

文部科学省マテリアル先端リサーチインフラ
第3回データ連携基盤委員会/第2回データ構造化委員会 合同委員会
議事要旨

開催日時：2021年9月13日 13:00～15:00

開催方法：オンライン会議

出席者：(敬称略)

運営機構

小出 康夫 (機構長)

データ連携基盤委員/データ構造化委員

吉川 英樹 (委員長/NIMS)、松波 成行 (副委員長/NIMS)、戸津 健太郎 (東北大学)、
田浦 健次郎 (東京大学)、加藤 剛志 (名古屋大学)、小寺 秀俊 (京都大学)、
加藤 幸一郎 (九州大学)、石川 亮 (東京大学)、湯川 博 (名古屋大学)、土屋 智由 (京都大学)、
富井 和志 (京都大学)、藤ヶ谷 剛彦 (九州大学)

陪席：

伊藤 聡サブPD、永野 智己PO

文科省 小川 浩司参事官付参事官補佐、文部科学省研究振興局よりその他複数名
ハブ機関

今野 豊彦 (東北大学)、落合 幸徳 (東京大学)、森山 和彦 (東京大学)、
センターハブ機関

三留 正則 (NIMS)、内堀 千尋 (NIMS)、その他複数名

配布資料：

資料 デ基 3-1 ハブ機関訪問報告

資料 デ基 3-2 データ構造化 進捗状況

資料 デ基 3-3 IoT ネットワークの説明会の再開催 (案)

資料 デ基 3-4 メーカー対応進捗

資料 デ基 3-5 データ構造化 WG：勉強会開催について

資料 デ基 3-6 加工・プロセスのデータ構造化の作業提案

資料 デ基 3-7 クラウドデータ構造化システムの設計方針

議事：

(1) ハブ機関訪問報告

資料デ基 3-1 に基づき、松波副委員長よりハブ 5 機関を訪問した結果について報告があった。

(2) データ構造化進捗状況

資料デ基 3-2 に基づき、松波副委員長よりデータ構造化の進捗状況について報告があった。

主な議論は以下の通りであった。(敬称略)

(加藤剛)

名古屋大学の FIB-SEM について、メタデータが×になっていた。テキストファイルのデータを渡したと思うが、そこから検出はできなかったということか。

(松波)

SEM の撮影条件は書かれていたが、FIB の条件が書かれていなかった。

(加藤剛)

電子線露光装置について 15 台が「不可」となっているが、ほかの機関ではダメだったということか。

(松波)

東北大学にあるエリオニクス社の装置しか見ていないが、その運用を参照したところ、電子線露光装置については、データ構造化の実作業は難しいと判断した。

(加藤剛)

名古屋大学は日本電子のものなので、状況は異なるかもしれない。

(松波)

名古屋大学様とは個別に打ち合わせをさせていただき、状況をお聞きしながら、また相談させていただきたい。

(3) IoT ネットワーク説明会のリピート開催

資料デ基 3-3 に基づき、松波副委員長より IoT ネット説明会についての開催報告があった。

(4) メーカー対応進捗報告

資料デ基 3-4 に基づき、吉川委員長よりメーカー対応進捗報告について報告があった。

主な議論は以下の通りであった。

(小寺)

メーカーと交渉をする場合、マテリアルリサーチインフラで生まれてくるデータなどが、いかにメーカーにメリットをもたらすか、という点はどのように説明しているか。

(吉川)

装置メーカーのデータにメタデータが付いて普及される。そしてリサーチインフラを通じて流通し共用される。ユーザーにとって、より再利用の価値の高いデータになる、という説明をしている。

(小寺)

メーカー側のメリットとユーザー側のメリットがある。このように使ったらこのような問題があった、このようなところに時間がかかったというようなデータがメーカーに戻されることで、メーカーと研究者は共同研究をしているところが多く、このようなフィードバックはメーカーの競争力向上や、開発の指標が出てくるといったことをメリットとして強調するのがよい。メーカーへのアプローチを共有できるとよい。

(松波)

各機関における機器の利用方法（装置の設定情報など）がメーカーもシェアされることであれば、利用方法に応じたユーザー対応や装置開発につながることでから、その点について企業は関心を寄せている。また、構造化されたデータが最新の駆動型研究のアルゴリズムにどう使われていくか、どのような python スクリプトがユーザー側や研究者がデータを蓄積していく動機付けとなるかなど、自社では獲得できないマーケット情報についてメーカーは非常に強い興味を持っていると感じた。

(吉川)

Rigaku では、「ユーザーがどのような装置の使い方を知っているかを知ることがメーカーの役に立つ、ユーザーがどのようなメタデータの出力を求めているかを知ることが製品開発に役に立つ。」と考えられている。

(藤ヶ谷)

メーカー対応のページで装置名が記載されているものは、これらの装置が協力してもらえるものか、メーカーが装置を持っているということか。

(吉川)

個別会議の表に記載された装置は、その装置を対象とした協力をお願いしている。合同会議の表に記載された装置はメーカーが取り扱っている装置を挙げたものなので、このうちどれが実際に協力してもらえるかは、これから決まることになる。装置で分けている理由として、メーカーによっては、装置の担当事業部が異なり、事業部によってデータのフォーマットへの考え方が違う場合もあるためである。

(5) データ構造化 WG：勉強会開催について

資料デ基 3-5 に基づき、データ構造化 WG の勉強会開催について松波副委員長より提案があった。

主な議論は以下の通りであった。

(藤ヶ谷)

参加するにあたり必要なスキルはあるか。

(松波)

座学形式で行うが、実作業のワークフローを体得してもらうことが目的。まずはデータ構造化 WG のメンバーを対象として考えている。Python の初心者でも構わない。敷居を低くするために、調査を終えたサンプルコードなどを共有できる仕組みを整えたいと考えているが、理解が難しいということをフィードバックをいただきながら今後改善を進めたい。11 月上旬を目途に

開催を予定しているので、多くの参加をお願いしたい。

(6) 加工・プロセスのデータ構造化の作業提案

資料デ基 3-6 に基づき、加工・プロセスのデータ構造化の作業について松波副委員長より提案があった。

主な議論は以下の通りであった。

(小寺)

横断関係のコーディネーターが東京大学と京都大学にいたので協力できる。

(吉川)

スケジュールはどのようになっているか

(松波)

加工 WG の委員選任から始め、10 月中旬をめどに 1 回目の打ち合わせを行いたい。

(7) クラウドデータ構造化システムの構築概況

資料デ基 3-7 に基づき、クラウドデータ構造化システムの構築概況について吉川委員長より説明があった。

主な議論は以下の通りであった。

(田浦)

まだ公開されていないデータを使って派生するデータを利用する権利は誰にあるのか。もともとのデータを利用できる人の範囲だと限定されすぎるのではないか。

(吉川)

非共有の段階ではデータを取った人の権利が 100%。共有になった段階では、データを取った権利に加えて利用者の権利も発生していると考えないと、共有が進まない。データ構造化という処理を行っている点は事業のユニークさ。構造化された後に利用価値が高まったものは、共有するために権利関係を定めたい。

(田浦)

個々のデータの貢献は薄まっているような、AI のモデルのようなものは事業内で広く使えるようにしないと利用促進が進まない。

(吉川)

利用促進のためにデータを登録した人を引用するなどのリスペクトする作業は必要。

(戸津)

完全公開のところは利用報告書も相当するのではないか。

(吉川)

利用報告書もデータをまとめたものとして共有公開の対象と考える。利用報告書とデータセットのエンバーゴの設定方針は変えたほうが良いのではないかと議論している。ユーザーによっては利用報告書の内容が薄く、すぐに公開できるものもあるため。

(加藤剛)

非共有に上げるデータは事業内共有に一定期間後は移行する前提で上げるということか。ある

いは移行に同意しない場合は非共有にも上げないのか。

(吉川)

未来永劫非共有というものは本事業のミッションから外れるので、事業内共有を認めた人にデータを登録いただく。データによっては特許の関係でエンバーゴに達しても共有できないものもあるが、エンバーゴの期間の延長で調整する。

(永野)

まず非共有、事業内共有、完全公開の境界をどのように線引きするか。議論によっては多様に考え得る話。非共有の場合でも別の研究チームと共有する可能性もあれば、事業内共有の場合においても事業内の切り方も様々ある。事業内と言ったときに事業に利用者として登録した人数、例えばナノプラでは万の単位となるし、また事業の提供サイド、参画者として登録している人は 1000 人いる。1000 という単位、万という単位で共有する、あるいはその中の線引きとして機関で共有する、横断技術で線引きもできる。この事業として、どこにどの線引きをするのが良いか、NIMS のシステムとしての作り方なども考慮し検討していただきたい。何が良いのかは現時点ではまだわかっていない。本事業が目指す趣旨として、データを利活用した研究開発を促進するやり方としてどこに線を引けばよいのか、よく議論していただきたい。また本事業では「利用内容を公開する」と定義している。利用内容とは旧ナノプラでいうと利用報告書となっている。旧ナノプラでいう利用報告書に相当するものの公開とデータの公開と二つある。利用報告書の在り方は見直しを求めており、他の委員会で議論している。データの在り方の公開はここで議論することになる。利用内容の公開を議論する委員会との連携をお願いする。

(吉川)

他の委員会との連携を今後行っていく。事業内共有の考え方は重要であると認識。非共有のところで「他チームとの共有」を入れた理由は、事業内共有をシンプルかつ有効にしたいと考えたため。共同研究レベルで行うような個別案件は非共有のフレームワークで行っていただく。事業内共有における粒度について引き続き議論をお願いしたい。

(永野)

事業内共有のところで、利用者間での共有が理解できない人がいるのではないかと。利用者は単に利用する人で、完全公開に見える。事業内とは何か、をよく議論していただきたい。

(小寺)

非公開にしている期間は論文特許に影響する。事業内で共有すると公開とみなされるので、その点も考慮した設計が必要。論文公開後は積極的に公開したい一方で、カギになるところは秘匿したいというユーザーもいる。

(吉川)

論文と特許は研究者として重要なので、それが出るまで共有が難しいことは理解できる。今はエンバーゴを設定し、その期間内に論文と特許を出していただき、出た後に初めて共有できるという考え方。エンバーゴを延長すれば論文や特許が出て共有できるデータと論文や特許になるデータとはならないけれど共有したいデータの切り分けは必要。一つの実験の中でも出せるデータと出せないデータはあるだろうが、システムの対応が難しいので、データセット単位でオープン・シェアをコントロールする。データの中で公開していいものと論文や特許がでるま

で秘匿したい場合はそれぞれのデータセットにする。

(今野)

まず、この事業のもう一つの骨子はナノプラ時代からの装置共用。他の委員会との整合性をとる意味から、装置共用だけのユーザーがいるという認識を持っている必要がある。この委員会では非共有に相当するかもしれないが、事業全体では装置を共用している前提があり、利用報告書も作成するし、利用料金に差をつける話も出ている。このような点に積極的なとらえ方をする必要がある。次にナノプラ事業をやってきて、企業の利用はトラブル対応やクレーム処理、試作目的などがあり現実的には共有はむずかしい。実施機関では装置の使い方やクリーンルームの使い方など必然にデータは共有している。実施機関とユーザーの間でデータを共有しているので、事業内共有は概念的、抽象的すぎる。最後に、データをどのように利用するか、どのように機械学習に落とし込むかという議論がない。データセットが機械学習に落とし込めるかわからない。データセットを準備する考え方に加え、どのように利用してもらうかを見据えながら、どのようなデータセットを出していくかという論点が必要と考える。

(加藤幸)

九州大学でもデータサーバーを用意しようとしているが、利用者について議論している。装置だけ、装置とデータ、に加え、データのみ使いたい人が出てくる。データだけの人に利用料を求める場合に、事業内共有や完全公開などの区別が現れるのかと感じた。データだけ利用したい人をどのように扱うか、がポイントと考える。データセットでコントロールするという話であったが、その粒度によって使う側の使勝手も変わるので、ユーザーを意識した設計の議論も必要。

(吉川)

データ利用だけのユーザーも大事なので、システムでは想定している。データセットに関しては、ビューアを用意してデータが一目でわかるような機能もある。またデータカタログも準備している。

(田浦)

事業内共有にスペクトルがあるのは認識の通りで、利用者を含むのか事業実施者のみなのか、データを提供した人と事業実施側の区別はなるべくないほうが良いと考える。エンバーゴについて、一定期間が経れば共有してほしいという事業の趣旨はあるが、プラットフォームにデータを提供する人を増やすというゴールからの検討も必要。一定期間を経ると共有が強制されるという状態を嫌がる人は、非共有の状態にすらならないという感じ方をするかもしれない。

(落合)

まず事業のメンバーというのはスペクトルが広く、ナノプラの業務実施者はエンバーゴ前にデータが見られないということになるか。実務的に、データを提供するといわれたものが有益なものであるかどうか、誰もチェックできなくなる。守秘義務を課して業務実施者はエンバーゴ前にデータが見られる必要があるのではないか。事業内共有の図で事業内メンバーへ共有となっているが、これは実質的に公開と同じではないか。ID やパスワードで制御してもその先は制御できない。次に集めたデータセットが有用かどうか、を懸念している。マテリアル DX プラットフォームでは 3 つの組織に加えて「データ創出活用型マテリアル研究開発プロジェクト」

が動いており、連携することになっている。このプロジェクトはマテリアルインフォマティクスを駆使して材料データを開発するグループであるので、どのようなデータが集まれば有益か、といった意見を聞くことはしたか。

（吉川）

事業内共有は事業の全利用者で共有するので限りなく公開に近く、場合によっては万人の単位で共有する場合もある。事実上公開では困る場合は、非共有としてメンバー内の共有を想定。できるだけ多くの人と共有する場合と、閉じた場合を両立するためにエンバゴを導入した。他のプロジェクトとの議論はできていないが今後検討する。

（松波）

論文等での調査は行っているが、他のプロジェクトとは打ち合わせの機会を持ちたい。

（今野）

東北大学でもフィージビリティスタディに参加しており草の根的に議論を進めている。他のプロジェクトと議論しようとするベクトルは重要と認識。

（小川）

フィージビリティスタディ内でも連携について同じような議論はある。今の議論はリサーチインフラ側から利用者のニーズをヒアリングするもの。一方でフィージビリティスタディの中でもリサーチインフラを将来どのように使えるか、どう連携するか議論は始まっている。現在は個々に話をしているところもあり、リサーチインフラで検討しているデータの共有について方向性をフィージビリティスタディに採択された5拠点の関係者と共有し、逆にどのようなデータが欲しいかといったことをフィージビリティスタディ側から提供することがあってもよい。打ち合わせの機会については調整したい。

最後に吉川委員長より、本委員会をほぼ3カ月に1回開催したい旨、また本委員会での議論の内容はハブ機関からスポーク機関との共有をしてほしい旨のアナウンスがあった。

以上

（議事要旨作成：内堀）