

## 認識能力を備えたデータ解析システム

木野 千晶<sup>†</sup> 鈴木 善雄<sup>†</sup> 武宮 博<sup>†</sup> 中島 憲宏<sup>†</sup>

日本原子力研究開発機構 システム計算科学センター<sup>†</sup>

### 1 緒言

流体・構造・化学反応など様々な物理現象を考慮した大規模数値シミュレーションの必要性・重要性は年々増している。大規模・複雑な数値シミュレーションからは膨大なデータが出力されるが、人間の認識能力の限界から見落としや誤解などヒューマンエラーが発生する可能性は否めない。

著者らはこのような大規模・複雑データ解析を支援するために、「認識能力を備えたデータ解析システム Cognitive methodology based Data Analysis System (CDAS)」を提案している[1]。CDAS の基本概念は、大規模・複雑データから物理的・工学的に有意義と判断される情報を見出し、ユーザーに提示することである。CDAS には有意義な情報を見出すための多種多様なデータ解析ワークフローがデータベースに蓄えられている。ユーザーは CDAS を利用することで、自身とは異なる観点から解析された情報を知ることができる。このような多角的な解析は、見落としや誤解などのヒューマンエラーを防ぐ上で重要である。結果的に大規模・複雑データにおいて、効率的・効果的なデータ解析の実現が期待される。

有意義な情報を見出すデータ解析を実現するには物理的・工学的意味に立脚したワークフローが必要である。本研究では物理的・工学的意味をシステムが認識可能な形式で動的に記述するため、科学概念語彙を策定する。本稿ではこの科学概念語彙のフレームワークに関して議論する。

### 2 科学概念語彙の策定指針

#### 2.1 科学概念語彙の設計要件

科学概念語彙が実際の研究で有効に活用されるためには以下の要件を満たす必要がある。

- (1) 汎用性
- (2) 高い記述力
- (3) 拡張性

数値シミュレーションが対象とする解析体系・条件は千差万別であり、あらかじめ特定することは難しい。よって、この語彙は解析体系・条件が異なる数値シミュレーション結果に対しても対応できるものでなければならない。このため、語彙の共通認識としてのオントロジー構造を確立する必要がある。

次に、数値シミュレーションでは極めて広範な科学概念を使用する。これらの概念を逐次語彙に置き換えた場合、語彙数は膨大な数に達することが予想される。よって、科学概念語彙は研究者が習得可能な有限個の語彙を用いて効果的に広範な科学概念が記述可能となるように設計されなくてはならない。

最後に、科学技術は日進月歩であり、常に新しい科学概念が生まれている。よって、この科学概念語彙は高い拡張性を持ち、必要に応じて新たな語彙を定義可能とするように設計されなくてはならない。

#### 2.2 情報記述モデル

科学概念語彙の情報モデルは RDF(Resource Description Framework)情報モデル[2]をベースに構築する。RDF は主語・述語・目的語の組み合わせ(トリプル)を記述単位として文を構成し、セマンティックな記述が可能である。RDF は常にノードとアークを意識した情報記述モデルを想定しており、同じ属性(述語)を持った主語や目的語を同じノードにまとめることができる。ノードによってまとめられた概念同士の関係性をアークによって定義することで、分散的な情報を柔軟に扱うことができる記述手法である。

このような情報記述モデルに立脚した RDF は科学概念語彙を構築する上で有効である。まず、RDF のネットワーク的情報記述モデルは、語彙が背景として持つオントロジー構造を記述するのに適している。また、ノードとアークを追加することで容易に知識を拡張することができる。さらに、RDF スキーマが定義されており、新たに語彙を定義することも可能である。以上から、この科学概念語彙を定義するのに非常に有効な情報モデルであると言える。

### 3 科学概念語彙の構造

#### 3.1 科学概念語彙の情報モデル

科学概念語彙は RDF の情報記述モデルに基づいて構築される。RDF は Web 上における情報の共有を想定して構成された情報モデルであり、リソースおよびリソース間の関係性を示すプロパティによって構成される。RDF ではリソースを Web 上で識別されるものとして定義している。本論文で提案している科学概念語彙ではリソースは(1)データ(2)メソッド(3)意味情報(結合先リソースアドレス)の3つの情報を持つものとして定義する。

データとはシミュレーションの結果データ解析において必要となる物理量分布や領域情報などである。データはリテラル、ファイルデータ、テーブルによって入力され、システム内の処理を通じて変更・生成される。

メソッドとはリソースが持つデータを操作するアルゴリズムである。数値流体シミュレーションを例にすれば、

Development of Cognitive methodology based Data Analysis System

Chiaki KINO<sup>†</sup> Yoshio SUZUKI<sup>†</sup> Hiroshi TAKEMIYA<sup>†</sup>  
Norihiro NAKAJIMA<sup>†</sup>

<sup>†</sup>Center for Computational Science & E-systems, Japan Atomic Energy Agency

温度依存性のある物性値の場合、温度と物性値表を読み取って、適当な物性を算出するアルゴリズムや、液相中に存在する気液界面を検出するアルゴリズムが想定される。リソースが持つメソッドへ入力されるデータはプロパティによって接続されたリソースが持つデータに限定される。すなわち科学概念語彙の文法とはリソースが持つメソッドのインタフェースを規定したものと言い換えることができる。

本研究では意味情報を「リソースと既存の知識体系との関係性を明確にすること」であると定義する。具体的には、全てのリソースを URN によって識別することで、他のリソースとの関係性(プロパティおよび結合先のリソースに関する情報)を明確にし、結合先のリソースアドレス情報として保持する。

例えば、「ある領域における物理量 2 次元分布を測定する」というメソッドを持ったリソース A があるとすると、この時、「ある領域」として「リソース(壁面)」が、「物理量」として「リソース(せん断応力)」が結合しているという「意味情報」をリソース A は持つ。これにより、リソース A は「壁面せん断応力分布を測定している」という物理的意味を保持することができる。

### 3.2 科学概念語彙のデータ構造

リソースが扱うデータ構造を考察するため、科学的情報のオントロジー構造を調べる。図 1 に数値流体シミュレーションを例とした科学的情報のオントロジー構造を示す。

科学的情報とは「ある観測空間の定量・定性的特徴」とまとめることができる。観測空間とは単なる時刻・領域の設定に留まらず、どのような環境・条件であるか(例えば境界条件や周辺圧力)、どのような流体を用いているか(例えば流体の物性値や特性)、どのような解析体系であるかなど、系全体が持つ広範な概念を含んでいる。

定量的・定性的特徴はさらに評価指標と判断基準に整理できる。評価指標は「1 次元速度分布」や「壁面せん断応力分布」といった物理量分布と「気泡」や「液滴」といった物理現象に分類できる。また、判断基準とは物理量分布が基準値よりも「高い・低い」など評価指標が有意味であるか否かを判定するために必要となる情報である。科学概念語彙のリソースはこれらに関するデータを保持することで物理的・工学的意味を形成していく。

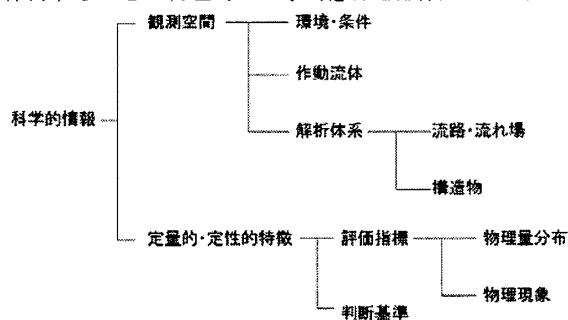


Fig.1 数値流体シミュレーションの科学情報に関するオントロジー構造

### 4 科学概念語彙の記述例

図 3 には数値流体シミュレーションに対する科学概念語彙で記述されたデータ解析ワークフローの記述例を示す。このワークフローでは 3 次元円柱領域における空間平均ボイド率の主流方向分布を測定するという解析内容

が記述されている。

まず<evaluation:target-region>以下で測定領域として主流方向に沿った直線(始点と終点)を定義する。

<evaluation:measurement-number>では測定点の点数を定義し、<evaluation:measurement-range>では測定点を基準として、空間平均を測定するための円柱を定義する。<evaluation:physical-value>では、用いる物理量は空間に占める気相の割合を示す VOF 値であることを示す。この VOF 値は urn:Physical-Value:VOF によって対応するデータファイルが定義される。

今回、科学概念語彙を認識し、解析を実行するプロトタイプシステムを製作した。そして、この記述例を用いて実際の数値流体シミュレーション解析結果に適用したところ、想定どおりの平均ボイド率分布の測定に成功した。ここで用いられている語彙は「測定領域」「1 次元分布方向」「物理量」など科学的意味を持った情報である。この記述例より、科学的意味を持った語彙を組み合わせ、科学的情報を抽出するワークフローが構成できていると言える。

```
<evaluation:target rdf:parseType="Resource">
  <evaluation:target-region>
    <shape:lineair>
      <shape:start-point rdf:parseType="Resource">
        <value:x> 9.33333E-04 </value:x>
        <value:y> 9.33333E-04 </value:y>
        <value:z> 1.40000E-03 </value:z>
      </shape:start-point>
      <shape:end-point rdf:parseType="Resource">
        <value:x> 9.33333E-04 </value:x>
        <value:y> 9.33333E-04 </value:y>
        <value:z> 8.17000E-02 </value:z>
      </shape:end-point>
    </shape:lineair>
  </evaluation:target-region>

  <evaluation:measurement-number> 5 </evaluation:measurement-number>

  <evaluation:measurement-range>
    <shape:cylinder>
      <value:radius> 1.0E-02 </value:radius>
      <value:height> 16.0E-03 </value:height>
      <shape:axis-direction rdf:parseType="Resource">
        <value:x> 0.0E0 </value:x>
        <value:y> 0.0E0 </value:y>
        <value:z> 1.0E0 </value:z>
      </shape:axis-direction>
    </shape:cylinder>
  </evaluation:measurement-range>
  <evaluation:target>
    <evaluation:physical-value>
      <rdf:Description rdf:about="urn:Physical-Value:VOF">
        </rdf:Description>
      </evaluation:physical-value>
    </evaluation:target>
  </evaluation:physical-value>
</evaluation:target>
```

Fig.2 科学概念語彙を用いた解析ワークフローの記述例

### 5 まとめ

本稿では科学概念語彙のフレームワークと、数値流体シミュレーションのオントロジー構造に基づく構築例を示した。今回の研究で得られた知見を元にプロトタイプシステムの有効性を検証することが今後の課題となる。

### 参考文献

- (1) 木野千晶 鈴木喜雄 西田明美 櫛田慶幸 林幸子 中島憲宏 “認識能力を備えたデータ解析システムの概念設計” 日本計算工学会論文集 論文番号: 20080018, 2008
- (2) World Wide Web Consortium (W3C): Resource Description Framework, <http://www.w3.org/RDF/> 2004