
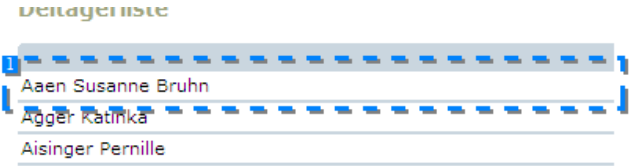


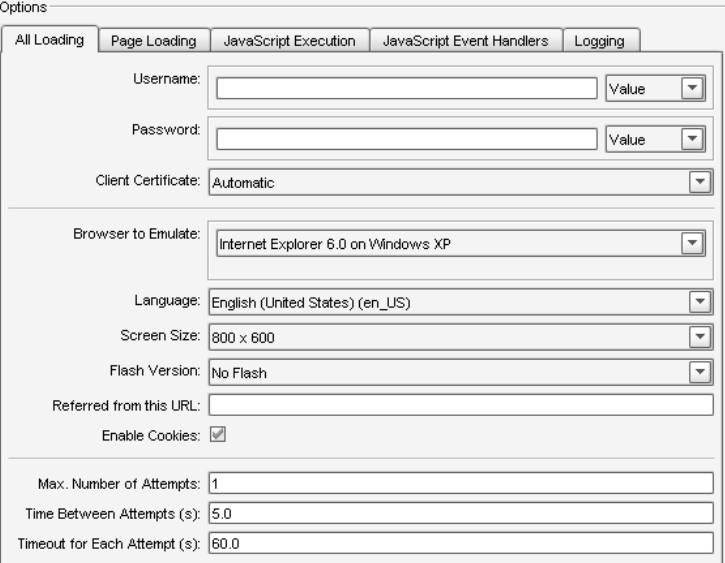
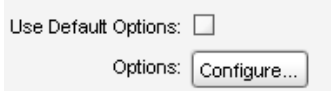


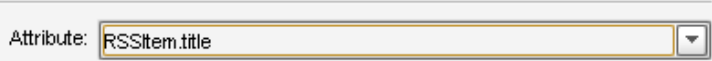
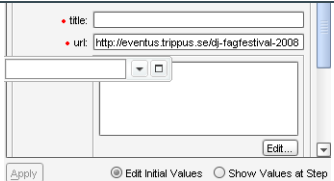

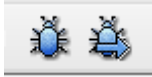
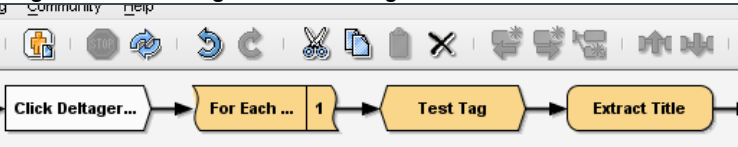
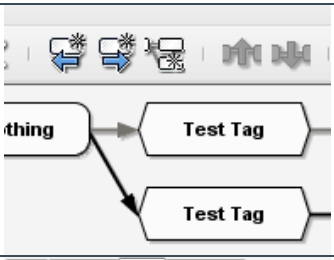

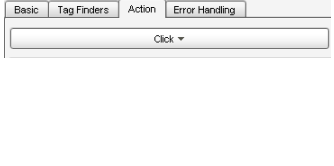
## HUSKESEDDEL

Hvad	Hvorfor	Hvordan
Hent OpenKapow	OpenKapow er et enkelt program til at bygge robotter for at skrabe data på nettet. Programmet er gratis at hente og bruge. Det findes i en linux- og windows-version. Virker linket ikke så søg blot efter openkapow og klik på download. Fra hjemmesiden er der også adgang til at hente færdige robotter og deltage i diskussioner for at få tips.	<a href="http://openkapow.com/blogs/getting_started/archive/2007/10/26/download-robomaker-6-3.aspx">http://openkapow.com/blogs/getting_started/archive/2007/10/26/download-robomaker-6-3.aspx</a>
Installer programmet	Alle hjælpeprogrammer er indeholdt. Følg bare vejledningen.	Klik på exe-filen.
Vælg en side, der skal skrubes	Sørg for at gennemse siden og vælg de felter ud, som du skal bruge. Lav notater om formater (tekst, tal, dato og lignende) Du kan altid revidere ting, men det er klogt at bruge fem-ti minutter på at få overblik over sidens indhold.	Klik rundt.
Åbn Robomaker		Klik på program-ikonet på programlinjen.
Vælg robottype	Start med RSS/Atom feed og klik dig igennem. No til input-values. Nu får du en robot frem med første og sidste step oprettet – load page og return item. I RSS-robot skal der udtrækkes noget i felter for både titel og URL. I programvinduet er der seks felter. Et øverst med robotens step, et med hjemmesiden i normalt format, et med hjemmesidens tags, et med hjemmesidens kildekode, et med beskrivelse af det fremhævede grønne step i robotten og endelig et med input- og output-values.	Select the type of service to create: <input checked="" type="radio"/> Create an RSS/Atom feed <input type="radio"/> Create a REST web service
Gem robot	Giv robotten et passende navn og gem den. Robomaker gemmer normalt robotterne i .../programs/...openkapow/project/library.	
Indsæt step	De vigtigste skridt er loop, test tag (conditions), click og extract. Du kan udvide det valgte område ved at holde alt-tasten nede og klikke i feltet med tags eller med kildekode. Når du er tilfreds med området, højreklikker du, og du kan nu vælge skridt.	 Klik på det sted på hjemmesiden i robomakers udsnit, som du vil bearbejde.



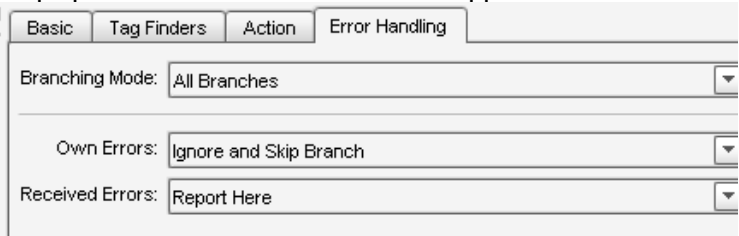
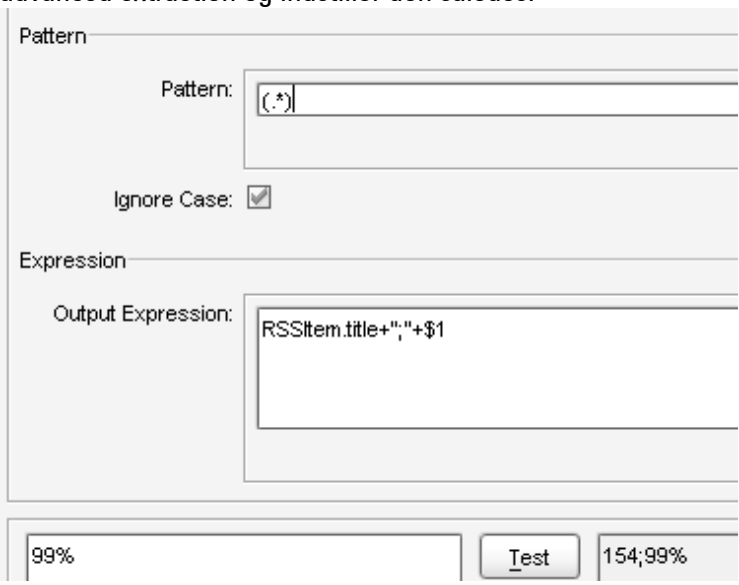

HUSKESEDDEL

<p>Loop</p>	<p>Oftest er det en tabel, og man looper over rækker i tabellen. Efterfølgende stiller man top og bund for loop i feltet action under beskrivelsen af robot-step.</p> <p>Det udvalgte område er nu i næste step markeret med en blå firkant på hjemmesiden og man kan efterfølgende klikke inde i den blå firkant for at indsætte steps med conditions, click eller extract.</p> 	
<p>Test tag</p>	<p>Definer om robotten skal teste for forekomsten af den pågældende tekst i html-koden eller ren tekst. Sæt .* før og efter test-ordene.</p> <p>Og du kan definere om robotten skal fortsætte eller stoppe, hvis indholdet er til stede i det valgte udsnit.</p>	
<p>Click</p>	<p>Højreklik på det link, som robotten skal trykke på. Du kan indstille load page, ved at gå ind i action-feltet og afklikke "Use default options" og trykke configure. Nu kan du sætte robotten til at vente længere på en side en default 60.0 sec(Timeout for Each Attempt), prøve flere gange (Max number of Attemts), du kan fjerne load frames (klik i Page Loading) og execute javascript (Javascript Execution) for at gøre robotten hurtigere.</p> <p>Er der passwords til en side, kan du indsætte det under Username og Password.</p> 	

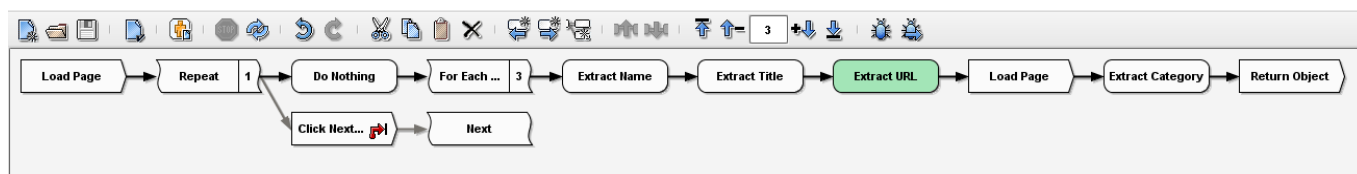
HUSKESEDDEL

<p>Extract</p>	<p>Højreklik på det indhold, der skal udtrækkes og definer om det er tekst eller tal. Vælg det felt, det skal gemmes i. Det vil så se sådan ud:</p> 	
<p>Edit initial values</p>	<p>Hvis et felt skal have en fast værdi i alle rækker er det lettest at gøre ved at klikke på Edit Initial Values og efter at have skrevet den faste værdi trykke Apply. Husk derefter også at trykke på Show Values at Step.</p>	
<p>Gem robotten</p>	 <p>Robotten kan for eksempel se sådan ud.</p>	
<p>Test robotten</p>	<p>Klik på debugger. Enten for at køre hele robotten eller for at køre den fra det sted, hvor den står i Robomaker. Det er i så fald ikonet med pilen på.                  Debuggeren stopper ved evt. fejl, og du kan rette dem, indtil den kører fejlfrit. Gem robotten før hver test i debuggeren. Efterhånden kan du lave ret komplicerede robotter, der kræver nogen finindstilling, før de kører godt.</p>	
<p>Kopier</p>	 <p>Ved at holde kontrol-tasten nede og klikke på de forskellige step, bliver de markeret. Man kan markere alt med control-A. Herefter kan man kopiere eller klippe alle step på en gang og sætte dem ind et andet sted.                  Ved at trykke på krydset sletter man step.</p>	
<p>Indsæt ekstra step eller ny gren.</p>	<p>Ved at klikke på ind af de tomme step-felter med pil på, indsætter du et ekstra step enten før eller efter det aktive step. Ved at klikke et tomt step-felt nedenunder et andet step, indsætter man en ekstra gren i robotten, for eksempel hvis man vil gøre en ting, når en test er positiv og en anden, når den er negativ.</p>	
<p>Sideskift 1: Repeat og Next</p>	 <p>Ofte skal man hente resultater ud fra flere sider. Det gør man ved at indsætte et blankt step. Under Action trykker man på Click, hvor der kan vælges step-type på samme måde som ved at højreklikke. Under Loops finder man både Repeat og Next.</p>	

HUSKESEDDEL

<p>Sideskift 2: Do Nothing</p>	<p>Der er indsat et felt med "Do Nothing" for at kunne teste, at sideskiftet faktisk virker ordentligt.</p>	
<p>Sideskift 3: Click Next</p>	<p>Under Tag finders skal man arbejde med at få defineret det helt rigtige. Her er det ved at skrive <code>.*Next.*</code> i Tag Pattern:</p>  <p>Tilsvarende skal man under Error Handling få den til at stoppe, når der kommer en fejl. Det betyder, at når der ikke længere er en pil på siden med teksten Next, så stopper robotten.</p> 	
<p>Kombiner udtræk af info I et felt</p>	<p>Nogle gange har du brug for at have mere end én information i et felt. Det gør du ved at bryde det ned i flere step og første gang udtrække første information. Anden gang udtrækker du anden information, tilføjer en converter (klik +), vælger advanced extraction og indstiller den således:</p>  <p>Oprindelig er tallet 154 trukket ud, og det nye ud træk er 99%. I pattern skriver du <code>(.*)</code> – det markerer hele inputtet. Og i Output Expression skriver du <code>RSSItem.title+";"+\$1</code> – det udtrækker det oprindelige plus et ; plus hele inputtet, dvs. 154;99%. Og sådan kan du fortsætte med flere og flere informationer i samme felt.</p>	 <p>Select a converter: Extraction: Advanced Extract</p>

## HUSKESEDDEL



Her ser du et eksempel på en robot, hvor der efter load page er indsat en bladrning mellem sider (repeat, click og next), derefter en loop over resultatet, hvorefter der udtrækkes diverse oplysninger, for derefter at klikke på en hjemmeside og udtrække en ny oplysning her, før alle oplysninger returneres.

En sådan model er typisk brugbar, når der er link i hver enkel linje og man skal ind på hvert link for at hente specifikke oplysninger om den enkelte post.

Hovedprincippet er, at mest muligt udtrækkes så tidligt som muligt i robotten.