	Trends Research ENabler for Design Specifications
	FP6-IST-2005-27916
Deliverable	**D 2.1**
Security Classification :	**PP**
Leading partner	**SERAM**
Issue Date	**28/06/06**
Version	**1**
Authors	**Didier Ziakovic, Guillaume Logerot, Jean François Omhover**
Approved by	**Améziane Aoussat, Carole Bouchard**
Date	**04/12/06**

DESIGN AND INNOVATION DATABASE, IMAGES AND WORDS DATABASE Workpackage 2 - Task 2.1

This document is the public report explaining how to create some images databases according to a list of sectors of influence. It also describes the content of the images and words database.

Acronym	TRENDS
List of participants	LCPI SERAM PERTIMM INRIA ROBOTIKER CRF (FIAT) STILE BERTONE UNIVERSITY OF LEEDS UNIVERSITY OF CARDIFF
Coordinator organization	LCPI SERAM : Laboratoire Conception de Produits et Innovation, SOCIETE D'ETUDES ET DE RECHERCHES DE L'ECOLE NATIONALE SUPERIEURE D'ARTS ET METIERS
E-mail contact person	carole.bouchard@paris.ensam.fr
Project Website	www.trendsproject.org
Project Type	STREP (Specific Targeted Research Project)
Contract number	FP6-IST-27916
Start Date	1 January 2006
Duration	36 months

0. INDEX

1. General Introduction

This document presents the deliverable D2.1 *Design and innovation database, images and words database.* In order to explain this procedure, the context of WP2 is described in the following part.

1.1 Work package 2 objectives: design of the system architecture

The objective of WP2 is to develop and validate the methodology and procedures for the Conjoint Trends Analysis (CTA). The work develops along a panel of tasks to set up procedures and techniques required obtaining the sectors of influence as outlined in the following list of sub-tasks. Relevant sources of inspiration and influences will be extracted at this stage, and the corresponding procedure will be explained. The inspirational process is considered as essential for the generation of creative solutions. The CTA method is based on considering that there is an underlying structure in designer information process which can be defined.

The application of the Conjoint Trends Analysis method will allow identifying the key elements and links for the software architecture definition. On the other side, general outlines of the interface design elements will be proposed in this work package in coherence with architecture and the needs analysis results.

To design the TRENDS software architecture, the work package activities will include:

- To elaborate an initial sociological and design trends database.
- To define a procedure for the identification of the sectors of influence.
- To define a procedure for the identification of the websites for the extraction of sociological and design trends.
- To define a procedure for the mono-sectorial mappings realization.
- To define a procedure for the ambience realization.
- To define a procedure for the palettes realization.
- To define a procedure for the statistics realization.
- To define interface graphic design specifications.
- To design the software architecture for TRENDS system software, presenting sociological and styling trends.
- To validate the software architecture with end-users.
- To define the choice of the communications protocols and data transfer functions, protocols and structures.
- To define elements that will be used: computer, processors, programming language.
- To review technologies and artificial techniques to explore and test.
- To define general scheme of possible algorithms to search.
- To define general scheme of algorithm for intelligent web crawling.

1.2 Description of work tasks T21 and T22

This report is related to the first phase with the following tasks of WP2:

- T2.1. Definition of the sectors of influence from an initial sociological and design trends database.
- T2.2. Definition of a procedure for the identification of the websites.

The first phase including T2.1 and T2.2 begins with an investigation about specific lexical information from websites showing the evolution of sociological values. From these results, relevant sectors of influence can be identified and new websites found as sources for the image extraction. One of the deliverables resulting from the tasks T1.1 and T1.2 is the deliverable D2.1 *Design and innovation database, images and words database,* which is presented in this report.

1.3 STRUCTURE OF THE REPORT

This document describes how to grab pages on the web and how to create design and images database. The goal of this grabbing is to set up a demonstration database: a sufficient amount of images, pages and texts that will enable us to lead the experiments that will follow. This full size database should be a good demonstrator of what the TRENDS database will be.

The first part of this document describes the software used to grab the demonstration database.

The second part shows the content of the demonstration database built from a selection of various websites. These websites were previously selected by design experts (see deliverable D2.2), and used in the grabbing phase as a list of URLs (web addresses).

The third part addresses the full size test images database grabbing. These databases are very important to identify problems and evaluate the solution to future image and text search engines and define the intelligent agent of work package 5. The quality of images could be evaluated by users.

2. PROCEDURE FOR WEBSITES AND IMAGES GRABBING FOR DEMONSTRATION DATABASE

2.1 INTRODUCTION

In order to gather the content of the database from the Internet, we used a software called “a crawler”. This software explores a list of websites that have been pre-selected (in our case, by the experts). Then, it makes a local copy of the sites: “.html” pages and images are copied on the computer’s hard drive. Such an operation is called “crawling” (or “grabbing”). The final result of this operation is called a *mirror*, i.e. a local copy of the websites that is stored on the hard drive and can be used anytime, even when the computer is not connected to the Internet.

This is a difficult task because web servers commonly use different scripted languages (PHP, ASP) and different communication protocols (http, https, ftp…) to carry and display the web pages to the user; by recomposing the pages, the crawler can restore the pages *as if* they were static pages – pure “.html” pages that can be directly displayed. This set of grabbed “;html” pages, and their attached images and texts, are to be stored in our database.

The crawler used for the demonstration database is called “HTTRACK” (Windows version).

2.2 BASIC PROCEDURE FOR THE DATABASE ELABORATION

The basic procedure for the database elaboration is explained in figure 1 below.

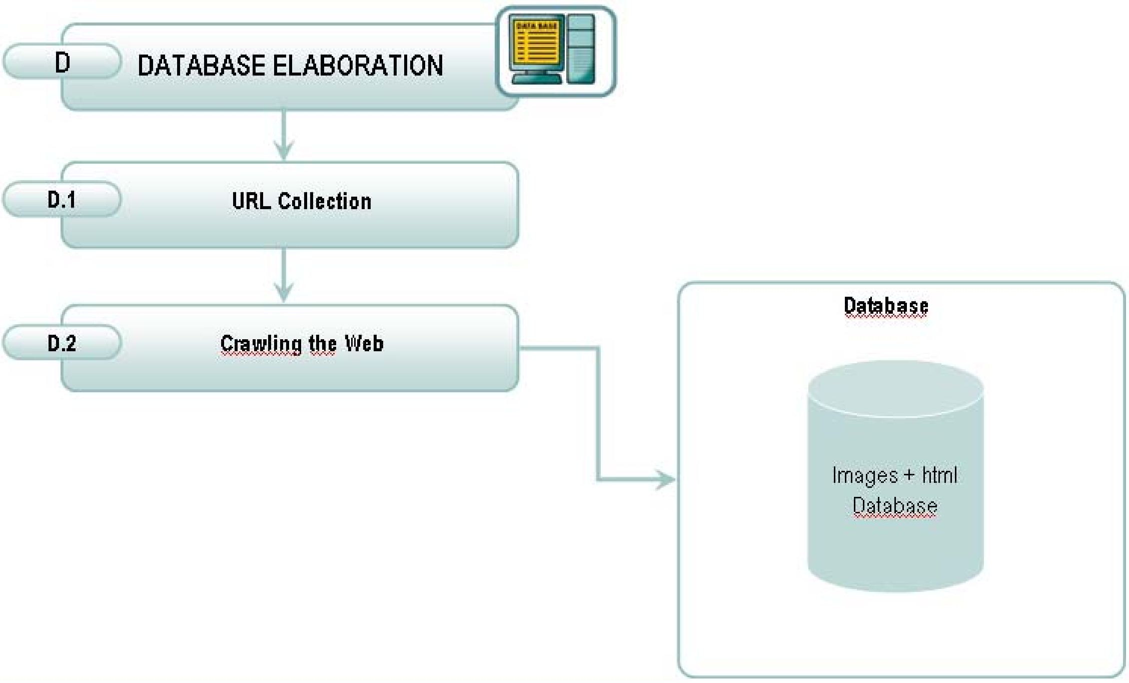


Figure 1: Database Elaboration Procedure

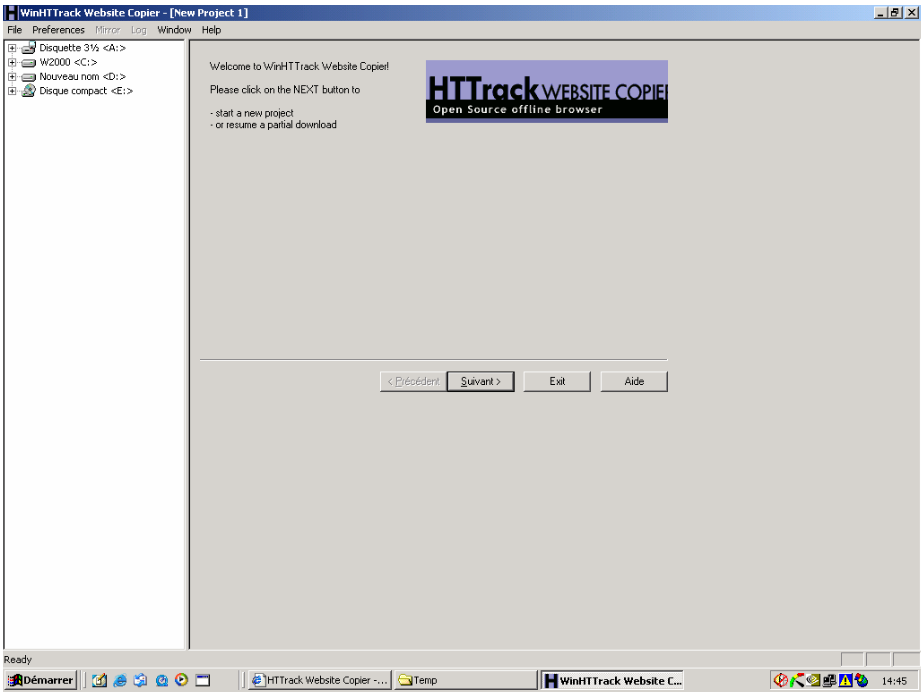
2.2.1 Subtask D1 : URL collection

Before crawling the web, we need to set up a list of web addresses .This list has been elaborated by experts of each domain (see deliverable D2.2) and is segmented by sectors of influence detected by them.

2.2.2 Subtask D2 : Configuration Of The Crawler

The resulting file can be used as an input in the starting addresses window of HTTRACK: these starting addresses will be used by HTTRACK as entry points for the grabbing .

D2.1 Click <Suivant> button



Welcome to WinHTTrack Website Copier!

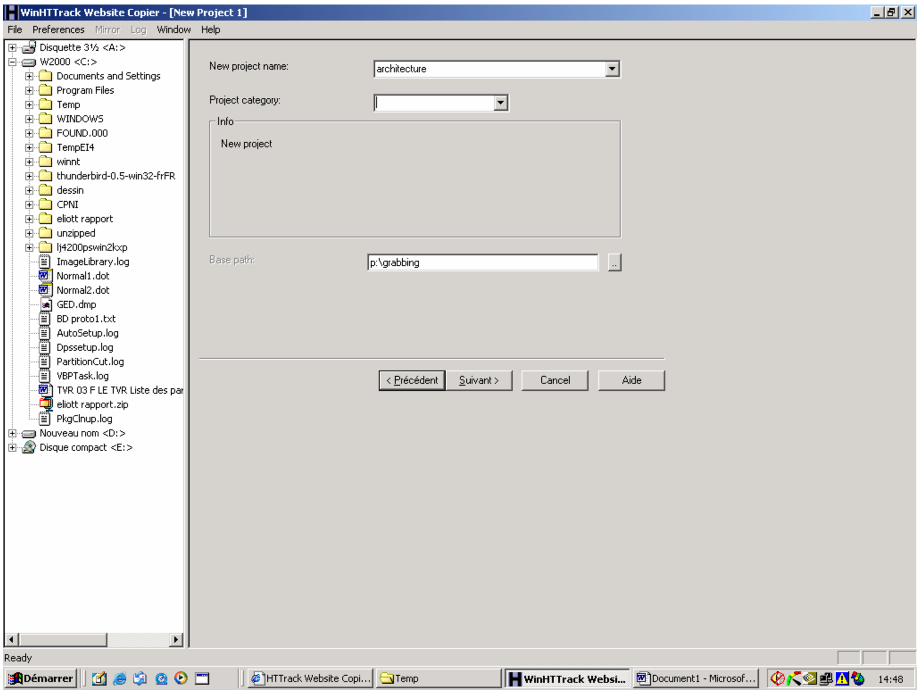
Please click on the NEXT button to

- start a new project
- or resume a partial download

Figure 2: First window of HTTRACK interface

D 2.2 Give a name to your project (*architecture* corresponding to the sector of influence studied for example). The software uses this name to create working folders and create a project where you can store all parameters of your search and use them later. HTTRACK can update the data after a first crawling. It can copy new web pages and modified web pages.

D 2.3 You can select the working folder *where all data will be put (example: p:\grabbing)*



New project name: architecture

Project category:

Info

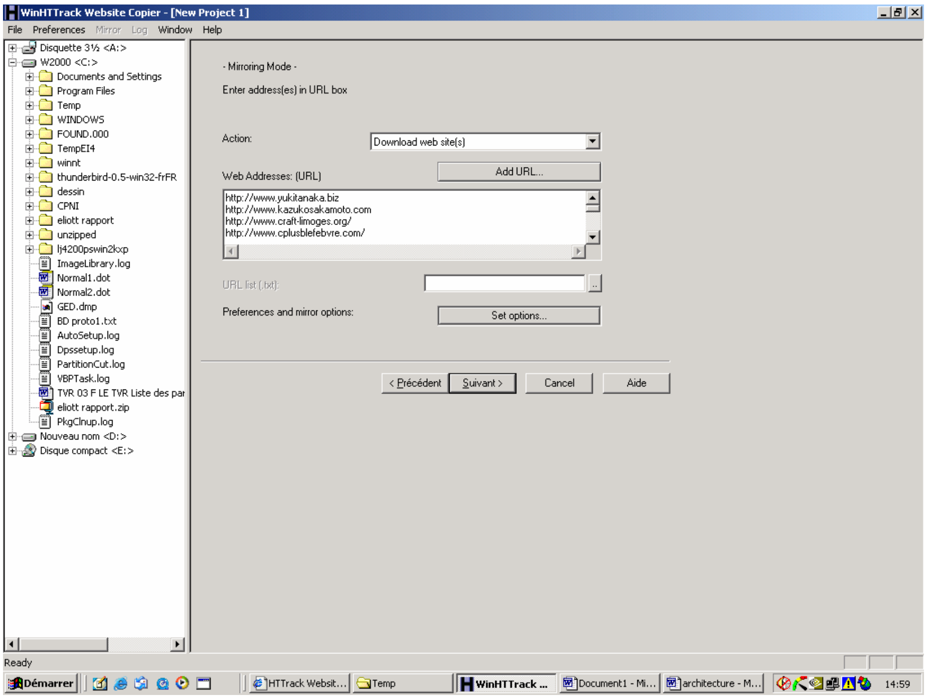
New project

Base path: p:\grabbing

Figure 3: Defining project HTTRACK interface

D 2.4 Click “Suivant*” button and paste the starting address list into the “Web Adresses (URL)” window

*nb: there is a bug in the translation of certain buttons of HTTRACK   Précédent=before, Suivant=Next, aide = help



- Mirroring Mode -

Enter address(es) in URL box

Action: Download web site(s)

Web Addresses: (URL)

http://www.yukitanaka.biz
http://www.kazukosakamoto.com
http://www.craft-limoges.org/
http://www.cplusblefebvre.com/

URL list (.txt):

Preferences and mirror options:

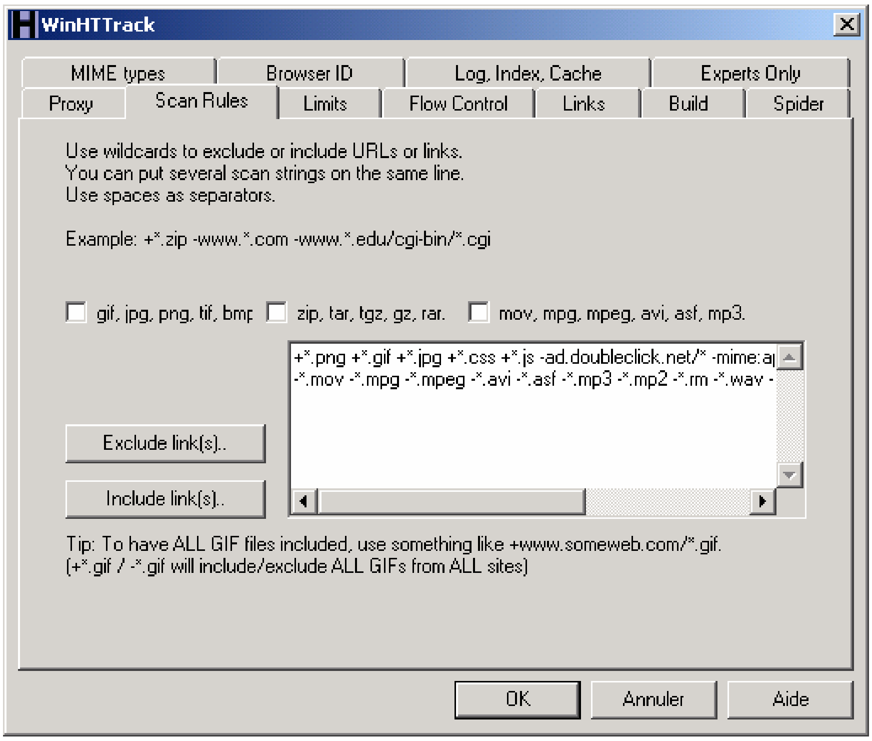
Figure 4: Starting addresses HTTRACK interface

D 2.5 Click “Set options” button

The various functions of options thumbnails are:

- **“Spider”**: this function allows to accept or refuse cookies, explore or not the java file, and respect or not the eventual robot exclusion protocol of the websites.
- **“MIME types”**: this function allows to create correspondences between files extensions and MIME type.
- “**Browser ID**”: this function allows to be identified by website administration with the type of browser used.
- “**Log, index, cache**”: this function allows to create log files, index files and to use or not cache files.
- **“Experts Only”:** this function allows to modify the rule of filtering exploration, and to activate a debug mode.
- “**Proxy**”: this function allows to define a proxy server.
- “**Scan Rules**”: this function allows to implement and modify filtering rules on files extensions (.jpeg,.gif,.mov,.mpeg,.zip,.tar…).
- “**Limits**”: this function allows to define limits in the internal and external depth of exploration, the size of files and websites, number of hyperlinks…
- “**Flow Control”**: this function allows to define the number of simultaneous connexions, the time out waiting, and the minimum band with.
- “**Links**”: this function allows to manage hyperlinks.
- “**Build**”: this function allows to define structure of the data, mirror of websites, or separate global folders for “;html” and images files.

*D2.6 Select "**Scan Rules**" thumbnail and select "zip, tar…" and "mov, mpg, mpeg…" cases, replace "+" sign by "-" to exclude this type of format. (as shown in the example below):*



Use wildcards to exclude or include URLs or links.
You can put several scan strings on the same line.
Use spaces as separators.

Example: +*.zip -www.*.com -www.*.edu/cgi-bin/*.cgi

☐ gif, jpg, png, tif, bmp ☐ zip, tar, tgz, gz, rar. ☐ mov, mpg, mpeg, avi, asf, mp3.

```
+*.png +*.gif +*.jpg +*.css +*.js -ad.doubleclick.net/* -mime:a|
-*.mov -*.mpg -*.mpeg -*.avi -*.asf -*.mp3 -*.mp2 -*.rm -*.wav -
```

Tip: To have ALL GIF files included, use something like +www.someweb.com/*.gif.
(+*.gif / -*.gif will include/exclude ALL GIFs from ALL sites)

Figure 5: Inclusion / exclusion files HTTRACK interface

*D 2.7 Click "Limits" thumbnail, and select "0" for the line "**Maximum external depth**"*

This command avoid that the crawler explores external links according to the starting address list*. Select 50 000 for the line "**Max transfer rate**", this command tune the band with to 50KB/s and 10 for the line ". **Max connections/seconds**",* this command allows HTTRACK to use 10 tasks at the same time.

The other fields are:

"**Maximum mirroring depth**": this parameter defines the number of links between the starting address and the furthest page grabbed. When this window is empty, the depth of crawling has no limit.

"**Max size of any HTML file (B)** ": this parameter defines the maximum size (in Bytes) of html collected files.

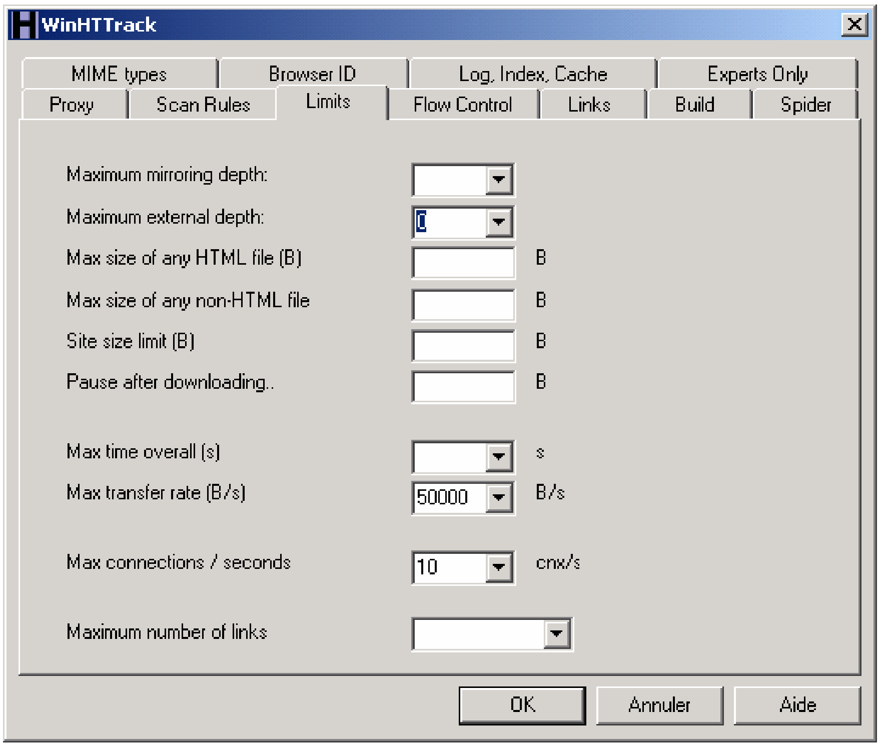
"**Max size of any non HTML file**": this parameter defines the maximum size (in Bytes) of other formats collected files.

"**Site size limit**": this parameter defines the maximum size (in Bytes) of an entire website.

"**Pause after downloading**": this parameter defines the maximum size (in Bytes) of an entire project.

"**Max time overall**": this parameter defines the maximum time of crawling (in seconds).

"**Maximum number of links**": this parameter defines the total number of links explored.



Maximum mirroring depth:

Maximum external depth: 0

Max size of any HTML file (B) B

Max size of any non-HTML file B

Site size limit (B) B

Pause after downloading.. B

Max time overall (s) s

Max transfer rate (B/s) 50000 B/s

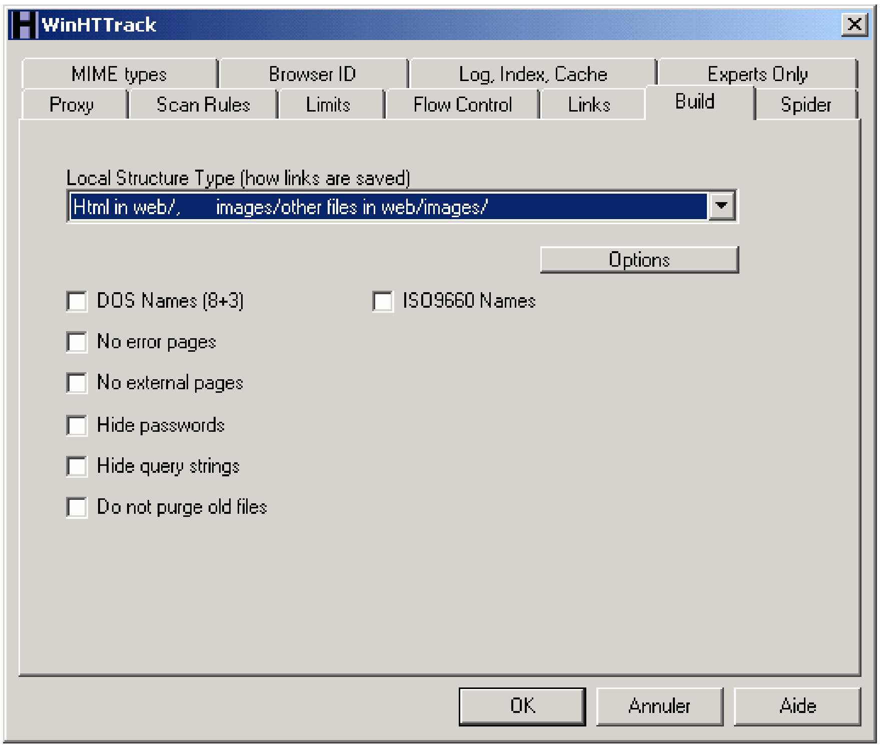
Max connections / seconds 10 cnx/s

Maximum number of links

Figure 6: Limits HTTRACK interface

D 2.8 Click "Structure" thumbnail and select "Html in web/, images/other files in web/images" line:, this action creates two separate folders for html and images. It could be useful to separate images from html files to get all images in the same folder in order to facilitate browsing.

This function allows to easily import or export the data.



Local Structure Type (how links are saved)

Html in web/, images/other files in web/images/

□ DOS Names (8+3) □ ISO9660 Names

□ No error pages

□ No external pages

□ Hide passwords

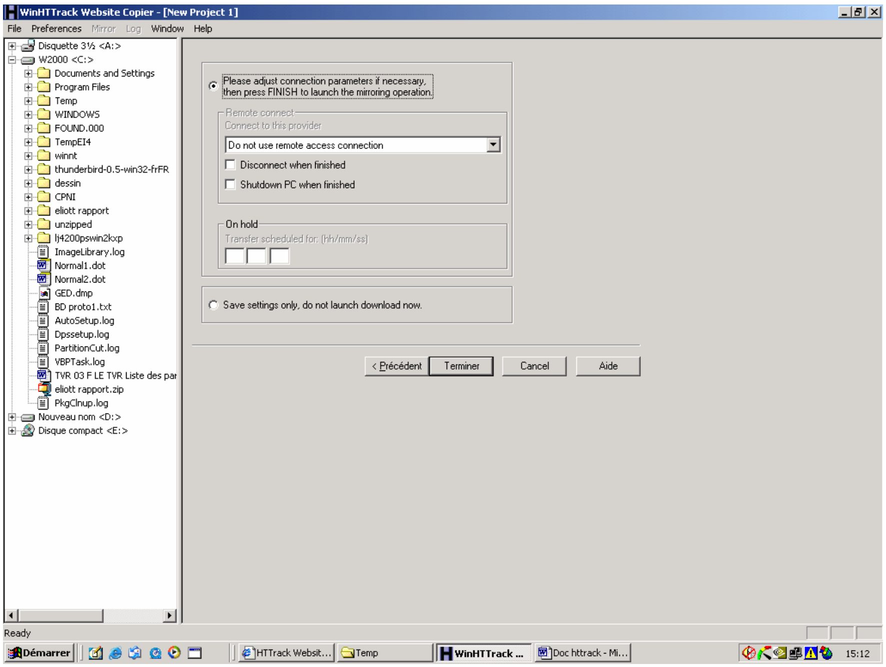
□ Hide query strings

□ Do not purge old files

Figure 7: Structure of data HTTRACK interface

D 2.9 Click "OK" button

2.2.3 Starting the crawler



● Please adjust connection parameters if necessary, then press FINISH to launch the mirroring operation.

Remote connect

Connect to this provider

Do not use remote access connection

□ Disconnect when finished

□ Shutdown PC when finished

On hold

Transfer scheduled for: (hh/mm/ss)

○ Save settings only, do not launch download now.

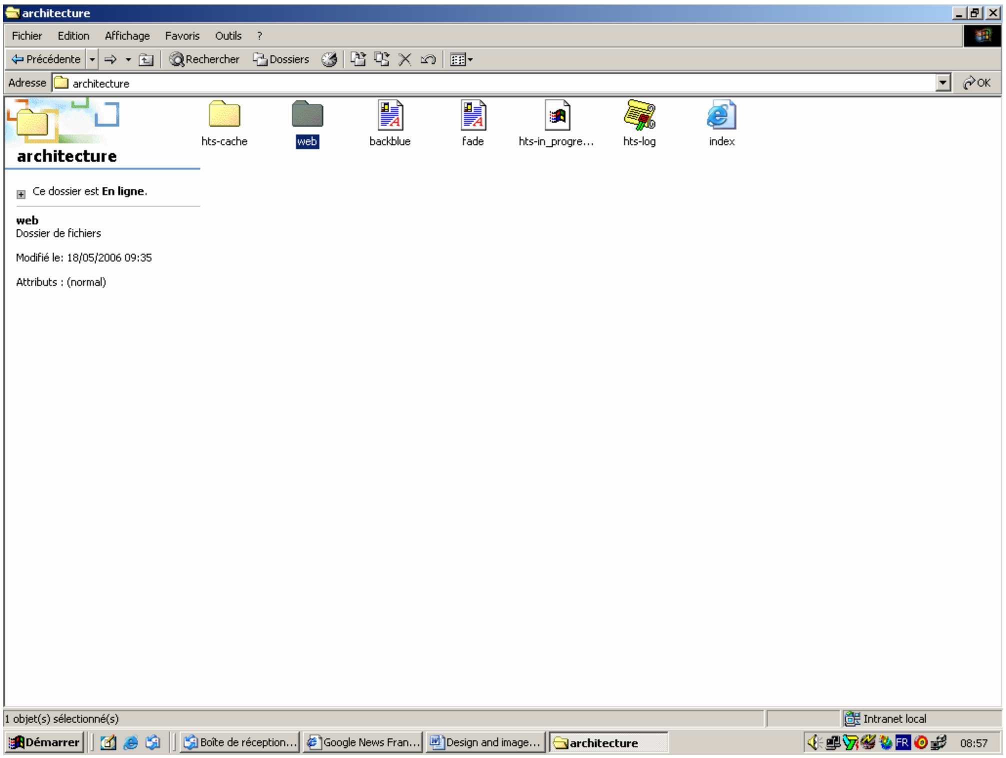
Figure 8: Final HTTRACK interface

D 2.10 Click “Terminer” (= finished, an other bug!) button to start the crawling.

2.2.4 Result of the crawler

After a certain time (24 to 72 h) of crawling, the folder web/images contains the images of the desired websites in the images folder. The user can browse these images files and select what he/she needs:

D 2.11 Open the web folder in the architecture folder



Ce dossier est **En ligne**.

web
Dossier de fichiers

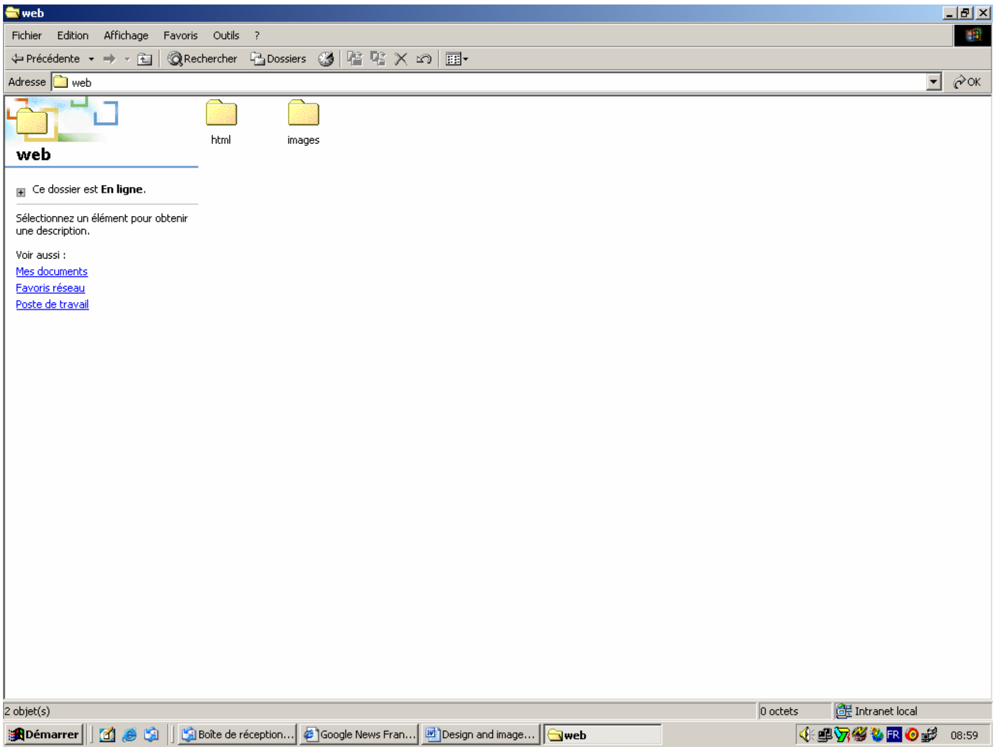
Modifié le: 18/05/2006 09:35

Attributs : (normal)

1 objet(s) sélectionné(s)

Figure 9: HTTRACK working folders

D 2.12 Open the images folder:



Ce dossier est **En ligne**.

Sélectionnez un élément pour obtenir une description.

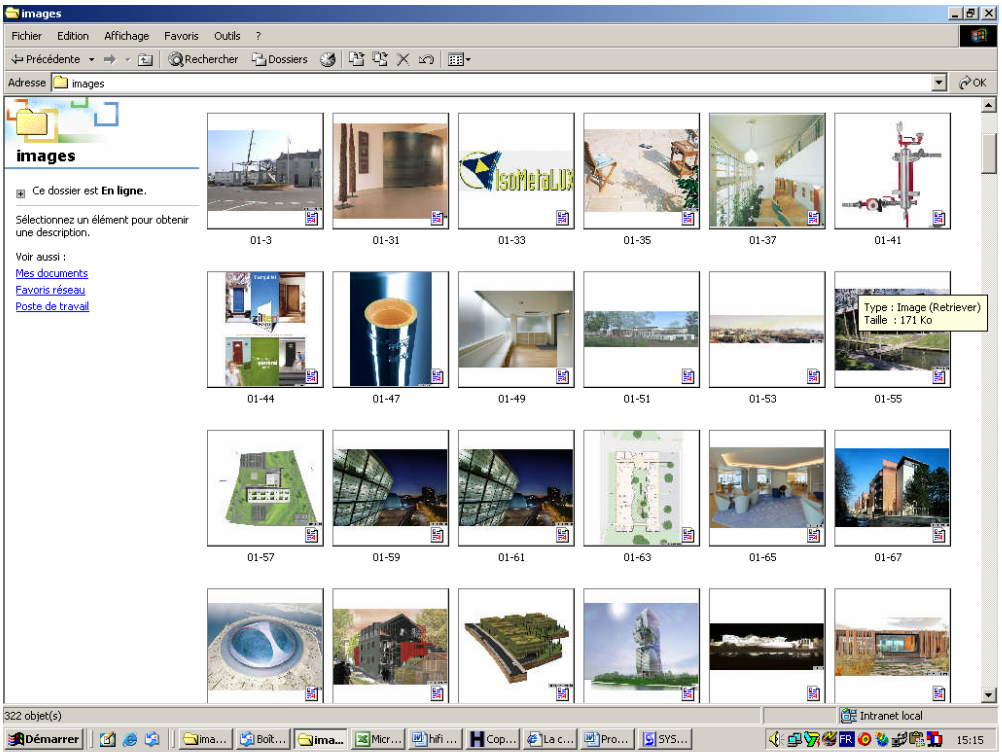
Voir aussi :

2 objet(s)

0 octets

Figure 10: HTTRACK images folder

The images folder contains images about "architecture" sector:



Ce dossier est **En ligne**.

Sélectionnez un élément pour obtenir une description.

Type : Image (Retriever)
Taille : 171 Ko

322 objet(s)

Figure 11: Examples of images from architecture sector, as results of the crawling

3. CONTENT OF DEMONSTRATION DESIGN & IMAGES DATABASE

Various sectors of influence were identified, such as automotive, aerospace, architecture, advertising, design, hi-fi, animals, plants, food, movie, cinema, music, travel, science fiction, virtual reality, as described in the deliverable 2.2. The grabbing of websites related to these sectors provides images databases as shown in the figures below:

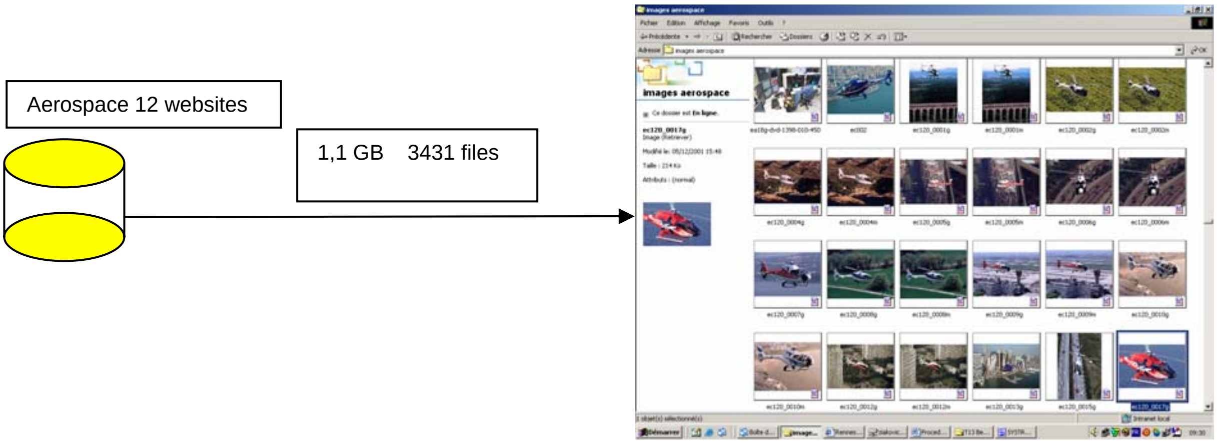


Figure 12: Aerospace images database

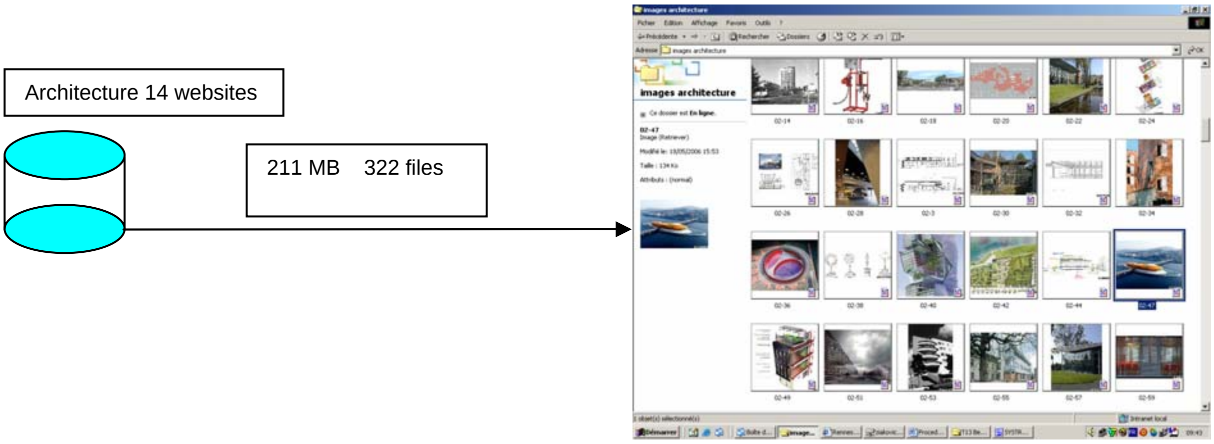


Figure 13: Architecture images database

GRABBING OF FULL SIZE TEST IMAGES DATABASES

4.1 PERTIMM SPIDER

To improve exhaustiveness and structure duplication, we will use a more powerful tool: Pertimm spider "Oggl". This spider is actually in a beta version so there is no interactive user interface but this should not be an issue with a crawler. In fact the main point is to quickly and stealthily get a good local mirror or the targeted websites or web pages.

So we use a configuration file which is really easy to write or modify using any text editor, along with a command line - to start the program.

4.2 RESULTS

As of now, only five of the seven themes have been retrieved (hifi and images are missing). But we can nevertheless start computing statistics on this basis. The grabbing was launched with 50 kB minimum size for images and 50 000 maximum documents for each website.

The time taken to grab the whole list is about 180 hours (7.5 days) on an 8MB bandwidth ADSL.

Data are stored on an external 500 GB hard drive.

4.3 GRABBED FILES STATISTICS

We can first see that out of 70 GB of grabbed documents, only 8 GB are images, which is a rather low ratio. Even if the average size of images is more than twice the average of html files size, it still represents a very low percentage (0.05%) of Internet documents.

The ".jpg" extension (JPEG coding) is the most common format for Internet images with a ratio of 12 over ".gif extension". This shows, not surprisingly, that if we want to get a huge amount of images on the Internet and store the documents to which they are linked, an even higher capacity of storage is needed.

We had some problems with this size of data to transfer files from internal hard drive to external hard drive. We encountered several crashes during the transfer, and even using "USB type 2" transfer technology, it tooks 10 hours to transfer 100 GB of data using Windows Backup procedure. In exploitation, we may face a real issue to update the data.

PS: documents processed by servers, like PHP, are transcribed to html during the grabbing, so they are included in the counting.

4.4 JPEG FILE SIZES

As JPEG is the most used file format for images on the Internet, we should estimate the distribution of file sizes of the grabbed samples:

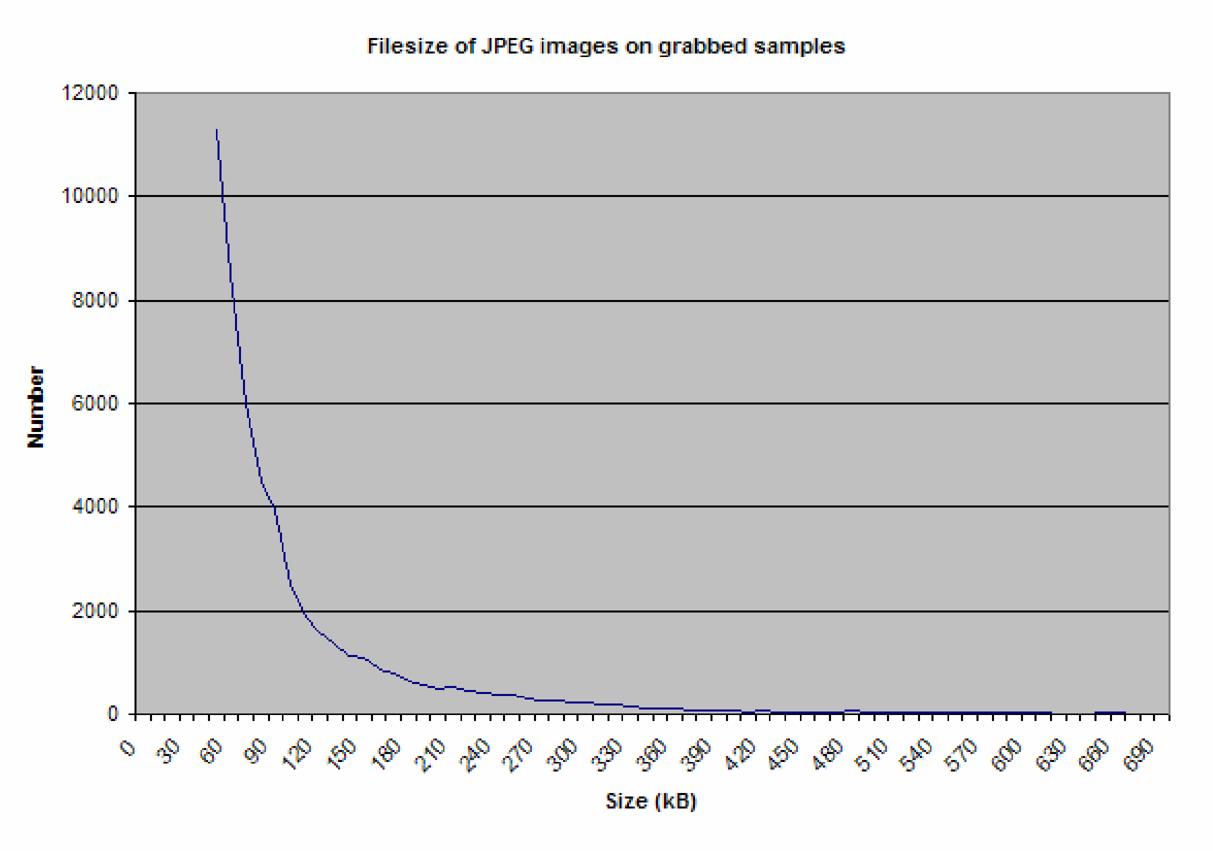


Figure 14: File size of JPEG images on grabbed samples

One can notice that we are missing most of the images with the 50 kB limit but the fact is that an image under this size wouldn't make any sense or would not be useful, e.g. little logos, part of bars, buttons, etc.). This indicates that we should be very careful in choosing that minimum size.

4.5 VOCABULARY STUDY

On the specific list called “values” we have tried to establish a typical vocabulary used by Internet websites which best represents the designer work.

The occurrence of words in these websites is the following:

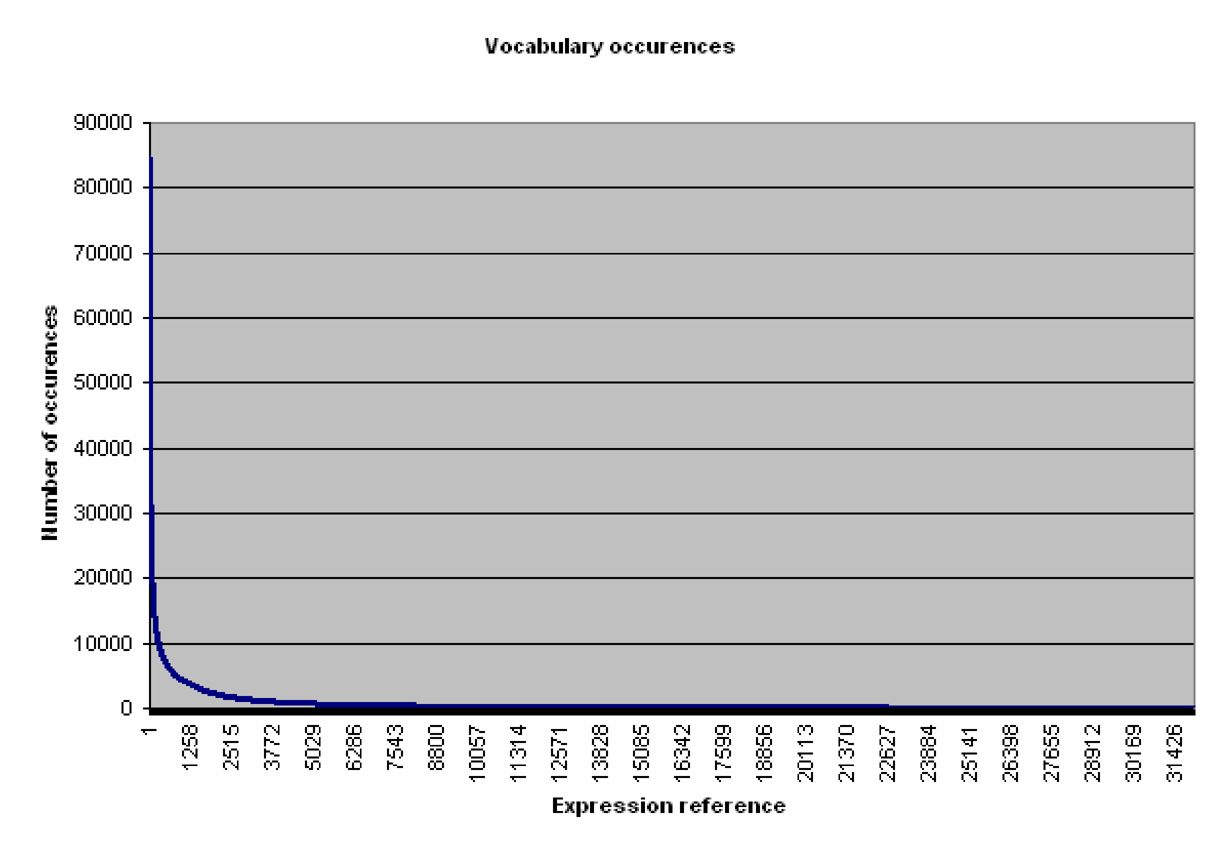


Figure 15: Vocabulary occurrences graph

As we can see, all average documents tend to make a heavy usage of just a few expressions and most of the other expressions are hardly used at all. Usually, the relevant information should be in the frequent expressions (i.e. between 300 and 3600 with an average use), but of course this is to be decided by the SERAM LCPI. This curve is the usual Zipf curve.

The word list we have drawn out has been submitted for further studies.

5. CONCLUSION

It is evident that Pertimm *Oggl* is five times better than HTTRACK (2 GB of images for HTTRACK, 10 GB for Pertimm Oggl). There are a lot of small JPEG files in the Web, which makes it not easy to find high-resolution images.

Categorizing is made by grabbing seven sources of separate starting Web addresses provided by experts. Images are put in different folders according to categorization. Pertimm may later use this categorization with its textual search engine to find images by category.

But a lot of sources of influences have the same websites sources, and the folders of results are intermingled, which makes it difficult to define categories in such a context. Metadata are contained in html files associated to images, so it is important that the intelligent agent of work package 5 extracts these metadata to allow the search engine to find images by textual or conceptual queries.

The users have an excellent sample of 95 000 images to appreciate what we can find on the web in the different sectors of influence.

6. LIST OF FIGURES AND TABLES

List of figures

7. GLOSSARY

ASP

Acronym for “Microsoft Active Server Pages”. Tool to combine HTML pages, script commands, and COM components to create interactive web pages or web-based applications that are easy to modify. The server-side scripting environment helps you create and run dynamic, interactive web server applications.

CRAWLER

A Web crawler refers to a computer program that automatically gathers and categorizes information on the Internet. Spider or grabber are synonyms of crawler.

COOKIES

Piece of information stored on users' computers by websites, in order to uniquely identify the user across multiple sessions.

CSS

Acronym for “Cascading Style Sheets”, used to format HTML, SGML and XML based documents.

FTP

Acronym for “File Transfer Protocol”. Protocol that allows to transfer files on the web between two computers.

HTTP

Acronym for “Hyper Text Transfer Protocol”. Protocol used by the web, it allows hyperlinks to work for every object of the web.

HTTPS

Acronym for “Hyper Text Transfer Protocol Security ”. Protected version of HTTP.

JAVA

Java programming language: object-oriented high-level programming language: Java applet, allows software to run in web browsers, and is accessible on most PCs.

JAVASCRIPT

Scripting language syntactically similar to Java programming language, but semantically different from it

MIME

Acronym for “Multipurpose Internet Mail Extensions” : Internet standard specifying message formats for transmission of various types of data by electronic mail.

PHP

PHP is an open-source, reflective programming language. Originally designed as a high-level tool for producing dynamic web content, PHP is used mainly in server-side applications.

PROXY

Proxy server, a computer network service that allows clients to make indirect network connections to other network services

SHOCKWAVE

File format which is used with flash software. It allows animation and is frequently used by car manufacturers to present new models on their websites.

THUMBNAIL

Thumbnails are reduced-size versions of pictures, used to make it easier to scan and recognize them, serving the same role for images as a normal text index does for words. Visual search engines and image-organizing programs normally use them, as can some modern operating systems or desktop environments, such as Windows XP, KDE, and GNOME. Note that while automatic thumbnailers reduce large pictures to a small size, the result may not be a quality thumbnail. Some inexperienced web designers produce thumbnails by simply reducing the dimensions of a large image using HTML coding, rather than using a smaller copy of the image. In practice the display size of an image in pixels should always correspond to its actual size, in part because one purpose of a thumbnail image on a web page is to reduce download time. The visual quality of browser resizing is also usually less than ideal. Reducing a significant part of the picture instead of the full frame can allow the use of a smaller thumbnail while maintaining recognizability. For example, when thumbnailing a full-body portrait of a person, it may be better to show the face slightly reduced than an indistinct figure. This has the disadvantage that it misleads viewers about what the image contains, so it is less well suited for searching or a catalogue than for artistic presentations.

URL

Acronym for “Uniform Resource Locator”.Address of each object of the web. Also called “web address”.