第 1 回データ連携基盤委員会議事要旨

開催日時：2021 年 5 月 12 日 17:00～18：40
開催方法：オンライン会議
出席者：

運営機構

小出 康夫（機構長)、出村 雅彦（NIMS）

データ連携基盤委員

吉川 英樹（NIMS)、戸津 健太郎（東北大学)、田浦 健次朗（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）

陪席：

文科省　曽根 純一ＰＤ、伊藤 聡サブ PD、永野 智己ＰＯ、田中 竜太ＰＯ
江頭 基参事官、小川 浩司参事官付参事官補佐、その他複数名

ハブ機関

今野 豊彦（東北大学)、幾原 雄一（東京大学)、土屋 智由（京都大学)、
村上 恭和（九州大学）

センターハブ機関

花方 信孝（NIMS)、三留 正則（NIMS)、松波 成行（NIMS)、内堀 千尋（NIMS）

配布資料：

【議事次第】第 1 回データ連携基盤委員会
【資料 1】データ連携基盤責任者リスト
【資料 2】データ構造化の進め方プレゼン資料
【資料 3】今後の予定＆連絡事項

議事：

(1) 挨拶及び趣旨説明（以下、敬称略）

運営機構長から趣旨説明が行われ、委員長は運営機構長が兼務することが提案され、認められた。

(2) 委員自己紹介

資料 1 に沿ってデータ連携基盤委員の自己紹介が行われた。

(3) データ構造化の進め方

資料 2 に沿って、NIMS 出村よりデータ構造化の進め方について発表があった。

(4) 討論・課題抽出

データ構造化の進め方について主な議論は以下の通りであった。

小寺委員：加工には一つのプロセスで 16 機関が 10~15 種類の装置を使うため横串を考える必要がある。このためハブで 2 機種を選ぶのは難しいのではないか。また作業人員と作業量に見合った予算をもらっていないのではないか。（データ構造化 WG の活動は）有志でできるような内容ではない。

→ 指摘いただいた問題は認識している。データを後から入力して連結させるのが一つのアイディア（小出）

→ 試料に ID を付与して複数の機関でとったデータを横串で見られるようにする。ハブが 2 機種選ぶのではなく、全体の中でどれから取り掛かると効果が得られるかという視点で運営機構が選定する。来年以降は数を増やしていく。（出村）

小寺委員：タスクフォースや WG において、それぞれのハブ＆スポークで対象となっている領域のデバイスや材料をきちんと決めて、それぞれの領域でどのように進めるかを議論したほうが効率的ではないか。分野会議を構成してもらいたい。

田浦委員：NIMS ですでに行った 18 機種のサンプルデータやコードは共有できるか。これまで NIMS で取り組んできた人はどのような専門性やポジションの方であるか。

→ これまで作ったものは即時提供可能。NIMS の定年制エンジニアが取り組んできた。エンジニアが装置の管理者や利用者とひざ詰めでどのようなデータを取ってるかどのような自動化が求められているかを聞き、装置メーカーとの交渉も別途行い、どの結果をデータ項目として抽出すればよいかを整理する。材料研究の経験や分析の経験があり、プログラミングもできる人材。任期制職員でプログラミングができる人材も雇用して進めている。（出村）

田浦委員：NIMS で取り組んできた人自身が研究者という訳ではない。

→ その通り。（出村）

戸津委員：PLC などによりローカルで動く装置も多いが、それに PC をつないで Python で抽出するイメージでよいか。

→　Python のプログラムを実際に駆動するのはデータ構造化システムが走るクラウド（クラウド環境はデータ中核拠点で用意）の上。ローカルな環境でプログラムは走らない。プログラムの動作を確認するためにローカルな PC で動かすことはありうる。（出村）

戸津委員：PC をつないでいない装置に PC をつないでデータを抽出する場合の PC はどちらが用意するのか。

→　ネットにつなぐところまではスポークやハブで用意。IoT のデバイスについてはセンターハブで一括購入。（出村）

戸津委員：ハブには今年度 2 個。スポークには来年度以降となるか。

→　データ構造化の予算が付くのはハブまでのため、コーディングの作業はハブまででではないかと理解。IoT を実際の装置につけるのはスポークを含めた装置管理者。（出村）

→　NIMS の経験でもコーディングは慣れてくるとペースが上がってくる。スタートは 2 機種でその後はスピードアップを期待している。クラウドのシステムは 22 年度立上げおよび性能検証から 23 年度試験運用、本格運用はまだ先。（小出）

→　年間 6 件~10 件くらいまでペースが上がる。同じ装置を使っていれば同じデータ構造化スクリプトが使用できるため対応可能な装置数はその分だけ増える。（出村）

加藤（幸）委員：NIMS で構築する全体システムにハブで作るスクリプトが集約され、実際の運用では構造化システムを作るのに必要な生データをハブやスポークがアップロードするイメージで良いか。

→　その通り。（出村）

加藤（幸）委員：各ハブでも DB を使いたい場合もあるかと思うが、ローカルでスクリプトを使ったり、他のハブのスクリプトを使ったりすることはできるか。出来上がったスクリプトの共有範囲は。

→　NIMS でもメーカーの承認が得られているものは公開している。オールジャパンでスクリプトを使えることは重要。囲い込みはしない。（出村）

田浦委員：先端リサーチインフラで提供している実験装置の上限は。

→　ナノプラットフォームでは 1000 台程度が登録されている。NIMS では IoT デバイスを 500 個準備。どのようにどこまで増やすかは今後委員会や WG などで議論する。（小出）

田浦委員：データのフォーマットは何種類あれば全部カバーできるか。

→　装置の分析をセンターハブで進めている。1000 台のうち研磨装置などを除き、さらに、重複を省くと 200 種類程度と見込んでいる。初年度 12 種類作る。また、NIMS はすでに 18 種類作成済みで、そのうちいくつかは転用可能。3、4 年でカバーするイメージ。200 台全部カバーするのか、費用対効果で数が決まるのかもしれない。（出村）

加藤（剛）委員：ハブ機関でデータを上げる 2 機種はセンター機関から指定されるのか。
→　これから調査し、構造化委員会の中で検討して決める。共通性のあるものなどを委員会で議論して決める。（小出）

加藤（剛）委員：こちらから提案することと理解。NIMS で行った 18 機種以外に 2 機種を選ぶのか。
→　NIMS には計測装置が主にあり、25 法人の中で有効性の高いものなど、広く考える。検討委員会で全体をみて決めていきたい。（小出）

加藤（剛）委員：18 機種についてはすでにスクリプトができているという理解でよいか。
→　装置からの自動翻訳に関しては、18 機種はできている。（出村）

加藤（剛）委員：18 機種以外の 2 機種をハブ機関が提案して、承認されれば作業を開始するということか
→　いくつ提案するのか等、装置リストに関する調査については運営機構で調整と理解している。（出村）

加藤（剛）委員：装置リストの提案を 6 月末までに行うのか
→　構想が合意されればタイムスケジュールを決める。（小出）

加藤（幸）委員：自動抽出は装置に関するメタデータ、テンプレートは試料に関するメタデータを抽出するためのツールという理解で良いか。
→　大きくはその理解でよい。測定の条件に関する情報が装置から取れない場合もある。それをテンプレートからとる場合もある。（出村）

加藤（幸）委員：データ構造化システム自体が裏に装置ごとのデータベースをもっていて、試料 ID で横串しに構造化されたデータベースが組みあがるのか。一般ユーザに公開されるのは構造化されたデータベースなのか。
→　ユーザは自身がタッチできるデータにアクセスできて、構造化されたデータについて検索したり、試料 ID で検索すると構造化済みのデータが自分のデータの場合だとダウンロードできる予定。装置については、装置マスタをシステムに持っている。IoT でつないだ場合にユーザがどのスクリプトを使って変換したらよいか選ぶのが大変なので、IoT でつないだ段階で装置が特定され、その装置で使用できるスクリプトを選べるように設計する予定。（出村）

加藤（幸）委員：ユーザが登録データで設定や条件を思い出したい場合、構造化されたデータに情報は入っているか
→　装置から吐き出されるファイルに書いてある場合は自動で翻訳されて、可読できる形で記録されることになる。NIMS ですでにいくつかの装置で開始しており、ユーザからの

評判が良い

小寺委員：データ構造の設計では、試料に関するメタデータでは、化学組成、結晶構造などが縦につながっている。評価・計測に関するメタデータでは評価・計測条件といったものが縦につながっている。関係するハブで、同じような悩みや関連するところを議論して、実際に出てくるデータをどのように使うかというシーンも考えながら、自動化と手作業の共通認識を作ったほうが作業が早いのではないか。
→　合成、加工のところでは議論になるかと思うが、仕訳を行う点について了解した。(小出)

(5) 各ハブ・スポークデータ構想との連携と課題抽出
主な議論は以下の通りであった。(敬称略)

幾原（東大)：これから情報基盤センターと各加工・計測を連携して進めていく。この構想でNIMS とスタートすることをこれから周知していく。

加藤（剛）委員：名古屋大学としては、一次データ、二次データ、実験ノートが共有化できて後で参照できるのが一番良いと考える。マテリアルリサーチインフラでは一次データを取りだすことを主点においてやっていくのだろうと理解。我々の構想とマッチしていると考える。実験ノート、二次データをまとめたものが本来の役に立つ構造化されたデータだと思う。そこまで到達できたらよい。実験の細かな条件など、技術職員に聞かないとわからないようなもの、その人がいなくなれば消えてしまうレシピがデータ化されていて、誰もが専門家のようにデータを使えるのが理想。

小寺委員：マルチマテリアルのデータベースは材料データ、プロセス、加工、界面特性、界面構造を進める中で様々な計測データがでてくる。その途中のプロセスで計測データがでてくる。そのデータをどのように構造化するか。タグをつけていくのだと考えている。高分子でも同様。制御 PC、測定 PC と手入力のデータをアップロードする。最初から全部はできないので、特徴的なプロセスを似た機関で共通認識をもちながら分担して、最初に議論すれば早い。微細加工ではデータや計測データはユーザにも入力を促してきた。このような経験があるので、ブレーンストーミングをすれば作れると思う。ハブとスポークを縦にデバイスや材料で並べ、横に構造解析、微細加工、物質合成で並べると全部網羅できていないことがわかる。不足するところはハブやスポークで補わないといけない。

加藤（幸）委員：装置ごとに標準的な使い方や付与可能なメタデータを定めていて、あるフォーマットで装置ごとにデータが出てくることを考えていた。標準化が難しいメタデータについては情報のテキスト化にかかわるガイドラインを作成して、将来は機械に読ませることも視野にガイドラインを作成して情報を残そうと考えている。資料 ID を使ってプロセス、構造特性を紐づけすることを考えていた。従って親和性は高いと考える。構造化したものをセンターハブに上げるのかと思っていたが、生データを上げるということが分かったので、検討していく。

戸津委員：ユーザにどのように利活用できるかを明らかにしたうえで協力をお願いしたい。試作回数をへらすなど具体的なメリットをアピールできればよい。企業データをいかに集めるかはポイント。どこまでシェアクローズドなデータとして扱えるのか、具体的に示すのもポイント。データ適用できない事例、成果の公開など議論が必要。持ち込みデータはクローズドになるので、切り分けなければならない。クローズドデータに触れずにシェアクローズドにもっていけることもあるのではないか。ID を振るのはありだけれど、匿名化も検討してはいかがか。

→　このような形でデータを収集するという方法を紹介した。これで一歩を踏み出したい。実際に行わなければならない作業とタイムスケジュールは別途進めていく。（小出）

田浦委員：アメリカでの MGI などでも、装置データの収集は取り組んでいるのか。世界動向を知りたい。

→　MGI のデータ収集の一つの柱は第一原理計算。ヨーロッパや日本でもやっている。装置から直接集めるという取り組みはユニーク。実験データを入力する取り組みはあるがデータ量も少なく、フォーマットも統制されていないため、横断的な使用ができない状況と理解。（出村）

→　実験データではＮＲＥＬ（National Renewable Energy Laboratory）が薄膜に関するデータを網羅的に集めて公開することが進んでいる。MGI ではマテリアルデータキュレーションというサービスをすることになっている。仕組みはあるがデータは集まっていない。NIST が中心に HTEMC（High- Throughput Experimental Materials Collaboratory）はデータを集めるというより、データを流通させるしくみ。我々の取り組みは他ではないという認識。（伊藤 聡）

今野（東北大）：１９８０年代に半導体デバイスではシミュレーションではかなりのことができていた。データの収集システムは素晴らしいものができているが、目的がなければならない。ターゲットを絞ってデータを収集しないといけない。典型的なプロセスなど、ターゲットがあってデータを集めたら義侠の人間も利活用できる。メタデータとして定義されているものが重要。それを定義するところに材料をわかっている人がかかわる必要がある。装置のスマートラボに陥ることを危惧。出口を見てデータを収集する方向になればよいと考える。

→　利活用を考え、ご指摘の議論は進めていく。（小出）

(6) その他

運営機構長から資料 3 に沿って今後の予定が説明された。

以上

（議事録作成：内堀）