第 1 回 SPARC Japan セミナー2020
「研究データ公開:フルオープンと制限公開の境界線」

農研機構統合 DB の構築と データ共有の取り組みについて

桂樹 哲雄
（農業・食品産業技術総合研究機構（農研機構））

講演要旨

近年、農業分野における研究環境の ICT 化にともなって電子データが急増している。多大な労力をもって生み出された研究データは、それ自体が貴重な財産であることから、これらを適切に収集・管理し、利活用することが研究活動を加速化し、従来の分野・領域を超えた学際的な研究を創出できると期待される。そこで、農研機構では研究データの適切な保存・管理・整理を目的としたデータ基盤として、「農研機構統合 DB」を構築し、2020 年度から試験運用を開始した。データ再利用の観点から、機構内のデータは基本的に機構職員が自由に参照できる仕組みを構築した一方、ライセンスや機密保持の観点から、参照制限をかけるものも存在する。
本講演では、農研機構統合 DB の概要を示したのち、オープン・クローズド戦略に基づいて実際にどのようなルールと仕組みでデータの共有を行っているかを紹介する。まだ手探りのところもあるため、本講演を通じて皆様のご意見を伺いたい。

桂樹 哲雄

2002年大阪府立大学工学部海洋システム工学科卒業、2005年同大学院工学研究科機械系専攻博士前期課程修了、2011年同航空宇宙海洋系博士後期課程単位取得退学、2014年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。日本学術振興会特別研究員、豊橋技術科学大学情報・知能工学系助教を経て、農業・食品産業技術総合研究機構農業情報研究センター データ戦略推進室主任研究員に着任。流体計算の分野で研究キャリアをスタートし、バイオインフォマティクスの分野で植物代謝の計算手法を開発、ケモインフォマティクスの分野で化合物の構造活性相関の研究に従事した。現在は農業データ利用促進のためのデータベースに関する研究・開発を行う。

農研機構統合 DB　開発経緯

私の所属するデータ戦略推進室は 2019 年 4 月に設置されました。そこから農研機構統合 DB を作ろうということで、2020 年 3 月に DB 運用ガイドラインを整備し、6 月から機構内限定で試験公開を始めました。農研機構統合 DB のデータを外部ユーザーと一部共有開始したのが今年 9 月です。2021 年 4 月から本格運用を開始予定です。

開発の背景としては、データがたくさんたまってきたことと、経年によってデータがどんどん散逸しているということがありました。職員が退職するとデータがどこにあるか誰も知らないということがよく起きてきたので、分野横断的かつ統一的なデータ基盤の構築を急務として、そのデータ基盤として農研機構統合 DB を構築しました（図 1）。コンセプトは「人をつなぐ、データをつなぐ」で、農研機構全体での研究データの共有・活用を目指し、研究データを通じて人をつないだり、また、われわれはグラフデータベースを導入していますので、それでデータ自体をつないだりしています。

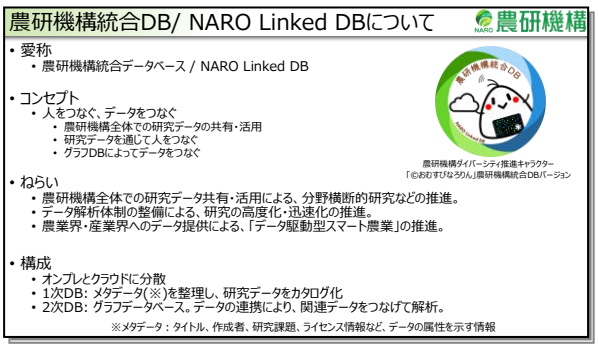
DB の構成としては、オンプレミスとクラウドに分散しており、DB を 1 次 DB と 2 次 DB に分類してい

ます（図 2)。1 次 DB は、メタデータを整理して研究データをカタログとして塊で置いておくものです。2 次 DB はグラフデータベースで、データを連携して関連データをつなげて解析できるようにするものです。1 次 DB にためたものから選抜して 2 次 DB に持っていくという構成にしています。農研機構内に 1 次 DB、クラウド環境に 2 次 DB を置いて、両方に機構内からも機構外部からも適切にアクセスできるように、ファイアウォールや認証システムなどを導入してアクセス制御をしています。

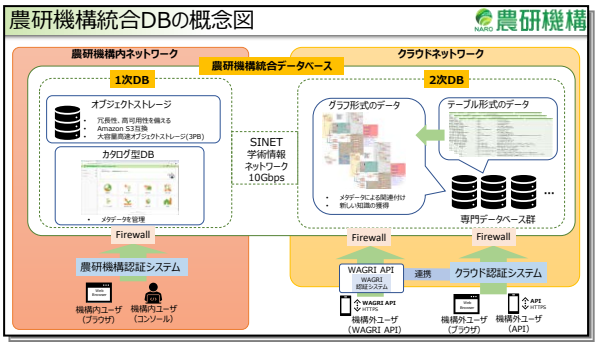
1 次 DB（カタログ型 DB）

1 次 DB は、カタログ型のメタデータを入れて管理するメタデータ DB と、そのデータ本体を置くオブジェクトストレージの二つで構成されています。データにメタデータを付けてカタログ型 DB で管理し、その実体を Amazon S3 互換オブジェクトストレージに格納するという設計になっています。

特長としては、機構内のデータを集約して安全に保管することで、研究者のデータ保存の負担を軽減し、



（図 1）



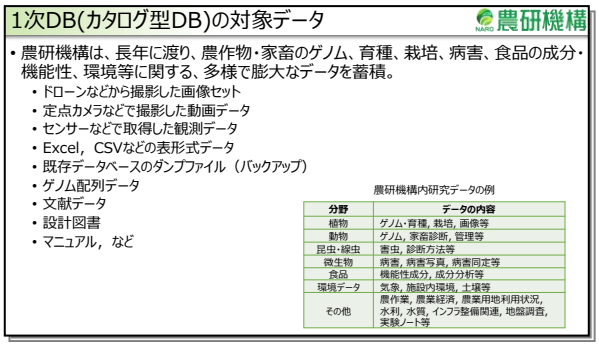
（図 2）

自分たちでハードディスクを買わずに済むということと、メタデータを整備することで機構内の有用な研究データの存在を可視化できるということがあります。メタ情報を基に検索可能で、他の人がどのような研究をしているかといったことから新たなテーマを発見したり、展開したりできます。掲示板機能によって研究者間のコミュニケーションを促進することも目指しています。また、柔軟なアクセス制御によってデータを安全に共有できます。

それから、完全に外部に出してしまうデータに関しては、別システムを立ち上げました。もし公開用のシステムに何らかの不具合で外部から不正に入られてしまっても、機密情報は出ないようにするという仕組みを持たせています。

この 1 次 DB には、農研機構に集まってくるあらゆるデータが入ります（図 3)。ドローンから撮影した画像セットや、定点カメラあるいはセンサーなどで取得した観測データなどがリアルタイムで入ってくることも想定して設計しています。もちろん Excel データやゲノム配列データ、それから文献データや設計図書も置いてもらおうと考えています。

データに付けるメタデータの項目も、昨年（2019 年)、一生懸命考えました。もちろんデータをただ置いておくだけでは他の人にはどのようなデータか分からないし、どんなライセンスかということも分からないので、そういう情報を整理しようということになりました。ただ、あまりたくさん付け過ぎると、実際に入力する人に、入力したくない、面倒くさいと言われるので、どれぐらいのデータであればみんな入れてく



（図 3）

れるか、そして検索にきちんと引っ掛かるかということを考えました。

設計方法としては、Dublin Core をベースにして、農研機構で必要と思われる項目を追加し、入力は煩雑にならないように語彙統制をしつつ、しかし柔軟に、項目は欲張らずというということでメタデータ項目を考えています。結果的には Dublin Core や JPCOAR と似た項目となりました。このメタデータは、これだけでは足りないということで、今後カスタムメタデータを付けられるように設計しようと、データベースの改修を行っているところです。

2 次 DB

2 次 DB については少しだけ説明します。グラフデータベースやテーブルデータベースとして Oracle のクラウドのデータベースを使いましょうということでデータを入れていっています。Linked Open Data などと連携できます（図 4）。

2019 年に私どもの部署ができてからデータを集め出し、今、その入れ物を農研機構統合 DB として構築したところで、これからどんどんデータが入ってくることを期待しているのですが、データの種類としてはあらゆるデータがあります。紙データ、リアルタイム計測データ、マルチメディアデータ、実験データなどがあるのですが、結局のところ、カテゴライズするとしたら、電子ファイルといってもテキストデータが入っているものと入っていないものに分かれます。入っていないものは 1 次 DB にメタデータを付けて放り込み、テキストデータがあるものに関しては全文検索ができるようにしようと考えています。また、テーブルデータやグラフデータは、2 次 DB で専門のテーブルデータなどを扱えるデータベースにしていますので、そこに入れていこうと考えています。

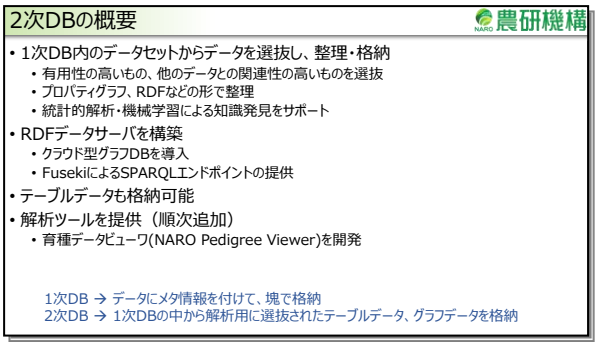
農研機構統合 DB のオープン・クローズド戦略

これは皆さんのご意見も頂けたらと思っているのですが、農研機構でも、オープンデータとシェアードデータ・クローズドデータを分けて考える必要があるという議論をしてきています（図 5）。

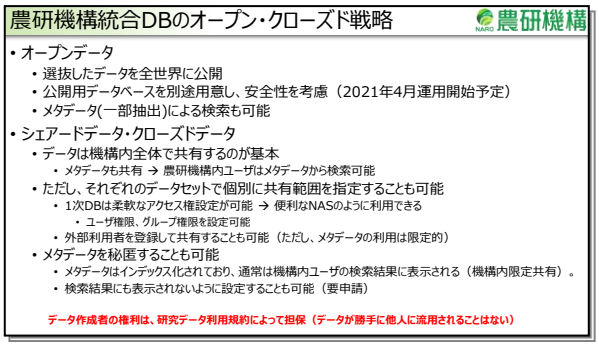
オープンデータに関しては、選抜したデータを公開することと、公開用 DB を別途用意して安全性を考慮することを考えています。メタデータに関しては一部だけ検索できるようにしようとしています。

シェアードデータ・クローズドデータに関しては、農研機構は国の機関ですし、機構のお金で作ったデータは機構全体で共有することを基本に考えていこうということになっています。これまで農研機構ではあまりそういう考え方が一般的ではなかったのですが、もうそういう時代ではないので、どんどん共有していきましょうということで動いています。メタデータも共有して、農研機構内のユーザーであればメタデータからデータを検索できるようにしようと考えています。

ただし、もちろん農研機構のデータを何でも共有していいわけではありませんので、それぞれのデータセットで個別に共有範囲を指定することもできるようにしています。1 次 DB は柔軟なアクセス権を設定して、ユーザー権限とグループ権限を設定することが可能ですので、便利な NAS のように利用できるということ



（図 4）



（図 5）

を宣伝しています。また、外部利用者（県の試験場の方、大学の先生など）を登録して共有することも可能です。ただし、その場合にはメタデータの利用を限定的にしようとしています。というのも、メタデータで検索できてしまうので、データが見えてしまうのです。メタデータが見えてしまうのはまずいということで利用を限定しています。

それに関連して、メタデータ自体を機構内のユーザーに対しても秘匿することを可能としています。というのは、メタデータはインデックス化されていますので検索でき、検索結果に一部のデータが表示されてしまいます。そこにデータが存在することが分かってしまうのも困るということがありますので、検索結果にも表示されないように設定することが可能です。これは申請ベースで受け付けています。

先ほど、農研機構の文化のようなことを少し言いましたが、やはりデータ作成者には、これからはデータを機関内で共有するのだといっても、勝手に使われてしまうのではないかと心配される方が非常に多いです。従って、データ作成者の権利は研究データ利用規約で担保されているのでデータが勝手に他人に利用されることはないということを強調しています。

公開用メタデータとしては、一部だけ公開するということで項目を選択しています。例えば、機構内の ID などは機構外の方には関係ないものなので出しません。問い合わせ先（連絡窓口）に関しては、データセットの責任者に直接連絡できるようにしようと考えていますが、これについてはまだ少し変わるかもしれません。今、内部で相談しているところです。

利用促進のための 3 要素

データベースは今やっと試験運用が始まったところなので、これからどんどん皆さんに使ってもらいたいと思い、利用促進のための 3 要素を考えました。まず、運用ガイドラインを策定しました。それから、データ利用規約も策定しました。また、AI スパコンを構築しましたので、それと併せて DB に関しても教育を行って広報活動をして、皆さんに使ってもらおうと考えています。

まず、運用ガイドラインでは、研究データの定義を、研究の過程でできたデータの全部としています。研究で発生したデータは機構全体にとって貴重な研究資産なので、個人やチームだけで保存するのではなく、機構が定めた安全な場所に永続的にバックアップし、分野横断的に使っていきましょうということを定めています。DB の運用開始後は、研究成果となるデータは原則機構内での再利用を認めるということを明記しています。ただし、秘匿性が高い場合など、別途認められた場合は利用範囲を限定できます。

そしてデータ利用規約では、「データを利用した研究について、学会発表や論文投稿等の外部発表を行う場合や特許提案等を行おうとする場合は、事前にデータ提供者に連絡し、著者、引用方法、謝辞、出願等について協議すること」「データ提供者は、提供したデータによってデータ利用者が被ったいかなる損害にも責任を負わないこと」というように、データを出した人が全ての責任を負うわけではないことを明記しているのがポイントです。

クローズドデータへのアクセス制御

統合 DB 全体をファイアウォールや Web アプリケーションファイアウォールできちんと保護しています。あるいは、IdP でアクセス管理しているのですが、それでユーザーごとやグループごとに権限を制御し、トークンを用いて認証・認可ができるようにしています。

実際の手順

実際に DB にデータが入ってきたときに、われわれが農研機構内でどうしているかということを紹介します。農研機構の研究は、大課題、中課題、小課題に分類されています。その中課題ごと、あるいは部署を横断するプロジェクトがある場合にはその責任者をオーナーとするフォルダを作ります。責任者がそのフォルダ内のデータに責任を持って必要なメンバーをユーザ

ーID やグループ単位で登録し、それぞれに権限を付与する形でデータを管理しています。

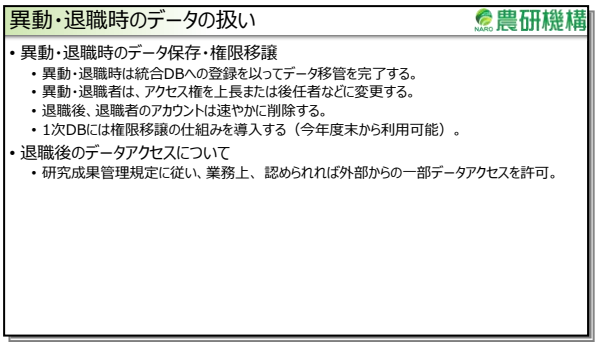
それから、メタデータのアクセス制御ももちろん行っていますし、先ほどから言っているように、秘匿したい場合には秘匿することも可能です。

異動・退職時のデータ保存・権限移譲

職員の異動・退職時は、データがどこかへ消えてしまうタイミングでもあるので、統合 DB への登録をもってデータ移管を完了することにしています（図 6)。異動・退職者はアクセス権を上長または後任者などに変更し、変更後は退職者のアカウントは速やかに削除するという運用方法を採ろうとしています。そうなると、退職者が持っている全てのデータのアクセス権を手で変更するというのは非常に大変ですので、その移動の仕組みを現在開発中です。退職後については、研究成果管理規定に従って、業務上、認められれば外部からのデータアクセスを一部許可することにしています。

まとめ

以上、農研機構統合 DB と、そのオープン・クローズド戦略の実際について紹介しました。取得されたさまざまなデータをより高度に利活用できるよう、今後も農研機構統合 DB を改良していきたいと思っています。



（図 6）