

経緯

# MLデータセット生成システム: データ出力

- 全登録データのスプレッドシート化 → NIMS Standard Spreadsheet  
最小構成要素はセル:

\* 項目名

\* 型

\* 単位

\* 値

1 サンプル 1 ライン、ただし、セル内の木構造も可能。

- システムから同項目のデータを回収、項目名/型/単位を統一

\* 項目間関係を辞書記述し、自動的に辞書解析、項目互換を行う

\* 項目間関係の記述形式も同一のスプレッドシート形式

- アドバンス機能としてシステムから指定項目に対する変換可能項目のデータを回収、変換出力

- 何をInput/Outputにするかはユーザーの決定

# MLデータセット生成システム: 過去のサービス構成

- gitlab
  - \* gitlab-ci-ymml
- Mathematica
- tq
- Jupyterhub
- gpg

ユーザーは git と gpg があればよい。  
当初は Mathematica をメイン、tq を補助的に利用する方針。

# 現状の方針

## MLデータセット生成システム: 現状の方針

- 解析クラスタ中心
  - \* 解析クラスタでtqを動作させることを第一目標とする
  - \* NIMS Standard Spreadsheetを出力する目的は変わらない
- tqをメイン、ソルバー (Mathematica 等) はユーザー環境を利用
- 署名や信頼性確保は他のサブシステムが行う
- データ形式定義、データ、辞書をすべて同一形式(言語)で記述



## tq: 要求

- データ構造 = データ形式定義 + 辞書参照 ( 木 ) を記述可能であること
- 辞書構造 = 知識構造 ( グラフ ) を記述可能であること

## tq: 文法

$\text{< tq >::}$  ①  $\text{< label >}$  ③  $\text{< reference >}$  ④  $\text{< operator >}$  ②  $\text{< name >}$  ②  $\text{< bind >}$   
( $\text{< tq >, ...}$ )

$\text{< tq >::}$   $\text{\# <: num :> \$\# <: num :> \$ <: alp :> \$ < name > [ <: num :> ]}$  ( $\text{< tq >, ...}$ )

e.g.

$\text{\#1 \$\#2 \$Op \$Name (\#2 Name2 [2])}$

$\text{\#1 \$\#2 \$Op \$Name @ \#2 Name2 @ \#2 Name2 (\#2 Name2, [2] @ (Length, Weight))}$



tq: セルの構成

(項目名, 型 (値, 単位))

e.g.

```
(Weight, Quantity(68, kg))  
(Comment, String(This is a comment line.))  
(Comment, String("A", "B", "C"))  
(No., Numeric(1))
```

## tq: 項参照

表現	参照部	被参照部	被参照部バインド表現
$\$ \# 1 f$	$\$ \# 1$		$\$ \# 1 f$
$(\$ \# 1, \# 1)$	$\$ \# 1$	$\# 1$	$(\$ \# 1 \textcircled{\#} 1, \# 1)$
$\$ \# 1 f(\# 1 g)$	$\$ \# 1$	$\# 1 g$	$\$ \# 1 \textcircled{\#} 1 g(\# 1 g)$
$\# 1 f(\$ \# 1 g)$	$\$ \# 1$	$\# 1 f$	$\# 1 f(\$ \# 1 g \textcircled{\#} 1 f)$

## tq: データバインド

表現	データ	データバインド表現
[1]	L, W, 22, 3, 21, 5	[1]⓪(L)
[2]	L, W, 22, 3, 21, 5	[2]⓪(L, W)
[2]([2])	L, W, 22, 3, 21, 5	[2]([2]⓪(L, W, 22, 3))
H1[2](H2[2])	L, W, 22, 3, 21, 5	H1[2](H2[2]⓪(L, W, 22, 3))

## tq: 項参照とデータバインドを使ったスプレッドシート の構成

データ: Length, Weight, mm, kg, 1, 2, 322, 4, 5, 68

入力形式定義: (#1[2], #2[2], [3](#4[2]))

出力形式定義: \$PI\$(\$#1, Quantity(\$#4, \$#2))

出力:

```
(
  (
    (Length, Quantity(1, mm)), (Weight, Quantity(2, kg))
  ),
  (
    (Length, Quantity(322, mm)), (Weight, Quantity(4, kg))
  ),
  (
    (Length, Quantity(5, mm)), (Weight, Quantity(68, kg))
  )
)
```

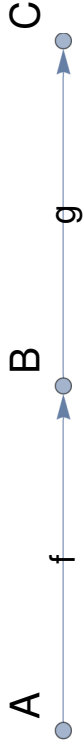
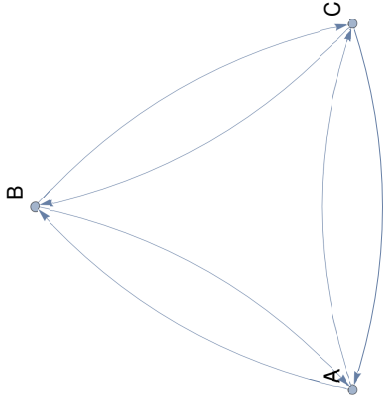
改行は便宜上

## tq: トリプル表現

	P(Arrow)	S(Dom)	O(Cod)
P(S,O)	P	S	O
(S(\$#1P),#1O)	P	S	O
S(\$#1P) / (S(\$#1P),)	P	S	<無名>
S(O)	<無名>	S	O
(,)	<無名>	<無名>	<無名>

tq: トリプルと頂参照を使ったグラフ表現

$\$G\$((\lt \text{オブジェクトリスト} \gt), (\lt \text{トリプルリスト} \gt))$

表現	グラフ構造		隣接行列
$((\#1A, \#2B, \#3C),$ $(f(\$ \#1, \$ \#2),$ $g(\$ \#2, \$ \#3)))$			<pre> ,2,3,,[f:6],7,8,,, ,,,,,[6-&gt;:\$#1:7-&gt;2],,,, ,,,,,[6-&gt;:\$#2:8-&gt;3],,,, ,,,3,4,,,,[g:9],10,11, ,,,,,,[9-&gt;:\$#2:10-&gt;3],, ,,,,,,,[9-&gt;:\$#3:11-&gt;4],, </pre>
$((\#1A, \#2B, \#3C),$ $((\$ \#1, \$ \#2),$ $(\$ \#1, \$ \#3),$ $(\$ \#2, \$ \#1),$ $(\$ \#2, \$ \#3),$ $(\$ \#3, \$ \#1),$ $(\$ \#3, \$ \#2)))$			<pre> ,,,,,[6-&gt;:\$#1:7-&gt;2],,,,,,, ,,,,,[6-&gt;:\$#2:8-&gt;3],,,,,,, ,,,,,[9-&gt;:\$#1:10-&gt;2],,,,,,, ,,,,,[9-&gt;:\$#3:11-&gt;4],,,,,,, ,,,,,[12-&gt;:\$#2:13-&gt;3],,,,,,, ,,,,,[12-&gt;:\$#1:14-&gt;2],,,,,,, ,,,,,[15-&gt;:\$#2:16-&gt;3],,,,,,, ,,,,,[15-&gt;:\$#3:17-&gt;4],,,,,,, ,,,,,[18-&gt;:\$#3:19-&gt;4],,,,,,, ,,,,,[18-&gt;:\$#1:20-&gt;2],,,,,,, ,,,,,[21-&gt;:\$#3:22-&gt;4],,,,,,, ,,,,,[21-&gt;:\$#2:23-&gt;3],,,,,,, </pre>

## tq: 辞書表現

グラフ:  $\$G\$((\langle \text{オブジェクトリスト} \rangle), (\langle \text{トリプルリスト} \rangle))$   
↓ そのまま!! ただし、辞書トリプルのSとOには木構造を許す  
辞書:  $\$D\$((\langle \text{オブジェクトリスト} \rangle), (\langle \text{トリプルリスト} \rangle))$

e.g.  
 $((\#1A, \#2B, \#3C), (f(\$ \#1, \$ \#2), \$X\$near(\$ \#2, \$ \#3),$   
 $\$def\$(\underline{f}, ((a, b), \$eq\$ (a, b))), \dots))$   
 $((\#1A, \#2B, \#3C), (f(\$ \#1, \$ \#2), \$X\$near(\$ \#2, \$ \#3),$   
 $\$def\$(\underline{f}, ((a, b), \$eq\$ (a, \$pow\$ (b, 2))))), \dots))$

トリプル 被定義項 定義

tq: 辞書表現(圏)

e.g.  
 $f : A \rightarrow B :: \text{\texttt{\$arrow\$(f(A),B)}}$   
 $g : B \rightarrow C :: \text{\texttt{\$arrow\$(g(B),C)}}$   
 $\text{gof} : A \rightarrow C :: \text{\texttt{\$arrow\$(g(f(A)),C)}}$



## 開発状況: 実装済み部分

- 木構造パーサー
- グラフ構造パーサー（ラベルによる項参照）
- データバインド
- テータリストの内積化
- アンパックス（木構造の平坦化）
- クォーテイング
- リテラライズ（オペレーション回避）