Parsing/interpreting Stata into SDTL

C2Metadata

Ole Voldsæter <ole.voldsater@nsd.no>

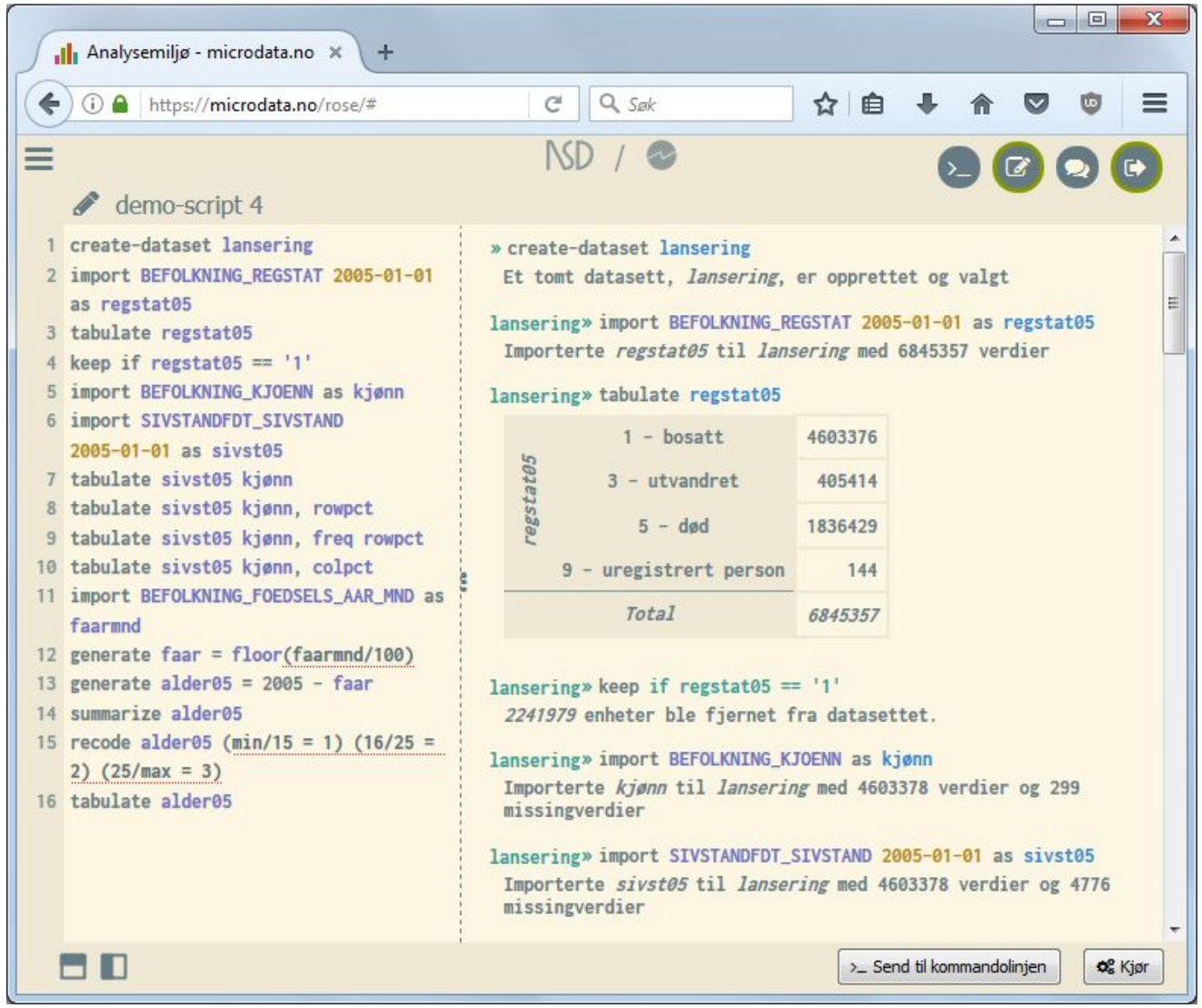
Ørnulf Risnes <ornulf.risnes@nsd.no>

Montréal 30 May 2018/IASSIST 2018

NORSK SENTER FOR FORSKNINGSDATA

Prior experience

Development of grammar and execution of transformational and analytical commands in RAIRD / microdata.no



NORSK SENTER FOR FORSKNINGSDATA

The Stata language

Mostly declarative

Extremely flexible and powerful

«Convenience syntax» with abbreviations

```
tab | tabu | tabul | tabula | tabulat | tabulate
```

Most «state changes» created by commands can be inferred from programs/scripts alone (but not all)

Has macros (statements that generate program snippets)

Not necessarily 100% consistent grammar between different commands(?)

NSD
NORSK SENTER FOR FORSKNINGSDATA

Handling state changes (interpreter)

```
DS1: V1 V2 V3 //initial «virtual» state
```

V1	V2	V3
2	3	0
1	4	0

```
>> drop V2
```

```
DS2: V1 V3 //new virtual state
```

V1	V3
2	0
1	0

```
>> generate V2 = V1 * 100
```

```
DS3: V1 V3 V2 //new virtual state
```

V1	V3	V2
2	0	200
1	0	100

NORSK SENTER FOR FORSKNINGSDATA

Macros

```
replace a = b + 5 //Regular command

foreach op in "+" "-" "*" {
  replace a = b `op' 5
}
```

...becomes...

```
replace a = b + 5
replace a = b - 5
replace a = b * 5
```

NORSK SENTER FOR FORSKNINGSDATA

The «Stata parser» is really…

- A parser

- An *interpreter* with support for on-the-fly macro precompilation

- Demo: https://tinyurl.com/stataparser

NORSK SENTER FOR FORSKNINGSDATA

Biggest remaining challenges

`egen`

`by/bysort`

`merge`

+ unknown unknowns

NORSK SENTER FOR FORSKNINGSDATA

Unsolvable problems

Commands that use data values as input

`reshape wide`

Stata 15 help for reshape

[D] **reshape** -- Convert data from wide to long form and vice versa

Syntax

```
    Overview

            long
        +------------+                           wide
        | i  j  stub |                    +----------------+
        |------------|                    | i  stub1 stub2 |
        | 1  1   4.1 |     reshape        |----------------|
        | 1  2   4.5 |   <--------->      | 1    4.1   4.5 |
        | 2  1   3.3 |                    | 2    3.3   3.0 |
        | 2  2   3.0 |                    +----------------+
        +------------+
```

Illustration from Stata 15 help

NORSK SENTER FOR FORSKNINGSDATA