

A collaborative environment for web crawling and web data analysis in ENEAGRID

*DATA 2017 - International Conference on Data Science, Technology and Application.
Special session on Knowledge Discovery meets Information Systems: Applications of Big
Data Analytics and BI - methodologies, techniques and tools (KOMIS)*

Madrid, 24-26 July, 2017

Ing. Giovanni Ponti, PhD / DTE-ICT-HPC, ENEA – Italy
**Santomauro G., Ambrosino F., Bracco G., Colavincenzo A. , De Rosa M., Funel A.,
Giammattei D., Guarnieri G., Migliori S.**



1101 0110 1100
0101 0010 1101
0001 0110 1110
1101 0010 1101
1111 1010 0000



Outline

- **The ENEA context**
 - ENEAGRID environment and CRESCO HPC clusters
- **Web Crawling**
 - Intro
 - Integration in ENEAGRID
- **Web Data Analysis**
 - Virtual Laboratory
 - Web Application
- **Tests**
 - Single and Periodic Snapshots
- **Conclusions**
 - Results & Future Developments



1 Headquarters;
9 Research Centers;
5 Laboratories.



Portici Research Center



ENEAGRID & CRESCO HPC Clusters

ENEAGRID

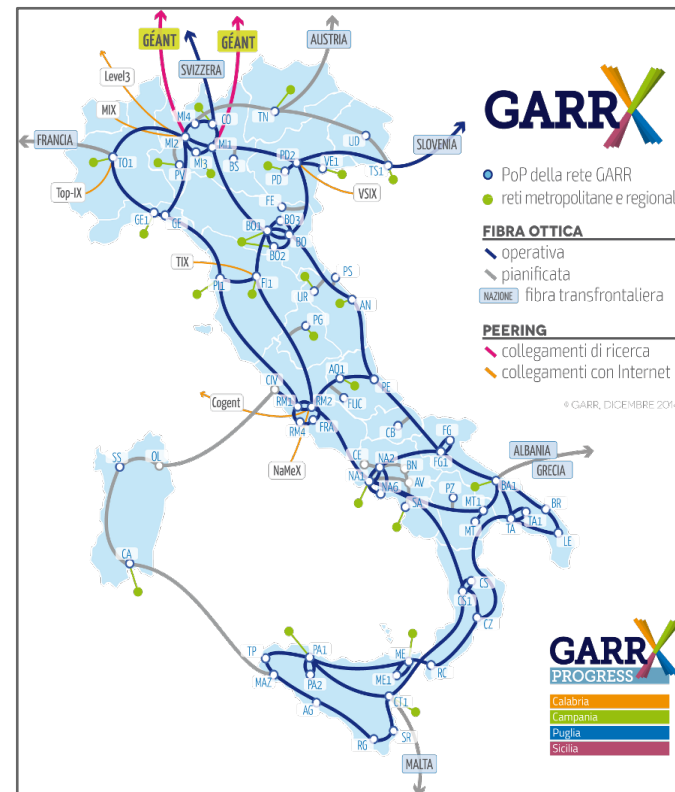
Computation & Storage **ENEA** distributed resources interconnected via **GARR** network.

CRESCO HPC Clusters:

- More than 8000 cores;
- Computing nodes:
 - *Linux x86_64*;
 - *Special systems (GPU, PHI)*;
- Storage resources:
 - *AFS* (distributed);
 - *GPFS* (parallel high-speed) ~1.5PB;
- Cloud computing facilities (*Openstack, VMWare*);
- 6 *Data Center* in ENEA (**Portici** is the main site).



<http://www.cresco.enea.it>



<https://www.garr.it>

Web Crawling

- Activity to browse *www* systematically and download web content;
- All the most popular search engines, such as *Google*, *Bing* and *Yahoo*, need to periodically download the content of a wide web space in order to update their indexes and then provide the most recent results for the querying tasks;
- Data are stored locally and processed to build indexes, statistics and to structure them;
- Web snapshots are typically stored incrementally and can be analyzed to discover changes, new contents and the evolution of the web.

Application contexts:

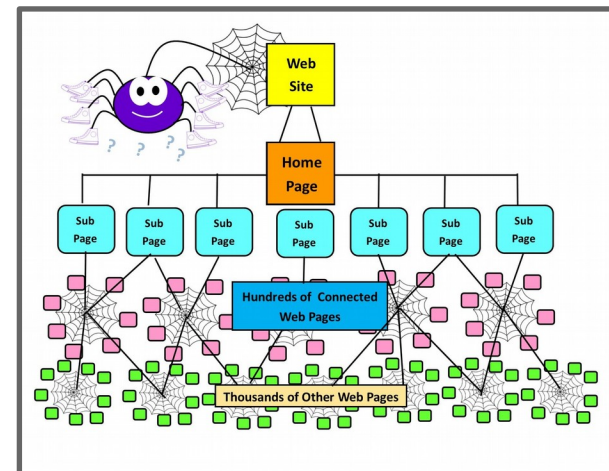
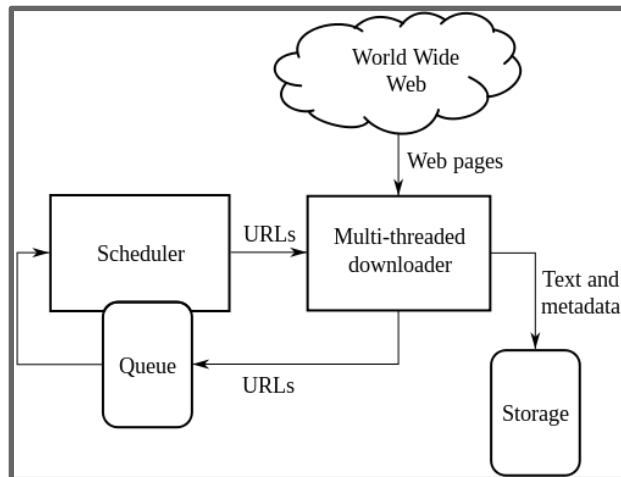
- Web Searching;
- Intelligence & security;
- Blog analysis;
- User behaviors;
- Marketing.





Setting up an environment for web crawling activity

Web Data Retrieving: systematically and automatically exploring a network content in order to create web data archives.



Main tasks

- Installing web crawling tools on **CRESCO Cluster** at *Portici Research Center*;
- Software integration into **ENEAGRID**;
- Definition and confinement of hardware infrastructure;
- Tuning of software parameters.



Problems and Regulations

- Search for best practices to avoid overloading and/or improperly using the network;
- Finding laws that manage the crawling process in order to comply with privacy and/or copyright.



Product Requirements

- Open source solutions;
- Easy integration into the ENEAGRID/CRESCO infrastructure;
- Flexible configuration;
- Tools and interfaces for monitoring;
- Web data storage format;
- Best performance.



BUBiNG

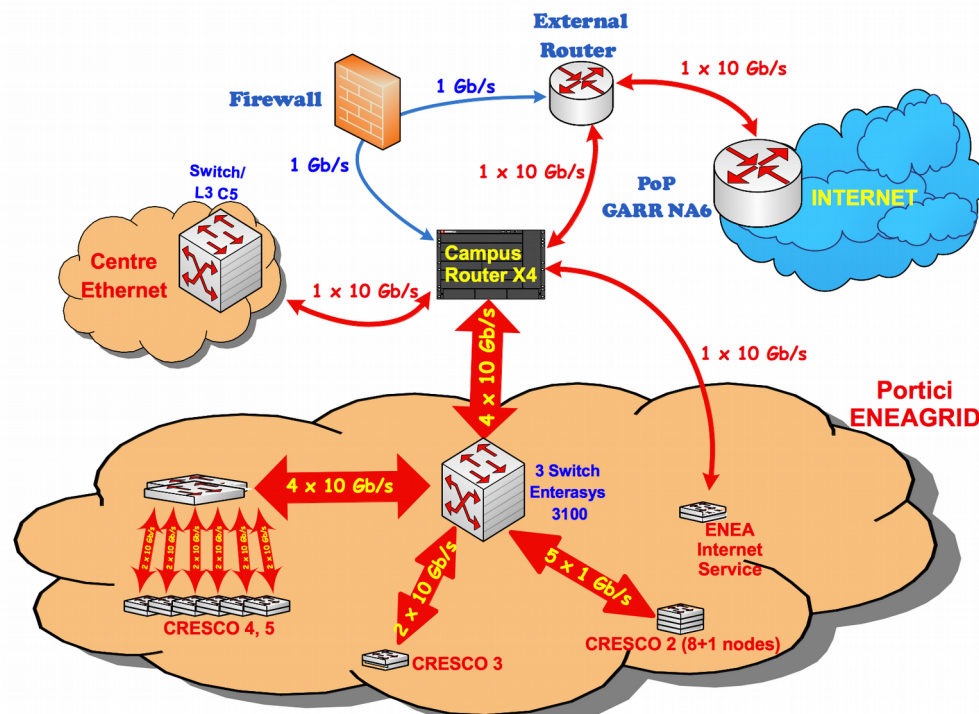
Hardware Solutions: Resources and Network



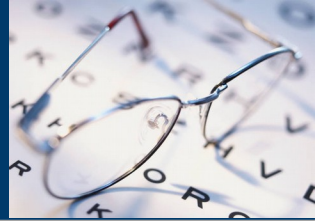
Resources features

- **8 (+1) nodes of CRESCO2:**
Processor:
2 Xeon Quad-Core Clovertown;
RAM:
16 GB;
- **Resources Scheduler:**
LSF 7.0.3;
- **Storage:**
GPFS 4.2.2;
- **Bandwidth:**
1Gbps.

Network schema



Next step



Integration of analysis tools for web content

Web Data Analysis:

- Installing some tools to index, archive and display the data;
- Integrating some instruments for data aggregation and topic detection;
- Investigate on quality of the web data.

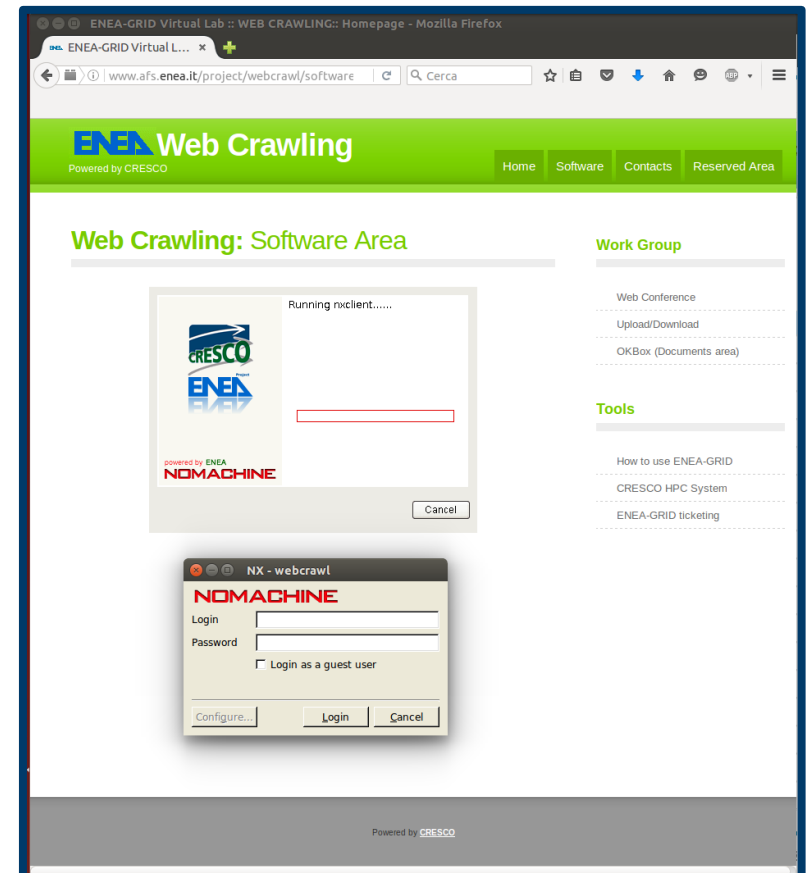
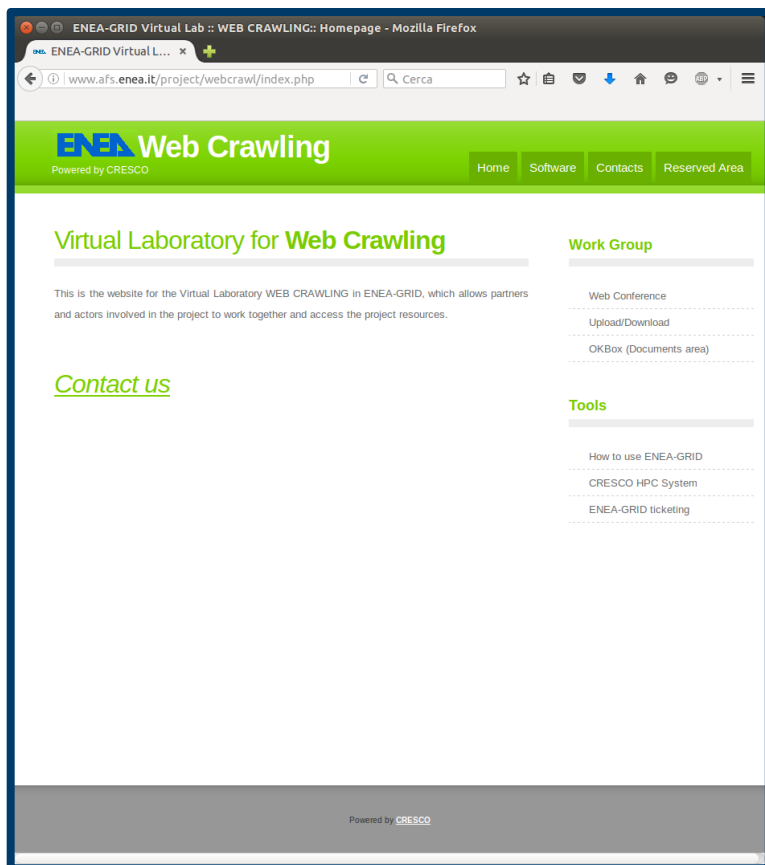
Main tasks

- Creation of **Web Crawling Virtual Lab** for collaborative work in scientific community;
- Development of graphical user interface (**GUI**) for Virtual Lab;
- Integration of *Apache Solr* for **querying** and **data analysis**;
- Integration of *OpenWayback* for data **displaying**;
- Integration of *Carrot²* for data **clustering**;
- Running **single** and **periodic** snapshots.

Virtual Lab



Methodology that allows to remotely use some research services provided by large facilities, which are unique for features and/or costs.

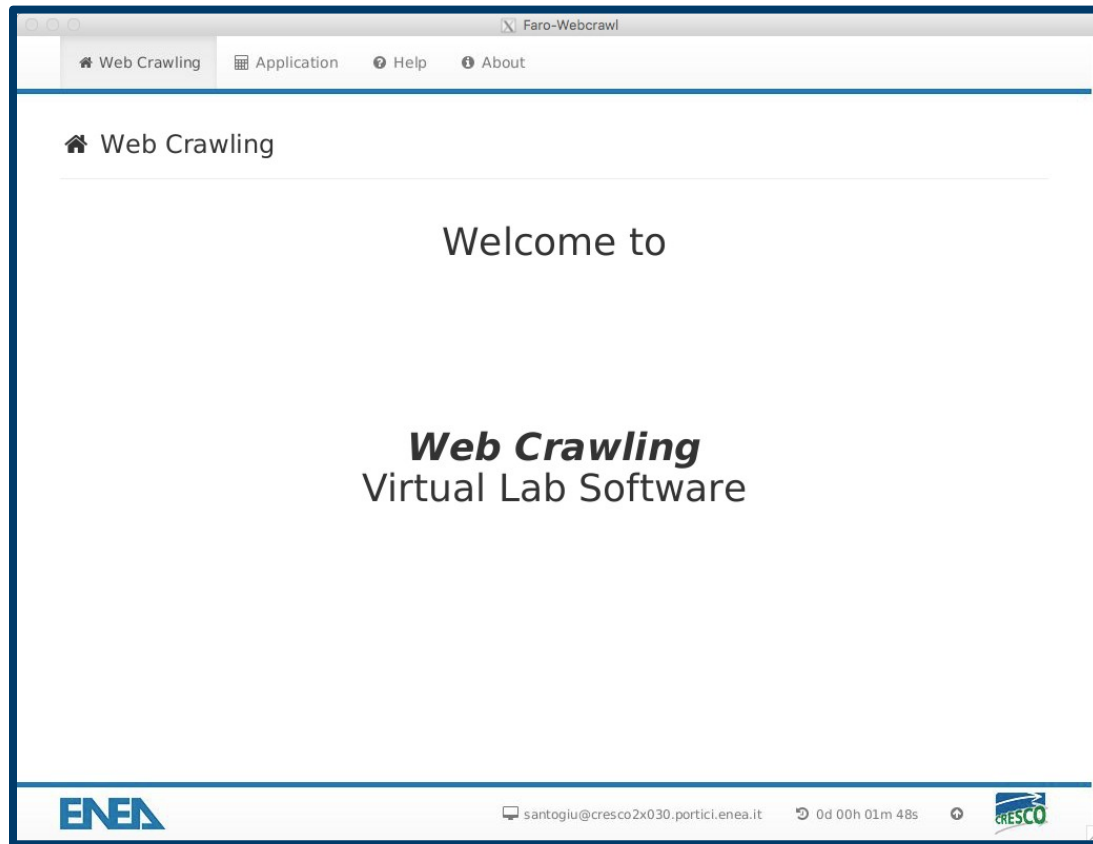


PONTI G. - A collaborative environment for web crawling and web data analysis in ENEAGRID
Madrid – 24-26 July 2017

Graphical User Interface



Web application, developed in *JavaFX*, which allows the user to easily interact with hardware and software resources dedicated to the web crawling project.



Welcome Screen of the application.

GUI: snapshot

The screenshot shows the 'Faro-Webcrawl' application window. The top navigation bar includes 'Web Crawling', 'Application', 'Help', and 'About'. The 'Application' tab is active, displaying a sub-menu with 'Snapshot', 'New Configuration', 'New Initial Seed', 'Monitoring', 'Statistics', and 'Analysis'. The 'Snapshot' sub-tab is selected, showing a form for creating a snapshot. The form is divided into sections: 'General informations' with fields for 'Snapshot Title', 'Comment', 'Running Time', and 'Begin Time'; 'Periodic snapshot' with fields for 'Snapshots Number', 'Frequency', and a dropdown for 'Select frequency type for snapshot'; 'Configuration' with a dropdown for 'Setting'; and 'Initial seed' with a dropdown for 'URL list'. The footer of the application shows the ENEA logo, the email 'santogiu@cresco2x030.portici.enea.it', a timestamp '0d 00h 16m 14s', and the CRESCO logo.

General informations	
Snapshot Title:	Insert title for snapshot set
Comment:	Insert comment for snapshot set
Running Time:	Set time span for snapshot (in second)
Begin Time:	Set begin time for snapshot

Periodic snapshot	
Snapshots Number:	Set number of snapshot sessions
Frequency:	Set frequency number for snapshot
Select frequency type for snapshot	

Configuration	
Setting:	Select configuration for snapshot

Initial seed	
URL list:	Select initial seed for snapshot

Features:

- *Single or periodic* snapshot submission;
- Options:
 - running time;*
 - begin time;*
 - software configuration;*
 - initial seed.*

Tab for snapshot submission.

GUI: new configuration

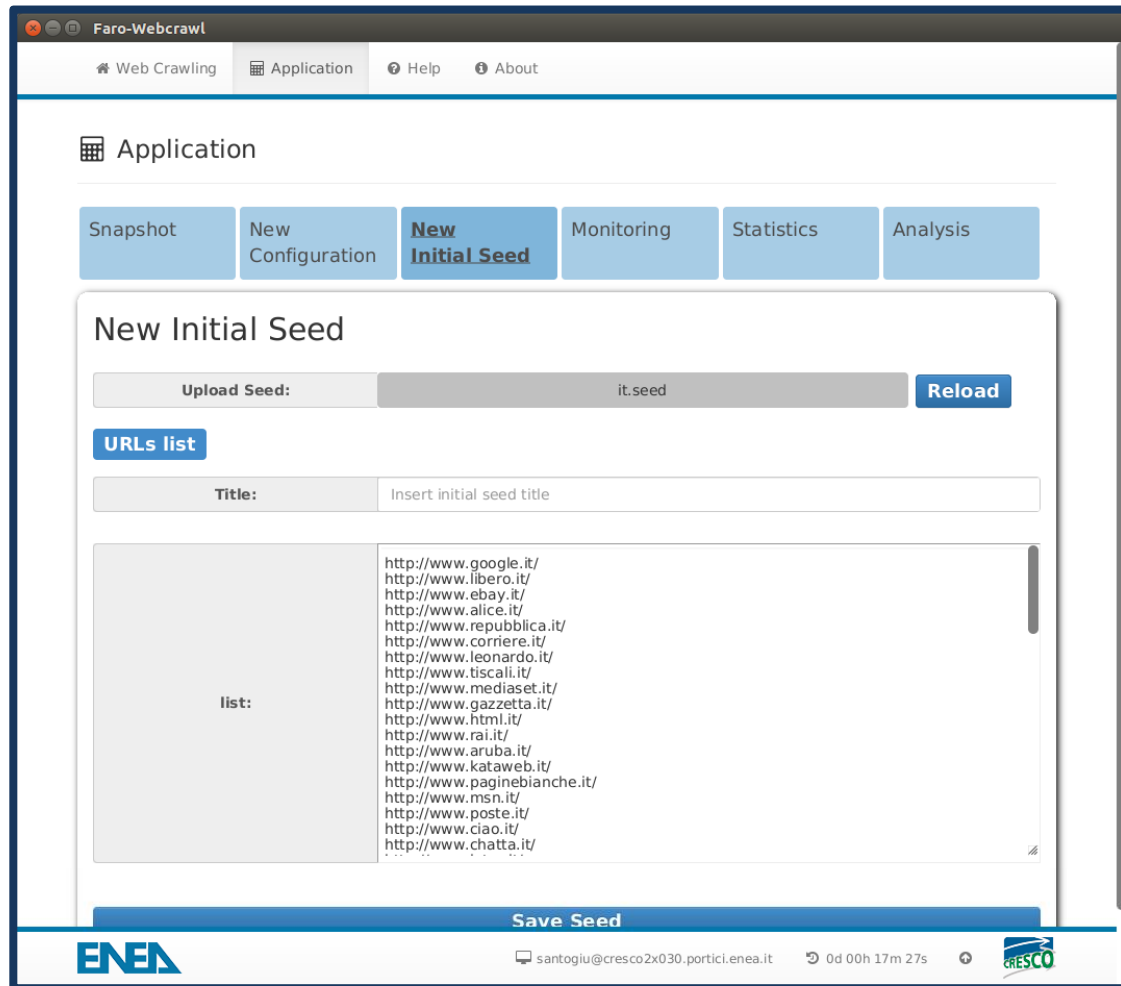
The screenshot shows the 'Faro-Webcrawl' application window. The top navigation bar includes 'Web Crawling', 'Application' (selected), 'Help', and 'About'. Below this, the 'Application' section contains several tabs: 'Snapshot', 'New Configuration' (highlighted), 'New Initial Seed', 'Monitoring', 'Statistics', and 'Analysis'. The 'New Configuration' form is divided into two main sections: 'General Options' and 'Software parameters'. The 'General Options' section includes fields for 'H/W environment' (a dropdown set to 'Select target machine (CRESCO by default)'), 'Title' (placeholder: 'Insert configuration title'), 'Comment' (placeholder: 'Insert comment for configuration'), 'Nodes' (placeholder: 'Set the number of nodes (8 by default, max 8)'), and 'Agents per node' (placeholder: 'Set the number of agents per node (2 by default, max 8)'). The 'Software parameters' section includes a 'Software' dropdown (set to 'Select software (BUBING by default)'), and fields for 'rootDir' (placeholder: 'Insert root directory (/afs/.enea.it/project/webcrawl/data/bubingdata_gpfs/datacrawl/nur)'), 'maxUrlsPerSchemeAuthority' (placeholder: 'Insert max urls per scheme authority'), 'parsingThreads' (placeholder: 'Insert parsing threads'), and 'dnsThreads' (placeholder: 'Insert dns threads'). At the bottom of the window, there is an ENEA logo, a user email 'santoglu@cresco2x030.portici.enea.it', a timestamp '0d 00h 28m 03s', and a CRESCO logo.

Features:

- Creating new settings for the software:
#nodes;
#agents;
#threads;
cache size;
etc.
- Upload stored configurations, modify and re-saved them.

Tab for configuration management.

GUI: new initial seed

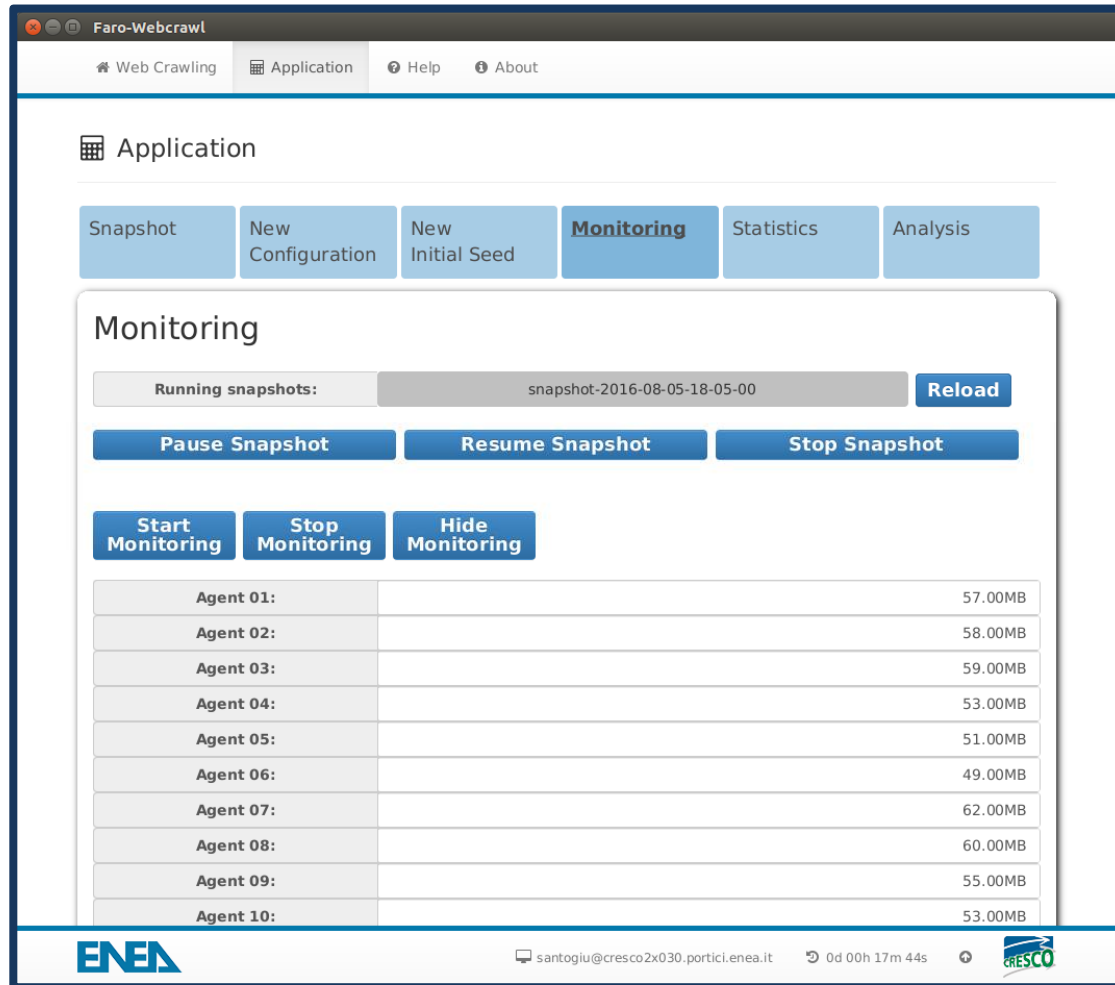


Features:

- Creating new initial URL lists;
- Upload stored initial seeds, modify and re-saved them.

Tab for initial seed management.

GUI: monitoring

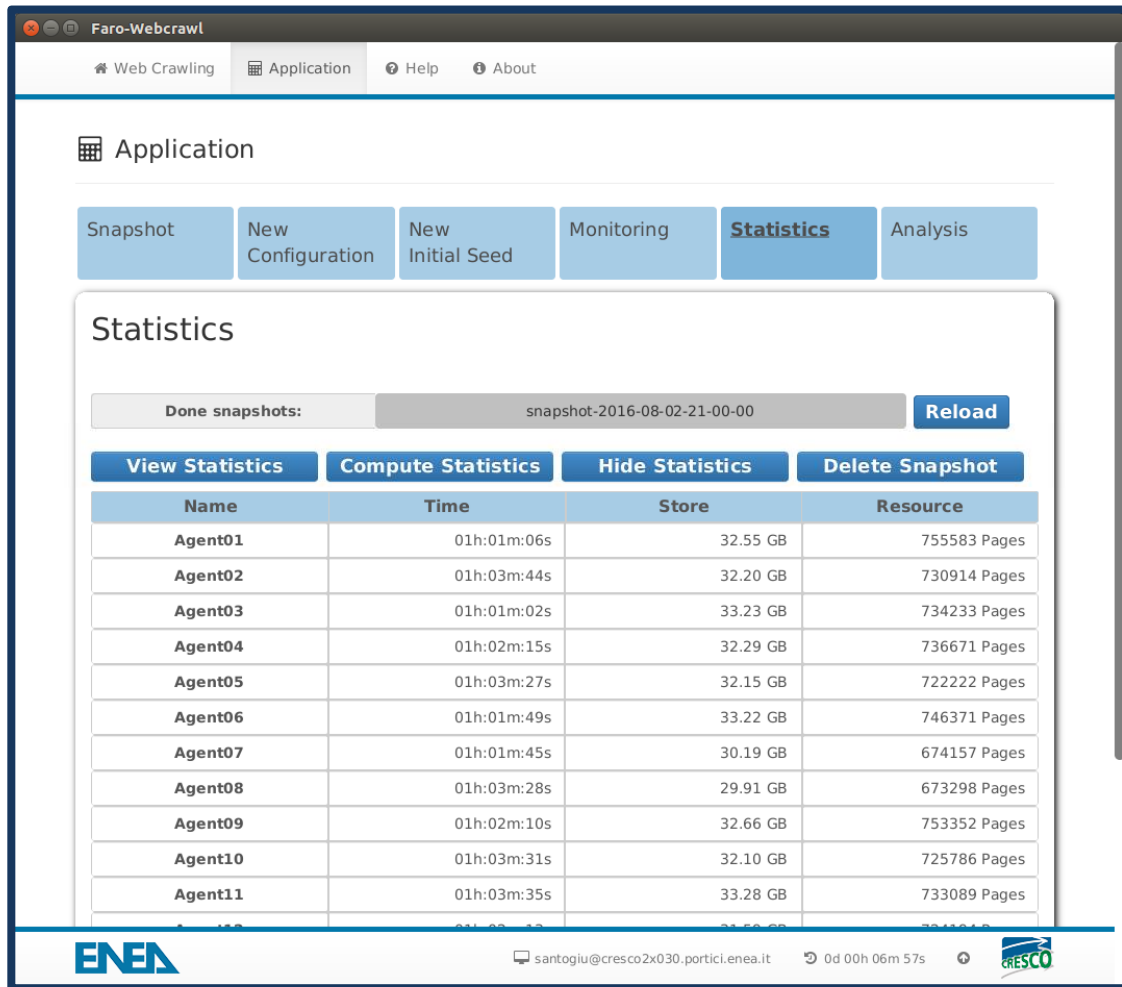


Features:

- Monitoring in real time the downloaded data amount for agent e total;
- Pause, restart and stop a running snapshot.

Tab for snapshot monitoring.

GUI: statistics



The screenshot shows the 'Faro-Webcrawl' application window. The top menu bar includes 'Web Crawling', 'Application', 'Help', and 'About'. The 'Application' tab is active, displaying a navigation bar with buttons for 'Snapshot', 'New Configuration', 'New Initial Seed', 'Monitoring', 'Statistics' (highlighted), and 'Analysis'. Below this, the 'Statistics' section shows a 'Done snapshots:' label with a dropdown menu displaying 'snapshot-2016-08-02-21-00-00' and a 'Reload' button. A row of action buttons includes 'View Statistics', 'Compute Statistics', 'Hide Statistics', and 'Delete Snapshot'. The main area contains a table with the following data:

Name	Time	Store	Resource
Agent01	01h:01m:06s	32.55 GB	755583 Pages
Agent02	01h:03m:44s	32.20 GB	730914 Pages
Agent03	01h:01m:02s	33.23 GB	734233 Pages
Agent04	01h:02m:15s	32.29 GB	736671 Pages
Agent05	01h:03m:27s	32.15 GB	722222 Pages
Agent06	01h:01m:49s	33.22 GB	746371 Pages
Agent07	01h:01m:45s	30.19 GB	674157 Pages
Agent08	01h:03m:28s	29.91 GB	673298 Pages
Agent09	01h:02m:10s	32.66 GB	753352 Pages
Agent10	01h:03m:31s	32.10 GB	725786 Pages
Agent11	01h:03m:35s	33.28 GB	733089 Pages

The footer of the application shows the ENEA logo, the email 'santoglu@cresco2x030.portici.enea.it', a timer '0d 00h 06m 57s', and the CRESCO logo.

Features:

- Display statistics about a terminated snapshot;
- Upload stored snapshot statistics, recompute them, or totally delete a snapshot from archive.

Tab for snapshot statistics displaying.

GUI: analysis, display & clustering

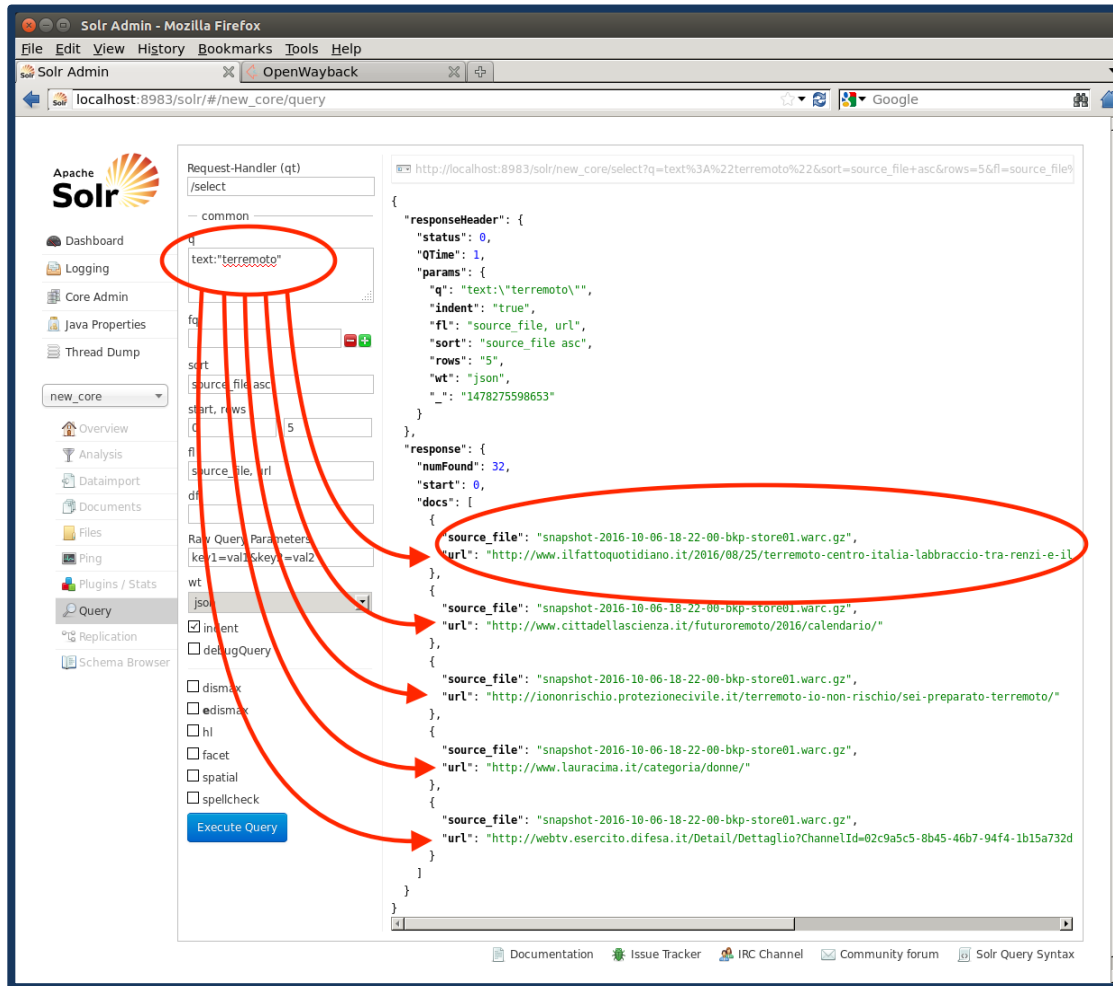


Features:

- Starting the analysis, displaying and clustering tools on terminated snapshot;
- **Solr** (querying);
- **OpenWayback** (display);
- **Carrot²** (clustering).

Tab for web data analysis, displaying and clustering.

Data Analysis: *Solr*



Features:

- Indexing of data collections;
- Text searching in documents;
- Content analysis.

Solr screen.

Data Display: *OpenWayback*

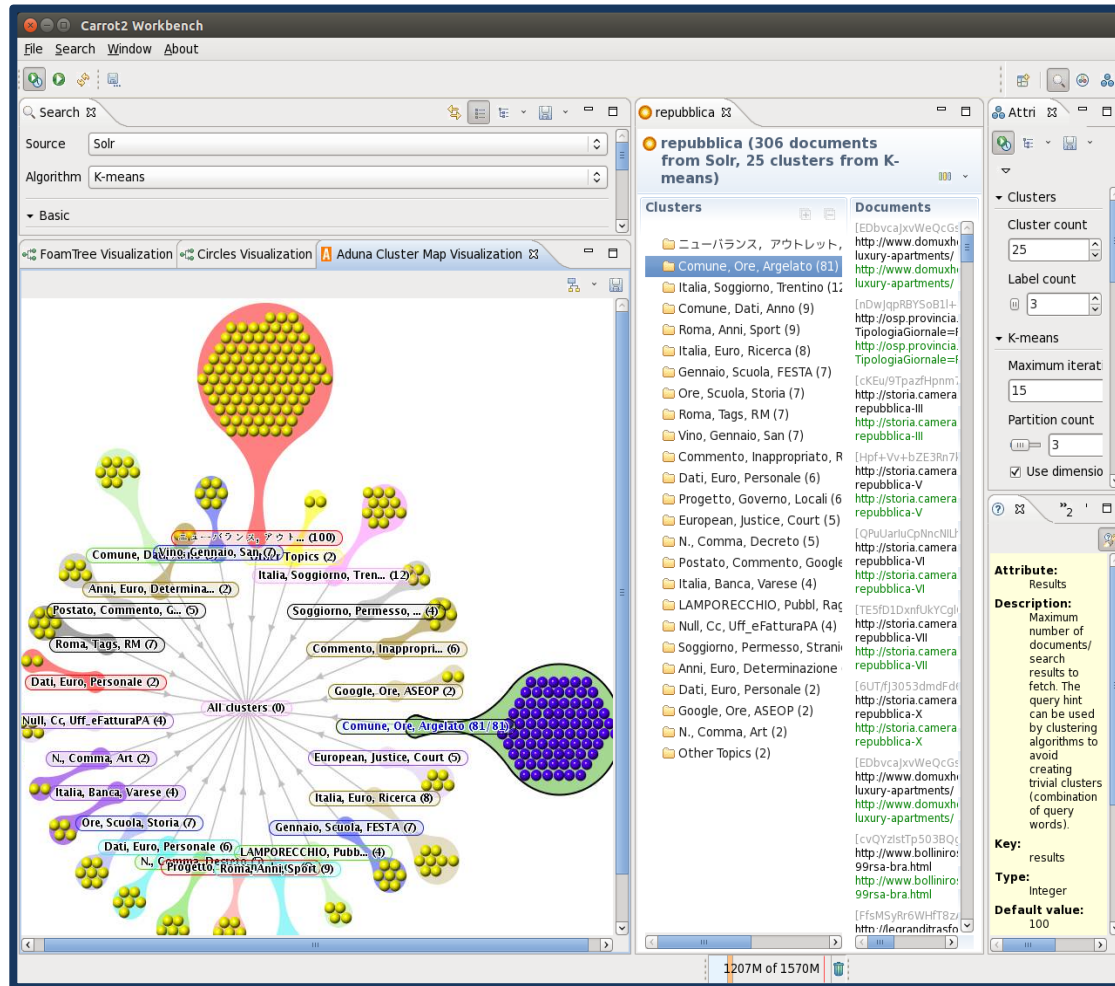


Features:

- Indexing of data collections;
- Documents searching for url and data;
- Web page display.

OpenWayback screen.

Data Clustering: *Carrot*²



Carrot² screen.

Features:

- Clustering of data collections;
- Different clustering techniques (*K-means*, *Lingo*, *STC*, ...);
- Showing results in different forms.

First Test: Single Snapshot (Efficiency and Robustness)



Number of agents: 16;

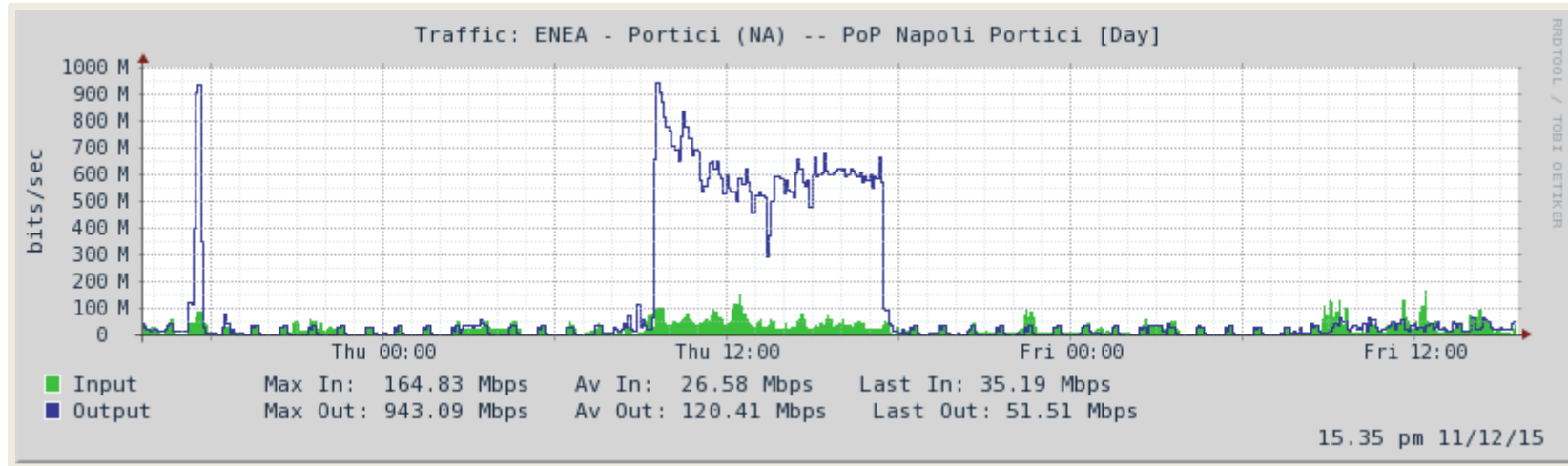
Running time: 8 h;

Amount of downloaded data: 2.94 TB;

Amount of downloaded resources: 66.806.790 Pages;

Data downloading speed: 850 Mbps;

Resources downloading speed: 2305 Pages/Sec.

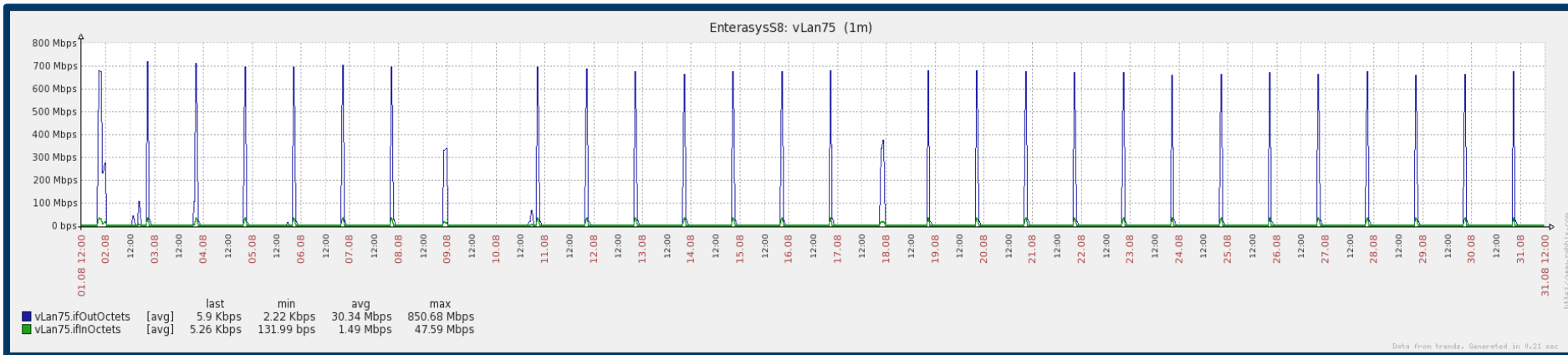


Network traffic measured by GARR on Napoli-Portici PoP.

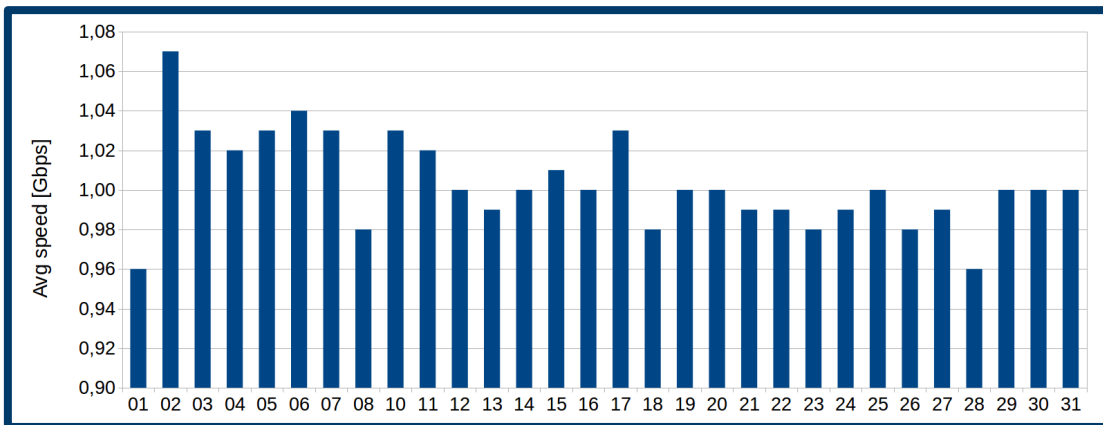
Second Test: Periodic Snapshots (Reliability)



Daily web crawling sessions, during **one month**, each of them kept alive for **one hour** (from 21:00 to 22:00), by considering only web pages from **.it** domain.



Network traffic in the snapshot period.



Average of download speed for each snapshot.

- 15 TB of downloaded data (484 GB/snapshot);
- 3,3 TB saved on the storage (111 GB/snapshot);
- **Downloading Speeds:**
 μ : 1,00 Gbps, σ : 0,0005 Gbps.

Conclusions

Results



- Tools are completely integrated in ENEAGRID infrastructure by the Web Crawling Project.
- The project allows collaborative work, thanks to its Virtual Lab, which provides a web graphical application in order to remotely use all instruments integrated in the infrastructure;
- The Virtual Lab also offers some post-crawling solutions, i.e. for indexing, querying, displaying and clustering the web data;
- The experimental results confirm the high quality of the configuration for the web crawling software from an efficiency, a robustness and a reliability point of view.

Future developments



Improve the indexing process:

- **To parallelize the indexing process** (currently the task is performed by a *single machine* - 4 hours of indexing for each one of crawling);
- **Image and multimedial content indexing** (currently *only text*)



Thanks for the attention

Ing. Giovanni Ponti
giovanni.ponti@enea.it