PanelWhiz Plugins: Automatic Vector-Oriented Data Cleaning for Large Scale Panel Datasets

Markus Hahn

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI Essen)

5th German Stata Users Group meeting
(April 2, 2007)

"Employed Again Last Year" (Source: GSOEP)

Year 2004	Year 2005
[1] First Job	[1] First Job
[2] Job After Break	[2] Job After Break
[3] Job With New Employer	[3] Job With New Employer
[4] New Job-Self Employed	[4] Company Taken Over
[5] Company Taken Over	[5] Changed Job,Same Firm
[6] Changed Job,Same Firm	[6] New Job-Self Employed

RWI ESSEN

"Number Employees in Firm" (Source: GSOEP)

Year 2004	Year 2005
[1] LT 5	[1] LT 5
[2] GE 5 and LT 20	[2] GE 5 and LT 20
[3] GE 20 and LT 100	[3] GE 11 and LT 20
[4] GE 100 and LT 200	[4] GE 20 and LT 100
[5] GE 200 and LT 2000	[5] GE 100 and LT 200
[6] GE 2000	[6] GE 200 and LT 2000
[7] Self-Employed	[7] GE 2000
	[8] Self-Employed

RWI ESSEN

What a plugin does

- Ensures time consistency of variables
- Data cleaning / Conversion (DM⇒EURO, ...)
- Creating new variables (real values, ...)

How a plugin works

- Executable ado-files for each item
- Plugin only describes what actions and when these actions are to be performed
- Plugin system executes the necessary actions
- Auto-documentation of what is done

RWI ESSEN

Example Plugin 1/2

```
program define bp_lglohn

*! ------------------------------------------------------------------------------
*! Plugin    : [lglohn] Brutto Lohn- und Gehaltssumme
*! Author    : Markus Hahn
*! Email     : hahn@rwi-essen.de
*! Web       : http://www.panelwhiz.eu
*! Copyright : Copyright 2006. All rights reserved. Unauthorized copying prohibited.
*! Version   : 1.0
*! Years     : a b c d e f g h i j k l
*! Possible  : a b c d e f g h i j k l
*! Changes   : a b c d e f g h i j k l
*! ------------------------------------------------------------------------------

pwplugin begin, item(lglohn) year(2004)
    .action.dm2euro

    .plugin.newitem
    .action.dm2euro
    .action.real, thisyear
pwplugin end

end
```

Example Plugin 2/2

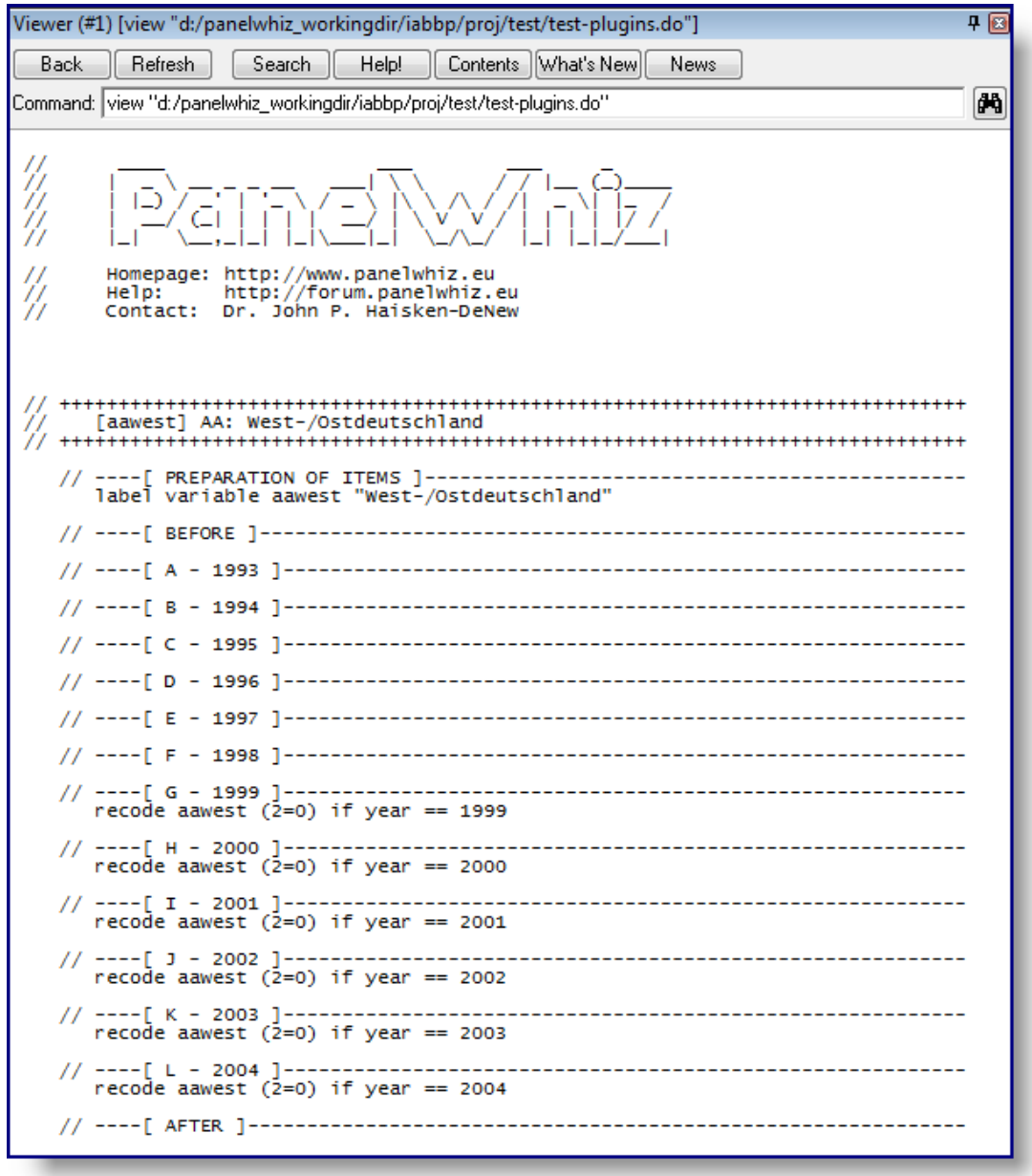
```
program define bp_aawest

*! ------------------------------------------------------------------------------
*! Plugin    : [aawest] AA: West-/Ostdeutschland
*! Author    : Markus Hahn
*! Email     : hahn@rwi-essen.de
*! Web       : http://www.panelwhiz.eu
*! Copyright : Copyright 2006. All rights reserved. Unauthorized copying prohibited.
*! Version   : 1.0
*! Years     : a b c d e f g h i j k l
*! Possible  : _ _ _ d e f g h i j k l
*! Changes   : _ _ _ _ _ _ g h i j k l
*! ------------------------------------------------------------------------------

pwplugin begin, item(aawest) year(2004)
    .action.recode (2=0), g h i j k l
pwplugin end

end
```

Automatically generated plugin-do-file (Auto-doc)



Why we do not use normal ado-files

- Relative high complexity of code (loops, ...)
  $\Rightarrow$ Not that easy to understand
- The code should be self-explanatory (transparency)
- Simpler code could is easier to process
  $\Rightarrow$ for generation purposes ("plugin wizard")
  $\Rightarrow$ for auto-documentation purposes beyond log-files

RWI ESSEN

Implementation of the plugin system

- Stata and Mata code
- Class system (`*.class`)
- Dialog boxes (`*.dlg`)
- On-the-fly generation of smcl-files (Output for Plugin Wizard)

RWI ESSEN

Recoding a variable

```
.action.recode (2=0), a b c d e
```

- Similar syntax to stata's recode
- Item variable gets recoded: 2 becomes 0
- Action is done for the waves a-e

Changing a variable

```
.action.replace wage/12, a b c d e
```

- Similar syntax to stata's replace
- Item variable gets divided by 12
- Action is done for the waves a-e

RWI ESSEN

Handling money variables

`.action.dm2euro`

- Item variable will be converted to EURO (years < 2002)

`.action.real, thisyear`

- Item variable will be converted to real values
- The deflator (price index) of the current year is used

RWI ESSEN

Label definition

.action.label var(Variable Label)

- Item variable gets new variable label

.action.label val(1 "Yes" 0 "No")

- Item variable gets new value label

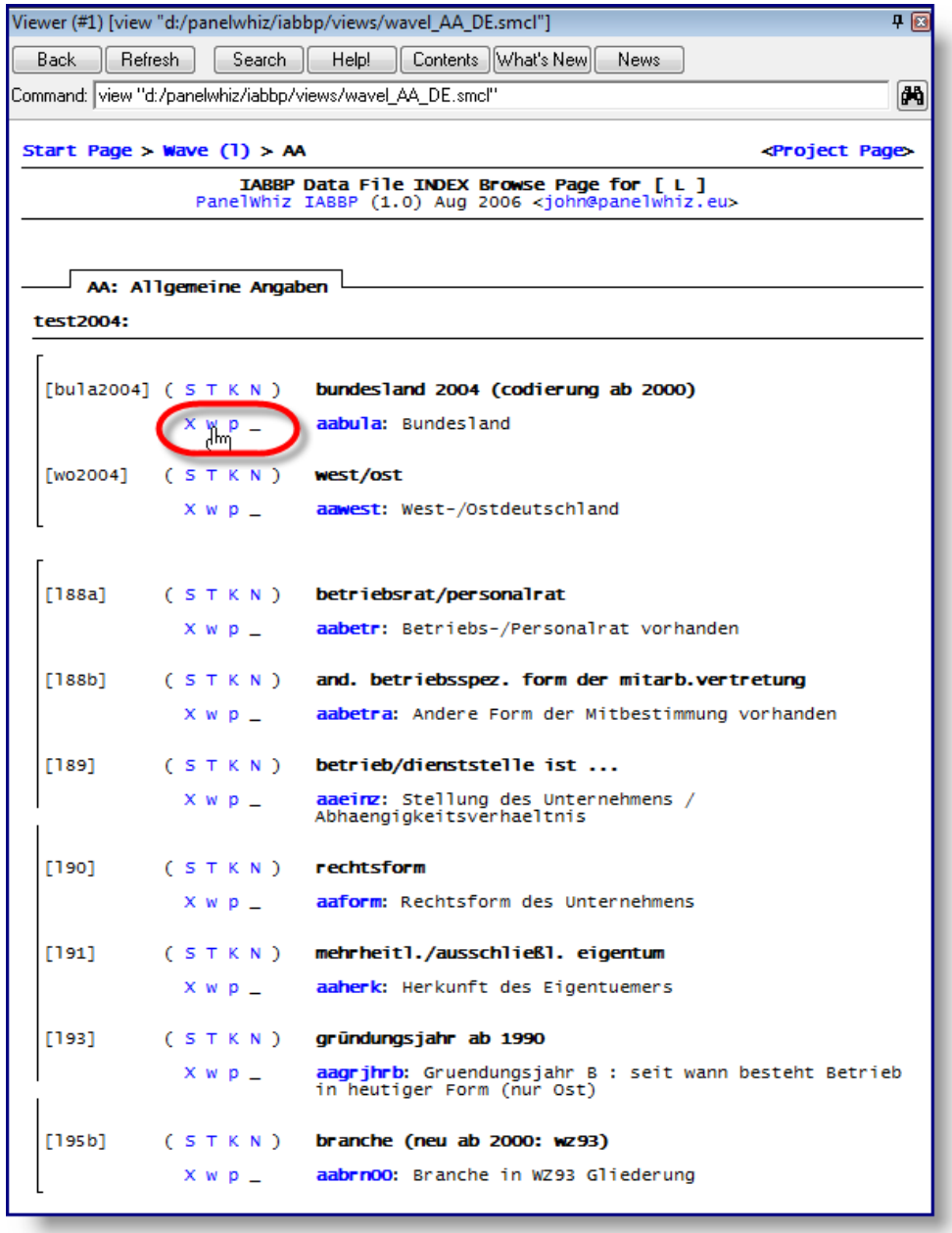
Creating new variables

`.plugin.newitem`

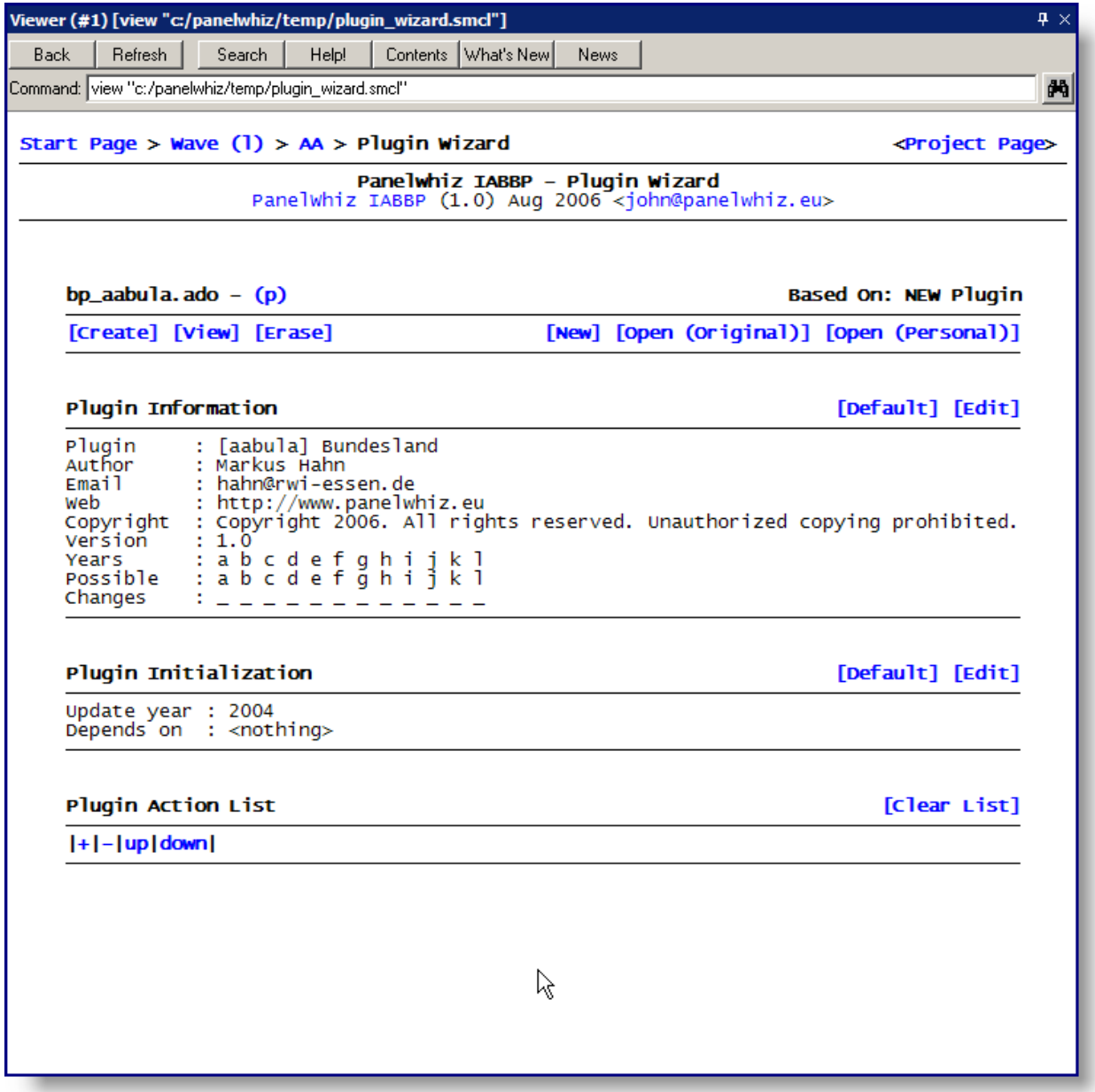
- A new variable will be created
- Variable name will be automatically assigned
- Following actions/commands will refer to new variable

Plugin Wizard

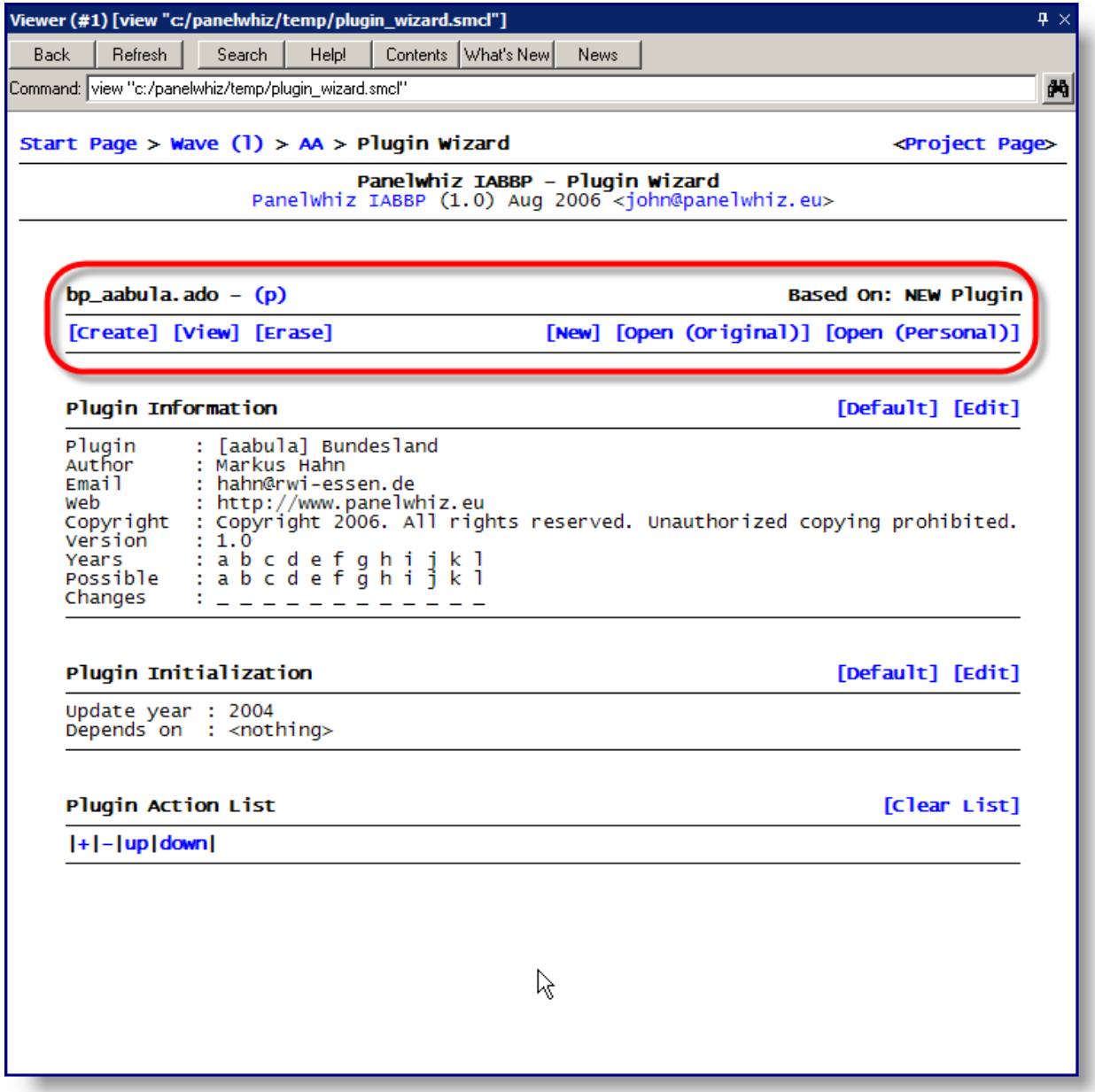
- Tool for creating plugins by "clicking"
  $\Rightarrow$ No manual typing necessary
- User interface: generated smcl-files + dialog boxes
- Supports the complete syntax of plugin code
- Once created plugins can be re-read



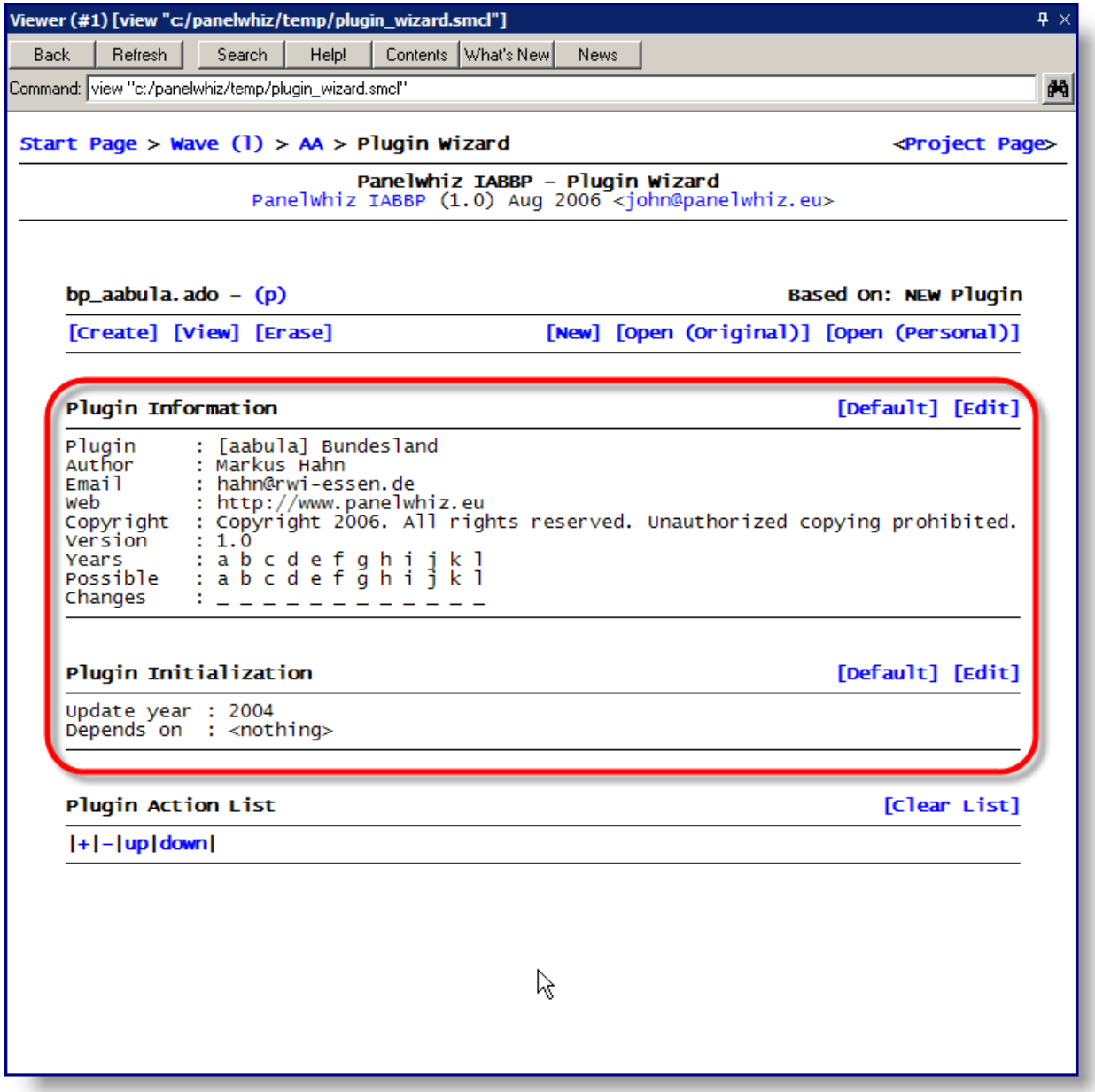
Data Browse Page: Start Plugin Wizard



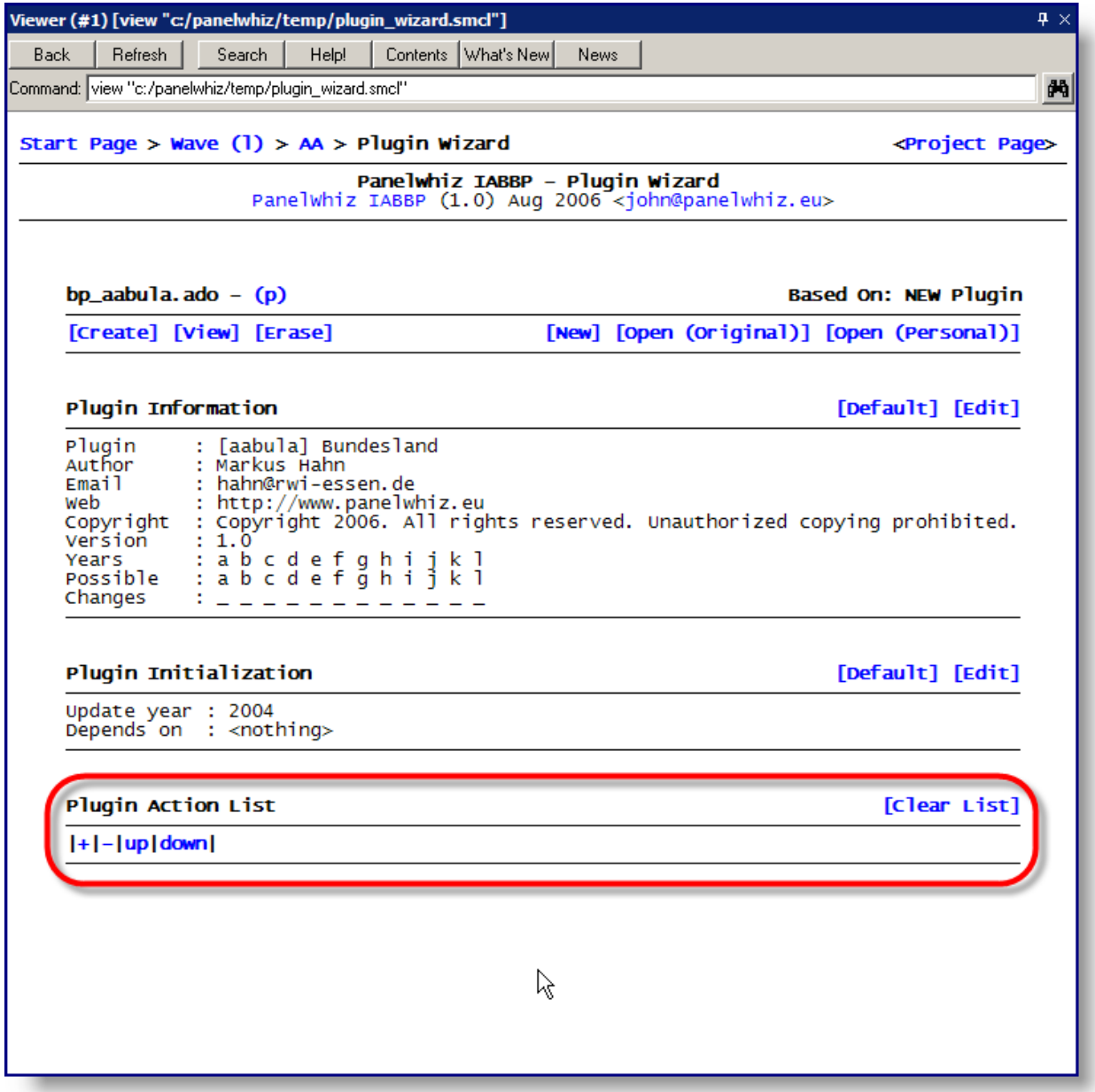
Plugin Wizard



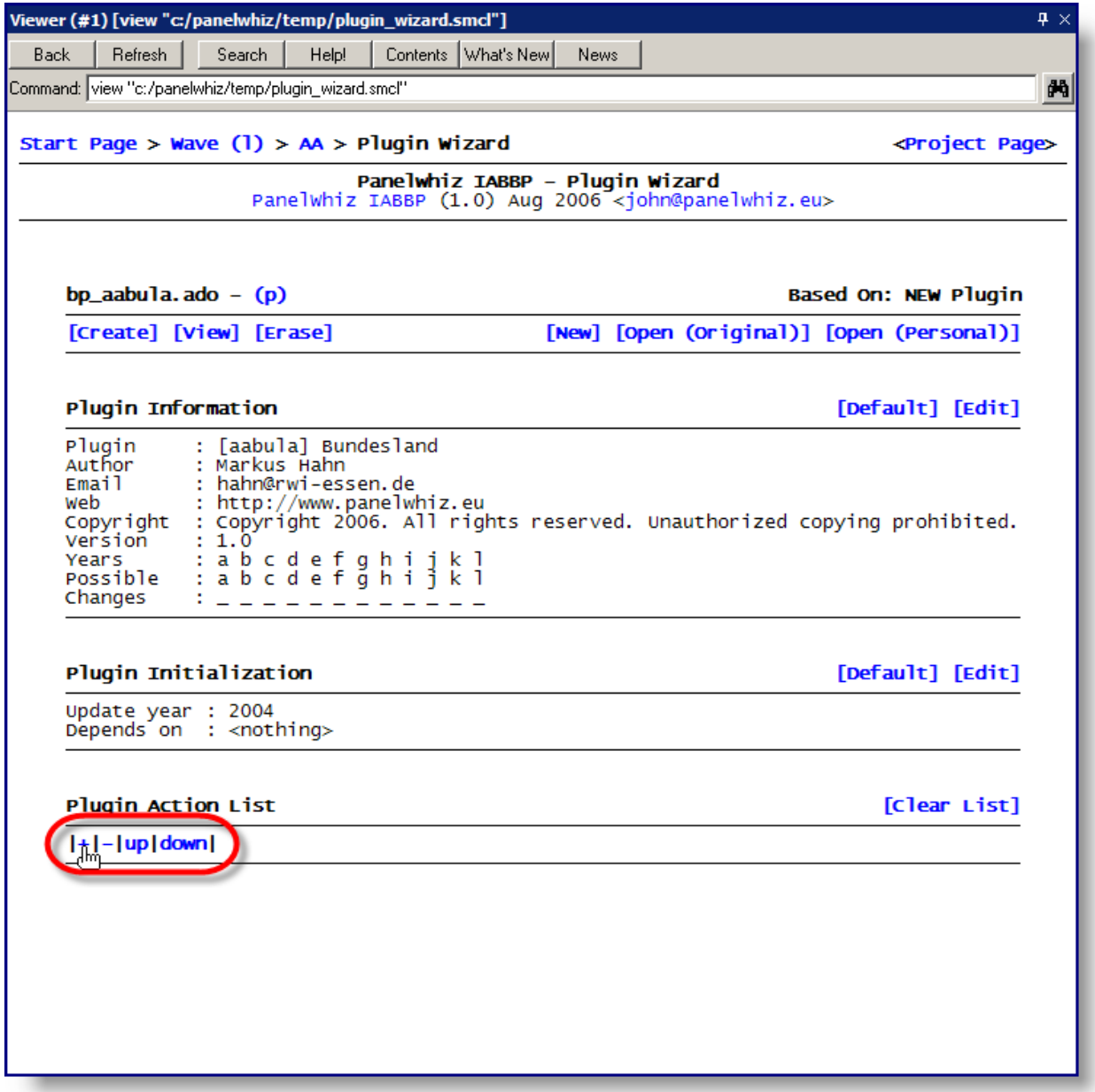
Create, view, erase, open, ...



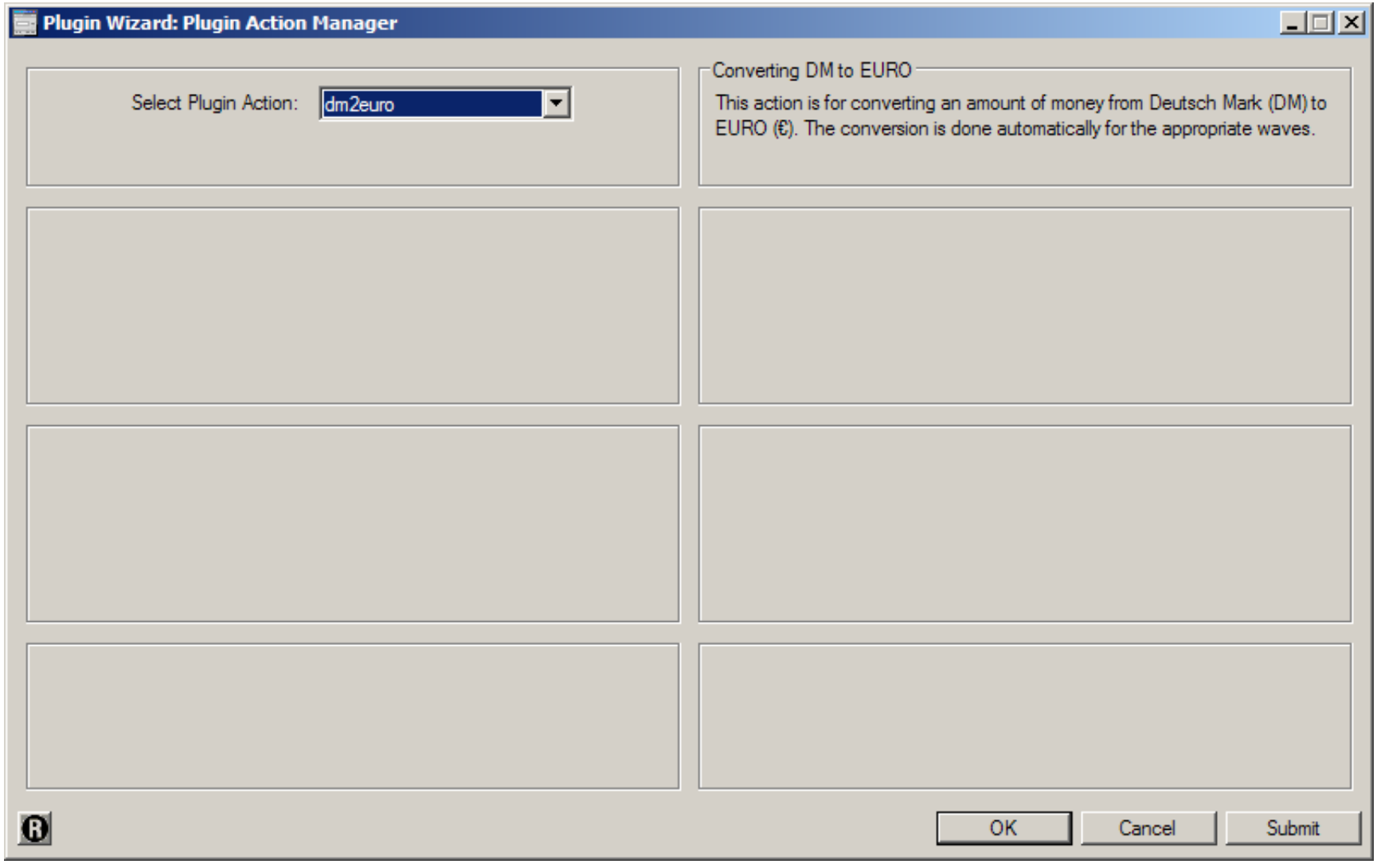
Information about plugin


Viewer (#1) [view "c:/panelwhiz/temp/plugin_wizard.smcl"]
Back
Refresh
Search
Help!
Contents
What's New
News
Command: view "c:/panelwhiz/temp/plugin_wizard.smcl"
Start Page > Wave (l) > AA > Plugin Wizard
<Project Page>
Panelwhiz IABBP - Plugin Wizard
Panelwhiz IABBP (1.0) Aug 2006 <john@panelwhiz.eu>
bp_aabula.ado - (p)
Based On: NEW Plugin
[Create] [View] [Erase]
[New] [Open (Original)] [Open (Personal)]
Plugin Information
[Default] [Edit]
Plugin : [aabula] Bundesland
Author : Markus Hahn
Email : hahn@rwi-essen.de
Web : http://www.panelwhiz.eu
Copyright : Copyright 2006. All rights reserved. Unauthorized copying prohibited.
Version : 1.0
Years : a b c d e f g h i j k l
Possible : a b c d e f g h i j k l
Changes : _ _ _ _ _ _ _ _ _ _ _ _
Plugin Initialization
[Default] [Edit]
Update year : 2004
Depends on : <nothing>
Plugin Action List
[Clear List]
|+|-|up|down|

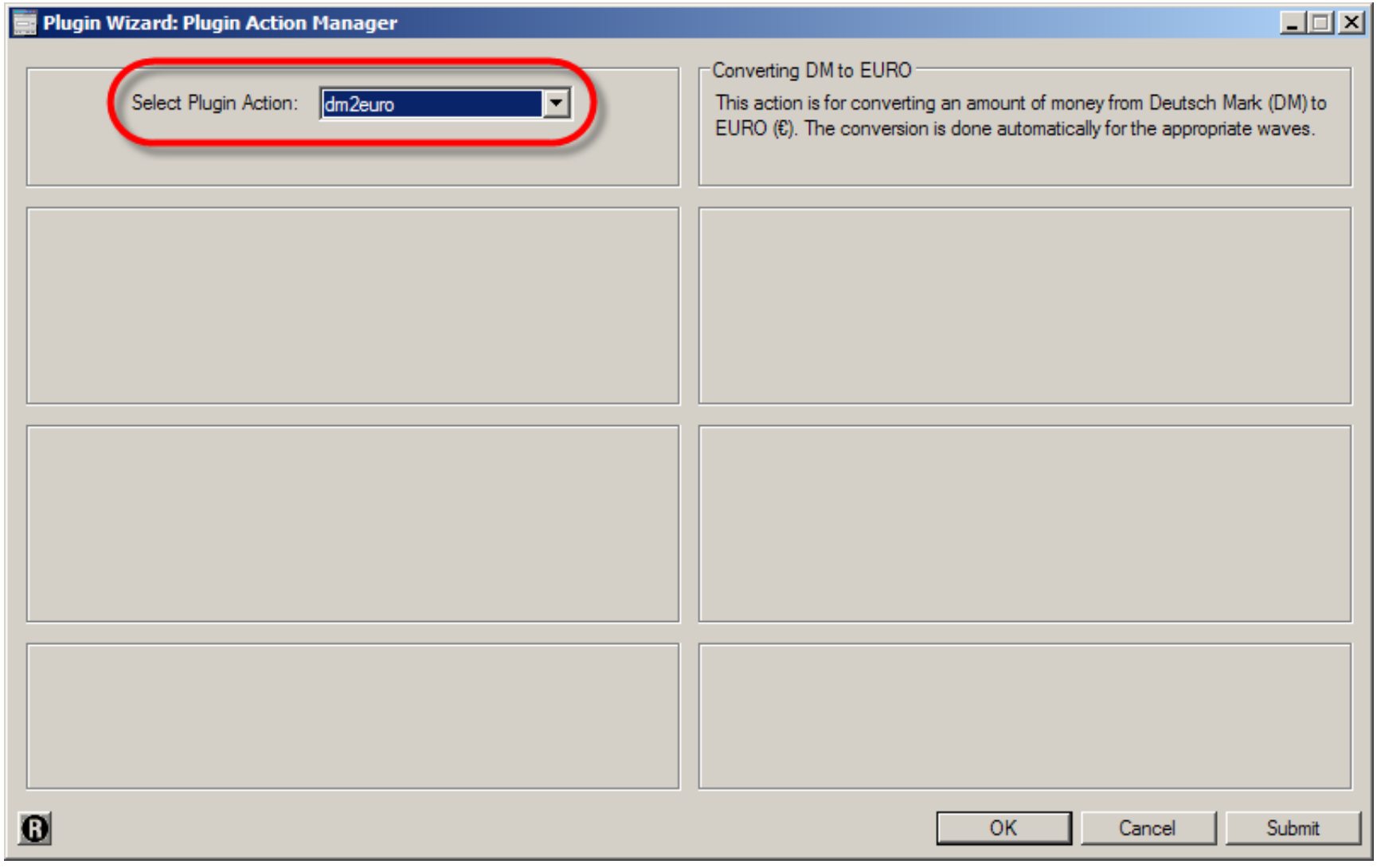
Action List



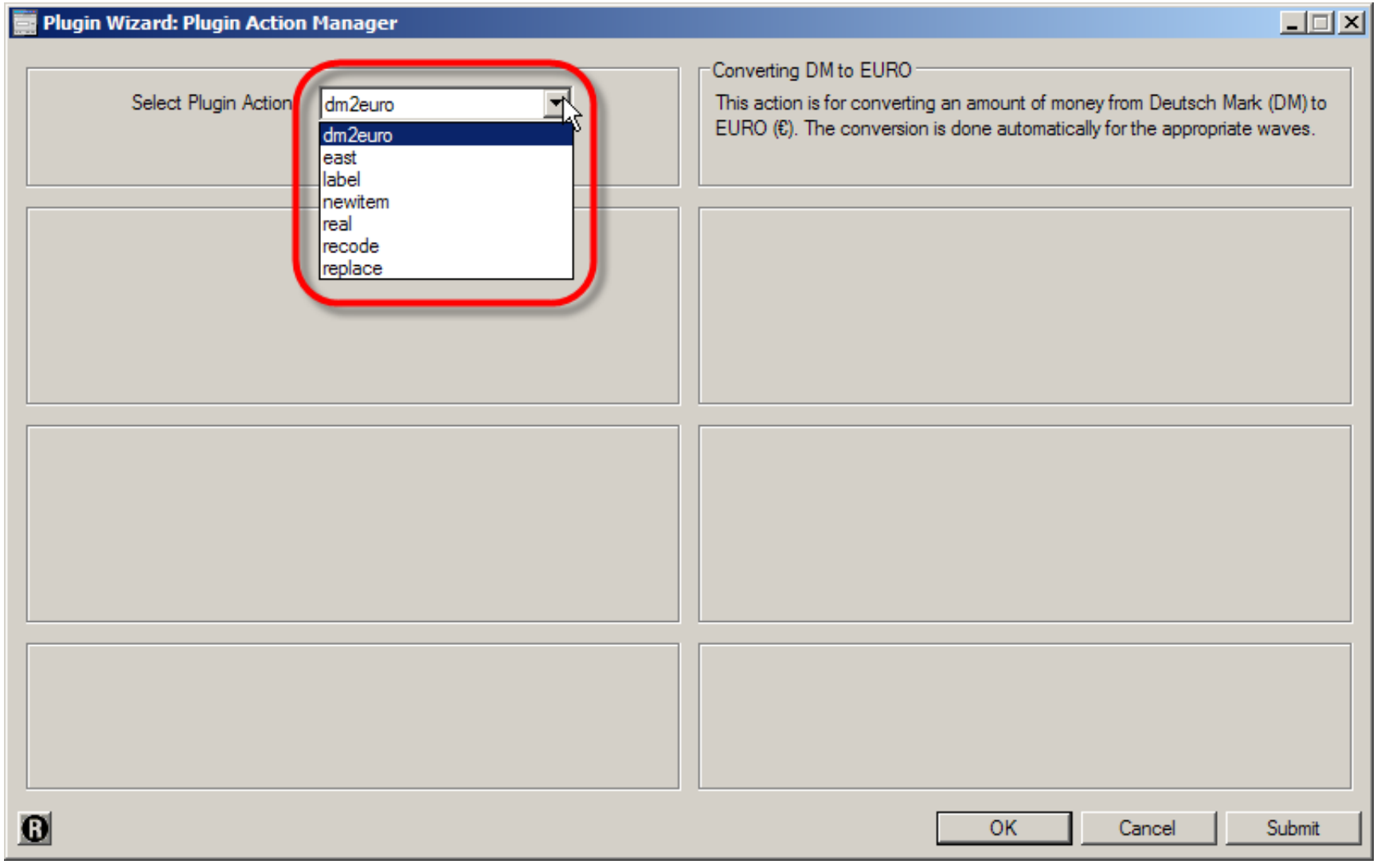
Add an action to the list



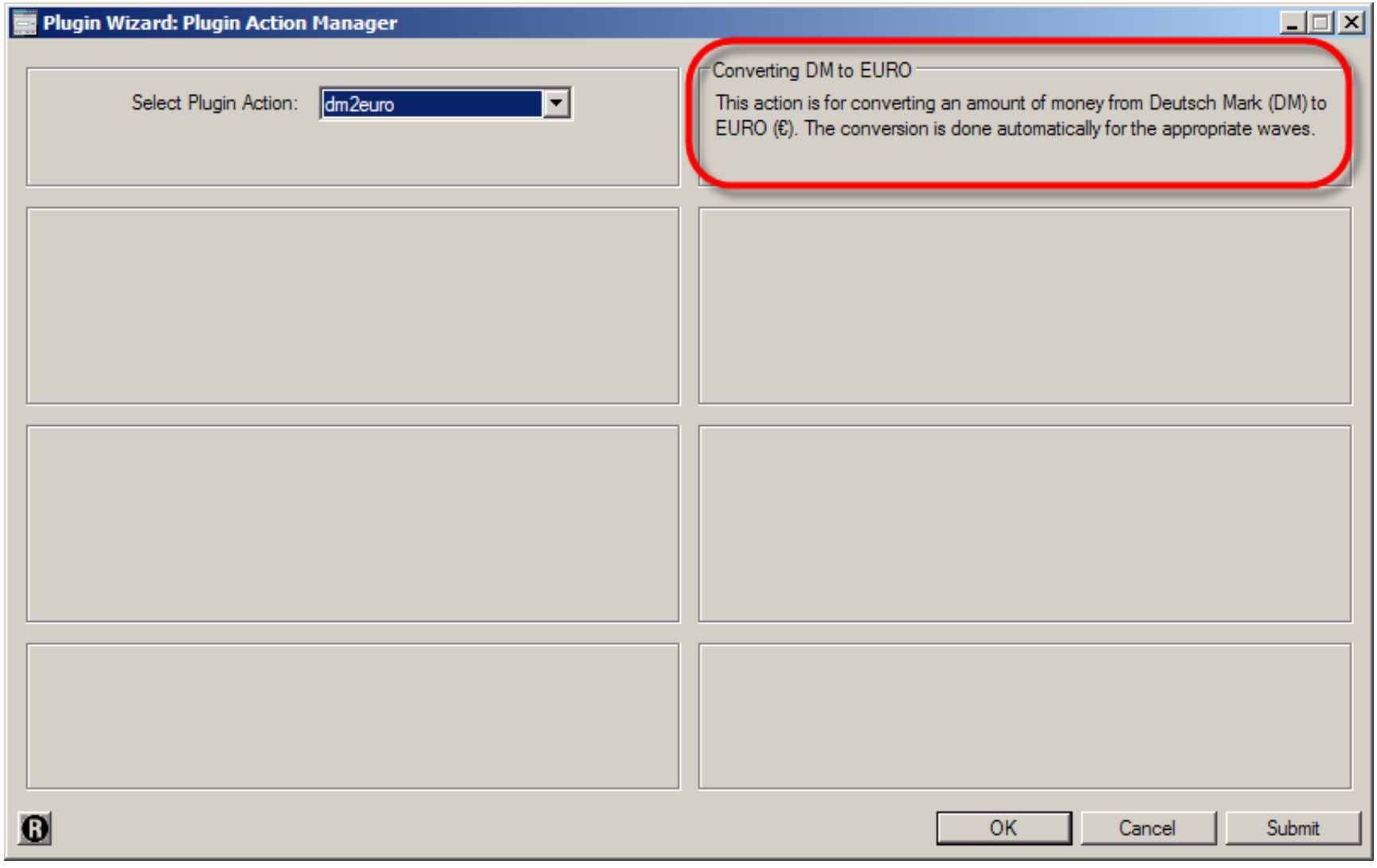
Plugin Action Manager



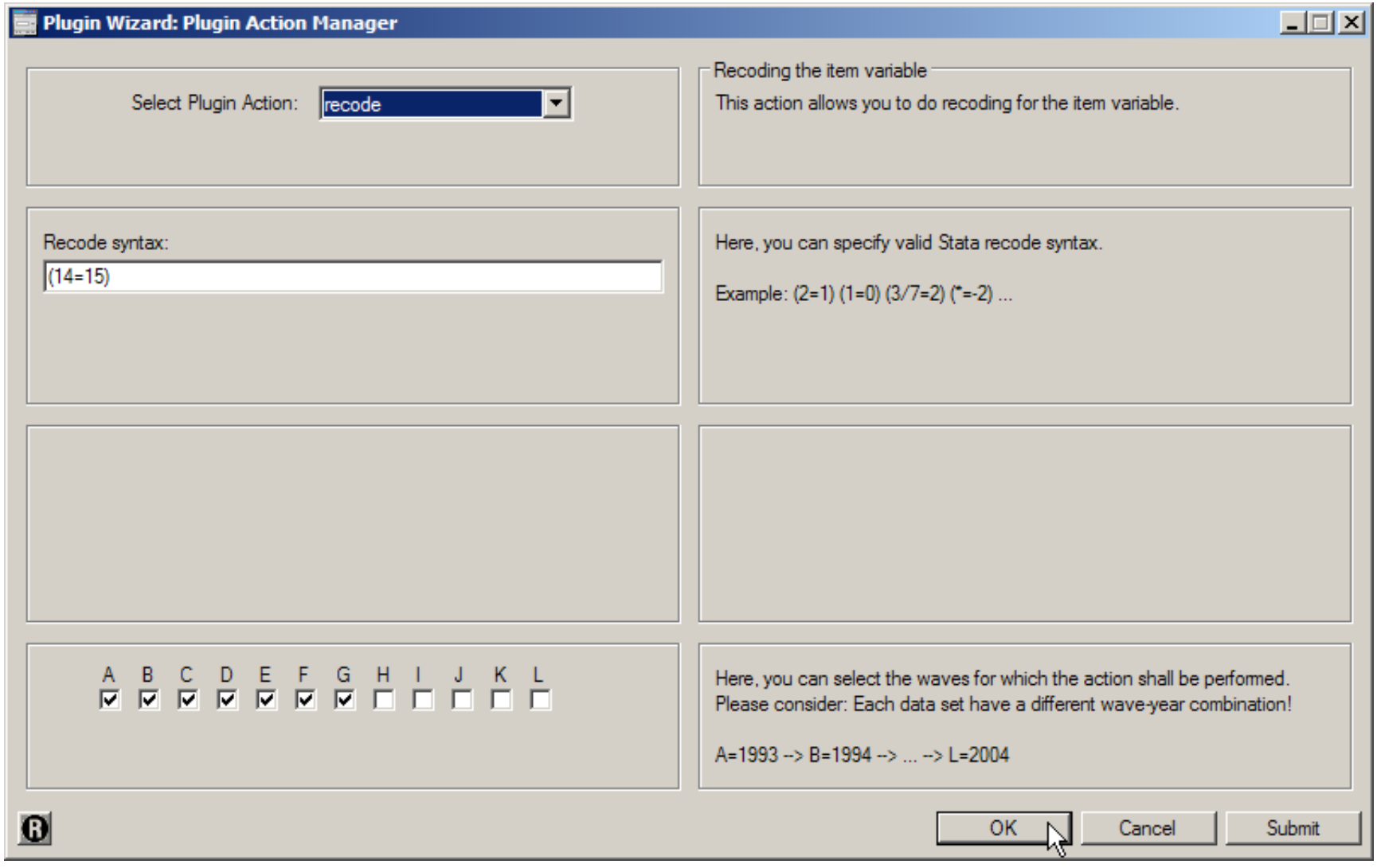
Select an action



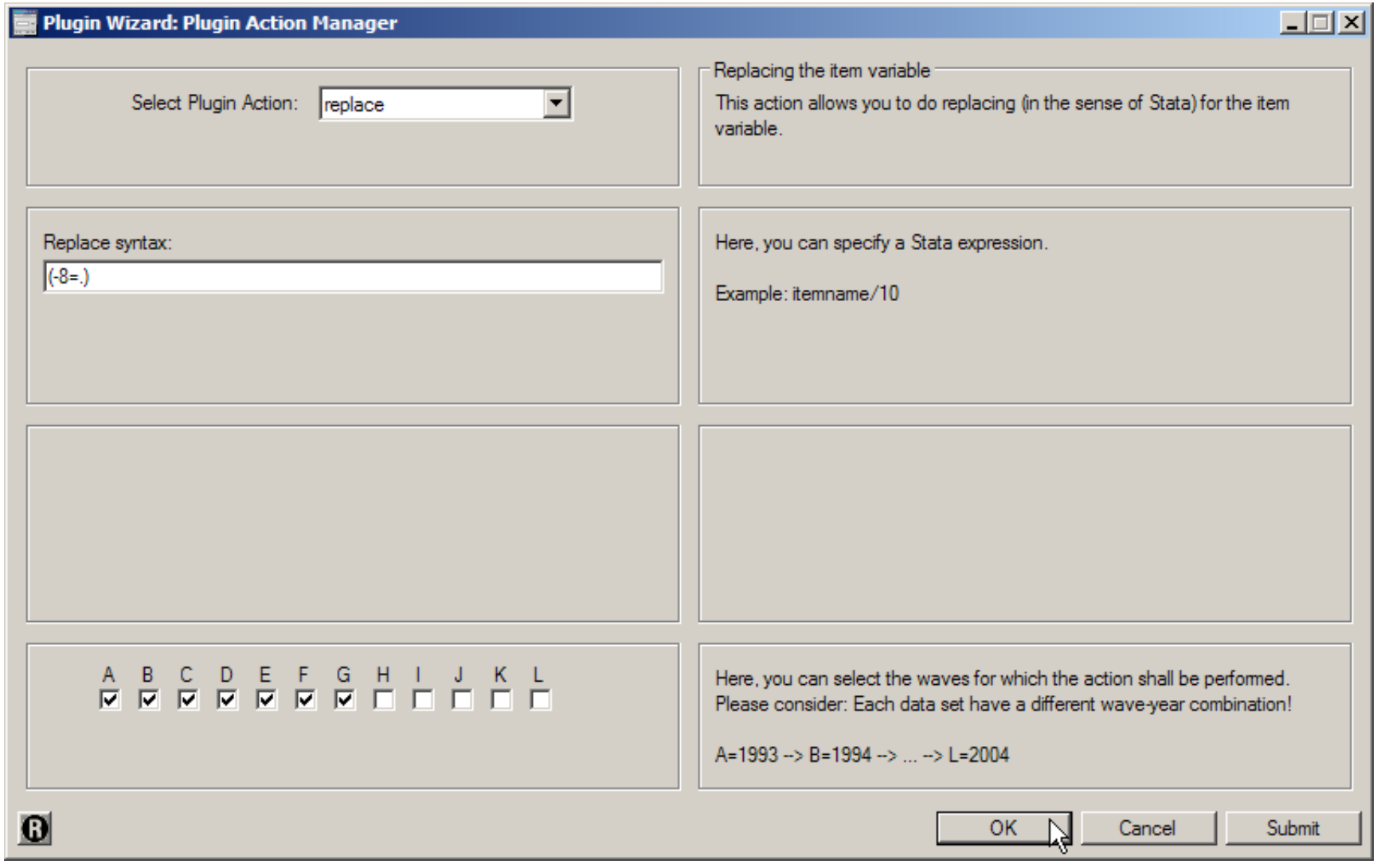
Select an action



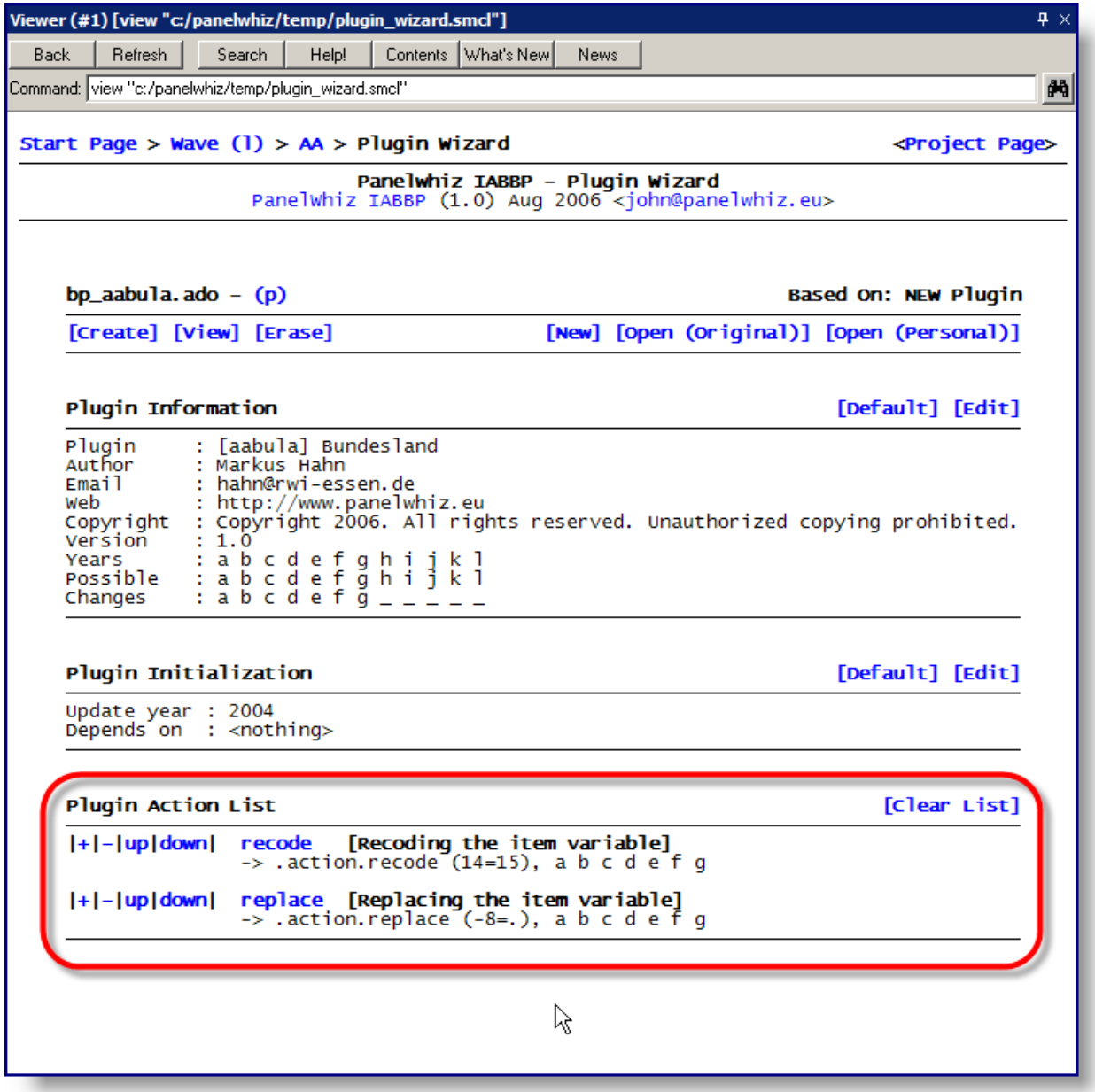
Description of action



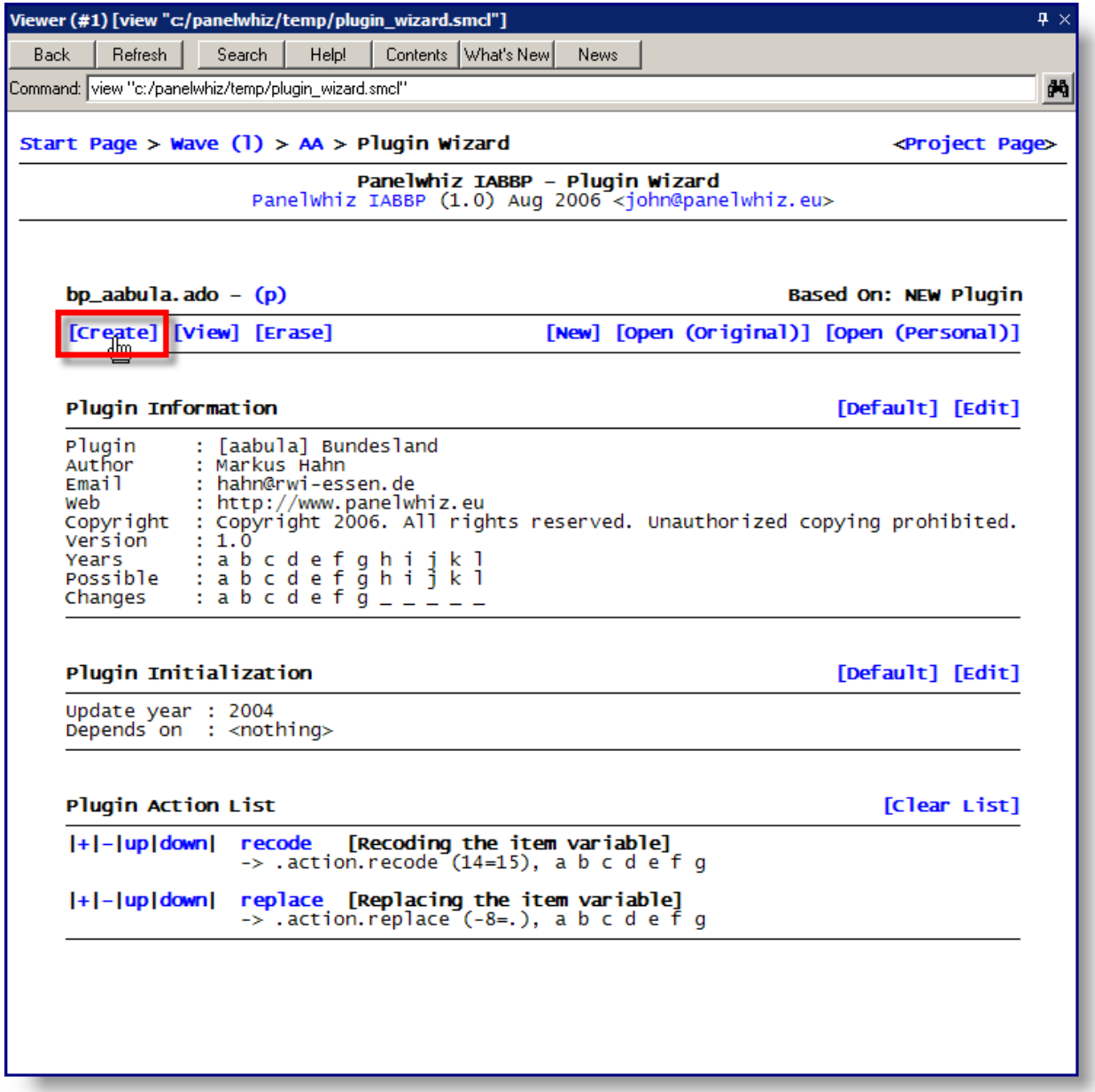
Action "recode"



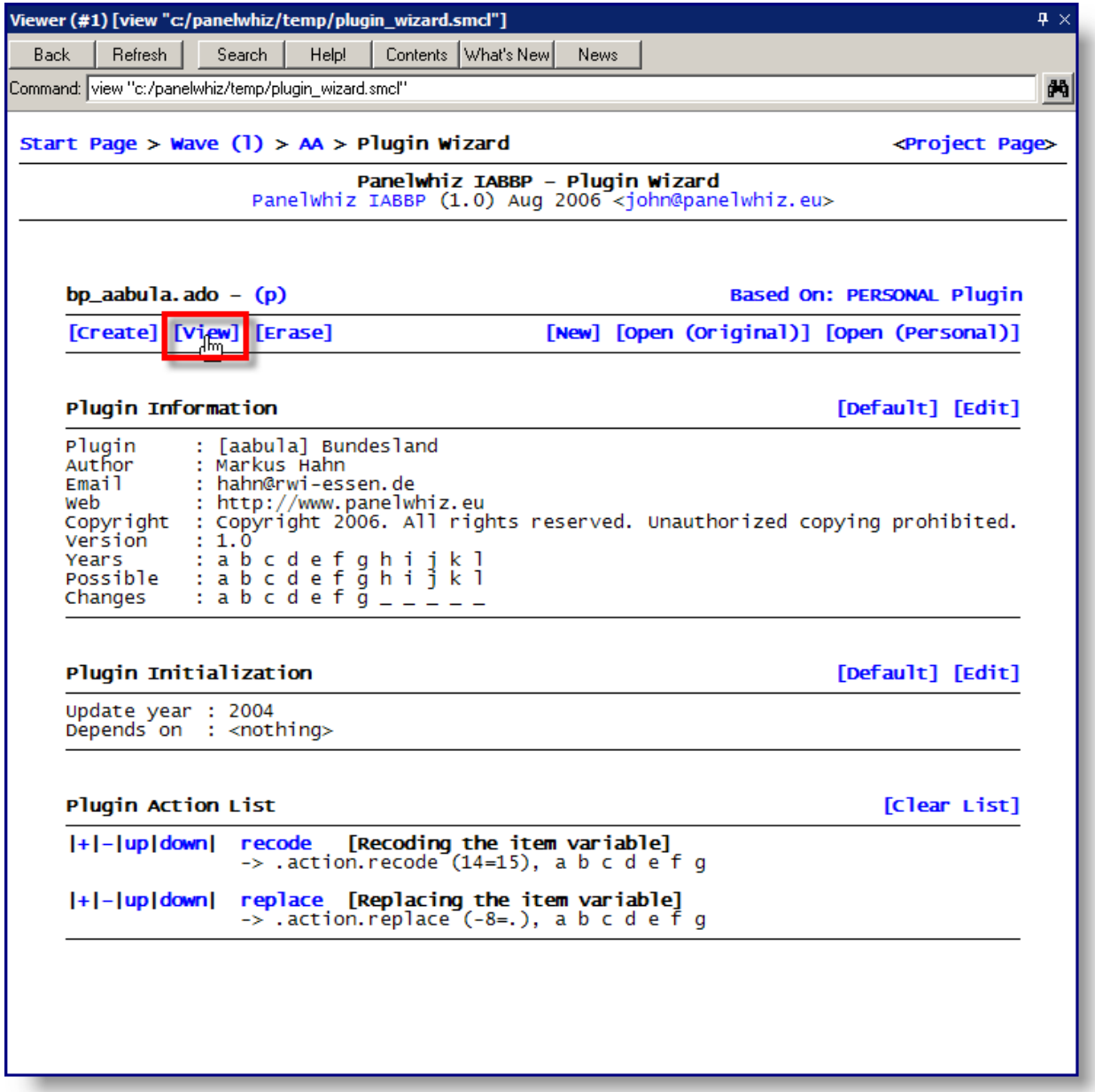
Action "replace"



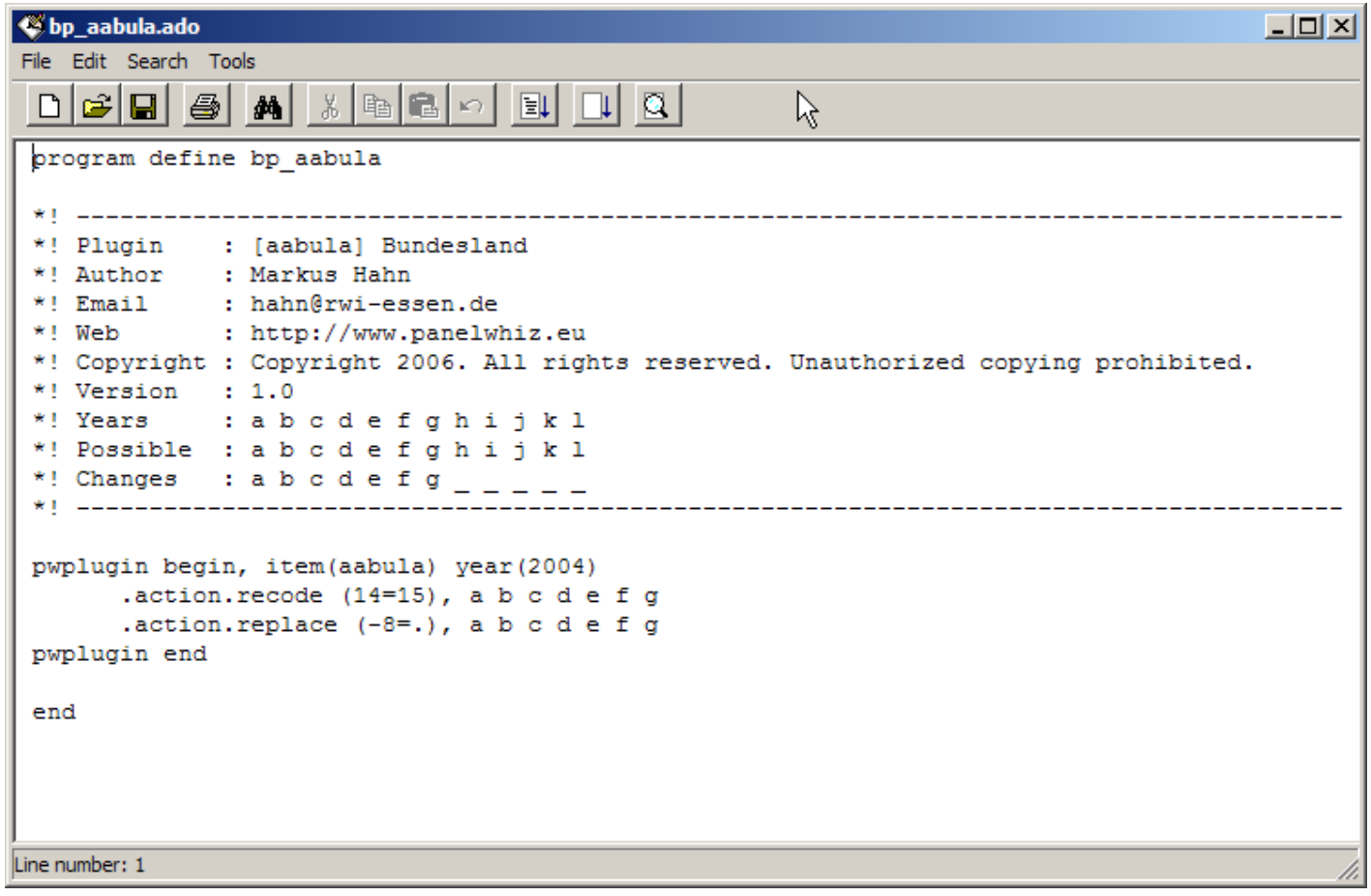
Two actions were added



Create plugin


Viewer (#1) [view "c:/panelwhiz/temp/plugin_wizard.smcl"]
Back
Refresh
Search
Help!
Contents
What's New
News
Command: view "c:/panelwhiz/temp/plugin_wizard.smcl"
Start Page > Wave (1) > AA > Plugin Wizard
<Project Page>
Panelwhiz IABBP - Plugin Wizard
Panelwhiz IABBP (1.0) Aug 2006 <john@panelwhiz.eu>
bp_aabula.ado - (p)
Based On: PERSONAL Plugin
[Create] [View] [Erase]
[New] [Open (Original)] [Open (Personal)]
Plugin Information
[Default] [Edit]
Plugin : [aabula] Bundesland
Author : Markus Hahn
Email : hahn@rwi-essen.de
Web : http://www.panelwhiz.eu
Copyright : Copyright 2006. All rights reserved. Unauthorized copying prohibited.
Version : 1.0
Years : a b c d e f g h i j k l
Possible : a b c d e f g h i j k l
Changes : a b c d e f g _ _ _ _ _
Plugin Initialization
[Default] [Edit]
Update year : 2004
Depends on : <nothing>
Plugin Action List
[Clear List]
|+|-|up|down| recode [Recoding the item variable]
-> .action.recode (14=15), a b c d e f g
|+|-|up|down| replace [Replacing the item variable]
-> .action.replace (-8=.), a b c d e f g

View plugin



Created plugin code

When a plugin is finally created...

- Can be used directly with PanelWhiz retrievals
- Can be shared with the PanelWhiz community
  - Embedded knowledge, which is respected by contract
    $\Rightarrow$ PanelWhiz Plugins must be cited
  - Modularized piece of code $\Rightarrow$ Shareable

RWI ESSEN

Plans for future releases

- Add more predefined actions
- More flexibility through "free coding"
- Extend the auto-documentation feature

RWI ESSEN