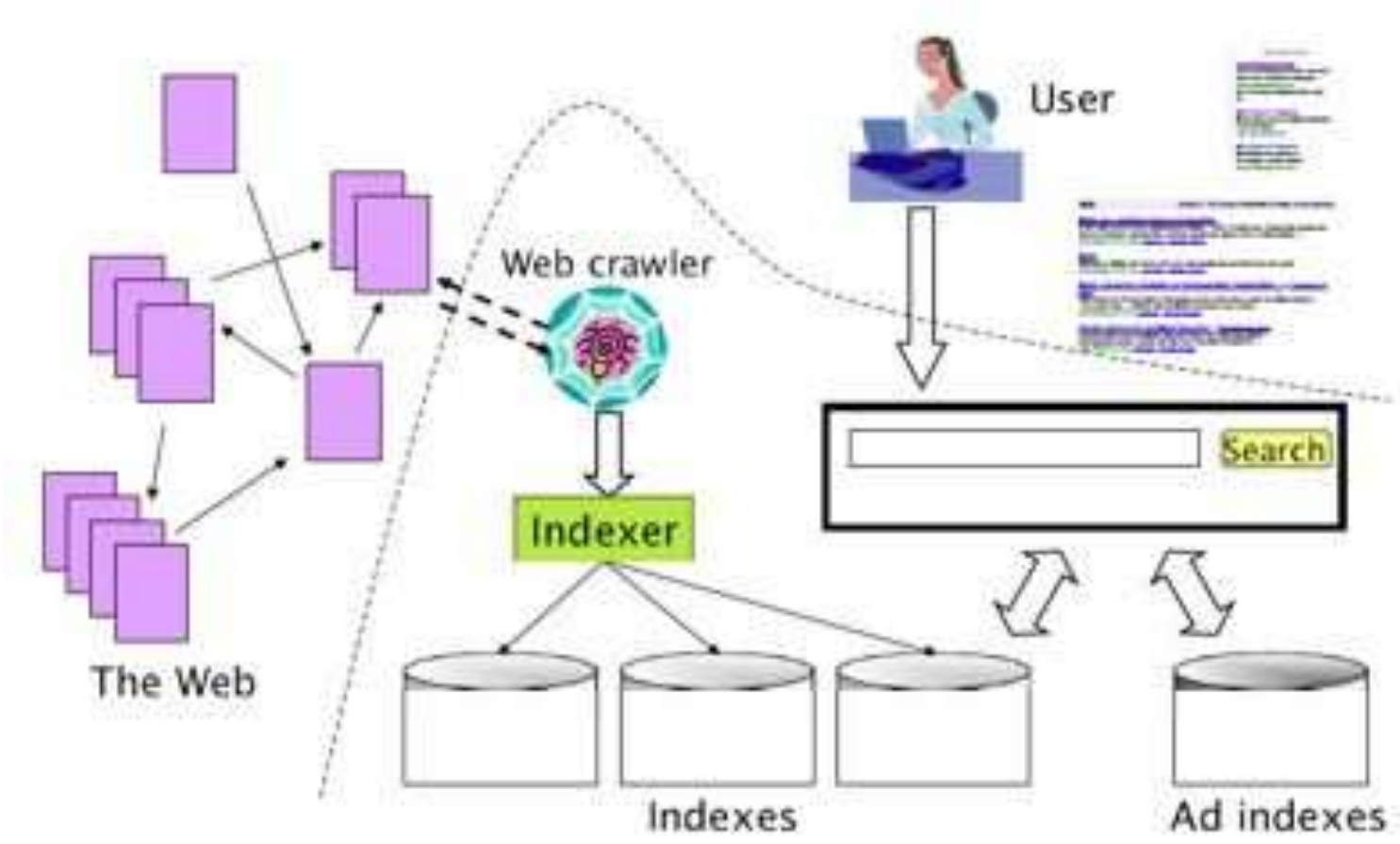
Sistemi informativi per la gestione d'azienda 2013

LABORATORIO
ESERCITAZIONE
WEB CRAWLER 2

Info

- Contacts
  - sandro.labruzzo@isti.cnr.it
  - claudio.atzori@isti.cnr.it
- WEB sites
  - http://www.nmis.isti.cnr.it/casarosa/SIA/

Architecture of a Search Engine



Open Source Crawlers

http://java-source.net/open-source/crawlers

- WebSPHINX - http://www-2.cs.cmu.edu/~rcm/websphinx/
- Jspider - http://j-spider.sourceforge.net/
- WebEater - http://sourceforge.net/projects/webeater
- WebLech - http://weblech.sourceforge.net/
- Arachnid - http://arachnid.sourceforge.net/
- JoBo - http://www.matuschek.net/software/jobo/index.html
- Web-Harvest - http://web-harvest.sourceforge.net/
- Crawler4j - http://code.google.com/p/crawler4j/
- Ex-Crawler - http://ex-crawler.sourceforge.net
- Bixo - http://openbixo.org
- Nutch - http://nutch.apache.org/

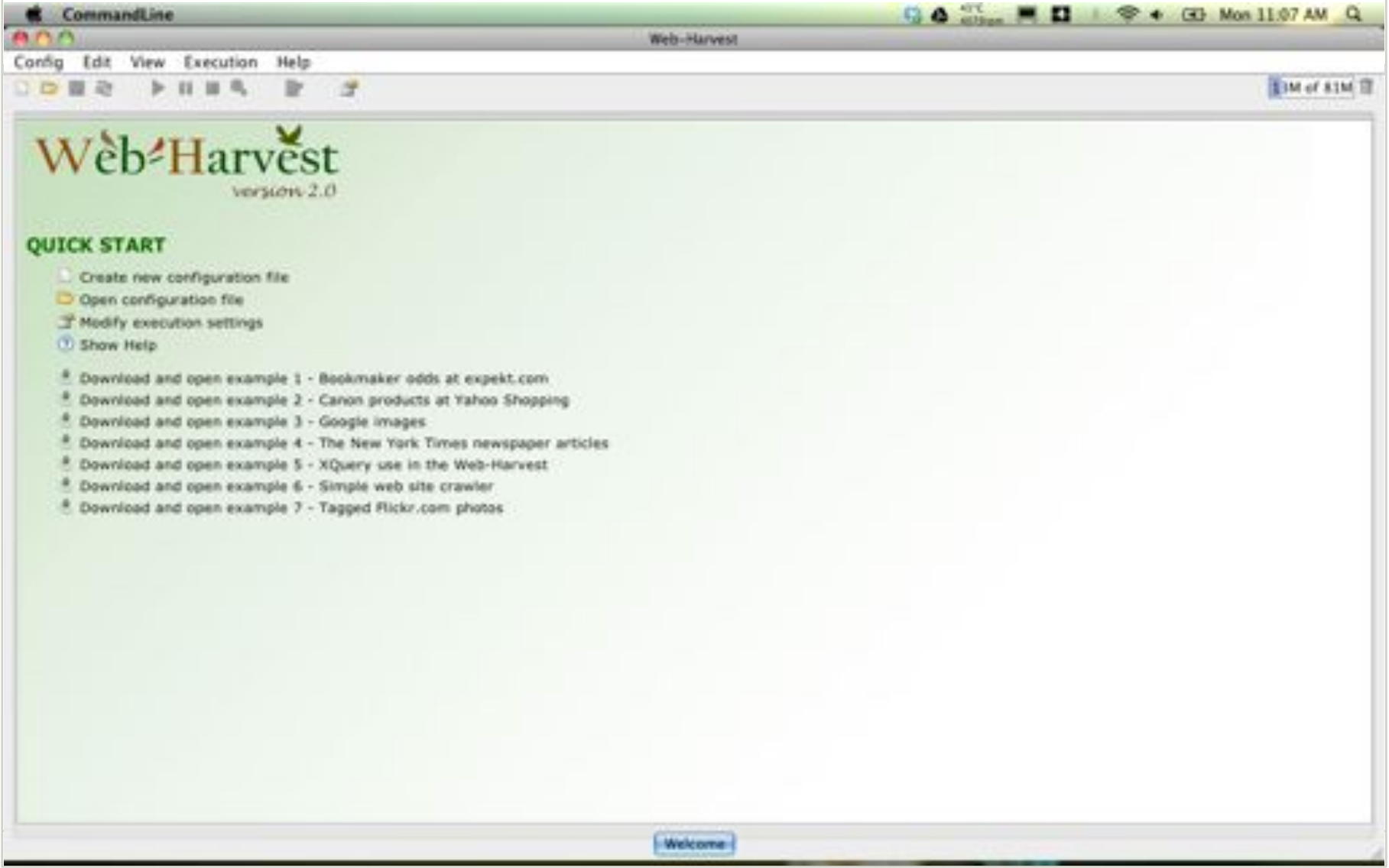
Web-Harvest

- http://web-harvest.sourceforge.net/
- Web-Harvest is an *Open Source Web Data Extraction tool* written in Java.
- Allows to collect desired Web pages and extract useful data from them.
- Leverages on well established techniques and technologies for text/xml manipulation such as
  - *XSLT*
  - *Xquery*
  - *Regular Expressions*
- Process of extracting data from Web pages is also referred as *Web Scraping* or *Web Data Mining*

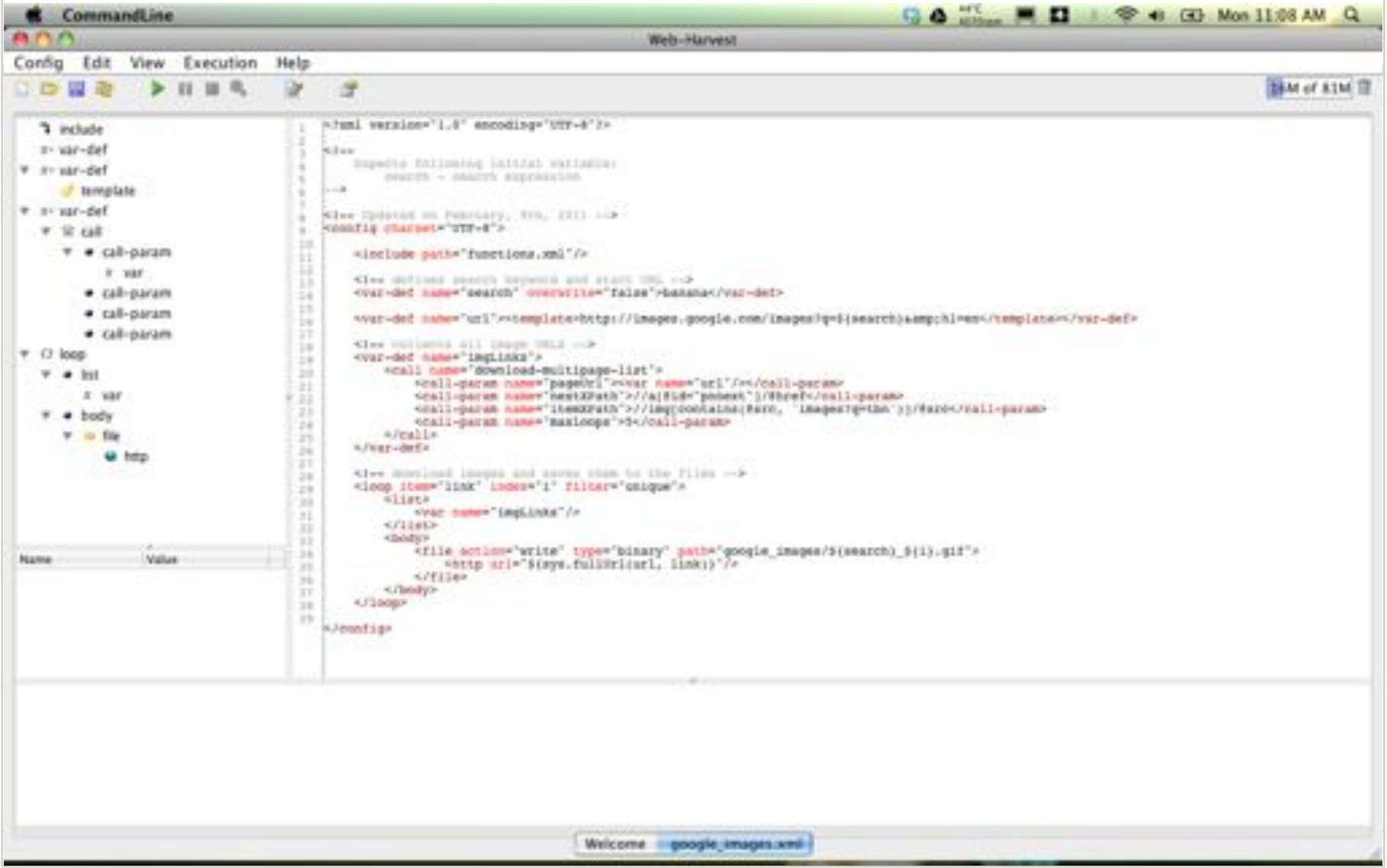
Web-Harvest

- Work defined by “recipes”, no more need to code Java (yay!)
- Recipes are described in a specific configuration language (XML)
- Recipes reference manual
  - http://web-harvest.sourceforge.net/manual.php
- Download URL:
  - http://web-harvest.sourceforge.net/download/webharvest2b1-exe.zip
  - extract
  - Run (java –jar <webHarvest.jar> | double click… | whatever…)

Web-Harvest


CommandLine
Mon 11:07 AM
Web-Harvest
Config Edit View Execution Help
Web-Harvest
version 2.0
QUICK START
Create new configuration file
Open configuration file
Modify execution settings
Show Help
Download and open example 1 - Bookmaker odds at expekt.com
Download and open example 2 - Canon products at Yahoo Shopping
Download and open example 3 - Google images
Download and open example 4 - The New York Times newspaper articles
Download and open example 5 - XQuery use in the Web-Harvest
Download and open example 6 - Simple web site crawler
Download and open example 7 - Tagged Flickr.com photos
Welcome

Web-Harvest


CommandLine
Mon 11:08 AM
Web-Harvest
Config Edit View Execution Help
include
var-def
var-def
template
var-def
call
call-param
var
call-param
call-param
call-param
loop
list
var
body
file
http
Name
Value
<?xml version="1.0" encoding="UTF-8"?>
<config charset="UTF-8">
<include path="functions.xml"/>
<var-def name="search" overwrite="false">banana</var-def>
<var-def name="url"><template>http://images.google.com/images?q=${search}&hl=en</template></var-def>
<var-def name="imgLinks">
<call name="download-multipage-list">
<call-param name="pageUrl"><var name="url"/></call-param>
<call-param name="nextXPath">//a[@id='pnnext']/@href</call-param>
<call-param name="itemXPath">//img[contains(@src, 'images?q=tbn')]/@src</call-param>
<call-param name="maxLoops">5</call-param>
</call>
</var-def>
<loop item="link" index="i" filter="unique">
<list>
<var name="imgLinks"/>
</list>
<body>
<file action="write" type="binary" path="google_images/${search}_${i}.gif">
<http url="${sys.fullUrl(url, link)}"/>
</file>
</body>
</loop>
</config>
Welcome
google_images.xml

Web-Harvest

```
<config charset="UTF-8">
    <var-def name="urlList">
        <xpath expression="//img/@src">
            <html-to-xml>
                <http url="http://news.bbc.co.uk"/>
            </html-to-xml>
        </xpath>
    </var-def>

    <loop item="link" index="i" filter="unique">
        <list>
            <var name="urlList"/>
        </list>
        <body>
            <file action="write" type="binary" path="images/${i}.gif">
                <http url="${sys.fullUrl('http://news.bbc.co.uk', link)}"/>
            </file>
        </body>
    </loop>
</config>
```

Web-Harvest

This configuration contains two pipelines.

```
<config charset="UTF-8">
    <var-def name="urlList">
        <xpath expression="//img/@src">
            <html-to-xml>
                <http url="http://news.bbc.co.uk"/>
            </html-to-xml>
        </xpath>
    </var-def>

    <loop item="link" index="i" filter="unique">
        <list>
            <var name="urlList"/>
        </list>
        <body>
            <file action="write" type="binary" path="images/${i}.gif">
                <http url="${sys.fullUrl('http://news.bbc.co.uk', link)}"/>
            </file>
        </body>
    </loop>
</config>
```

1

2

Web-Harvest

The first pipeline performs the following steps:

- HTML content at *http://news.bbc.co.uk* is downloaded,
- HTML cleaning is performed on downloaded content producing XHTML,
- XPath expression is searched for, giving URL sequence of page images,
- New variable named "urlList" is defined containing sequence of image URLs.

1

```
<var-def name="urlList">
    <xpath expression="//img/@src">
        <html-to-xml>
            <http url="http://news.bbc.co.uk"/>
        </html-to-xml>
    </xpath>
</var-def>
```

Web-Harvest

The second pipeline uses result of the previous execution in order to collect all page images:

- *Loop* processor iterates over URL sequence and for every item:
- Downloads image at current URL,
- Stores the image on the file system.

2

```
<loop item="link" index="i" filter="unique">
    <list>
        <var name="urlList"/>
    </list>
    <body>
        <file action="write" type="binary" path="images/${i}.gif">
            <http url="${sys.fullUrl('http://news.bbc.co.uk', link)}"/>
        </file>
    </body>
</loop>
```

Web-Harvest: task

- foreach example in examples
  UNDERSTAND example
  RUN example
  WRITE description(example)

- function description
  foreach tag in example
    UNDERSTAND tag (http://web-harvest.sourceforge.net/manual.php)
  return **detailed pseudocode** of the example

Email examples.zip to claudio.atzori@isti.cnr.it
[subject: SIA webharvest <name.surname>]

Hint: examples provide some comments … consider integrating yours