# HTML-aware tools for Web data extraction

Thesis presentation

**Student:** *Xavier Azagra*

**Supervisor:** *Andreas Thor*



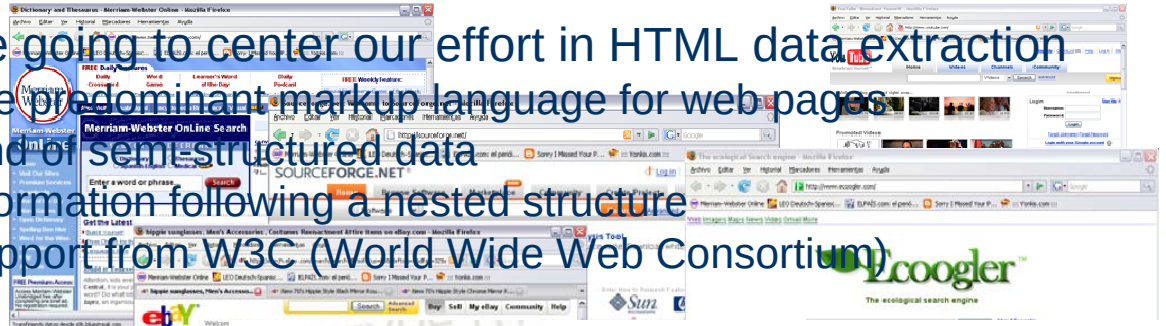
UNIVERSITÄT LEIPZIG

# Table of contents

- Introduction
- Data Extraction Process
- Data Extraction Tools
- Realized tests
- Future Work

# Introduction

- We are going to center our effort in HTML data extraction
  - The predominant markup language for web pages
  - Kind of semi-structured data
  - Information following a nested structure
  - Support from W3C (World Wide Web Consortium)

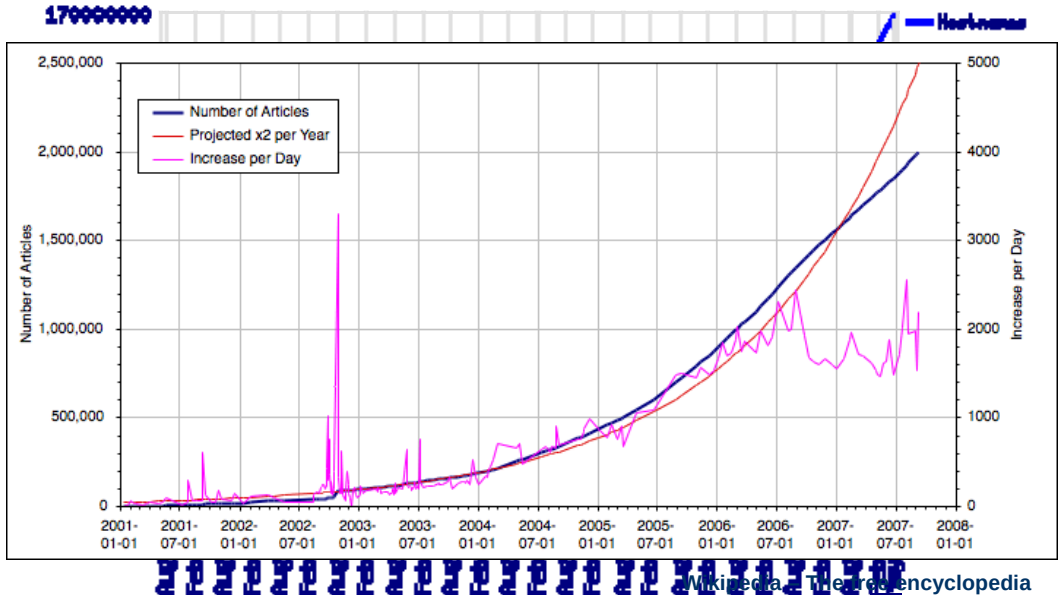



# Introduction

## Internet growth

**168** Million sites

**1400** Million of Internet users

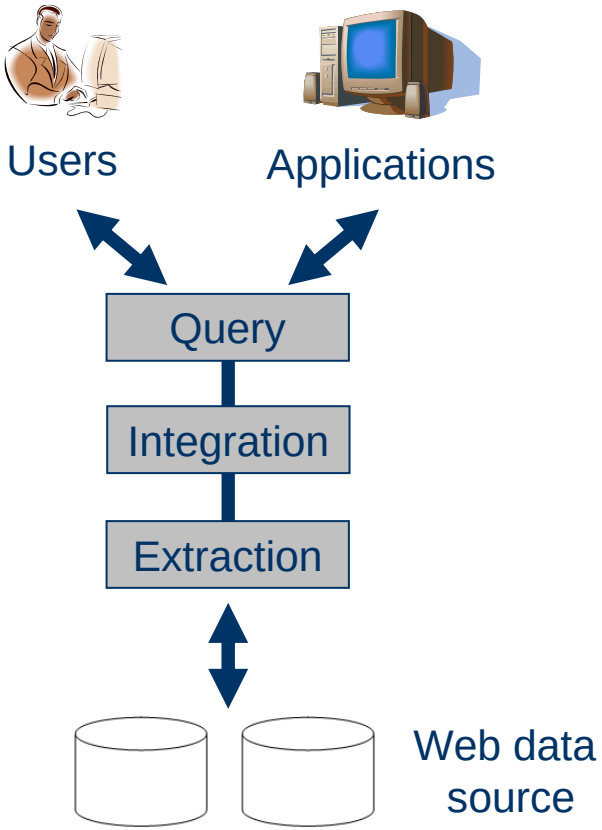



Wikipedia, the free encyclopedia

May 2008 Web Server Survey - www.netcraft.com

# Introduction

## Purposes of Web data extraction




- Get information from the Web to be used in other areas or by applications
- Information retrieval ( e.g. Feeds, Web search engines…)
- Let the user to access particular data from the Web
- Economical issues ( e.g. stock market, shopping comparison…)

# Data extraction process

## Main problems

- Internet was designed as a source of data for a human use. Problems appear when we want to extract data from HTML

- Data not presented in HTML format:
  - Password protected sites
  - Cookies
  - Sessions ID’s
  - Javascript
  - Dynamic content

- Deep resources:
  - Unlinked content
  - Contextual web
  - Limited access content

# Data extraction process

## Types of content

### Free text

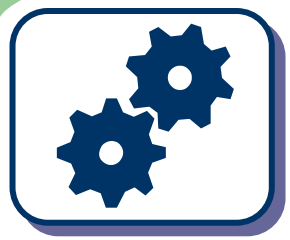
- Natural language texts
- Patterns involving syntactic relations between words or semantic classes of words

### Structured text

- Textual information following a predefined strict format
- Use of the format description

### Semi-structured text

- Between unstructured collections of textual documents and fully structured tuples of typed data
- Extraction patterns are often based on tokens and delimiters



# Data extraction process

## Ways to perform data extraction

**Manual**

- Precise
- Treat elements individually

**API**

- Specific Web Sites
- Limited specifications

**Wrapper**

- Set of methods
- Independent of source

**Manual**

- Ad hoc code
- Not trivial
- Error-prone

**Semiautomatic**

- Support tool
- GUI support
- Less Error-prone

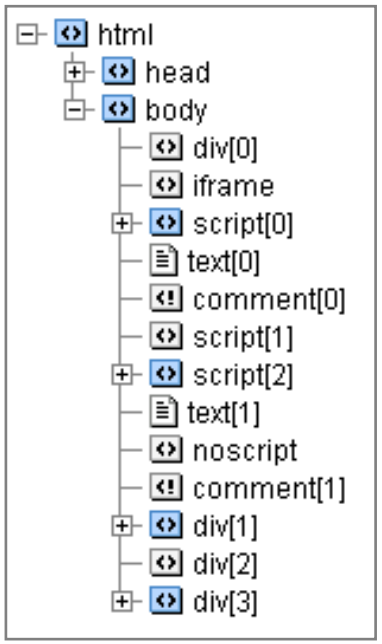
**Automatic**

- Machine-learning techniques
- Supervised learning

# Data extraction process

## HTML structure for data extraction

- When speaking about HTML-aware tools, before performing the extraction process, these tools turn the document into a parsing tree




- Each node represents a tag
- Outer tags are leaves
- Expressions to navigate through all the hierarchy

```
html.body.div[1].div[1].div.div[1].div[5].div.div[0].span.text
```

- Maximum precision is found on the content of a leave

# Data extraction process

## HTML problems to extract data (I)

- Presentatior
  - Logic, si
  - Unorgan
- Bad constru
  - Bad plac
  - Repeate
  - No close
- Nested data
  - Element

contain diff

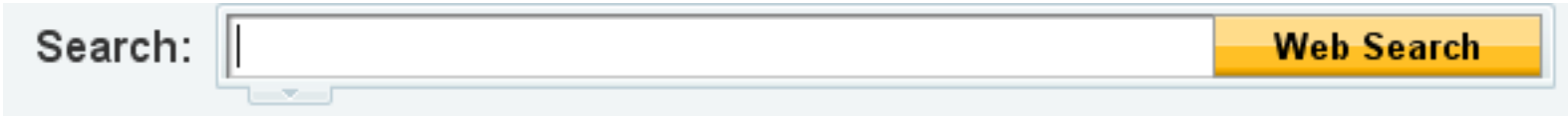

orrect extractions

t by element could

# Data extraction process

## HTML problems to extract data (II)

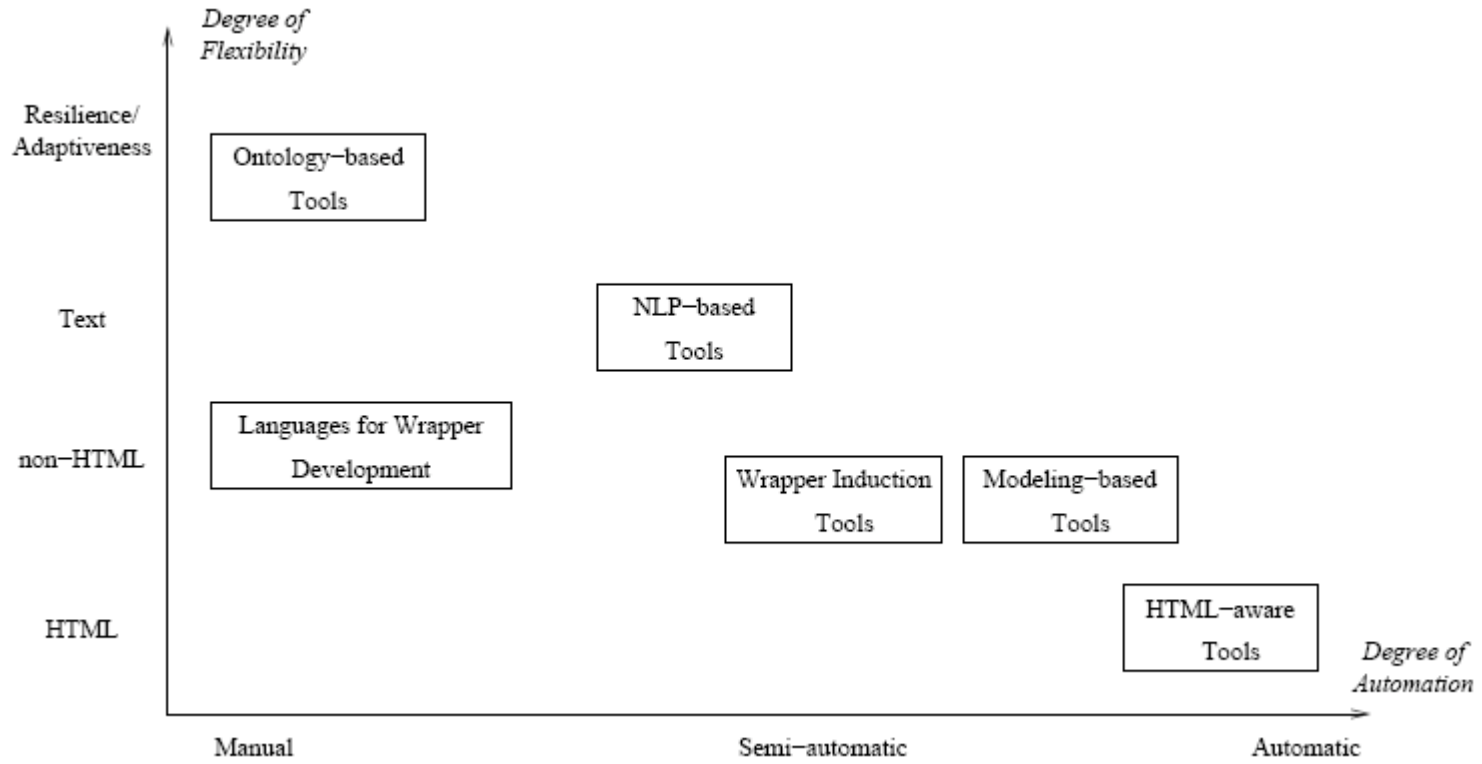
- Problems choosing the correct Web page source example
  - Content structure could change depending on some factors
  - Example: Result page of Web Search Engines




- Problems using scripts or dynamic content
  - Hidden or changing information
  - Syntax different to HTML
  - Javascript, PHP, AJAX or Flash


# Data extraction tools

## Taxonomy (I)



# Data extraction tools

## Taxonomy (II)

**Languages for wrapper development**

- Assist wrapper construction
- Alternatives to general purpose languages

**Ontology-based**

- Extraction relying directly on the data

**NLP-based**

- Based on syntactic and semantic constraints

**Wrapper induction**

- Rules derived from a given set of training examples

**Modeling-based**

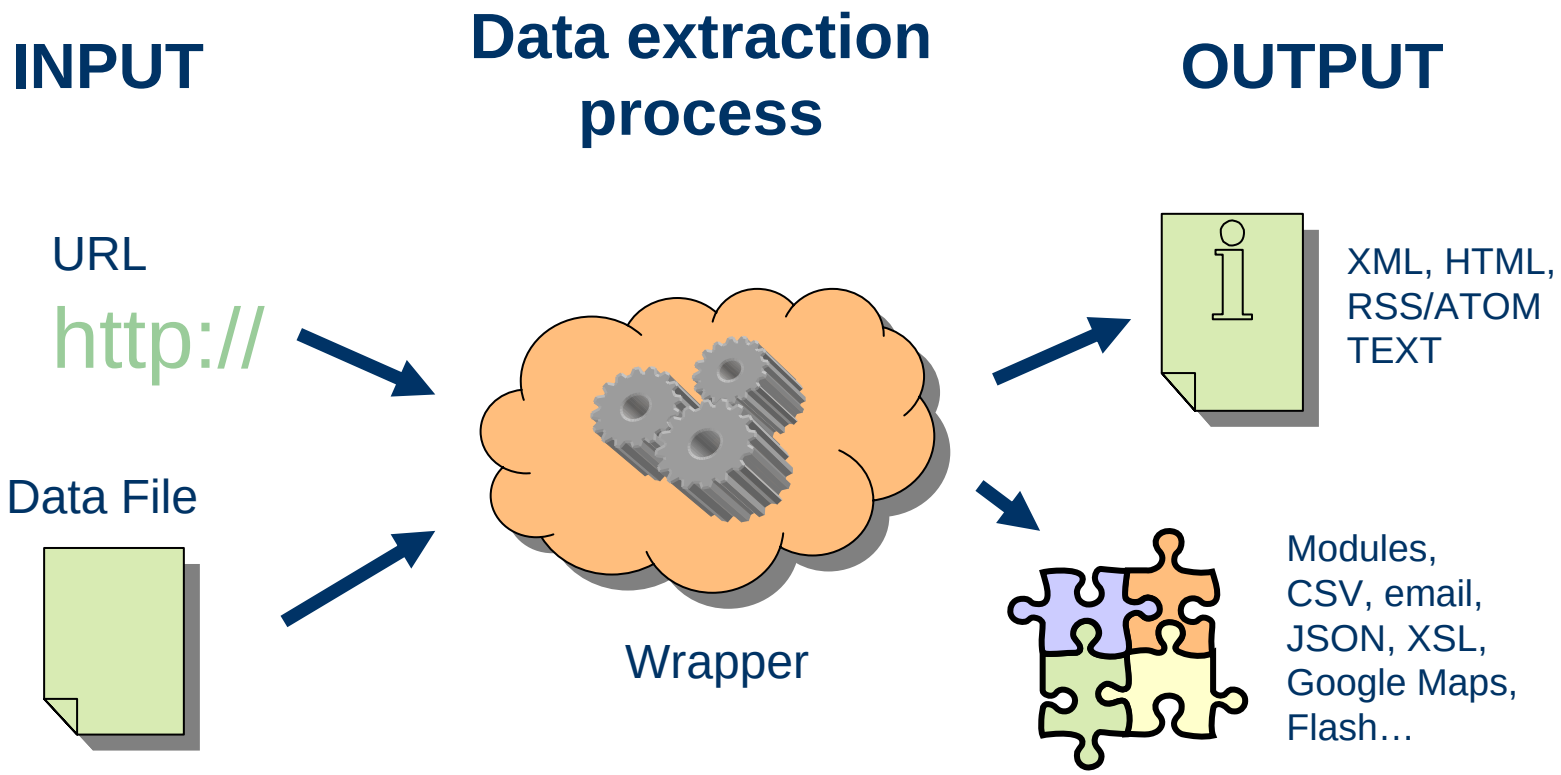
- Try to locate in Web pages portions of data that implicitly conform to a structure

**HTML-aware**

- Rely on inherent structural features of HTML documents

# Data extraction tools

## Flow of data



# Data extraction tools

## Structure

- 10 HTML-aware tools
- Categorization of this tools using several criterias
- Test-bench scenarios

# Data extraction tools

## Used HTML-aware tools

- Dapper
- Robomaker
- Roadrunner
- XWRAP
- Lixto
- Webharvest
- Goldseeker
- WinTask
- Automation Anywhere
- Web Content Extractor

- Commercial and non commercial tools
- Shell and GUI support tools
- Screen scrapping and non screen scrapping tools
- Linux and Windows tools
- ...

# Data extraction tools

## Structure

- 10 HTML-aware tools
- Categorization of this tools using several criterias
- Test-bench scenarios

# Data extraction tools

## GUI

### ■ No GUI

- Shell commands
- Configuration files and coding
- Input files
- Roadrunner

### ■ Integrated browser

- Direct Interaction between the tool and the navigation browser
- Visualize information of the Web elements
- Lixto, Robomaker, Web Content Extractor

### ■ Web browser

- Loads Javascript and Dynamic content
- Separation between the tool and the window browser
- Automation anywhere, Wintask

# Data extraction tools

## Resilience

- **Capacity of continuing to work properly in the ocurrence of changes in the pages for which they are targeted**

- Common changes to:
  - the data
  - the structure
    - Add, erase or modify elements
  - the visual design
  - introduce new technologies (AJAX, PHP, Javascript…)

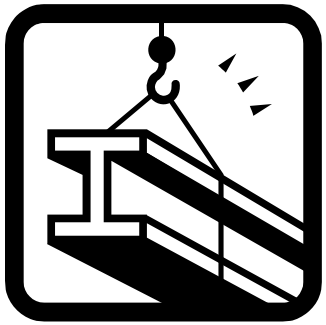
- The resilience grad varies depending the used tool

# Data extraction tools

## Adaptiveness

- Grade of a wrapper for built pages of a specific Web source on a given application domain to work properly with pages from another source in the same application domain

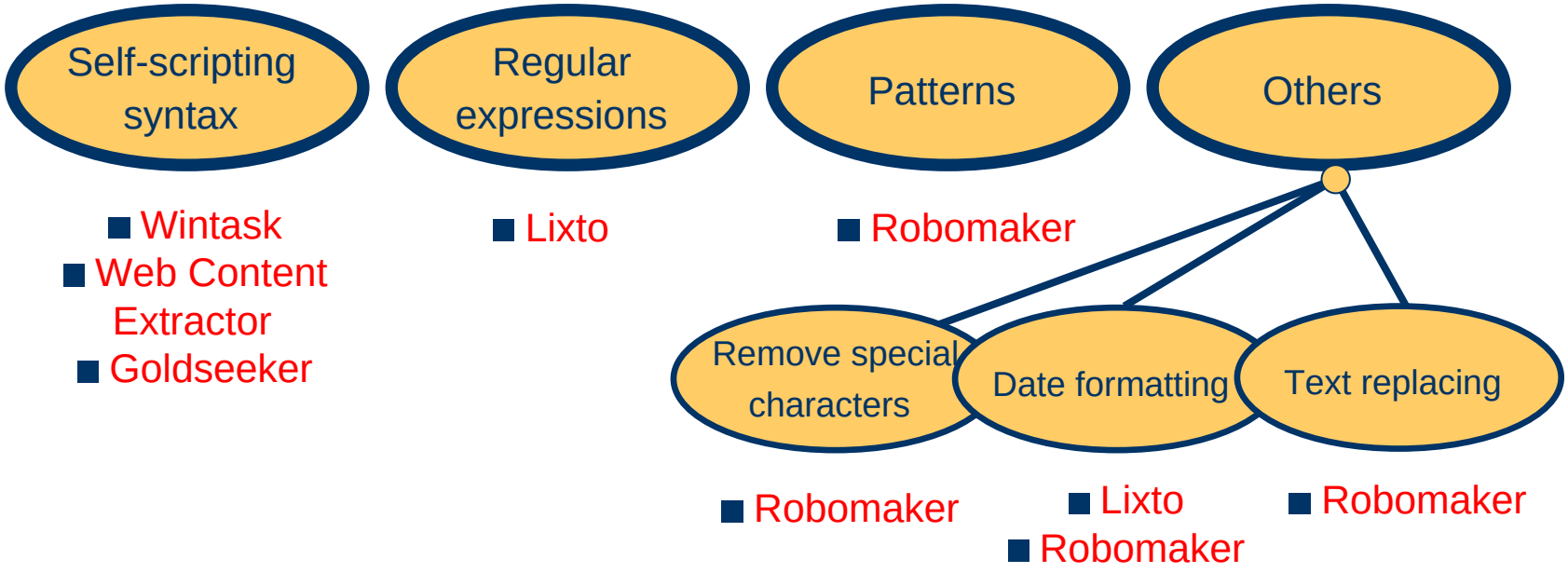
- From all of the taxonomy of web data extraction tools only the Ontology-based tools feature fully resilience and adaptativeness properties



# Data extraction tools

## Scripting and expressions

- The atomicity of the HTML parsing tree is found in a leave (outer tag)
- Necessity to extract information in a more precise way



# Data extraction tools

## Input variables

- In some cases we need input variables to realize searches through Internet:
  - Ebay
  - Web search engines
  - Youtube
  - Amazon
  - ...
- We want to extract data from the resulting pages, we need tool support
- Robomaker, Dapper, Lixto, Wintask

# Data extraction tools

## Input/Output formats

| | Input Formats | Output Formats |
|---|---|---|
| **Dapper** | HTML | XML, RSS, HTML, Modules, Atom Feed, CSV,JSON,XSL, YAML, email |
| **Robomaker** | HTML | RSS/Atom Feed, REST Web Service, Web Clip |
| **RoadRunner** | HTML | XML, HTML |
| **XWRAP** | HTML | XML |
| **Lixto** | HTML | XML |

| | Input Formats | Output Formats |
|---|---|---|
| **WebHarvest** | HTML | XML |
| **GoldSeeker** | HTML and documents | Text |
| **WinTask** | HTML and documents | File, Excel, DB |
| **Automation Anywhere** | HTML and documents | File, Excel, DB, EXE |
| **Web Content Extractor** | HTML | File, Excel, DB, SQL script File, MySQL script File, HTML, XML, HTTP submit |

# Data extraction tools

## General features (I)

| | Interface | Complexity | Resilience | Execution time | Free |
|---|---|---|---|---|---|
| **Dapper** | Internet browser | Low | Good | Very Good | Yes |
| **Robomaker** | Program GUI, Internet browser | Medium | Very good | Very Good | Yes |
| **RoadRunner** | Linux Shell | Medium | Poor | Good | YES, GNU GPL License |
| **XWRAP** | Internet browser | Medium | Good | Good | Yes |
| **Lixto** | Program GUI, Internet browser | Medium | Good | Very Good | No, requires license |

# Data extraction tools

## General features (II)

| | Interface | Complexity | Resilience | Execution time | Free |
|---|---|---|---|---|---|
| **WebHarvest** | Program GUI | High | Good | Good | Yes |
| **Goldseeker** | Internet browser | Medium | Good | Poor | Yes, GNU LGPL License |
| **Wintask** | Program GUI, internet browser | Medium | Poor | Good | No |
| **Automation Anywhere** | Program GUI, Internet browser | Low | Poor | Good | No |
| **Web Content Extractor** | Program GUI, Internet browser | Low | Poor | Poor | No |

# Data extraction tools

## Advanced characteristics

| | Input variables | Scripts usage | Non static content pages | More than one page | Javascript or Dynamic content |
|---|---|---|---|---|---|
| **Dapper** | Yes | No | Yes | No | Good |
| **Robomaker** | Yes | Yes | Yes | Yes | Good |
| **Roadrunner** | No | No | No | No | Poor |
| **XWRAP** | No | No | Yes | No | Poor |
| **Lixto** | Yes | Yes | Yes | Yes | Good |
| **WebHarvest** | No | No | Yes | No | Poor |
| **Goldseeker** | No | Yes | Yes | No | Poor |
| **Wintask** | By script | Yes | No | Yes | Good |
| **Automation Anywhere** | No | No | No | Yes | Good |
| **Web Content Extractor** | No | No | No | No | Good |

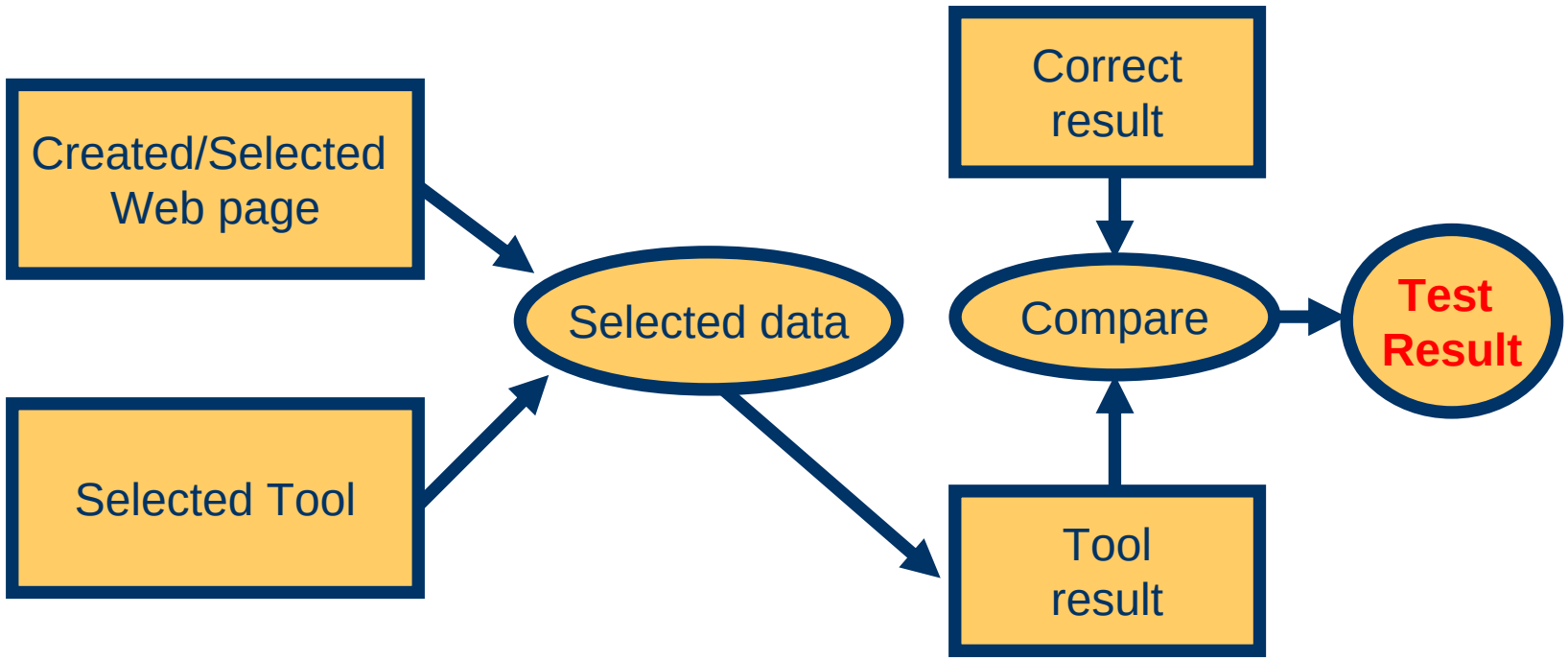
# Data extraction tools

## Structure

- 10 HTML-aware tools
- Categorization of this tools using several criterias
- Test-bench scenarios

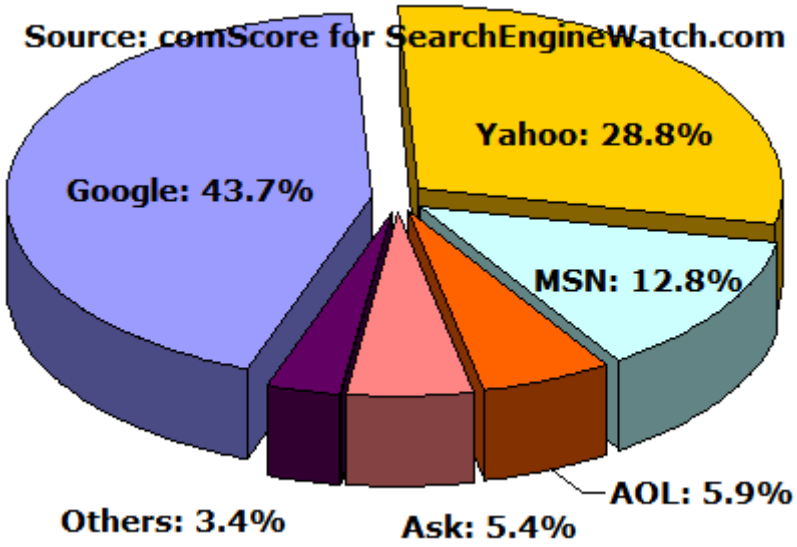
# Realized tests

## Methodology



# Realized tests

## Web search engines (I)




- One of the most used resources of the Web
- Use of input variables and dynamic result pages
- Yahoo! Search uses a live search input form

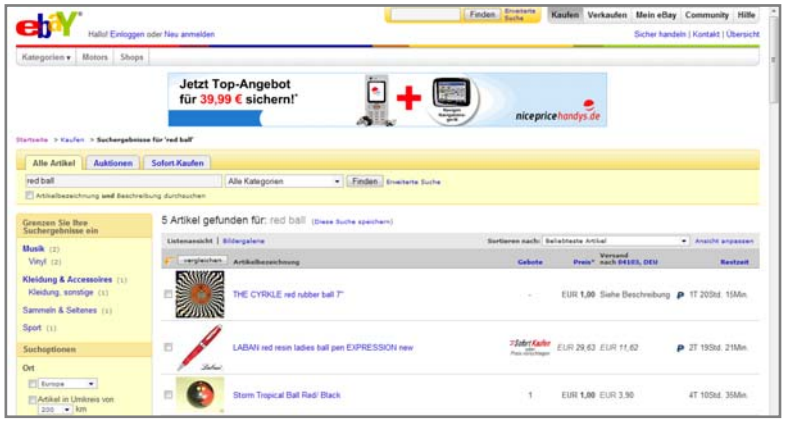
# Realized tests

## Web search engines (II)

| | Google Search | Yahoo! Search | MS Live Search |
|---|---|---|---|
| Dapper | ✔ | ✔ | ✔ |
| Robomaker | ✔ | ✘ | ✔ |
| Lixto | ✔ | ✔ | ✔ |
| WinTask | ✘ | ✘ | ✘ |
| Automation Anywhere | ✘ | ✘ | ✘ |
| Web Content Extractor | ✔ | ✔ | ✔ |

# Realized tests

## Ebay

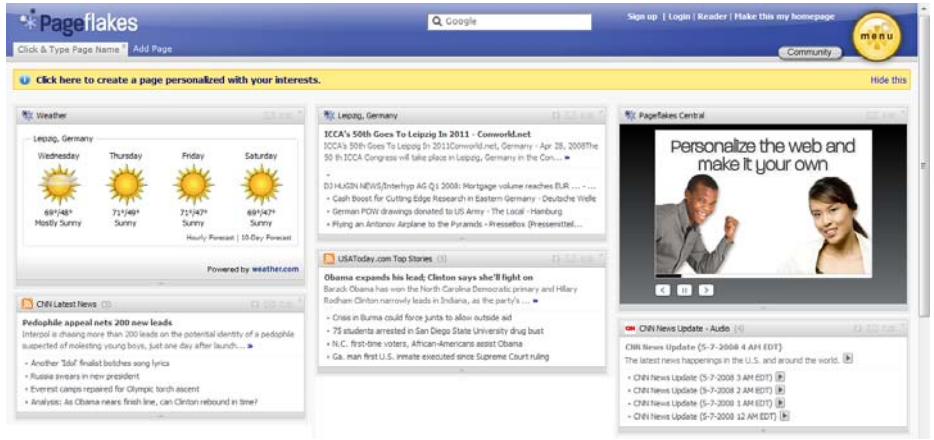



- The most important auction shop of Internet
- Use of input variables and dynamic result pages
- Fields containing variable content

| | Ebay search |
|---|---|
| Dapper | ✔ / ✘ |
| Robomaker | ✔ |
| Lixto | ✔ |
| WinTask | ✘ |
| Automation Anywhere | ✘ |
| Web Content Extractor | ✔ |

# Realized tests

## Dynamic content Web pages




Pageflakes

- AJAX based start page
- Use of Dynamic content and personalized user modules

| | Pageflakes |
|---|---|
| Dapper | × |
| Robomaker | × |
| Lixto | × |
| WinTask | × |
| Automation Anywhere | × |
| Web Content Extractor | × |

# Realized tests

## Resilience tests (I)

**1-** Obtain a result page of Amazon.com
**2-** Download the source page and related files
**3-** Upload to a test server
**4-** Configure tools to extract 4 fields: title, book format, new price and valuation

1. 


**The Jungle (Enriched Classics)** by Upton Sinclair (**Mass Market Paperback**) April 27, 2004)

Buy new: **$5.95** 54 Used & new from $2.67

Get it by **Wednesday, April 30** if you order in the next **10 hours** and choose one-day shipping.
Eligible for **FREE** Super Saver Shipping.

★★★★☆ 6)

**Excerpt** - page 7: "... THE **JUNGLE** Jurgis," of all men, to Jurgis Rudkus, he with the ..."

**Surprise me!** See a random page in this book.

**Books:** See all 128,055 items

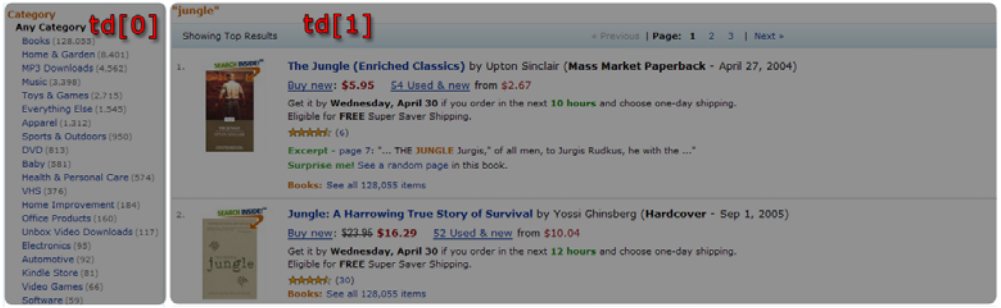
amazon.com

# Realized tests

## Resilience tests (II)

- Deleting content
- Modifying CSS style tags
- Duplicating extracted data
- Changing order of extracted data

- Deleting content Example:




| | Erase td[0] |
|---|---|
| Dapper | ✔ |
| Robomaker | ✔ |
| Lixto | ✔ |
| Web Content Extractor | ✘ |

# Realized tests

## Precision tests (I)

LIST OF PUBLISHED BOOKS

| Name | Author | Last Published edition |
|---|---|---|
| How to begin with Computers | Andrew Moss | 1998-07-07 First edition |
| Spain. The guide | Roberto Díaz | 1995-02-04 Second edition |
| The book of Manchester United | John Henley | 2003-06-18 First edition |
| How to survive in Africa | Kate Nebit | 1991-01-25 First edition |
| Red apple, blue sky | Marko Owen | 2006-12-07 Second edition |
| Love in the mountain | Katja Müller | 2000-05-19 Fourth edition |
| Bash programming guide | John Harker | 2001-11-23 Second edition |
| The 100 best horror films | Jack Ismay | 1995-04-22 Second edition |
| Speak freanch in 1 month | Henry Petit | 1997-03-19 First edition |
| Welcome to the reality | Robert Morel | 2005-10-10 First edition |
| Discrete mathematics | Vera Beltran | 1999-30-05 Second edition |
| Planes and boats | Naomi Michel | 1997-08-03 First edition |
| Second world war image collection | Juan Espada | 2002-03-12 Third edition |
| Discovering Poland | Anja Tomaka | 2003-06-22 Second edition |

# Realized tests

## Precision tests (II)

- Done three different modifications to the source page with different characteristics to:
  - Extract data from formatted text
  - Extract data using styled text (class attribute)
  - Extract data from CSV formatted text

# Realized tests

## Precision tests (III)

- Example: Extracting data from CSV source

| | All the information of the last published edition | Date of the last publication | Year of the last publication | 2 last digits of the year of the last publication |
|---|---|---|---|---|
| Dapper | ✗ | ✗ | ✗ | ✗ |
| Robomaker | ✔ | ✔ | ✔ | ✔ |
| Lixto | ✔ | ✔ | ✔ | ✔ |
| WinTask | ✗ | ✗ | ✗ | ✗ |
| Automation Anywhere | ✗ | ✗ | ✗ | ✗ |
| Web Content Extractor | ✔ | ✔ | ✔ | ✔ |

# Future work

- Given a Web source which features the tool accomplish. Useful to find the most suitable tool
- Testing with non visual GUI tools
- Realize a detailed document that contains all the realized work
- ...

# Thanks for your attention!