LumberJaczk

Web data acquisition and on-top-of-Web applications

Viktor Zigo

About

Viktor Zigo

Lixto Software GmbH.

Senior software architect

DBAI, TU Wien

PHD student

LumberJaczk

Semantic Web

- **Web of data and relations** vs. web of proprietary documents
- Annotated content in machine readable representation and common ontology
- Data interchange, integration and inference

Uhm, cooool but…

LumberJaczk

Semantic Web ???

- Inherently agnostic Web segments
  - prices, services, news, competition
- Google ignores it
  - not visible = 80% fake
- No “semantic browser”
- Behavioral aspects
  - navigation, interaction,
- Complexity
  - WS are cool, REST and XmlRpc are way cooler
- Legacy Web, Web 2.0

To Do

- Web-2-Sweb conversion
- Data integration and aggregation
- Adding new functionality upon Web

  largest information data-source

Web monitoring, data searching or meta-search, making Web more structured and semantic, integration of Web enabled apps, content aggregation, Web tests automation, screen scraping, web extraction, Web remixing or mash-up

LumberJaczk

An Example – Simple Meta Search

Flight Search Application:

"Search for the best flight offers according to user preference in real-time out of several airlines"

User : *From, To, Date*

easyJet.com

SKY EUROPE

RYANAIR.COM

197.00	EUR	Wed. 17.05. 07:15	Airberlin AB 8893
197.00	EUR	Wed. 17.05. 15:55	Airberlin AB 8855
197.00	EUR	Thu. 18.05. 07:15	Airberlin AB 8893
221.75	EUR (GBP)	Tue. 16 May 06 11:10,	Air Berlin
231.09	EUR (GBP)	Tue. [illegible]ay 06 11:30,	Air Berlin
280.73	EUR (GBP)	[illegible]. 16 May 06 15:55,	Air Berlin
401.46	[illegible]	Tue. 16 May 06 14:40,	Czech Airlines CSA
411.[illegible]	EUR (GBP)	Tue. 16 May 06 19:30,	Austrian

AlephZarro

LumberJaczk

Process Stages

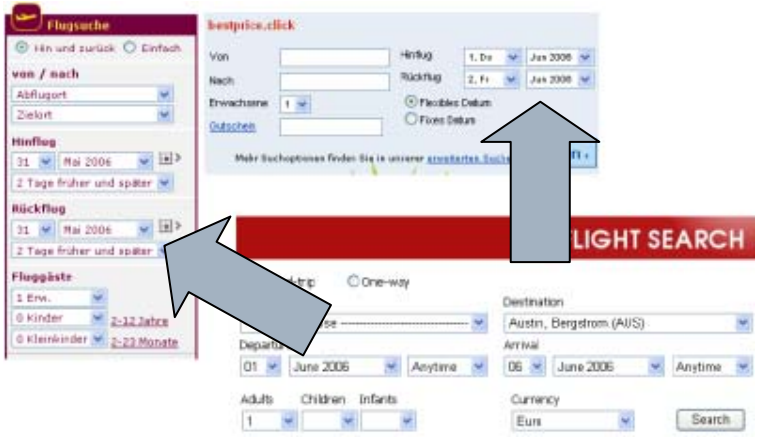
1. User query mapping to portal search forms
2. Web navigation and query setting
3. Parallel execution
4. Results synchronization
5. Data extraction (navigation)
6. Data unification, merging, cleaning
7. Presentation and Syndication

LumberJaczk

Query Mapping

Map user query to Portal Specific Queries

- Hard-coded
- Automatic on-fly mapping



Problems :

- M :N mapping
  `2006-01-03`
  `3rd, January, 2006`
- syntactic, semantic
- Mapping of intervals (prices, age)
- Units
- Multilingual mapping, synonyms, free text

AlephZarro

LumberJaczk

Deep Web Navigation

Types :

- Crawling
- Given navigation
- Auto navigation

Obstacles :

- Dynamic content
- Sessions, state-full
- Authorizations,HTTPS
- Proxies
- AJAX
- Changing structure
- Page interaction
  - Forms
  - In-page navigation
  - Flash

Common Wrapping Techniques

- Techniques
  - Structural (tree)
  - Syntactic (regexp)
  - Tokenization
  - Text analysis
  - Visually based

Wrapper Generation

- Hardcode
- Visually Design
- Supervised learning
- Automatic

  (pattern recognition, templates)

LumberJaczk

Normalization, transformation and data-cleaning

- Inverse process of query mapping
- Common output schema (ontology)
- Value mapping

  `L.S.O, Symphonic Orch London, Orchester, London`
- Duplicate entries, data clustering

LumberJaczk

Syndication

- Format
  - XML, RDB mapping, RDF, PDF, text….
- Publishing
  - Portals, legacy app, CMS, RSS, DB …

LumberJaczk

Axe the Web!

http ://lumberjaczk.org

Viktor Zigo, May 17th 2006

LumberJaczk Project

"An open-source technology that enables better ways to get and use data and information on the Web."

Pragmatic approach :

light-weight, flexible, end-to-end, get things done first

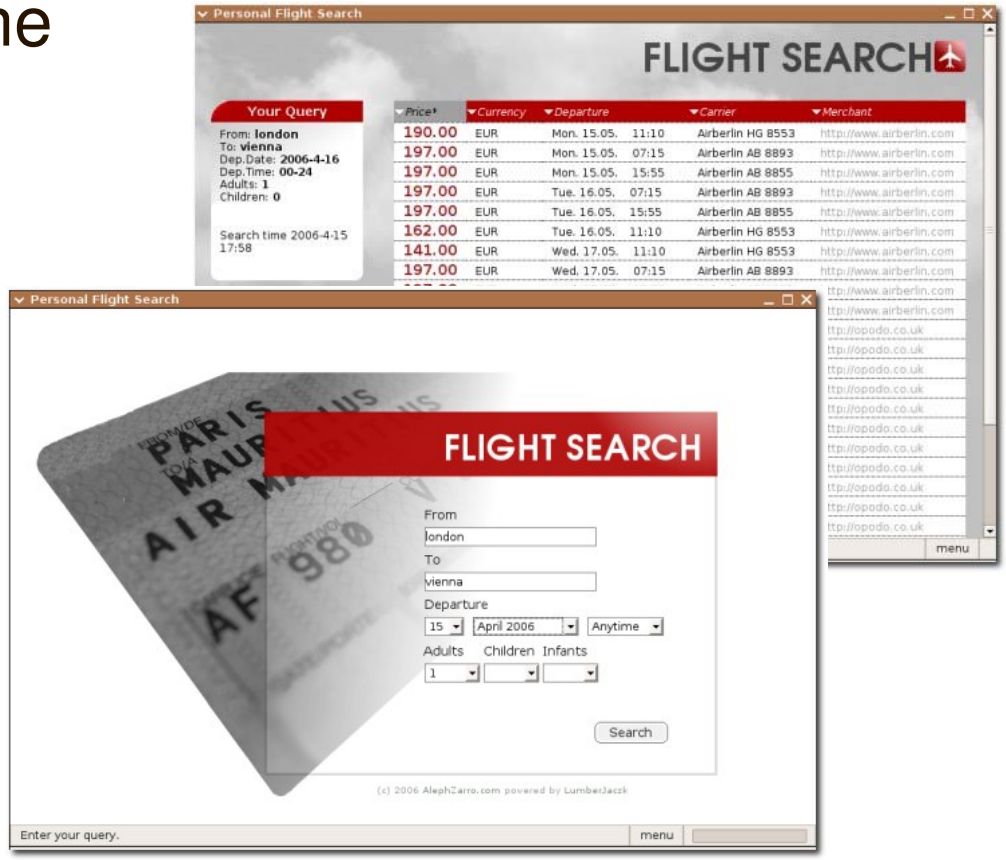
- Aug 2005 : Started
- May 2006 : usable, rolling out

Tech background

- Mature and rich Mozilla code-base, open source
- Many similar applications :
  - MIT PiggyBank, Chickenfoot, GreaseMonkey
  - Test4Web, Selenium, Solvent
  - AJAX Toolkit Framework (IBM)
  - FF extensions
- Real-world experience with data extraction

Appetizer – Personal Flight Search

- Running standalone **Flight Search** application
- Download & run
- 2 days (incl. GUI)



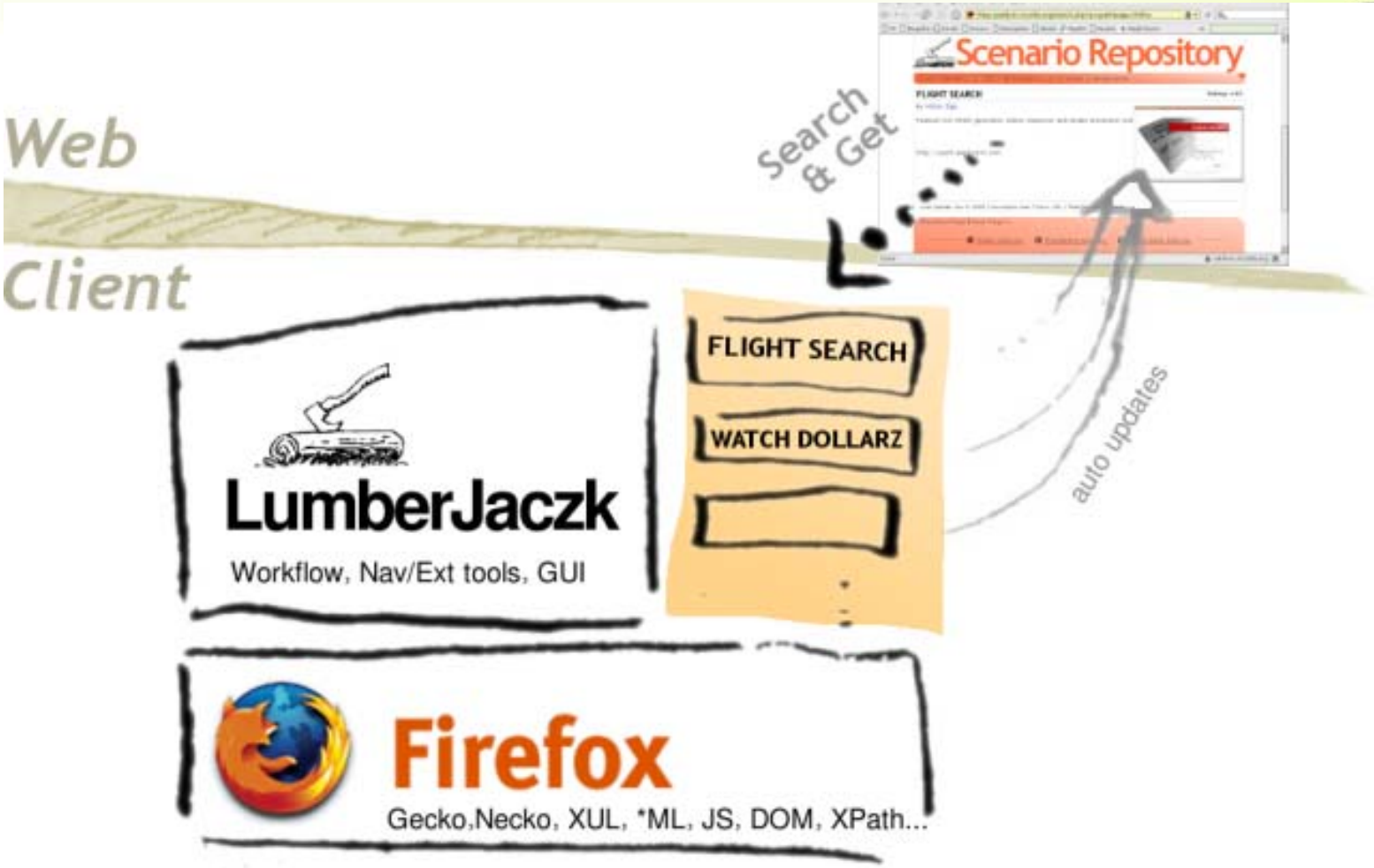
LumberJaczk

What a heck is this Lumber …?

A **generic workflow engine** and framework…

- …primarily **client-side**
- …for **execution of apps** (*scenarios*)
- …operating **on top of the Web**
- …provided in very **portable packages**
- …with **rich GUI** and **user Interactions**

Architecture and Principles



Generic Player

Get the generic player once, either

- as a **standalone app**
- or in your Firefox **browser**

You can do this then :

1. Browse web repositories
2. Install them from web by single click
3. Customize, manage and run them from within your client (or browser)
4. Get automatically the updates

Standalone Generic Player

- Launch and manage all the installed scenarios from your desktop
- Manage your recent runs and results

Execution Modes

- standalone scenario (desktop)
- standalone generic player (desktop)
  - runs any installed scenarios
- single scenario browser extension
- generic player as a browser extension

Scenario properties

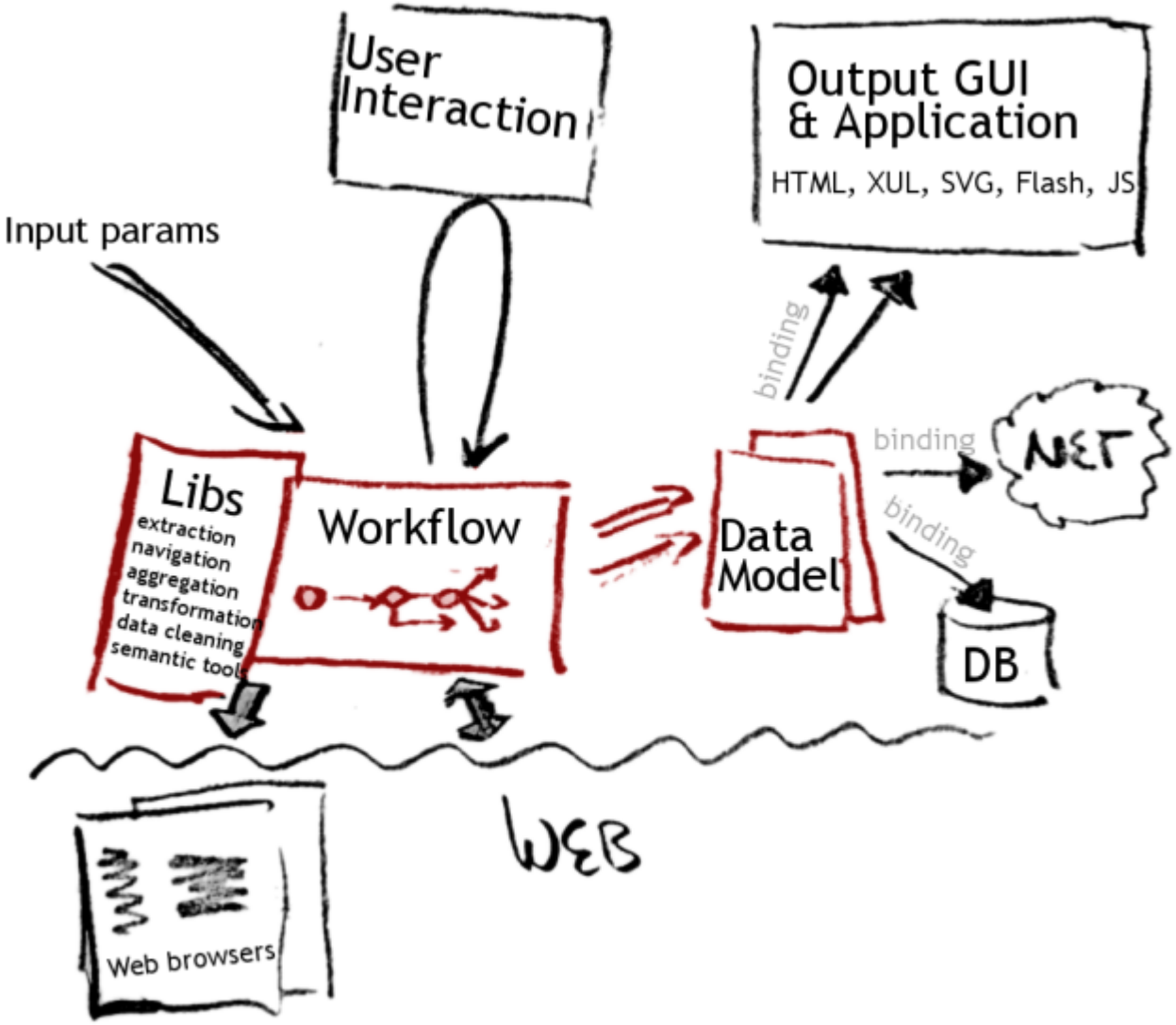
It is a single application LumberJaczk runs.

- small (kB), compact, and self-contained
- installs by finger-click
- installs directly from internet web page
- automatically updated
- dynamic scenario federations
  - e.g. simply add new sources to your app
- might be signed
- platform independent

Scenario features

- application business logic
- configuration GUI
- output data schemas
- output application (GUI + logic)
  - presentation in GUI
  - filtering, graph generation, system integration, etc.
- löalizätion

How does it work ?



LumberJaczk

Scenario GUI …

….is rich… more than rich !!!

It can be a blend of

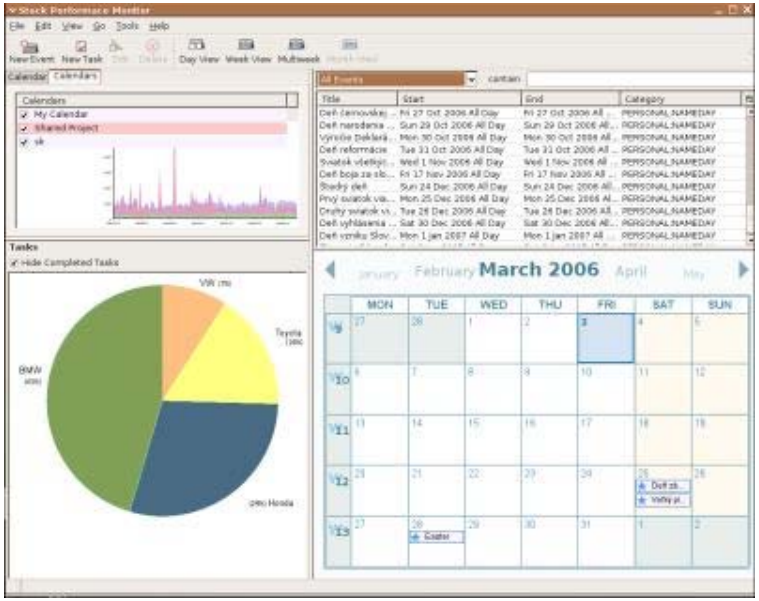
- **HTML, SVG**
- **Flash**
- **native GUI**
- **integration to existing Web page**

LumberJaczk

Native GUI Scenario

- Native widgets
- Application logic
- Graphs and charts

Business events monitoring example

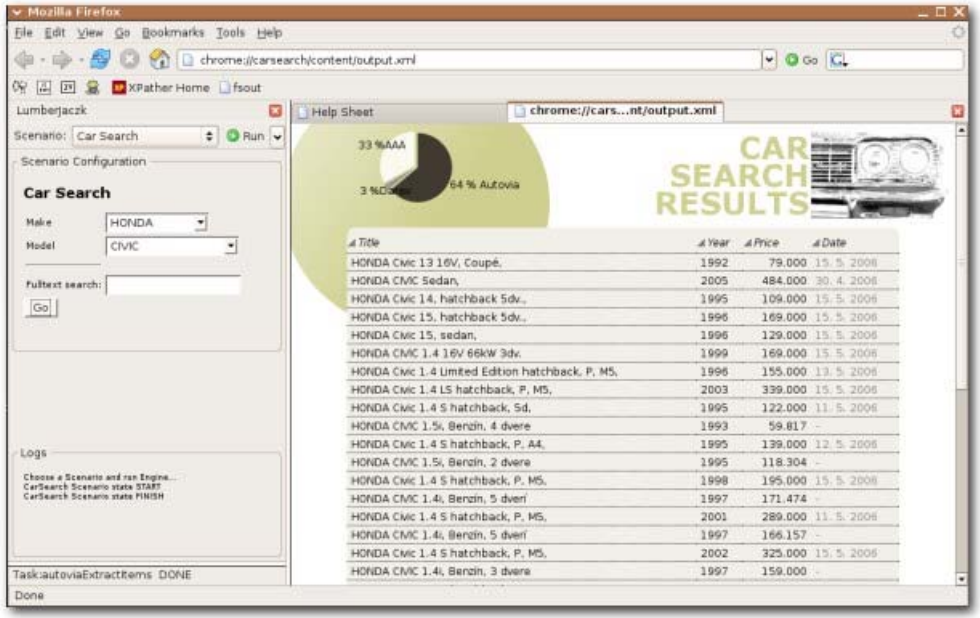


AlephZarro

In-Browser Scenario

- up and running in a browser by 2 clicks
- launched from generic player
- HTML, native, Graphs possible
- Integration to 3rd party Web

Car Search scenario example



Live and Dynamic

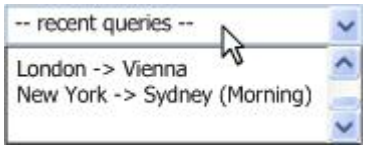
Execution experience can be very live and dynamic

- watch the progress
  - search, workflow, communication, extraction
- live results modification
  - adding new, making more precise
- on fly generation of graphs
  - cool eh ?
- stop, pause, resume, inspect, and interact !

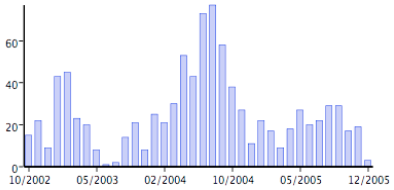
It lives with you

The scenario player can :

- remember your inputs and configurations
  - you have run recently
  - you have predefined



- keep the state and results from the previous executions
  - and combine them for example (graphs)



Getting Technical

Viktor Zigo, May 17th 2006

LumberJaczk

Engine

- Generic workflow engine
- Scriptable actions
- Execution Context
- Output Data Models – XOutput
- Navigation, Extraction, transformation libraries

LumberJaczk

Workflow engine

- dynamic flow decision
- synchronization on any custom event
- parallel execution (forking)
- native iterations (for-each)
- sub-processes (nesting)
- arbitrary scriptable action in any state (JS)
- interpreted
- arguments, shared variables
- static function, custom action libraries

Actions

- JavaScript (Java)
- complete Mozilla control
- networking, windows, DOM, XPath, RegExp, XML, XSL, SOAP, WS, XmlRpc
- Libraries for navigation, extraction, and data integration
- XOutput
- interaction with Scenario GUI
- error handling

LumberJaczk

Workflow Script

```
  // output.set('in_carrier', in_carrier);
   output.set('link', transformedLink);
   output.set('merchant','http://opodo.co.uk');
   output.link('flight',transformedLink, true);
}

//*****************************************************************************
//********** AIRBERLIN *****************
function airberlin(){
   lib.addTab('http://www.airberlin.com/site/index.php?LANG=eng'); //ONLINE
   return syncOn( window , 'airberlinFillForm', 500)
}

function airberlinFillForm(){
   var root = lib.assertFirst( lib.xpathFrames( "/html/body/table[1]/tbody/tr[2]/td[2]/t

   var input = lib.assertFirst( lib.xpath(root, "./tbody/tr[1]/td/table/tbody/tr[2]/td[1]/s
   var option=lib.select(input, parameters.from, Lib.MATCH_TOP | Lib.MATCH_VISUAL

   input = lib.assertFirst( lib.xpath(root, "./tbody/tr[2]/td/table/tbody/tr[1]/td/table/tb
   lib.click(input);

   input = lib.assertFirst( lib.xpath(root, "./tbody/tr[5]/td/table/tbody/tr[2]/td[1]/selec
   option=lib.select(input, parameters.to, Lib.MATCH_TOP | Lib.MATCH_VISUAL);
```

Data Output - XOutput

- native XML generation
- simplified model schemas
- linking with XLink
- unique XOutput technique
- parallel data generation
- per-partes flushing
- direct binding to output device (GUI)
  - HTML, XUL applications, HTTP, SOAP, XMLRPC, WS

LumberJaczk

Further usage

Bottom-up approach

- Generic client application workflow engine
- Rule-based extraction engine
- Hierarchical extraction engine (XML)

Status

- Usable system
- Preparation for OS release
- Building web site and infrastructure

Web : http ://lumberjaczk.org

Interested ?

interested@lumberjaczk.org

Related Projects

Viktor Zigo, May 17th 2006

Related and New Projects

- Public web repository of Scenarios
- Generic Wrapping
- Collaborative semantic database
- Server side solutions
- Data merging and cleaning toolset
- Visual Scenario Development

Scenario Repository Portal

- Public repository for LJ scenarios
  - Publishing /review / share process
  - Multi-dimensional categorization, tagging
  - Scenario Marketplace

Example :

https ://addons.mozilla.org

LumberJaczk

Generic Wrapping I

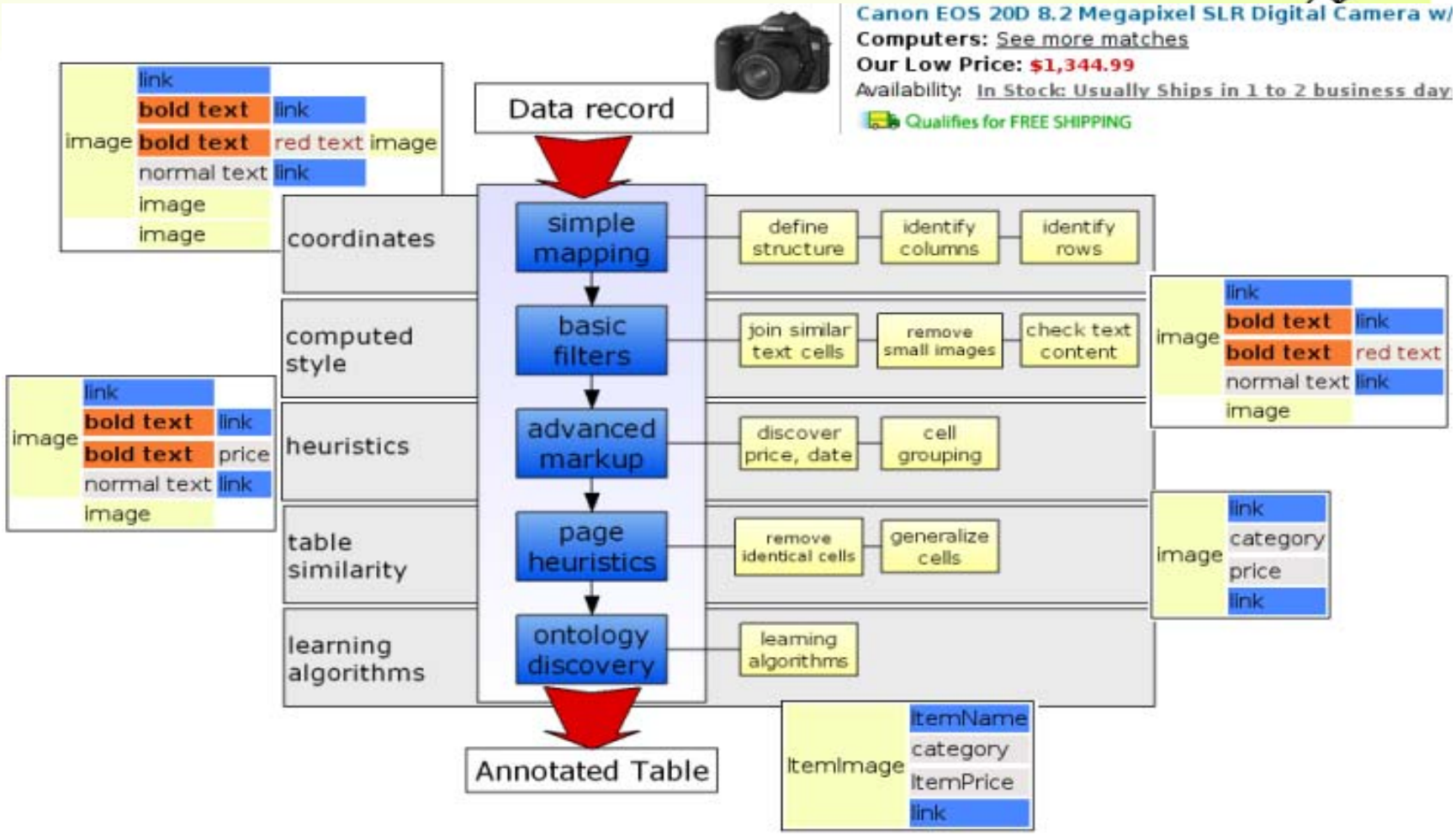
Visually based semantic annotation

(Peter Szniek)

- Problems of tree-based /syntactic wrapping
  - Strong noise, dynamic, complex
- Visual Elements (WYSIWYW)
  - HTML, PDF → **meta-table**
  - Robust and simple structure for data extraction

LumberJaczk

Generic Wrapping I - schema



AlephZarro

LumberJaczk

Generic Wrapping II

Inductive Wrapper Generation of Wrappers

(Michal Ceresna)

- Applies machine learning
  - Tree-aligning
  - Attribute classification
  - Set covering machines

Collaborative Semantic Database

- Scenario users supply and update the shared data storages
- In turn, the scenario users query also the data storages
- Data storage search engine
- Data schemas and ontologies mapping
- Data integration and inference

LumberJaczk

Server-side solutions

- Farming
  - Scalability, load-balancing, failover
- Portal integration
  - Web access to server-side scenarios

LumberJaczk

Misc

- Data cleaning toolset
- Advanced browser controls
- Internet Explorer hacking ☺
- Scenario development tools
- Core
  - Semantic Data Models – RDF, Sesame
  - Profiles, history support
  - XUL Runner
  - Scenarios Lego
  - Crawling

LumberJaczk

Contact

http ://lumberjaczk.org

Viktor Zigo

Interested ?

interested@lumberjaczk.org

AlephZarro