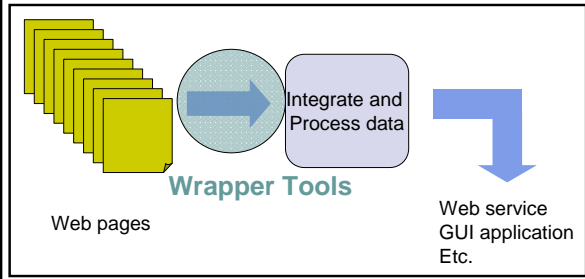


The Big Picture

Data Integration Application



How are they made? (In general)

- Attempt to find structure in a webpage
 - Regular expressions, DOM structure, etc.
- Exploit this structure to extract data
 - Look for a pattern, pull out information
- Allow you to define the format of extracted data
 - XML, HTML, etc.

Open Kapow

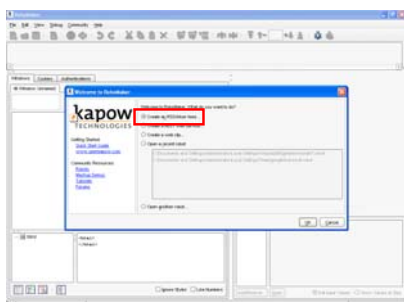
- <http://www.openkapow.com>
- Wrapper generation tool
 - Robots/Agents
- Relies on the DOM structure of a page
- Mimics a person browsing the web
- RoboMaker tool for visual generation
- Ability to
 - Publish robots for everyone to use
 - Use other published robots
- Three types of robots: RSS, REST, WebClips

Open Kapow: RSS Agent

- Grab top story from Digg
 - Modified to get all front page stories from Digg
- Extract title, url, description
- Show how to iterate, correct iterations
- Contend with errors (missing data)
- Publish online and call

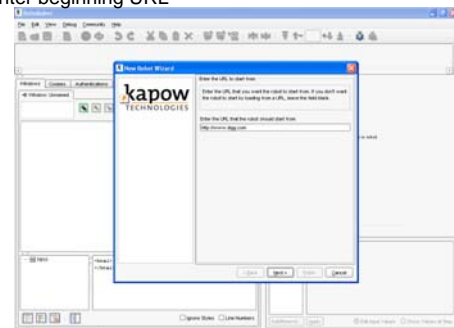
Digg.com RSS Agent

Select the Agent type



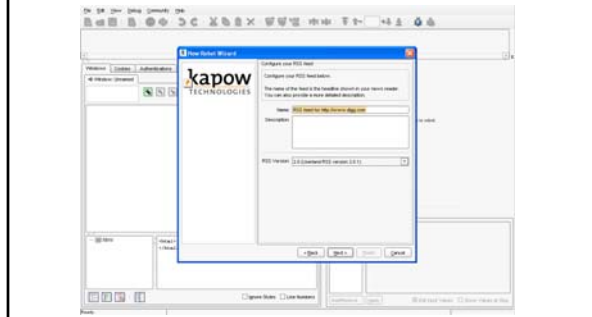
Digg.com RSS Agent

Enter beginning URL

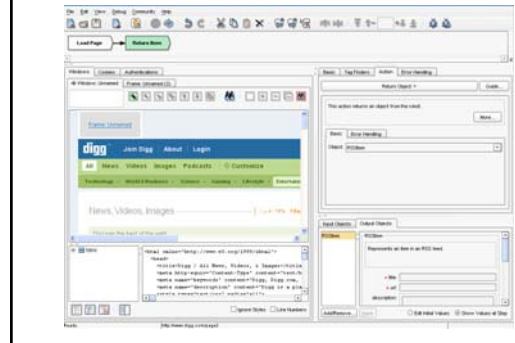


Digg.com RSS Agent

Describe your agent

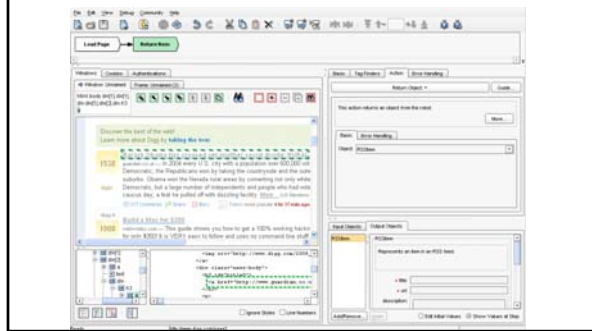


Digg.com RSS Agent

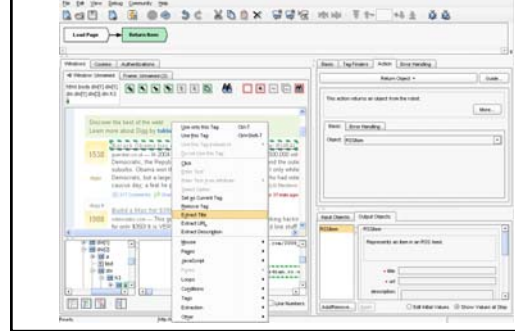


Digg.com RSS Agent

Select text to extract the title

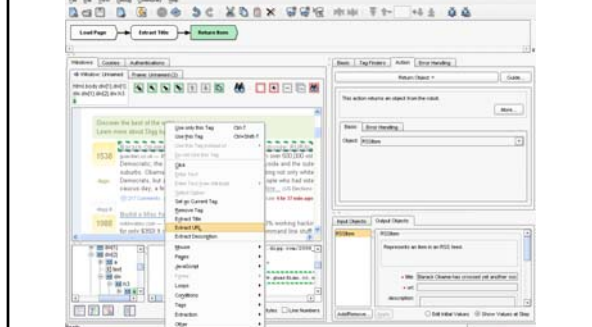


Digg.com RSS Agent



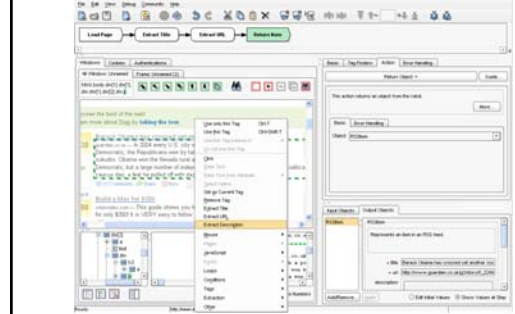
Digg.com RSS Agent

Extract the article URL



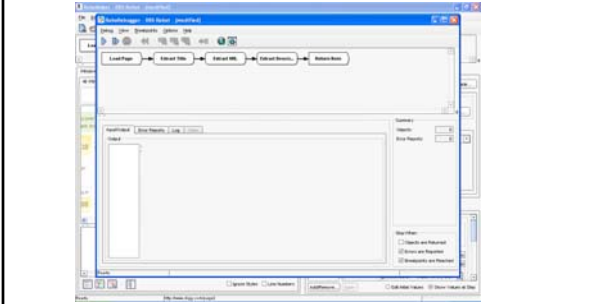
Digg.com RSS Agent

Extract article description

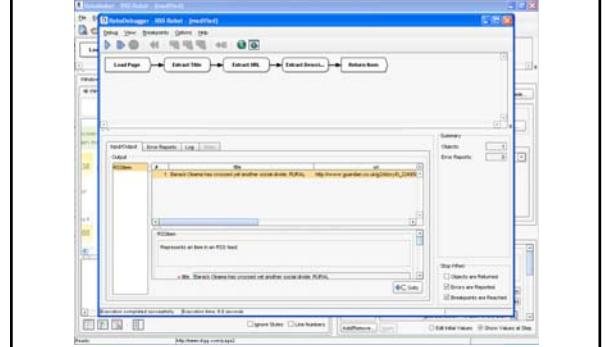


Digg.com RSS Agent

Run debugger to test your robot

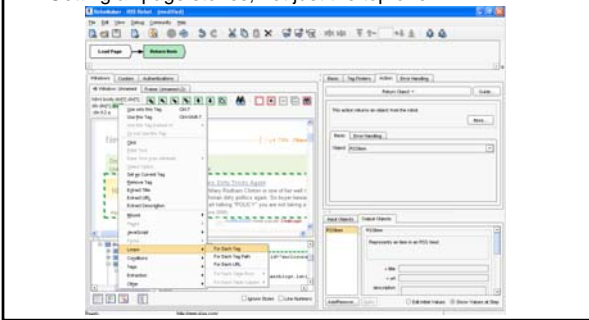


Digg.com RSS Agent

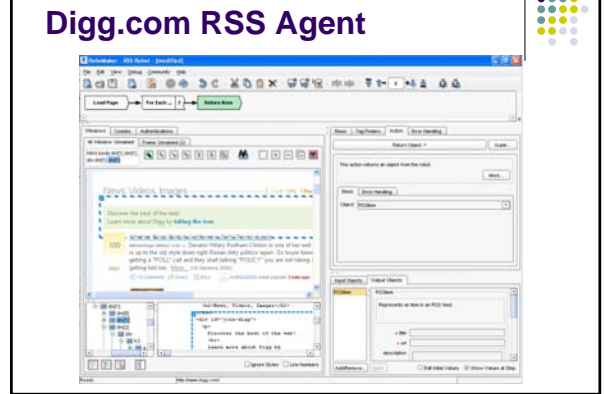


Digg.com RSS Agent

Getting all page stories, not just the top one

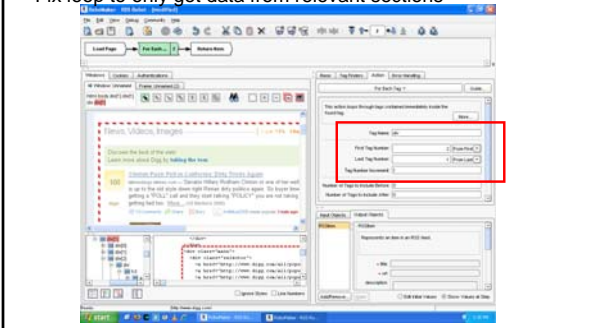


Digg.com RSS Agent



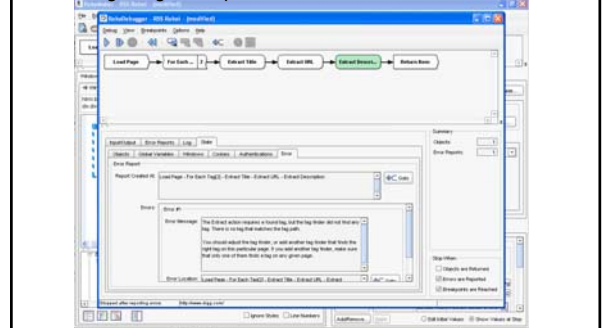
Digg.com RSS Agent

Fix loop to only get data from relevant sections



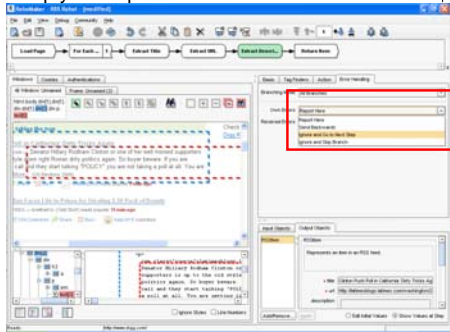
Digg.com RSS Agent

Errors getting descriptions for all articles



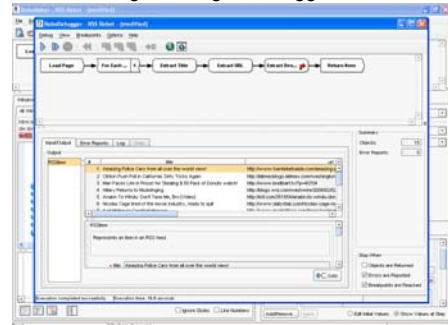
Digg.com RSS Agent

Allow empty descriptions



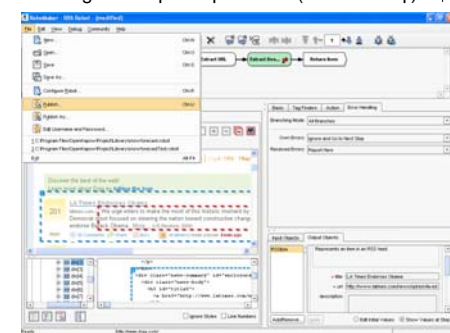
Digg.com RSS Agent

Now we have an agent that gets all digg stories



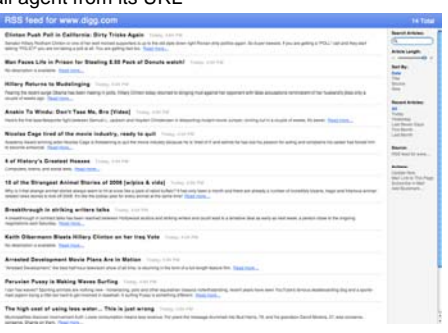
Digg.com RSS Agent

Publish the agent to openkapow.com (account req)



Digg.com RSS Agent

Call agent from its URL

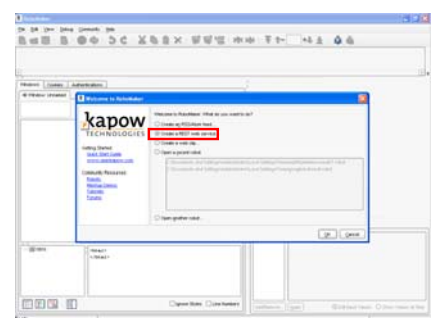


Open Kapow: REST Agent

- Snowforecast.com
- Take URL as input
- Go through all resorts and get resort info
- Use URL to load details page
 - Get weather
- Return as XML (or HTML, JSON, CSV)
- Add group name to divide into elements
- Publish and use

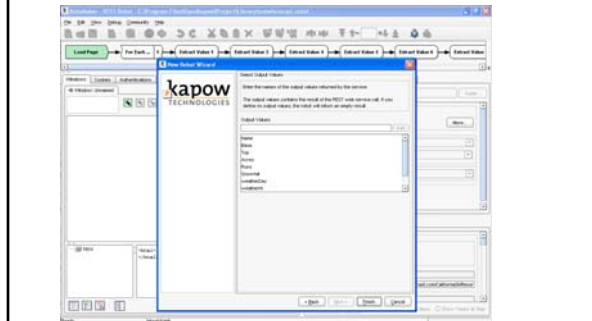
REST Agent

Create a REST web service robot

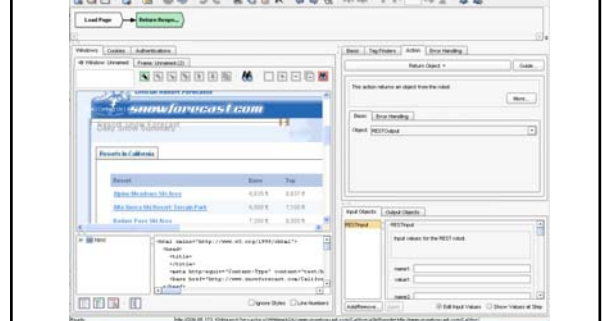


REST Agent

Define output variables

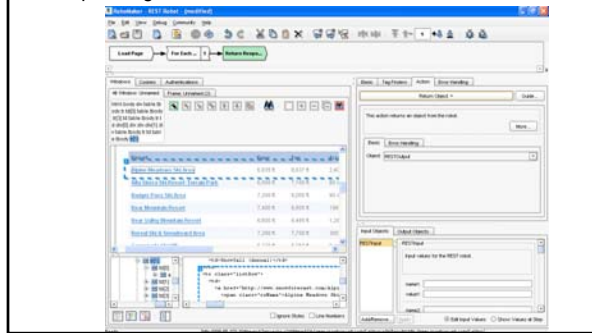


REST Agent



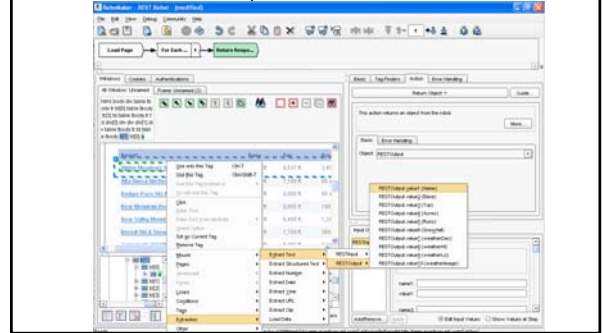
REST Agent

Loop through all resorts

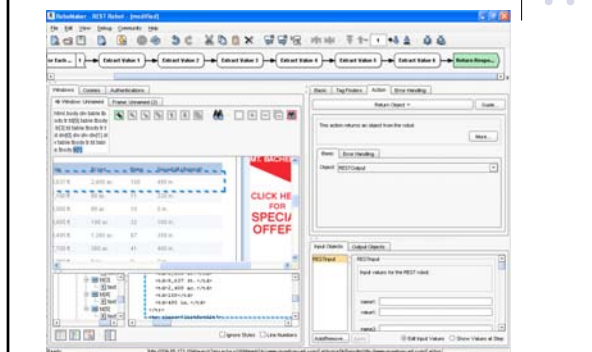


REST Agent

Extract name, base, top, acres, run, snowfall

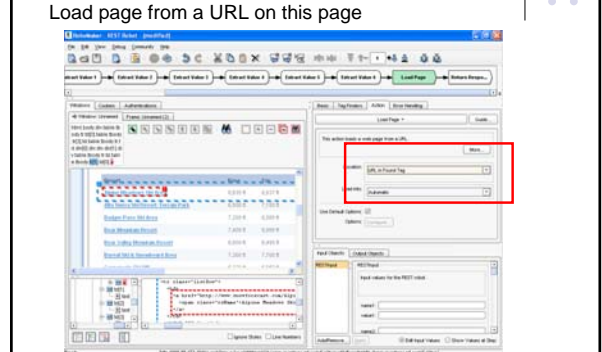


REST Agent

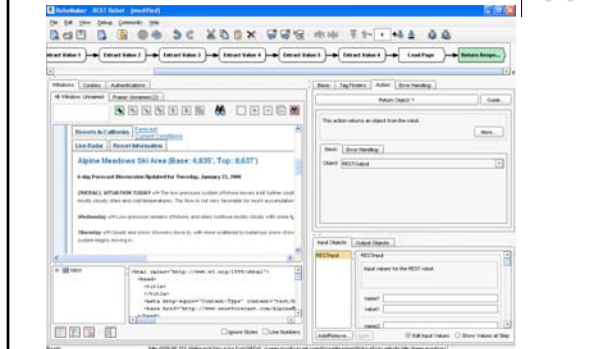


REST Agent

Load page from a URL on this page

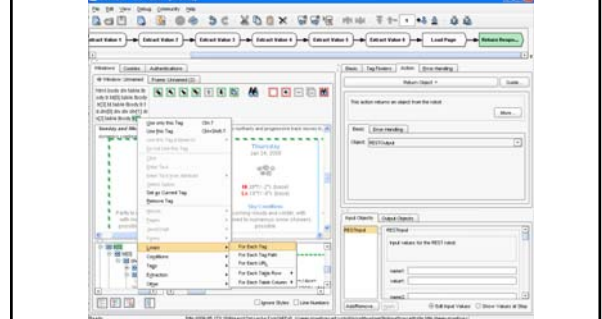


REST Agent



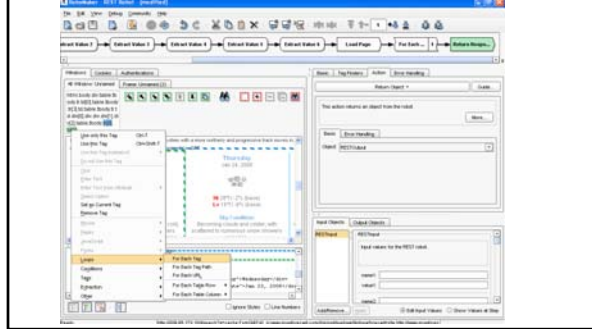
REST Agent

Loop over both rows of weather



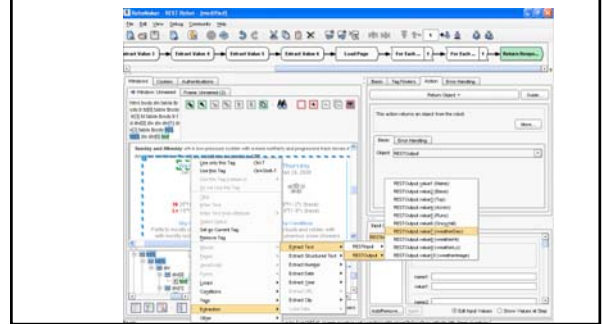
REST Agent

Loop over three columns in each row

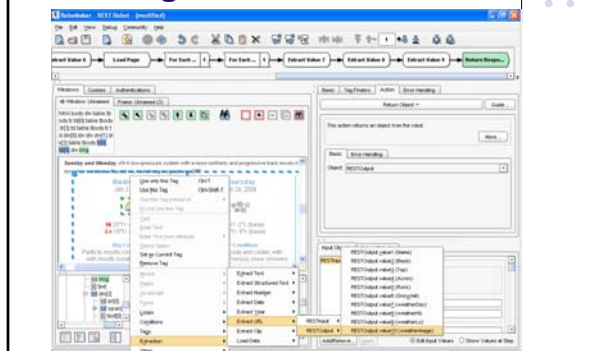


REST Agent

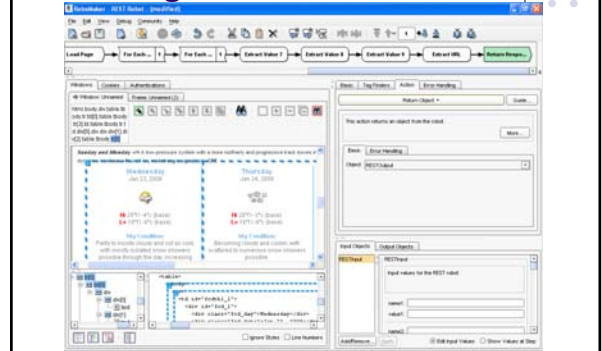
Extract weather for each day



REST Agent

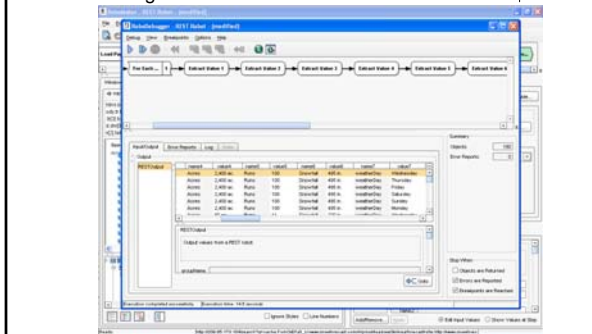


REST Agent



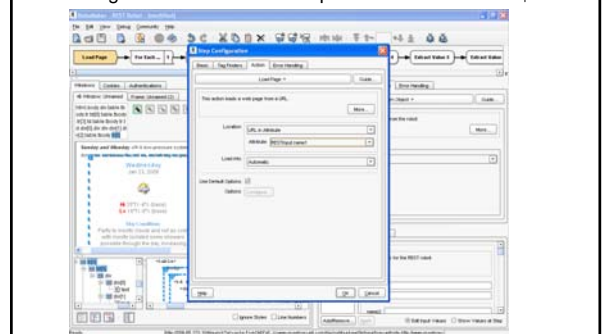
REST Agent

Test the agent



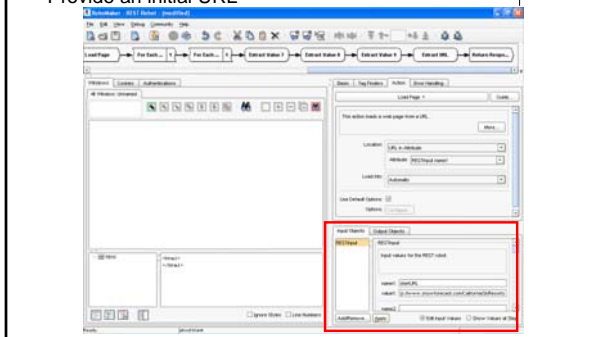
REST Agent

Alter agent to take a URL as input



REST Agent

Provide an initial URL



REST Agent

Publish agent and run it



REST Agent

Results, not divided into separate elements



REST Agent

Assign output group name to divide into elements

