tq開発の全体像と経緯

2022-03-25 天野晃（材料データプラットフォームセンター）

ゴール

以下を同一の枠組みで行う仕組みづくりを行う。

- データフェデレーション
- 機械学習の定義
- 辞書とそこからの知識抽出

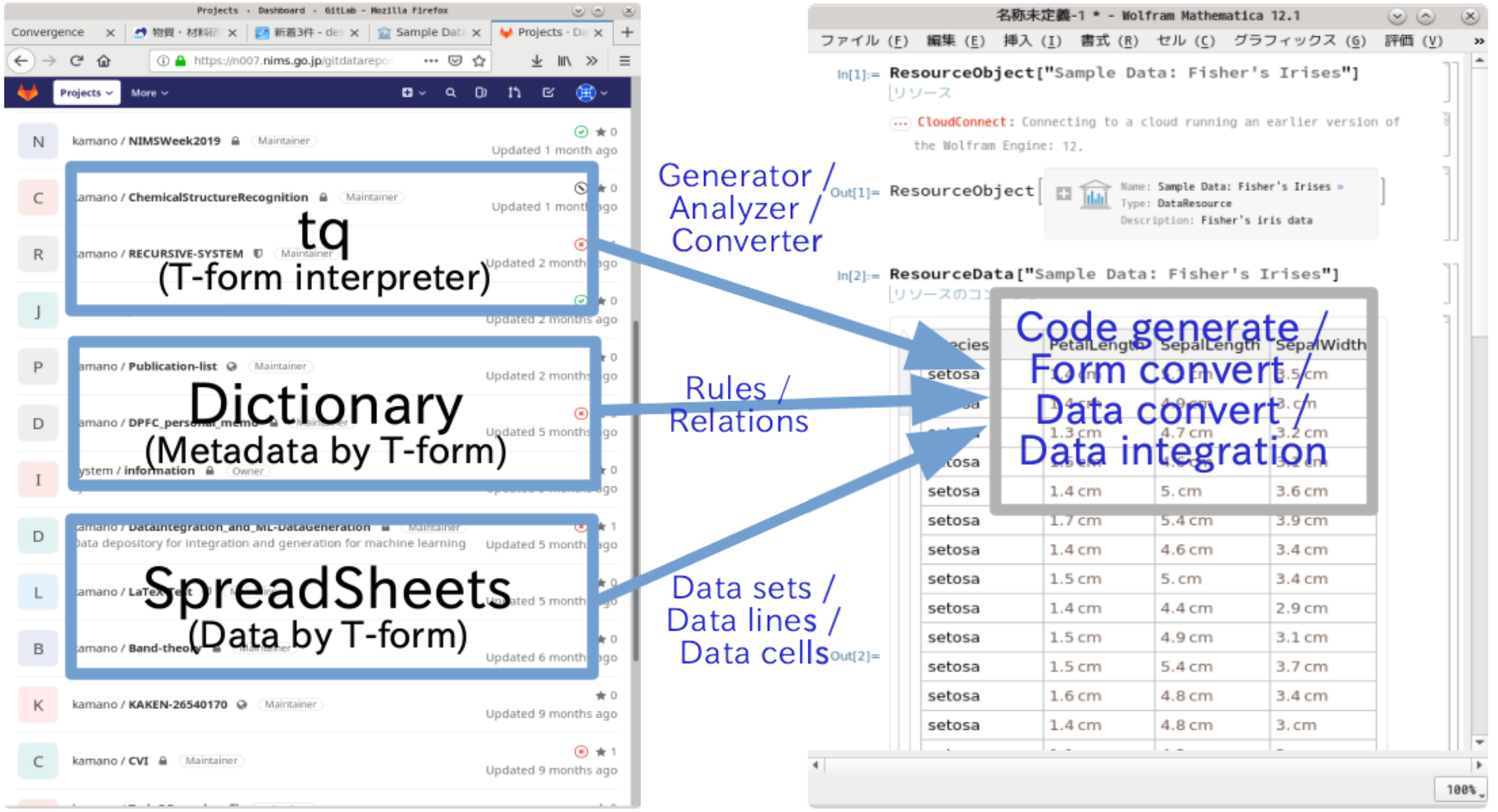
目的

以下を同一の枠組みで行う仕組みづくりの一環として、
その**記述**ための**形式言語**を定義する。

- データフェデレーション
  - NS3 (NIMS Standard Spread Sheet)
  - CSVからNS3への変換
  - S式型言語間の変換
- 機械学習の定義
  - ネットワーク（そのまま記述可能）
  - データ構造（NS3を利用）
- 辞書とそこからの知識抽出
  - 数式
  - 定義

- 目的としていなかったが派生したもの
  - 有機化合物記述

出口のイメージ


tq
(T-form interpreter)
Dictionary
(Metadata by T-form)
SpreadSheets
(Data by T-form)
Generator /
Analyzer /
Converter
Rules /
Relations
Data sets /
Data lines /
Data cells
Code generate /
Form convert /
Data convert /
Data integration
Data repository
Solver

開発の経緯 (1/7)

- 様々な分野背景を持つ材料科学において、
- 知識とデータを統一的にオペレート可能で、
- データ結合やコード挿入が可能な、
- 完全に形式化された言語が必要である

いままでにも試みられたことであるが、うまくいっていないように見受けられる。
個別の問題には対応できているものの、統一された枠組みが存在せず、トランスレーション困難であることがうまくいっていないように見える原因と考える。
記述に目を向ければ、その理由は、実装言語の特徴を吟味せずに使用し、必要な機能を複雑な実装や表現を用いて実現していることにある。

開発の経緯 (2/7)

本研究は、

- 記述対象となる知識やデータの構造
- 上記の記述に必要な言語機能

を吟味するところから始まる。

開発の経緯 (3/7)

知識の構造

知識

- 藤原[1]によれば、知識構造を表すには少なくとも均質化2部グラフ構造が必要である。
- 知識構造を表せたとして、知識の総体が無矛盾で整合であることは証明されていないため、知識の意味処理には問題を含む。

言語

- 均質化（2部）グラフを表現可能な形式言語の定義と、その構造を解釈できるパーサーの実装を行う。
- 均質化グラフ構造のパーシング（インタープリター）と意味処理（ソルバー）は分断する。※

※とはいえ、便利さには勝てず一部の意味処理を実装することとなる。

[1] 藤原譲: 「情報学基礎論の現状と展望」, 情報 知識学会誌, Vol. 9, No. 1, pp. 13–37, 1999.

開発の経緯 (4/7)

データの構造

単一要素 ⊆ ベクトル ⊆ 行列 ⊆ テンソル ⊆ 木 ⊆ グラフ ⊆ 均質化(2部)グラフ

開発の経緯 (5/7)

均質化(2部)グラフ

$E \subseteq 2V$

$V = V \cup E$

$E = E \cup V$

$\sigma: L \rightarrow E \cup V$

- 多項性:一つの E により任意の数の V のリン クが可能(ハイパーエッジ)である
- 双対性(均質化):オブジェクトがV、E、V∧E で有りうる、かつタイプを判定可能である
- 入れ子(内部構造):オブジェクトが V ∧ E の とき、V に内部構造を持ちうる(E の内部構造は自明)

直接的に処理可能な既存言語なし

開発の経緯 (6/7)

パーサーとソルバーの分離

- 前述の通りパーサーがない
- ソルバーはS式を受け入れ可能な既存システムを使えば良い
- そのためにはコンバーターが必要

言語（パーサー/コンバーター）開発

開発の経緯 (7/7)

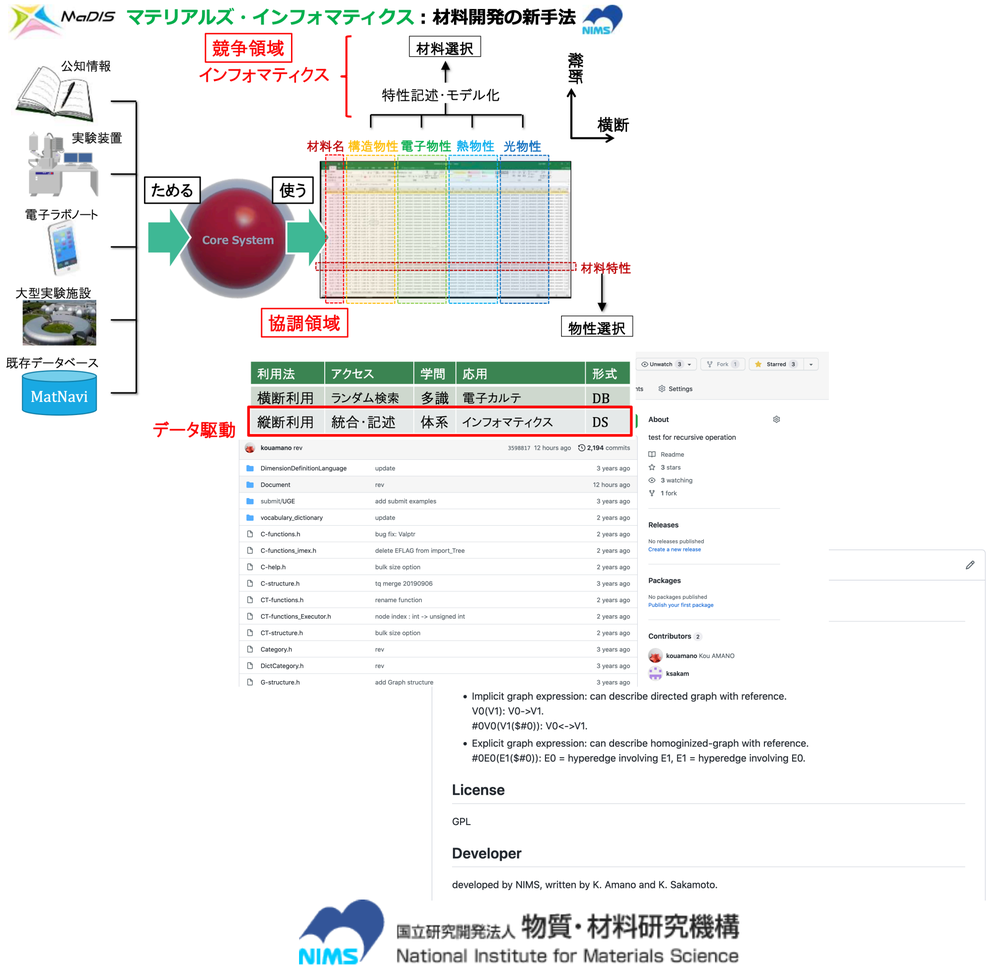
- S式
- 超簡潔
- 関係性記述に特化

E-CELL

NIAS 生物研
国立研究開発法人
National Institute of Agrobiological Sciences
農業生物資源研究所

2005

E-CELLのラッパー言語を基盤とする



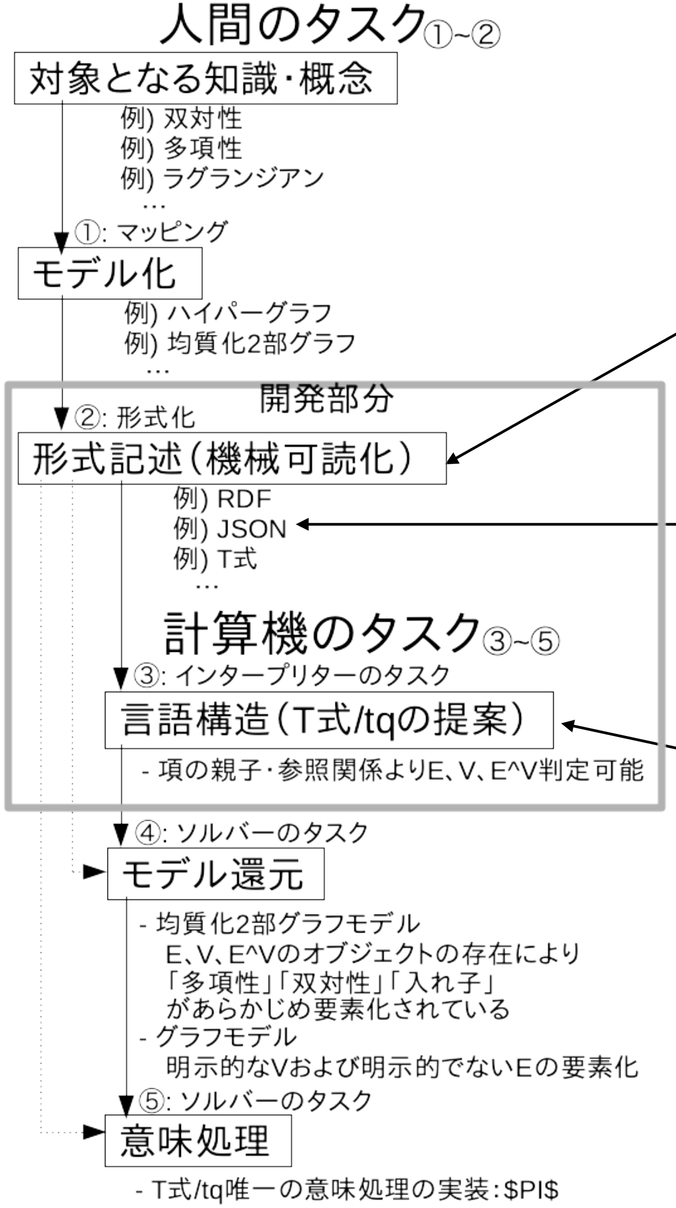
NIMS 国立研究開発法人 物質・材料研究機構
National Institute for Materials Science

2018

NS3生成のための言語として開発を開始

2021

言語のコンセプト (1/5)



言語定義（T式） ← 坂本さんの中心議題

- 言語をT-formと命名
- パーサーを T-form Query-language (tq) と命名

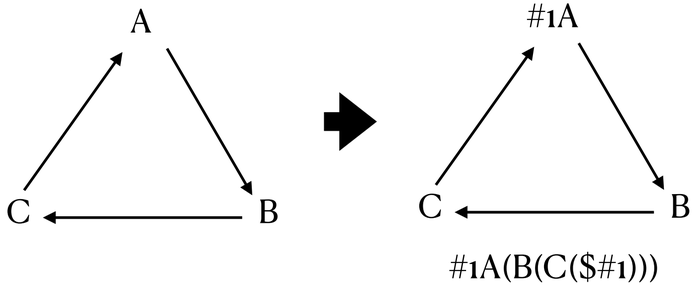
コンバーター

パーサー/
インタープリター

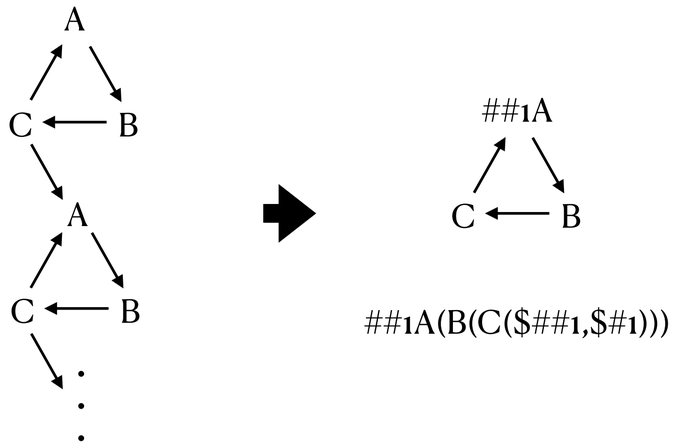
言語のコンセプト (2/5)

問題1：「木構造」である言語からどのように「グラフ」を表現するか？

答え：ラベル「#、##」と参照「$#、$##」をつかう。



サイクリック構造



無限再帰構造

(+サイクリック構造)

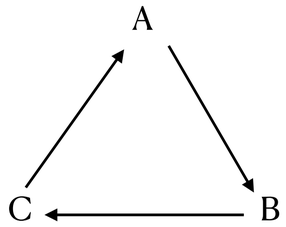
※ここでのラベルはスライド9のラベル"L"とは異なる。

※この時点では「均質化グラフ」構造は表現できていない。

言語のコンセプト (3/5)

問題2：グラフ表現の解釈に二つの方法が派生する。

答え：どちらをつかうかはユーザーに任せる。混ぜない限り問題にならない。

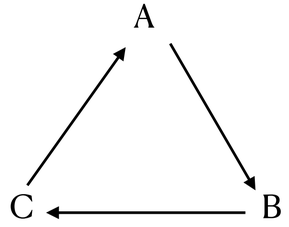


#1A(B(C($#1)))

implicit表現:

エッジを明示しない

(ハイパーエッジが使えない)



((#1A,#2B),($#2,#3C),($#3,$#1))

explicit表現:

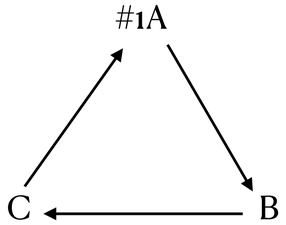
エッジを()で明示する

(全体をまとめるルートエッジが必要)

言語のコンセプト (4/5)

問題3：均質化グラフにおいてV、E、V∧Eをどのように判別する？

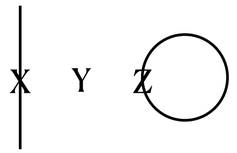
答え：explicit表現で均質化グラフの記述が可能かつ木構造を読み解けば判別可能。



A(B(C($#1)))

明示されているA、B、CがV、
無名エッジが3つある。

implicit表現:
(簡単、、、)
explicit部分がV、親子の接続がE
均質化グラフの表現はできない



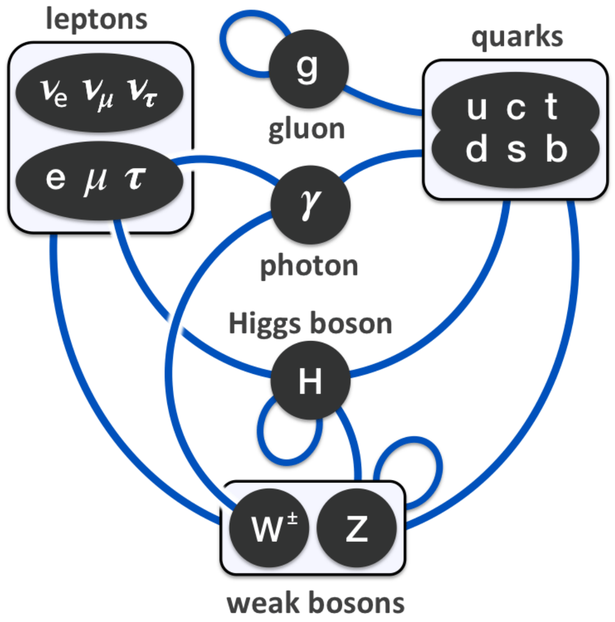
(X(),Y,#1Z($#1))

XがE、YがV、ZがV∧E。

explicit表現:
リーフノードであればV、
リーフノードでなければE、
リーフノードかつリーフノードでない場合はV∧E

言語のコンセプト (5/5)

コンセプト実証: 標準模型の構造をtqで記述




```
(
#1leptons(#7(E,M,T),(nE,nM,nT)),
#2quarks(u,d,c,s,t,b),
#3weak bosons(#8W,Z),
$#3($#1,$#2,$#3),
$#4(#4g,$#2),
#5p,
$#5($#2,$#7,$#8),
$#6(#6H,$#2,$#7,$#3)
)
```

言語仕様の詳細と実装

坂本さん

- 言語設計
- パーサー設計
- 実装の進捗

成果

- Kou Amano and Koichi Sakamoto. tq: A Comprehensive Disciplinary Language for Materials Science. NIMS week Day 2. 東京国際フォーラム 2019-10-30.
- 天野晃, 坂本浩一, 鈴木晃, 松田朝彦, 鈴木伸崇. 材料科学分野のための総合的記述言語tq. 情報メディア学会第21回研究会, 東北福祉大学 ステーションキャンパス, 2019-11-02.
- 天野晃, 坂本浩一. 知識情報構造の形式記述と解釈可能性についての展望. 情報知識学会誌 Vol. 31 No. 1 p.71, 2021.