースによっては、データベースの設計自体はちゃんと自分たちでやって、こっちの実装は外注するといったようなことをやります。基盤は基本的にRDBを使っているところが多いとは思います。我々のところは知識グラフをやっているので、今、Virtuosoとか、そういう知識グラフというか、RDFを扱えるようなデータベースを使っています。ちょっと効率化とかの問題もあるので、また別のやつを試していますが、そこら辺は、我々の中の技術者も含めて、外注の業者さんとも話し合いながらやっているというような形になっています。
だから、データベースと我々が言っているのは、知識として抽出した後のものをちゃんと管理して使えるようにしているというところをデータベースと言っています、ということでお答えになっていますか。
【千葉委員】ありがとうございます。しばしばハードウエアの話はこういうところでよく出るんですけど、ソフトウエアもだんだん陳腐化してきて、動かないという……。
【五斗先生】そうですね。あります。
【千葉委員】ことがありがちなので、ちょっと伺いたかったんですね。以上です。
【尾上主査】ありがとうございました。五斗先生ありがとうございました。
それでは、続きまして、マテリアル分野として、物質・材料研究機構より出村先生からプレゼンテーションをお願いいたします。同様に15分程度御発表いただきまして、その後10分程度の質疑応答とさせていただきたいと思います。
それでは、よろしくお願いいたします。
【出村先生】NIMS、出村でございます。本日、オンラインで失礼いたします。皆様よろしくお願いいたします。
マテリアル分野から、我々が今までやってきたことを中心に共有したいと思っております。タイトルに「蓄積から活用に向けた課題」と入れておりまして、これからどういうふうに活用を広げていくかというところが今課題でございます。
政策的なところを申し上げますと、文部科学省のマテリアルDXプラットフォーム事業というものがございまして、これは3つの事業で構成されています。象徴的にデータを「つくる」とか「ためる」とか「つかう」というキーワードでまとめておりますけれども、左側にあるのが、省略形だけ言いますとARIMという事業でして、この中では、基本的に共通設備を皆さんに提供するということが軸に1つあって、その共通設備を使ったユーザーのデータを集めてくると。真ん中の上にありますデータ中核拠点、MDPFでは箱を用意して、ユーザーからのデータをお預かりする。また、後ほど説明するようなデータベースを独自に構築したりしています。
さらにこういう基盤的なものを使って、フラッグシップのいわゆるデータ駆動による成果をいろんな材料領域で出していこうというのが、最後、DxMT、ディーマテと我々呼んでいますけれども、こういう構成で主には材料分野のナショナルプロジェクトが進んでいる状況です。
まず、NIMSにおけるデータ蓄積の取組ということで、我々世界最大級の材料データベースを構築してきました。まず、これ最近、私いろんな種類に分けて説明するときに、いわゆる論文から専門家がキュレーションしたようなタイプと、それから、リファレンスデータというふうに、ちょっとこれ私の定義で呼ばせていただいていますけれども、意図的にこういう段取りでデータを取りますというこ

とを決めて、実験にしろ、計算にしろ、データを系統的につくっていくというような種類、この2種類にまずは分けて説明をしたいと思います。
論文からキュレーションするものでは、無機材料の結晶構造を収集しているもの、これに論文の中に材料の特性があればそれも収録すると。それから高分子の同じくポリマーの構造を収録しているもの、これいずれも世界最大のものになっています。
最近はこれに目的に応じて、例えば電池の研究をしたいということになれば、特に電池の固体電解質や正極材料、こういったものについての構造と特性のデータを集めるというようなデータベースの構築をしたり、あと、Starrydataというデータベースのフレームワークを用意して、こちらはプロットを全部データ点に読み取るという結構凝ったことをしていまして、これも特定の分野、例えば熱電材料、電池材料、磁石材料、これニーズに応じて機動的にデータがつくれるようになってきています。
リファレンスデータのほうでは、データベースをつくる仕事というよりは、もともと、例えば金属材料の信頼性を評価するための試験を我々長年やっておりまして、例えば40年近くずっと引っ張り続けるような試験をやっています。こういうデータをデータシートという形でまとめて出版していたんですけれども、これをデジタル化する形で、皆さんに使っていただくというようなタイプになっています。
そのほか、いわゆる計算によってデータをつくる。これ欧米ではマテリアルズプロジェクトといいまして、いわゆる第一原理計算で完全結晶の構造をたくさん計算するというようなプロジェクトでデータが国際連携でつくられていますけれども、我々のところでも、そういうものに協力しながら、高精度な、例えばその中に入っていないようなフォノンの計算であるとか、合金系の計算をして、これらをデータベースとして提供するというようなことをやってございます。
また、他機関、JAXAさんから頂いたデータを我々のほうでデータベース化して提供するというようなこともやっている次第です。
3番目の領域として我々最近取り組んでいますのが、日々の研究からデータを直接蓄積すると。いわゆるワーキングデータを取り込もうということをやっています。ワーキングデータには、いわゆるうまくいかなかった失敗データが含まれていると。機械学習をしようとすると、成功したケースだけではなくて失敗したケースもやはり同時に学ぶことが重要なので、ぜひこういうものを集めていきたい。
NIMSで様々、2017年から、一種の社会実験的に研究者に協力してもらっていろんな方法を試したんですけれども、その中で知恵として出てきたものが、我々独自に開発をしたデータ構造化・収集のためのRDEという名前のシステムです。これは計測から出てくるファイルと、あと、カスタム化した入力フォームを処理するデータセットテンプレートというものを自在にアダプターのように付け加えることができて、様々な装置や様々な研究目的に自在に対応できるような仕組みになっています。
ユーザーは自分の使うテンプレートを選んで、装置からデータをアップロードすると。そうしますと、そのテンプレートに入っているPythonスクリプトに従って、構造化、例えば表形式でデータを取り出したりとか、あるいは計測ファイルのヘッダー部分に通常よく書かれている計測条件のようなものを機械可読可能なJSON形式、あるいは読み方も、機械の装置メーカーに独自な、ちょっとほかの人からじゃ分からないような形式ではなくて、一般的な用語で整理されるような形に翻訳をして集めると。こういうものをデータベースの中に一貫して格納するというような仕組みになっています。
ポイントは、装置や研究ワークフローごとに構造化のためのテンプレートを設計するという非常に手間のかかることがあるわけですけれども、これ幸い、2023年度から開始をしまして、今、3年度経過をして、マテリアルの分野の中の割と標準的なものとして認知が進んできております。
特に、冒頭申し上げたARIMという設備を共用する中では標準システムとして採用されておりまして、この中に参画している26機関の1,030台の装置が既にテンプレート化されていまして、この1,030台の装置を利用されているユーザーはこのテンプレートを使ってRDEの中にデータを預けていただくというようなことをしてございます。
預かったデータは、2年のエンバーゴ期間を経て、広域的なシェアにする。後で少し御説明しますけれども、そういうような取組でデータの共用化というものを推進するということをやっています。
データのファイル数だけいいますと、400万で、これ大体毎年アベレージで100万ファイルから150万ファイルがこの3年の間にたまってきているという状況です。
特にARIMは、先ほど申し上げた装置のデータ中心なんですけれども、幾つかの研究室では、研究で生まれるプロセス条件から装置のデータまで全部ひもづけるような形でRDEをいわゆるデータマネジメントのシステムとして使っていただいています。
これはNIMSの磁性を研究している高橋さんからお借りしてきたスライドなんですけれども、成膜条

件とそれから電子顕微鏡の実験結果、それから磁力測定の結果がひもづいている。さらにデータを処理する部分に、例えば自動セグメンテーションのようなプログラムをつけておくとか、あるいは、このグラフも大事なんですけど、ここから飽和磁化とか、研究に使う特徴量があって、そういうものを自動で判定して取ってくるようなものを付け加えることで、今まで人手でやっていたところをかなり軽減できる。かつ、一定の形で必ずデータがたまっていくので、学生さんが卒業した後も、データを探して、あれどこ行った、これどこ行ったということにならないという点でも、入れるまでちょっとハードルが高いんですけど、入れた後は非常に便利に使っていただいています。
我々はこのマテリアルDXプラットフォーム全体でデータのエコシステムをつくっていきたいと考えていまして、ARIMから例えばデータが入ってきますと。エンバーゴ期間を経て皆さんで使える状態になります。これにもちろんユーザー御自身のデータも預けてもらって、さらに冒頭説明したNIMSの長年蓄積してきたデータ、これもバルクで使えるようにする。このために我々クラウドのデータ基盤を用意しておりまして、この上で機械学習等ができるpinaxという名前のデータ解析基盤も用意してございます。この上でしたらこういうデータを統合的に使っていただいていいですよということで、昨年の12月に一般公開した次第です。
さらに、共用化したデータですね。広域シェアになったデータを、これは有償でライセンスするサービスも去年の9月から始めております。こういういわゆる論文に載ってない実験データをライセンスするというのは多分私たちが調べた中では世界で初めての試みになっているんじゃないかなと思います。
まとめていきますと、我々のデータは、左からしっかり集めているようなタイプのリファレンスデータ、実験や計算、それから、ある程度論文の中でこされたデータ、それから、全然こされてなくて、多様で、品質もそういう意味ではいろいろあるという研究ワーキングデータというような3つの構成になってございます。
こちら、参考までに、今フルオープンになっているものや有償に提供しているもの、それから閲覧はできるんですけれども、データセット全体としては別の契約が必要ですというものに分けて表示をしてございます。
さらに我々、リファレンスデータの革新として、自動自律型の実験にも取り組んでおりまして、ロボットや、あるいは自動的な計測装置を組み合わせることによって、様々、自律的な研究を進め始めているんですけど、それを別の面で見ると、系統的なデータを取る非常によいデータ工場としても位置づけている次第です。
ここから少し、私ども、今日の機会で考えを整理したところを共有したいなと思っています。
まず、マテリアルの分野で、データ時代というか、データ駆動の研究が本格化したのは、2010年、本格化というか、入り始めたのが2010年の半ばぐらいじゃないかなと思います。そういうわけで、この10年ぐらい、データ時代にようやく入った。
データ時代以前から我々データベースを集める仕事をやっていたわけですけれども、これは主な用途としては、人が参照するためのものでした。ですので、データベースのアプリケーションソフトウエアとしては、ウェブを通じて検索や閲覧ができるようなものになっていると。その上で、データの価値としては、研究開発をするときに、例えば事前に調べたり、補助的な情報として人が見て、ものを判断するというものでした。
データ時代に入ってきまして、AIの解析や機械学習というものがだんだん主目的になってきて、利用形態も検索・閲覧ではなくて、データセット丸ごと使わせてくれというような要求がどんどん出てきて、NIMSはどうしてオープンにしないんだと言われて、いやいや、検索・閲覧はオープンですよと言ってきたんですけど、話がかみ合わなくて、なるほど、データセットかということで、データセットの提供の仕方を我々整え始めたというような感じです。
AI for Scienceに入ってきますと、いわゆる知識生成というものが主流になってくると、我々、これからですけど、捉えていまして、主な利用形態としては、データそのものではなくて、むしろ知識とか、それを学んだAIを提供していくという時代に入ると思っています。
どちらにしても、データがAIの学習資源になるということは変わらないので、バルクのデータそのものが競争力を持つ時代にいきなり入ってきまして、それで我々、何周も遅れているんですけど、こつこつためてきたデータがいきなり価値を持ち始めて、それをどうやって皆さんに提供するかということに直面しているわけです。
我々、そういうことを考えていって、全世界オープンの世界ももちろんオープンサイエンスとしてあるわけで、これも大事にしていきますけれども、一方で、各機関や企業の中でクローズドにして競争力の源泉になるようなものもあります。その間として、クローズドシェアといいますか、広域でデー

タをシェアしていく。しっかりと登録・認証された方に対して選択的に提供していくというような、こういう領域が材料の分野では必要なのではないかということで、先ほど来、広域シェアのものを、例えばライセンスで提供を始めましたというようなものもその1つになっています。
データの活用に向けてどういうふうにやっていったらいいかということが私たち非常に悩みどころになっていまして、データを単に提供するところからやっぱり知識として提供していくということが重要だと思っています。1つは、利用ハードルの高さで、ここはAIエージェントによる支援を期待しております。
もう一つが、材料の探索空間の大きさに比してデータははっきり言えばまだまだまばらで、少ししかない。データそのものをぴたっと渡すのではなく、そこからくみ出した知識として渡すことで、いろんな未観測領域についても対応できるようにしていくというのが大事で、これからのデータ基盤の仕事ではないかと考えています。
ということで、我々としては、AI for Materialsと呼んで、今、データとAIと自動実験を組み合わせてマテリアルイノベーションを起こしていくような、こういうAIエージェントやマテリアル基盤をつくっていきたいと思っています。
最後にNII様への期待ということで、我々マテリアル分野でドメインのデータの創出・蓄積をやってまいりました。学術基盤としてNII様に期待するところは、やはり1つは分野横断のデータ連携のところ。もう一つが、AI知識基盤の部分ですね。マルチモーダルな基盤モデルをどうつくっていくかという部分。それと最後にデータ基盤や信頼保証技術、認証や認可の技術、また、データ処理も我々非常に高速にデータ処理をしないといけないので、データベース技術としてどういうものが標準的で、そして非機能的な要件を満たすにはどのようなデータベース技術が必要かと、あるのかということをぜひNIIの方々に教えていただけるとありがたいなと思っています。
時間が来ていますので、まとめは読み上げませんけれども、最終的には柔軟なデータ蓄積、それから戦略的なオープン＆クローズド、AI知識と統合したような形で次世代の研究基盤をつくっていき、我々の産業、我々の領域でいきますと、部素材産業の競争力強化、アカデミアの研究力向上に我々としては資していきたいと思っている次第です。
以上でございます。
【尾上主査】出村先生、ありがとうございました。それでは、ただいまの御説明に対しまして、御質問等ございましたら挙手にてお知らせいただければと思います。
いかがでしょうか。
若目田委員、どうぞ。
【若目田委員】御説明ありがとうございました。民間から見ても、マテリアル領域は非常に重要な領域だと認識しておりまして、最後のほうにコメントありました、データを渡すのではなく、データから出た価値・知識を共有していくという世界感は全くそのとおりだなと思っておりまして、民間でも、マテリアル領域に関していうと、コンフィデンシャルコンピューティングですかね、秘密計算であるとか、そういう別の技術等々の活用というものが非常に期待されているわけですけども、NIMSさんの中で、そういうコミュニケーションコンピューティングであるとか、データは秘匿したまま価値は共有するような取組というものが具体的に行われているのであれば教えていただきたいなと思います。
【出村先生】まず非常にプリミティブな技術ですけれども、いわゆるフェデレーションラーニングで、8機関、材料メーカー、それから重工メーカー、それから電力会社、NIMS、それから、日本原子力研究機構を入れまして、皆さんのお持ちの、クリープデータというんですけれども、金属の信頼性に関するデータについて、データを秘匿した状態で共通モデルをつくるということをやった経験がございます。
もう一つは、産総研と協力をしていまして、産総研側でもう少しアドバンストな秘密計算技術を、いわゆる分散秘密化をやっておりまして、それについて我々のところにも分散計算のためのワークステーションのサーバーを置いて一緒に遠隔地で検証実験をするということをやったことがございます。
【若目田委員】ありがとうございます。NIIさんの中でも秘匿の基盤というものが開発の項目に挙がっていたと思うので、そういうものとの連携の中に、何か個別要素なのか、この辺、うまく連携ができるといいかなと思いました。ありがとうございます。
【尾上主査】江村委員、どうぞ。
【江村委員】ありがとうございます。うまく質問し切れないところがあるんですけど、AI for Scienceの時代になって、それから自動実験をやるという形になったときに、今まではデータの種類

とか量とか規模とかというのが結構勝負になっているようなイメージがあったのが、今度、実験のほうもリアルタイムで回っていくようになると、データの使い方そのものが変わってくるんじゃないかなと思うんですけど、その辺について何かお考えがあったらお伺いしたいと思うんですけど。
【出村先生】ありがとうございます。まさに我々もそう考えています。具体的には、自動自律実験は探索範囲の中を探すのは大変上手ですけれども、探索範囲を決めるというところが大仕事になっていて、そこにこれまで蓄積していきたいろんなデータであるとか、材料学の情報を賢く学習したような、そういうマテリアル基盤モデルが、探索範囲の適切な設定に支援してくれるというような形でデータの価値が変わってくるんじゃないかなと我々も見ています。
そういう意味では、大きなループがあって、どこを狙っていったらいいのか。どこを狙っていったらいいかが決まったら、あとはロボットがその中の局所最適、局所最適という言い方はよくないですね、その中の最適化問題をリアルワールドでやる。また、そのデータが大きなループの中に戻ってきて、探索範囲設定をより賢くできる。何かそういうことがこれから起こるんじゃないかと我々期待しています。
【江村委員】分かりました。大変分かりやすい動きで、やっぱり変化感、大事ですよね、これからね。
【出村先生】そうだと思っています。
【江村委員】ありがとうございました。
【尾上主査】続きまして、石田委員、どうぞ。
【石田委員】九州大学の石田です。2つほど質問をさせてください。まず3ページのところで、データベースの構築について、学術論文からキュレーションしていってデータを抽出していくというお話があったと思うんですけれども、これがどれぐらいの規模で行われているのかということをお聞きできたらと思います。これは多分データの質にも関係してくると思うので、それをお聞きしたいというのが1点目でございます。
それから2点目は、この状態というのは、系統立てて、非常にすばらしいデータの収集、それからデータベースの構築というのを行っていらっしゃるなということがあったんですが、これは先生の実感というか、感覚で構わないんですが、マテリアルサイエンスの分野だからできたものなのか、それとも、他分野でもこういった取組というのはできる可能性があるのかというようなところをもし思うところがございましたら教えていただきたいなと思います。
以上です。
【出村先生】まず論文からデータをキュレーションするときの規模感、AtomWork-Adv、PolyInfo、Battery、Starrydata、いずれも、少しサイズはいろいろあるんですけど、最大でも10名から15名ぐらいでやっています。我々NIMSのサイズからすると、結構かけているなという感じなんですが、実は世界的に見ますと、いわゆる米国の化学界を母体とするケミカル・アブストラクト・サービス、CASと言われている会社がありまして、試薬にもCASナンバーというのがつけられていると思うんですけど、彼ら、同じように論文からいわゆる化合物の情報を集めています。一説によると、バイトも含めて、バイト的なものも含めて、3,000名近いキュレーションでやっていると聞いていまして、私たち、結晶とか高分子というふうにぎゅっと絞り込んでやっているので、何とか彼らから見てもある程度魅力があるものをつくれていますけど、ちょっと物量ではかなわないというようなイメージです。こういう論文、それからリファレンスデータというのはライフの世界等でもあるんじゃないかなと思いますけれども、特にワーキングデータについては、これはかなり、本当にこういうことを集めて意味があるのかとか、うまくいくのかというのは、かなり皆さん懐疑的な中で、とにかくやってみようという、かなり力技で、いわゆる国家プロジェクトとして構成してやってきたということで何とかやれています。
そういう意味では、我々集めたデータがちゃんと知識化して価値を持つということを示さないと、これやっぱり先続かないということで、マテリアルだからできたのか、そういうふうにえいやと踏み込んだからできたのか、そこはちょっと私のほうでは分かりかねるんですけど、実態としてはそんな感じでございます。
【石田委員】ありがとうございました。
【尾上主査】続きまして、千葉委員、どうぞ。
【千葉委員】千葉です。先ほどと似た質問になるんですけれども、特にマテリアルの分野では長年データを蓄積されていた経験があってすばらしいと思うんですが、逆にいわゆるデータベースは動かすためにはソフトウエアが必要で、やっぱりソフトウエアはどんどん陳腐化していくと思うので、それを長期間運用していくノウハウといいますか、知見があればぜひお話しいただきたいと思うんです

が、いかがでしょうか。
【出村先生】ありがとうございます。私たちもまあまあ大人数にウェブでアクセスに応えていかないといけないので、実証のところは外部にお願いをしていまして、そういう意味ではかなり我々も外部に依存している体制になっていて、かなりお金もかかるというところが悩みです。先生のおっしゃった部分は、データベースのソフトウエアだけではなくて、それを支えるミドルウエアとか、OSとか、そのレベルでも適切にパッチ当てていってフォローしていくという意味でかなり正直苦労しながらやっています。
そういうところが、もしNIIさんで、こういうタイプのデータベースだと、これがパフォーマンスがちゃんと出るよというところをやっていただけるようであれば、我々はドメインの材料に集中してやれるなというような期待はございます。
特に、最近は、NIIへの期待のところですね。ちょっと下に書いたんですけれども、がちっと構造化を最初から決め切れないものが研究現場でたくさんあるので、後から構造化したいと。でも、入れるときにドロップボックスみたいに適当に入れちゃうと、これ結構、どこに何が入っているかというところから始めるの大変なので、やっぱりある程度整頓して入れるんですけど、後から構造化するという、そういういわゆる我々が調べたところによると、データレイクでもなく、データウエアハウスでもなく、レイクハウス系の技術が最近あって、こういうものを我々有望じゃないかと思っているんですが、NIIの専門家でそういう方がいたら、ぜひ一緒にやっていただけるとありがたいなみたいな、例えばそういうこともあります。
【尾上主査】それでは、最後に吉田委員、お願いいたします。
【吉田委員】出村先生、どうもありがとうございました。今日、バイオの話、データベースとちょっと対比していろいろ思ったところが、バイオの場合は、遺伝子とか、たんぱく質とか、疾患とか、化合物とか、何かいろいろ異なるデータベースがうまくつながる要素があるわけなんですね。一方で材料の場合は、一見すると、無機のデータベースと高分子のデータベースとか金属のデータベース、なかなかデータベース間の統合というか、バイオみたいな形でいろんなインテグレーションができると何かすごく面白いなと思った点と、あとは、バイオの世界で進んでいるようにデータベース間の国際連携みたいな、今日NIMSのデータベースが中心で、あとは、ARIMとのデータベースが、国内のデータが中心だったと思うんですけど、同時に世界各国でいろんなデータベースができていて、そういうデータベース間の国際連携みたいなビジョンもあり得るんじゃないかなと思うんですけど、それについてもしお考え等あれば、御意見をお聞かせいただけると幸いです。
【出村先生】ありがとうございます。最初の点、おっしゃるとおりで、今参考スライドを皆さんにお配りしてないものを初めて示しておりますけれども、材料は、対象とかプロセスのやり方とか使う目的が様々なので、かなり統合するといっても、高分子と無機でどうするんだろうって非常に悩むところです。まだ答えがしっかりあるわけではないんですけれども、実は1980年代に材料という分野が新たに再定義されるときに、物理とか化学とか機械工学とか、いろんな近接領域の先生方で集まってアメリカで教育のために議論した中で、共通のフレームワークというものができました。
それが非常にプリミティブなんですけど、プロセスと構造、ストラクチャーとプロパティ、特性と材料を使っている環境における性能、パフォーマンス、この4つのどこかで皆さん研究しているねと。そういうふうにカテゴライズすることができるというのがありまして、我々も、いろんな材料があるんだけれども、つくり方と構造とそれから特性と性能、こういう形で材料のレイヤーを見たときに、特に構造の部分は、簡単に言ってしまえば、原子の並べ方とその不均一の問題なので、ここはかなり共通性を高められるポイントだと思っています。
ですので、分子の構造であるとか、原子の構造であるとか、あるいはここにちょっとちらっと書いていますけど、もう少しメゾスケールな不均一の情報、こういうところをモデルにぐっと持っていくことで、いろんな分野を統合したような知識にできるんじゃないか、そういうことができると非常に面白いんじゃないかと思っています。
次、国際連携については、ぜひ我々も進めたいと思っています。一番突破口になるのはやっぱり計算データだと思っていまして、計算データは物理の方が中心にやられていて、マインドもオープンサイエンスで、かなり、道具もそうですし、計算したデータそのものも、みんなで同じ計算する必要ないよねということで、共通化、連携が進んでいますので、我々も、チャンスがあって、例えばそういう計算データをつくるようなアクティビティーができた場合には、そういう計算データをできるだけオープンでつくると。データのツールもオープンにしていって、国際的にやる中に一緒に入ってやっていくということができたらいいかなと思っています。
実験データのところはまだいっぱいいろいろ議論があるところだと思いますけど、いろんなチャレン

ジをしていくべきかなと思っています。ありがとうございます。
【吉田委員】実験データに関しては多分NIMSが世界をリードしているんじゃないかなと思っていまして、そうすると、そこのフィールドだと、逆に日本が世界をインテグレーションリードできるポジションを取れるんじゃないかなと思っています。
【出村先生】ありがとうございます。いろいろとぜひ議論させていただければと思います。
【尾上主査】出村先生、ありがとうございました。続きまして、大型実験装置、放射光施設としてSPring-8に関する状況をお伺いできればと思います。理化学研究所より初井先生に御参加いただいております。同様に15分御発表いただいた後に10分の質疑応答とさせていただければと思います。
初井先生、どうぞよろしくお願いいたします。
【初井先生】ありがとうございます。理化学研究所の初井と申します。このたびはこのような機会をいただきましてありがとうございます。AI for Scienceといったところというのは幅広うございますので、今回お示しするのはごくごく一端というところにはなるかと思いますけれども、状況について御報告と取組状況を御報告できればなと思っております。よろしくお願いいたします。
2ページ目見えていますでしょうか。SPring-8ですけれども、一周1.5キロの周長を持つ加速器から高輝度のX線を生成する施設になっております。
非常に大きな施設になっておりますのは理由がございまして、より短波長のX線を生成するには大型の施設が必要だということで、東北に設置されていますナノテラスがより長波長のところを担当するのに対して、短波長のところを担当する施設ということでSPring-8が稼働しております。
延べ利用者数は1万4,000人日というところが1年間当たりというところで、ユニークユーザーで1万人ぐらいがいて、大体3年間で使われる方が1万5,000人ぐらいがいて、入れ替わり立ち替わりというところで使われていて、約2割が民間利用というところになっております。
利用分野は非常に幅広くありまして、生命科学（薬学）、化学、材料科学、エネルギー、文化財、地球科学、あるいは素粒子実験みたいなものもやっておりますし、非常に大型の動物とか、そういったものの技術を利用して、最近ではコンクリートとか、そういったものの研究も行われております。
SPring-8-IIに向けまして、SPring-8の役割を拡大するようにというような、いろんな答申が出ておりまして、これまでの募集型の利用以外に国の国家戦略に合致したような利用を拡大するということを要請を受けておりまして、具体的には半導体戦略や国土強靱化、あるいは農学・食料安全保障などの利活用というのが最近増やしてきているというところになります。
能力的には2029年度にSPring-8-IIに高度化するというところで、高輝度性という意味では100倍の性能向上が見込まれております。
この施設のデータ基盤に関しましては、分野ごとにデータの特性あるいはオープン／クローズ方針、解析・ワークフローが違うということで、分野スペシフィックなものというよりは、各分野を支える共通の基盤が必要というところで、その部分について、共用補助金によって整備がなされております。
最初の設置の段階では、補正予算によってデータセンターが設置されまして、現在、大体5本分ぐらい、全体で60本弱のビームラインがあるんですけれども、5本分ぐらい、データ帯域換算で約25％程度について措置がされているというところになっております。
データセンターは、したがいまして、どちらかというと、データの流通と呼んでおりますが、共同研究者間でのデータがスムーズにやり取りができるですとか、実験中のデータ解析の支援、データを持ち帰って解析して、もう1回実験しに来るということを避けて、その場で解析して、次の実験に移っていくといった、良質のデータをいかに早く確実に得ていくかといったところの支援をする基盤として位置づけられております。
したがいまして、できないことが結構ありまして、例えば長期保存ですとか、大規模のデータ解析については、HPCIのリソース、HPCIストレージで長期保存あるいは富岳等を使った大規模データ解析というところで、我々のほうでそういった連携が簡単にできるような基盤を提供しているというところになっております。
民間企業さんですと、パブリッククラウドを利用したデータ解析というのも盛んに行われております。
さらに、最近ですが、GakuNin RDMの連携サービスも試行しておりまして、GakuNin RDMで研究室データを管理されている大学の先生方におかれましては、SPring-8のデータも同じGakuNin RDM上から確認とか管理ができるというような状況になってきております。
具体的な例を若干御説明したいと思います。1つは、データセンターを利用した大規模計算とその場測定を用いて新しい材料を創出するというような研究です。こういった材料、特に多元材料探索では

組合せ爆発が起きて、実験のみでは探索効率が低いということで、大規模な理論計算と実験の統合が重要であると考えられております。
SPring-8の長所は、大量の試料あるいは条件で測定ができるというところで、1日に数千から数十万のデータの取得が可能です。実際の装置が、ちょっと見にくいかもしれませんけど、右の写真にありまして、こちらのほうは完全自動になっておりまして、サンプルをカセットに入れますと全て自動で、調整も含めて自動で行われて測定されると。データが自動的にデータセンターで格納されていくというようなシステムになっております。
御紹介する事例ですと、第一計算によって大規模な構造予測を用いて探索を行い、ある程度スクリーニングした上で、放射光X線開設によって高速にスクリーニングして、実験データで得られた構造から候補物質を選定、さらにその結果を得て新規セシウム塩化物の合成を行ったところ、そういったものができましたというような事例になっております。
次に、こちらのほうではX線の画像検出器が使われているんですけれども、データに関しましては、画像検出器が重要ですので、画像検出器を若干トレンドを御説明したいと思います。
X線画像検出器の開発トレンドですけれども、横軸が運転、運用を開始した年で、縦軸が、左のほうが1時間当たりのデータ、右のほうが年間のおおよそのデータというところになっております。
SPring-8に隣接しております施設SACLAでのデータというのは大体、検出器あたり年間2ペタバイトぐらいだったんですけれども、SPring-8になりますと、検出器あたり年間ペタバイトとか、そういったところが出てくるというようなことになっておりまして、今後、SPring-8-IIといったところになりますと、一つの検出器から年間エキサバイトとか、そういった非常に巨大なデータが出てくるということになります。
トレンドとしましては、10年間で100倍といったトレンドになっております。ムーアの法則等を見ますと10年間に40倍とか、そういった速度ですので、X線画像検出器の出力帯域トレンドは、トランジスタ密度のトレンドを上回る速度で向上しております。
ネットワーク等の他のIT技術はこれよりも遅いという場合もございますので、放射光施設でのデータ基盤ではR&Dが非常に重要になっていて、特に圧縮技術を含む広帯域のデータ処理パイプラインの開発が重要だと考えられています。我々も含め世界的に、そのような理解の下に開発を行っているということになります。
具体的な実装例を1つ御紹介いたします。この事例では我々が開発したCITIUSという検出器を利用して1日2.3ペタバイトといったデータが出てくる実験系についてのデータ処理パイプラインの実装例です。
独自開発のFPGA演算加速ボードを使いまして、リアルタイムにデータ前処理して情報抽出・データ圧縮を行っております。その後、自動データ転送とデータセンター内にあるクラスタによって即時解析が行われて、そのときの測定条件によって正しくデータが取れているか、あるいは十分なデータの統計があるかどうかが判断できるようになっております。
ユーザーとの応答においては、理化学研究所のR-CCS（計算科学研究センター）の支援を受けまして、OpenOnDemandを導入しておりまして、ビームラインの端末のブラウザからで確認はできるということになっております。
実験条件、あるいは解析履歴を含む構造化されたデータとしてデータセンター内に保存がされます。
今後ですけれども、現在、このパイプラインについては、米国のアルゴンヌ研究所のAPS施設が興味を持っていて、我々に支援を求めてきたということがありまして、このデータ処理パイプラインの技術供与と立ち上げを行っております。
また、APS施設と共同でAIによるリアルタイム解析、あるいは実験制御を見据えた基盤技術を開発しています。
米国Genesisミッションでは、新しいデータ形式等も議論されていると聞いておりまして、そういったところでのデータレベルでの相互運用というのを今後検討するというようなお話をしているところです。
国家課題解決に向けたトップダウン型の戦略的利用というところで、幾つかあるわけなんですけれども、そういったものの中でAIの利用というところを少しお話をしたいと思います。
これらの解析にはデータが非常に大きいということもありまして、富岳ないし富岳Nextでのデータ解析というところが前提となっています。
現在、コンクリートやアスファルトの解析を行っているわけなんですけれども、そういったものですと1日大体0.5ペタバイトぐらいのデータが出てまいります。これらの実験データを実際に計算機の中で像再構成を行って、三次元の計測データに変換をいたします。ここの部分は富岳の非常に大きなメ

モリーを使うことによって非常に高速にできるようになっています。
その後で、R - CCSのWahibらが開発したビジョンAIを使いまして、セグメンテーションを行い、セグメンテーションの結果から劣化の様子を見るといった知見を現在得ようとしております。既にアスファルト等ですと、約2年から5年程度の長寿命化ができる見込みが出てきておりまして、年間、非常に大きな予算削減効果があると期待されているところです。
もう一つは開発中の半導体への応用で、回折顕微鏡というのがございますが、その場合、1日に30ペタバイトといった非常に巨大なデータが出るというところで、これをリアルタイム圧縮をいたしまして、富岳に転送し、富岳上で三次元の像再構成、さらには領域分割を行うということを今検討しています。
こういったAIによってSPring-8の解析能力を強化するという観点での課題について本日は2点お話ししたいと思います。
1つは、富岳といった外部の計算機にデータを移動するというところで、ビームライン、データセンター、富岳と3か所のストレージをデータが移動していると。いわゆるステージングがユーザーさんの利便性を損なっていると。計算しようと思うとデータがないので、転送を待つとか、そういったことで、データの場所と計算内容をよく考えないと実際に計算できないといったところが課題になっています。
もう一つは、セキュリティーで、先端半導体や電池開発などでの産業利用で、高度なセキュリティーが企業様との連携では要求されています。既存インフラでは対応ができないというところで、例えば、民間からは民間クラウド事業者のベストプラクティスと同等の運用サービスが求められていて、サーバー設置室の警備員配置ですとか、試料あるいはデータの監視システム、あるいはユーザー領域の物理層での隔離オプションといった事項に加えて、監査ですとか補償対応といったところが求められています。
また、これらの研究において実験自体をLLMによって補佐するということもテクニカルにはできるようにきているんですけれども、そういった場合に機密保持に対応できるようなローカルの高度なLLMが必要というところが課題になっています。
1つのやり方として今検討しているのは、SPring-8から富岳へのデータ転送として、ステージング、エッジ、データセンター、富岳と3か所にストレージがある現状を、NTT様と共同研究しています IOWNを使ったデータストレージ技術というもので解決しようとしています。この場合、富岳のストレージに直接書き込むということを今前提にR&Dを行っています。
この場合、IOWNのオールフォトニクスネットワークという特徴を利用して、特定波長を独占的に利用できるということもメリットとして挙げられます。専用線と同様に、物理的に通信経路を隔離できるというところで、これだけでセキュリティーが全て解決できるわけではないですけれども、非常にセキュリティー的にはメリットのあるネットワーク構成になっていると思っています。
データ基盤の現状として、幾つかのデータベース等のお問合せをいただいています。データベースに関しましては、施設としてはデータベースの作成等は携わってないんですけれども、ユーザーコミュニティーのそういった活動を支援するといったことを行っております。SPring-8のデータ基盤としましては、オープン・クローズはユーザー、正確には実験責任者が判断できるようなシステム構築を行っております。
これに対応してデータ基盤というのは、データ流通あるいは実験中のデータ解析に特徴があるようなリソースになっています。
データ管理について、先ほども申し上げましたけども、GakuNin RDMというものを使えるようにしましたので、学術でGakuNin RDMの利用が増えてくれば、こういった利用も増えてくると思っております。
データ構造化の取組は非常に重要で、AI可読な形式でのデータ保存というのが順次共用装置について対応しているという状況になっております。
さらにAI for Scienceに向けた課題という観点で幾つか申し上げたいと思います。
1つは、利用者へのファーストタッチというところで、特に放射光施設を使ったことのない方にとりまして、国内の施設に多数あるいは多様な分析装置、分析技術があるということで、それ自体はいいんですけれども、どれを最初に試したほうがいいのかといったところが非常に難しい、という指摘を受けております。こういった部分をAI等のコンシェルジュ機能でサポートすれば、放射光施設のより効率的な活用が期待できるというふうに、今、放射光の分野のコミュニティー全体として考えているというところになります。
認証について、SPring-8では独自の認証基盤を現在運用しているところですけれども、なかなかいろ

んな新しいサービスに対応できないということで、現在更新を検討しています。SPring-8には民間ユーザーが2割ぐらいいますので、民間ユーザーを含む対応が必須になっています。こういったものが国内の施設で共通の認証基盤として使えれば非常にありがたいというところで、NIIさんの認証基盤等がもし使えるようになれば非常にありがたいなと思っています。
ネットワークに関しましては、先ほど若干IOWNの話をさせていただきましたけれども、IOWNがSINETにオプションとして利用できるようになれば、いろんな施設がAIの計算リソースを持つHPCIの計算リソースにつながって非常にありがたいのではないかなと思っています。
最後にAI向けの計算リソースですけれども、例えばセグメンテーションだけをとりましても、非常に大きなリソースが必要だと現在見積もっておりまして、これらについての計算量の削減、R&Dも必要なんですけれども、いずれにしても相当程度の大きなリソースが必要と考えているところです。
以上です。ありがとうございます。
【尾上主査】初井先生、ありがとうございました。ただいまの御説明に関しまして、御質問等ございましたら挙手でお知らせいただければと思います。
少し私のほうから御質問させていただきます。最後、ここで出ているスライドのところなんですけども、先ほど実際に富岳等でも処理をしようと思うと、3つのストレージを渡り歩くことになってしまって、ここで書いていただいているネットワーク等で、IOWNでダイレクトにデータセンター、そこに対する処理になっていくというようなお話があったと思います。
これはSPring-8として見たときには、もちろんデータとして、非常に国家課題のものだとすごく大きいというお話があったと思うんですけども、単発の実験だと、データがすごく小さいものもあるのか、概して、あるいはSPring-8からSPring-8-IIになっていくときに、全般的に増えていく方向だと考えればいいのかというところを教えていただければと思います。
【初井先生】ありがとうございます。データのサイズに関しましては、考え方として2つある。考え方として、まずデータの類型というのが重要だというところで、必ずしもその類型に当てはまる例ばかりではないんですが、大まかに分けて2つの類型がございまして、今日最初にお示ししたX線の回折のデータの場合は、1日に数万とかいうデータが出てくるんですけれども、それぞれのデータは、この画像を、データを処理して、一番右の回折のプロファイルにまで解析した後でデータセンターに送るということで、一つ一つのデータは数百キロバイトというところで、非常に多数のデータが出るんですけれども、1個1個は小さいというデータになります。
この場合はトータルとしてもストレージを圧迫するようなデータではないんですが、データマネジメントですとか、それから、たくさんあってもリアルタイムに何らかの示唆を与えるような解析が働いてないと実験がうまくいかないということで、AIも含めたデータの解析支援というのが重要になっています。
もう一つは、いわゆるイメージングですね。本当に画像を撮りますという場合は、こういったコンクリートとか半導体の場合には画像を撮っていますので、なかなかデータの圧縮が難しいというところで、それでも最近は圧縮ができるようになってきています。実績でいいますと、コンクリート等のいわゆる実像の場合ですと10分の1ぐらい、下の半導体のほうですと1,000倍とか、非常に効率がよくなっておりまして、エッジでの圧縮というのが非常に有効な技術になろうとしています。エッジでの圧縮ができる場合は、圧縮した後で富岳に送るといったことを想定していて、それでも普通の実験よりは大きいわけですけれども、いわゆるHPCIストレージとか、そういったものの許容できるようなサイズにはなろうとしているというところになっております。
【尾上主査】ありがとうございました。そういう類型に応じたシステムの構成というのが必要になると思いました。ありがとうございます。
そのほか。石田委員、どうぞ。
【石田委員】九州大学の石田です。データの共有という観点からお聞きしたいんですけれども、ここのユーザーの方々は多分それぞれの実験の目的でデータを取られていると思うんですけれども、中には、例えば1つの実験をやって、その結果をみんなでデータを、結果をシェアするというような形のものもあるのか、やはりそれぞれ皆さん別個でやられているという形なのかという、その辺のデータの共有具合についてお伺いできればと思います。
【初井先生】ありがとうございます。まずデータセンターに接続しているビームラインに関しましては、測定したデータが、測定した直後にデータアクセス権限が設定されておりまして、共同研究者として申請された方はすぐにアクセスできるようになっております。
さらに、最初の申請時には想定していなかったんだけれども、実験をする中で、例えば情報科学の先生とか、そういった方が追加で必要になったという場合には、そういった方を実験責任者が指定をし

て、アクセスが後ほどできるというようなシステムになっております。
それから、データベースを最終的に構築するというところですけれども、研究データからデータのクオリティーチェックをして、実際に解析のクオリティーを確認してデータベースに登録するという流れがあるかと思うんですが、例えばたんぱく質の構造解析に関しましては、大阪大学がそういったデータベースを運用しているんですけれども、そういったものが一続きの流れでできるようになっています。
ただし、登録とかまで、少なくとも登録申請までは自動でしたいとか、もう少しITによる自動化等の支援が欲しいというようなお話もちらちら出てきておりまして、そういったものに関しては施設側で支援をしていくというような体制になっております。
【石田委員】ありがとうございました。
【尾上主査】林委員、どうぞ。
【林委員】NTTの林です。御説明ありがとうございました。データ転送のところで、特定波長を独占的に使っているという話があり、そこでIOWNの低遅延性も利用することで、高性能でというところを御説明いただいたかと思います。なので、低遅延性が大事ということは、結構リアルタイム性、重要かなと思うので、万が一のときの冗長構成とかも取られているんじゃないかなと思うんですけれども、そこのどういう方針かという辺りをちょっと教えていただけるとありがたいです。
【初井先生】ありがとうございます。あまり細かい説明ができておらず申し訳ありません。ネットワークとしての低レイテンシというのがどこに効いてくるかといいますと、富岳に置いてあるストレージを例えばSPring-8からファイルのリスティングをしたりとか、そういったときに、低遅延でストレージのファイルシステムのデータベースにアクセスできる必要があるんですけれども、そこの部分でやはり通常のネットワークですとなかなかリスティングができないとか、そのファイルがあるかどうかがすぐ分からないとかというところでデータ書き込み・読みだし速度がかなり遅くなってしまうということがございます。そういったところがIOWNで解消できるということを念頭に今R&Dをしているというところになります。
実際の実験のニーズとして、ロボットのようにリアルタイムで何か判断をするというところはほとんどなくて、大体秒から分ぐらいのレイヤーでデータ解析が行われれば実験はスムーズに行くというところになっておりまして、そういった意味では、レイテンシがSPring-8側の何かの判断に直接効いているというところではありません。
それから、波長で分離するところなんですけれども、やはり半導体の企業さんは2つに分かれまして、1つは、敷地の外にデータもサンプルも出しませんという会社がありまして、そういったところは放射光を扱わないということになってしまうんですが、やはり放射光の重要性を鑑みて、例外として試料ないしデータを出しましょうという会社もあります。そういった場合には、やはり監査とか、そういったものに加えて、原理的にはアタックされにくいといったことが要求されておりまして、可能な限り富岳とSPring-8を専用線で結んでほしいという具体的な御要求がございました。
専用性で結ぶというところは、費用を度外視すれば可能なんですけれども、やはり全体としてはなかなか受け入れがたい費用になってくるというところで、IOWN等の技術で、ほかの利用とも共有をしながら、必要に応じて通信経路を隔離できるといった技術が非常に有望だというところで、今、NTTさんと、今のIOWNのサービスではそういったものはないんですけれども、そういったサービスが技術的には可能性があるということで御相談をしているというところになります。非常に期待しています。
【林委員】御説明ありがとうございます。今、研究段階でそういう方向性を狙っているということで理解いたしました。ありがとうございます。
【尾上主査】最後、千葉委員、お願いいたします。
【千葉委員】今日のこのお話は、データの扱いについて伺いたいんですけれども、基本的には装置、SPring-8から出たデータを処理して、解析結果を出すまでのデータの扱いが重要という理解だと思うんですが、合っていますでしょうかということをまず伺いたいんですけど。長期間保存しておいて、ほかのプロジェクト外の人が利用するようなデータではないという理解でよろしいでしょうか。
【初井先生】学術としては全体としてそこまでいわゆるデータライフサイクルのマネジメントが重要だということは理解をしております。
一方で、施設としてどの部分を責任持って今やっているかといいますと、やはりデータの再利用とか、そういったところは、富岳、非常に大きなデータですので、富岳でのHPCIのストレージを活用してやっていただくとか、そういったことで、二重投資にならないように、SPring-8のサイトではSPring-8のサイト内の計算リソースでしかできないところにフォーカスをして、それ以外の部分につ

いては既に国の政策でそういう役割を担っておられるHPCI、あるいは今日のNIMSのお話もありましたけど、マテリアルに関してはNIMSのマテリアルプロジェクトにそのプロジェクトに参画されている方がデータを登録していくとか、そういったところを御支援するということで整理をして行っているということになります。
【千葉委員】そうすると、研究管理という観点で今日のお話を整理すると、データをなるべく、いわゆるSPring-8のサイトから、データセンターから富岳へコピーしたり、そういうコピーのオーバーヘッドを減らしていくのが大事という、そういうまとめ方をしてもよろしいでしょうか。こちらの部分のポイントといたしましては。
【初井先生】そうですね。ITシステムとしてはそうなんですけれども、運用してみて分かったのは、やっぱり人間がついてこないというところなんですね。結局いろんなところにデータがあって、データの場所が違うと、計算ができる種類が違うということ自身が、やはり我々のところのユーザーさんは、例えば本当にマテリアルですとか、ITの方ではないので、非常に使いづらいと。結果として利活用を損なってしまうというところで、やはりデータが、SPring-8の場合は、バーチャルにでも1つの場所にあって、そこにあるデータが必要に応じていろんな計算ができるというのが必要で、そうしますと、実装としてはやはり下のような形で、巨大な計算リソースが必要ですので、データと計算リソースが近接しないとユーザーの対応を満たせないというところなので、データの転送が重要になってくると。
【千葉委員】ありがとうございました。
【尾上主査】初井先生、ありがとうございました。すいません。司会不手際で既に20分ビハインドになっております。この後はNIIのほうで行っていただきました海外調査や国内インタビューの内容について合田先生のほうから御紹介いただきます。よろしくお願いいたします。
【合田先生】NIIの合田でございます。時間もないので手短に。
これまで4回にわたりまして私どもの基盤の現状と今後の計画についてお話しさせていただいたところでございまして、また本日も分野の先生方から数多くの御期待の声をいただきまして誠にありがとうございます。
今日お話ししたいのは2点でございまして、1点目が、これまでの議論の中で、どうしてもNIIの話ばかりではなくて、もうちょっと広めに見たAI for Scienceに必要な基盤という視点できちっと議論すべきとコメントいただいておりますので、それに対する私どもの回答といいますか、御説明をさせていただきたいと思います。
2点目が、冒頭にもございましたけれども、私どももいろんな分野の方々とお話をさせていただいておりまして、その状況を簡単に御説明させていただきたいと思っております。
では、最初の1点目ですけども、この絵はこの会議で初めてお見せしますけども、これまでAI以前のオープンサイエンスにおいてどう基盤を連携していくかということをこれまで私ども考えてきたわけですけども、その絵を俯瞰的にまとめたものでございます。
左下にデータ創出。データ創出の基盤というふうにAI for Scienceの基本方針には書かれていますけども、いわゆるここから実験装置等からデータが生まれてきたりですとか、またデータを生み出すのは実験装置だけではなくて、いわゆる人文系にあるように、人の作業によって生まれるデータというのもございます。
こういった様々なデータがここで生まれて、それが右側の計算資源に送られて、そこで分析や解析が行われると。さらにはそれが上のデータ基盤のところに蓄積されまして、いろんな利活用が進むというようなことがこれまでオープンサイエンスの中で目指したことでございます。
これから先、AI for Scienceの時代になると、特に上のデータのところにAIというものが加わりまして、データ基盤に蓄積されたデータといったものがAIによって活用されて、そこで新たな価値が生まれる。具体的には、上にありますけど、知識基盤機能と書いておりますけども、研究者がAIを使って、環境を使って様々な研究支援を受けることができるですとか、分野に特化したモデルの構築もできるようになる。こういったことができるようになるのがAI for Scienceだろうと思います。
結果的にデータのAIからデータ創出基盤にも矢印が伸びまして、研究者の日々の活動を支援したりですとか、今日もお話ございましたけども、実験装置の制御ですとか自動化にも資するということで、こうやって実験、データを創出する側も高度化すると。さらにそこから新しいデータが生まれて、このサイクルがぐるぐる回っていくと。これがまさに目指すべき姿であろうと我々考えているところであります。
左上に書かせていただきましたけれども、データの創出、計算、データ・AIの基盤が三位一体となってエコシステムを構築することが重要と考えております。こういった世界をつくるために私どもこれ

まで説明してきたことが上の部分になりますけども、従来の研究データ基盤に対して、AIの知識基盤の機能、AIの機能といったものを強化してつけることによって、上にある「AI対応研究データ基盤」というものをつくろうというのが今目指しているところでございます。この名称はまだ中でも検討中でございまして、これからよい名前を考えていければと考えております。
この絵がNII主体の絵とお叱りいただいた絵ですけども、先ほどの考え方をこれまでの議論に当てはめると、こういう青い矢印のようになるだろうということで追加させていただきましたので、お時間のあるときに御覧いただければと思っております。
では、2つ目の話で、分野からの期待ということで、これまで私ども、研究分野の方々、さらにHPCIの計算資源のコミュニティーの方々からのヒアリングですとかディスカッションを通して得た話を今日簡単に御紹介したいと思います。特に研究分野からの期待については、今日お話しいただいた方々とかぶるところもあるんですけども、どちらかというと補完的に御覧いただければと思っております。
まず、研究分野のデータ基盤のコミュニティーとして、ARIM、DDBJ、NanoTerasu、SPring-8、DIASの方々からAIに対する取組ですとか、NIIの今検討しております基盤に対する期待といったものをいろいろ議論させていただきました。そこでの議論を簡単にまとめるとこのスライドに書かれているとおりでございまして、大きく4つの期待があると我々考えております。
1つが、AIのモデル・データ処理の高度化でありまして、各分野で特化型のモデルをつくるときの構築ですとかチューニングといったものを一緒にやるという期待をいただいておりますし、メタデータの生成ですね。こういったデータの生成の自動化の話もありますけど、こういった部分。さらには、知識グラフ、今日もお話ありましたけども、これを活用することで検索・比較・発見といったものの高度化をするといったようなことで御期待の声をいただいているところであります。
次が右側の研究・実務の自動化・支援サービスでありまして、これはまさに、今いろんな研究者、学生さんも使い始めていると思いますけれども、研究のアシスト、レポートを作成する支援ですとか、今日もありましたが、実験の自動化ですね。また、今、様々な分野で問題になっているのが、課題や論文の審査の負荷でございまして、こういったものの業務の支援に対する御期待も非常に高いと考えております。
一方で、インフラに関する期待というのも大きくて、セキュリティーが担保されたローカルなAIの環境と。AI Sovereigntyという声もありますし、やはりきちっとセキュリティーが担保されて、中が見える形でのAIを動かす環境というのは重要である。特に機微性の高いデータを扱うときは重要であるという声をいただいていたりですとか、国産のモデルやOSS、信頼性の高いモデルといったものを使いたい。
また、計算資源についてもやはり課題であると考えております。
最後に右下、ここの部分は我々も再認識したところでありますけども、人材育成ですとか、知見の共有であります。今、とかくこういった議論をすると、インフラですとか、特にハードウエア、計算資源に議論が及びがちですけども、こういった基盤資源を活用するためには、高度なスキル、ノウハウが必要でございまして、こういった部分の専門知識の蓄積や共有ですとか、協働での人材育成といったものも大きな課題であると考えておるところでございます。
ここから先何ページかは各機関からヒアリングさせていただいた内容のサマリーがございますが、今日は時間の都合で割愛いたしますので、お時間のあるときに御覧いただければと思います。
続きまして、計算資源、特にこれはHPCIのコミュニティーの方々からの期待というのも簡単に述べさせていただきたいと思います。
1つ目がHPCIコンソーシアムです。HPCIコンソーシアムというのは、HPCIの資源を構成する基盤センター等の組織です。あと、ユーザーの組織ですね。コミュニティーから成る組織でございまして、まさにHPCIを構成する機関とユーザー機関の声がここにまとまっているとお考えいただければと思います。
HPCIコンソーシアムでは毎年、提言、ユーザビリティ等に関する提言のまとめを行っております。これまでの提言を改めてサマライズして、今求められている基盤の機能といったものをまとめたのがこのスライドであります。
1つ目、高速ネットワークが必要であると。これはもちろんのことではありますけども、今日も幾つか御議論ありましたけども、同一のユーザーIDでワンストップで資源を利用できる。つまり、認証連携、シングルサインオンといったような御期待というのも高いと思っておりますし、また産業界との連携も同様にしたいという期待の声も高まっていると思います。
また、最後、これも一つ今日話があったと思いますけれども、いわゆる計算機と実験装置がばらばら

にあるのではなくて、実験装置やIoTのデータを外部データベースと計算資源が直接連携してリアルタイムな処理を可能にする環境といったものも重要であるという声もまとめられているところであります。
2つ目がHPCI計画推進委員会、これは文科省の下につくられておりますHPCIの計画をまとめる委員会でございますけども、こちらにおいても、先ほどと同じように、ネットワーク、認証、基盤の連携といったようなことの重要性というのがうたわれているところであります。
続きまして、我々、同時に海外の同様のプラットフォーム、公的機関におけるAI for Scienceを見据えた研究のインフラといったものの動向調査を昨年度行いましたので、こちらについても簡単に御紹介したいと思います。
こちらについては、5月12日に開かれました情報委員会の中でも報告しておりますので、今日はサマリーだけを御紹介したいと思います。ページ何ページか飛びますけれども、すみません、ページ数見えませんけれども、この部分ですね。これは海外の例えばEOSCですとか、アメリカのアメリカンサイエンスクラウドのような、やはりAI for Scienceを見据えて基盤整備を行っている組織に対してヒアリングも含めた調査を行った結果であります。
この中で、海外が目指している方向性としては、この表の左側のカラムにありますけど、大きく3つあります。1つはフェデレーション。つまり、計算資源ですとか、データ基盤、こういったもの、認証連携も通じて、統合化して使っていくということ。
2つ目がAI駆動型研究を支えるデータ・エコシステムの構築ということで、先ほど最初にもお見せしましたけれども、データが出てきて、計算して、基盤モデルを使ってというエコシステムをつくることの重要性。
最後が、こういった基盤を効率的に運営するということであります。
こういった調査から見えて、日本の取り組むべき具体策を右側にまとめておりますけれども、まさに私、先ほどお話ししてきたことで、資源連携しましょう。きちっとモデルを活用できて、エコシステムをつくりましょう。やはり持続的かつ発展的な運営のための体制強化をしましょうということがこれからの我が国においても必要であると考えているところであります。
こちら最後ですけれども、こういったことを踏まえますと、やはりデータの創出、計算、データ・AIの3要素が密接に連携するエコシステムの構築というのがこれから不可欠でありますし、研究分野ですとかHPCのコミュニティーからもこれにつながる期待をいただいていると思います。
また、海外動向への即応や国際競争力の維持においてもこのエコシステムが重要でありまして、こういったAI for Scienceのための基盤の構築がこれから急務であると考えているところでございます。
以上で説明を終わります。
【尾上主査】 合田先生、ありがとうございました。ただいまの御説明に関しまして、御質問等ございましたら、挙手してお知らせいただければと思います。いかがでしょうか。
よろしいですか。今回、5ページ目にAI for Scienceのための基盤連携という形で、過去、ずっと6ページ目の図を出していただいた、一部の図を出していただいてたものを、もう少し広い全体を見たようなこういう図をつくっていただいているというところで、こういうようなところで研究自体の、今回の特にAI for Scienceのプロセス全体を俯瞰していただいているという形かと思います。
千葉委員、どうぞ。
【千葉委員】 千葉ですけれども、ありがとうございます。伺いたいのは、今日の前半のお話だと、やっぱり各分野ごとに共有のためのデータベースですとか、すいません、公開するデータベースをつくるということは結構やられているというお話だったんですけど、そうすると、やっぱり国全体の5ページの絵みたいな場合の共通基盤として用意する部分と、あと、それから分野別のデータベース、あるいはデータ公開基盤登録システムは、どういうふうにすみ分けていけばよいとお考えでしょうか。つまり、全部を基盤で、なるべく共通化、ソフトウエアも、特に、私、ソフトウエアは大変だと思っているんですけども、そこを共通化する方向で頑張るのか、それとも、そういう部分は、ハードウエアは共通のものを提供できるかもしれませんが、その上ものの、先ほどもあったミドルウエアから上は各分野で整備されていくべきなのか。その辺の見通しといいますか、どのようにお考えでしょうか。
【合田先生】 御質問ありがとうございます。こういう議論をすると、全部共通化して1つにしてしまえという意見もあるんですけれども、それは現実的ではなくて、やはり今日のお話にもありましたけれども、いろいろな研究分野によっては、本当長年の蓄積で、研究者の声も反映させた、すばらしいデータ基盤が出来上がっているわけですね。それは恐らくその分野の研究者にとって一番使いやすいものですので、それはきちっと生かして、それらをどう連携するかといことを考えるというのが重要

だろうと考えております。
そのときに、やはり今、千葉委員おっしゃったように、共通化できるとできないところというのはきちっと議論をした上で整理することが必要で、例えば今日も幾つか言及されておりましたけども、認証ですね。そういった部分は恐らく共通化できますし、すべきところでありまして、そこは共通化しつつ、分野に特化した機能と分野等でもいろいろつくっていくという姿がいいのではないかと考えております。
【千葉委員】私が思うに、お話聞いていて思ったんですげ、共通化できる部分というのはやっぱり限られてくるので、ある程度は分野ごとに整備していかなければいけないと思うんですが、一方で、今までは各分野ごとでばらばらに、予算の配分なんか、国家予算の配分なんかも各分野でばらばらにやられていたように私には見えるんですが、システムはばらばらだったとしても、やっぱりある種の共通化といいますか、コントロールと言うと言葉が悪いんですけれども、協調しながら運用していくという体制づくりは要るんじゃないかなとは思いましたという、最後は意見になってしまいますが、すいません、以上です。
【山本学術基盤整備室長】ありがとうございます。基盤室でございます。今日、分野の連携の話もいただきましたけれども、分野の状況も踏まえながら、国としてRDCと分野のデータ基盤をどう連携してやっていくかというのは非常に重要でございますので、海外のようなプラットフォームを参考にしながら、予算要求もそういうものを含めてパッケージ化して要求できるところはしたいということで、AI for Scienceを見据えて考えておりますので、そういったところでこのワーキングの報告書を参考に進めていきたいと思っております。ありがとうございます。
【尾上主査】ありがとうございます。お時間が、すいません、超過しておりますが、ここで事務局のほうから取りまとめ報告書の素案のポイントについて御紹介いただければと思います。よろしくお願いいたします。
【麻沼参事官補佐】事務局でございます。本日お時間もございませんので、次のスケジュールを紹介させていただきたいと思います。資料6の下のほうございますけれども、次回は第6回目、6月16日火曜日、16時から18時を予定しております。こちらで審議まとめ案について集中的に御審議をいただければと考えております。
また、取りまとめ後ですけれども、情報委員会のほうへも御報告をと考えておりますので、御承知おきいただければと思います。
2ページ目が審議まとめ素案の構成をお示ししておりまして、後ほど先生方にメールにて御意見を賜れればと考えておりますので、どうぞ御協力のほどよろしくお願いいたします。
本日は、説明のほうは割愛をさせていただければと思います。
以上でございます。
【尾上主査】ありがとうございます。構成案等出していただいていますが、何かどうしても言っておくべきことがあれば、もしありましたら挙手いただければと思いますが。後ほどメールでよろしいですかね。
ありがとうございます。委員の皆様、最終取りまとめに向けて引き続き御意見等賜れればと思いますので、どうぞよろしくお願い申し上げます。
最後に事務局から連絡事項等あればお願いいたします。
【麻沼参事官補佐】ありがとうございます。取りまとめに向けましては、次回まであまりお時間もございませんので、メールにて改めて御連絡をさせていただきますが、どうぞ御協力をよろしくお願いいたします。
以上でございます。
【尾上主査】ありがとうございます。それでは、本日の議題はここまでとなりますので、これにて閉会とさせていただきます。どうもありがとうございました。次回もよろしくお願いいたします。
── 了 ──
お問合せ先
研究振興局参事官（情報担当）付学術基盤整備室
（研究振興局参事官（情報担当）付学術基盤整備室）
ページの先頭に戻る
文部科学省ホームページトップへ