Avec le soutien de l'Union européenne et de la Région wallonne

Web Data Extraction with Retroweb

Fabrice Estiévenart (fabrice.estievenart@cetic.be)

Rencontres Mondiales du Logiciel Libre - Bordeaux - 8th July 2010

www.cetic.be

Your connection to ICT research

CETIC, overview

- ICT research centre
- Created in 2001
- Initiatited by 3 universities
- Connection between Industry & Research
- 3 departments, 40 researchers
- Contribution to Regional Economic Development
- International focus – European Research Area

SEVENTH FRAMEWORK PROGRAMME

Ponte

www.cetic.be

CETIC, the ICS team

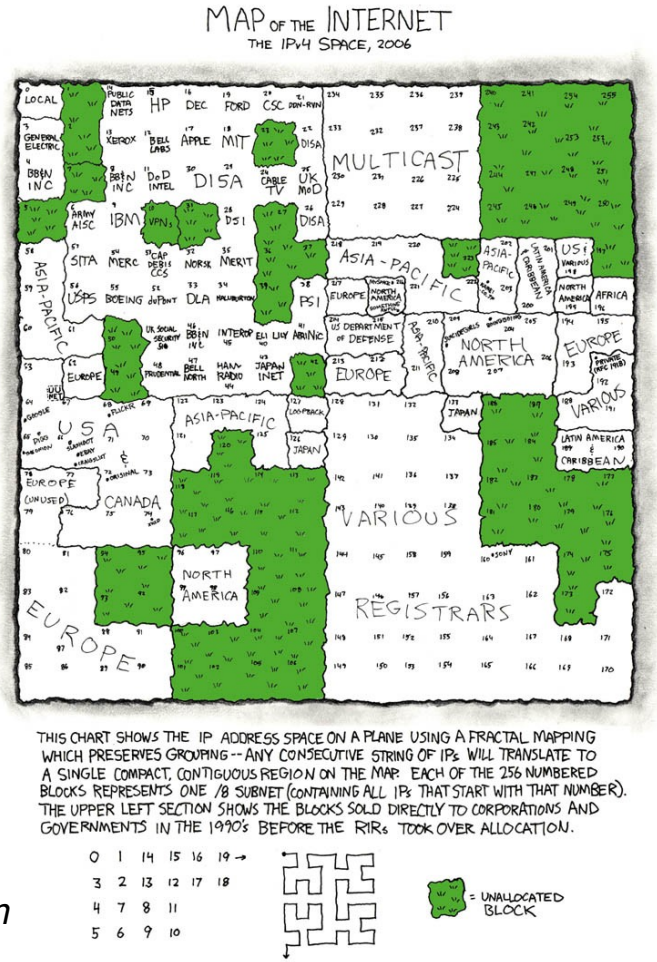
- Knowledge Extraction from Unstructured Content
  - Web wrapping
  - Document clustering
  - Text mining
- Search Engines
  - Crawling
  - Text extraction
  - Analysis / Indexing
  - Search
- Semantic Web
  - Ontology and terminology engineering
  - Micro-formats

www.cetic.be

Context & motivations

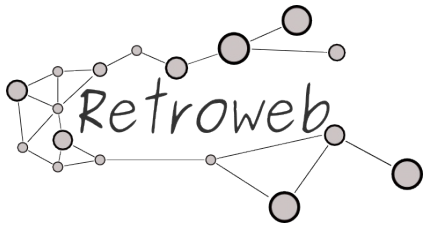
Internet

- Data of interest are hidden within HTML
- HTML has weak semantic
- Web pages are constantly changing



Source : xkcd.com

Biggest challenges in web wrapping

- Collecting/organizing relevant documents
  - Intelligent crawling
  - Web services
- Locating data of interest within documents
  - RegExp
  - Element structure → 

  - External resources

Retroweb, overview

Tool for web data extraction (web wrapping)

Retroweb-GUI: semi-automated definition of extraction rules, visual process

Retroweb-Wrapper: web data extraction from extraction rules

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T.+32 71 490 700 • F.+32 71 490 799 • info@cetic.be

Retroweb, terminology

Web Page

http://www.imdb.com/title/tt0821642

http://www.imdb.com/title/tt0858486

http://www.imdb.com/title/tt0458525

Page Type

imdb-movie

Page Component

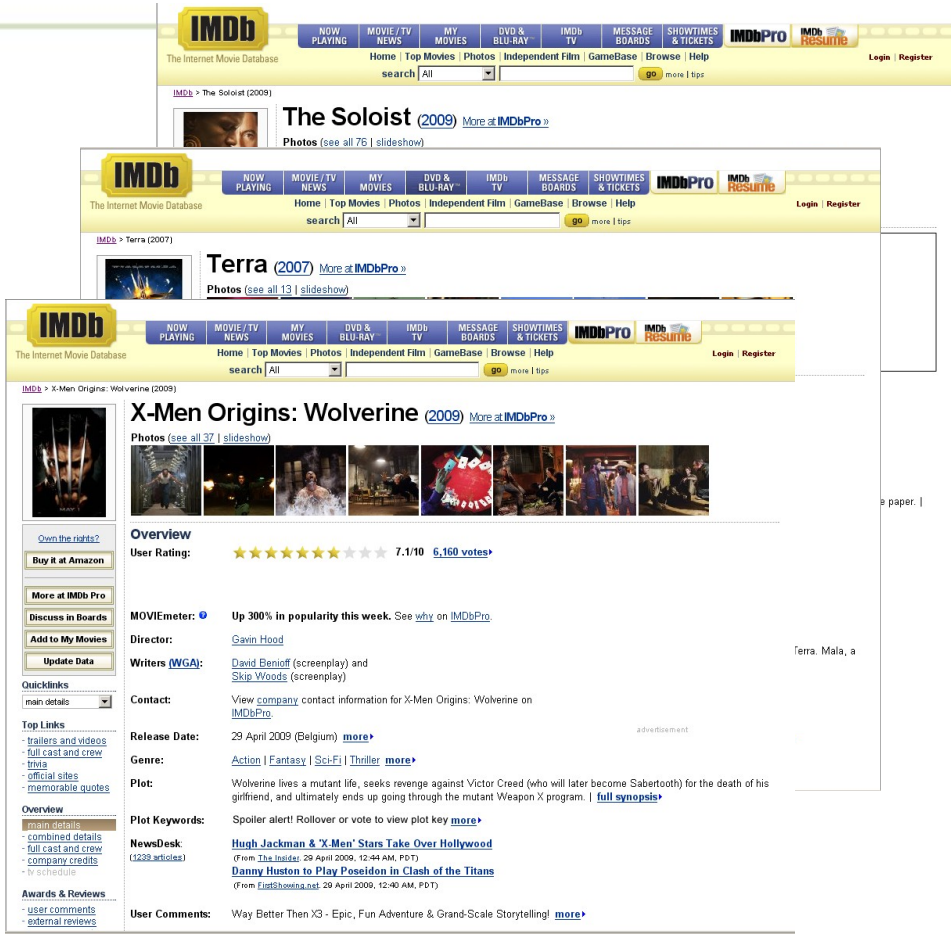
title

tagline

actors

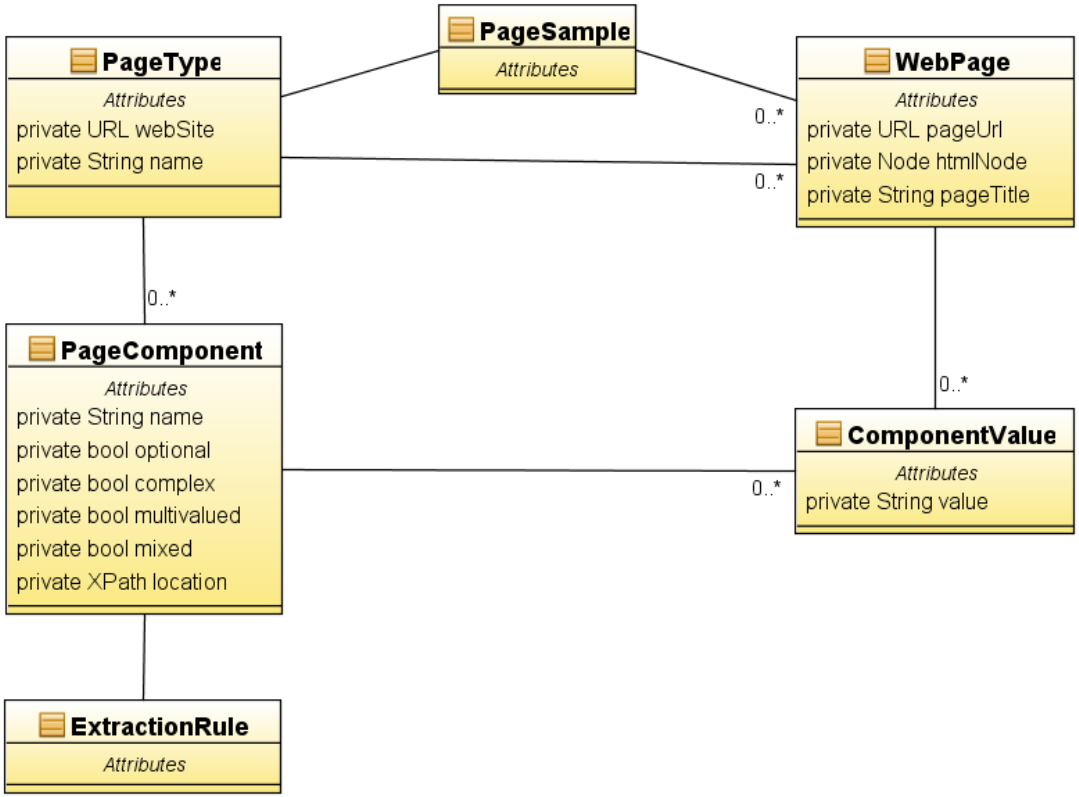
Component Value

The soloist



www.cetic.be

Retroweb, model



The wrapping process

Retroweb GUI

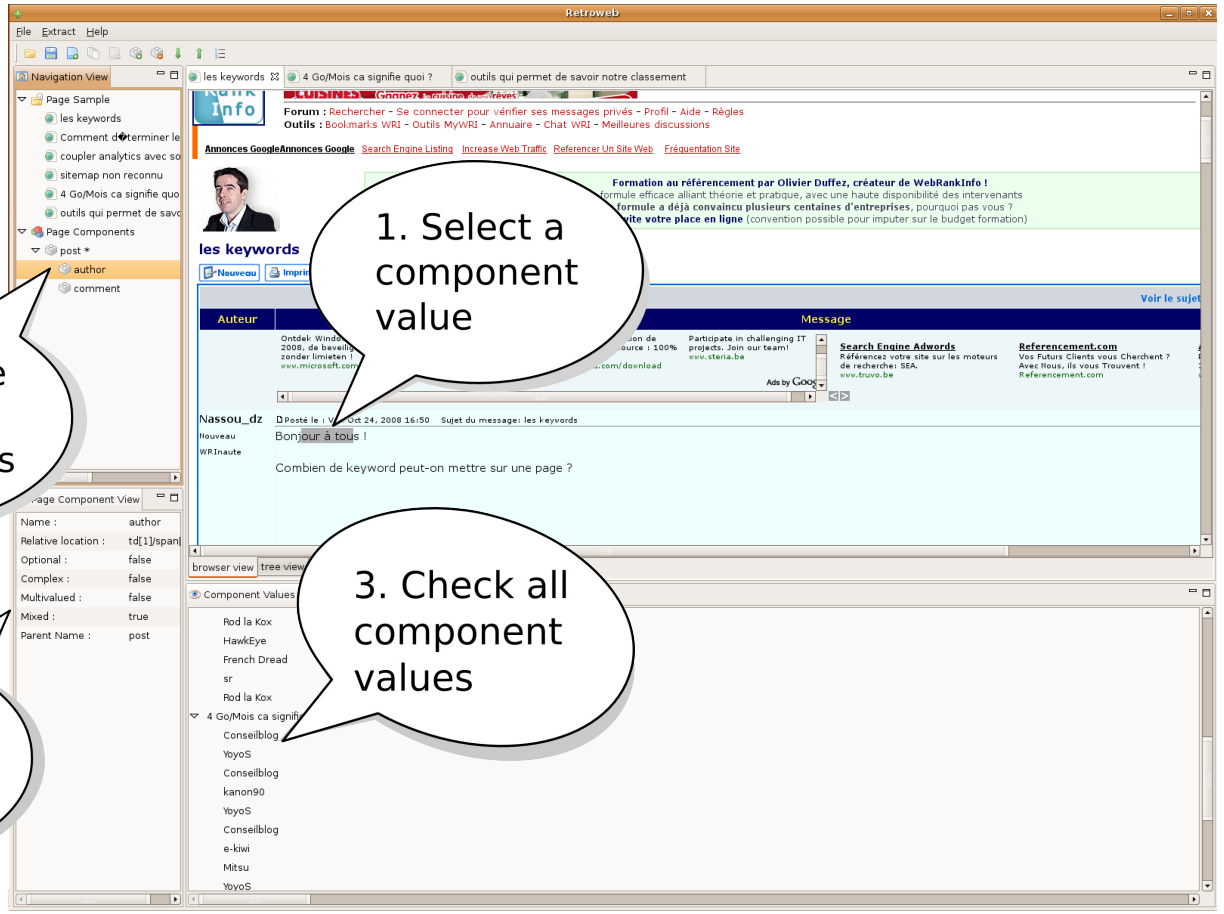
- Select a page sample
- Select values and name page components
- Refine extraction rules
- (Re)structure page components
- Check component values

Retroweb Wrapper

- Extract to XML

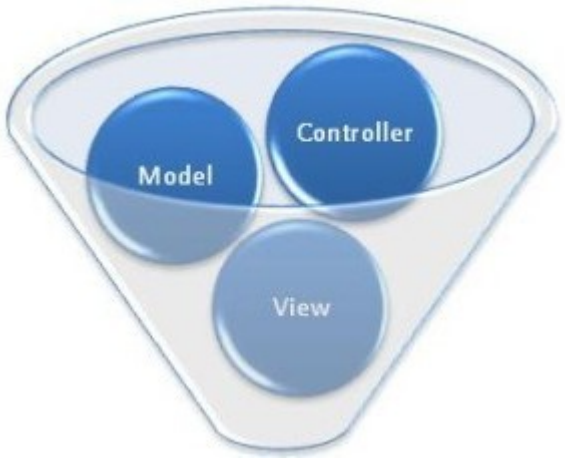
CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T.+32 71 490 700 • F.+32 71 490 799 • info@cetic.be

Screenshot



www.cetic.be

Technologies



www.cetic.be

Retroweb, benefits

Easy: no need to learn a specific language

Flexible: only data of interest are extracted

Robust: extraction rules are defined from a set of pages

Extensible: based on standards (XML, XPath) and open-source (Affero GPL v3)

Portable: GNU/Linux, MS-Windows

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T.+32 71 490 700 • F.+32 71 490 799 • info@cetic.be

Some applications

Web sites reverse engineering

Search engines

Competitive intelligence

Semantic annotation of corpus

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T.+32 71 490 700 • F.+32 71 490 799 • info@cetic.be

Some similar applications

- Lixto Visual Developer
  - Spin-off from the Wien University
  - Services in web intelligence
  - Visual tool based on Eclipse-RCP
- Dapper
  - Free web application for data extraction
  - Not extensible

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T.+32 71 490 700 • F.+32 71 490 799 • info@cetic.be

Wanna be involved ?

- Centre d'Expertise en Logiciel Libre à Vocation Industrielle (CELLAVI)
- https://forge.pallavi.be

CELLAVI

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T.+32 71 490 700 • F.+32 71 490 799 • info@cetic.be

Next steps

- Search engine integration
  - Web pages collection and clustering
  - Semantic indexation
  - Advanced search
- Error detection
  - Rules maintenance
- Semantic import/export
  - Micro-formats (RSS, FOAF, ...)
  - Ontology population

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T.+32 71 490 700 • F.+32 71 490 799 • info@cetic.be

cetic

Avec le soutien de l'Union européenne et de la Région wallonne

Thanks for your attention

Questions?

www.cetic.be

Your connection to ICT research

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be