WEB SCRAPING IN GERMAN PRICE STATISTICS

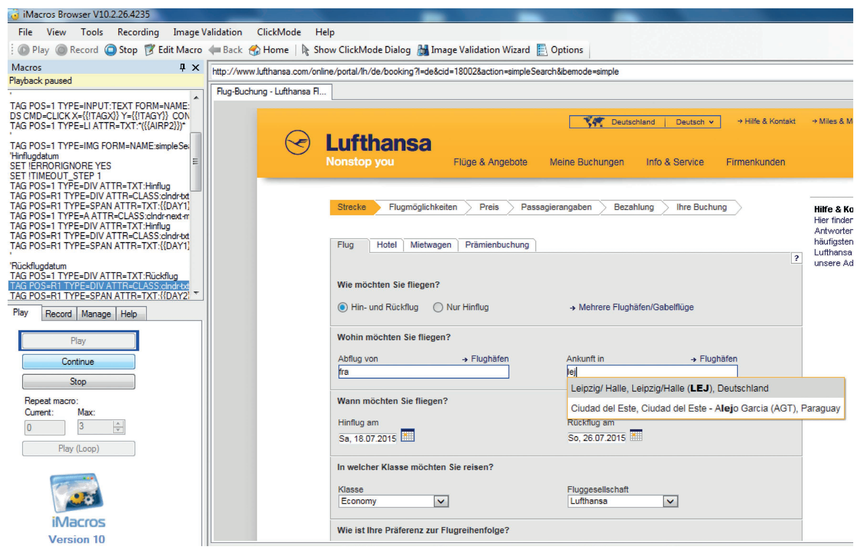
Dorothee Blang, Federal Statistical Office of Germany

Aims of the feasibility study

- Is it possible to automize price surveys via internet by imitation of manual collection?
- Is this an efficient survey method?
- What are the advantages / disadvantages?

IT-infrastructure and applied tools

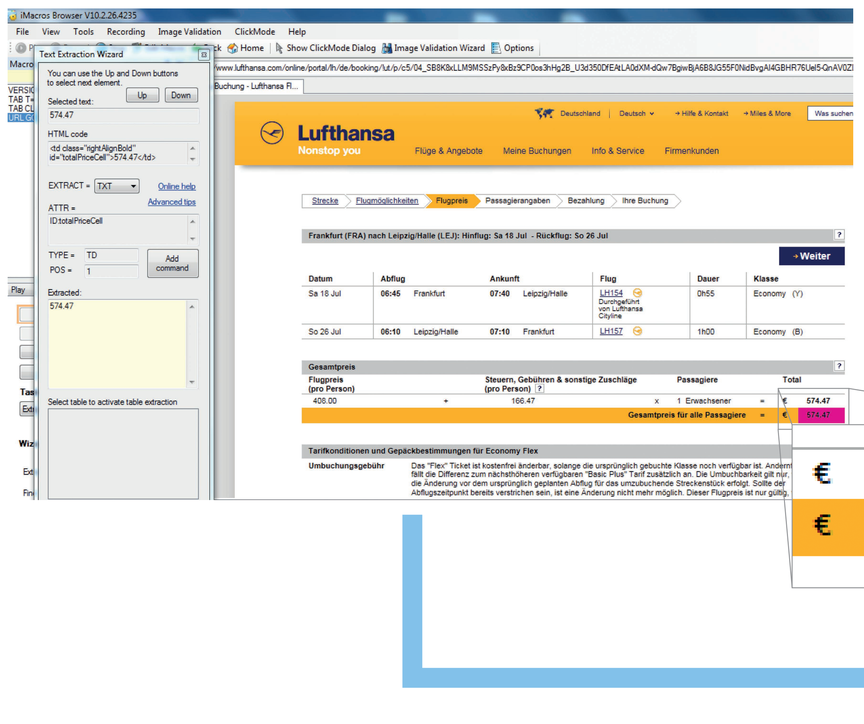
iMacros – recording scripts, form filling



iMacros – code example for flights

```
recorded:
    URL GOTO=http://www.lufthansa.com/
    TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
    ATTR=ID:fld20 CONTENT=fra
    TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
    ATTR=ID:fld22 CONTENT=lej
    TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
    ATTR=ID:fdepdateDisplay CONTENT=Do,‹SP›20.12.2012
edited:
    URL GOTO=http://www.lufthansa.com/
    TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
    ATTR=ID:fld20 CONTENT={{AIRP1}}
    TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
    ATTR=ID:fld22 CONTENT={{AIRP2}}
    TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
    ATTR=ID:fdepdateDisplay ={{DATE1}}
```

iMacros – data extraction



MySQL Database: Input Data for iMakros Data extraction

departureAirport	arrivalAirport	tarif	daysToDeparture	nights	outTime1	outTime2	retTime1	retTime2
FRA	MUC	E	30	2	10:00:00	15:00:00	10:00:00	15:00:00
FRA	BER	E	30	2	10:00:00	15:00:00	10:00:00	15:00:00
FRA	HAM	E	30	2	10:00:00	15:00:00	10:00:00	15:00:00
MUC	BER	E	30	2	10:00:00	15:00:00	10:00:00	15:00:00
MUC	DUS	E	30	2	10:00:00	15:00:00	10:00:00	15:00:00
MUC	HAM	E	30	2	10:00:00	15:00:00	10:00:00	15:00:00

MySQL database: Output Data

airp_from	airp_to	date1	date2	airline1	airline2	price	i_erh_dat	Time1	Time2	rate
Frankfurt	Amsterdam	Mo 29 Jun	Di 7 Jul	LH992	LH987	313.21	28.06.2015	13:00	10:10	Business (P) Economy (W)
Frankfurt	Barcelona	Mo 29 Jun	Di 7 Jul	LH1130	LH1125	139.9	28.06.2015	11:00	10:30	Economy (T) Economy (T)
Frankfurt	Barcelona	Mo 29 Jun	Di 7 Jul	LH1130	LH1127	139.9	28.06.2015	11:00	12:45	Economy (T) Economy (T)
Frankfurt	Barcelona	Mo 29 Jun	Di 7 Jul	LH1128	LH1125	139.9	28.06.2015	12:55	10:30	Economy (T) Economy (T)
Frankfurt	Berlin, Tegel	Mo 29 Jun	Di 7 Jul	LH182	LH185	289.22	28.06.2015	11:45	11:45	Economy (U) Economy (W)
Frankfurt	Berlin, Tegel	Mo 29 Jun	Di 7 Jul	LH186	LH185	289.22	28.06.2015	13:45	11:45	Economy (U) Economy (W)
Frankfurt	London, Heathrow	Mo 29 Jun	Di 7 Jul	LH906	LH903	442.55	28.06.2015	12:00	10:30	Economy (M) Economy (T)
Frankfurt	London, Heathrow	Mo 29 Jun	Di 7 Jul	LH906	LH905	467.55	28.06.2015	12:00	11:30	Economy (M) Economy (S)
Frankfurt	London, Heathrow	Mo 29 Jun	Di 7 Jul	LH908	LH903	442.55	28.06.2015	14:00	10:30	Economy (M) Economy (T)
Frankfurt	Paris, Charles de Gaulle	Mo 29 Jun	Di 7 Jul	LH1034	LH1029	428.12	28.06.2015	12:10	10:40	Economy (H) Economy (V)
Frankfurt	Paris, Charles de Gaulle	Mo 29 Jun	Di 7 Jul	LH1034	LH1031	428.12	28.06.2015	12:10	11:30	Economy (H) Economy (V)
Frankfurt	Paris, Charles de Gaulle	Mo 29 Jun	Di 7 Jul	LH1036	LH1029	428.12	28.06.2015	13:45	10:40	Economy (H) Economy (V)

Results

- Yes, for a lot of products it‘s a feasible solution
- It‘s efficient and can help to increase number of price observations
- Development requires profound programming skills
- Website changes occur irregularly, thus
  - Service has to be available at any time
  - Work load for service cannot be predicted

Conclusions

- Implementation in daily production is intended
- Allocation of stuff resources for support is an essential precondition

Federal Statistical Office of Germany, 30th Voorburg Group meeting, Sydney, 21th – 25th september 2015