計測標準フォーラム第21回講演会
計量標準・計測におけるDX －信頼性確保・データ活用に向けて－
2023年10月24日（火）

「計測分析を未来につなぐための データの在り方と標準化」

九州工業大学・情報工学研究院・物理情報工学研究系
安永卓生

本日の話題

1. 背景:未来につなぐためのデータの在り方

2. 標準化への試み：データフォーマットの概要
   計測分析のデータ構造
   　０，１層の定義
   独立可用性の担保
   　汎用的表現，データファイルの挿入，名前空間，シソーラス
   計測分析のモデル化
   トレーサビリティ
3. データ作成及び変換のガイドライン
   ユースケース，開発事例，利用事例

背景：ビッグデータとしての計測分析データ

- ビッグデータの特徴：3V
  - Volume：データ量の多さ
    - 1D/2D/3D/4D/3D+3Dベクトル or スペクトル+1D　など
    - 自動計測・分析による大量のデータ
    - データの分析，強化学習への利用
  - Velocity：データの生成速度や流入の速さ
    - 計測そのものの高速化
    - 自動合成，自動計測，センサー群によるリアルタイム・データの輩出，自動分析
  - Variety：　データの種類の多様性
    - 構造化データ：「キー」が固定化したRDBのようなデータベース，テーブル型データ（CSV)
      - 事前定義，一貫性がある上で，変更に制約がある
    - 非構造化データ：テキスト，画像，動画
    - 半構造化データ：構造化はされているが，固定的でないデータ（JSON/XML/YAML)
      - キーと値（属性付）のペア：柔軟性，階層性，検索可能（XML to CSV）

背景：ビッグデータとしての計測分析データ

- ビッグデータの特徴：＋3V
  - Veracity：データの真実性，品質
    - データの完全性　→　独立可用性に係わる部分，Open/Closeと秘匿化
    - 正確性，整合性：独立可用性に係わる部分
      - データの整合性を意識した構造化の必要性
    - データの信頼性　→　不確かさに関わる情報
    - メタデータによるデータの背景の理解による信頼性の向上
  - Value：データから得られる価値
    - 目的の明確化とデータ収集にかかるコストと価値　→　事例：材料開発のためのデータ価値
    - データクレンジング（品質の保証）
    - 分析などのアプローチの更新によるデータ価値の創造　→　必要になるデータは変化
    - 適切なツール，スキルなど　→　OpenなAPIなどの必要性
  - Variability：データの変動性，不規則性
    - 計測分析条件の変化，機器の違い・変化によるデータの変動
    - 計測分析の進化によりデータの品質が変化　→　変化に対応できるフォーマット

背景：未来につなぐための計測分析データ

- 多様で，複合的で，持続的に開発が進む計測・分析　→　変化できる半構造化データ
  参考：個別の計測分析機器のフォーマットなどは既提案
  AniMLなど

- 複合した計測分析の全データの統合　→　計測分析のモデル化（プロトコル）とログの記録
  - 多様なデータセットに，メタデータを紐付し，構造化できる形で提供

  例：材料開発における計測分析：MIのためには複数のプロセスの全データの包括
  原料・材料の製造
  →　（複数の）計測分析に対応した試料の前処理
  →　（複数の）計測分析の結果，条件，試料情報
  →　（複数の）計測分析を（統合した）後処理・分析

- データの改ざん　→　データの唯一性の保証と改ざん検知
  - 品質確保のためにも異常データは必要だが，偽データはいらない

- オープン／クローズ戦略　→　データの完全性と秘匿化の両立
  - 秘匿化により，データの完全性の可能性を高め，独立可用性を担保

標準化への試み

データフォーマット(MaiML)の概要

JIS（原案）として審査中

計測分析装置の分析データ共通フォーマット

Common format for metrological analysis data

Measurement **A**nalysis **I**nstrument **M**arkup **L**anguage

計測分析の包括フレームとしてのデータフォーマットの提案

計測分析，前処理，後処理などのプロセス（ワークフロー，プロシージャー）を表現し，
データの独立可用性（全データを含）を保証するデータフォーマット

- 可視化：XML準拠（フォーマットの意味そのものも定義：半構造化データ）
  - スキーマの提案，構造化データへの変換方法(XML to CSVなど）の提案
  - 変化に対応できるフォーマット
- 再現性（ワークフロー／プロシージャー）の表現：計測分析のモデル化
  - ペトリネット型表現（非同期的，分散システムの数学表現）：メタデータの計測分析のフロー上での時間的な位置づけ
- 汎用的データ表現（変化する，多様なデータの表現）
  - 属性（データ型），値，データの意味（メタデータの意味付けは変化に対応）
  - 外部データの挿入指定：（規定された）多様なフォーマットは外部ファイル
- トレーサビリティ
  - XES型表現：データ取得時を含め，ログの記載（それぞれの単体の計測分析のプロセスのログの記載）
- 唯一性の保証
  - UUIDの採用
  - ハッシュデータによるデータの改ざん検知
  - 秘匿化によるデータの完全性の保証

XML（eXtensible Markup Language）

- ワード，Excelなどでも利用されている半構造化データ
  - 参考：.docxを.zipに変更して，開いてみてください。

```
<photo>
        タグ名      属性          値
        <acc     units=“kV”> 200.0  </acc>
        <mag> 30000 </mag>
        <position units=“um”> 310.0  405.1 253.0 <position>

        <quality> 10 </quality>
</photo>
```

書き方の作法
（どのようなタグ・属性を
どの順番で記載するか
記載の個数0以上，1，1以上）
はスキーマで定義可能

階層的に記載可能
タグ名，属性を使って値の
意味付けを行うことが可能

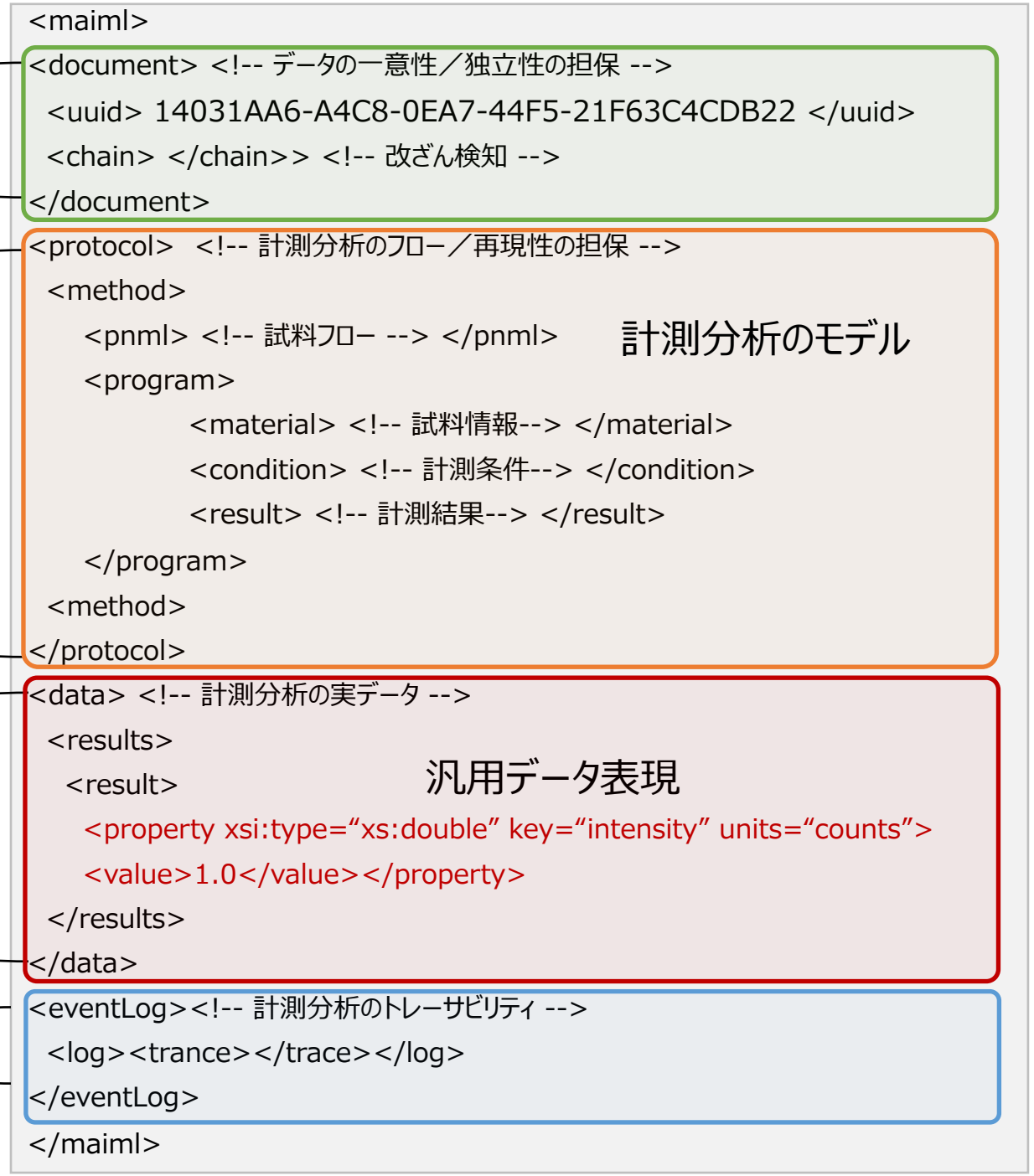
情報を記載，記載なしの選択
可能

参考：CSV(comma separated values)ファイル（構造化データ）
（項目行を名書くことが必須ではない，後から付け加えることは困難）

加速電圧, 倍率, 撮影場所x(um)，撮影場所y (um),撮影場所(um），質
200.0, 30000, 100, 200, 310.0, 405.1, 253.0, 10
200.0, 30000, 100, 200, 620.0, 675.1, 833.0, 8
…

JIS(案):MaiML

- 計測の包括的表現
  - 一意性+5W1H
  - 計測分析のフロー
  - 実データ
  - トレーサビリティ
    - XES表現
- データの汎用的表現
  - 型・意味・値
- 計測分析のモデル
  - ペトリネット表現
    - 試料・条件・結果
- 改ざん防止・秘匿化



計測分析の包括的表現：第０，１層

- データ構造：半構造化データ
  - サイバー空間で，データ自身の情報（定義）をもつ表現（読み取れる表現）
    - XML型（ASCII表現）

- 計測分析データのもつべき情報：第0～4階層で定義
  - <maiml>

計測分析の実施

設計

- **<document>**
  - データの唯一性，生成の起源（機器，データの所有者など），他のデータとの連携（改訂／改ざん検知）
- **<protocol>**：クラス
  - <method>
    - データ取得のための計測分析のフローのモデル化と表現（条件，試料，結果，及び命令）
- **<data>**：インスタンス
  - <results>
    - 実際の計測分析（条件，試料，結果）のセット
- **<eventLog>**
  - <log>
    - 各計測分析の命令（操作）の実行時のログ（計測分析の操作の遷移：完了など）

データの独立可用性の担保：方針

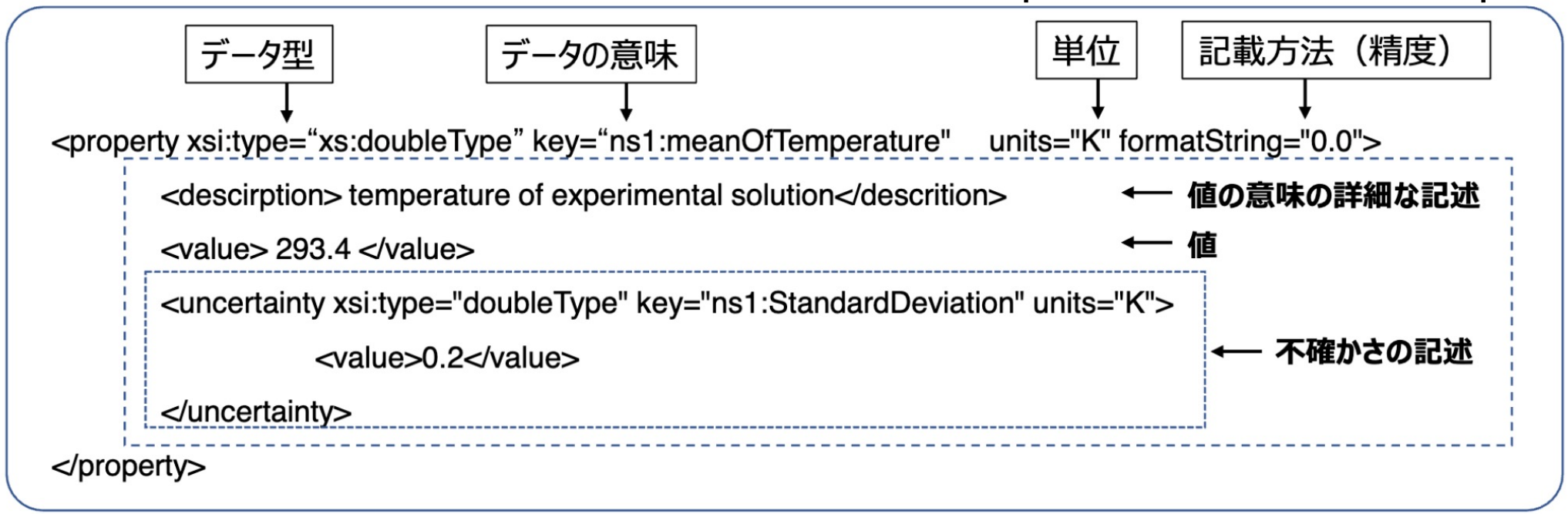
- メタデータの意味(key）の値は，JISでは定義せず，記載方法を規定
  - 計測分析およびその前後のプロセスを含めて全体を包括するため
    → 新しい機器，複合機器の開発も進むことから，メタデータの意味定義をしていくと継続的な変更が必要となる。自由なメタデータの記載
    → ユーザー規定型のメタデータ（キー，値，型などの属性値）
  - すでに他のJIS/ISOなどで規定されている用語も多い
    → 名前空間を利用して，JIS，メーカーなどの引用元を指定して利用できる。
    → シソーラスはオントロジー表現をつかって別途提供
    （附属書Ａ，Ｂ）
- データ型と値，不確かさの記載方法を定義
  - データアナリストは，型と値が分かればよい
    → アノテーションは，専門分野のユーザーにとって重要で，方言も多く，進展も多い

データの独立可用性の担保：汎用表現

- データ構造:多様なデータおよびメタデータへの対応
  - 汎用データの取扱：ネスト構造による階層性実現　→　ハッシュ型／構造体
    - <property>　単独の値
      - データ型：属性　xsi:type
      - 値：　<value>要素
      - 意味付け：属性　key

データ解析にとって重要

対象分野にとって重要



データの独立可用性の担保：他との連携，配列

- 他のデータ構造
  - 外部データの挿入：多様なデータフォーマットに対応，それらを包括するフォーマット
    - <insertion>：外部ファイル型の表現など

```
<insertion>
        <uri>http://www.ns1.co.jp/terminology/1.0.0/measurement.rdf</uri>
        <hash> e034136c7997579f235aab2c95ab32cbed801f7c2fa37709b3cbef81591564b6</hash>
        <format>text/rdf</format>
</insertion>
```

ファイルの挿入
ファイルの所在
ファイルのハッシュ値
ファイルのフォーマット

  - 配列型の値:<content>　配列型の値

```
<content xsi:type="decimalListType" key=”ns1:AnalogVoltage" units="V" size="3" axis="x”>
        <value>5.00 6.00 7.00</value>
        <uncertainty xsi:type="decimalListType" key="ns1:StandardDeviation" units="V">
                <value>0.12 0.25 0.31</value>
        </uncertainty>
</content>
```

配列の表現

附属書A（参考）の情報：名前空間の利用

計測分析に関わるメタデータの記載方法

名前空間を利用した，キー（値の意味付け）の収集

企業ごと，ユーザー毎の方言を吸収

附属書Ｂで示す，オントロジー表現で，標準との関係性を表現

```
<!-- 省略 -->
<material xmlns:exm="http://www.example.com/maiml/material#">
<!-- 省略 -->
        <property xsi:type="stringType" key="exm:sampleLotNumber" >
                <description>Lot Number of Sample</description>
                <value>00203080156T</value>
                <!-- 省略 -->
        </property>
        <!-- 省略 -->
</material>
<!-- 省略 -->
```

附属書Ｂ（参考）の情報：シソーラスの記載法

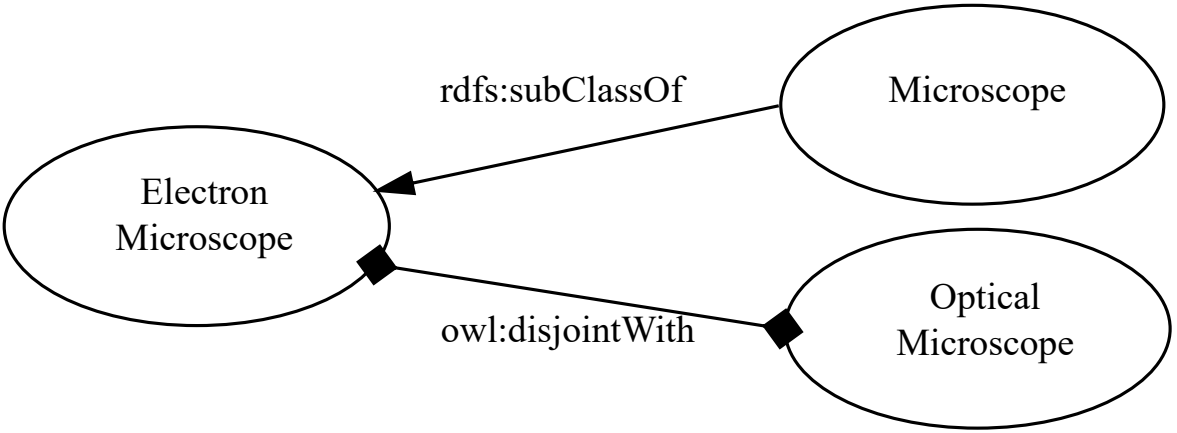
計測分析用語の関係性の表記方法：分析共通用語シソーラス

・key属性などで用いる用語集とそのオントロジー表現の事例を参考

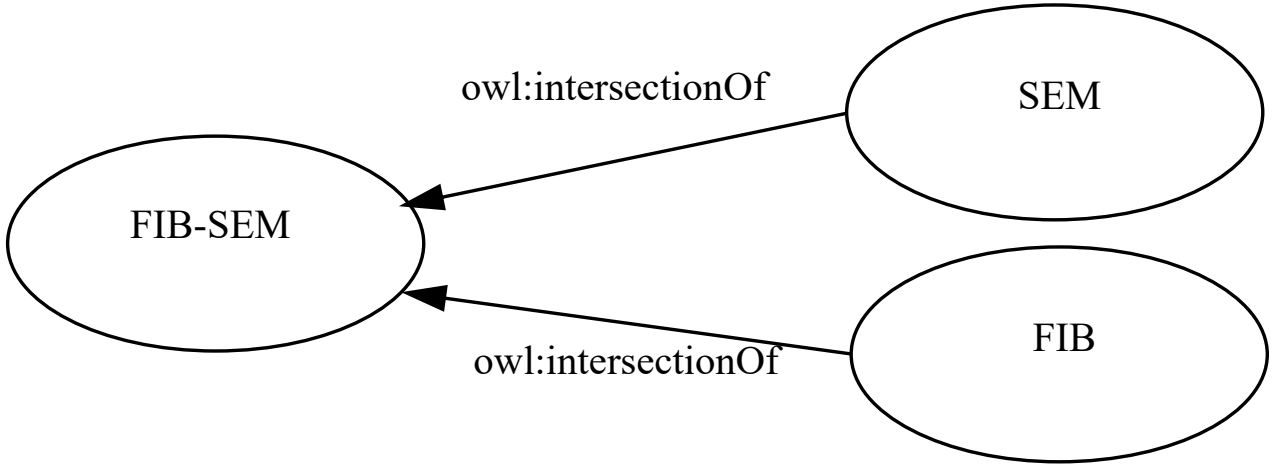
述語の定義は，規定のものを使い，JISでは規定しない。

メーカーやユーザーが指定した用語（名前空間）を，オントロジーで接続

電子顕微鏡は顕微鏡の部分集合であるが，
光学顕微鏡とは異なる



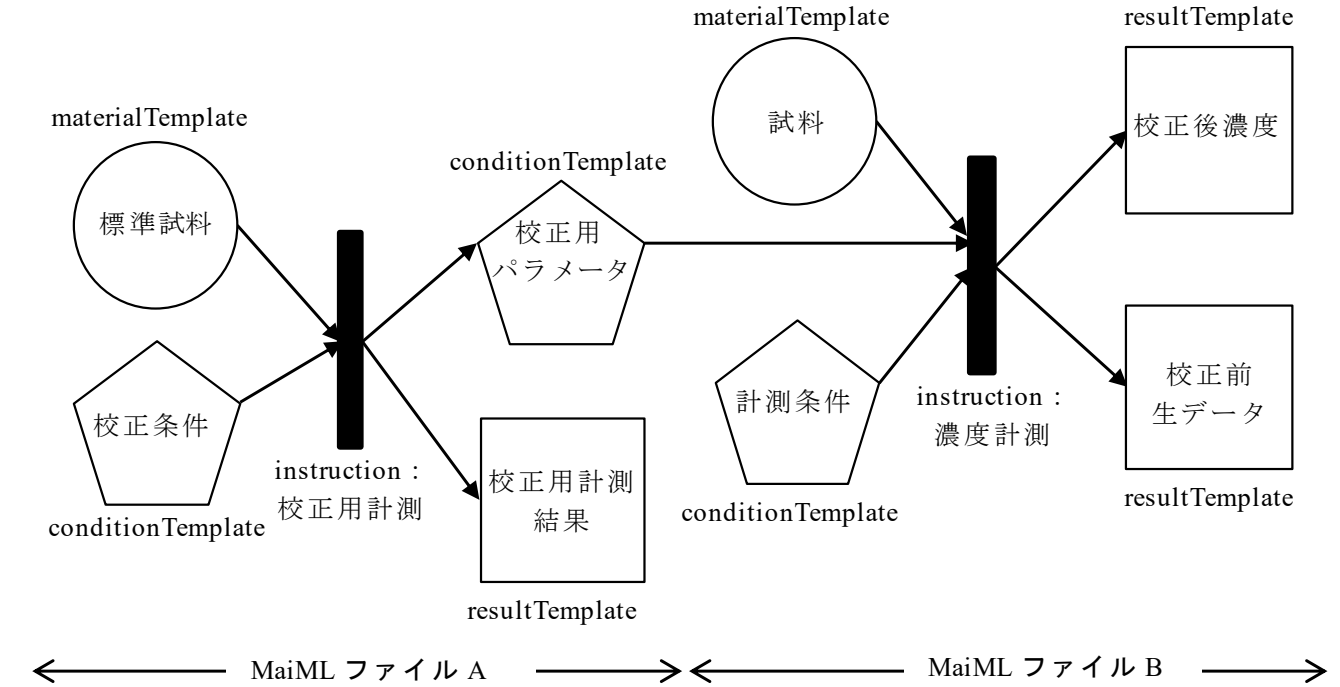
FIBSEMは，SEM及びFIBが結合したものである



計測分析のモデル化：条件／材料／出力／操作

- 計測分析データのモデル化
  - 条件：計測分析パラメータ
    - <conditionTemplate>
    - <condition>
    - 同一条件の計測分析
  - 材料：物理的実体を伴うもの
    - <materialTemplate>
    - <material>
    - 同一試料の計測分析
  - 結果：計測分析の情報
    - <resultTemplate>
    - <result>
    - 計測分析の出力
  - 命令：計測分析の操作
    - <instruction>

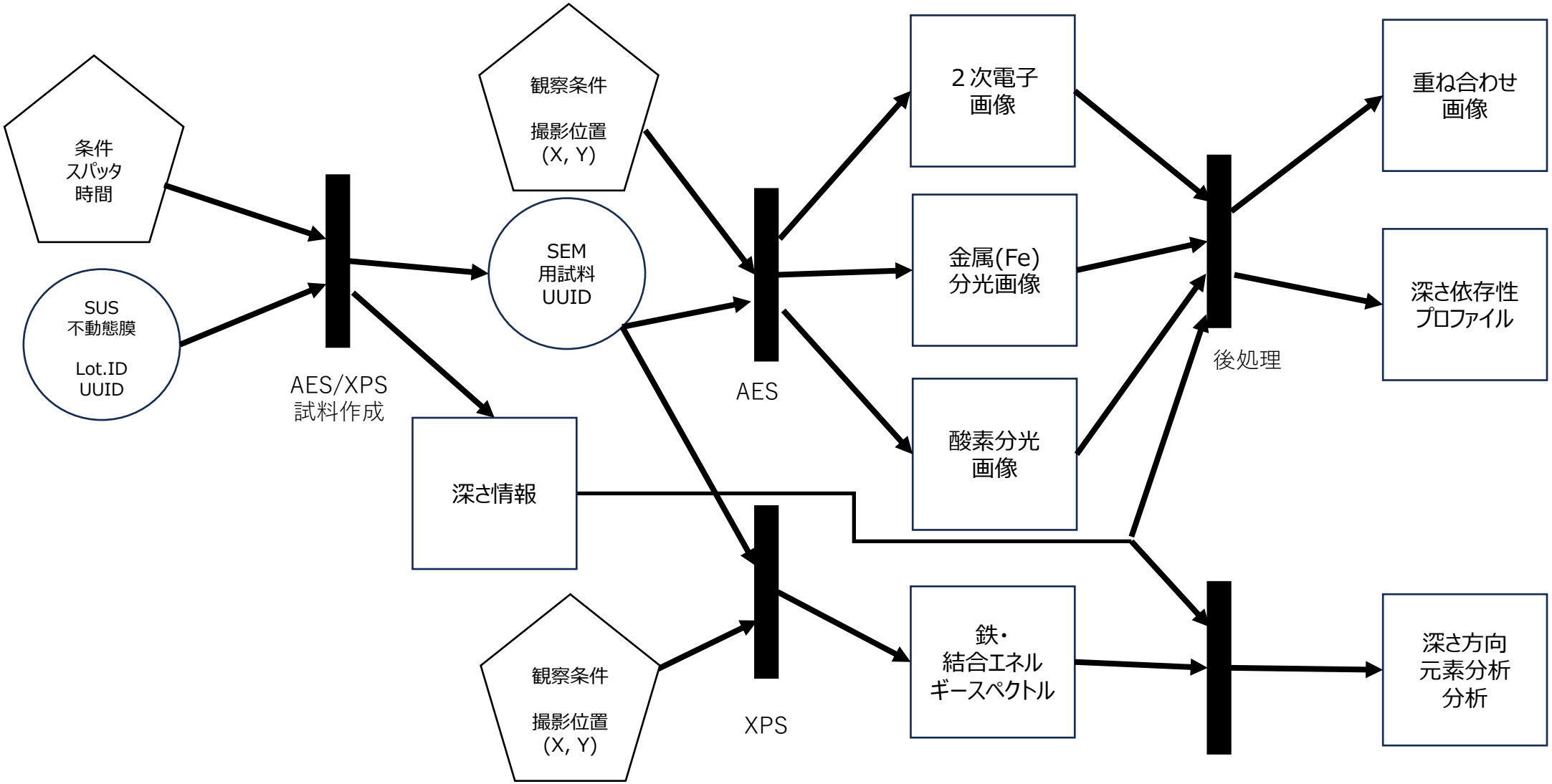
校正線を用いた濃度計測の事例



注記 英語表記は，要素名をいう。日本語表記は，それぞれの要素で記載しているコンテンツの意味を示してい

対応例：オージェ電子分光法(AES)+XPS

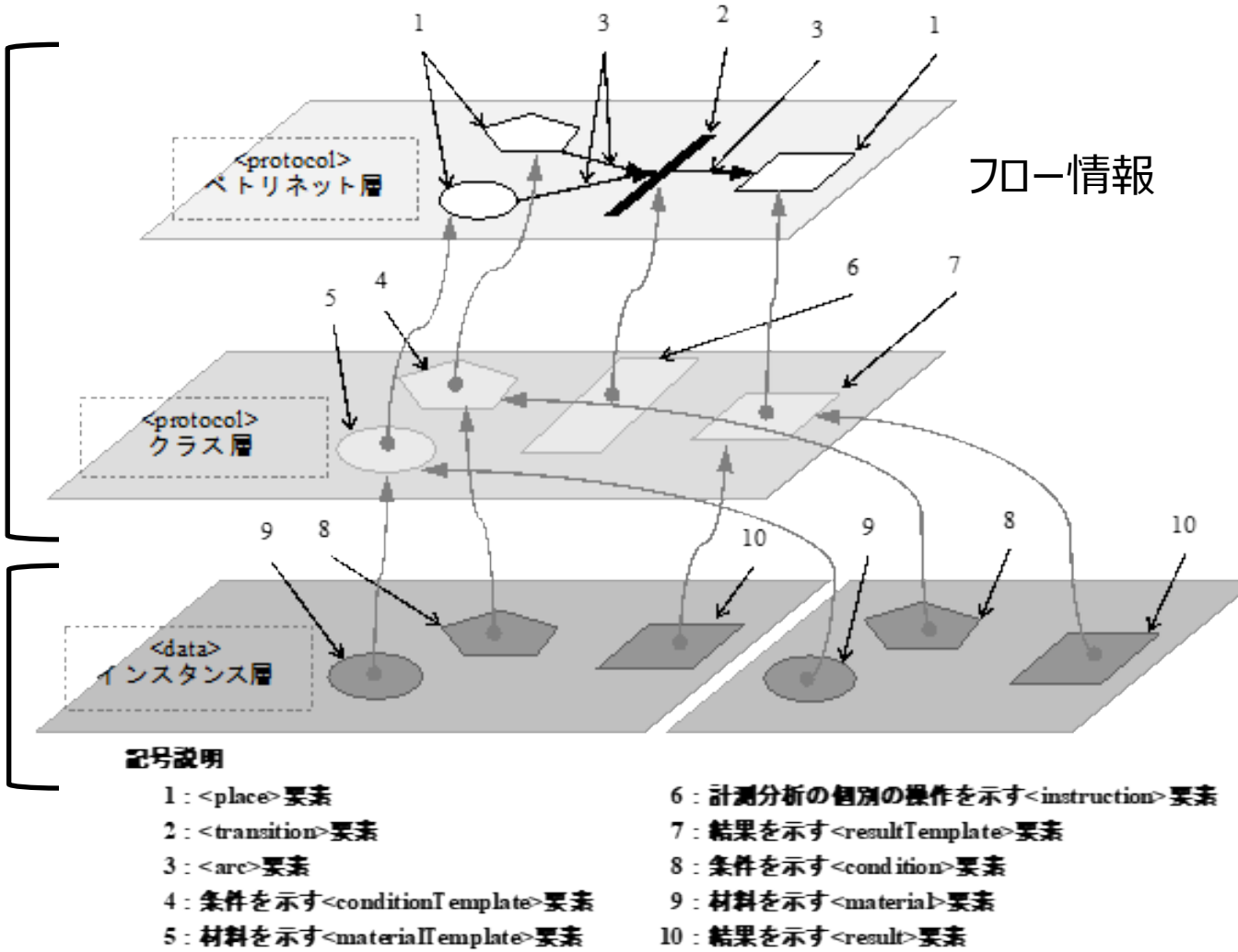
AESによる微小領域マッピング：SUS不動態膜の分析



参考：XPS･AESによる深さ方向分析の比較(B0211)（財）材料化学技術振興財団 参照:https://www.mst.or.jp/casestudy/tabid/1318/pdid/35/Default.aspx

計測分析のモデル（クラス：設計）と実体

- モデル(Template)：<protocol>
  - <xxxxTemplate>要素を使ってクラスとして記載（デザイン）
    - condition/material/result
      （条件）　（材料）　（結果）
  - ペトリネット層<pnml>内で
    - <place>要素
      - 試料／条件／結果
    - <transition>要素
      - 操作：<instruction>要素
- 実体(Instance)：<data>
  - <xxxx>要素を使って記載
- （必要であれば）モデルと実体との接続
  - <id>及び<ref>要素を用いて接続



類似の計測分析＋附属書C（参考）の情報

「類似の計測分析の工程」に関する記載方法の例

<templateRef>, <instanceRef>を用いて，類似のテンプレート／実体を参考

類似：ほとんど同じ条件，材料，結果ではあるが，
同一，若しくは同一とは言えない場合の記載

参考：同一の試料の場合は，UUIDが共通

事例：温度だけが条件として異なる場合の条件の書き方：<conditionTemplate>要素

```
<conditionTemplate id="conditionTemplate_p_m1">
        <uuid>CONDITION-TEMPLATE-P-M1-UUID-HERE</uuid>
        <!-- <tenplateRef>要素で参照されたテンプレートとは，UUID値が異なる。 -->
        <!-- コンテンツが記載され，異なるテンプレートを意味する。 -->
        <property xsi:type="xs:doubleType" key="ns1:temperature" units="K" formatString="0.0">
                <!-- 値を変更した要素又は追加した要素として指定する。 -->
                <value> 293.4 </value>
        </property>
        <placeRef id="placeRef_p_ms0" ref="place_p_ms0"></placeRef>
        <templateRef id="templateRef_p_m1" ref="conditionTemplate_p_m0"></templateRef>
</conditionTemplate>
```

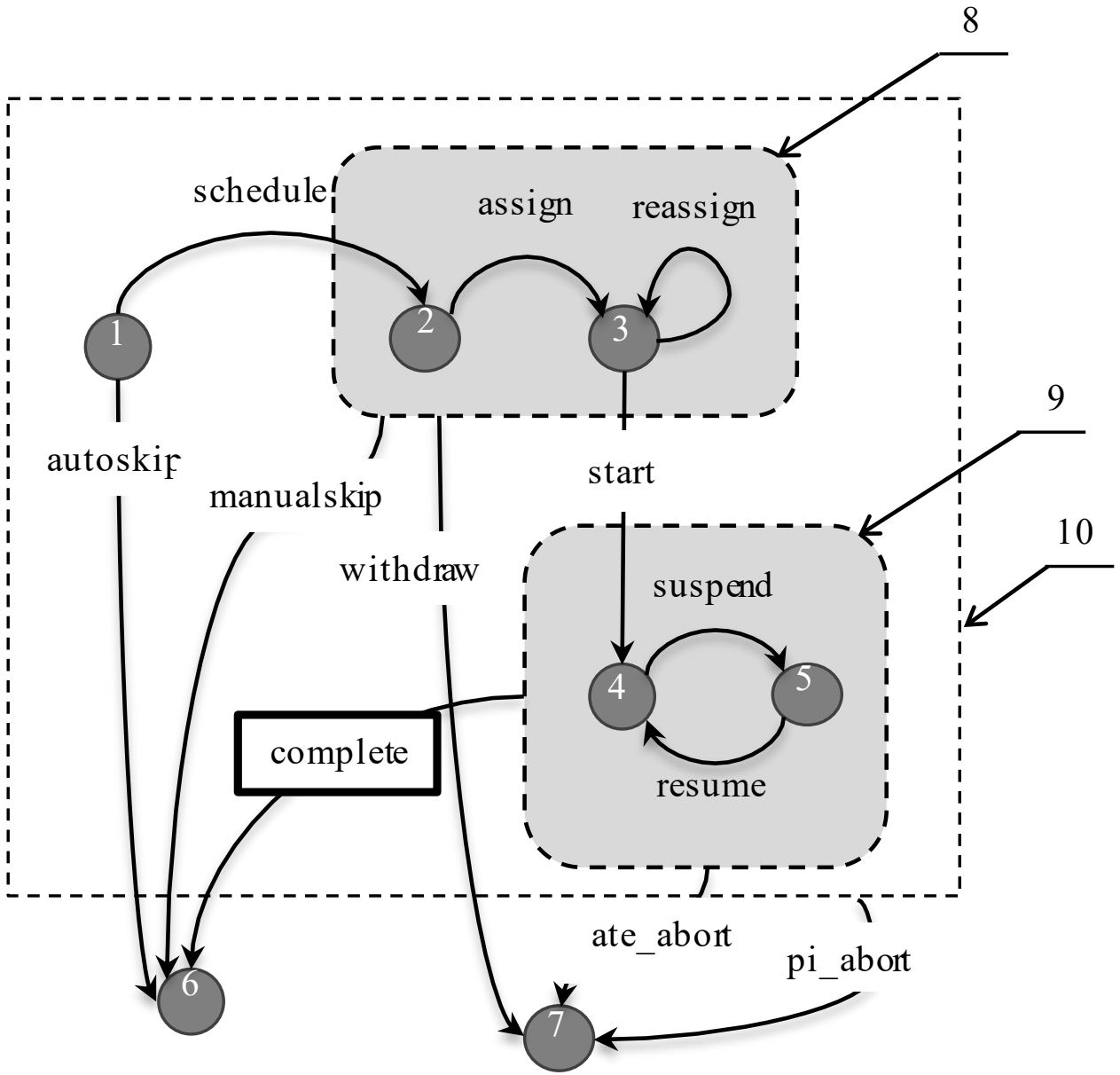
トレーサビリティ

- 計測分析ログの記載：
  <eventLog>
  - <instruction>
  - /<method>
  - /<results>などに対応して記載

  - 計測分析の操作における状態(state)の遷移として
    日時と合わせて記載

- プロセスのモデル化
  - complete/start/assign…

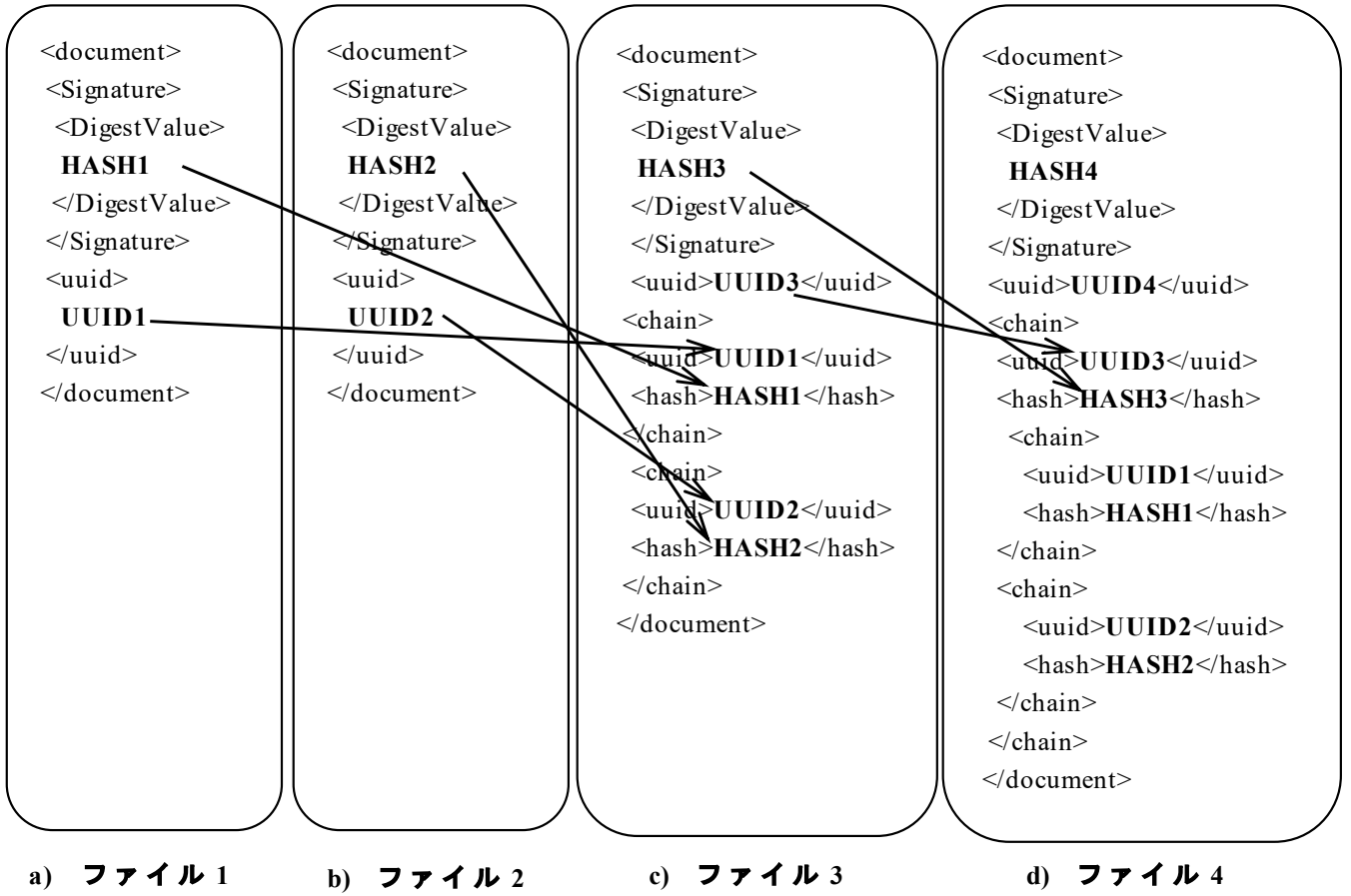
計測分析のプロセスログのモデル



改ざん防止

- 改ざん検知
  - **UUID**による一意性の保証
  - **<chain>要素**による，複数のファイルの連結による改ざん検知の導入
  - <parent>要素による，改訂版の連携

過去のファイルの
ハッシュ値
UUID値
を未来のファイルがもつことで，
過去のファイルが改ざんされたことが検知可能



オープン・クローズ戦略と独立可用性

- 全てのデータを含むことが期待
  - オープン／クローズの制御が必要
- 秘匿化
  - **XML暗号化**の制限付導入
    - 制限：データのフレームワークと計測分析のプロセスは可視化出切ることを保証
  - 制限を利用したプレースの分割

```
<material id="specimenForSEM-00" ref="specimenForSEM"
 xmlns:exm="http://www.example.com/maiml/material#">
    <uuid>1a74526e-ee89-4c5a-a9fa-d6efc70ccc66 </uuid>
    <name>exm:Sample</name>
    <description>Sample for SEM</description>
   <!-- 省略 -->
    <property xsi:type="stringType" key="exm:LotNumber">
      <value>1234502345SA</value>
    </property>
   <!-- 省略 -->
</material>
```

試料情報のうち，
UUIDを除く全てを秘匿

```
  <material id="specimenForSEM-00" ref="specimenForSEM"
 xmlns:exm="http://www.example.com/maiml/material#">
    <uuid>1a74526e-ee89-4c5a-a9fa-d6efc70ccc66</uuid>
    <EncryptedData>
   <!-- 省略 -->
    </EncryptedData>
</material>
```

一部の秘匿は不可：<material>要素を秘匿レベルで分ける

共通データフォーマット対応ガイドライン

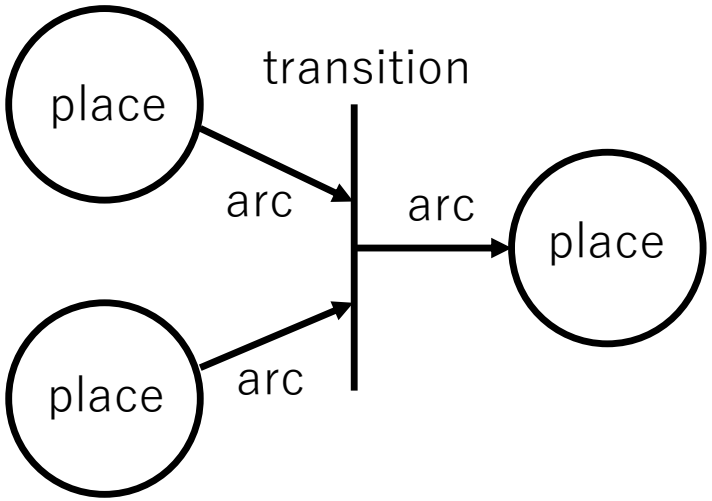
利用方法，作成方法に関するガイドライン

MaiMLの考え方，利用法，データ事例，開発プログラム例，FAQ

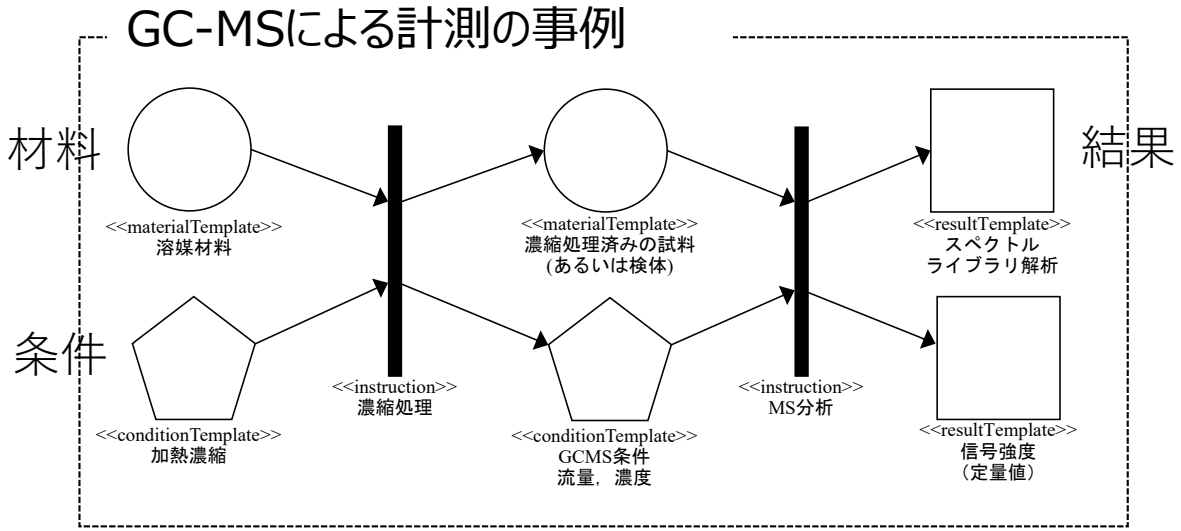
本日，一部の事例を記載

ガイドライン：要素の基本的な考え方

1. はじめに
2. 全体像
3. MaiMLで利用される要素とその記載方法
   XMLによる表現，グローバル要素の持つ意味
   ペトリネットによる計測分析などの一連の工程の記載



設計



ガイドライン：ファイル利用のユースケース

4. ファイル全体の構成

ユースケース１：データを含むもの

ユースケース２：計測分析の設計書のみ

<results>要素：一度の操作に関連した要素
<condition>要素
<material>要素
<result>要素

ISO9001における単一プロセスとプロトコル（ペトリネット）の関係

操作
ログのeventに対応

ガイドライン：汎用的なデータの表現

5. 汎用データコンテナ
各種表現事例

秘匿化ができる単位は，
汎用データコンテナの内部の全ての要素
属性は，秘匿化できない

試料の情報を組み合わせて１つのデータ（構造体）で表現したもの

```
<property xsi:type="propertyListType" key="ns1:SampleInformation">
  <property xsi:type="stringType" key="ns1:SampleId">
    <value>S123-00001-001#001</value>
  </property>
  <property xsi:type="stringType" key="ns1:SampleName">
    <value>不純物XXX含有率検査検体</value>
  </property>
  <property xsi:type="stringType" key="ns1:SampleType">
    <value>Unknown Sample</value>
  </property>
  <property xsi:type="unsignedByteType" key="ns1:SampleTypeNumber">
    <value>0</value>
  </property>
</property>
```

この値だけの
秘匿化も可能

全ての要素の
秘匿化

ガイドライン：他のフォーマットとの連携

6. 外部ファイルの引用方法

外部ファイルを利用するユースケース

- 高速処理を意図したバイナリファイル又は巨大なファイルを取り扱う
- 計測分析条件の一部又は全部をベンダー特有のファイルとして取り扱う
- 複数のMaiMLファイルから同一のファイルを引用する
- ネット上で，施錠，パスワードなどでアクセス制限を行うファイルを引用する

外部ファイルの指定方法

URI: ファイルの保管場所（ローカル，ネット上）
ハッシュ：ファイルのダイジェスト値（同一ファイル，改ざん無しであることの保証）
UUID: MaiMLファイルなど，UUIDをもつファイルで有る場合は記載（ファイルの一意性の保証）
フォーマット：多目的インターネットメール拡張（MIME）として指定されたメディア型を参考

ガイドライン：開発の事例

7. 機密性・安全性を確保したデータ収集・管理技術
8. シソーラスとオントロジー
9. 計測分析におけるMaiMLファイルのためのコンバータ作成フローの事例
   ワークフローの作り方の詳細

10. 計測分析に関わるユースケース
    pythonによるcsvファイルへの変換法
    keyを頼りにしたデータの取り出し方　など

11. よくある質問（ＦＡＱ）
    Java/C#/pythonなどを使った開発事例

補遺A MaiMLで使用する文字のガイドライン

補遺B XML修飾名の概要と命名のガイドライン

今後必要であるのは，APIとして公開されたもの

```
import pandas as pd
import glob
import xml.etree.ElementTree as ET

maimlfiles = glob.glob(' path/to/*')
fileNum = len(maimlfiles)
index = 0
HEADERS = ['key', 'description']
row = []
# Read from maiml
for i in range(fileNum):
    maiml = maimlfiles[index]
    index += 1
    maimltree = ET.parse(maiml)
    root = maimltree.getroot()
    # to row data
    for e in root.iter('{http://www.maiml.org/schemas}
            property'):
        key = e.attrib['key']
        description = ""
        for child in e.iter('{http://www.maiml.org/schemas}
            description'):
            description = child.text
        row.append([key, description])
# Write to CSV
df = pd.DataFrame(row, columns=HEADERS)
df.to_csv ('path/to/file.csv')
```

最後に

- 今回を含めて，様々な機会に御質問をうけた内容などを元に，FAQ(Frequently asked question)を充実させていきます。
- 国際標準化も進めています。そちらへも繁栄したいとおもいます。
- 様々な利用例を想定していますが，是非御一緒に考えていきたいと思います。
- いつまでも計測分析データが利用されるために，何が必要なのか，現状なにが不足しているのか，それらを考えて行きたいと思います。