高付加価値科学データ創出を指向した 研究データ管理プラットフォームのアーキテクチャ

菊地伸治, 門平卓也, 鈴木峰晴, 内藤裕幸

国立研究開発法人 物質・材料研究機構(NIMS), 統合型材料開発・情報基盤部門(MaDIS)

〒305-0047 茨城県つくば市千現1-2-1

E-mail: KIKUCHI.Shinji@nims.go.jp, KADOHIRA.Takuya@nims.go.jp, SUZUKI.Mineharu@nims.go.jp, NAITO.Hiroyuki@nims.go.jp

あらまし 近年, クラウドコンピューティング, 機械学習の成熟化に伴い, 材料科学分野におけるデータ利活用は益々高い需要にある. そのためには研究開始の創出期から実用に向けた適用応用期迄に創出される一連の研究データをシームレス・高品質に管理・提供出来る研究データ管理プラットフォームの実現が求められる. 本稿では, 国立研究開発法人 物質・材料研究機構で設計開発を進める材料データプラットフォームのアーキテクチャの上位機能層の構成について概説する. そこでは来歴管理や, データ識別・一意性を実現するための PID サービス, 機構内に広く存在する計測装置・システム等から大規模にデータ収集を目指すアダプタ, オントロジ管理等, を含んで構成され, マイクロサービス概念を適用した柔軟性を合わせ持つ.

キーワード 科学データ管理, アーキテクチャ

Architectural design of the research data management platform oriented to generating the value added science data

Shinji Kikuchi, Takuya Kadohira, Mineharu Suzuki, Hiroyuki Naito

National Institute for Materials Science(NIMS),

Research and Services Division of Materials Data and Integrated System (MaDIS)

1-2-1 Sengen, Tsukuba, Ibaraki, 305-0047 JAPAN

E-mail: KIKUCHI.Shinji@nims.go.jp, KADOHIRA.Takuya@nims.go.jp, SUZUKI.Mineharu@nims.go.jp, NAITO.Hiroyuki@nims.go.jp

Abstract In recent years, data sharing, utilization and reuses in Material Science have been in high demand, responding to the maturation of Cloud Computing and Machine Learning. Accordingly, it has been required to realize a Data Platform to manage research data, that can manage and provide sets of the research data generated from the research initiation phase to practical applications continuously with high quality. This paper presents the outline of architectural aspects of the upper functional layer in the Data Platform that has been designed and developed at the National Institute for Material Science, Japan. That includes Data Provenance, PID Service for realizing identification of research data individually with its own uniqueness, Adapter for collecting data from distributed measurement devices widely in a large scale and Ontology Management, under applying the conceptual Micro Service.

Keywords Science Data Management, Architecture

1. 緒言

材料科学分野では近年, 世界規模で Material Informatics と呼ばれる材料科学とデータ科学を融合した取り組みが進展している. Material Informatics とは,蓄積された膨大な実験データ, 計算機能力の向上により算出可能となった膨大な計算データを入力として統計学, パターン認識, AI 等のデータ解析技法を

用いてプロセスと特性間, 異なる特性間に成り立つ法則性を抽出・発見, 予想を可能とすることで, 新たな材料開発を加速することを含意している[1],[2]. 特に近年ではテキストマイニングで得られる大量のデータに機械学習や深層学習を加えた材料探索の開発等も進められている[1],[3]. 材料科学から他領域に目を転じると, 予てから Scientific Workflows として天文学, バイオ領域等で類似の取り組みが存在しており, いずれの分野でも大規模データの集積, 機械学習の適用により新たな知見の獲得を着実に進めて来ている[4]. また古くは製造業における Computer Integrated Manufacturing として発展してきた階層型アーキテクチャとも構造的なアナロジーが存在する[5]. この様に広く散在する大量なデータをシステマティックに収集・集積の上, 機械学習の適用により新たな知見を獲得するパラダイムはクラウドコンピューティング, 機械学習の成熟化に伴い, 益々深化している. その結果, 研究開始の創出期から実用に向けた適用応用期迄に創出される一連の研究データをシームレス・高品質に管理・提供出来る研究データ管理プラットフォームの構築と実現が分野横断で進展している.

本稿では, 材料科学に纏わる各種データを“つくる”,“ためる”,“使う”,“公開する”という4機能が相互に関係した Material Informatics 環境の実現に向けて, 国立研究開発法人 物質・材料研究機構で設計開発を進める材料データプラットフォームについてのシステムアーキテクチャ, 特に上位機能層の構成について概説する[6]. そこでは他分野と共通な技術要素のみならず, 材料科学分野固有の要件も考慮して様々な技術要素を取り込んでいる. 前者に関する具体的要素としては, 従来からの強い要求がある研究データの来歴管理[7], それを支える研究データの識別・一意性管理のための PID サービス, データ信頼性保証の仕組み等であり, 後者には多様で領域特有性の強いデータをハンドリングするため, オントロジを含んだメタデータの流通, 計測装置・システム等から大量の研究データの収集を実現する柔軟性を持つアダプタ等である. 当該プラットフォームでは機械学習それ自体を実施する訳ではないが, 多様な研究データ群から意味的に整合させた上で統合する専用の Data Set 生成機構の開発等, 半ば挑戦的な試みも含む. 更には多様な領域を扱う故に自ずと内在する構成的複雑性, 研究における異なる進化過程も柔軟に扱い, 各関連システム群の自律性を尊重して統制, 逓増的に機能拡張が出来ることを考慮してマイクロサービスの概念も取り入れている. 当該分野固有事項を捨象, システム構成論の立場を強調すれば, “高付加価値科学データ創出”を指向した研究データ管理プラットフォームと定義することも出来る.

以下, 本稿の構成を述べる. 続く2章では, 当該プラットフォームに関してマクロ的視点で概説する. アーキテクチャの決定要因となった要求事項, 背景等を概説の後, 上位機能層の静的構造について概説する. そこでは扱う情報モデルや, 流通のためのメタデータ形式も言及する. 3章では主要要素群の内, 特記するべき要素群について概説する. 具体的には, 柔軟性を持つアダプタ, PID サービス, データ信頼性保証である. 4章では一部考察を含んで暫定評価を加える. 具体的には概説に至る迄の具体的制約事項や, 進化に対する見通し, 他研究プラットフォームとの比較の上で当該プラットフォームの位置付け, 大局的な要求に対する見解等である. 最後の5章で結言とする.

2. マクロ的視点から見たアーキテクチャ

2.1. 背景と概要

一般にシステムアーキテクチャでは, 静的構造(設計時構成)と, 動的構造(実行時構成), 外部に対する振る舞い, 品質特性の４つの独立した側面で定義される[8]. 静的構造(設計時構成)と動的構造(実行時構成)では, 機能層に応じて下位構造であるハードウエア, ネットワーク基盤等と, 上位構造である業務アプリケーション, サービス等に分化される. 本稿では上位機能層の静的構造(設計時構成)を中心に概説する. それ以外の３点, 並びに下位構造についても基本設計フェーズでは無論, 検討しているが, 本稿では背景事項として扱う.

以上の論点に限定する背景には, 特に品質特性, 外部に対する振る舞い等については, 当該プラットフォームが逓増的に発展する性格を持ち, 試行的な適用の上で運用に移行しなければ, 最終的・確定的な設計解には至らない, と言う状況もある. 例えば, 材料分野の多様性に従い, 基本設計フェーズ迄に計算量の実デマンドを高精度で見積もれる訳ではない. 安全側に見積もると遊休状態を招くリスクも懸念される. このため, サイジングプロセスを確実に実施出来ない等の制約故に, 非機能要件を含んだ品質特性は簡単に定義出来ない. 基本設計フェーズの最終盤になった段階でも, 将来的に発生し得る種々デマンドに柔軟に対応出来, 容易に拡張出来る仮想化環

境の導入する, との基本方針以上を説明出来る段階にはない. 同様なことは業務プロセス定義についても言える. オントロジの様な情報資産が発展途上で, かつ接続するシステム群が比較的少数の中では, 当該プラットフォームが提供するサービスそれ自体が試行的な側面を持つことも拭えない. このため, 業務プロセスについても試行適用～課題抽出～更新・最終確定の一連のサイクルを完了する迄は粗いユースケースのみを定義出来るに留まり, 確定的なプロセスを決定する迄には至らない. 以上の様な背景から, 本稿では静的構造(設計時構成)に関する概説に留め, それ以外については別の機会に報告する.

静的構造(設計時構成)を説明するには, ソフトウエア要素の構成・配置, 並びに情報モデルの論点が重要になる. 特に後者については構成要素の記述と伴に, 当該プラットフォームが扱うデータの意味的構成も含める.

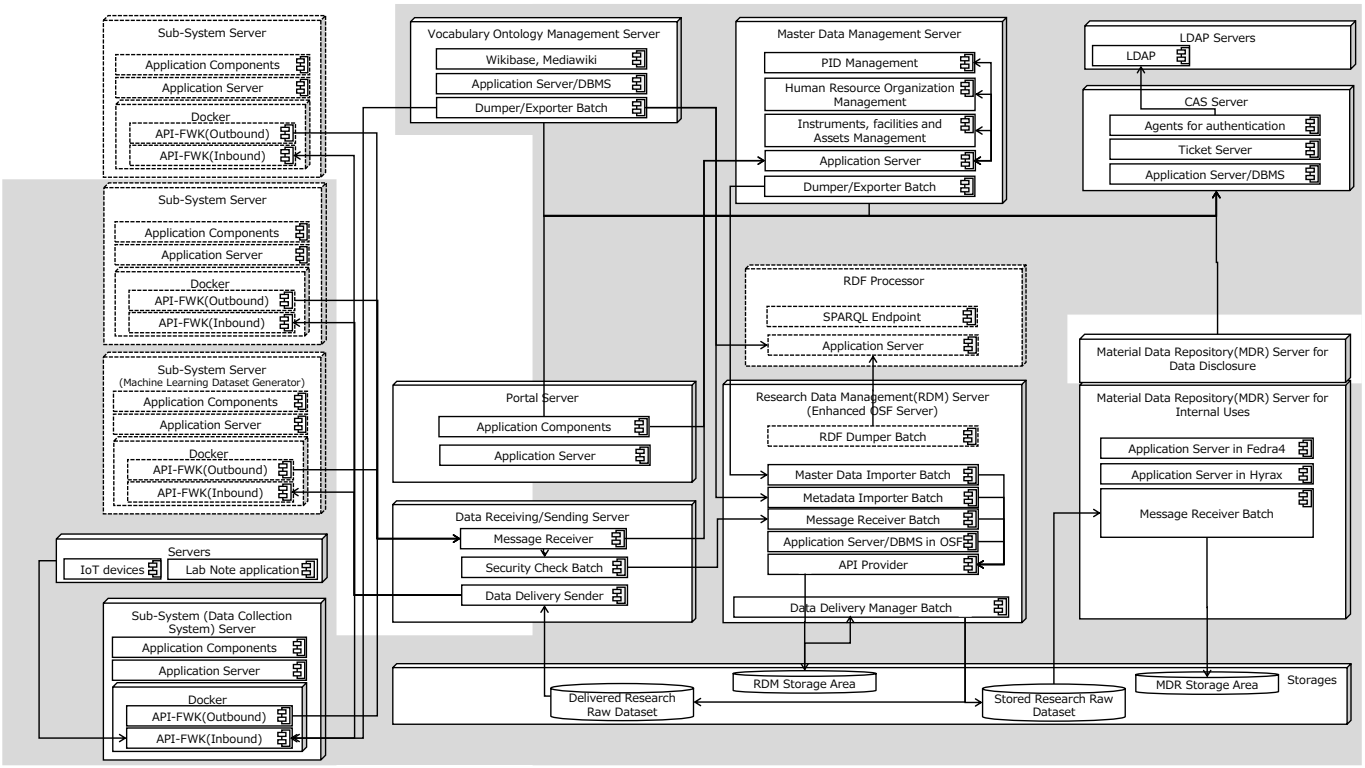

図.1. UML 配置図による全体アーキテクチャ

2.2. ソフトウエア要素の構成・配置の側面

図.1.は当該プラットフォームの上位構造である業務アプリケーション群の構成要素を記した UML(Unified Modeling Language)配置図になる. 一つの長方形はサイトを意味し, これらサイト間の関係は, 呼び出し方式の指定の代わりに有向線で記し, 呼び出し関係(依存関係)を意味する. また図中で灰色で塗り分けられている部分は, 当該プラットフォームを構成するネットワークの内, セキュアな内部セグメント領域であり, それ以外は DMZ 上のセグメント等に相当する. 図中の各要素は, 概ね基本設計書フェーズで定義されている構成要素名を記しているが, より読者の理解を得易い様に本稿では一部名称を変更, 更に計画された将来拡張部も含んでいる.

この図.1.を機能的に捉えると２つの構成に分化される. 図中の下半分は業務プロセスを扱う機能群に相当し, 左側から Data Collection System(DCS)と称する研究データの発生源・これらの統制機能, Research Data Management(RDM)に代表される研究データ管理機能, そして Material Data Repository(MDR)の様な研究データの再利用のための公開機能であり, これらがシームレスに連携する. これは前述の“つくる”,“ためる”,“使う”,“公開する”に相当する機能群となる. これに対して図中上半分は, この業務プロセスを支えるための情報資源管理機能, ユーティリティ機能に相当し, RDF による検索等の再利用基盤も含んで構成される.

これに対して, 当該図.1.を新たにデータ流通の点から眺め

ると別の分類も定義される. それは研究データを“ためる”ために流通統制・管理を行う中核コアのシステム HUB と称する部分と, 研究データを“つくる”,“使う”に相当し新たな付加価値を生み出す一連の周辺のサブシステム群となる. この場合, 中核コアのシステム HUB により必要とされる研究データが供給される. 前者の中核コアのシステム HUB に相当する機能は, Research Data Management (RDM) Server, Data Receiving/Sending Server の部分であり, 共通的に利用される後述マスタデータ, 辞書データを供給する意味では Vocabulary Ontology Management Server, Master Data Management Server 等も含む. 更にはシングルサインオンの機能を提供する CAS Server, LDAP Server 等もこの中核コアのシステム HUB に含まれる. これに対してサブシステム群は, 当該図.1.での左側で定義される一連の Sub-System Server 群であり, 前述中核コアのシステム HUB とは, 柔軟性を持つアダプタである後述 API-FWK を介して自律的に連携する.

最後に, 当該配置図は, 基本設計フェーズの初期段階から多様で変動する要求を取り入れて逓増的に進化して来ており, 未だに潜在的な更新可能性は否定出来ない. 例えばワークフロー機能についても当初, 定義されていたが, 基本設計フェーズの途上で複数要因により, その導入を取り止めている. この背景については4章に概説する. 以下, 各構成要素を, 表.1.に記す.

表.1. サイト,ソフトウエア要素の定義

サイト名, ソフトウエア要素	定義
Research Data Management (RDM) Server	周辺システムに相当する各種サブシステム(Sub-System Server)群から生成される研究データ(詳しくは 2.4 節で説明するメタデータ, 保管対象データ)を集積・管理の上で 流通管理を行う機能であり, 中核コアのシステム HUB の中心機能である. これと伴に研究データの来歴管理も行う. 実装に当たっては OSS の OSF(Open Science Framework)を採用しており, バックエンドサーバとして機能する[9]. この為に後述マスタデータの取り込みや, 内部データを RDF 化してダンプする複数のバッチ機能も配置している.
Data Receiving/Sending Server	周辺の各種サブシステム(Sub-System Server)群から生成される研究データを集配信する機能である. 2.4 節で説明する保管対象データのセキュリティ検証, マルウエア対策を実行する機能を含み, 研究データの品質の維持管理を担う基盤としても機能する. 各種サブシステム(Sub-System Server)群間で流通・交換する保管対象データは, 数～数十 G バイトになる場合もあり, それらを送受信する際にネットワーク帯域を消費してしまう懸念もあるため, 保管対象データを分割して流通させる機能も実装する.
Portal Server	利用者が当該プラットフォームを操作する上で GUI を提供する機能である. 周辺の各種サブシステム(Sub-System Server)群へリンクも含み, シングルサインオンによりシームレスな操作を提供する.
Sub-System Server, API-FWK	研究データを“つくる”,“使う”に相当し, より付加価値の高い研究データを生み出す一連のサブシステム群である. 材料科学の分野毎に多様なレガシーデータベース, 情報システムが存在している. 具体的には各種計測装置から集積した研究データを管理するシステムや, シミュレーションのライブラリ・計算データの管理, 文献データベース, 高分子データベースの様な専門データベース群である. これらは各々, その時々にそれら自身への要求事項に従い開発されたもので, 実装環境・言語に標準・統一性を期待することは出来ない異種システム群である. この様なサブシステム群からデータを移送するため, その通信処理を標準化しサブシステム群と連携する専用アダプタが API-FWK である. これは, 従来の EAI(Enterprise Application Integration)におけるアダプタ類の役割に相当し「データ形式」や「通信規約」等を標準化し, サブシステム群を通信に関する煩雑な手続きから解放する目的で開発されている. 詳細事項は, 3.1 節で説明する.
Sub-System Server, (Data Collection System)他	Laboratory automation の進展に応じて, 実験に利用される計測機器類は材料科学の分野でも高度化が進んでいる. これにより, 機器自身が計測結果等の研究データの生成, 保存, 管理等を行っている. 更には実験ノートのタブレット化等, ペーパレス化も同時に進行している. このサブシステムは Data Collection System(DCS)と称し, 実験計測の現場近くに配置される標準システムであり, 各種の研究データ・計測データの収集・集積を担う. このシステムは製造業で Computer Integrated Manufacturing として発展してきたショップフロアー制御と同水準に位置付けられ, 4 階層機能で定義されるリファレンスモ

	デルでは他システム連携を担う第三階層の機能に相当する[5].
Sub-System Server, (Machine Learning Dataset Generator)	Research Data Management (RDM) Server に集積・管理された保管対象データの相当数の配信を受け, そのデータ内で保管される計測装置・シミュレーションプログラムが生成する機械可読データ, もしくは意味解釈のアノテーションを付与した抽象化データ群を抽出, セマンティクスに基づき新たに意味的統合をして, 機械学習プログラムへのインプット形式に成形するための Data Set 生成の処理系である. 意味的統合を実施するため,挑戦的な機能であるが, アーキテクチャ上では周辺サブシステム群の一つとして扱われる.
RDF Processor	Research Data Management (RDM) Server に集積・管理されたメタデータや内部データである来歴管理情報を RDF 形式でダンプした結果を取り込み, 多様な関連検索機能を提供する. この機能は, 現在, 実装に向けて具体化しておらず, 将来, 拡張実装する予定である.
Material Data Repository (MDR) Server	Research Data Management (RDM) Server に集積・管理された保管対象データを“公開する”際に利用する機能である. 保管対象データは, Research Data Management (RDM) Server を介した流通段階では, 後述 2.4 節で記す様にアーカイバで一ファイルに統合・圧縮されて集配信される. それに対して公開段階では斯様な形式では利用者観点での検索等に耐えられないため, 統合・圧縮を解き, 利用者が直接閲覧出来る様な形式に戻す. 更に提供される研究データの品質保証を目指して発行元の電子署名等も付与される.
Storages	巨大で多様なデータ類を保管する領域であり, 仮想化ファイルシステム上に, 各システム要素がマウントすることでそれぞれのファイルを管理する予定である. 詳細説明は割愛する.
Vocabulary Ontology Management Server	後述表.2.で定義される《Abstract》辞書定義のインスタンス群を管理するシステムであり, 特定領域で標準的に利用される語彙・オントロジ類を中央集権的に管理する. 実装に当たっては実績のある Wikibase を用い, 他 Server 群にそれらを提供・配信する周辺機能が付与されている.
Master Data Management Server	後述表.2.で定義される《Abstract》マスタデータのインスタンス群を管理するシステムであり, 人員, それに関連する組織, 更には認可, 装置等のマスタデータ類を中央集権的に管理している. 特に人員に関して当該プラットフォームでは, 各利用者は複数 ID 体系でログイン出来ることを目指しており, 当該研究機構の ID, 学術認証 (https://www.gakunin.jp/), 一部制約を設けるが ORCiD(https://orcid.org/)の利用を想定している. このため複数 ID 体系で表現される同一人物を名寄せして管理する. 他 Server 群に API, ファイル交換を介して, それらマスタデータのインスタンスを提供・配信するが, メンテナンス用の GUI 等も保持する. 更に, この Server では, 当該プラットフォーム内で研究データの識別・一意性を保障するため PID(Persistent IDentifier)でもある種々の識別子を取得・管理する機能も含む.
CAS Server, LDAP Server	利用者がログインする際にシングルサインオンを実現するために採用しているものである[10]. 前述の様に各利用者は複数 ID 体系でログイン出来ることを目指しているが, LDAP Server は当該研究機構の ID 情報を管理するものであり, セキュリティレベルに応じて複数インスタンスが配置される予定である.

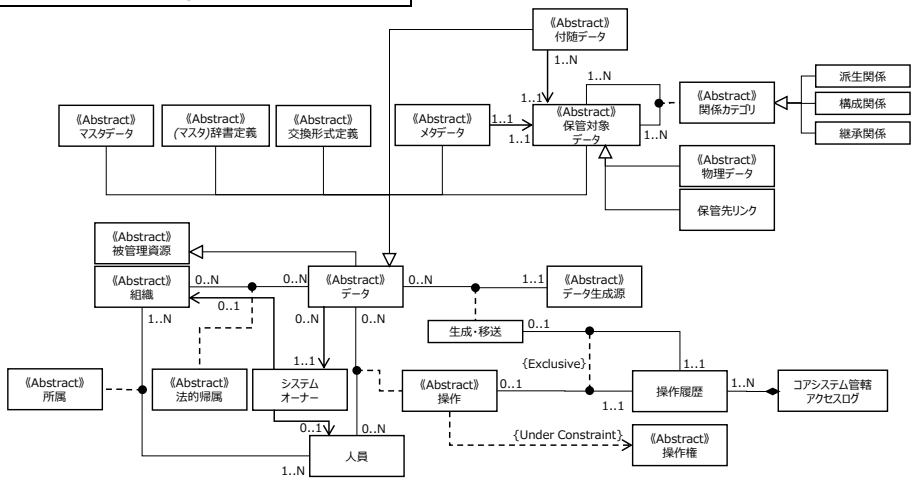

図.2.当該プラットフォームの上位機能で扱うべき情報モデル

2.3. 情報モデルの側面

図.2.は当該プラットフォームの上位構造が扱うべき情報モデルを UML クラス図で定義したものである. この図は情報モデルとしては最もプリミティブなものに限定しており, この情報モデルを継承・拡張することで実装に至る細部が概ね定義される. 本稿では紙面の関係上, 情報モデルの細部については割愛する. このクラス図から理解される様に《Abstract》データクラスを軸として２分野に分化される. 図.2.の下部は業務プロセスに関連し, 上部はそのデータクラスを特化したメタデータ等に継承される. これは, 前述図.1.の２つの構成部分にそれぞれ対応する. 表.2.に主要なクラス群を定義する.

表.2. 情報モデル内の各クラスの定義

クラス	定義
《Abstract》データ	当該プラットフォームで管理する全てのデータを抽象化したクラスである.
人員	所属に関係しない人員のクラスである. 当該情報モデル上では明示的に表現されていないが, 個人を特定出来る項目と, それ以外の一般的な属性項目を分離して管理する.
《Abstract》組織	法人組織や部門等, 人員が配置されるものを汎化した抽象クラスである.
《Abstract》データ生成源	《Abstract》データを生成するあらゆる機器・装置を抽象化したクラスである.
《Abstract》操作	人員が《Abstract》データに対して施す処理内容を汎化したものでありマニュアル登録等の具体的な操作行動となる.
《Abstract》マスタデータ	《Abstract》データの一つの実装インスタンスであり, 共通的に利用されるものを「マスタ」の接尾辞を付与して管理している. 特に《Abstract》辞書定義他以外のものを扱う抽象クラスである. 特化される対象は「ソフトウエア資産」等も想定したが, 実際には「人員・組織・認可」と「装置・資産」等のみが定義されている.
《Abstract》辞書定義	《Abstract》データの一つの実装インスタンスであり, 語彙に関する辞書の具体的実装を束ねる抽象クラスである. 特定領域で標準的に利用されるべき語彙・オントロジを共通的に管理しているものを「辞書」と称するが, この抽象クラスは, それらを汎化して定義される抽象クラスである. 主に RDF で記述されて辞書インスタンスとして特化される. 物質辞書, 特性辞書, 計測辞書等に特化される予定であるが, 未だに調整途上にある.
《Abstract》交換形式定義	メッセージ等の交換形式の総称であり, 汎化した抽象クラスである.
《Abstract》メタデータ	《Abstract》データを継承するもので, 《Abstract》保管対象データを記述するために指定項目を持つデータである.
《Abstract》保管対象データ	《Abstract》データを継承するもので, メタデータで記述され参照・提供される実際のデータであり, 電子署名を付与する対象である. 複数の要素ファイル群で構成される場合は, アーカイバで一ファイルに統合・圧縮されて管理・保管される. 物理データとして添付される場合, もしくは規模が大きい故に保管先リンクのみが URI で記述される場合の２つの特化ケースを想定している.

2.4. 扱うデータの意味的側面

図.3.は, 当該プラットフォームで扱うデータ・記述群を内容・抽象度に応じて分類し, それをどの様な形式のデータ表現に対応付けるか, を記載した概念図である. 図中左側のマトリックスは, その中心概念を表現しており, 材料科学分野で生成される諸研究データの分野内容を横軸に, その記述抽象度を縦軸にして分類したものである. 当該マトリックスの最上位行は, 研究データそれ自体の生成日付や作成者等のいわゆる研究データに関する書誌情報に相当し, 抽象度が最も高く分野に関係無く全てのユースケースで共通的に利用される. その意味で「必須共通メタ」として定義される. サムネイル画像を除き, 当該必須共通メタデータ以下の行は, 分野内容毎に分類される. 分野内容は材料科学の論点でその研究データの特徴を記述するための項目群であり, 「物質材料」と「合成・プロセス」を与件とした結果, 得られる「特性」を記載することを基本とし, その際の「計測手法」, もしくはシミュレーション等により評価されることも考慮して「計算」の記述群を含む. この５要素を選択的に利用させることで, 多様な材料科学研究の方法・ユースケースに対して可能な限り共通的に適用されることを目指している. これに対してマトリックスの縦軸に相当する記述抽象度は, 計測装置・シミュレーションプログラム等自体が生成する機械可読のバイナリデータを最下層とし, 人による可読性を向上する目的で抽象化・アノテーション化を図ったものが, より上位に位置付けられ

る.

このマトリックス定義に基づき, 当該プラットフォームで扱う研究データ群を如何に物理的にマッピングするかが決まる. 具体的には「必須共通メタ」である最上位層と分野内容の最も基本的なデータ・記述群に関しては, 図.4.の形式で定義されるメタデータ形式で記載される. それ以外のデータ・記述群については, メタデータと伴に流通・保管する「保管対象データ」に含まれる. この保管対象データは, 表.2.で定義される《Abstract》保管対象データクラスのインスタンスに相当し, 複数の要素ファイル群で構成される場合, アーカイバで一ファイルに統合・圧縮される. このため, 任意の保管対象データには内部構造を持ち, 図.3.の右下部で定義されるディレクトリ構造にて, それら要素ファイル群が管理される.

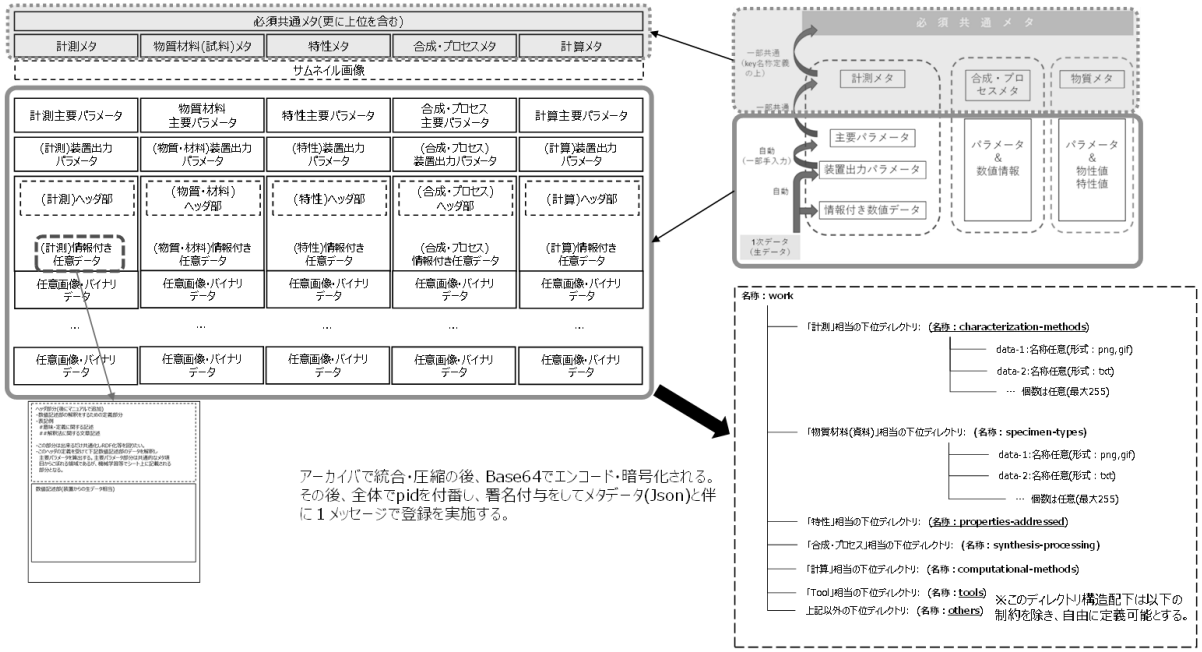

図.3.　当該プラットフォームで扱う研究データの構造に関する概念図

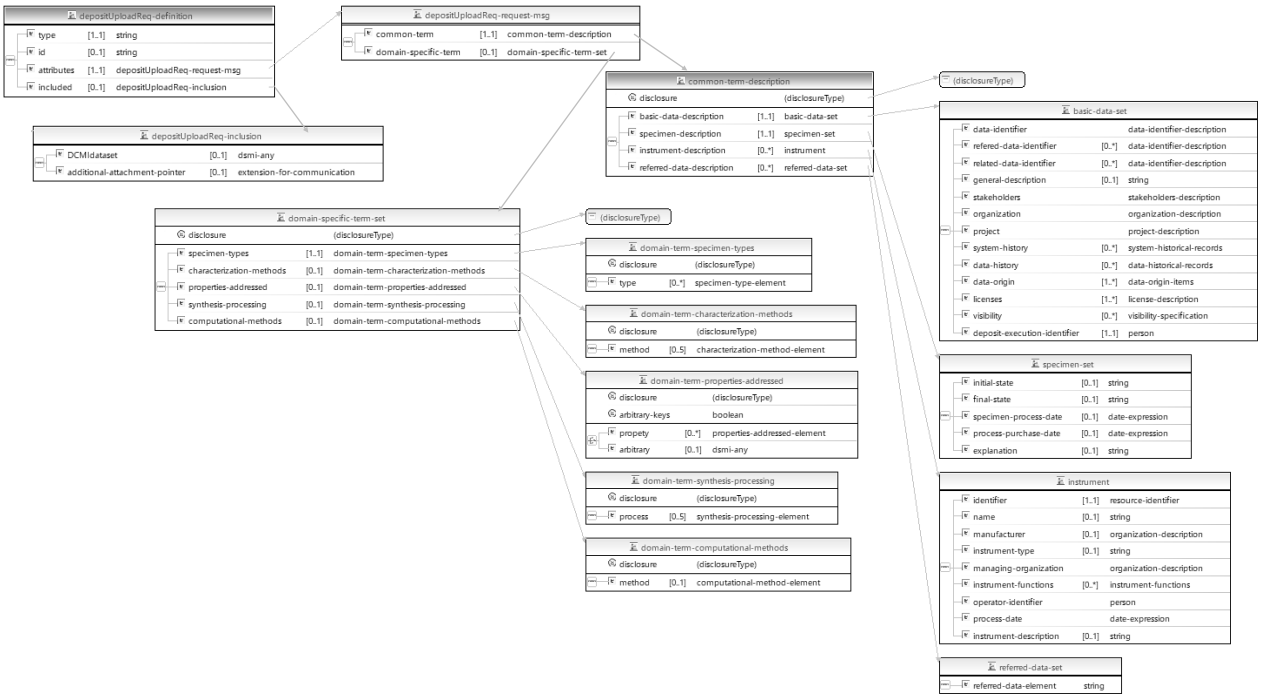

図.4.　メタデータの形式定義(抜粋)

図.4.は当該プラットフォームで流通・交換・保管されるメタデータの主要構造である. 当該プラットフォームのメタデータは, XML Schema で定義の上, Json 形式に形式変換して利用される. このメタデータは流通・交換・保管の各々で利用されるため多面的な要素を持ち, 単に保管対象データ記述の側面だけではなく, 研究データ登録のための通信メッセージとしても機能する. 保管対象データ記述としては, 研究データそれ自体の書誌情報, 材料科学の論点でその研究データの特徴を記述するための前述分野内容に基づく項目群も含んで構成される. 以上から複雑な構造を内包する.

図.4.の最左上位の要素が depositUploadReq であり, これが記述の最上位になる. この要素とその構成要素である depositUploadReq-request-msg 等は研究データ記述の側面よりも, それを交換する上での通信メッセージの側面が強く, それが故に標準的な JSON:API v 1.0の記述流儀を継承している [11]. depositUploadReq-request-msg 要素は, 二つの構成要素である common-term-description と domain-specific-term を含む. 前者の common-term-description 要素は研究データに関する書誌情報を記述するためのものであるが, 一部には材料, 並びに装置に関するサマリー・概要情報を含む. 当該 common-term-description 要素の主要部は basic-data-set 要素であり, これは当該プラットフォーム内で長期に渡り, 研究データを識別・一意性を保障するための記述で, PID(Persistent IDentifier)でもある data-identifier 要素や, そのデータの生成母体となったプロジェクト識別子 project 要素, 生成日時等の履歴情報である data-history 要素を含んで構成される. 対して後者の domain-specific-term 要素は, 前述分野内容を具体化した項目群であり, 計測(domain-term-characterization-methods), 物質材料(domain-term-specimen-types), 特性(domain-term-properties-addressed), 合成・プロセス(domain-term-synthesis-processing), 計算(domain-term-computational-methods)の各記述要素を含んで構成される. 前述の様に当該メタデータは, 多様な材料科学研究の方法・ユースケースに対して可能な限り共通的に適用されることを目的としているため, 前述５要素は選択的に利用される.

3. 主要要素

本章では, 特記するべき主要要素群について選択の上で概説する.

3.1. 柔軟性を持つアダプタ

これは表.1.で定義される API-FWK のことであり, 「データ形式」や「通信規約」等を標準化し, サブシステム群を通信に関する煩雑な手続きから解放する目的で開発されている. その具体的な手続きとは, (1)流通・交換する要素ファイル群をアーカイバで統合・圧縮と取出し, (2)流通時の暗号化・復号化, (3)ウイルス対策, (4)電子署名付与・検証, (5)メタデータとの関連付け, (6)研究データの識別・一意性を保障する PID(Persistent IDentifier)識別子の取得・付番等である. その構成概要を図.5.に記す.

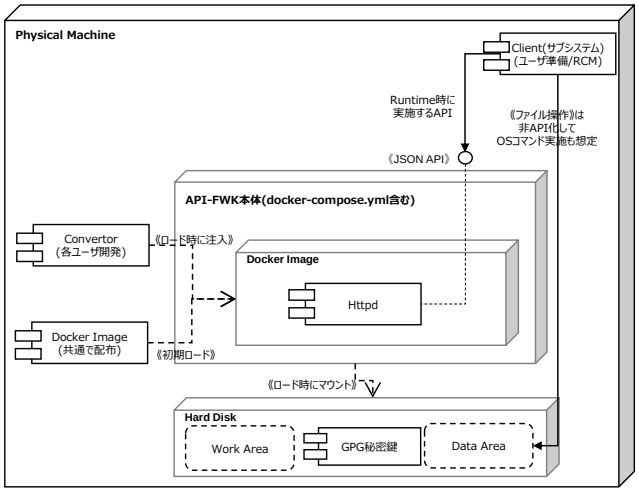

図.5. アダプタの構造

このアダプタは, それ自体が一つの独立したシステムとして機能し, docker 環境内に組み込まれ, 各サブシステムが配置される物理・仮想マシン上に配備される. そしてクライアントであるサブシステム・外部システムは, このアダプタに対して WebAPI を介して必要な操作を行う. それによりアダプタの通信制御を行い, 必要な研究データを当該プラットフォームの中核コアのシステム HUB 機能である Research Data Management (RDM) Server にアップロードする. 斯様な通信二重化を適用した背景には, (1)各種サブシステム(Sub-System Server)群が実装環境, 言語に関する標準性・統一性を持ち得ない異種システム群であること, (2)言語依存の組込みライブラリを提供し, 各サブシステムを改修する方式を採用すると,

システム拡張の都度, 実装言語毎のライブラリを開発・強化せざるを得ないリスクが伴い, 予算上の制約がある中では困難であることに依る. この為には, 言語独立な共通 API を提供する方が望ましい.

このアダプタと連携する任意サブシステムでは2.4.節で定義されるメタデータ形式と保管対象データの生成とアップロードを実施する必要がある. しかし各種サブシステム(Sub-System Server)群が異種である程, 益々, 固有のデータ管理形式・表現形式が存在する. その為, 固有の形式から2.4.節で定義されるメタデータ形式と保管対象データへのマッピング・変換を図る必要がある. 当該アダプタを利用して研究データをアップロードする場合, メタデータ形式へのマッピング・変換は, 図.5.中の Convertor 部分をオーバーライトすることで対応する. Convertor 部分のディフォルト実装は, 2.4.節で定義されるメタデータ形式の検証・受信を中心としているが, 当該アダプタを適用する際のシステムインテグレーション時にカスタマイズも許容する. 標準性・統一性が弱い異種システム群に対しても, 共通的に適用出来るためには, 当該アダプタ自身に柔軟な複数のカスタマイズポイントを持たせる必要がある. この Convertor 部分のオーバーライトはその代表例である.

前述の様に, 各種サブシステム(Sub-System Server)群で流通・交換させる保管対象データは, 数～数十 G バイトになる場合もあり, それらを送受信する際にネットワーク帯域を消費してしまう懸念もある. このため, 保管対象データのサイズ等を評価判定の上, 分割配信するか否かを選択する, ある程度の自律性を持った機能も含んでいる. 更には前述 Vocabulary Ontology Management Server で中央集権的に管理される語彙・オントロジ類についても, このアダプタにて一度配信を受け, 各種サブシステム(Sub-System Server)群からの問い合わせにより, それらを提供する. これらにより, 当該アダプタは単に通信に関する煩雑な手続きからサブシステム群を解放するだけではなく, 研究データに関する意味的統合や, 通信の自律的制御等のより高度なエージェント機能の役割も担っている.

3.2. PID サービス

周辺の各サブシステム(Sub-System Server)群が生成する研究データを集積・保管の上で集配信・流通管理を行う際, それらの来歴を管理するには, 当該プラットフォーム内で個々の研究データの識別・一意性を確保する必要がある. また新たな分析作業に於いて, 複数研究データ群を統合して比較可能とするためには, 試料や装置等についても識別・一意性を持った識別子を付番の上, それら識別子間の関係を妥当な形で管理する必要がある. この目的のため, 識別・一意性を保障し PID(Persistent IDentifier)として管理するための機能を当該プラットフォーム内で Master Data Management Server に配置し, サービスとして提供している. サブシステム(Sub-System Server)群はアダプタである前述 API-FWK を介して, それらの PID の識別子を取得, メタデータ等に記すことで, 意味的整合性を確保の上, 研究データ間の統合可能性を実現する.

3.3. データ信頼性保証

提供される研究データについて, 当該プラットフォーム内で決められた手続きで処理していることの保証, かつ公開においても同様の保証をするため, 保管対象データに対して電子署名を付与する. これは当該プラットフォームの信頼性を確立し, ブランド価値を引き上げる意味でも重要である. 要求定義フェーズ当初は, 改ざん防止, 更には研究データを共有する組織間での信頼性醸成の目的でブロックチェーン技術の適用も提案されたが, 実際の運用範囲が限定されている中では, 殆ど効果を持たないため, この方法は取り止め, 公証に近い概念として電子署名の付与を行う. 具体的には研究データをアップロードした際の受領確認や, 公開する際に Material Data Repository (MDR) Server で出典保証をする意味で電子署名を付与することである. 基本設計フェーズ当初, 公開鍵基盤を適用することで設計を進めて来たが, 公開鍵証明書認証局の設置に対するコストと効果のアンバランスから GNU Privacy Guard 方式の採用に変更した.

4. 暫定的な評価

前章迄に概説した当該プラットフォームは2018年度に概ねの基本設計フェーズを終え, 2019年度から本格的な実装をする予定である. 本章では基本設計フェーズを経て現構成に至る迄の制約事項や, 進化に対する見通し, また暫定評価として他研究プラットフォームとの比較の上で当該プラットフォームの位置付け, 大局的な要求に対する見解について述べる.

裾野の広い材料科学分野の研究データ管理で特記するべき点としては, 利用する実験・計測装置, 扱う手法, 試料, 並びに課題等が極めて多様であるため, 研究データの管理に向けた共通的な業務プロセスのモデル定義に大きな困難が伴うということである. 関連するオントロジや語彙等の情報資産が稚拙で発展途上にある段階では, 益々, その傾向が助長される. このため, 接続されるシステム数を少数に限定した上で, 当該プラットフォームが提供するサービスを試行的に適用, それを段階的に繰り返すことで発展させるスパイラル的な戦略を取る必要がある. これを受けて業務プロセスは, 試行適用～課題抽出～更新・最終確定の一連のサイクルを完了する迄, 粗いユースケースのみが定義され, プロセス成熟度・練度が発展途上に留まる傾向が続くことも考えられる. 領域固有の斯様な困難さの中で実施した基本設計フェーズでは, 実際, 概説した構成に至る迄, 何度も大きな要求変更と仕様変更, 構成の再設計を繰り返している. これは, 単に要求定義対象だけに起因するとは限らない. 扱う領域が広範囲に渡る程, それを実現する参加ベンダ数が増え, その責任分界点を明確化するプロジェクト管理上の困難や, 利用ソフトウエア固有の課題も指数関数的に増大する. 例えば, スパイラル的な発展を目指す上で妥当と思われ, セキュリティ隔離を容易にするワークフロー機能については, 基本設計フェーズ当初は検討されたが, 採用するソフトウエア数と依存関係の増加が逆に責任分界点の曖昧さを助長することを理由に, 結局は採用を見送り, より全体のマイクロサービス的な性格が強まった.

前述の天文学, バイオ領域等で類似の取り組みや, 構造的アナロジーを持つ Computer Integrated Manufacturing 等は凡そ二十余年に渡り段階的発展をした歴史を持つのに対して, レガシーシステム群と新規機能を統合する当該プラットフォームは, 新たなクラウド環境, IoT 環境等も考慮して全体構想を垂直立ち上げで定義の上, 各種設計を行う必要があり, 必然的に高い困難さを内包している. アナロジーを持った前例プラットフォーム群が長期間のスパイラル的発展をしている事実に基づけば, 当該プラットフォームの実装・運用も逓増的構築に依ることが, 必然的に求められる.

最後に暫定評価として, 他の研究プラットフォームと比較の上で, 当該プラットフォームの位置付けを明示した上で, 要求に対する見解を述べる. 図.6.は ACM(Association for Computing Machinery)にて発刊された Scientific Workflows の現状に関するサーベイ論文で, 帰納的に定義された典型的ワークフローの概念モデルである[4]. 図中でグレーに塗り潰した領域は当該プラットフォームが受け持つ機能領域と重なる部分であり, Data Mining 領域の大部分については表.1で説明された機械学習プログラムへのインプット形式の Data Set を生成する Machine Learning Dataset Generator 機能とその後続機能が担う. この点で Material Informatics を成熟させて行くには, 前述 API-FWK を介在させた, より多くのシステム統合を, 継続して実施していくことが必須となる. 当該プラットフォームは, その要求に応じられるだけの基盤的な要素技術と現実的なアプローチを, 潜在的に十分に内包している.

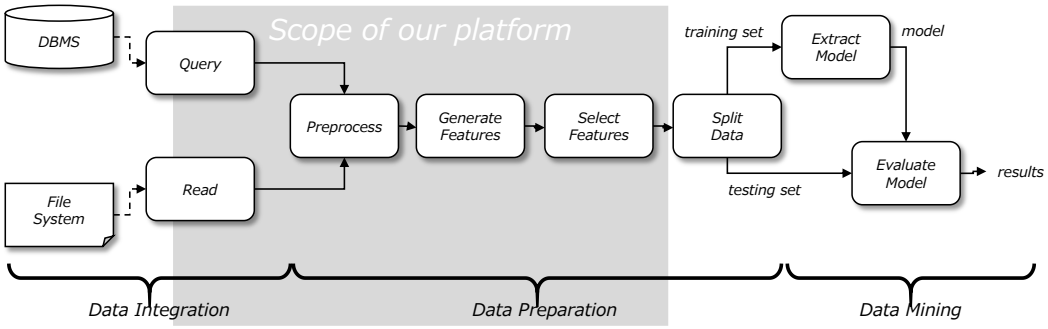

図.6. 帰納的に定義された Scientific Workflows の典型的ワークフローの概念モデル[4]

5. 結言

本稿では国立研究開発法人 物質・材料研究機構で設計開発を進める材料データプラットフォームのアーキテクチャ, 特に上位機能層の静的構造について概説した. 今後は, 基本設

計フェーズの設計解を具体的に実装することになるが, 前述の通り, 必然的にその実装・運用では, 逓増的に進めることになる. 本稿の執筆に当たり, 基本設計フェーズに携わった諸氏に感謝と伴に御礼を申し上げる.

文　献

[1]知京豊裕, ‘マテリアルインフォーマティクスの現状と課題～海外の動向と日本の挑戦’ ,情報知識学会誌, 2017 Vol.27, No.4, Pp.297-304, 2017.

[2]上島伸文, 及川勝成‘技術解説～計算材料科学・工学の最新動向’, 電気製鋼, 第87巻1号, 2016年, Pp.21-26.

[3]Y.Liu, T.Zhao, W.Ju and S.Shi, ‘Materials discovery and design using machine learning’. Journal of Materiomics, Vol.3. 2017, Pp.159-177, 2017.

[4]C.S. Liew, et al.,‘Scientific Workflows: Moving Across Paradigms’. ACM Computing Surveys, Vol. 49, No. 4, Article 66, 2016.

[5] T.J.Williams, 'A Reference Model for Computer Integrated Manufacturing from the Viewpoint of Industrial Automation', Proceddeings of 11th IFAC World Congress on Automatic Control, Tallinn, 1990 - Volume 5, Tallinn, Finland, 1990.

[6]NIMS Now, 2019.1月 号, https://www.nims.go.jp/publicity/nimsnow/vol19/hdfqf100000aoslh-att/hdfqf100000aosp0.pdf

[7]S.S.Sahoo, A.Sheth and C.Henson, ’Semantic Provenance for eScienece’,IEEE Internet Computing, July/August, 2008, Pp.46-54.

[8]N.Rozanski and E. Woods, ‘Software Systems Architecture: working with stakeholders using viewpoints and perspectives’, Pearson Education Inc. 2005.（邦訳：榊原彰監訳, 牧野祐子訳,‘システムアーキテクチャ構築の原理～IT アーキテクトがもつべき３つの思考’,翔泳社,2008）

[9] https://osf.io/

[10] https://www.apereo.org/projects/cas

[11] https://jsonapi.org/format/#status