CLARIN
Common Language Resources and Technology Infrastructure

The proliferation of CMDI components and profiles

Twan Goosen & Dieter Van Uytvanck
Max Planck Institute for Psycholinguistics
Dieter.VanUytvanck@mpi.nl
CMDI fuure workshop, Utrecht
2013-10-14

CMDI providers

- Currently, about 14 CMDI-providers: see Center Registry:
  - https://centerregistry-clarin.esc.rzg.mpg.de/
- All CLARIN B-centres (currently about 20 candidates) will provide it
- At this moment (public!):
  - 133 profiles
  - 772 components

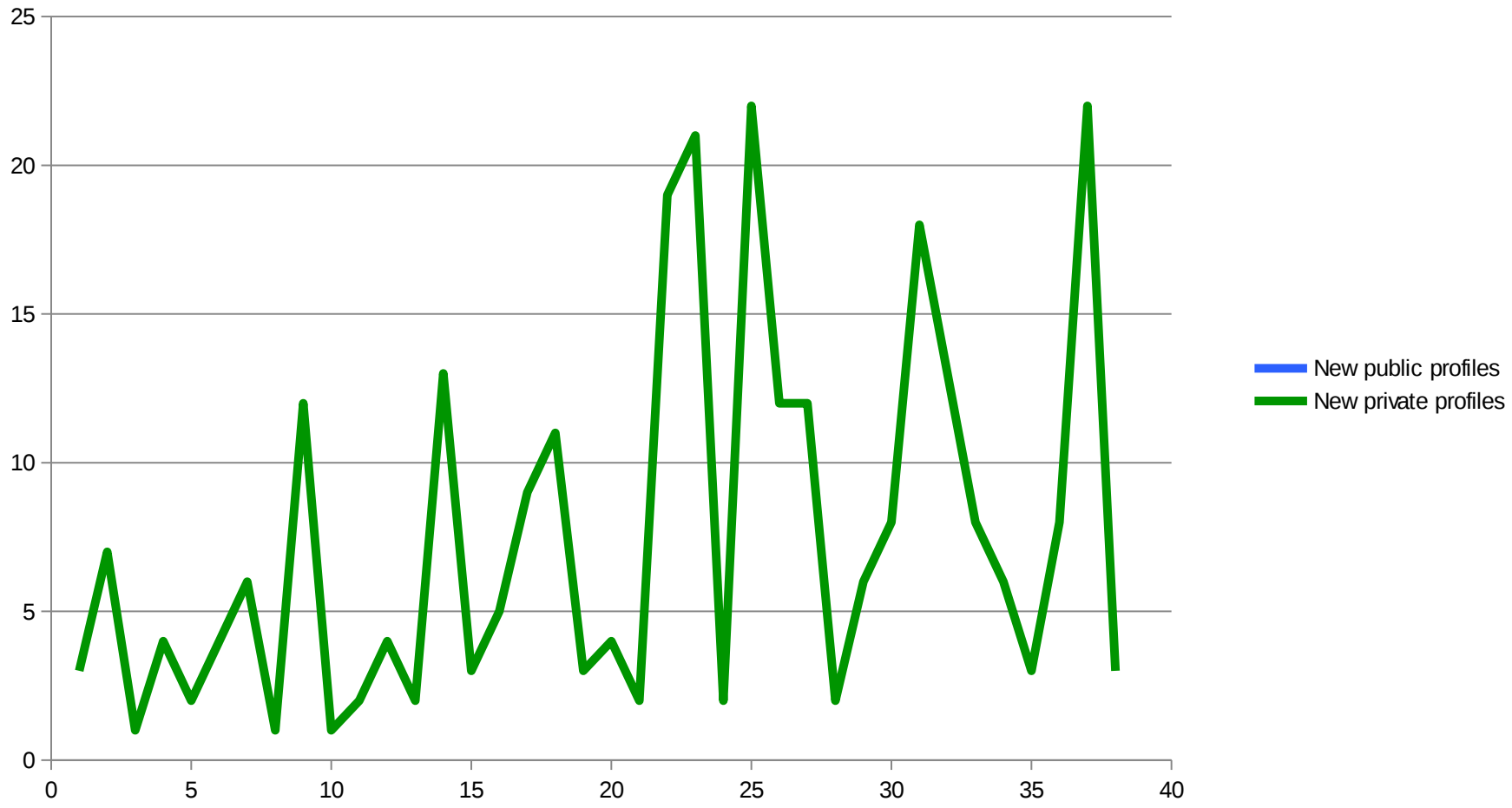
Proliferation by example

- Component search on „location“: 17 components
- Profile search on „corpus“: 29 profiles

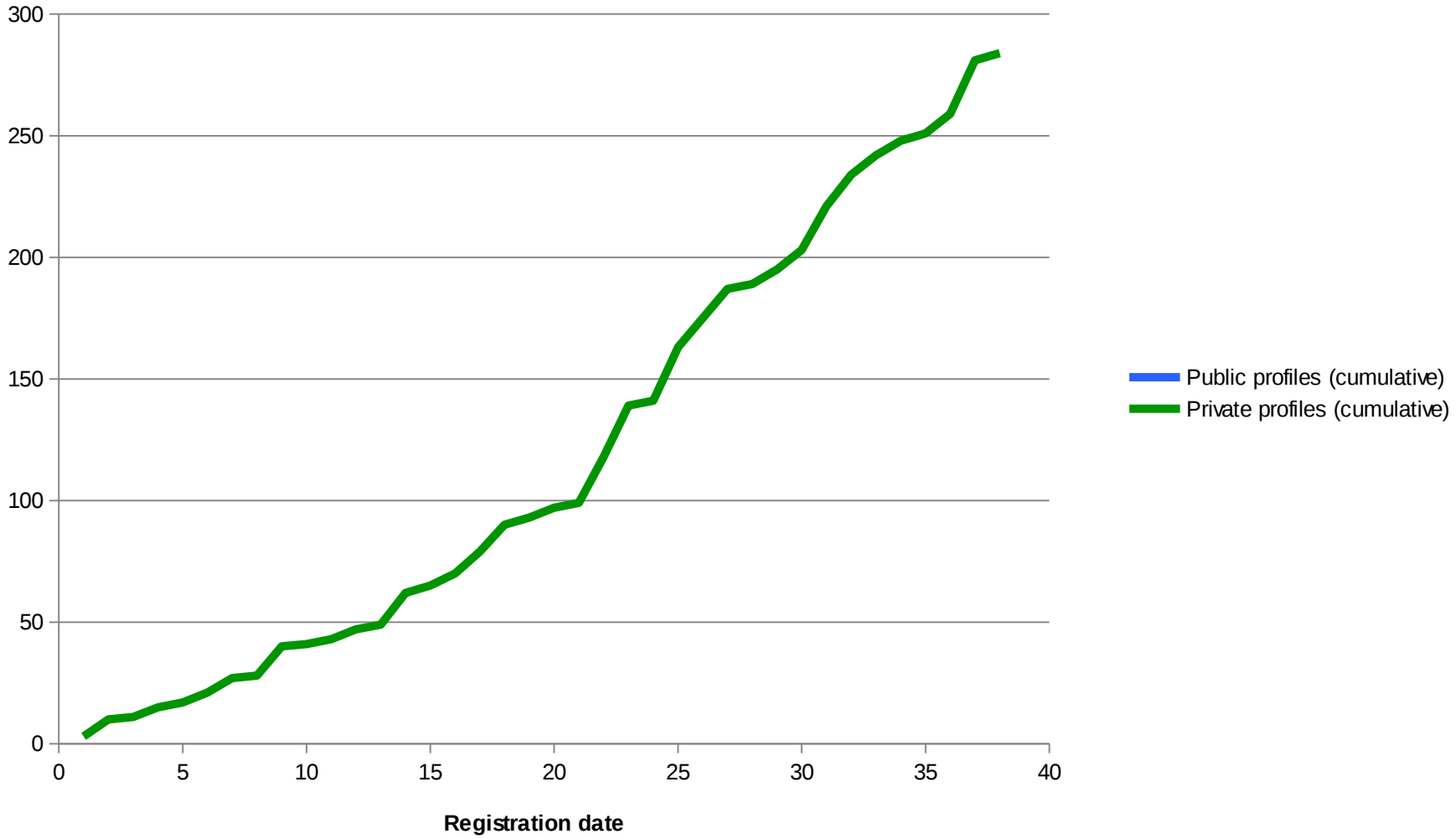
Some structural numbers

- Data collected 7 October 2013
- Data based on component registry database on catalog.clarin.eu
- Private/public reflects current state
- Currently deleted profiles and components are ignored

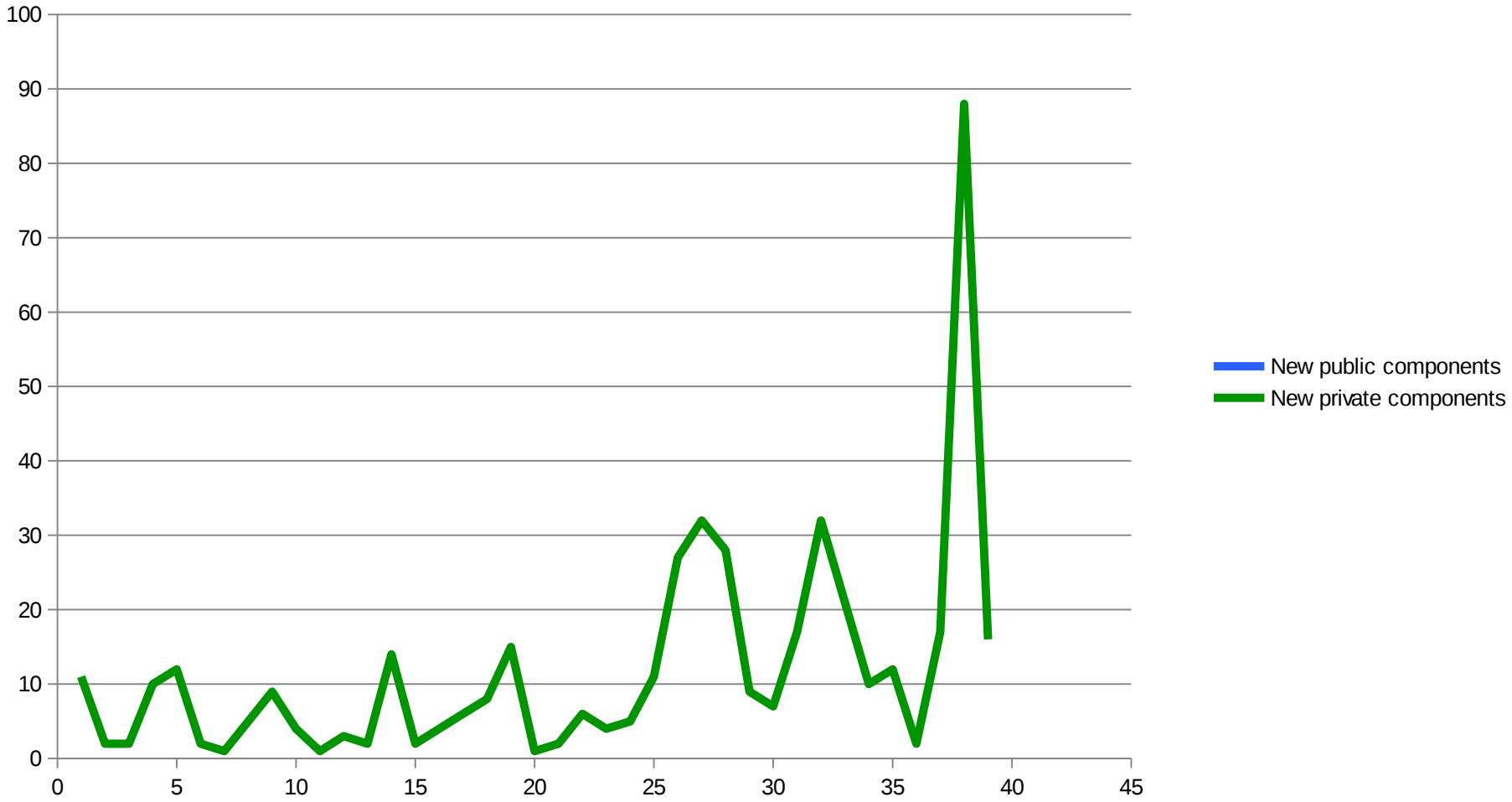
New profiles over time



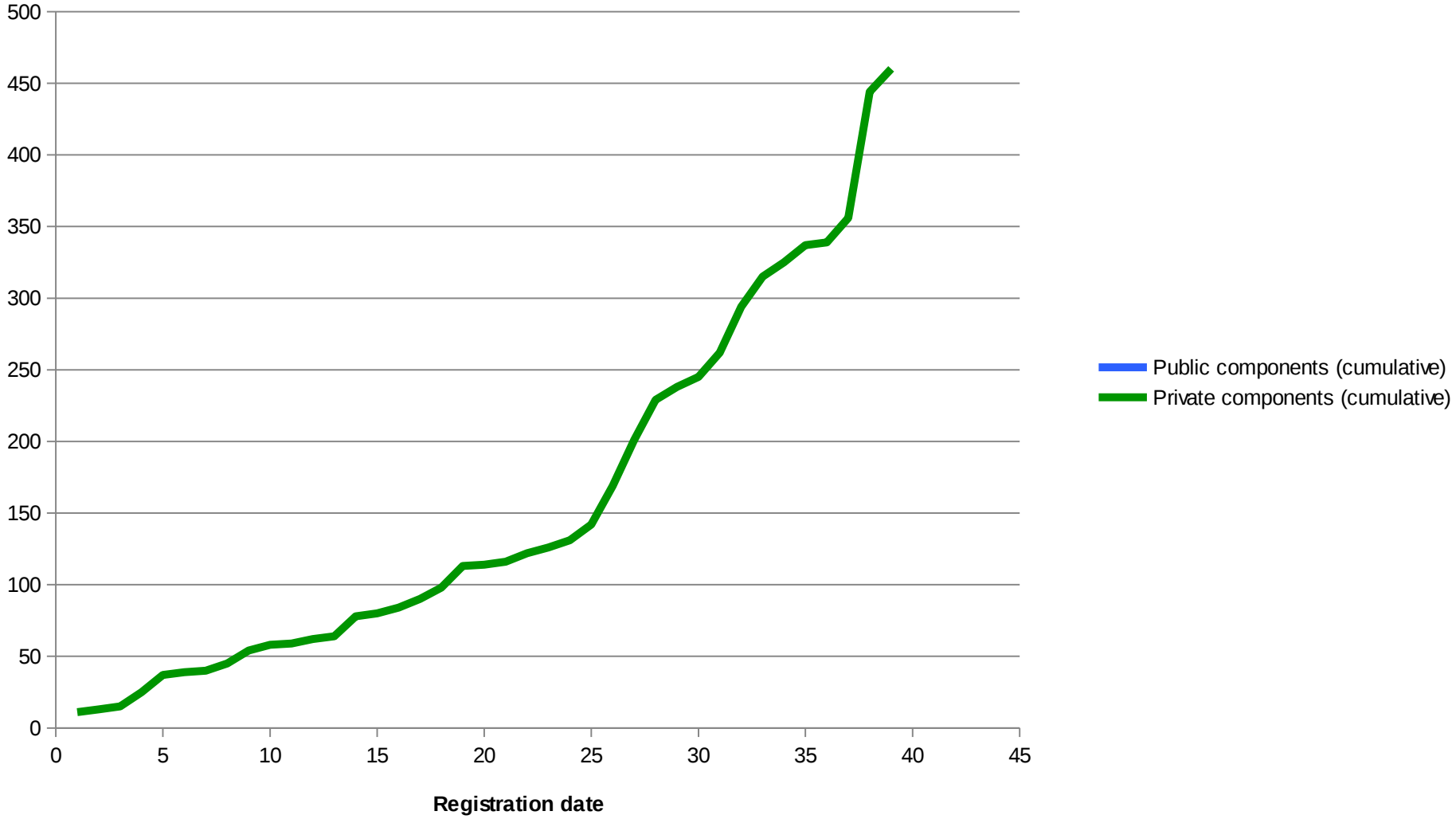
Total profiles over time



New components over time



Total components over time



Top 10 profile creators

	#public profiles
Folkert de Vriend	24
nalida	23
Dieter Van Uytvanck	12
Thomas Eckart	7
Eric Sanders	6
Griet Depoorter (INL)	5
CLARIN (cmdi@clarin.eu)	4
Penny Labropoulou	4
Peter M. Fischer	3
Peter Withers	3

	#private profiles
Marc Kemps Snijders	33
Twan Goosen	16
KB Editor Clarin	16
Henk Matthezing	14
Dieter Van Uytvanck	11
Kirsten Vis	10
gusapoap@gu.se	10
CLARIN (cmdi@clarin.eu)	7
Guido van Dongen	7
paucas	7

Top 10 component creators

	#public components
nalida	173
Penny Labropoulou	147
CLARIN (cmdi@clarin.eu)	120
misutka	36
vdelint	33
Eric Sanders	31
Dirk Roorda	20
Dieter Van Uytvanck	20
mwindhouwer	20
Maaske Treurniet	17

	#private components
Eline Westerhout	51
Guido van Dongen	46
DK-CLARIN User	36
Hanna Hedeland	27
Twan Goosen	26
lucea-groep	17
gusapoap@gu.se	17
Dieter Van Uytvanck	15
Rosemary Orr	14
Menzo Windhouwer	14

Suggestions for further analysis

- Clustering of
  - profile * component
  - component * component
  - component * datcat

Some SMC statistics

- Source: <http://clarin.aac.ac.at/smc-browser/>:
- created 2013-10-05+02:00
- Profiles 158
- Components 6681 distinct Components 994
- Elements 15103 distinct Elements 2817
- Elements with DatCats 2109
- Elements without DatCats 719
- ratio of elements without DatCats 25.52 %
- used Concept 636
- available Concepts (in Metadata profile or used in CMD) 1148

Conclusions

- The anticipated proliferation is taking place
- Only a part is visible: mind the private components and profiles
- Detection can be partially done automatically...
- ... but not the rationalisation
- Time to take an active role in this!