

Machbarkeitsstudie zur Automatisierten Preiserhebung im Internet

Das Projekt MAPI

Machbarkeitsstudie zur **A**utomatisierten **P**reiserhebung im **I**nternet

- **gehört zum Themengebiet Multipurpose Price Statistics**
- **Teilfinanzierung von EUROSTAT zu 90%**
- **Projektdauer: 01.01.2012 bis 31.03.2014**

Zielsetzung des Projekts

- **Es soll untersucht werden, inwieweit Preiserhebungen im Internet mithilfe automatisierter Verfahren (sogenanntes web scraping oder screen scraping) möglich und sinnvoll sind.**
- **Aufwand und Fehleranfälligkeit manueller Preiserhebungen im Internet können gesenkt werden**
- **Ersparnis ist hauptsächlich bei Erhebungen zu erwarten, die wiederholt in kürzeren Zeitabständen durchgeführt werden**
 - **Verbraucherpreisstatistik**
 - **Kaufkraftparitäten nur bei bestimmten Erhebungen (Flugreisen, Hotels)**

Verwendete Software



iMacros

- **Spezielles Aufzeichnungstool für Web Scraping**
- **Einfache Aufzeichnung, ähnlich wie Excel Makros**
- **Es wird Code generiert, welcher nachträglich bearbeitet werden kann**
- **Ausfüllen von Formularen, Anklicken von Links, Extrahieren von Inhalten**

Java

- **Abbildung von Programmlogik**
- **Bereinigung von Daten**
- **Auslesen von Daten aus der Datenbank**
- **Daten in Datenbank ablegen**

iMacros – Skripte aufzeichnen, Formular

ausfüllen

The screenshot shows the iMacros Browser interface recording a flight booking on the Lufthansa website. The browser window displays the Lufthansa flight booking page with the following details:

- Page Title:** Lufthansa - Flug-Buchung ...
- URL:** NUQyMEICR09HMVY3Mk9VNS8wWwGNEQzqzMjUwMDAyL3NhLnNpbXBsZVNhYXJjaC9NYXhpbnWl6ZWQvRGVmYXVsdAII/?l=de&cid=18002&p=LH&s=DE
- Navigation:** Deutschland | English → Hilfe & Kontakt
- Header:** Lufthansa Nonstop you. Buchen & Planen, Top-Angebote, Info & Service, Miles & More.
- Progress Bar:** Strecke (selected), Flugmöglichkeiten, Preis, Passagierangaben, Bezahlung, Ihre Buchung.
- Form Fields:**
 - Wie möchten Sie fliegen?** Hin- und Rückflug Nur Hinflug → Mehrere Flughäfen/Gabelflüge
 - Wohin möchten Sie fliegen?**
 - Abflughafen: fra → Städte
 - Zielflughafen: lej → Städte
 - Dropdown: Leipzig, Leipzig/Halle (LEJ), Deutschland
 - Buttons: schließen
 - Wann möchten Sie fliegen?**
 - Hinflug am: Do, 20.12.2012
 - Rückflug am: So, 30.12.2012

The iMacros interface on the left includes a menu (File, View, Tools, Recording, Image Validation, ClickMode, Help, Extend Upgrade Protection Cover), a toolbar (Play, Record, Stop, Edit Macro, Back, Home, Show ClickMode Dialog, Image Validation Wizard, Options), and a script editor with the following code:

```
VERSION BUILD=8021970
TAB T=1
TAB CLOSEALLOthers
URL GOTO=http://www.lufthansa.com/oni
TAG POS=1 TYPE=INPUT:TEXT FORM=II
TAG POS=1 TYPE=INPUT:TEXT FORM=II
TAG POS=1 TYPE=INPUT:TEXT FORM=II
TAG POS=1 TYPE=INPUT:TEXT FORM=II
```

Below the script editor are controls for Play, Record, Edit, and a list of tasks including 'Wait a bit' and 'Do it!'. There are also wizard buttons for Extract (Text, Image) and Find (Image Validation).

iMacros – Beispielcode

Aufgezeichnet:

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:fdepdteDisplay CONTENT=***Do,<SP>20.12.2012*

Bearbeitet:

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:fdepdteDisplay ={{***DATE1***}}**

iMacros – eine Auswahl treffen

Bitte wählen Sie Ihren Hin- und Rückflug aus. ?

Günstigster Preis Datum ändern

Anzeige nach:
Reisezeit ▼

Gesamtpreis für Hin- und Rückflug ¹⁾ -
1 Erwachsener

Hinflug	+	Rückflug	=	Gesamt:
---		---		---

²⁾

Hinflug

Frankfurt (FRA) Nach Leipzig/Halle (LEJ) : Do 20 Dez

Von	Nach		Reisezeit	Economy Saver i	Economy Basic i	Economy Flex i	Business Basic Plus i
07:00 Frankfurt	08:00 Leipzig/Halle	LH154 durchgeführt von Lufthansa Cityline	1h00		<input type="radio"/> €114.23	<input type="radio"/> €191.23	<input type="radio"/> €291.23
09:10 Frankfurt	10:10 Leipzig/Halle	LH156 durchgeführt von Lufthansa Cityline	1h00		<input type="radio"/> €153.23	<input type="radio"/> €191.23	<input type="radio"/> €291.23
12:55 Frankfurt	13:55 Leipzig/Halle	LH158	1h00		<input type="radio"/> €138.23	<input type="radio"/> €191.23	
17:05 Frankfurt	18:05 Leipzig/Halle	LH160	1h00		<input type="radio"/> €138.23	<input type="radio"/> €191.23	
21:50 Frankfurt	22:50 Leipzig/Halle	LH164	1h00		<input type="radio"/> €138.23	<input type="radio"/> €191.23	<input type="radio"/> €291.23 8 Plätze
17:35 Frankfurt	18:25 Düsseldorf	LH084	3h00	<input checked="" type="radio"/> €86.70	<input type="radio"/> €126.70	<input type="radio"/> €332.70	<input type="radio"/> €522.70 1 Platz
19:35 Düsseldorf	20:35 Leipzig/Halle	LH2782 durchgeführt von Eurowings					

iMacros – Datenextraktion

The screenshot shows the iMacros Browser interface with a Lufthansa flight booking page. The browser title is "iMacros Browser V8.02.1970". The address bar shows a URL with a search parameter. The page content includes the Lufthansa logo and navigation tabs: "Strecke", "Flugmöglichkeiten", "Flugpreis", "Passagierangaben", "Bezahlung", and "Ihre Buchung". The "Flugpreis" tab is active, displaying flight details for Frankfurt (FRA) to Leipzig/Halle (LEJ).

The "Text Extraction Wizard" window is open, showing the following configuration:

- Selected text: 288.79
- HTML code: <TD id=totalPriceCell
- EXTRACT = TXT
- ATTR = ID:totalPriceCell
- TYPE = TD
- POS = 1
- Extracted: 288.79

The flight details table is as follows:

Datum	Abflug	Ankunft	Flug	Dauer	Klasse
Do 20 Dez	07:00 Frankfurt	08:00 Leipzig/Halle	LH154 Durchgeführt von Lufthansa Cityline	1h00	Economy (W)
So 30 Dez	10:55 Leipzig/Halle	12:05 Frankfurt	LH161 Durchgeführt von Lufthansa Cityline	1h10	Economy (H)

The price breakdown table is as follows:

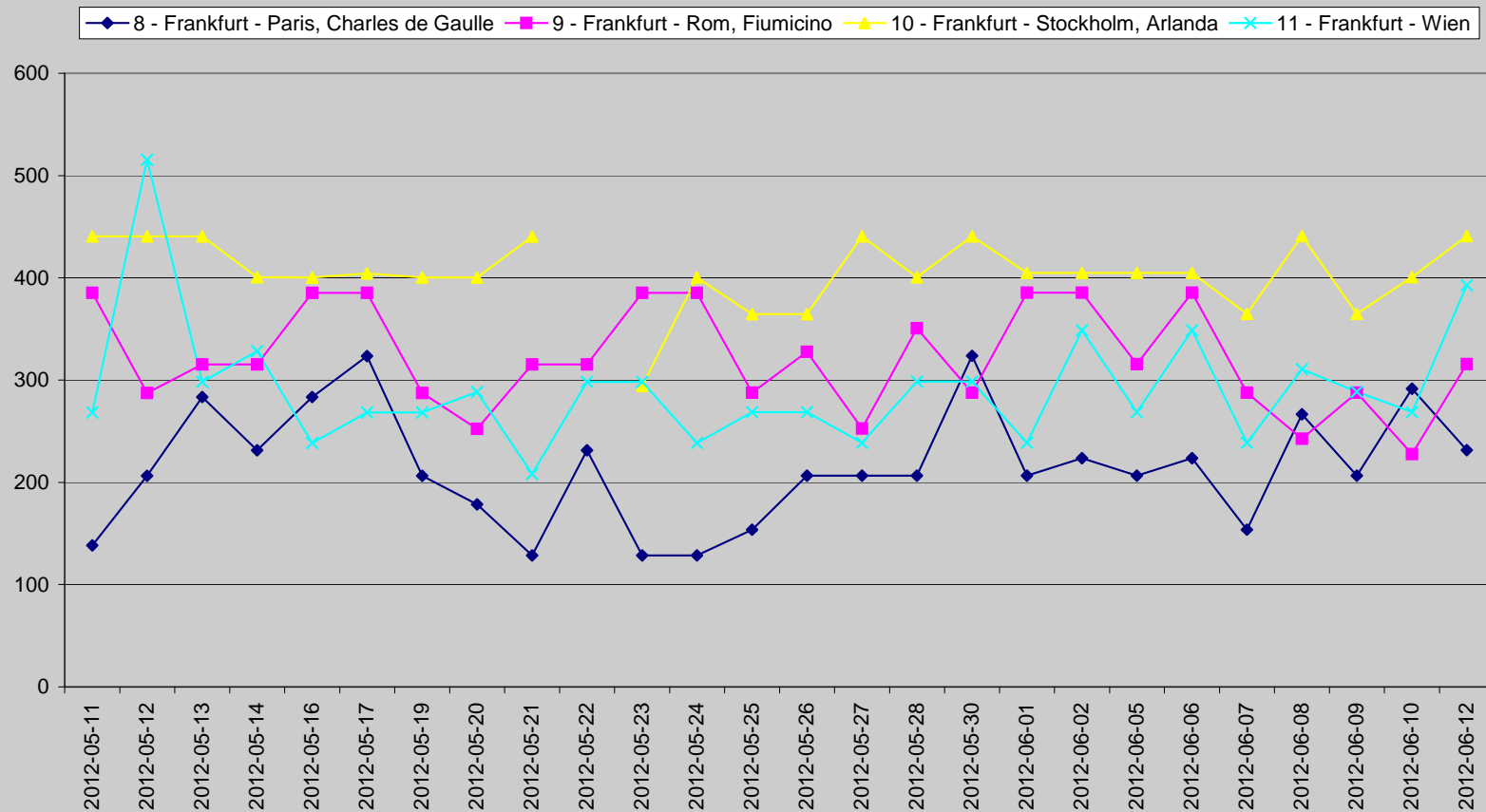
Flugpreis (pro Person)	Steuern, Gebühren & sonstige Zuschläge (pro Person) [?]	Passagiere	Total
148.00	+	140.79	x 1 Erwachsener = € 288.79
Gesamtpreis für alle Passagiere			= € 288.79

Below the price breakdown, there is a note: "Sie können sich dieses Angebot bis zum : 21.11.12 09:00 sichern, indem Sie die Option "Reservierung mit Preisgarantie" auf den folgenden Seiten auswählen."

MySQL Datenbank

airp_from	airp_to	date1	date2	airline1	airline2	price	i_erh_dat	Time1	Time2	rate
Frankfurt	Barcelona	Do 25 Okt	So 28 Okt	LH1128	LH1125	325.91	26.08.2012	13:15	10:00	Economy (S) Economy (S)
Frankfurt	Barcelona	Do 25 Okt	So 28 Okt	LH1128	LH1127	325.91	26.08.2012	13:15	12:30	Economy (S) Economy (S)
Frankfurt	Berlin, Tegel	Do 25 Okt	So 28 Okt	LH182	LH185	183.42	26.08.2012	12:55	11:45	Economy (W) Economy (S)
Frankfurt	Berlin, Tegel	Do 25 Okt	So 28 Okt	LH182	LH187	183.42	26.08.2012	12:55	12:45	Economy (W) Economy (S)
Frankfurt	Berlin, Tegel	Do 25 Okt	So 28 Okt	LH178	LH185	207.42	26.08.2012	10:15	11:45	Economy (V) Economy (S)
Frankfurt	London, Heathrow	Do 25 Okt	So 28 Okt	LH906	LH905	208.17	26.08.2012	12:45	11:50	Economy (E) Economy (T)
Frankfurt	London, Heathrow	Do 25 Okt	So 28 Okt	LH906	LH907	208.17	26.08.2012	12:45	13:40	Economy (E) Economy (T)
Frankfurt	London, Heathrow	Do 25 Okt	So 28 Okt	LH908	LH905	208.17	26.08.2012	13:45	11:50	Economy (E) Economy (T)
Frankfurt	Paris, Charles de Gaulle	Do 25 Okt	So 28 Okt	LH1036	LH1029	286.65	26.08.2012	13:35	10:40	Economy (K) Economy (Q)
Frankfurt	Paris, Charles de Gaulle	Do 25 Okt	So 28 Okt	LH1034	LH1029	291.65	26.08.2012	12:00	10:40	Economy (L) Economy (Q)
Frankfurt	Rom, Fiumicino	Do 25 Okt	So 28 Okt	LH232	LH233	393.6	26.08.2012	10:45	13:30	Economy (T) Economy (Q)
Frankfurt	Rom, Fiumicino	Do 25 Okt	So 28 Okt	LH232	LH231	643.6	26.08.2012	10:45	10:15	Economy (T) Economy (Y)

Niedrigster Flugpreis (Lufthansa)



Bisherige Arbeiten

Kaufkraftparitäten

- Flüge bei Expedia und Lufthansa
- Hotels bei HRS
- Internationale Versandhändler (Esprit, Zalando, H&M, Zara, Deichmann und Adidas) in mehreren europäische Länder

Verbraucherpreisindex

- Versandhändler: Otto
- Online Apotheken
- Mietwagen
- Bahn
- Städtereisen

Was ist zu beachten

Ausgangsdaten

- Vorgaben müssen teilweise für maschinelle Zwecke „übersetzt“ werden, z.B. Preise für Flüge um die Mittagszeit ermitteln → Flüge mit Abflugzeit von 10.00 – 14.00.
- Für die automatisierte Verarbeitung angelegte Daten liefern bessere Ergebnisse als Weiternutzung von Daten, die für die manuelle Erhebung verwendet werden.

Datenextraktion

- Extraktion von Einzelinformationen schwierig, wenn diese nicht systematisch im HTML Code abgelegt sind.
- Bei Preisen gab es hier bisher keine Schwierigkeiten.
- Mengenangaben, Merkmale können abhängig von der Website nicht immer sauber extrahiert werden.

Was ist zu beachten

Technische Probleme

- **Abwehrmaßnahmen gegen Web Scraping sind möglich.**
- **Bei Änderungen auf den Websites muss die Software angepasst werden.**

Rechtliche Aspekte

- Auch die Inhalte von Datenbanken sind urheberrechtlich geschützt sofern wesentliche Teile betroffen sind.
- Das zielgerichtete Auslesen von Preisen ist unproblematisch.
- AGBs die Web Scraping untersagen sind nur wirksam wenn Sie vorher ausdrücklich akzeptiert wurden.
- Es werden nur Daten ausgelesen die ohne die Akzeptierung von AGBs erreichbar sind.
- Es dürfen keine technischen Hürden umgangen werden (z.B.: CAPTCHAS)
- Web Scraping ist aus diesen Gründen z.B. bei Ryanair nicht möglich