農研機構

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」
2020年10月02日

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

国立研究開発法人 農業・食品産業技術総合研究機構（農研機構）
農業情報研究センター データ戦略推進室
主任研究員 桂樹哲雄

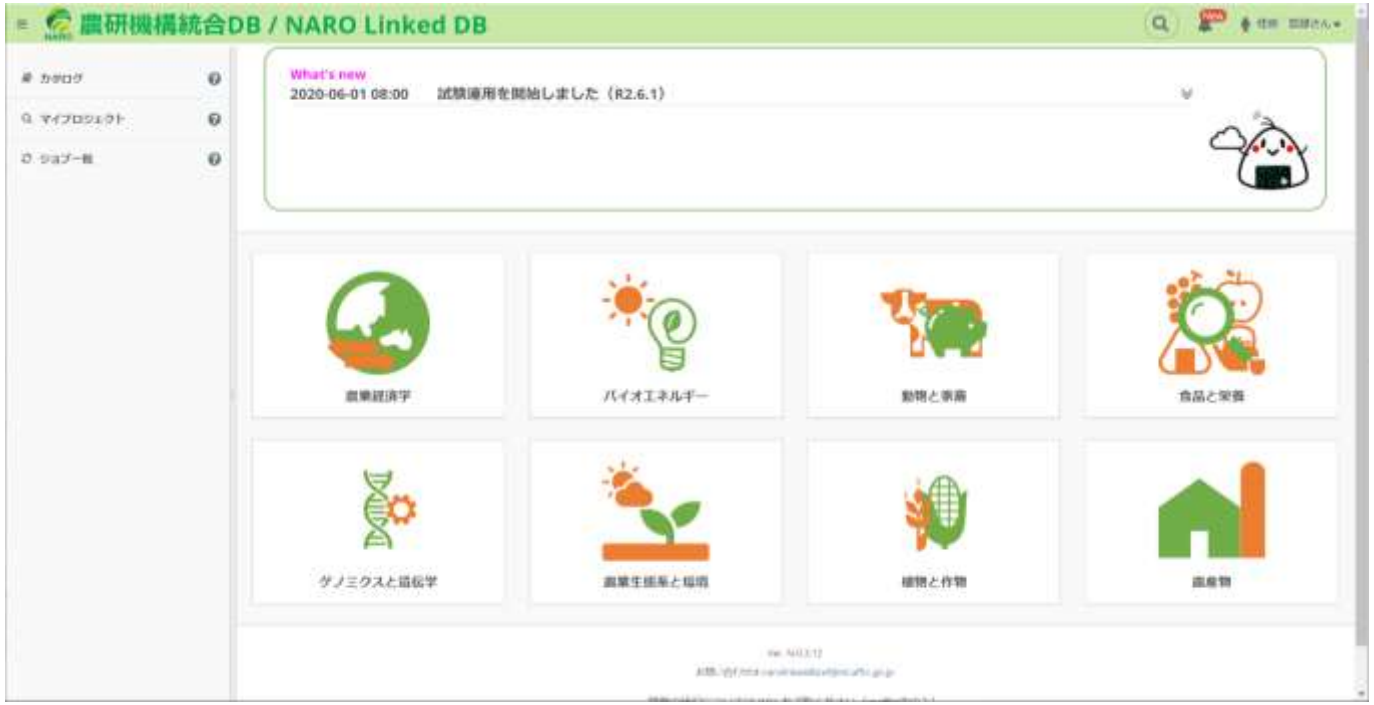
NARO

今日の話題

- 農研機構統合DB
  - 開発経緯
  - 1次DB
  - 2次DB
- 農研機構統合DBにおけるオープン・クローズドデータの実際
  - 運用ガイドライン
  - データ利用規約
  - 実際の運用
  - 退職時の扱い

データ戦略推進室 / 農研機構統合DB

- 2019年04月　　データ戦略推進室設置
- 2020年03月　　DB運用ガイドライン整備
- 2020年06月　　農研機構内限定試験公開開始
- 2020年09月　　農研機構外部ユーザとのデータ共有開始
- 2021年04月　　本格運用開始予定



農研機構統合DBのトップ画面

開発の背景

- 急増するデータ、散逸するデータ
  - [電子データ]農業分野における研究環境のICT化にともなう電子データが急増
    - 研究データの適切な保存・管理・整理が必要
    - 統計解析や機械学習等のデータ科学への適用に期待
  - [紙データ]経年により、機構内の貴重なデータが散逸しつつある

- 分野横断的かつ統一的なデータ基盤の構築が急務
- データ基盤として「農研機構統合DB」を構築する
  - 機構内の全研究データを農研機構統合データベース「NARO Linked DB」に一元的に集約

農研機構統合DB/ NARO Linked DBについて

- 愛称
  - 農研機構統合データベース / NARO Linked DB

- コンセプト
  - 人をつなぐ、データをつなぐ
    - 農研機構全体での研究データの共有・活用
    - 研究データを通じて人をつなぐ
    - グラフDBによってデータをつなぐ

農研機構ダイバーシティ推進キャラクター
「©おむすびなろりん」農研機構統合DBバージョン

- ねらい
  - 農研機構全体での研究データ共有・活用による、分野横断的研究などの推進。
  - データ解析体制の整備による、研究の高度化・迅速化の推進。
  - 農業界・産業界へのデータ提供による、「データ駆動型スマート農業」の推進。

- 構成
  - オンプレとクラウドに分散
  - 1次DB: メタデータ(※)を整理し、研究データをカタログ化
  - 2次DB: グラフデータベース。データの連携により、関連データをつなげて解析。

※メタデータ：タイトル、作成者、研究課題、ライセンス情報など、データの属性を示す情報

農研機構統合DBの概念図

農研機構

農研機構内ネットワーク

クラウドネットワーク

農研機構統合データベース

1次DB

オブジェクトストレージ

- 冗長性、高可用性を備える
- Amazon S3互換
- 大容量高速オブジェクトストレージ(3PB)

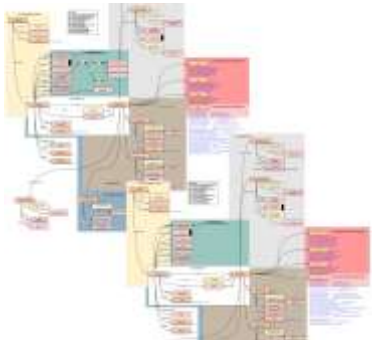
カタログ型DB

- メタデータを管理

SINET
学術情報
ネットワーク
10Gbps

2次DB

グラフ形式のデータ



- メタデータによる関連付け
- 新しい知識の獲得

テーブル形式のデータ

専門データベース群

Firewall

農研機構認証システム

Web Browser

機構内ユーザ
（ブラウザ）

機構内ユーザ
（コンソール）

Firewall

WAGRI API

WAGRI
認証システム

連携

クラウド認証システム

WAGRI API
HTTPS

機構外ユーザ
（WAGRI API）

Firewall

Web Browser

機構外ユーザ
（ブラウザ）

API
HTTPS

機構外ユーザ
（API）

1次DB(カタログ型DB)の概要

- 構成
  - メタデータDB（カタログ型DB） + オブジェクトストレージ
    - データにメタデータを付けてカタログ型DBで管理
    - データの実体はAmazon S3互換オブジェクトストレージ(3PB)に格納
- 特長
  - 機構内のデータを集約、安全に保管
    - 研究者のデータ保存の負担を軽減
  - 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
    - メタ情報を元に検索可能。新たな研究テーマを発見・展開
    - 掲示板機能により、研究者間のコミュニケーションを促進
  - 柔軟なアクセス制御により、データを安全に共有
    - 機構内ユーザは、機構内ネットワークから全員が利用可能
    - 機構外利用者も登録により利用可能
  - 外部公開データ用に別システムを用意(来年度より)

1次DB → データにメタ情報を付けて、塊で格納
2次DB → 1次DBの中から解析用に選抜されたテーブルデータ、グラフデータを格納

1次DB(カタログ型DB)の対象データ

- 農研機構は、長年に渡り、農作物・家畜のゲノム、育種、栽培、病害、食品の成分・機能性、環境等に関する、多様で膨大なデータを蓄積。
  - ドローンなどから撮影した画像セット
  - 定点カメラなどで撮影した動画データ
  - センサーなどで取得した観測データ
  - Excel，CSVなどの表形式データ
  - 既存データベースのダンプファイル（バックアップ）
  - ゲノム配列データ
  - 文献データ
  - 設計図書
  - マニュアル，など

農研機構内研究データの例

分野	データの内容
植物	ゲノム・育種，栽培，画像等
動物	ゲノム，家畜診断，管理等
昆虫・線虫	害虫，診断方法等
微生物	病害，病害写真，病害同定等
食品	機能性成分，成分分析等
環境データ	気象，施設内環境，土壌等
その他	農作業，農業経済，農業用地利用状況，水利，水質，インフラ整備関連，地盤調査，実験ノート等

2次DBの概要

- 1次DB内のデータセットからデータを選抜し、整理・格納
  - 有用性の高いもの、他のデータとの関連性の高いものを選抜
  - プロパティグラフ、RDFなどの形で整理
  - 統計的解析・機械学習による知識発見をサポート
- RDFデータサーバを構築
  - クラウド型グラフDBを導入
  - FusekiによるSPARQLエンドポイントの提供
- テーブルデータも格納可能
- 解析ツールを提供（順次追加）
  - 育種データビューワ(NARO Pedigree Viewer)を開発

1次DB → データにメタ情報を付けて、塊で格納
2次DB → 1次DBの中から解析用に選抜されたテーブルデータ、グラフデータを格納

2次DB（グラフデータベース）利用のイメージ

- グラフデータベースでデータをつなげる！（開発中）
  - 関連情報をつなげて検索する
    - 分類を用いた検索
      (例：果樹→かんきつ→うんしゅうみかん→山下紅早生の品種情報を検索)
    - キーワードを用いた検索（同義語にも対応）
      (例：βクリプトキサンチン→カリカキサンチンの成分情報を検索)
    - 上位概念、同位概念、下位概念による検索
      (例：山下紅早生→かんきつ→かんきつの出荷額を参照)
  - Linked Open Data(LOD)と連携できる
    - 食品成分表から栄養情報を得る（食品LOD）
      (例：山下紅早生→うんしゅうみかん→100g当たりのビタミンCの含有量を参照)
    - 作物の出荷情報、地域情報を得られる（統計LOD）
      (例：山下紅早生→平成28年産特産果樹生産動態等調査→県別出荷量→中国地方での栽培可否の判断に用いる)

- 食品LOD (公立はこだて未来大学) http://osakanabiyori.github.io/RDF/
- 統計LOD (政府統計の総合窓口 e-Stat) https://www.e-stat.go.jp/

APIによる操作

- API群(※)を構築中（進行中）
  - ドローン、定点カメラからの画像、センサーからのデータの逐次登録
  - 解析用にパラメタを指定して条件に合ったデータを取得
  - トークン認証、Oauth2.0認可に対応
  - Javascriptを用いたツールを試作

- 農業データ連携基盤WAGRI APIとの連携
  - WAGRIを通じたデータの「連携」・「共有」・「提供」を可能とする
  - 連携APIを試作
    - 統合DBに対して、認証を通してデータのアップロード、ダウンロードが可能
  - トークン認証、Oauth2.0認可に対応(WAGRI、NAROの2段階)
  - 今後拡充

WAGRI APIのロゴ

※API: アプリケーションプログラミングインターフェース。ソフトウェア同士が情報をやりとりするためのインタフェース

農研機構統合DBのオープン・クローズド戦略

- オープンデータ
  - 選抜したデータを全世界に公開
  - 公開用データベースを別途用意し、安全性を考慮（2021年4月運用開始予定）
  - メタデータ(一部抽出)による検索も可能
- シェアードデータ・クローズドデータ
  - データは機構内全体で共有するのが基本
    - メタデータも共有 → 農研機構内ユーザはメタデータから検索可能
  - ただし、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可能
    - 1次DBは柔軟なアクセス権設定が可能 → 便利なNASのように利用できる
      - ユーザ権限、グループ権限を設定可能
    - 外部利用者を登録して共有することも可能（ただし、メタデータの利用は限定的）
  - メタデータを秘匿することも可能
    - メタデータはインデックス化されており、通常は機構内ユーザの検索結果に表示される（機構内限定共有）。
    - 検索結果にも表示されないように設定することも可能（要申請）

データ作成者の権利は、研究データ利用規約によって担保（データが勝手に他人に流用されることはない）

利用促進のための3要素

1. 運用ガイドライン策定
2. データ利用規約策定
3. AI/DB教育による広報活動

クローズドデータへのアクセス制御

- 統合DB全体
  - Firewallによる保護
  - Web Application Firewallによる保護
  - 経路制御
    - ソースIPによる制限
    - Application Gateway(リバースプロキシ)による制御
  - IdPによるアクセス管理
    - ユーザごと、グループごとの権限制御
    - トークンを用いた認証・認可
- 1次DB内
  - データセット毎のアクセス管理
  - メタデータへのアクセス制御
- 既存公開サーバ内
  - 各データへのアクセス管理

異動・退職時のデータの扱い

- 異動・退職時のデータ保存・権限移譲
  - 異動・退職時は統合DBへの登録を以ってデータ移管を完了する。
  - 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する。
  - 退職後、退職者のアカウントは速やかに削除する。
  - 1次DBには権限移譲の仕組みを導入する（今年度末から利用可能）。
- 退職後のデータアクセスについて
  - 研究成果管理規定に従い、業務上、 認められれば外部からの一部データアクセスを許可。

まとめ

- 農研機構統合DBについて紹介した。
  - カタログDBとしての1次DB, テーブル・グラフデータを扱う2次DB
  - メタデータによるデータ検索が可能
  - 農研機構内部ユーザだけでなく、外部ユーザも申請により利用可能
  - 柔軟なアクセス制御が可能
  - データの共有が容易
- 農研機構統合DBにおけるオープン・クローズド戦略の実際について紹介した。
  - 運用ガイドライン
  - データ利用規約
  - アクセス制御の実際
  - 退職時の扱い
- 取得された様々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。