

**運営機構プレゼン資料**

**マテリアル先端リサーチインフラ  
第1回データ連携基盤委員会**

**2021年5月12日**

**運営機構長**

## データ構造化の進め方

- ① **装置・計測メタデータは自動抽出**：装置/計測ファイルに記述されている実験パラメータを自動的に抽出して登録（以下、「自動翻訳」と称する）
- ② **ユーザー入力はテンプレート化**：ユーザーに入力してもらうしかない試料情報は、材料分野毎にある程度統制するためにテンプレートを活用して登録
- ③ **後から付随するデータを加える仕組み**：後から別に分析・評価したデータをユーザーに追加してもらう仕組みを導入。サプリメントファイルとして添付。様々なタイプのデータが予想されることから、データ構造化は行なわない。

### 先端リサーチインフラ事業の構想から、データ構造化はハブ機関の役割

- データ構造化を共通化・整理するための「**データ構造化検討委員会**」、実務作業のための「**データ構造化WG**」を運営機構に設置する
- 準備作業を**6月末**までに終えたい

次のページ以降、各項目ごとに説明します。

# 装置メタデータ自動翻訳に関する進め方

## 《ハブ機関の役割》

NIMSの実績：18機種

- 内容： 翻訳スクリプトの作成.
- 目安： 1ハブで2021年度2機種分（6ハブ×2機種＝12機種を目標に）
- 期日： 今年度末（2022/3末）

## 《センターハブ機関の役割》

- 2021年度中にデータ構造化のためのシステムを開発
- 2022年4月からシステムに翻訳スクリプトを実装開始
- 2週間/1機種を目安に秋までに12機種の実装完了を目標

## 補足説明：

- 自動翻訳：装置のファイルに書かれているメタデータをプログラムで抽出．装置メーカーに情報提供を協力して頂く必要あり．（詳細5ページ）
- プログラム言語：python（ハブ機関でデータ構造化コードを共有することを想定．基本言語を統一することで共有や学び合いがしやすくなる）

2022年度は当該12種類についてIoT（P14以降の別資料）を介したデータ転送試験

## データ構造化実績としての装置リスト（18機種）

	装置一般名称	装置メーカー	機種
1	電子プローブ微小部分分析	日本電子	XM15003-0003
	電子プローブ微小部分分析	日本電子	XM809001-0001
2	走査型オージェ電子顕微鏡	日本電子	JAMP-9500F
3	グロー放電質量分析法	VG	VG9000
4	FIB-SEM	FEI	Helios G4UX & 650
5	高精度自動四探針シート抵抗計測器	ナプソン	Model RT-3000(S)/RG-80N
6	薄膜X線回折装置	BRUKER	D8-GADDS
7	蛍光X線組成分析装置	島津製作所	uEDX1400
8	DART-MS 質量分析	島津製作所	MSIR03
9	高速分注装置	ガイガー	CERTUS
10	X線回折装置	リガク	SmartLab
11	大気中光電子分光装置	理研計器	AC-5
12	硬X線光電子分光法（SPring-8 BL15XU）	Scienta Omicron, Inc.	HEA-4MS249
13	X線光電子分光法（SPring-8 BL23）	Scienta Omicron, Inc.	EA125（VAMASフォーマットで出力されたデータを構造化）
14	X線光電子分光法	Thermo Fisher Scientific	Sigma Probe
15	走査透過電子顕微鏡	Thermo Fisher Scientific	Titan
16	X線光電子分光法、反射電子エネルギー損失分光法	ULVAC-PHI	ESCA5500
17	X線光電子分光法	ULVAC-PHI	Quantum2000
	X線光電子分光法	ULVAC-PHI	QuanteraSXM
18	X線吸収分光法（SPring-8 BL14B2）	SPring-8 JASRI	検出系はJASRIによって構築
作業中	硬X線光電子分光法（SPring-8 BL46XU）	Scienta Omicron, Inc.	R4000

# データ構造化の作業のながれ

## ① 装置メーカーへの協力依頼

- バイナリー変換ツールの提供
- 以下の②・③の作業への協力

運営機構主導でメーカー  
協力を取り付ける

経済産業省、ユーザー産業界の協力も取り付け  
オールジャパンの事業として装置メーカー・装置機種を拡大

NIMSでの協力依頼取り付け状況（※）

- XRD: RIGAKU
- XPS: ULVAC-PHI
- AES: JEOL, ULVAC-PHI

※今後、随時拡大予定

- ② 装置メーカー協力体制のもと翻訳すべきメタデータ、数値データを抽出
- ③ メタデータを記述する装置メーカー公認の語彙の確認
- ④ データ様式の確認と自動翻訳スクリプトの開発

ハブ機関の役割

## 装置メタデータ自動翻訳に関する進め方[2]

### 《アクションプラン》

FY2021/7月～： 各ハブで実施。（3～4ヶ月で1機種）

- ・ 6月末までに**体制や人材（※）**のご準備をお願いします。
- ・ 類型や標準作業書はセンターハブから提供予定。
- ・ 自動翻訳コーディングに関する勉強会開催も想定

FY2021/6月末まで： 装置リスト選定、優先順位と担当分を配分、それをもとにデータ構造化をすすめる順番を策定。

**※体制や人材に求められるもの：** 当該装置に精通している。データ再利用の観点で必要なデータ項目を特定できる。プログラミングに精通している。詳しくは13ページ参考。

→ 自動翻訳データ項目についての共通化は**データ構造化検討委員会**にて整理

→ 装置リスト集約、優先順位確認、担当配分等を**データ構造化WG**で作業

データ構造化検討委員会・WGについては後述

## ユーザー入力統制のためのテンプレートに関する進め方

- ユーザーに入力してもらったデータ登録の様式を作成。装置・計測ファイルには記述されていない試料情報などの記入欄を設計。
- データ項目の設計が重要。対象材料・分野に応じて複数のテンプレートの設計が必要（テンプレート設計と呼びます）
- 設計されたテンプレートをコーディングし、データ構造化システムのGUIを構成する主要な要素とします（テンプレートコーディングと呼びます）
- テンプレートコーディング言語： HTML

### 《ハブ機関の役割》

- 内容： 機構長が認定したテンプレート設計についてコーディング作業。
- 目安： 自動翻訳と対となる2機種 （6機関×2テンプレート=12を目安）
- 期日： 今年度末（2022/3末）

### 《センターハブ機関の役割》

- テンプレート実装： 2022年4月以降にセンターハブがデータ構造化システムへ実装

テンプレートに関する用語の整理：

- ① テンプレート設計：ユーザーが入力するためのデータ項目を設計すること
- ② テンプレートコーディング：GUIを構成する要素として、テンプレートをHTMLで書き下すこと
- ③ テンプレート実装：コーディングしたテンプレートをデータ構造化システムに実装すること

# ユーザー入力統制のためのテンプレートに関する進め方 [2]

## 《課題》

- テンプレートは、材料種・材料分野ごとに多様化する
- 無制限にテンプレートが増えると、構造化の役目を果たさなくなる
- ある程度の共通化、整理が常に必要

## 《進め方の工夫ポイント1：データ構造化検討委員会によるテンプレートの整理》

- テンプレートの共通化・整理を検討するためのデータ構造化検討委員会を設ける。
- 当該委員会にてテンプレートの候補を比較検討し、先端リサーチインフラ共通のテンプレートとして機構長に提案する。
- 提案を受けて機構長が共通テンプレートを認定する。

## 《進め方の工夫ポイント2：データ構造化WGでのテンプレート設計作業》

- テンプレート設計の実務を実施するためのデータ構造化WGを設ける。
- 機構長の指示のもと、NIMS提供のテンプレート設計をもとに検討し、不足分を担当を決めて設計



# データ構造化を進めるための委員会・WG

## データ構造化検討委員会：

- データ構造化に関する事項を機構長の指示のもと検討する
- メンバー：各ハブ機関1名、有識者若干名（データ利活用・ユーザーの観点）
- 検討内容：
  - ユーザー入力を統制するためのテンプレート設計を整理し、先端リサーチインフラで共通して使用するものを機構長に提案する
  - 装置から自動翻訳で抽出する項目を整理して機構長に提案する

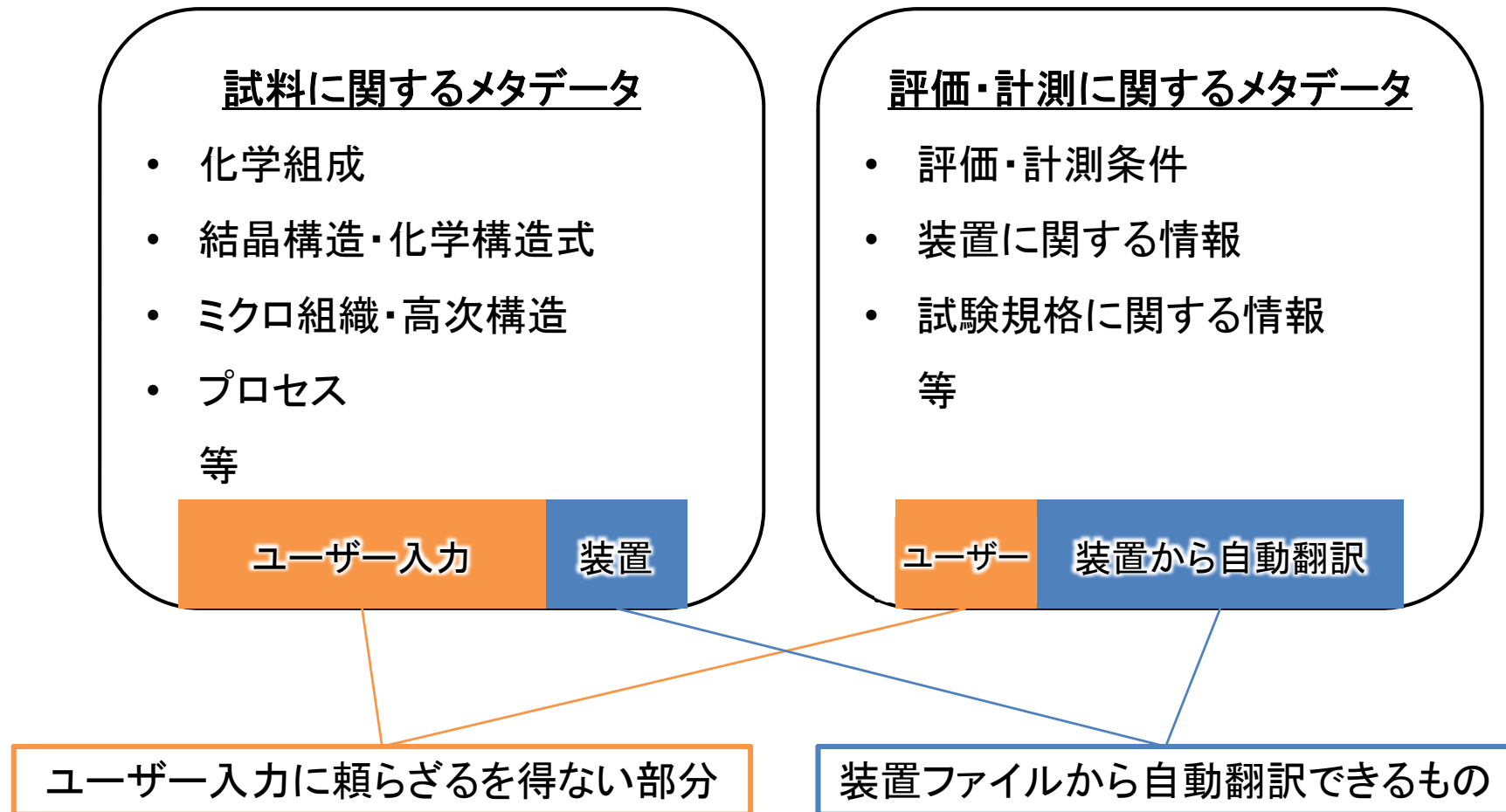
## データ構造化WG：

- データ構造化に関する実務を機構長の指示のもと実施する
- メンバー：ハブ機関・スポーク機関からの有志
- 作業内容：
  - ユーザー入力を統制するためのテンプレート候補を設計する
  - 装置から自動翻訳で抽出する項目を共通化する
  - 装置リスト集約、優先順位確認、担当機関調整を実施

## 付随データに関する進め方

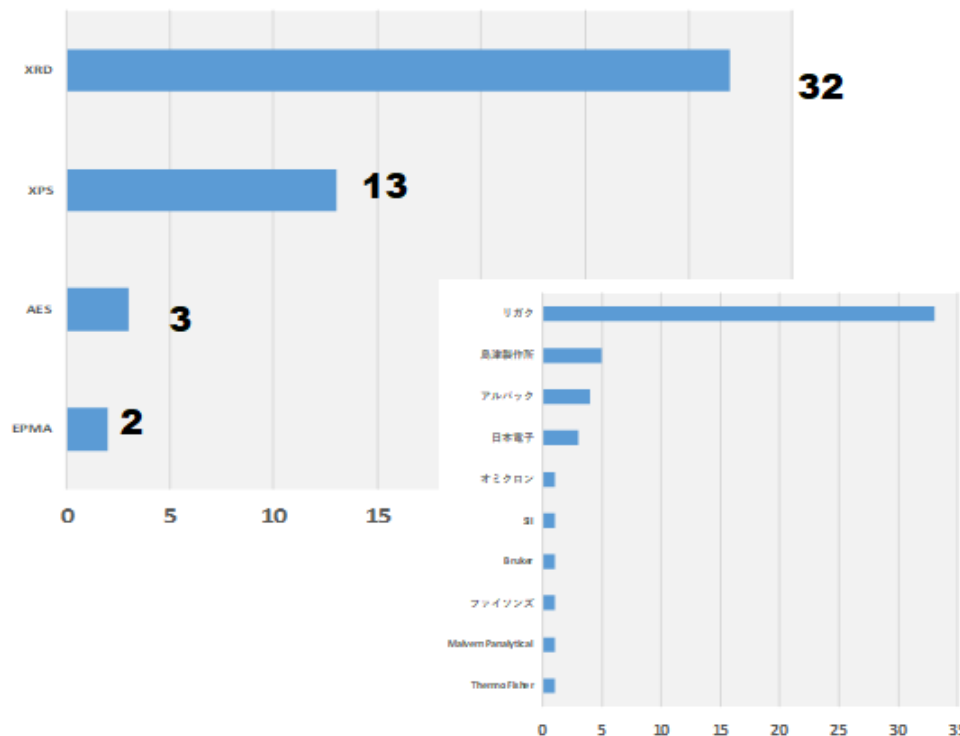
- 後からユーザーが分析・評価したデータを追加できるようにする
- 構造化データの参考となる実験に係るファイルやドキュメントを「**後から**」アップロード.
- 登録先はシステムの「付随データセット」という保存場所
- アップロードするファイルの種類は、ユーザーの裁量に任せる. 例えば、付随データを説明するテキストファイルを添付することも可能。
- ただし、情報セキュリティの関係上、アップロードできるファイルの種類には制限をかける予定. ex) txt, csv, png, jpeg, マークダウン文書
- ユーザーにどのように付随データを出していただくかについては運営機構でインセンティブ等の仕組み考える必要あり. また、その実現のための規約等については運営WG（仮称）で議論。

## 参考：データ構造の設計について



## 参考 各機関で重複する装置の例

- 機関が異なっても、汎用的な装置(XRD)やメーカーは同種である装置も多い。  
→ 共通する機器のpythonコードはコードシェアすることで運用効率を高められる



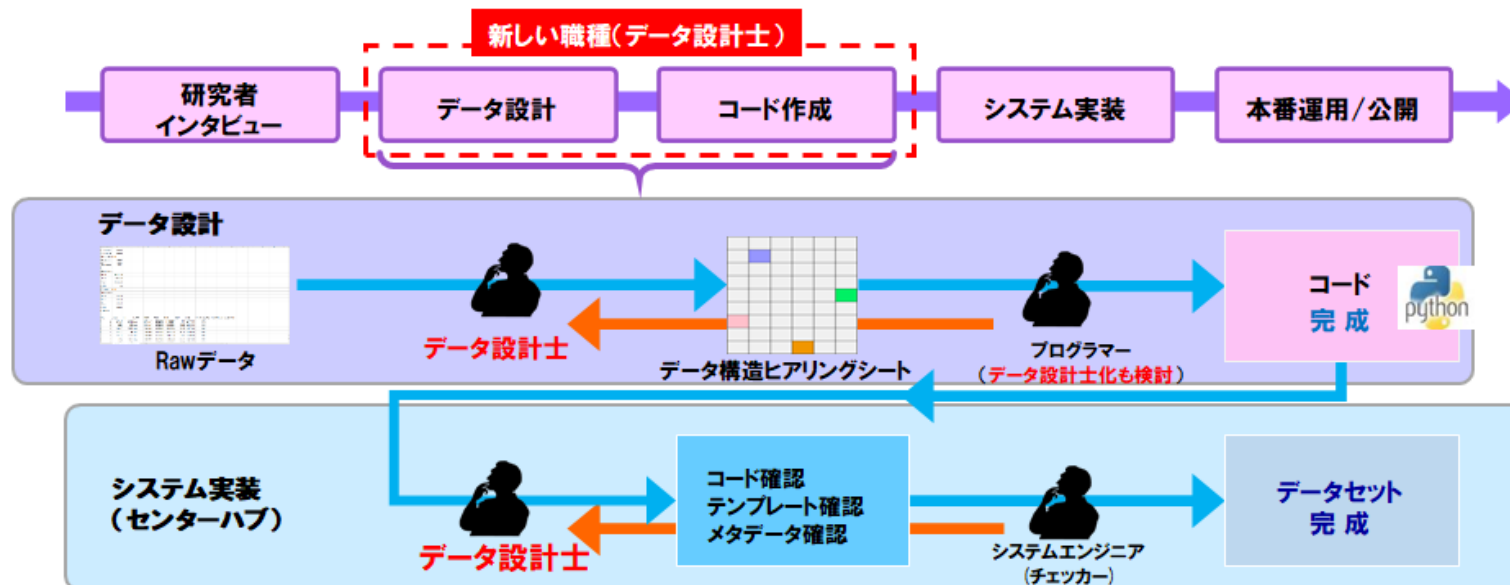
拠点	PF	装置	メーカー名	台数
NIMS	微細構造解析	AES	アルバック	1
北海道大学	微細構造解析	AES	日本電子	1
JAIST	分子物質合成	AES	アルバック	1
信州大学	分子物質合成	EPMA	島津製作所	1
JAIST	分子物質合成	EPMA	日本電子	1
名古屋大学	微細加工	XPS	thermofischer	1
AIST	微細加工	XPS	島津製作所	1
名古屋大学	微細加工	XPS	島津製作所	1
広島大学	微細加工	XPS	島津製作所	1
東京大学	微細構造解析	XPS	アルバック	1
AIST	微細構造解析	XPS	島津	1
JAEA	微細構造解析	XPS	放射光設備	1
NIMS	微細構造解析	XPS	放射光設備	1
北海道大学	微細構造解析	XPS	日本電子	1
九州大学	分子物質合成	XPS	アルバック	1
分子研	分子物質合成	XPS	オミクロン	1
JAIST	分子物質合成	XPS	島津製作所	1
JAIST	分子物質合成	XPS	ファイソズインスツルメンツ	1
AIST	微細加工	XRD	リガク	2
京都大学	微細加工	XRD	リガク	1
豊田工業大学	微細加工	XRD	リガク	1
名古屋大学	微細加工	XRD	リガク	1
広島大学	微細加工	XRD	リガク	1
JAEA	微細構造解析	XRD	放射光設備	1
NIMS	微細構造解析	XRD	放射光設備	1
東京大学	微細構造解析	XRD	リガク	6
九州大学	分子物質合成	XRD	Bruker	1
JAIST	分子物質合成	XRD	Malyvern Panalytical	1
JAIST	分子物質合成	XRD	リガク	4
NAIST	分子物質合成	XRD	リガク	2
大阪大学	分子物質合成	XRD	リガク	1
千歳科学技術大学	分子物質合成	XRD	リガク	1
名古屋工業大学	分子物質合成	XRD	リガク	1
名古屋大学	分子物質合成	XRD	リガク	3
分子研	分子物質合成	XRD	リガク	4
AIST	微細加工	XRF	Si	1
広島大学	微細加工	XRF	リガク	1
AIST	微細構造解析	XRF	島津	1
JAEA	微細構造解析	XRF	放射光設備	2
NIMS	微細構造解析	XRF	放射光設備	1
分子研	分子物質合成	XRF	日本電子	1
NAIST	分子物質合成	XRF	リガク	1
名古屋大学	分子物質合成	XRF	リガク	1
千歳科学技術大学	分子物質合成	小核散乱	リガク	1

### 要望事項

データ設計士のプログラム(コード)のコードシェアを通じて参画機関で共有できること

## 【参考】 データアーキテクトの人材像 (装置データに関して)

- 装置管理者(研究者)とシステムエンジニア/プログラマーをつなぐ「データ設計士(データアーキテクト)」という新しい人材が必要



- ✓ 計測機器に応じたファイルフォーマットを熟知していること
- ✓ データそのもののアーキテクチャを行い、データの統合、クレンジングの指針を考えられること
- ✓ 好ましくはそのようなクレンジング・「前処理」のコードを記述できること
- ✓ メタデータなどの概念を把握し、データベースの運用においても知見を有すること
- ✓ データ連携にかんするライセンスについても知見を有すること

# 取り組み事例

# アウトライン

1. データ構造化の意味
2. データ構造化の過去の事例紹介
3. データ構造化のシステム設計の基本方針

(注意) 本資料は関係者のみの閲覧/配布として頂きますようお願い致します

# データ構造化の意味

## 1. データ科学的な見地

- ✓ 説明変数と目的変数を明示的に定める
- ✓ 説明変数は、実験のログ、グラフ、もしくはスペクトルや画像から特徴量を数値化

## 2. データベース工学的な見地

- ✓ 数値データ部はソフトウェア互換性の高い csv等の表形式のテキストデータに変換  
バイナリデータの場合は可読性のある汎用的な圧縮方式でバイナリ化する
- ✓ 実験（プロセス・計測）パラメータ部をスキーマを定めてデータベースに登録
- ✓ データの意味が直感的にわかる参照図（サムネイル）を付す

## 3. 上記に共通して関係するデータの寿命化の見地

- ✓ 将来装置メーカーのサポートが切れてもデータの価値が失われないよう判読性の高い語彙や形式を使ってデータが記録される
- ◆ できるだけ人の手間を省いて構造化するための自動化の工夫を行う
- ◆ 装置メーカーの協力も得て自動化のワークフローを製作し、共用する。



# NIMSにおけるデータ構造化の事例

# 実験データのデータベース化ワークフローの自動化 によって作成したデータセットの例

The screenshot displays the Research Data Express (RDE) Viewer interface. At the top, there is a header with the RDE logo (Research Data Express) on the left, a search bar with the text "Search", and a user profile "hyoshi" on the right. Below the header is a blue banner with the text "Research Data Express (RDE) Viewer".

Below the banner, there is a search filter section with the text "Filter by Dataset\_ID, Title, Description, Keyword" and a "Clear" button. Below this is a "Filter By:" dropdown menu.

The main content area shows a list of datasets. The first dataset is "AES Reference Spectral Set" with Dataset\_ID: 56338950. It is described as "JAMP9500Fによる参照オーグեսペクト..." and "Reference AES spectral set obtained with JAM...". It has 1209 Data, edited by OGIWARA, Toshiya, and 201 View | 41 Download. The tags are "measurement", "AES", "Auger", "reference", "spectrum", "surface anal...", "quantitative ...", and "qualitative a...". It was updated on 2021-03-31 11:40:10.

The second dataset is "Monochromated EELS of LiCoO2 and related materials" with Dataset\_ID: 55679817. It is described as "Liイオン 2 次電池の正極材料として用いら..." and "EEL spectra of LiCoO2 and related materials f...". It has 10 Data, edited by KIMOTO, Koji, and 58 View | 26 Download. The tags are "measurement", "EELS", "monochrom...", and "Li battery". It was updated on 2020-08-17 11:33:20.

The third dataset is "XAFS data" with Dataset\_ID: 64567741. It is described as "XAFSソサエティデータ x線吸収分光スペ..." and "XAFS society data For development of materi...". It has 0 Data, edited by ISHII, Masashi, and 0 View | 0 Download. The tags are "measurement", "X-ray", "Spectroscopy", "Absorption", "Local struct...", and "Standard sa...". It was updated on 2021-03-09 09:33:00.

# データ登録の単位となるデータセットとは？

- 特定の目的やテーマに沿って収集されたデータ群。
  - 1) 統計処理や機械学習等にも利用しやすいようにデータが構造化された状態となっている。
  - 2) データの属性や意味がわかるデータカタログを備える。
- データセットはIoTを設置した機器単位とする。
- データセットの開設は、機器利用受付で受け付けた利用課題単位（課題番号）とする。

The screenshot displays the Research Data Express (RDE) Viewer interface. At the top, there is a search bar and a 'Filter by Dataset ID, Title, Description, Keyword' section. Below this, a 'Filter By' section is visible. The main content area shows three data sets, each with a title, description, and a 'measurement' button. The first dataset, 'AES Reference Spectral Set', is highlighted with a red dashed box and a pink label 'データセットの単位'. The second dataset is 'Monochromated EELS of LiCoO2 and related materials', and the third is 'XAFS data'. Each dataset also includes a 'View' button and a 'Download' button. The interface is clean and professional, with a blue header and a white background.

**データセットの単位**

**AES Reference Spectral Set**  
Dataset ID: 56338950  
JAMP9500Fによる参照オージェスペクトル...  
Reference AES spectral set obtained with JAMP...  
1208 Data | Edited by OGIMURA, Toshiya  
0 View | 41 Download  
measurement  
AES Auger reference spectrum  
surface anal. quantitative qualitative  
Updated: 2023-03-31 11:40:10

**Monochromated EELS of LiCoO2 and related materials**  
Dataset ID: 55679817  
リチウム二次電池の材料として用いら...  
EEL spectra of LiCoO2 and related materials f...  
10 Data | Edited by KIMOTO, Koji  
58 View | 26 Download  
measurement  
EELS monochromated battery  
Updated: 2020-08-17 11:33:20

**XAFS data**  
Dataset ID: 64567741  
XAFSソサエティデータ吸収分光スペ...  
XAFS society data for development of materi...  
0 Data | Edited by ISHII, Masamichi  
0 View | 0 Download  
measurement  
X-ray Spectroscopy Absorption  
Local structure Standard sample  
Updated: 2023-03-09 09:33:00

# データカタログについて

- 記述項目は『データセットの概要』、『データセットの諸元』、『メタデータリスト』からなる
- どのような公開方式にしたいかなどのライセンスもここで決めておく。

公開方式

## DPFC-DCS 公開データカタログ

(機能性材料研究拠点 (RCFM) ナノ電子デバイス材料グループ内限定)

作成日:2020-09-14

### ■データセット概要

データセット名	XRF Heatmap for inorganic sputter sample
概要	有機半導体の組成傾斜がある主に二元系スパッタ薄膜 (X:基板 X 方向, Y:基板 Y 方向, Z: カウント数) の二次元ヒートマップ図。
作成者	渡田貴弘
公開範囲	機能性材料研究拠点 (RCFM) ナノ電子デバイス材料グループ内限定
ライセンス	論文前, 公開前のデータにつき作成者の許可必要
言語	日本語
使用装置	島津製作所 XRF $\mu$ EDX-1400

### ■データセット諸元

データ数	XRF ( $\mu$ EDX-1400) に設置の $10^4$ より数度追加される
データの種類	XRF 二次元マッピング測定
raw データ属性	<ol style="list-style-type: none"> <li>拡張子: txt</li> <li>ファイルサイズ: 10~20KB</li> <li>欠損値: なし (特に定められていない)</li> </ol>
raw データのデータ構造	<ul style="list-style-type: none"> <li>ファイル内部は XML 構造表記がなされている。</li> <li>[FileInfo], [ResMap_Common], [ResMap_Element] の3つのブロックからなる。</li> <li>[FileInfo] は計測の書込情報の項目。[ResMap_Common] はマッピングを行った基板の座標情報。[ResMap_Element] は蛍光 X 線の計測データが格納されている。</li> <li>[ResMap_Element] の配下には、[NetCounts] のサブブロックには意味の強度、[Conc] には濃度分率に変換された百分率がしめされている。</li> </ul>
output ファイル属性	<ul style="list-style-type: none"> <li>raw データからは xrf_heatmap.py を通じて以下の6つのファイルが出力される。欠損値処理は特に起こっていない。</li> <li>→ csv 形式 →</li> <li>①ファイル名_netcount_original.csv ファイルサイズ: ~10KB</li> <li>②ファイル名_netcount_convert.csv ファイルサイズ: ~10KB</li> <li>③ファイル名_conc_original.csv ファイルサイズ: ~10KB</li> </ul>

### ■メタデータリスト

Viewer における英語語彙は出力パラメータの表記をそのまま転載している。日本語語彙は DPFC で任意に定めたものであり計測メーカーの丁来の語彙表記ではないことに留意の上、計測パラメータの理解の参考とすること。

	出力計測パラメータ	日本語語彙	英語語彙	単位
1	[FileInfo]			
2	SampleName	サンプル名	SampleName	
3	AnalType	解析手法	AnalType	
4	Comment	コメント	Comment	
5	ParFLN	解析グループ	ParFLN	
6	MeasDateStr	測定日時	MeasDateStr	
7	Operator	オペレーター	Operator	
8				
9	[ResMap_Common]			
10	StartX	開始位置 X 座標	StartX	mm
11	StartY	開始位置 Y 座標	StartY	mm
12	EndX	終了位置 X 座標	EndX	mm
13	EndY	終了位置 Y 座標	EndY	mm
14	PosZ	試料高さ	PosZ	
15	PointX	X 方向ポイント数	PointX	
16	PointY	Y 方向ポイント数	PointY	
17	Step	ステップ数	Step	
18				
19	[ResMap_Element]			
20	CompName	測定サンプル名	CompName	
21	ProcType	計測種類	ProcType	
22	Elem	観測元素種	Elem	
23	Line	観測 X 線名	Line	

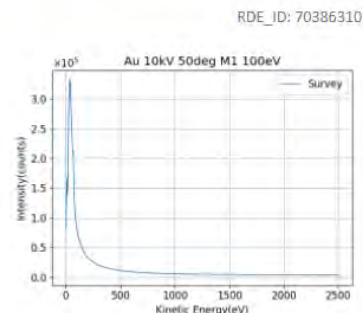
データ利用者は、実際のデータを見なくてもどのようなデータがあるかが分かります

# 標準スペクトル（電子分光）の登録事例

## データセット内のデータのリスト(Gallery形式)の例

A なお Tree表示も可能

Au-Angle-data135.A.zip

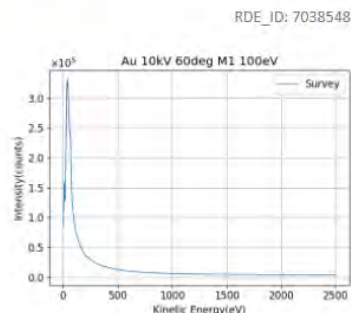


電子線加速電圧：10kV、電子線入射角：  
法線から50°、分光器のモード：M1  
(100eV)、エネルギー分解能：0.14%

分光法 オージェ電... 角度分解測... Au

Registered: 2021-04-23 14:37:28

Au-Angle-data134.A.zip

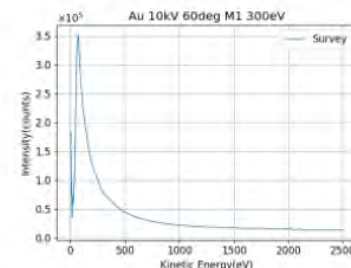


電子線加速電圧：10kV、電子線入射角：  
法線から60°、分光器のモード：M1  
(100eV)、エネルギー分解能：0.14%

分光法 オージェ電... 角度分解測... Au

Registered: 2021-04-23 14:36:06

RDE\_ID: 70384656

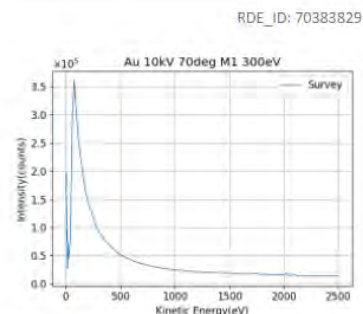


電子線加速電圧：10kV、電子線入射角：  
法線から60°、分光器のモード：M1  
(300eV)、エネルギー分解能：0.21%

分光法 オージェ電... 角度分解測... Au

Registered: 2021-04-23 14:34:44

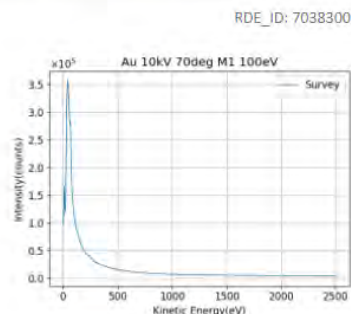
Au-Angle-data132.A.zip



電子線加速電圧：10kV、電子線入射角：  
法線から70°、分光器のモード：M1  
(300eV)、エネルギー分解能：0.21%

分光法 オージェ電... 角度分解測... Au

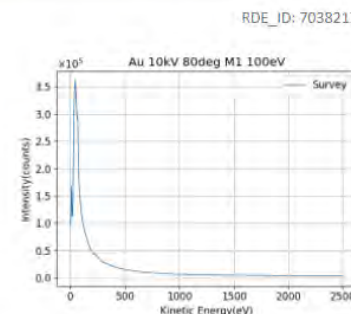
Au-Angle-data131.A.zip



電子線加速電圧：10kV、電子線入射角：  
法線から70°、分光器のモード：M1  
(100eV)、エネルギー分解能：0.14%

分光法 オージェ電... 角度分解測... Au

Au-Angle-data130.A.zip



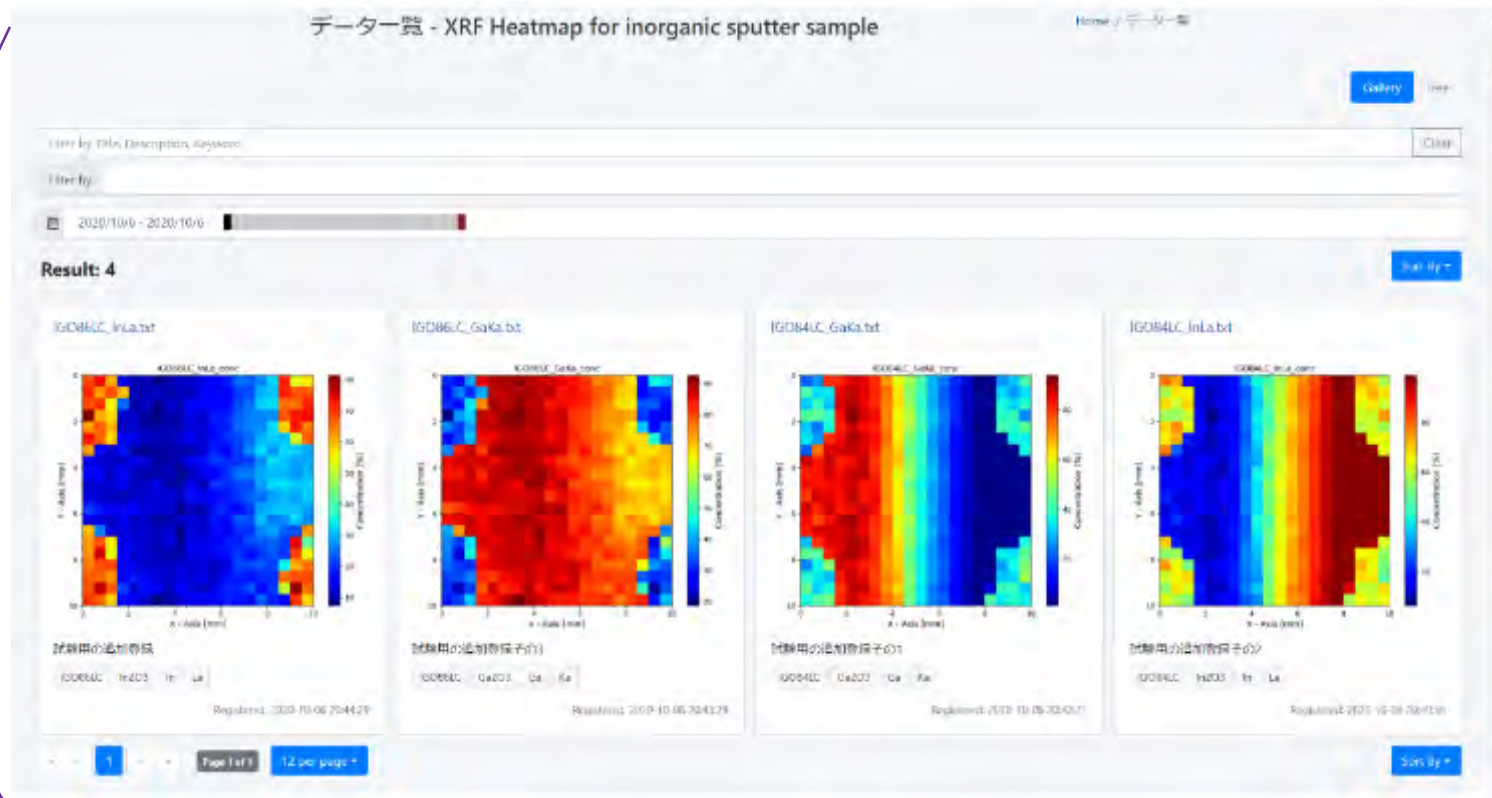
電子線加速電圧：10kV、電子線入射角：  
法線から80°、分光器のモード：M1  
(100eV)、エネルギー分解能：0.14%

分光法 オージェ電... 角度分解測... Au



# 組成傾斜試料の蛍光 X 線画像の組成二次元マッピング (Gallery形式) の例

- 同一のフォーマットで系統的に画像が蓄積されるため、長期的な傾向把握も可能となる



# 標準スペクトル（電子分光）の登録事例

## データの表示画面のトップ

Au-Angle-data135.A.zip

電子線加速電圧：10kV、電子線入射角：法線から50°、分光器のモード：M1 (100eV)、エネルギー分解能：0.14%

View: 5 Download: 3

RDE\_ID: 70386310

Registered on 2021-04-23 14:37:28

Created by OGIWARA, Toshiya

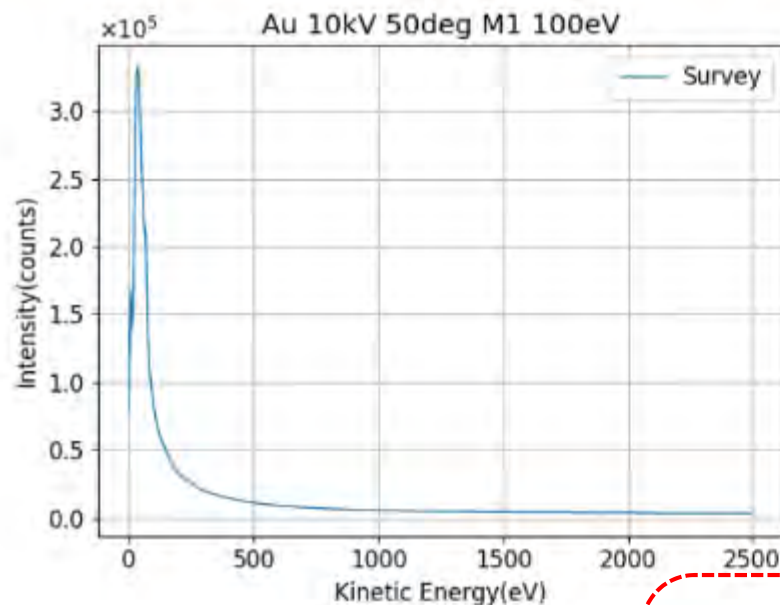
Contact: KAWADA, Satoshi

キーワード

Basic Information

数値データのグラフ化

分類 (Taxonomy) のメタデータ



Taxonomy Meta

Method category

分光法

Method sub-category

オージェ電子分光法

Measurement method

角度分解測定

General name

Au

Method category

分光法

Method sub-category

オージェ電子分光法

Measurement method

角度分解測定

General name

Au

Download

データのダウンロード

Raw Data:

Raw Data ☐

Numerical Data:

CSV(FNO) ☐

Metadata:

Metadata ☐


Download ☐

Download All ☐


# 標準スペクトル（電子分光）の登録事例

付随するデータ（試料の外観写真やSEM画像）


Annotation Files



Au-Photo1.png



Au-Photo1.png



Au-Photo2.png

Download ☐



# 標準スペクトル（電子分光）の登録事例

## 実験パラメータ（試料情報を含む基本情報）

Basic Term		
Common Term		
File Name	Description	Data Origin
Au-Angle-data135-A.zip	電子線加速電圧：10kV、電子線入射角：法線から50°、分光器のモード：M1 (100eV)、エネルギー分解能：0.14%	experiments
Technical Category	Contact Person	Creator
measurement	KAWADA, Satoshi	OGIWARA, Toshiya
Instrument Term		
Instrument_name	maker	NIMS_management_number
MAS_AES_JAMP_9500F	日本電子	90007009
instrumentation_department	Method Category	
材料分析ステーション	分光法(オージェ電子分光法)	
Specimen Term		
Specimen_ID	Description	General name
7440-57-5	高純度化学から購入した板状試料	Au
Measurement Term		
Method category	Method sub-category	Analysis field
分光法	オージェ電子分光法	定性分析, 定量分析
Measurement environment	Measurement date	Instrumentation site
真空中	2021-04-21	NIMS千現地区 精密計測実験棟Room#111

試料情報

# 標準スペクトル（電子分光）の登録事例

## 実験パラメータ（実験の再現性の確保に必要なパラメータ）

Primary Term		
Analyser detection solid angle[sr]	Analyser_type	Measurement method
31.5	CHA	角度分解測定
Operator identifier	Operation Date-time[Year]	Operation Date-time[Month]
Ogiwara	2021	04
Operation Date-time[Day]	Operation Date-time[Hour]	Operation Date-time[Minute]
21	17	20
Operation Date-time[Second]	Probe energy[keV]	Probe current[A]
16	10.00	$1.013 \times 10^{-8}$
Analyzer Mode	Analyser pass energy[eV]	Abscissa start[eV]
CAE	100	0.00
Abscissa end[eV]	Abscissa increment[eV]	Collection time[ms]
2500.00	1.00	30
Total Acquisition Number	Probe scan mode	Probe diameter[um]
200	Spot	10
Upper left x coordinate	Upper left y coordinate	Lower right x coordinate
0	0	63
Lower right y coordinate	Probe polar angle to sample normal[deg]	Comment
-63	50.0	Au 10kV 50deg M1 100eV
Neutralization active mode	Data type	Analysis chamber pressure when measurement finished[Pa]
inactive	Wide	$1.80 \times 10^{-7}$

# 標準スペクトル（電子分光）の登録事例

メタデータは、内部的には XMLデータベースとして記録されているが、ユーザは 以下のような csvファイルとしてダウンロードする

Primary_Measurement	Data_size	データサイズ	byte	271872
Primary_Measurement	File name	ファイル名		Image Of EELS 300kV monoTEM O-K
Primary_Measurement	Origin of spectrum intensity	スペクトル強度の原点		0
Primary_Measurement	Scale of spectrum intensity	スペクトル強度のスケール		0.22111663
Primary_Measurement	Unit of spectrum intensity	スペクトル強度の単位		e-
Primary_Measurement	Energy dispersion per channel	エネルギー分散	eV	0.25
Primary_Measurement	Number of channels	チャンネル数	pixel	2048
Primary_Measurement	Dynamic range of spectrum intensity	スペクトル強度のダイナミックレンジ	byte	4
Primary_Measurement	Detector	検出器		US1000FTXP 1
Primary_Measurement	Binning	ビニング		{1,1}
Primary_Measurement	Total integration time	総計測時間	s	200
Primary_Measurement	Number of frames	積算回数		200
Primary_Measurement	Measurement starting time	測定開始時刻		16:39:15
Primary_Measurement	Spectrometer entrance aperture	分光器入射絞り	mm	2.5
Primary_Measurement	Drift tube enable	ドリフトチューブ設定		TRUE
Primary_Measurement	Drift tube voltage	ドリフトチューブ電圧	V	520
Primary_Measurement	Magnetic sector setting	磁場セクター設定	eV	520
Primary_Measurement	High tension offset voltage	加速電圧オフセット電圧	V	0
Primary_Measurement	High tension offset enable	加速電圧オフセット設定		1
Primary_Measurement	EELS spectrometer name	EELS装置名		GIF Quantum ERS
Primary_Measurement	GMS version for acquisition	データ取得GMSバージョン		2.32.888.0
Primary_Measurement	GMS version for saving	データ保存GMSバージョン		2.32.888.0
Primary_Measurement	Magnification	倍率		1300
Primary_Measurement	Acceleration voltage	加速電圧	V	300000
Primary_Measurement	Post-specimen lens setting	結像系レンズ設定		DIFFRACTION
Primary_Measurement	Pre-specimen lens setting	照射系レンズ設定		TEM MICROPROBE
Primary_Measurement	Microscope information	電子顕微鏡情報		FEI Tecnai Remote
Primary_Measurement	Microscope operation mode	電子顕微鏡モード		DIFFRACTION
Primary_Measurement	STEM camera length	STEMカメラ長	mm	0
Primary_Measurement	File name2	ファイル名2		EELS 300kV monoTEM O-KandNi-L (

# 装置が出力するデータファイルから実験パラメータを自動抽出する際の課題

- (1) 装置が出力するデータファイルに埋め込まれた実験パラメータを記述する語彙が、装置メーカー固有の略語で書かれており、意味が不明。
- (2) 装置が出力する（装置メーカー固有の形式をもつ）バイナリデータのファイルに埋め込まれた実験パラメータは豊富であるが、バイナリデータの内部構造が公開されていないため実験パラメータが読み取れない。  
装置がもつバイナリデータのエクスポート機能を使って変換されるテキストファイルには、ごく一部の実験パラメータしか含まれていない。
- (3) 装置が出力するデータファイルに埋め込まれた実験パラメータの表記で、値はあるものの、値に対応するキー名がなく、物理単位が書かれていない。
- (4) 装置が出力する（装置メーカー固有の形式をもつ）データファイルが、以下の例のような複雑な内部構造をもっており、装置メーカーの説明無くファイルを見ただけではデータに埋め込まれた実験パラメータを正確に読み解けない。
  - 値と物理単位の記載位置が大きく離れている
  - データファイルの中にデータの内部構造を読み解く制御変数が存在する
- (5) 装置が出力するデータが複数個のファイルに分割されており、データに埋め込まれた実験パラメータの情報が分散している。

# 実験装置が出力するデータの構造化のワークフロー

## 装置メーカー固有の生データの バイナリファイル

## 装置メーカーが提供する変換プログラム

語彙が判読化された後に  
key-value形式のXML DBと  
して記録されたメタデータ

汎用的なテキスト/バイナリのファイル

メタデータや数値データを抽出し、  
語彙やフォーマットを判読化する  
python スクリプトを製作し共用  
する

判読化されたデータ

データをダウンロードしたユーザーによる  
フォーマットのカスタマイズを想定した  
Schema on Read 型のデータ登録

変換後の数値データ  
(ソフトウェア互換性や  
人による判読性に優れる)

[illegible]

```

import os.path
import csv
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.ticker import ScalarFormatter

parser = argparse.ArgumentParser()
parser.add_argument("file_path")
options = parser.parse_args()
readfile = options.file_path
#readfile = "50P159.109.csv"
name, ext = os.path.splitext(readfile)

with open(readfile, 'r') as f:
    reader = csv.reader(f)
    index = 1
    count = 0
    flag = False

```



# NIMS内における加工データの登録の事例

## FIB-SEM複合装置

### FEI Helios G4UX & 650

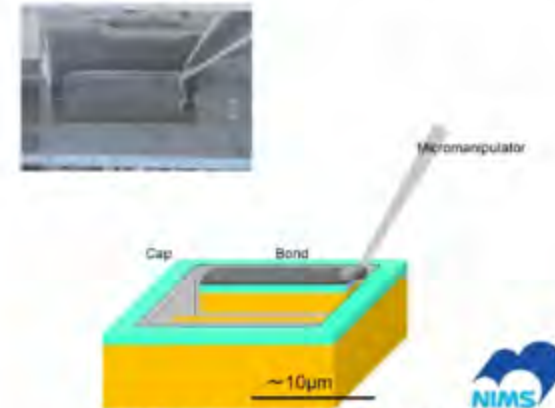


FEI Helios G4UX

走査型電子顕微鏡(SEM)と集束イオンビーム(FIB)とデポジション、マイクロマニピュレーターによりのTEM試料作製やアトムプローブ試料作製に加え、Spectroscopy)分析、EBSD(Electron BackScatter

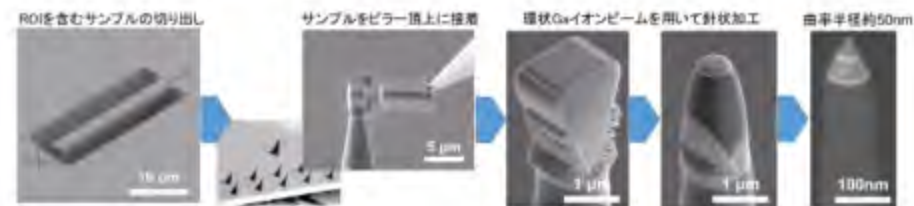
## FIB-SEM複合装置を用いたTEM試料作製の例

FIB-SEM複合装置を用いたマイクロサンプリング法により、試料の任意領域からTEM試料を作製することができます。結晶粒界、複相組織の異相界面、多層膜の界面、デバイスの任意領域のTEM観察が可能です。以下に任意領域からTEM試料を作製する一例をアニメーションで示しています。



Copyright © Magnetic Materials Unit, NIMS.

## FIB-SEM複合装置を用いた3DAP試料作製手順



- ✓ イオンビームの加速電圧，電流値，および加工マスク条件で形状をコントロール
- ✓ 素材による最適な加工条件やプロセス時間などの制御は匠の技になっている

# 加工データのユーザー入力テンプレート例

- WEB画面の感覚で触れる非ネットワーク環境でも動作する **Local-HTML** の**実験記録シート**.
- IoTでデータと一緒に送付：**加工形状（TIFF）**と**加工条件**が一体となって蓄積.

FIB加工記録（3次元アトムプローブ向け）

FIB 20200911

① 記入年月日 年/月/日

② NIMS user ID 選択してください

③ データセット名 FIB for 3DAP

④ 責任者(Responsible Person) 大久保 良雄

⑤ 計測の目的・説明（任意） 加工材料に起因するダメージ

⑥ 試料名（必須） 20201024FIB加工材料の加工条件が不明なため

⑦ 組成（任意） NIMSが提供しているデータベースを参照

⑧ 試料の説明（任意） 外部調査 参照 何も記入が不要

固有情報

加速電圧 [keV]	30	keV
束流値 [pA]		pA
マスク内径 [um]	1.5	um
マスク外径 [um]	5.0	um
加工時間 [sec]（自由記載）		sec

ファイル情報

ファイルを選択 選択されていません

## 基本情報

- ① 記入年月日
- ② データ投入者 NIMS-ID
- ③ データセット名
- ④ データ責任者

## サンプル情報

- ⑤ 計測の目的・説明
- ⑥ 試料名（番号）
- ⑦ 試料一般物質材料名
- ⑧ 試料の説明（option）

## カスタム項目（固有情報）

- ・加工のマスクの選択条件
- ・イオンビームの設定条件

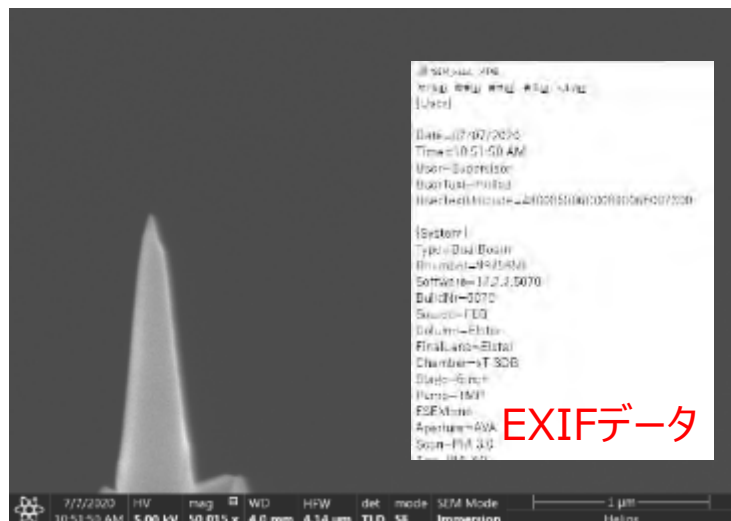
これまで作業者の**感と経験**によっていた項目

# 加工データ登録時の生データからのメタデータの自動抽出の事例

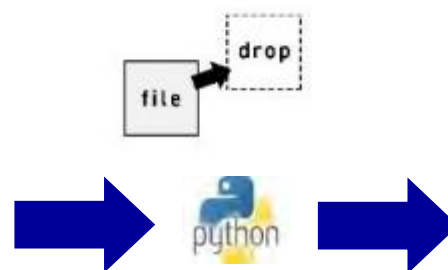
**要望**： FIBによる試験片の加工条件がとらえられていなかった。加工に必要なマスク条件も記録したい。

**適用**： TIFFファイルに埋め込まれているEXIF（メタデータ）から加工条件に重要なメタデータのみを自動抽出。

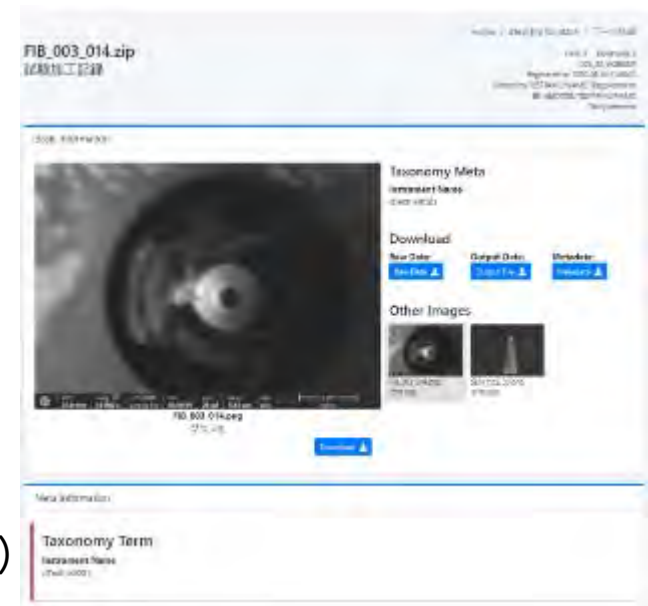
## rawデータの種類（SEMのTIFFファイル）



## Rawデータ データ送付状



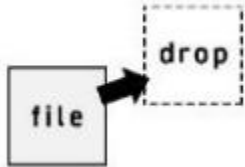















- EXIFデータの読み出し
- 指定の測定条件の抽出
- 記録台帳化（データセット）





# 超電導線材開発のデータ設計の事例

- プロセス工程ごとのプロセス条件のデータ構造化 → 説明変数（データ探索空間）の作成
- I-V特性のデータ構造化 → 目的変数

課題番号 (データセット名)	線材名	実験ID	説明変数 (特徴量)		目的変数
			冷間加工	熱処理 (温度, 時間, ステップ)	I-V特性 (Jc)
MgB2の検討	MgB2	001			
	MgB2	002			
	MgB2	003			
REBCOの検討	REBCO	004			
	REBCO	005			
	REBCO	006			

1. プロセスのログデータをデータ構造化 → ログデータから特徴量を定める
2. 機械学習モデルでJc推定モデルを作成する
3. プロセス条件をベイズ最適化で実験計画する

# 超電導線材I-V特性のユーザー入力テンプレート例

- WEB画面の感覚で触れる非ネットワーク環境でも動作するLocal-HTMLの実験記録シート.
- データと一緒に送付： 固有情報と一緒にデータベースに記録・蓄積される.

V.V. v202103-0

## 超電導 I-V特性評価 Ic-B(T)用（磁場印加角度固定）

- ①
- ②
- ③
- ④
- ⑤
- ⑥
- ⑦
- ⑧

記入年月日	<input type="text"/>	<input type="text"/>	<input type="button" value="日"/>
NIMS user ID	<input type="text" value="選択してください"/>		
データセット名	<input type="text" value="I-V Characteristic for Superconductor"/>		
タリ責任者(responsible Person)	<input type="text" value="北本朝品"/>		
登録開始の日時・設備(任意)	<input type="text" value="登録時間: 2021年3月10日"/>		
材料名(任意)	<input type="text" value="MgB2/CuO系で覆われている、1-1' スーパーヘテロ構造のMgB2/CuO系"/>		
組成(任意)	<input type="text" value="Hf/NbとNbと混合した、Cuと混合した化合物"/>		
試料の説明(任意)	<input type="text" value="手製、両面コーティング済"/>		

---

**共有情報**

磁場印加角度	<input type="text"/>	degree
線材の断面積	<input type="text"/>	mm <sup>2</sup>
銅化層コイルの断面積	<input type="text"/>	mm <sup>2</sup>
線材形状	<input type="text" value="テープ"/>	

ファイル情報

選択されたファイル

Submit

ファイルをデータ送付状とリンクさせるボタン

## 基本情報

- ① 記入年月日
- ② データ投入者 NIMS-ID
- ③ データセット名
- ④ データ責任者

## サンプル情報

- ⑤ 計測の目的・説明  
⑥ 試料名（番号）  
⑦ 試料一般物質材料名  
⑧ 試料の説明（option）

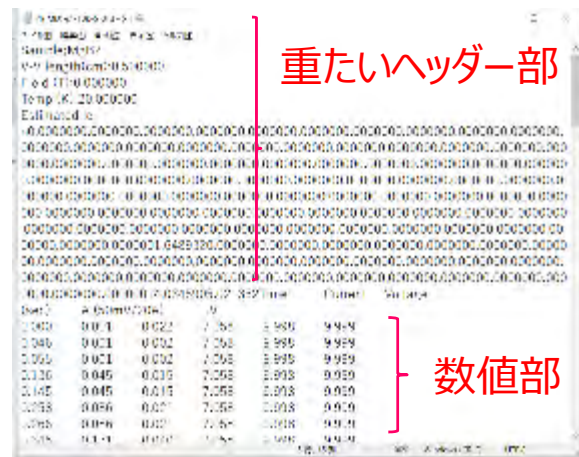
## カスタム項目（固有情報）

- ・磁場印加角度
- ・線材の断面積（ $I \rightarrow J$ への変換のとき）
- ・コアの断面積
- ・線材形状（選択式）

# 超電導線材におけるI-V特性データの構造化ならびにデータセットの構成

rawデータの種類 (テキストファイル)

重たいヘッダー部



数値部

Over	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP	EQ	ER	ES	ET	EU	EV	EW	EX	EY	EZ	FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL	FM	FN	FO	FP	FQ	FR	FS	FT	FU	FV	FW	FX	FY	FZ	GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY	GZ	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM	HN	HO	HP	HQ	HR	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP	IQ	IR	IS	IT	IU	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	IJ	JK	KL	KM	KN	KO	KP	KQ	KR	KS	KT	KU	KV	KW	KX	KY	KZ	LA	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO	LP	LQ	LR	LS	LT	LU	LV	LW	LX	LY	LZ	MA	MB	MC	MD	ME	MF	MG	MH	MI	MJ	MK	ML	MM	MN	MO	MP	MQ	MR	MS	MT	MU	MV	MW	MX	MY	MZ	NA	NB	NC	ND	NE	NF	NG	NH	NI	NJ	NK	NL	NM	NN	NO	NP	NQ	NR	NS	NT	NU	NV	NW	NX	NY	NZ	OA	OB	OC	OD	OE	OF	OG	OH	OI	OJ	OK	OL	OM	ON	OO	OP	OQ	OR	OS	OT	OU	OV	OW	OX	OY	OZ	PA	PB	PC	PD	PE	PF	PG	PH	PI	PJ	PK	PL	PM	PN	PO	PP	PQ	PR	PS	PT	PU	PV	PW	PX	PY	PZ	QA	QB	QC	QD	QE	QF	QG	QH	QI	QJ	QK	QL	QM	QN	QO	QP	QQ	QR	QS	QT	QU	QV	QW	QX	QY	QZ	RA	RB	RC	RD	RE	RF	RG	RH	RI	RJ	RK	RL	RM	RN	RO	RP	RQ	RR	RS	RT	RU	RV	RW	RX	RY	RZ	SA	SB	SC	SD	SE	SF	SG	SH	SI	SJ	SK	SL	SM	SN	SO	SP	SQ	SR	SS	ST	SU	SV	SW	SX	SY	SZ	TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK	TL	TM	TN	TO	TP	TQ	TR	TS	TT	TU	TV	TW	TX	TY	TZ	UA	UB	UC	UD	UE	UF	UG	UH	UI	UJ	UK	UL	UM	UN	UO	UP	UQ	UR	US	UT	UU	UV	UW	UX	UY	UZ	VA	VB	VC	VD	VE	VF	VG	VH	VI	VJ	VK	VL	VM	VN	VO	VP	VQ	VR	VS	VT	VU	VV	VW	VX	VY	VZ	WA	WB	WC	WD	WE	WF	WG	WH	WI	WJ	WK	WL	WM	WN	WO	WP	WQ	WR	WS	WT	WU	WV	WW	WX	WY	WZ	XA	XB	XC	XD	XE	XF	YG	YH	YI	YJ	YK	YL	YM	YN	YO	YP	YQ	YR	YS	YT	YU	YV	YW	YX	YY	YZ	ZA	ZB	ZC	ZD	ZE	ZF	ZG	ZH	ZI	ZJ	ZK	ZL	ZM	ZN	ZO	ZP	ZQ	ZR	ZS	ZT	ZU	ZV	ZW	ZX	ZY	ZZ																																																																																																																																																																																	
0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010	0.011	0.012	0.013	0.014	0.015	0.016	0.017	0.018	0.019	0.020	0.021	0.022	0.023	0.024	0.025	0.026	0.027	0.028	0.029	0.030	0.031	0.032	0.033	0.034	0.035	0.036	0.037	0.038	0.039	0.040	0.041	0.042	0.043	0.044	0.045	0.046	0.047	0.048	0.049	0.050	0.051	0.052	0.053	0.054	0.055	0.056	0.057	0.058	0.059	0.060	0.061	0.062	0.063	0.064	0.065	0.066	0.067	0.068	0.069	0.070	0.071	0.072	0.073	0.074	0.075	0.076	0.077	0.078	0.079	0.080	0.081	0.082	0.083	0.084	0.085	0.086	0.087	0.088	0.089	0.090	0.091	0.092	0.093	0.094	0.095	0.096	0.097	0.098	0.099	0.100	0.101	0.102	0.103	0.104	0.105	0.106	0.107	0.108	0.109	0.110	0.111	0.112	0.113	0.114	0.115	0.116	0.117	0.118	0.119	0.120	0.121	0.122	0.123	0.124	0.125	0.126	0.127	0.128	0.129	0.130	0.131	0.132	0.133	0.134	0.135	0.136	0.137	0.138	0.139	0.140	0.141	0.142	0.143	0.144	0.145	0.146	0.147	0.148	0.149	0.150	0.151	0.152	0.153	0.154	0.155	0.156	0.157	0.158	0.159	0.160	0.161	0.162	0.163	0.164	0.165	0.166	0.167	0.168	0.169	0.170	0.171	0.172	0.173	0.174	0.175	0.176	0.177	0.178	0.179	0.180	0.181	0.182	0.183	0.184	0.185	0.186	0.187	0.188	0.189	0.190	0.191	0.192	0.193	0.194	0.195	0.196	0.197	0.198	0.199	0.200	0.201	0.202	0.203	0.204	0.205	0.206	0.207	0.208	0.209	0.210	0.211	0.212	0.213	0.214	0.215	0.216	0.217	0.218	0.219	0.220	0.221	0.222	0.223	0.224	0.225	0.226	0.227	0.228	0.229	0.230	0.231	0.232	0.233	0.234	0.235	0.236	0.237	0.238	0.239	0.240	0.241	0.242	0.243	0.244	0.245	0.246	0.247	0.248	0.249	0.250	0.251	0.252	0.253	0.254	0.255	0.256	0.257	0.258	0.259	0.260	0.261	0.262	0.263	0.264	0.265	0.266	0.267	0.268	0.269	0.270	0.271	0.272	0.273	0.274	0.275	0.276	0.277	0.278	0.279	0.280	0.281	0.282	0.283	0.284	0.285	0.286	0.287	0.288	0.289	0.290	0.291	0.292	0.293	0.294	0.295	0.296	0.297	0.298	0.299	0.300	0.301	0.302	0.303	0.304	0.305	0.306	0.307	0.308	0.309	0.310	0.311	0.312	0.313	0.314	0.315	0.316	0.317	0.318	0.319	0.320	0.321	0.322	0.323	0.324	0.325	0.326	0.327	0.328	0.329	0.330	0.331	0.332	0.333	0.334	0.335	0.336	0.337	0.338	0.339	0.340	0.341	0.342	0.343	0.344	0.345	0.346	0.347	0.348	0.349	0.350	0.351	0.352	0.353	0.354	0.355	0.356	0.357	0.358	0.359	0.360	0.361	0.362	0.363	0.364	0.365	0.366	0.367	0.368	0.369	0.370	0.371	0.372	0.373	0.374	0.375	0.376	0.377	0.378	0.379	0.380	0.381	0.382	0.383	0.384	0.385	0.386	0.387	0.388	0.389	0.390	0.391	0.392	0.393	0.394	0.395	0.396	0.397	0.398	0.399	0.400	0.401	0.402	0.403	0.404	0.405	0.406	0.407	0.408	0.409	0.410	0.411	0.412	0.413	0.414	0.415	0.416	0.417	0.418	0.419	0.420	0.421	0.422	0.423	0.424	0.425	0.426	0.427	0.428	0.429	0.430	0.431	0.432	0.433	0.434	0.435	0.436	0.437	0.438	0.439	0.440	0.441	0.442	0.443	0.444	0.445	0.446	0.447	0.448	0.449	0.450	0.451	0.452	0.453	0.454	0.455	0.456	0.457	0.458	0.459	0.460	0.461	0.462	0.463	0.464	0.465	0.466	0.467	0.468	0.469	0.470	0.471	0.472	0.473	0.474	0.475	0.476	0.477	0.478	0.479	0.480	0.481	0.482	0.483	0.484	0.485	0.486	0.487	0.488	0.489	0.490	0.491	0.492	0.493	0.494	0.495	0.496	0.497	0.498	0.499	0.500	0.501	0.502	0.503	0.504	0.505	0.506	0.507	0.508	0.509	0.510	0.511	0.512	0.513	0.514	0.515	0.516	0.517	0.518	0.519	0.520	0.521	0.522	0.523	0.524	0.525	0.526	0.527	0.528	0.529	0.530	0.531	0.532	0.533	0.534	0.535	0.536	0.537	0.538	0.539	0.540	0.541	0.542	0.543	0.544	0.545	0.546	0.547	0.548	0.549	0.550	0.551	0.552	0.553	0.554	0.555	0.556	0.557	0.558	0.559	0.560	0.561	0.562	0.563	0.564	0.565	0.566	0.567	0.568	0.569	0.570	0.571	0.572	0.573	0.574	0.575	0.576	0.577	0.578	0.579	0.580	0.581	0.582	0.583	0.584	0.585	0.586	0.587	0.588	0.589	0.590	0.591	0.592	0.593	0.594	0.595	0.596	0.597	0.598	0.599	0.600	0.601	0.602	0.603	0.604	0.605	0.606	0.607	0.608	0.609	0.610	0.611	0.612	0.613	0.614	0.615	0.616	0.617	0.618	0.619	0.620	0.621	0.622	0.623	0.624	0.625	0.626	0.627	0.628	0.629	0.630	0.631	0.632	0.633	0.634	0.635	0.636	0.637	0.638	0.639	0.640	0.641	0.642	0.643	0.644	0.645	0.646	0.647	0.648	0.649	0.650	0.651	0.652	0.653	0.654	0.655	0.656	0.657	0.658	0.659	0.660	0.661	0.662	0.663	0.664	0.665	0.666	0.667	0.668	0.669	0.670	0.671	0.672	0.673	0.674	0.675	0.676	0.677	0.678	0.679	0.680	0.681	0.682	0.683	0.684	0.685	0.686	0.687	0.688	0.689	0.690	0.691	0.692	0.693	0.694	0.695	0.696	0.697	0.698	0.699	0.700	0.701	0.702	0.703	0.704	0.705	0.706	0.707	0.708	0.709	0.710	0.711	0.712	0.713	0.714	0.715	0.716	0.717	0.718	0.719	0.720	0.721	0.722	0.723	0.724	0.725	0.726	0.727	0.728	0.729	0.730	0.731	0.732	0.733	0.734	0.735	0.736	0.737	0.738	0.739	0.740	0.741	0.742	0.743	0.744	0.745	0.746	0.747	0.748	0.749	0.750	0.751	0.752	0.753	0.754	0.755	0.756	0.757	0.758	0.759	0.760	0.761	0.762	0.763	0.764	0.765	0.766	0.767	0.768	0.769	0.770	0.771	0.772	0.773	0.774	0.775	0.776	0.777	0.778	0.779	0.780	0.781	0.782	0.783	0.784	0.785	0.786	0.787	0.788	0.789	0.790	0.791	0.792	0.793	0.794	0.795	0.796	0.797	0.798	0.799	0.800	0.801	0.802	0.803	0.804	0.805	0.806	0.807	0.808	0.809	0.810	0.811	0.812	0.813	0.814	0.815	0.816	0.817	0.818	0.819	0.820	0.821	0.822	0.823	0.824	0.825	0.826	0.827</

# データ構造化のシステム設計の基本方針

# 本事業を実施するにあたっての背景

## (1) 実験装置の種類の多さ

「小型装置から大型装置まで」、「汎用装置から最先端装置まで」という多様な実験装置を対象とする。装置の出力データに含まれる実験パラメータや数値データは、装置メーカー固有の語彙や形式をもつため、装置メーカーへのこれら装置出力データの内部構造や語彙に関する情報の公開への働きかけが非常に重要。

## (2) 材料研究テーマの種類の多さ

各ハブ機関が担当する主たる材料研究分野ごとに、ユーザの多様な材料研究テーマを対象とするため、データ構造化のテンプレートのバラエティが増えることが予想される。

## (3) 実験装置や材料研究テーマ毎のデータ数やデータ量のバラエティの多さ

1回の実験で得られるデータの総サイズが、1MB以下から数GBを越えるものまで多様なサイズのデータを扱うため、登録するデータの選び方や選んだデータをいつ登録するかの判断にバラエティが生じる。

## (4) FAIR原則が浸透していない状況下での多数のユーザによるデータ登録作業

産官学の2万人規模のユーザを対象とし、FAIR原則（Findable, Accessible, Interoperable, Reusable）がユーザに浸透していない状況下で、FAIR原則に沿ったデータ登録に伴う手間をユーザに行って頂く必要がある。

# 本事業を実施するにあたっての背景を元にした メタデータ入力の方

メタデータ登録時の自由度を確保しつつ、登録されたメタデータの（永い年月を経ても判読可能という意の）長寿命化を図る

- メタデータの登録形式は、「統制」と「自由」のバランスを図る。
- メタデータを記述する語彙の判読性を確保する。

メタデータの登録者の負担をできるだけ少なくする

- 一人のユーザが一度にすべてのメタデータを付与するのではなく、複数の段階に分けてメタデータを登録する。
- 生データからのメタデータの自動抽出を行う。自動抽出のスキプトの製作にあたっては、必要に応じて装置メーカーの協力を得る。製作したスキプトは共用する。

# データ構造化を行うシステムの設計方針①

## (1) メタデータの2種類の登録方式

(a) FAIR原則 (Findable, Accessible, Interoperable, Reusable) の Findable を実現するための統制された共通のメタデータ

- 文献であれば書誌情報や Dublin coreで、一般的には5W1Hで、理解されるメタデータ
- リレーショナルデータベース (RDB) に登録

(b) FAIR原則の Interoperable、Reusable を実現するための自由度のある研究用のメタデータ

- 実験パラメータや試料の材料情報などの研究用のメタデータ
- テンプレート化されたメタデータファイルとして登録
- メタデータファイルは検索対象となる

## データ構造化を行うシステムの設計方針②

### (2) データセットとデータの2段構造でのデータ登録とメタデータの付与

- データは、(ある目的をもって集めたデータ群の意としての) データセットに属するとする。  
ID番号付けされた各利用課題を、データセットの単位とする。
- データセット単位で、アクセス者の認可制御、共用ライセンスの付与、(データセットに付与するメタデータの意としての) データカタログの付与、データの閲覧/ダウンロードを行う。
- データに付与するメタデータは、データカタログからの差分情報として記録する



## データ構造化を行うシステムの設計方針③

### (3) 実験装置が出力する生データに埋め込まれたメタデータの自動抽出

- 実験装置が出力する生データに埋め込まれた実験パラメータのメタデータを自動抽出するための python スクリプトを製作して共用する。
- 上記の生データにどのような実験パラメータがどのように埋め込まれているかの情報を必要に応じて装置メーカーに依頼して提供して頂く。

### (4) メタデータを記述する語彙の判読化とその機能のシステムへの組込 (自動翻訳)

- 定義が必要な語彙については、その定義をデータカタログに記載する。
- 生データに埋め込まれていた実験パラメータを判読化する場合、必要に応じて装置メーカーの協力を頂く。

## データ構造化を行うシステムの設計方針④

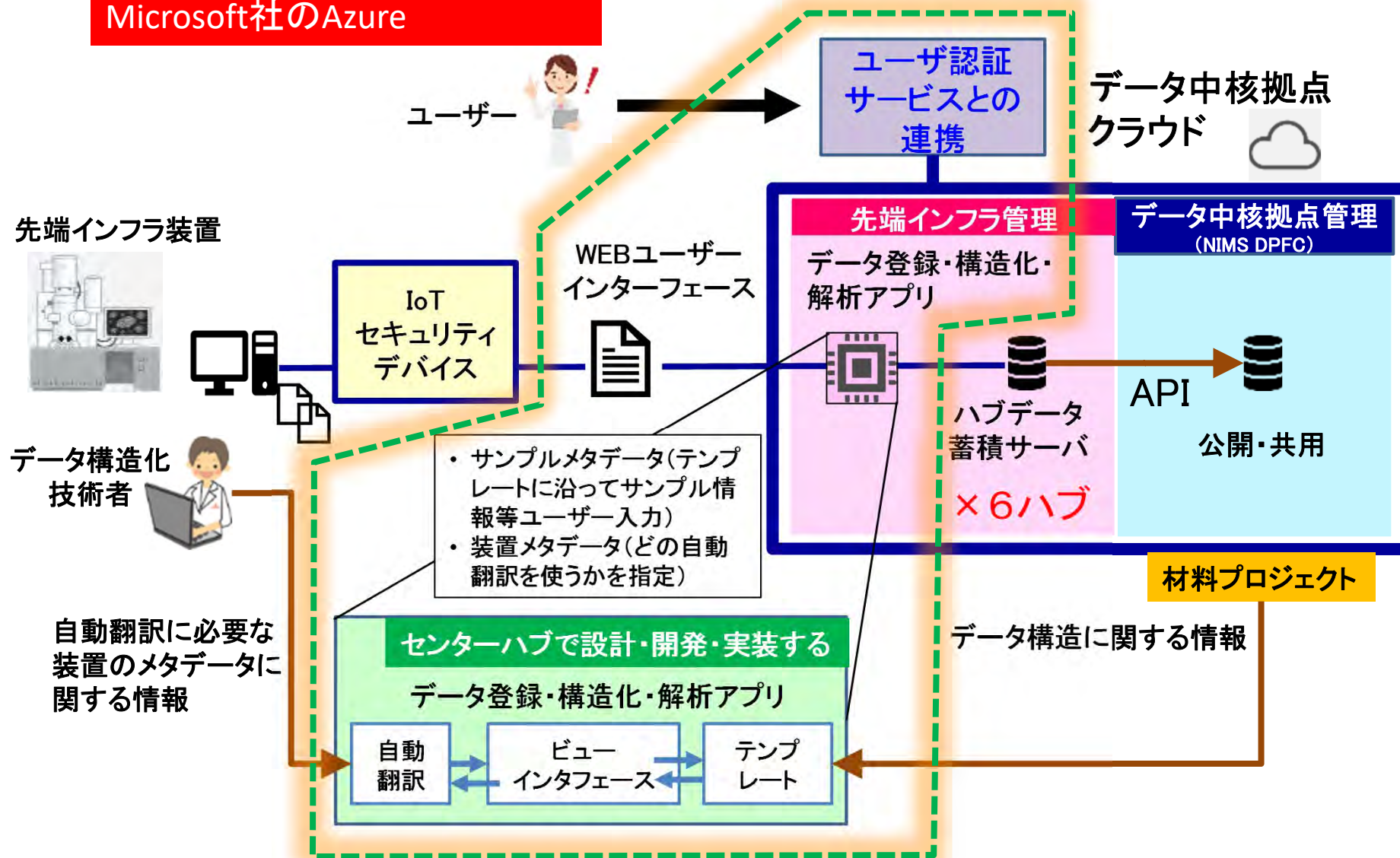
### (5) 実験装置が出力する生データ中の数値データの判読化

- ソフトウェア互換性の高いテーブル形式のcsvファイル等のテキストデータとして記録する。  
生データから（実験パラメータと共に）数値データを抽出する python スクリプトを記録し共用することで、数値データの構造情報を共有する。
- ファイルサイズが巨大でテキストデータでの記録が困難なバイナリデータについては、HDF5のように装置メーカー独自のソフトウェアを使わずとも解凍できる汎用の方式で圧縮したバイナリデータとして記録することが望ましい。

# データ登録・統合の効率化（データ構造化）システムの設計 開発の範囲

2021年度に設計開発する  
データ構造化システムの範囲

データ中核拠点のクラウドは  
Microsoft社のAzure



# データ蓄積の一次目的と二次目的の区別

システム設計の技術的な検討に関わる一次目的（データの共用）と

運用方針の検討に関わる二次目的（データの利活用）

を区別した議論を今後させて下さい。

# データ蓄積の一次目的（データの共用）

マテリアル先端リサーチインフラで生成した実験データを（データ生成者以外の研究者にも）共用できる形式にして永い期間蓄積する。



そのためには、

1. **検索**：検索に必要な最低限度のメタデータをデータに付与し、データが行方不明にならない。メタデータとして最も重要な項目は、データの生成者で、生成者のユーザ認証とID登録が必要。名寄せの課題については要検討。
2. **判読**：データが（メタデータの記述に使われる語彙および出力結果の数値データの表現について）最低限の判読性をもつ。生データのファイルの読取りと判読に関する実験装置メーカーのサポートが将来無くなるリスクも想定する。判読化作業の実施にあたって人による手間を少なくするための自動処理の工夫を行う。
- ★ 3. **安全**：情報セキュリティ上安全にデータを蓄積する。クラウドサーバへアップロードしたデータについてウイルス混入のリスクを最小化する。
- ★ 4. **契約 & 利用ライセンス**：データ提供者とのデータ登録に関する契約。データが共用になる際の利用のライセンスをデータ提供者が設定し利用に資する。
5. **ライフタイム**：データは、生成から処理/解析を経て共用され成果物として公開されるライフタイムを持つ。ライフタイムを考慮したメタデータの段階的な編集や（利用報告書を含む）複数のデータ間の紐づけを検討。データ蓄積のコスト推定の必要あり。

星印に関係する部分をこのあと説明いたします

## データ蓄積の二次目的（データの利活用）（案）

マテリアル先端リサーチインフラで生成した実験データを（データ生成者以外の研究者にも）共用できる形式にして永い期間蓄積することで、蓄積された多数のデータ自身を研究の対象や手段とするデータ駆動型研究の基盤を構築し、かつ共用実験装置の支援の効率化と高度支援者の人材育成を図る。

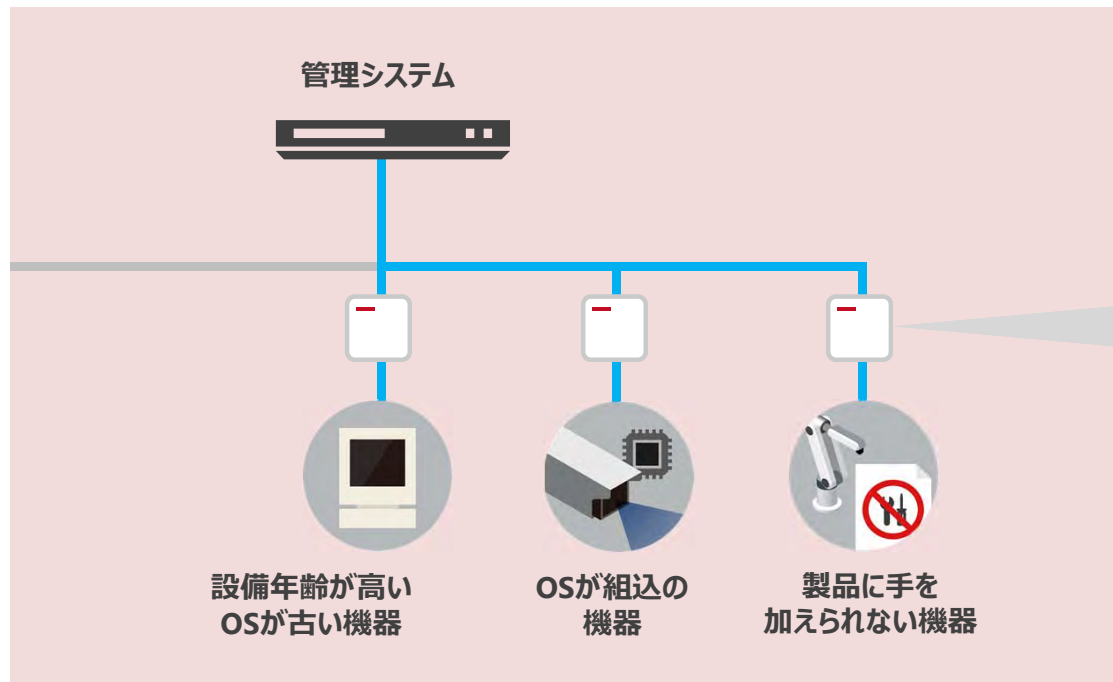
そのためには、「他者がデータを解読できるが、（例えば物質名が不明なため）そのデータの利用価値を想像できない」、という意味のデータ判読化のステージから、「他者にもそのデータの利用価値を想像できる」という意味でのデータの構造化に進むことが必要。以下の目的ごとにデータの構造化の指針が異なる。

1. 材料研究テーマごとのデータ構造化：材料研究テーマごとにデータに付与するメタデータの項目と記述形式および出力結果の数値データの記述形式についてテンプレートと python スクリプトを使って共通化する。
2. 計測、合成、加工の実験手法や装置の種類ごとのデータ構造化：実験手法や装置の種類ごとにデータに付与するメタデータの項目と記述形式および出力結果の数値データの記述形式についてテンプレートと python スクリプトを使って共通化する。

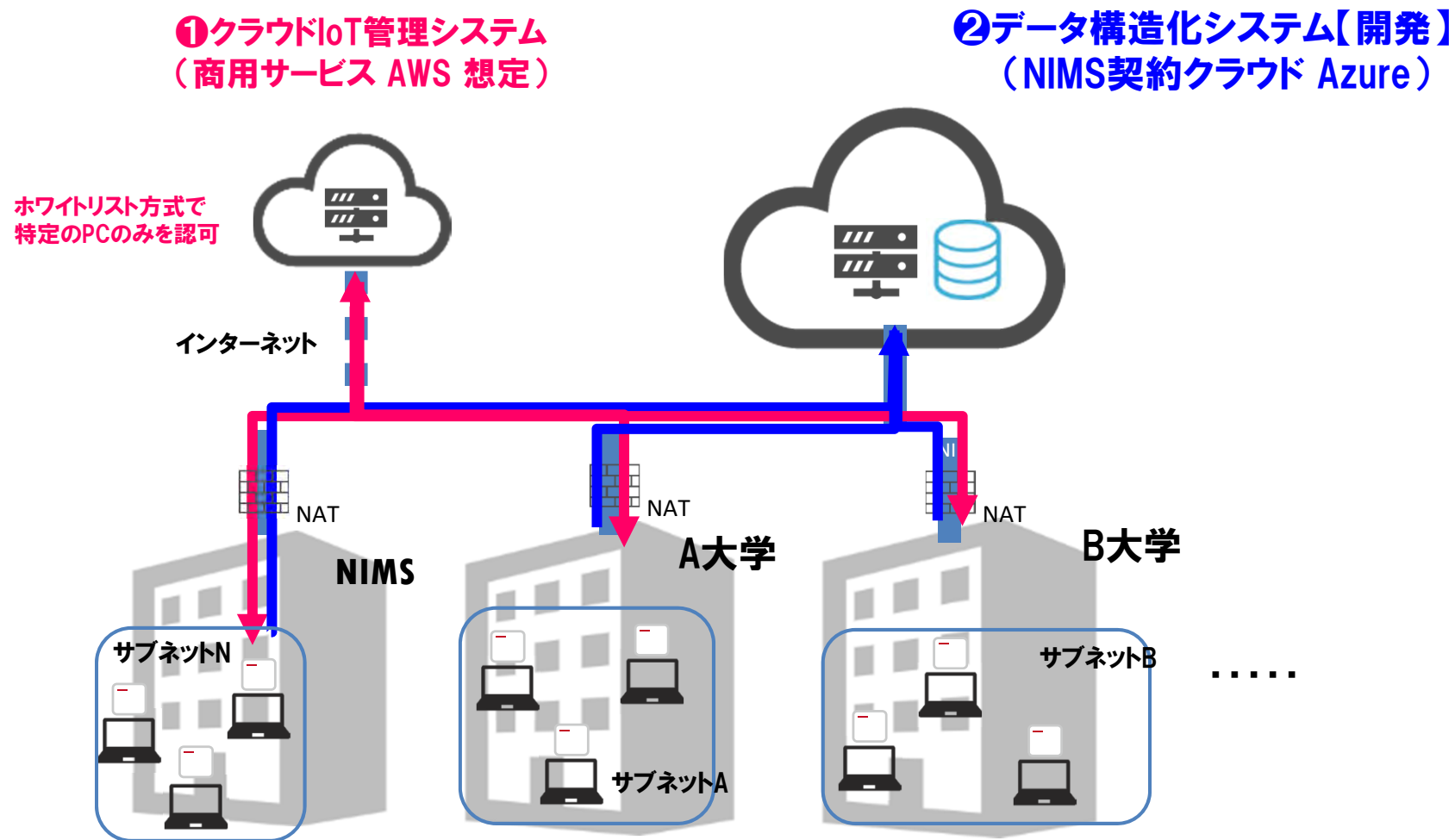
# IoTセキュリティデバイスについて

## 【例】 東芝のIoTセキュリティデバイス: CYTHEMIS

- イーサネットにデバイスを外付けするだけでセキュリティ対応
- 既存環境に影響を与えずに、システム全体をセキュア化
- 2021年度センターハブで500台整備予定(ただし、台数を調整する可能性あり)
- 様々なセキュリティ機能を外出しにし、エンドポイントのイーサネットの口に外付けデバイスを挟むことで既存システム全体をセキュアにする



# ネットワーク全体構成 インターネット経由でアップロード



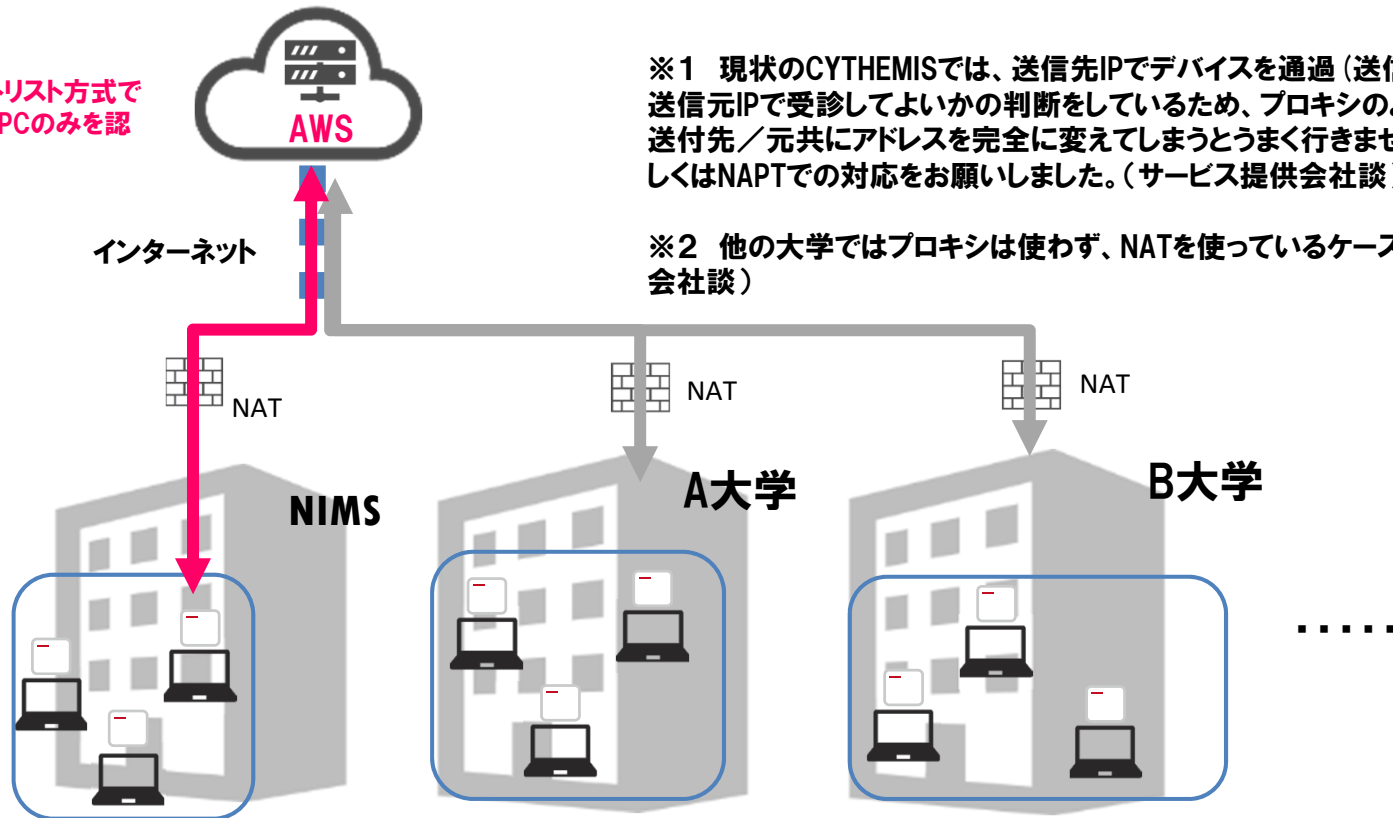
NAT(Network Address Translation): IPアドレスを変換するルーター



## R3年度の予定事項

### ①クラウドIoT管理システム (商用サービス想定)

ホワイトリスト方式で  
特定のPCのみを認  
可



- [1] NAT (※1)でインターネット経由でつなぐ
- [2] 各大学は各大学とサービス提供会社との間で  
接続可否を調査する(※2)

※1 現状のCYTHEMISでは、送信先IPでデバイスを通過(送信)させてよいか、送信元IPで受診してよいかの判断をしているため、プロキシのように一旦受け取って、送付先／元共にアドレスを完全に変えてしまうとうまく行きません。そのため、NATもしくはNAPTでの対応をお願いしました。(サービス提供会社談)

※2 他の大学ではプロキシは使わず、NATを使っているケースが多い(サービス提供会社談)

必要なもの: NAT機能をもつブロードバンドルーターに接続(設置)させていただくこと

## 利用者の認証と認可のフロー（案）

### 1. 利用者自身による利用者IDの取得

本システムを利用するための利用者ID（仮称DPF ID）を利用者が取得する。なお、ID取得を希望する人には誰でもDPF\_IDが発行されるものとする。

### 2. 各機関における利用者の契約確認

利用者からの利用課題申請を受理し、申請内容の承認をして頂いた各ハブまたはスポーク機関にて、利用申請者とその共同実験者のデータの登録と共用に関する利用規約や約款等の契約の確認をして頂く。

### 3. 各機関担当者による利用課題番号に関する情報を、利用者IDに付与し、課題IDと利用者IDを紐づけて承認

上記の利用課題申請を承認をした各機関の担当者の方が、利用申請者とその共同実験者のDPF\_IDの属性として、承認された利用課題番号に関する情報を材料データプラットフォーム認証システム（以下、認証システム）に入力する。

### 4. システム管理者が利用課題番号ごとにクラウド上にデータセットを開設する

上記の認証システムに登録された利用課題番号の属性情報を使ってシステム管理者がデータセットを開設する。

### 5. 認証システムの属性情報に基づいてデータセットの認可制御を行う

上記で開設されたデータセット単位で認証システムの属性情報をクラウド上の各アプリケーションにプロビジョニングをしてアクセス権限の認可制御を行う。

# 利用者の認証と認可のフローにおける課題

## 1. 利用者の契約確認の方法

利用申請者とその共同実験者のデータの登録と共用に関する利用規約や約款等の契約の確認の方法は、利用者からの利用課題申請を受理し承認をした各ハブまたはスポーク機関にお任せします。

なお、契約では以下の点を利用者に承諾して頂く必要があります。

- 利用課題と無関係なデータのアップロードをしない
- 故意に不正なメタデータの入力や不正なデータのアップロードをしない
- データの共用ルールに同意する
- データをアップロードするPCが限定されている場合はそれに従う。
- .....

## 2. 同一人物が、複数のDPF\_IDを取得した時の名寄せ

同一人物が複数の利用IDを取得することがシステム上、可能なため、本人による名寄せ等のために、同一性を確認する参考情報として、科研費番号(e-Rad)や世界研究者番号（ORCID）などの外部IDを追加登録することを推奨します。

# 利用者の認証と認可のフロー（案）



設備の管理者  
課題の管理者（課題承認）



設備の利用者（運営機関の人も  
含む当該機関利用者、大学利用  
者、企業利用者、個人利用者）



Step 1: 利用者は、DPF ID登録（NIMS提供システム）



Step 2: ハブ・スポーク各機関担当者は、ID登録者の利用課題における  
契約内容と属性の確認



- ① プロジェクト名（課題や研究プロジェクトのID）
- ② 氏名（科研費ID、ORCID等の参考記載を推奨）
- ③ 所属（ベタ）
- ④ プロジェクト有効期間（開始～終了＋猶予期間）

## DPF利用者の登録・認証・認可にかかる原則

1. 利用者が、複数の課題やプロジェクトを持つことOK
2. 機関担当者は、同一利用者が複数の設備利用により、複数IDを持つ時、  
単一化する承認権限あり。

ただし名寄せは完全ではないとの理解により：

3. 登録・公開データセットは、常に利用者IDと紐付く
4. DPFサービス（データ共用等）は、サービス側で認可制御する

# 情報セキュリティ上の懸念と対策

## 懸念事項

クラウドにアップロードする全てのファイルについて、クラウドの入口でウイルスチェックを実施しますが、全てのウイルスを完全に検知できるわけではありません。

## 上記の懸念事項に対する対策

ウイルスの感染リスクの高い（日々の業務として電子メールの添付ファイルを扱い種々のSNSにアクセスする）汎用PCを使ってバイナリの生データやスクリプトファイルを含む実験データをクラウドにアップロードすることを避け、感染リスクがほぼ無いと考えられる専用のPCだけからこれらのデータをアップロードして頂く対策を行います。

ここで言います専用PCは、通常は実験装置を制御するPCとなります。

# 情報セキュリティ上の安全対策（案）

## 1. クラウドにアップロードする際のウイルスチェック

クラウドにアップロードする全てのファイルについて、ウイルスチェックを実施します。

## 2. 実験装置の制御PCからの実験データのクラウドへのアップロードに対する特別な安全対策

- バイナリの生データやスクリプトファイルを含む実験データをクラウドにアップロードする際に、アップロードする（通常は実験装置を制御する）PCを指定して頂き、そのPCだけからのデータを（IoTセキュリティデバイスを経由して）クラウド側で受け取ります。
- IoTセキュリティデバイスは、AWS上の管理サーバで監視します。

## 3. クラウドに登録されたデータの検索/閲覧/ダウンロード

利用者は、アクセス権限があるデータセットについては、データセット内のデータの検索/閲覧/ダウンロードができます。データのダウンロードにあたっては、利用者の認証認可は行いますが、ダウンロードするPCの制限はしません。

# 予備資料

## データセットのタイプ

	データセットタイプ	インフォマティクスの種類	ステークホルダーの主なメリット
A	加工・プロセス型	プロセス・インフォマティクス	ナノブラのサービス提供機関のメリット:ノウハウ蓄積とユーザーサポートの迅速化。人材育成に役立つ。
B	構造解析・リファレンス型	計測インフォマティクス	ナノブラの機器利用者のメリット:データ解析の迅速化。人材育成に役立つ。
C	特性・性能規定型	狭義のマテリアルズ・インフォマティクス	ナノブラの(機器利用者とは異なる)データ利用者のメリット:マテリアルズ・インフォマティクスによる材料探索を可能にする
D	計算機型	マテリアルズ・インフォマティクス	大型計算機などから出力される理論計算結果のデータに対応する。
E	成果報告書型	テキストデータマイニングを使ったマテリアルズ・インフォマティクス	ナノブラの(機器利用者とは異なる)データ利用者によるテキストデータマイニングの実施を可能にする。 ナノブラのサービス提供機関のメリット:機器利用者の成果の統計データが得やすくなる

データセットのタイプに応じたリポジトリ様式の構築, ならびにユーザーフレンドリーなるGUIの設計を想定.



## *CYTHEMIS™* 3つの独自性

東芝インフラシステム様資料

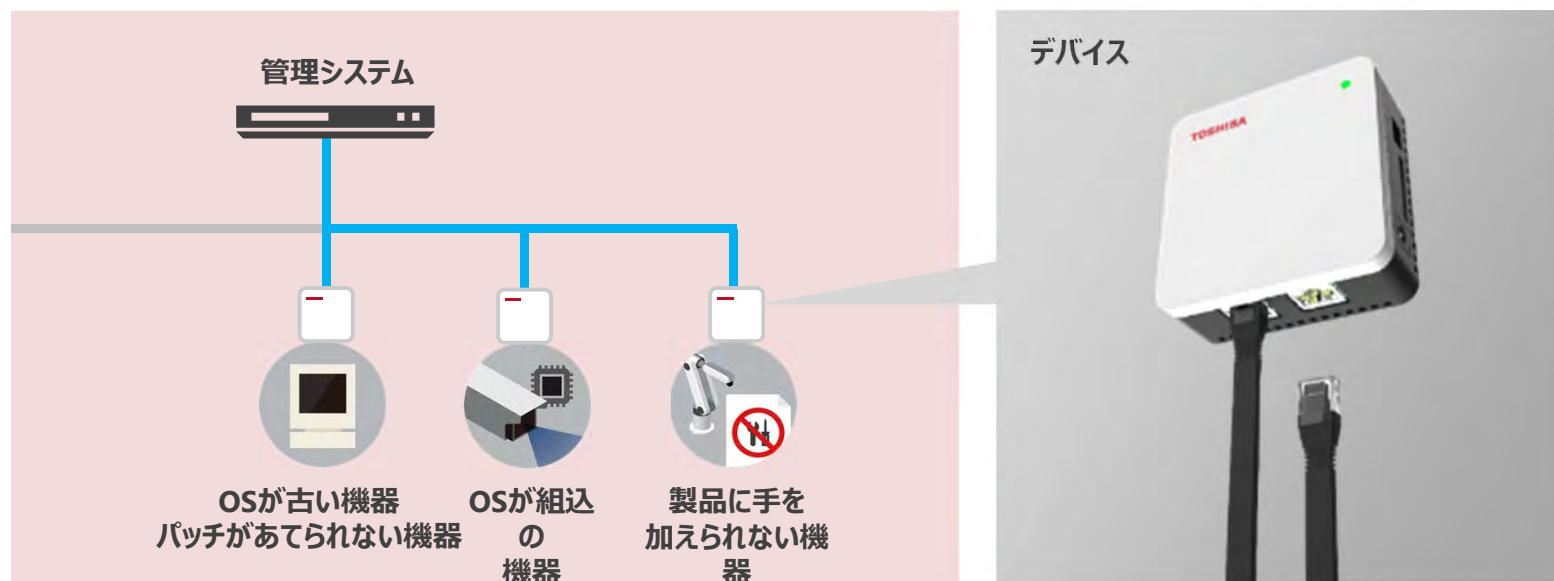
1. 外付けだからレガシーに対応
2. ネットワークの見える化
3. 信頼できる通信しかしない

## 独自性1. 外付けだからレガシーに対応

東芝インフラシステム様資料

### イーサネットにデバイスを外付けするだけでセキュリティ対応

エンドポイントには一切手を加えず、イーサネットの口の外付けデバイスを挟むだけで、エンドポイントを様々な脅威から守ります。それにより、これまでつなげられなかったレガシーな機器でもネットワーク化できます。



・機器には手を入れずに、レガシー機器をネットワーク化

## 独自性2. ネットワーク接続の見える化

東芝インフラシステム様資料

### レガシーな機器を含むネットワーク全体を見える化・一元管理

不正なアクセス・挙動を検知・防御し、感染を未然に防止します  
デバイスを設置したエンドポイント一式を見える化、一元管理し、ネットワークの状態をお知らせします



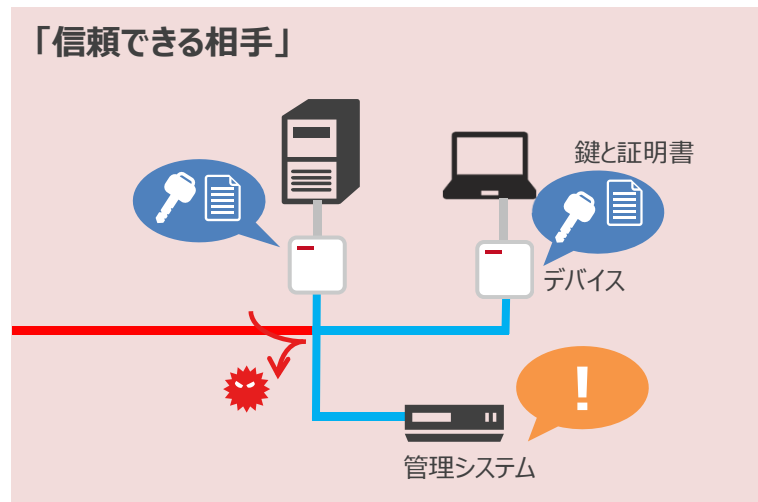
・エンドポイント一式を管理でき、異常時には検知・防御できます

## 独自性3. 信頼できる通信しかしない

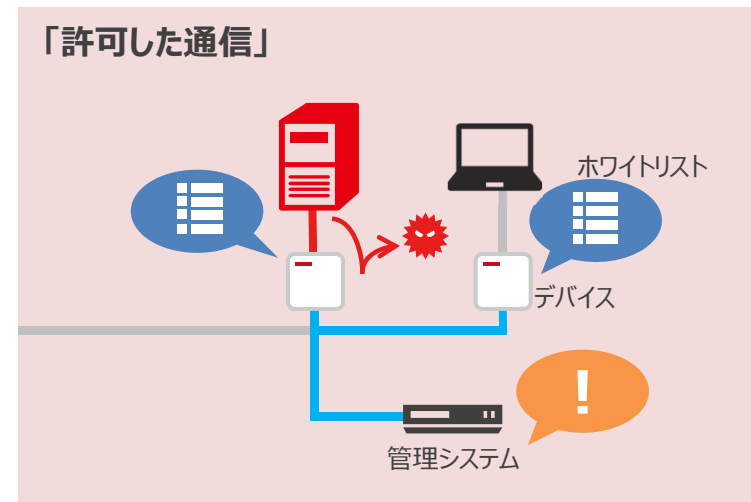
東芝インフラシステム様資料

### 「信頼できる相手」「許可した通信」のみ通信可能

デバイス間の相互認証・通信暗号化によって  
信頼できる相手としか通信しません  
正規に払い出された鍵と証明書を持った機器のみ通信可能です



どんな通信が正しいかあらかじめ学習し  
ホワイトリストを作成  
内部感染が起きた時も許可した通信しかさせません



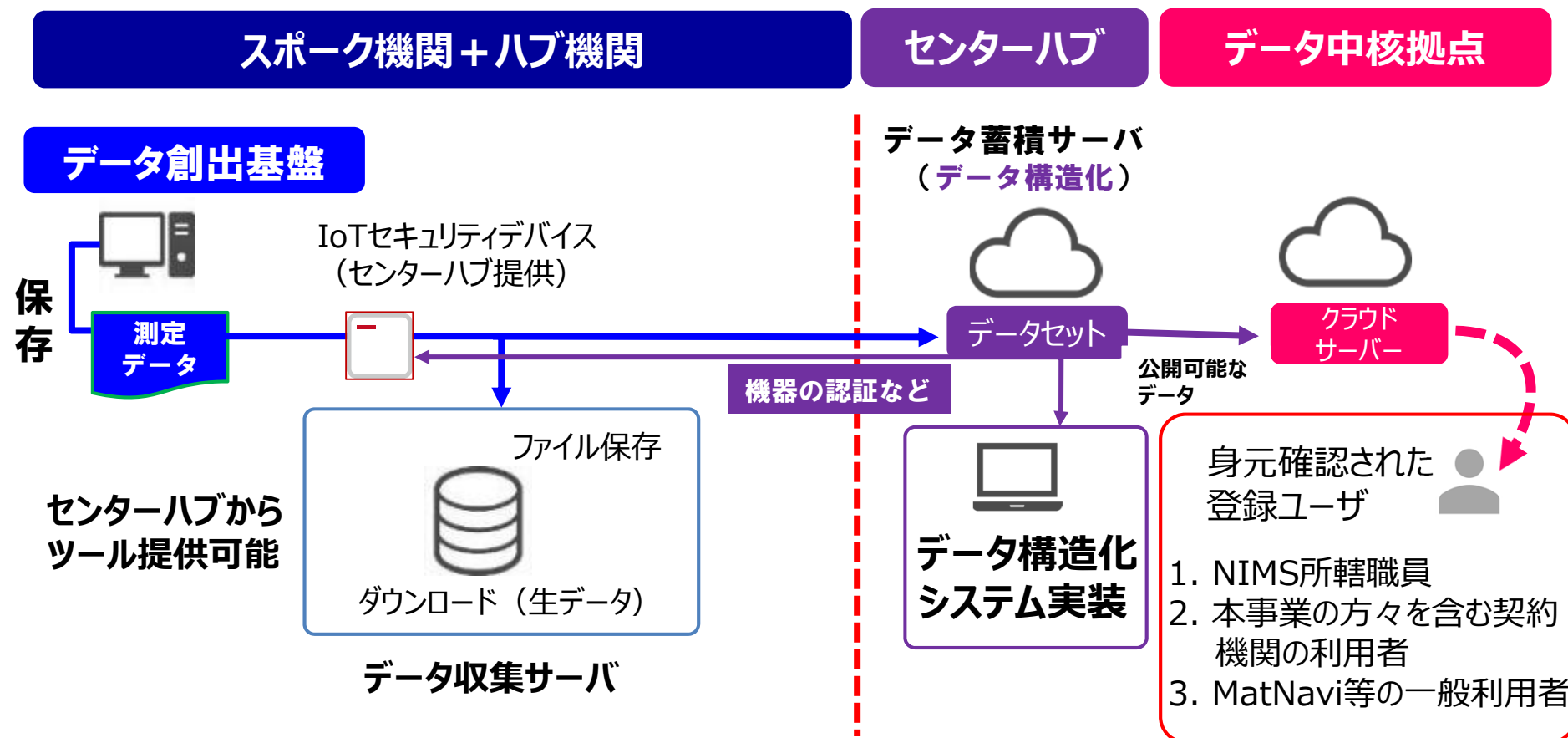
- ・マルウェアや不正なアクセスを感染前に検知・防御
- ・万が一、USBメモリ等で内部感染が起きた時も、感染拡大をブロック

## システムで必須とする構成要素



この構成要素の「要件定義」を行う

# 2021～2023年までのデータ収集の取り組み



- 2021年度 : IoTセキュリティデバイス(センターハブより提供)およびサーバー設置(ハブ&スポーク機関)を推進
- 2021年末 : データ構造化システム納入
- 2022年度 : データ構造化システムの立上げと性能検証(ハブ・スポーク機関想定)
- 2023年度 : データ構造化システムの試験運用(ユーザ想定)
- 2024年度 : データ構造化システムの本格運用