How do I get the soccer result as soon as possible?

“A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner.” (Kobayashi and Takeda, 2000)

Evaluation Process

- Tasks
  - Creation of Web Scraper Description
  - Creation of Functional Specification
  - Creation of Benchmark Catalogue
  - Comparison of Existing Web Scraper
  - Validation which Scraper is applicable in which areas
  - Evaluation which Scraper fits the need best
  - Definition of further Research Assignments

- Output
  - Web Scraper Description
  - Functional Specification
  - Benchmark Catalogue
  - Scraper Comparison
  - Benchmark Results
  - Topics for further Research Assignments

Results based on the Evaluation

General Strengths
- Good for gathering lots of information
- Various output formats of the crawled information
- Good support for different forms of navigation
- Quite easy to use
- Good support for proxy servers

General Weaknesses
- No quality assurance
- No intelligent adjustment
- Problems with extracting information from flash pages
- No prevention algorithm against crawler traps
- No support for CAPTCHAs
- Not possible to respect the robot exclusion standard

Comparison of the evaluated Tools

Conclusion and Further Work

General
- Scrapers are good for collecting data – but for nothing else
- Works smoothly as long as the structure of the website does not change
- Minimalization of the available risks by configuring different web sites
- Data processing and quality assurance system needs to be programmed
- Configuration/Maintenance effort vs. benefit?
- No assurance that quality is better than from feed providers

Open Questions
- How often do big sporting websites change their structure?
- How many sports and leagues should be supported?
- Is there a data processing and quality assurance system which could be used

Further Application Areas
- Fixtures/Results
- Names of Players
- Live bet analysis of competitors (coverage, availability,..)
- Odds comparison

Unstructured information can be scraped and saved in an appropriate file format like XML.