OSINT Documentation

Quick Start Guide

Introduction

The EMM OSINT Suite is a software package to help you search the Internet or a collection of files on disk. You can download search results from web search engines and then find relevant documents in the set of downloaded documents. In addition you can import documents from local disk for analysis. The software can process the most popular file text and binady formats, such as HTML, PDF, MS Office and others.

The core of the software is the entity extraction module which matches text locations against pre-defined patterns for different type of entities, such as person, organisation and place names, credit card numbers, VAT identifiers, URLs, etc.. User defined patterns can be added to find investigation specific entity types, such as number plates or tax identifiers.

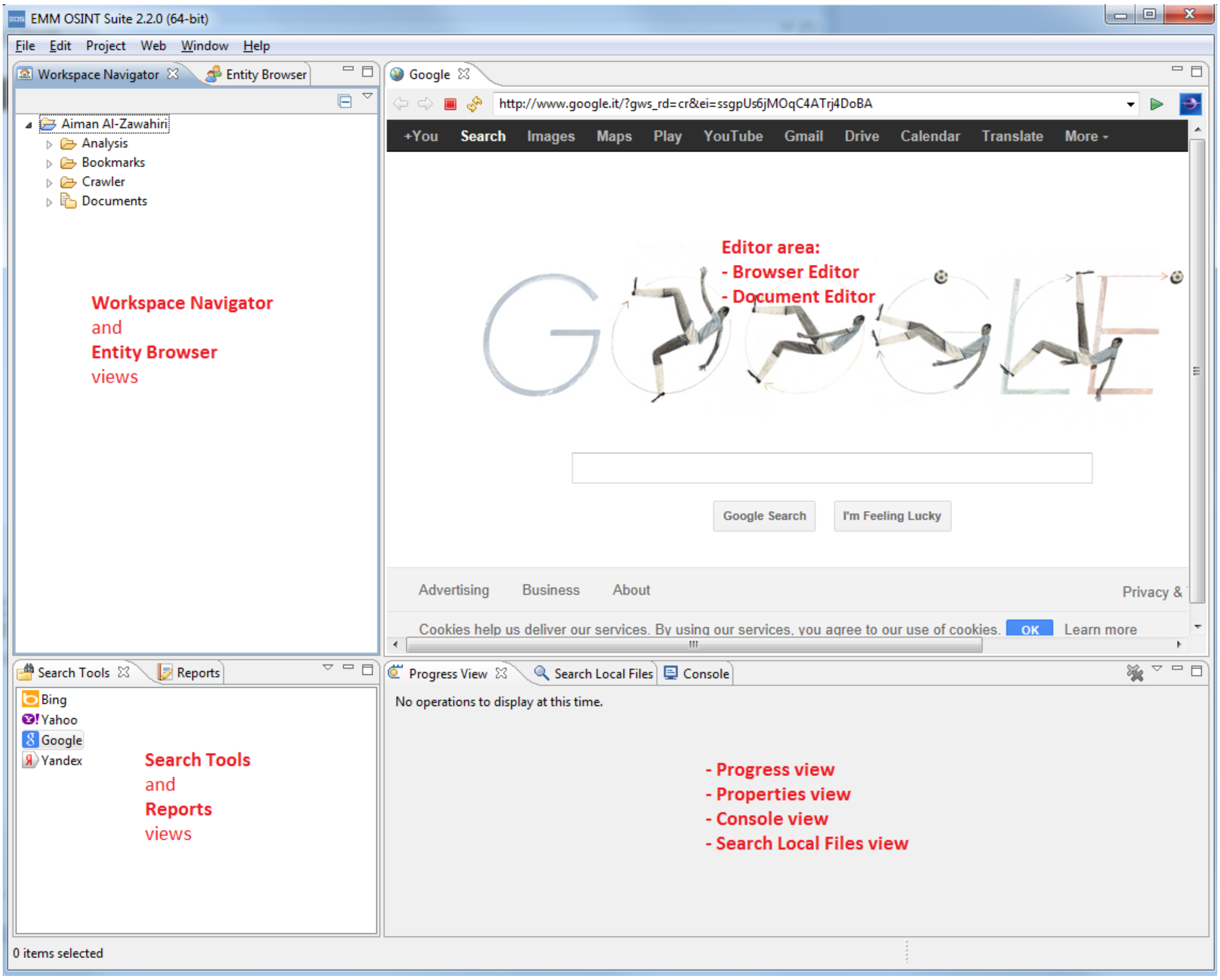
Running the application

- Running on Windows
- Running on Linux (Ubuntu)

Getting to know the application window

The application window consists of four main areas:

- The left view area with the **Workspace Navigator** view and **Entity Browser** view in the upper-left corner and the **Search Tools** view and **Reports** view in the lower-left corner
- An editor area in the centre of the window (empty at startup), this area hosts multiple editors which are usually used to edit files or contain web browser windows
- A bottom view area with multiple views, such as a the **Progress View** the **Properties** view and the **Console** view
- The right view area containing views related to the open editor (hidden or empty at startup)



Cookies help us deliver our services. By using our services, you agree to our use of cookies.

No operations to display at this time.

0 items selected

Workspace Navigator and Entity Browser views

Editor area:
- Browser Editor
- Document Editor

Search Tools and Reports views

- Progress view
- Properties view
- Console view
- Search Local Files view

The **Workspace Navigator** view shows all user created files. The root of the workspace is the chosen workspace root. The files within the workspace directory are organised in project folders (we call these projects **Case Projects**). The **Workspace Navigator** provides file operations such as cut, copy, paste, delete and rename to manage all user files.

The **Entity Browser** view shows you the extracted entity information and allows you to find documents where a specific entity was found. (In the beginning this view is empty, we first need to add some documents and run the entity extraction).

The editor area in the centre of the application window hosts multiple editors at the same time (after startup it is empty). It is used to edit files of the workspace or to interact with the web using a browser view.

The different views in the bottom view area and in the right view area show information related to files and objects selected in other views or editor windows (**Properties** view) or show status information (**Console** view) or the progress of background operations (**Progress View**).

✓ You can close views and later open them again by using the main menu: **Window > Show View** and then select the desired view

Performing an Internet Search and Downloading of Bookmarks

Carrying out the following steps, we could perform a search on the Internet using the EMM-OSINT suite:

1. Creating a Case Project
2. Performing an Internet search
3. Downloading of Bookmarks

Performing Entity Extraction and reviewing results

Once we have the documents downloaded under the *Documents* folder, the next step is to run the Entity Extraction process in order to analyze and detect the entities found in each document. Once the Entity Extraction process has finished, then we can search for any entity by using the Entity Browser view or show their relationships with other entities by using the Graph view.

1. Performing the Entity Extraction
2. Using the Entity Browser view
3. Using the Graph view

Creating a Report of the data

Reports are a way to export analysis data from a Case Project. The data can be exported into human readable formats, such as HTML, or machine readable formats such as tsv files (tab separated value - to be imported into MS Excel).

The EMM-OSINT Suite contains a function to create reports of the extracted data. A report can be used to export data from the software and open it in another program. The system can use a report template (there are a few basic ones already included), enriches the template with the extraction data and produces an output file, such as a HTML file. Therefore, there are two possibilities for creating a Report of the data:

1. Generating a Report
2. Creating a Custom Report

Tutorials

Adding a Custom Entity Type

The system provides a way to add additional custom entity types to the basic predefined types that already ship with the software (e.g. person, organization, location). A custom entity type is an additional type which extracts data not covered by the predefined types in the system.

This tutorial describes how to add a custom entity type for ***Swedish number plates.***

Introduction

Vehicle registration plates of Sweden are used for most types of vehicles and have three letters first and three digits after, if read from the left. The combination is simply a serial and has no connection with a geographic location, although the last digit shows what month the car has to undergo vehicle inspection. Vehicles like police cars, fire trucks, public buses and trolley buses use the same type of plate as normal private cars, and cannot be directly distinguished by the plate alone. Military vehicles have special plates.

The only possible coding to be seen by looking at the plate alone is when the vehicle must undergo inspection. The last digit of the plate denote this.

Last Digit	Inspection Month	Inspection Period
1	January	November-March
2	February	December-April
3	March	January-May
4	April	February-June
5	July	May-September
6	August	June-October
7	September	July-November
8	October	August-December
9	November	September-January
0	December	October-February

All letters in the Swedish alphabet are used, except the letters I, Q, V, Å, Ä and Ö. 91[1] letter combinations are not used since the may be politically offensive or otherwise unsuitable.

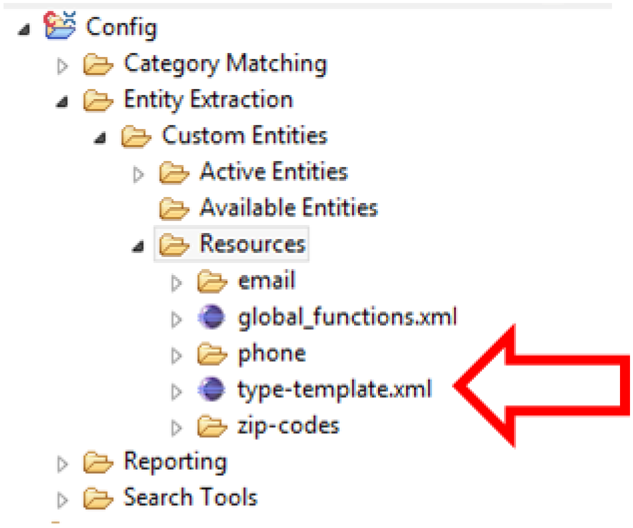
(Source: Wikipedia: Vehicle Registration Plates of Sweden)

To add a custom entity type to the system to recognize Swedish number plates, perform the following steps:

1. Creating a Configuration Project
2. Creating a new custom entity type definition file in the Entity Extraction folder.
3. Editing the new custom entity type definition file

Creating a new custom entity type definition file

To create a new custom entity type definition, copy the *type-template.xml* file to the *Active Entities* folder:



The *Active Entities* folder contains all active custom entity type definitions. You can move them to the *Available Entities* folder to temporarily deactivate them.

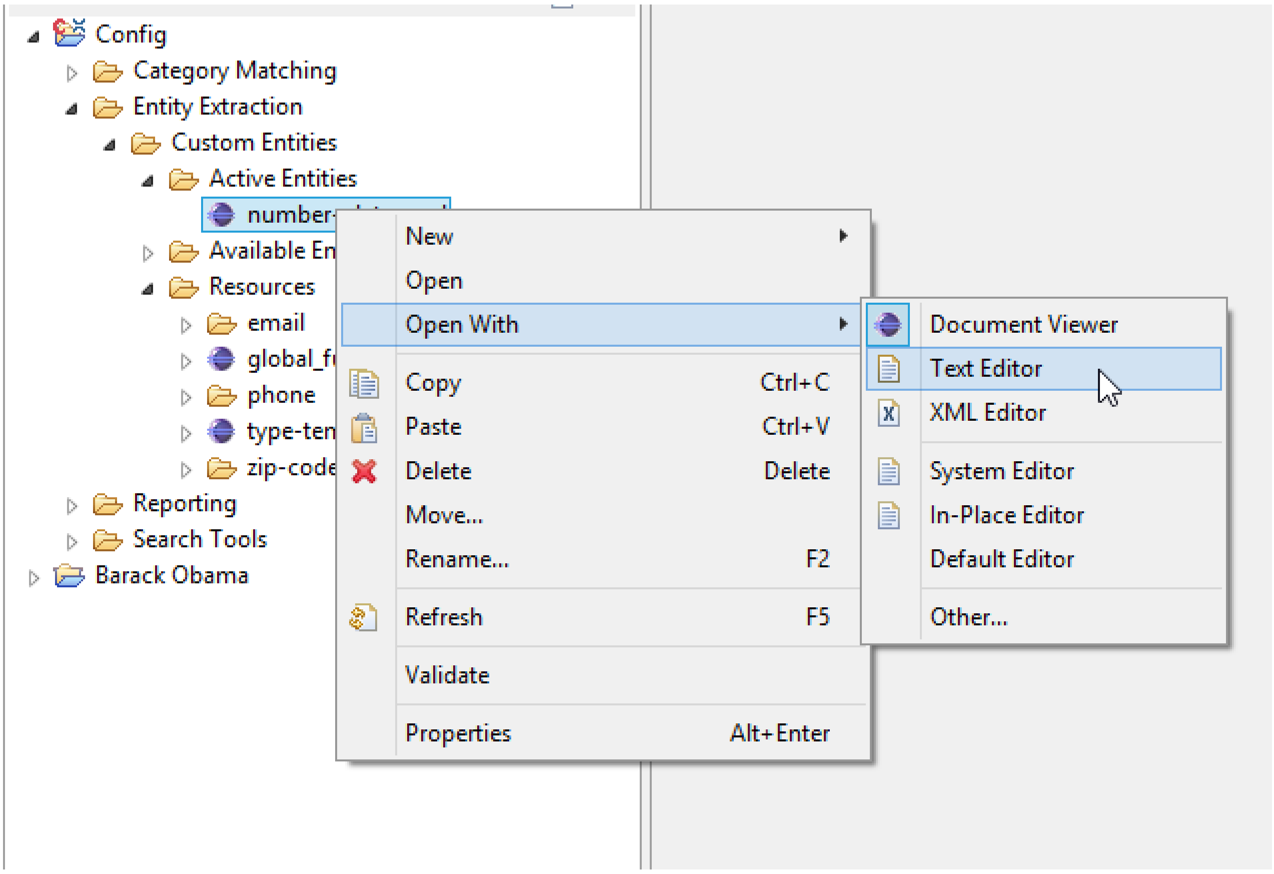
Inside the *Active Entities* folder, rename the newly copied file:

- Right click the *type-template.xml* file, then select Rename...
- Enter the name number-plates.xml and confirm.

Editing the new custom entity definition file

In the Active Entities folder,

- Right click the *number-plates.xml* file, then click on **Open With > Text Editor**.



The file will be opened in text editor in the editor area. The file contains a lot of comments to explain how to fill in the different tags.

Defining the Entity Type

In the text editor navigate after the <declaration></declaration> section and add a type entry for the new entity type as follows:

```
<type id="pn" description="number plate"/>
```

- "**id**" is mandatory; it is a code of at most two characters, and it must be unique in the OSINT Suite namespace. The letters p, o, u, t are already used for the internal types. We suggest choosing a two-characters code which is not yet used in any other custom entity definition xml file
- "**description**" is mandatory; it is a free text to describe the data type, this description is show in the user interface to denote the entity type.

Defining the Pattern to match the entity

In the text editor navigate to the <expressions> tag and add a new <expression></expression> child tag to hold the pattern definition as follows:

```
<expression>
 <regex><![CDATA[[A-ZÅÄÖ]{3}[ \-]?<000-999>]]></regex>
 <description>swedish plate number, example AAA-111</description>
 <output key="type" value="pn"/>
 <output key="country" value="sweden"/>
</expression>
```

- **<regex>** is mandatory; it defines (when mode is not set, or when mode="basic" is set) (in the cdata section) the regular expression pattern, expressed according to the syntax of dk.brics.automaton library [http://www.brics.dk/automaton/doc/dk/brics/automaton/RegExp.html]
- **<description>** is optional; it is a free text to describe the pattern
- within an **<expression>** tag, 0 or more **<output>** tags can be specified (likely at least 1); each <output> adds a piece of information as meta data to the tag which defines the found entity in the meta data of the file.
- **there must be** an **<output>** label which identifies the data type of the term that matched the pattern; in the above example this is <output key="type" value="pn"/> which tells the system that any term matching that pattern is of data type "pn" (plate number)

Pattern Syntax
By default the system uses the syntax of the BRICS library (see http://www.brics.dk/automaton/doc/dk/brics/automaton/RegExp.html) which is less expressive than the normal *java.util.regex* package. If you want to use the full *java.util.regex* syntax please set the mode attribute to "groups": **<regex mode="groups">**

This pattern will match three uppercase letters, optionally followed by a space or a dash, followed by a number between 000 and 999. This is a list of example terms that would match the above pattern:

```
WNF766
WNF 766
WNF-766
```

Understanding the definition of a Custom Entity Type

As can be seen, the pattern (regular expression) used as example for recognizing Swedish number plates and defined within the <regex> label is

`[A-ZÅÄÖ]{3}[ \-]?<000-999>`

Three main parts can be detected in this regular expression:

Pattern	Meaning
`[A-ZÅÄÖ]{3}`	This pattern will match three uppercase letters (from A to Z including the letters Å, Ä or Ö)
`[ \-]?`	Optionally (symbol ?) can follow a space or a dash
`<000-999>`	Necessarily is followed by a number between 000 and 999

There are many regular expression testers available on the Internet that allow testing our regular expressions (regexpal, regexr, etc.)

Under the hood, the system matches this expression to all text contents of the files being processed by the entity extraction. If some term matches, then the system adds a meta tag to the meta data of the file which is an xml element such as:

```
<emm:custom type="pn" country="sweden" name="the term that matched the pattern"
pos="the position of the term that matched the pattern" id="an unique identifier
for name">the term that matched the pattern</emm:custom>
```

In our above example, let's say we have a document which contains some text interspersed with some number plate terms as follows:

.... WNF766 ... ADE-683 ... OWA 882 ...

If the Entity Extraction process analyzes this text, it will produce the following tags to be included in the meta data of the file:

```
<emm:custom type="pn" country="sweden" name="WNF766" id="7">WNF766</emm:custom>
<emm:custom type="pn" country="sweden" name="ADE-683" id="8">ADE-683</emm:custom>
<emm:custom type="pn" country="sweden" name="OWA 882" id="9">OWA 882</emm:custom>
```

In other words, the system thinks it has found three different number plates (see the different id values), even though they are only spelled slightly differently and describe the same number plate. In order to overcome this problem we need to output the matched terms in a standardised form.

Further improving the custom regular expressions

In this section we will show a how to customize the regular expressions used for recognizing entities in the EMM-OSINT Suite.

Following the example above about recognizing ***Swedish number plates**,* the following steps are explained:

- **Standardizing the output name** of the entity. This would be interesting to apply when OSINT finds out different entities that are not spelled exactly in the same way, but they refer to the same entity
- **Defining capturing groups in the pattern**. The current library in OSINT for regular expressions doesn't support capturing groups. To avoid that, we can set the feature "mode" on the <regex>
- **Using keywords as patterns**. Basically, a text file containing the keywords is used to define the exact terms to recognize
- **Using a script to generate the patterns on-the-fly**. Following a Java style of coding, OSINT allows defining own scripts to recognize entities

Standardizing the output name

The output key called "name" of the entity recognized by OSINT can be standardized. Returning to the example of the regular expression (pattern) defined to recognize Swedish number plates:

```
<expression>
 <regex><![CDATA[[A-ZÅÄÖ]{3}[ \-]?<000-999>]]></regex>
 <description>swedish plate number, example AAA-111</description>
 <output key="type" value="pn"/>
 <output key="country" value="sweden"/>
</expression>
```

the output in OSINT for a text including "... WNF766 ... WNF 766 ... WNF-766 ..." would be:

```
<emm:custom type="pn" country="sweden" name="WNF766" id="7">WNF766</emm:custom>
<emm:custom type="pn" country="sweden" name="WNF 766" id="8">WNF 766</emm:custom>
<emm:custom type="pn" country="sweden" name="WNF-766" id="9">WNF-766</emm:custom>
```

Notice how there are three different entities with different "name" output keys. If we want to standardize this key in order to OSINT shows the same entity (*WNF766* for instance) for the three cases, we should add the "name" output key within the definition of the pattern as follows:

```
<expression>
<regex><![CDATA[[A-ZÅÄÖ]{3}[ \-]?<000-999>]]></regex>
<description>swedish plate number, example AAA-111</description>
<output key="type" value="pn"/>
<output key="country" value="sweden"/>
<output key="name"><![CDATA[
     name = term.replaceAll("[ \\-]", "");
     return name;]]>
</output>
</expression>
```

Now, if the Entity Extraction module processes the text again, it will now produce the following meta tags:

```
<emm:custom type="pn" country="sweden" name="WNF766" id="7">WNF766</emm:custom>
<emm:custom type="pn" country="sweden" name="WNF766" id="7">WNF 766</emm:custom>
<emm:custom type="pn" country="sweden" name="WNF766" id="7">WNF-766</emm:custom>
```

As shown, now the output for the "name" key is unified for all the entities.

Defining capturing groups in the pattern

By default, the current library in OSINT for regular expressions doesn't support capturing groups within the definition of the pattern. To avoid that, we can set the feature "mode" on the <regex> as follows:

```
<expression>
  <regex mode="groups"><![CDATA[([A-ZÅÄÖ]{3})[ \-]?(<000-999>)]]></regex>
  <description>swedish plate number, example AAA-111</description>
  <output key="type" value="pn"/>
  <output key="country" value="sweden"/>
</expression>
```

Notice how the "**mode**" key is added to the <regex> label and set up to "**groups**" in order to allow the definition of groups (using **brackets**) in the regular expression.
Then, we can use the defined groups for further processing, referring to them as *groups[1], groups[2]*, etc.

In the example above, if OSINT finds the entity "WNF766" in the text, the variable "groups[1]" would refer to the string matched by the first group defined in the regular expression (i.e. "WNF"), while "groups[2]" would refer to the string matched by the second group defined ("766"). Therefore, these "groups" might be used within a script that we can also define for generating patterns on-the-fly, as explained below.

Using keywords as patterns

In OSINT we can use keywords as patterns by using an external text file in which the keywords are included. To use this feature we have to set the "mode" key within <regex> as follows:

```
<expression>
  <regex mode="file"><![CDATA[keywords.txt]]></regex>
  <description>swedish plate number, example AAA-111</description>
  <output key="type" value="pn"/>
  <output key="country" value="sweden"/>
</expression>
```

Therefore, we have to define within the CDATA section the relative path (starting from where the XML file which defines the custom entity is loaded) to a file which contains keyword terms. The file which contains keyword terms must be a **text file**, in **UTF-8 format**. It is important to note that each line (which is not an empty line nor a comment line) is considered a keyword term, case sensitive, and it will be matched as-is (no need to escape the special characters). An example of the "keywords.txt" file would be:

```
#comment lines begin with # and are ignored
#empty lines (as the one below) are also ignored
#if possible, the first and the last line of the file
#should be either an empty line or a comment line
#keywords start here
WNF766
wnf766
WNF-766
wnf-766
```

Taking into account the "keywords" file above, for the example including the text "... WNF766 ... WNF 766 ... WNF-766 ...", OSINT would produce two xml elements as such:

```
<emm:custom type="pn" country="sweden" name="WNF766" id="7">WNF766</emm:custom>
<emm:custom type="pn" country="sweden" name="WNF-766" id="8">WNF-766</emm:custom>
```

ⓘ Whitespaces are allowed inside the keywords defined within the "keywords" file

Using a script to generate the patterns on-the-fly

Following a Java style of coding, OSINT allows defining own scripts to recognize entities. To use this feature we have to set the "mode" key with the value "script" within the <regex>, as follows:

```
<expression>
  <regex mode="script"><![CDATA[
      List<String> ls = new ArrayList<String>();
      ls.add("WNF[ \-]?[0-9]{3}");
      ls.add("UCD[ \-]?[0-9]{3}");
      return ls;
    ]]></regex>
  <description>swedish plate number, example AAA-111</description>
  <output key="type" value="pn"/>
  <output key="country" value="sweden"/>
</expression>
```

As can be seen, the script (in the <regex> CDATA section) is in the **Java language**. Some features related to the "script" mode are:

- the script can reference the path from where the XML definition file is loaded as "**resourcespath**" ("resourcespath" is a String)
- the script **must return** a List<String>, where each element in the list is a regular expression pattern

For the example above, the script will recognize entities such as "WNF777", "WNF-876", "WNF 987", "UCD465", "UCD-999" or "UCD 112".

Creating effective search queries

In order to find relevant pages on the internet it is important to have some background information about how a search engine works. On this page we summarise important tips how to effectively search on Google. The described techniques work on other internet search engines as well.

How does Google Search works?

Well, we don't know exactly, since this is Google's trade secret. However, we can make some assumptions based on Google's documentation and practical experience.

When you do a Google search you are not searching in the entire web, you are searching on the **Google Index**. Google uses software programs called web spiders that go through web pages following the links and storing all the information across hundred of thousands of machines they have. Currently, many billions of pages are stored by Google. When we type a query and hit return, the Google software

searches its large index to find every page that includes those terms. The index which is queried may be different
But, how does Google decide which few documents I really want? By taking into account over 200 parameters for each document in its index:

- How many times does this page contain your keywords?
- Do the words appear in the title of the page? in the url?
- Does the page include synonyms for those words?
- Is this page from a quality website or from a spam one?
- What is its **page rank**? (formula invented by Google that measures the importance of a web page by looking at how many outside links point to it and how important those links are)
- etc.

They combine all these factors to produce the overall score of each page and return the search results.

Tips for making a good search query in Google

- Choose **words** for your query that you think **will appear on the result page**
- Avoid using words that do not describe the concept you are looking for and use specific words (not common or generic)
- **Word order matters**, as well as every word appears in the search query also matters
- **Capitalization does not matter**
- **Punctuation often does not matter** (appart from special cases, such as "Google+" or the programming language "C++"
- **Spelling matters,** however Google might offer corrections to mispelt words

Power searching with Google

- **quotes**: use quotes to search for a phrase that will appear exactly as is in the results. Words can be used before or after the quoted phrase
- **site**: return results from the specified site only
- **filetype**: return files of the extension you specify. NO space between filetype, the colon, and the extension. Some file types: txt, pdf, swf, xlsx, gif...
- **minus (-)**: Eliminate irrelevant results. There must be a space before the minus sign. There must not be a space between the minus sign and the word you want to eliminate.
- **OR**: Use OR to include more than one way of expressing an idea.

Exercise C1-2

- Search for drug trafficking on The Guardian newspaper website
- Search for pdf documents contain the exact phrase "Al Qaeda"
- Search for documents about terrorist attacks or plans but not containing words such as "United States" nor "Iraq"

Related information:

- http://www.powersearchingwithgoogle.com
- Power Searching with Google Quick Reference

Editing the current Name Variant Database

The EMM OSINT Suite uses a large database of named entities containing mainly persons and organizations (see the Name Variant Matching concept for more information).

The suite allows editing its Name Variant Database in order to add new entities or modify existing ones, keeping the current keys assigned for each entity.

The Name Variant Database in OSINT is composed of **four columns** as follows:

- **KEY**. The import process does not take the values under this column into account because it already assigns its own primary key for each new entity. However, this column must appear as the first column within the TSV file, although their values are discarded.
- **PID (Profile Identification)**. It is the identification value of an entity. This numerical value is very important. For variants (different names for the same entity) that belong to the same entity, this value must be identical. The first occurrence found in the TSV file will be considered as the canonical entity (original name form of the entity) and the following ones as variants of this canonical form (see the example below).
- **TYPE**. It is the type of the entity. OSINT accepts four main entity types:
  - **o**, for organizations
  - **p**, for persons
  - **t**, for toponyms (locations)
  - **u**, for unknown types of entities
- **VARIANT**. It is the form or name of the entity, exactly written as you want that the process matches it in the documents.

Next, an example of a excerpt from the Name Variant Database in OSINT is shown:

key	pid	type	variant
2	11	p	Aaron Albert

3	11	p	A. Albert
4	11	p	A. M. Albert
5	21	o	Chad Calvin Christian
6	21	o	CCC
7	21	o	C.C. Christian
8	41	t	Milano
9	61	u	Harold Hugh
10	61	u	Henry Hugh

In this example, it can be observed how the entity *Aaron Albert* (person) has a PID value of 11. The first occurrence would be the canonical (original) form for that entity, whereas the next ones found with the same PID (*A. Albert*, *A. M. Albert*) are considered as variants of that canonical form. However, all these occurrences (variants) represent the same entity in real life (the person *Aaron Albert*). Another example in the table is the organization called *Chad Calvin Christian* (PID 21). As can be seen , there exist one canonical form and two variants (*CCC*, *C .C. Christian)* for this entity. Finally, we find the entity *Milano* (PID 41) with only one variant (the canonical form) and one entity of unknown type (*Harold Hugh*) with two variants.

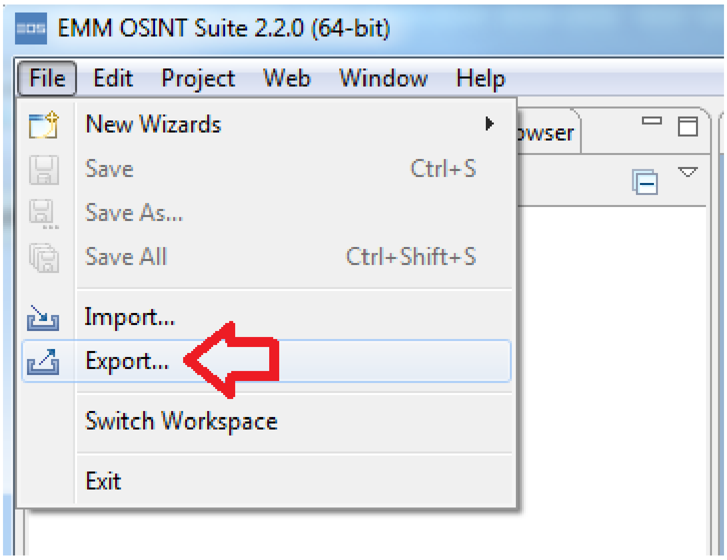
It is important to note that it should be used a **UTF-8** flat file with **TSV (Tabular Separate Values)** format.

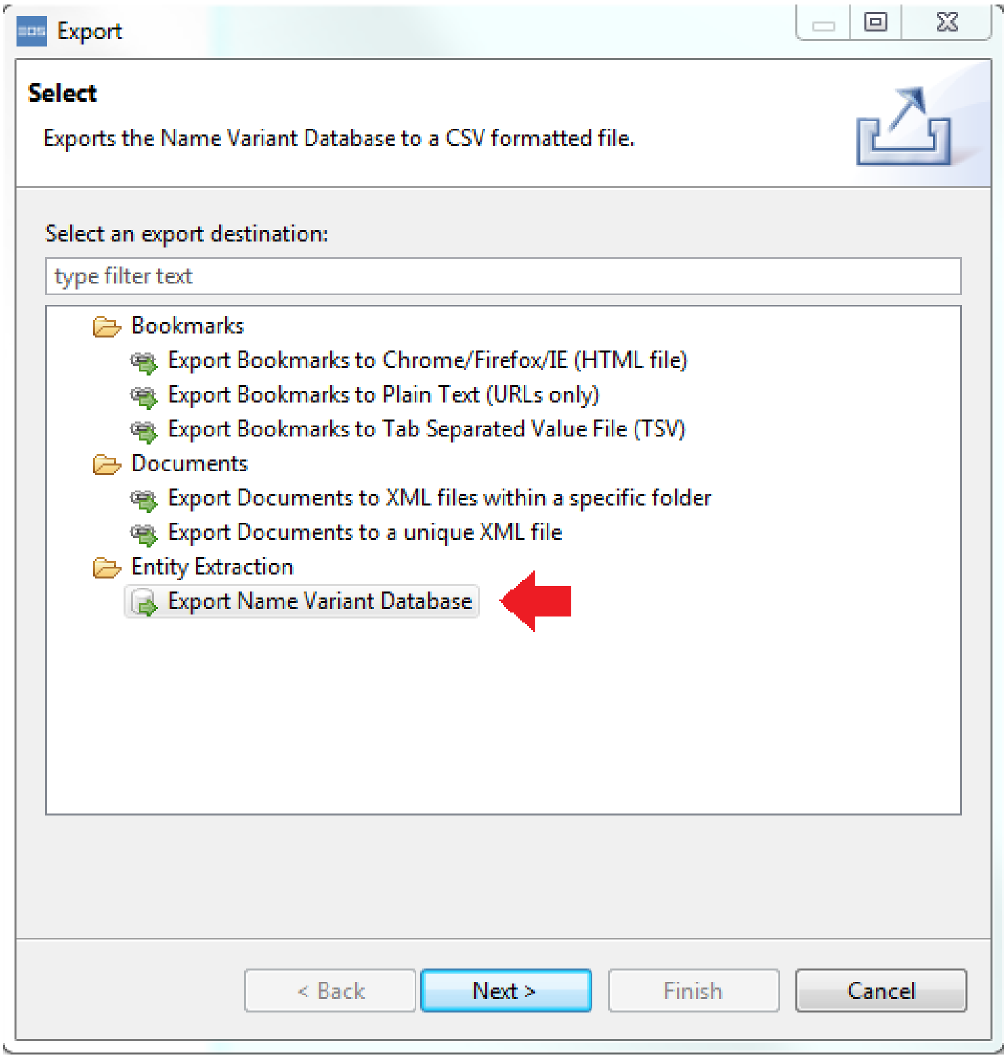
The procedure of editing the current Name Variant Database should be done in three steps:

1. Export the current Name Variant Database to a flat file
2. Open the database file and add new entities or modify existing ones. It is important to note that the database file **must be saved in UTF-8 format** and always keeping the structure of **four columns**, separating them by the tabular character.
3. Import the updated database file into OSINT

Exporting the current name variant database to a flat file

- Open EMM OSINT Suite and click in the main menu on **File > Export > Entity Extraction > Export Name Variant Database**





Export

Select

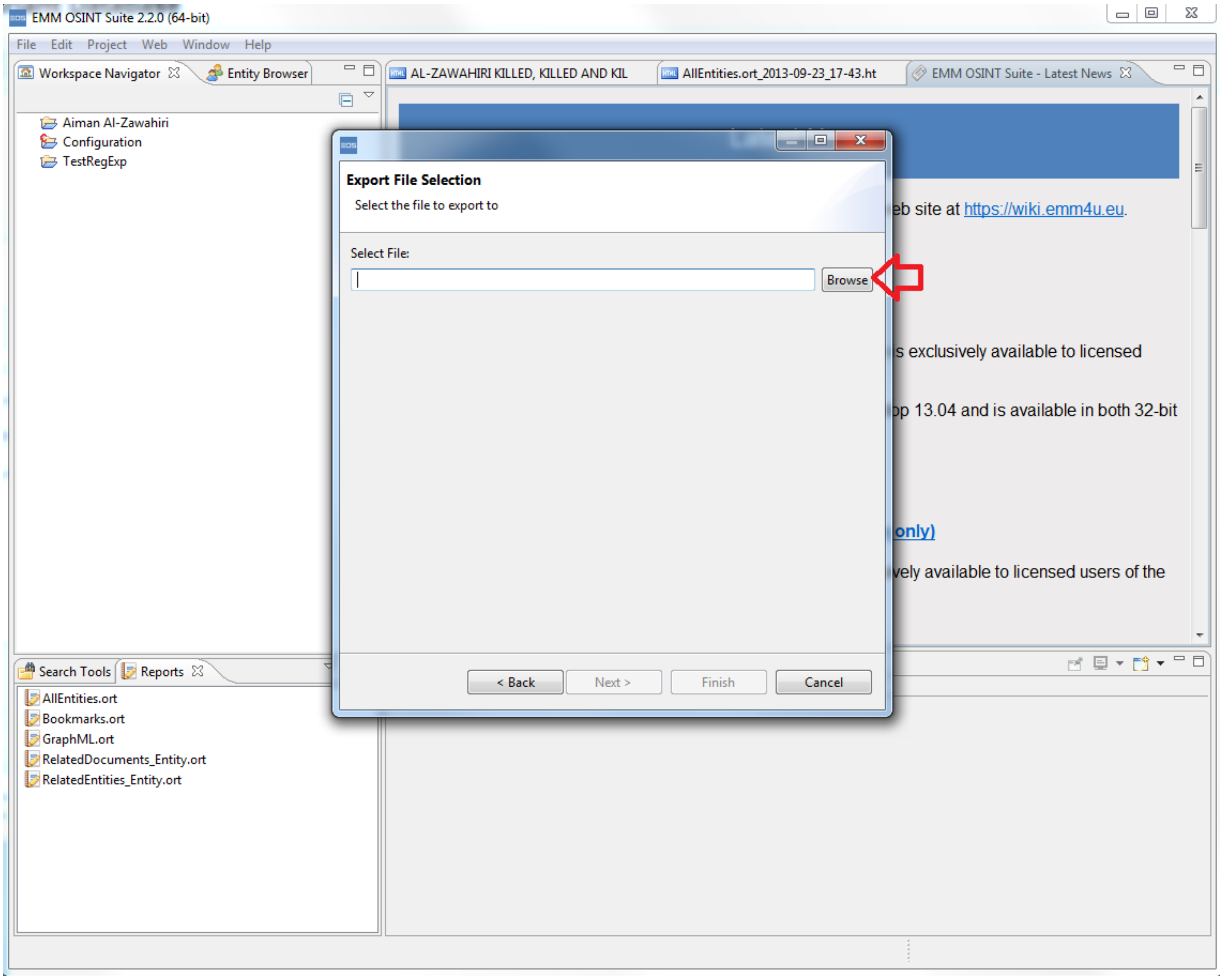
Exports the Name Variant Database to a CSV formatted file.

Select an export destination:

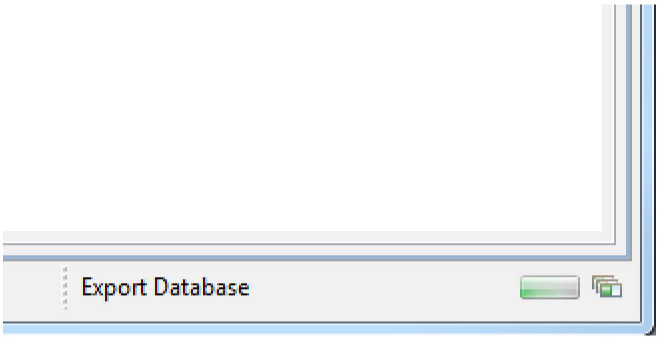
type filter text

- Bookmarks
  - Export Bookmarks to Chrome/Firefox/IE (HTML file)
  - Export Bookmarks to Plain Text (URLs only)
  - Export Bookmarks to Tab Separated Value File (TSV)
- Documents
  - Export Documents to XML files within a specific folder
  - Export Documents to a unique XML file
- Entity Extraction
  - Export Name Variant Database

- Click on **Next** and then **Browse** in order to **select the file** in your computer in which exports the current Name Variant Database. You can use any name for this file.


EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help
Workspace Navigator
Entity Browser
Aiman Al-Zawahiri
Configuration
TestRegExp
AL-ZAWAHIRI KILLED, KILLED AND KIL
AllEntities.ort_2013-09-23_17-43.ht
EMM OSINT Suite - Latest News
Export File Selection
Select the file to export to
Select File:
Browse
< Back
Next >
Finish
Cancel
Search Tools
Reports
AllEntities.ort
Bookmarks.ort
GraphML.ort
RelatedDocuments_Entity.ort
RelatedEntities_Entity.ort

- Finally, click on **Finish** and a progress bar on the right-bottom of the window will be shown. It might take few seconds depending on the size of the current database. The process of exporting can also be followed in the **Progress view**



The exporting process will finish when this progress bar disappears.

Opening the database file and adding new entities or modifying existing ones

The new export file generated is a flat file composed of **four columns** separated by the tabular character (see above).

Any text editor can be used to modify the database file (TextPad for Windows, WordPad, ...).

Once we have added or modified the new entities, as explained above, the database file **must be saved in UTF-8 format** and always must keep the four columns format.

Importing the updated database file into OSINT

See Importing a new Name Variant Database to import the updated database file.

Importing a Name Variant File

The EMM OSINT Suite uses a database of named entities containing mainly persons and organizations (see Name Variant Matching for more information). Sometimes it might be interesting to use your own database with specific entities to be matched by the entity extraction process. This tutorial shows you how to import your own data.

Creating an import file with custom name variants

Basic Requirements

- The file needs to be encoded in UTF-8 (refer to Encoding a File in UTF-8)
- The file must not have empty lines

Format of the import file:

The file is a TSV (tab- separated values) file. The file should contain **four columns**:

- **KEY**. The import process does not take the values under this column into account because it already assigns its own primary key for each new entity. However, this column must appear as the first column within the TSV file, although their values are discarded.
- **PID (Profile Identification)**. It is the identification value of an entity. This numerical value is very important. For variants (different names for the same entity) that belong to the same entity, this value must be identical. The first occurrence found in the TSV file will be considered as the canonical entity (original name form of the entity) and the following ones as variants of this canonical form (see the example below).
- **TYPE**. It is the type of the entity. OSINT accepts four main entity types:
  - **o**, for organizations
  - **p**, for persons
  - **t**, for toponyms (locations)
  - **u**, for unknown types of entities
- **VARIANT**. Is the name variant of the entity. The matching of the name variant is not exact but matches according of some rules.

Matching the name variant against the real text

The import file contains name variants as the fourth column. These variants are matched against the real text using the following rules:

Rule	Description	Example
Lower case matches both cases	If the name variant is imported as lower case, it matches both upper and lower case in the text	Name variant "procter and gamble" matches "Procter and Gamble" and "PROCTER AND GAMBLE"
Upper case matches only upper case	If the name variant contains upper case characters, these characters will only match upper case characters in the text.	Name variant "Procter and Gamble" matches "Procter and Gamble" **but not** "procter and gamble" Name variant "PROCTER AND GAMBLE" matches only "PROCTER AND GAMBLE"
Some characters are ignored	There are a number of special characters which will be ignored. These characters are "." (dot), "&" (ampersand), ":" (colon) and "-" (dash).	Name variant "Procter & Gamble" matches "Procter & Gamble", "Procter - Gamble", "Procter:Gamble", "Procter Gamble", etc.
Using wildcard character '%'	The percentage character will match zero or more characters.	Name variant "Procter%" matches "Procter & Gamble", "Procter - Gamble", "Procter:Gamble", "Procter Gamble", etc.

Using wildcard character '_'	The underscore character matches any single character	Name variant "Procter_Gamble" matches "Procter&Gamble", "Procter-Gamble", "Procter:Gamble" but matches not "Procter & Gamble" (first whitespaceis taken up by wildcard character)

Here is an example of a TSV file used for importing new name variants into OSINT:

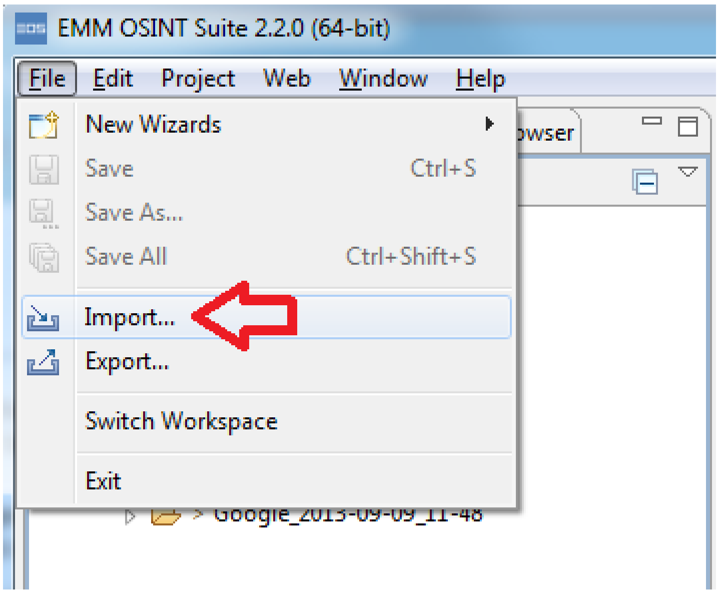
Key	Pid	Type	Variant
2	11	p	Aaron Albert
3	11	p	A. Albert
4	11	p	A. M. Albert
5	21	o	Chad Calvin Christian
6	21	o	CCC
7	21	o	C.C. Christian
8	41	t	Milano
9	61	u	Harold Hugh
10	61	u	Henry Hugh

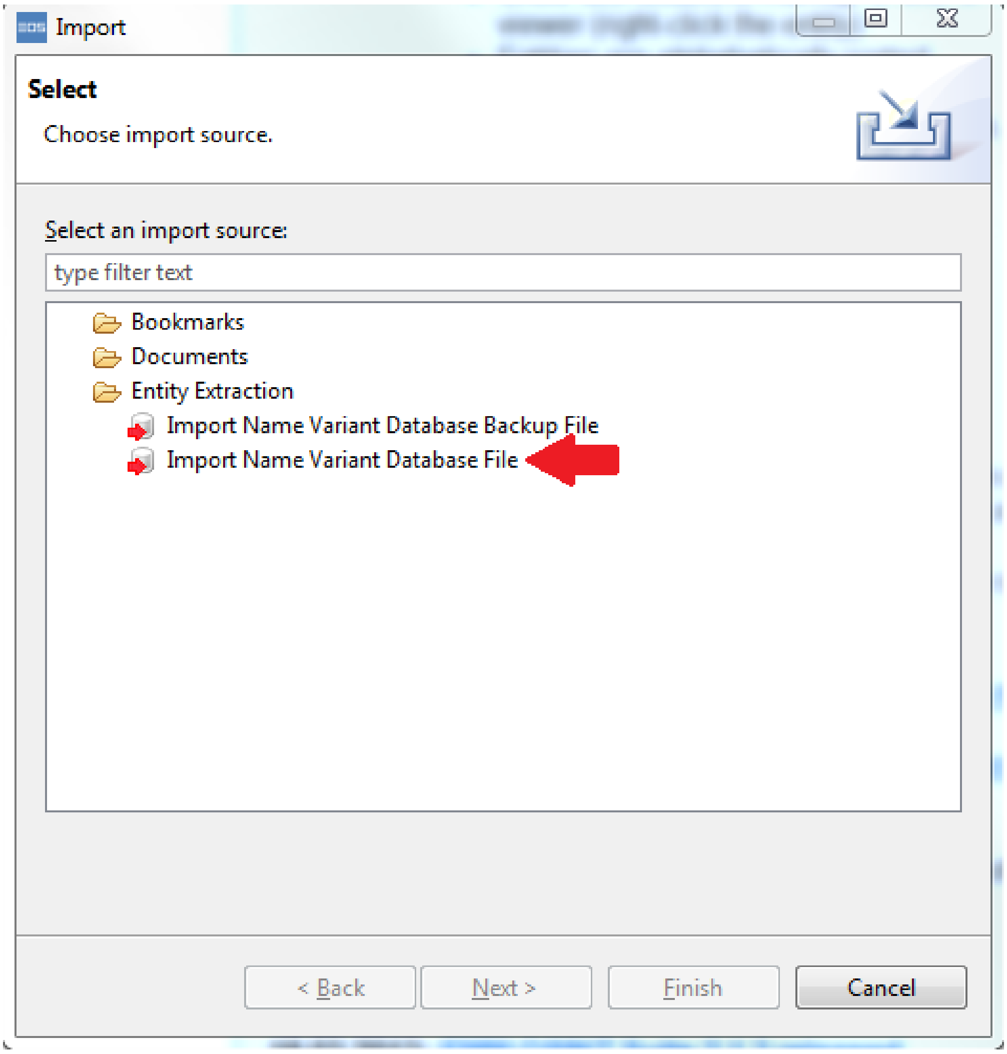
In this example, it can be observed how the entity *Aaron Albert* (person) has a PID value of 11. The first occurrence would be the canonical (original) form for that entity, whereas the next ones found with the same PID (*A. Albert*, *A. M. Albert*) are considered as variants of that canonical form. However, all these occurrences (variants) represent the same entity in real life (the person *Aaron Albert*). Another example in the table is the organization called *Chad Calvin Christian* (PID 21). As can be seen , there exist one canonical form and two variants (*CCC*, *C .C. Christian)* for this entity. Finally, we find the entity *Milano* (PID 41) with only one variant (the canonical form) and one entity of unknown type (*Harold Hugh*) with two variants.

Importing the Name Variant File

Perform the following steps to import the name variant file:

- Open the EMM OSINT Suite and click in the main menu on **File > Import > Entity Extraction > Import Name Variant Database File**





- Next click **Browse** and **select the TSV file** to import database from. Then click **Finish** to perform the import.

Finally the system imports the new database and prints a status message in the **Console** view.

> Before starting the import process, the system generates automatically a backup of the current database under the workspace folder
> under *<workspace>./metadata/.plugins/it.jrc.osint.extract/entity_20_backup_<date>.h2.db* .
>
> If something goes wrong, close the application and copy this backup file back into place over *entity_20.h2.db*.

New user-friendly format of the import file

> **Upcoming Feature**
> The new format of the import file will be introduced with version 2.4 of the software

In order to facilitate encoding of name variants a more user-friendly format of the import file will be introduced shortly

Basic Requirements

- The file needs to be encoded in UTF-8 (refer to Encoding a File in UTF-8)
- The file must not have empty lines

Format of the new import file:

The file consists of multi-line blocks, where each of such blocks provides the canonical and variant forms of one single entity.

Each block starts with a line including the canonical form and the corresponding entity type information, which are separeted with a tab. All subsequent lines in the same block
start with a tab, which is followed by variant form of the current entity.

Here is an example of an import file that contains 3 blocks corresponding to three entities.

World Anti-Doping Agency **ORG-PP**
Agenzia Mondiale Antidoping
Agence Mondiale Antidopage
Weltantidopingagentur
Agência Mundial Anti-Doping
Amsterdam Airport Schiphol **LOC-FA**

Aéroport de Schiphol
Aeroporto di Amsterdam
Schipol International Airport
Aeroportul Internaional Schipol
George W. Bush **PER**
George Walker Bush
President Bush
President George W. Bush
George W Bush
Bush the Younger
George Bush Jr.

The first block contains 5 name variants (including the canonical form) of the entity *World Anti-Doping Agency* (of type **ORG-PP** - political/public organisation). The second block
contains 6 name variants of the entity *Amsterdam Airport Schiphol* (of type **LOC-FA** - facility). The third block contains 7 variants of the entity *George W. Bush* (of type **PER** - person).
More information on new entity types will be provided shortly.

Please note that in case of ambiguous entities (i.e., entities that can have more the one type) one introduces a separate block of variants for each specific entity type.

Using the Duplicate Bookmark Detection

The EMM OSINT Suite automatically detects duplicate bookmarks. This is a useful feature to:

- Combining the search results from multiple search engines
- Repeating a search and discover new search results

As an example we search for a prominent person and combine the search results into a single folder.

This tutorial describes an example scenario to demonstrate the feature.

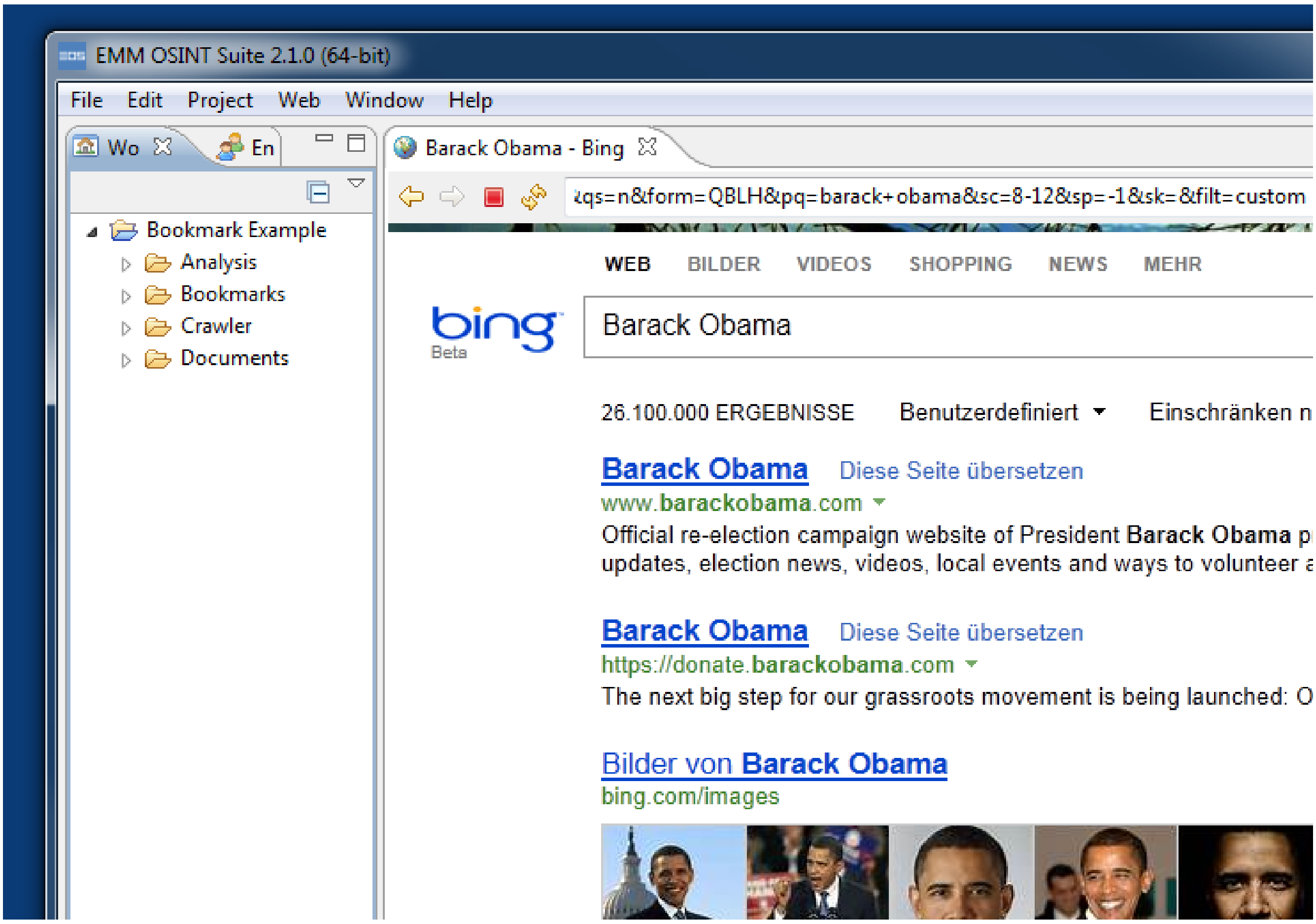
Perform the following tasks:

- Opening EMM OSINT Suite
- Creating a Case Project
- Perfoming an Internet search
- Organizing Bookmarks

Performing an Internet Search

To perform a search do the following:

- Double Click on Bing in the Search Tool View, the application opens a browser window pointing to http://www.bing.com
- Search for Barack Obama:

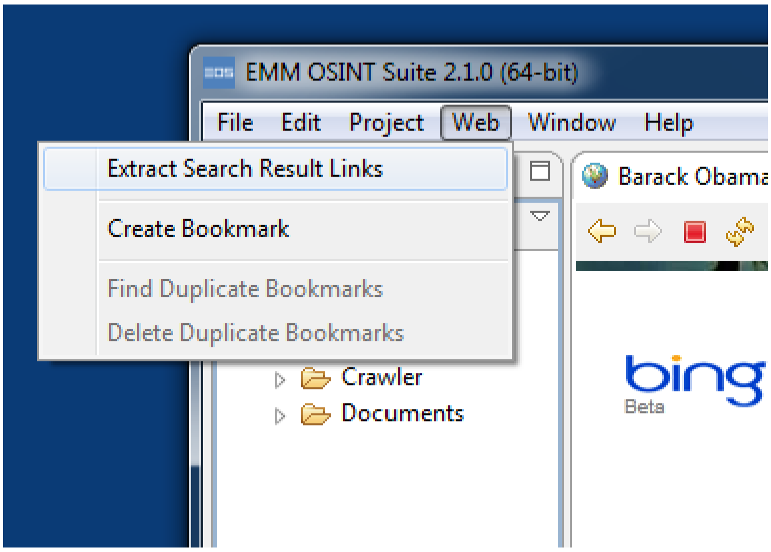

EMM OSINT Suite 2.1.0 (64-bit)
File Edit Project Web Window Help
Wo
En
Bookmark Example
Analysis
Bookmarks
Crawler
Documents
Barack Obama - Bing
&qs=n&form=QBLH&pq=barack+obama&sc=8-12&sp=-1&sk=&filt=custom
WEB BILDER VIDEOS SHOPPING NEWS MEHR
bing Beta
Barack Obama
26.100.000 ERGEBNISSE
Benutzerdefiniert
Einschränken n

Barack Obama Diese Seite übersetzen
www.barackobama.com
Official re-election campaign website of President Barack Obama p
updates, election news, videos, local events and ways to volunteer a

Barack Obama Diese Seite übersetzen
https://donate.barackobama.com
The next big step for our grassroots movement is being launched: O

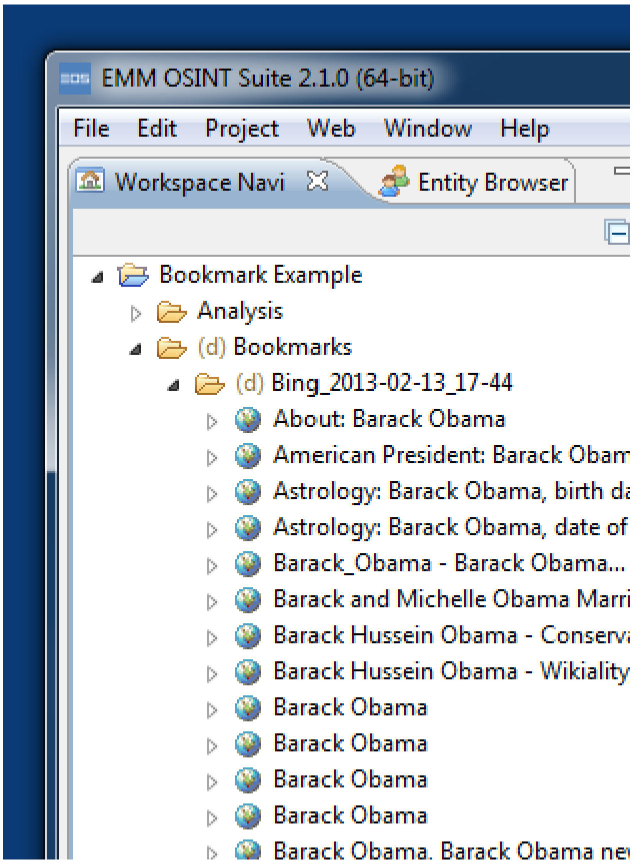
Bilder von Barack Obama
bing.com/images

- Select from the main menu Web > Extract Search Result Links to extract the result links of the search and to store them as bookmarks in the project:


EMM OSINT Suite 2.1.0 (64-bit)
File Edit Project Web Window Help
Extract Search Result Links
Create Bookmark
Find Duplicate Bookmarks
Delete Duplicate Bookmarks
Crawler
Documents
Barack Obama
bing Beta

The system extracts the first 100 result links (you can change the maximum number of extracted links under Window > Preferences > OSINT > Link Extraction). The links are stored in a newly created folder ("Bing_<Time Stamp>") under the bookmarks folder of the case project.

> *Each result link is stored in a bookmark file which is a XML file containing the following data: URL, Title, creation time stamp, search query, search engine). If you open the Properties View which is located next to the Console View (under the browser window) and select a bookmark file you can see the data contained in the file.*

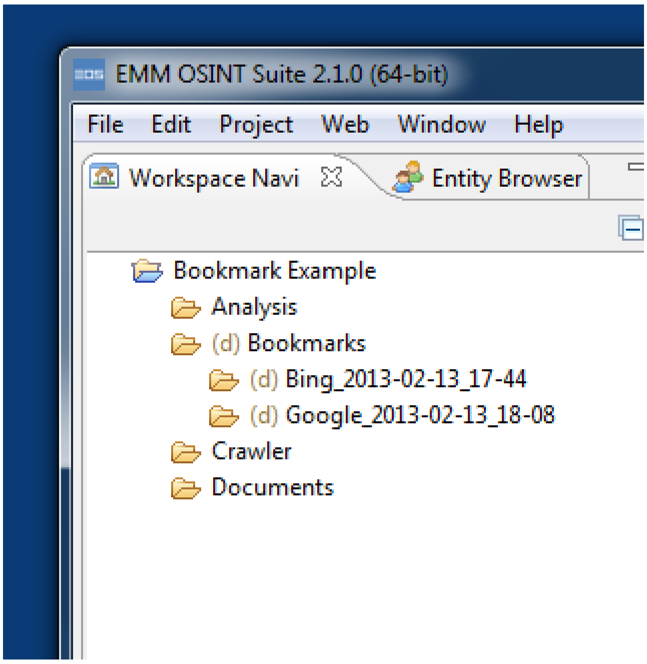

EMM OSINT Suite 2.1.0 (64-bit)
File Edit Project Web Window Help
Workspace Navi
Entity Browser

- Bookmark Example
  - Analysis
  - (d) Bookmarks
    - (d) Bing_2013-02-13_17-44
      - About: Barack Obama
      - American President: Barack Obam
      - Astrology: Barack Obama, birth da
      - Astrology: Barack Obama, date of
      - Barack_Obama - Barack Obama...
      - Barack and Michelle Obama Marri
      - Barack Hussein Obama - Conserva
      - Barack Hussein Obama - Wikiality
      - Barack Obama
      - Barack Obama
      - Barack Obama
      - Barack Obama
      - Barack Obama, Barack Obama ne

In this example the system has already detected a duplicate bookmark which is marked with a (d) in front of its title.

> *By default the system shows the title of the bookmark file and not the file name. You can change this behaviour if you click on the small triangle in the upper right corner of the workspace navigator view and select "Customize View". Then change to the Content tab and switch off "Bookmark Title", confirm with "OK".*

Now we search again but we use Google as search engine. As the end result we have two sub folders in our bookmarks folder:


EMM OSINT Suite 2.1.0 (64-bit)
File Edit Project Web Window Help
Workspace Navi
Entity Browser

- Bookmark Example
  - Analysis
  - (d) Bookmarks
    - (d) Bing_2013-02-13_17-44
    - (d) Google_2013-02-13_18-08
  - Crawler
  - Documents

The first folder contains the Bing search results, the second one contains the Google search results. Before we download the pages behind these search results, we want to eliminate duplicate bookmarks.

> ⓘ **Duplicate Bookmark**
> A bookmark file is considered a duplicate of another bookmark, if it contains the same URL and has a newer creation time stamp.

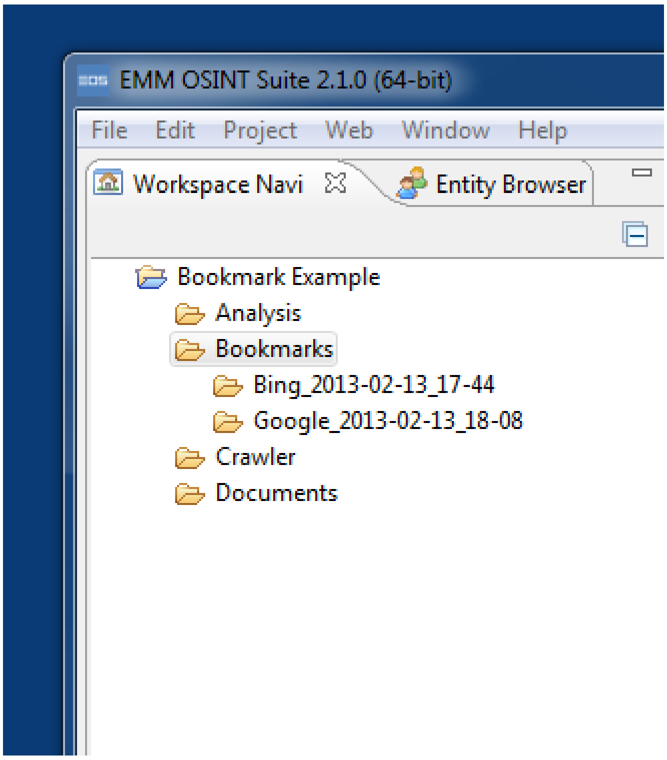
Organizing Bookmarks

In order to organize the bookmark files and delete duplicate bookmarks, do the following:

- Click on the Bookmarks folder in the Workspace Navigator
- Right click and select Organize Bookmarks > Delete Duplicates

The deletion is only performed on the selected folder and its sub folders. It does not affect bookmarks stored above the selected folder.

The system now deletes all duplicate bookmarks from the Bookmarks folder and shows the results in the Console view. The markup of duplicate bookmarks dissapears from the bookmarks folder and its sub folders:



The duplicate bookmark detection is also helpful if you want to repeat a search over time (for example each week) and see which search results are new. If you simply delete the duplicates only the new results remain.

Tasks

Crawling a Web Site

Prerequisite: Creating a Case Project

See also: Setting HTTP Proxy Information

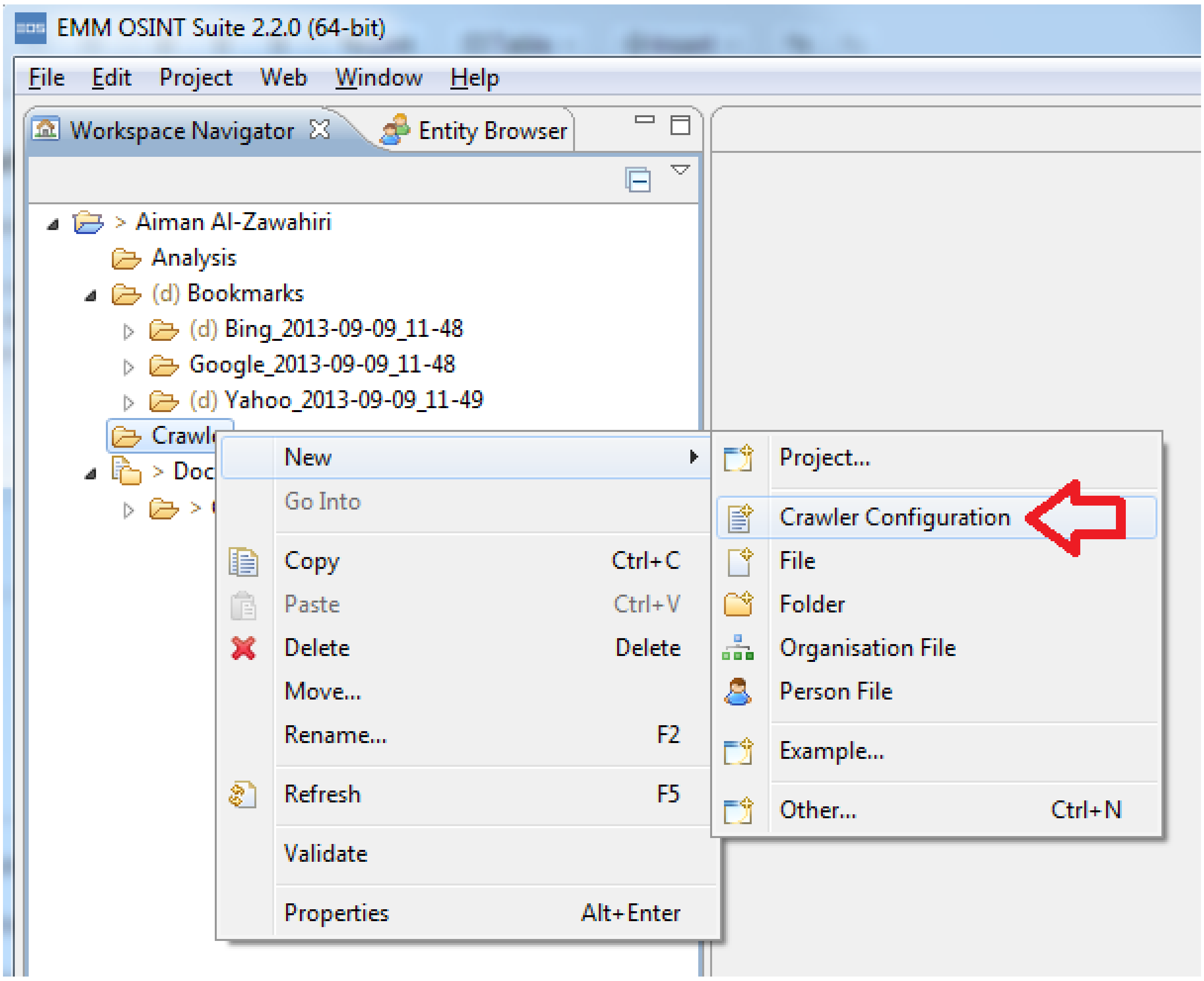
The application contains a crawler component which can be used to crawl (often also called "spider") a targeted web site.

The crawler component starts at a set URL and then follows the links on this web site until a predefined depth has reached.

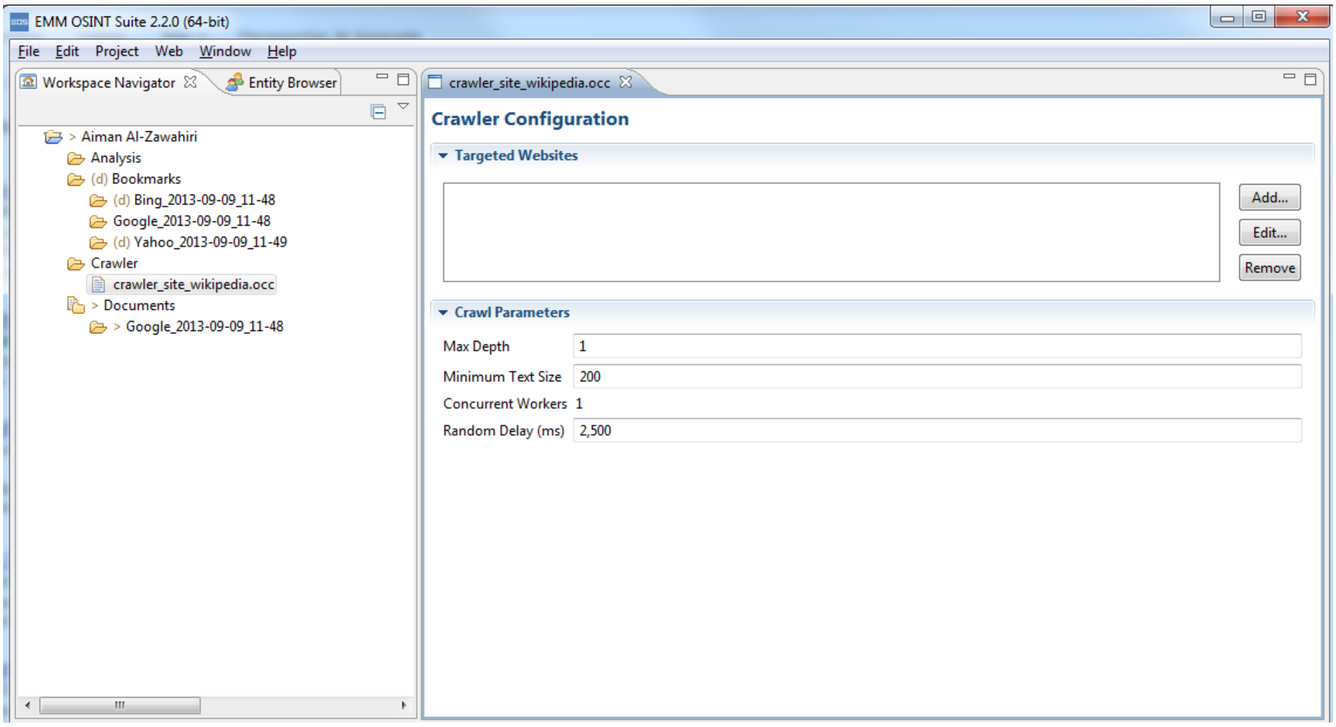
Creating a Crawler Configuration

The crawler component needs a configuration file which defines the starting URL and some parameters. To create one, do the following:

- In the **Workspace Navigator** view expand the *Crawler* folder and right-click the Crawler folder, then click **New > Crawler Configuration**. An **OSINT Crawler Configuration** creation dialog opens.



- In the **OSINT Crawler Configuration** dialog enter a file name and click **Finish**.The configuration file is created and opened in a Crawler Configuration editor.


EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help
Workspace Navigator
Entity Browser
> Aiman Al-Zawahiri
Analysis
(d) Bookmarks
(d) Bing_2013-09-09_11-48
Google_2013-09-09_11-48
(d) Yahoo_2013-09-09_11-49
Crawler
crawler_site_wikipedia.occ
> Documents
> Google_2013-09-09_11-48
crawler_site_wikipedia.occ
Add...
Edit...
Remove

Crawler Configuration

Targeted Websites

Crawl Parameters

Max Depth 1

Minimum Text Size 200

Concurrent Workers 1

Random Delay (ms) 2,500

In the Crawler Configuration Editor the following parameters can be set:

Name	Type	Description
Targeted Websites	list of URLs	The list allows you to add start URLs of web sites which should be crawled.

Max Depth	number	The maxium number of links to follow, default is 1 (crawling all pages connected to the start URL)
Minimum Text Size	number	The minimum extracted text size, default is 200 characters. If a page has less text, it will be ignored.
Concurrent Workers	number	The number of concurrent worker threads doing downloads, default is 1. This should be set to a very low number to avoid being black listed.
Random Delay (ms)	number	The waiting time between requests done by the worker threads. Default is 2500 milliseconds.

Adding a Web Site for crawling

In the Crawler Configuration editor do the following:

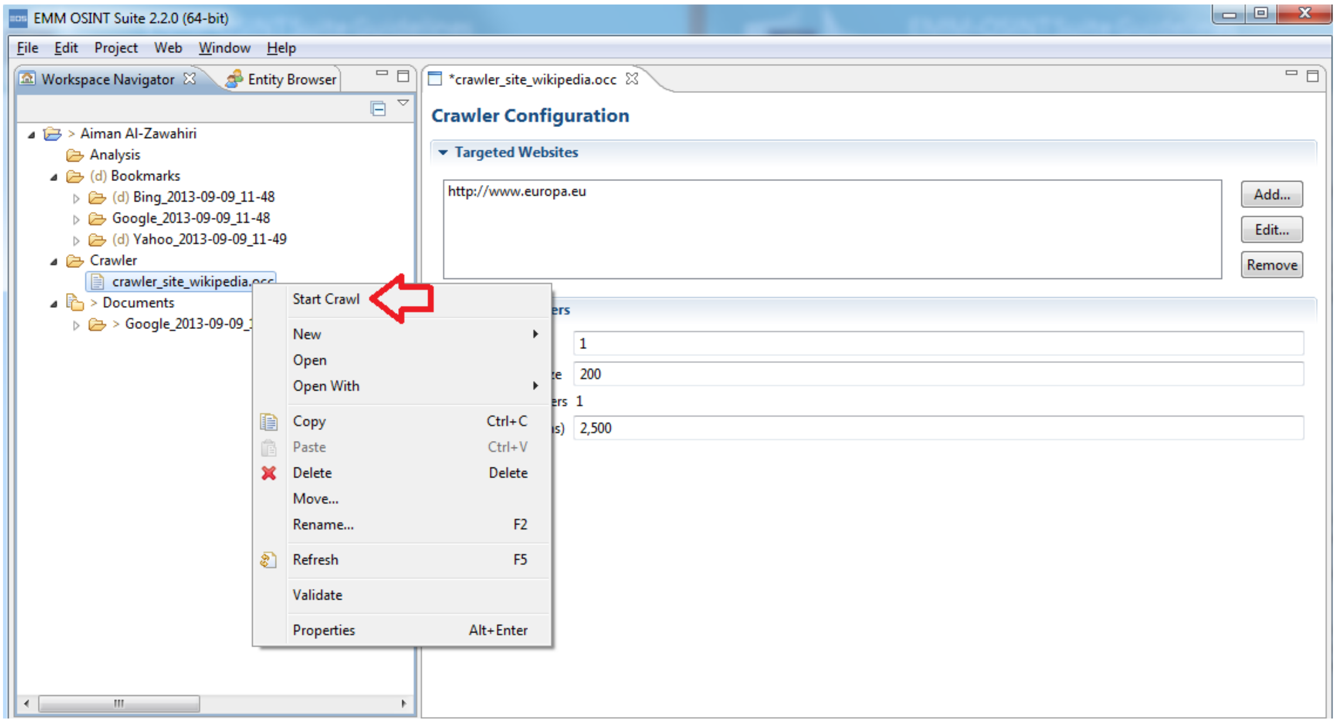
1. In the **Targeted Websites** section click **Add**. The **Add a new URL** dialog opens.
2. In the **Add a new URL** dialog enter a full URL, such as *http://www.europa.eu* and click OK. The added URL appears in the **Targeted Websites** list box.

To save the Crawler Configuration Editor use the main menu and click **File > Save**.

Performing a Crawl

To perform the crawl using the Crawler Configuration file created do the following:

1. Right-click on the Crawler Configuration file and click **Start Crawl**.



http://www.europa.eu

1

200

1

2,500

Crawling Speed
Using the default settings, the crawling is rather slow. This avoids being "black listed" on the target site.

The crawler module starts in the background, it creates for all crawled pages Bookmark files in the Bookmarks folder.

In order to download the pages, please refer to Downloading of Bookmarks

Downloading Speed

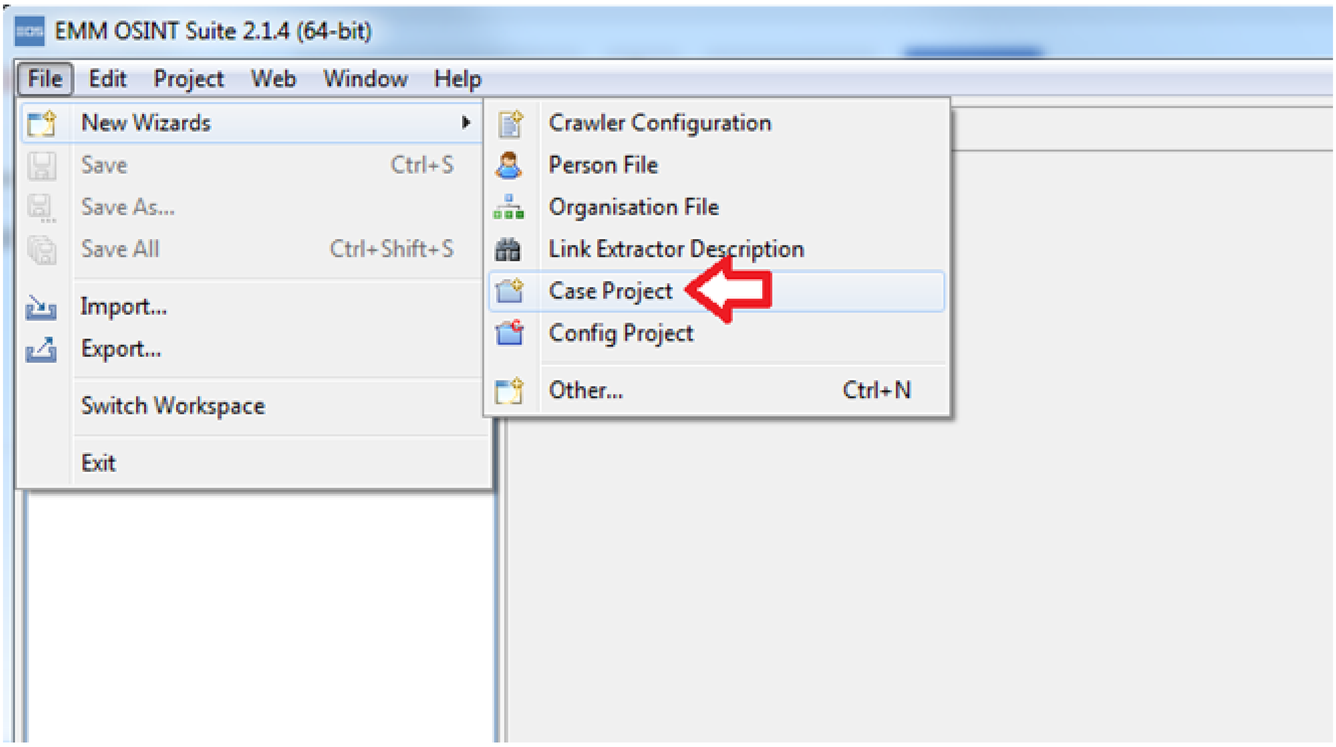
Like the crawling also the download of pages refered to by the Bookmark files is done in a slow fashion.

Creating a Case Project

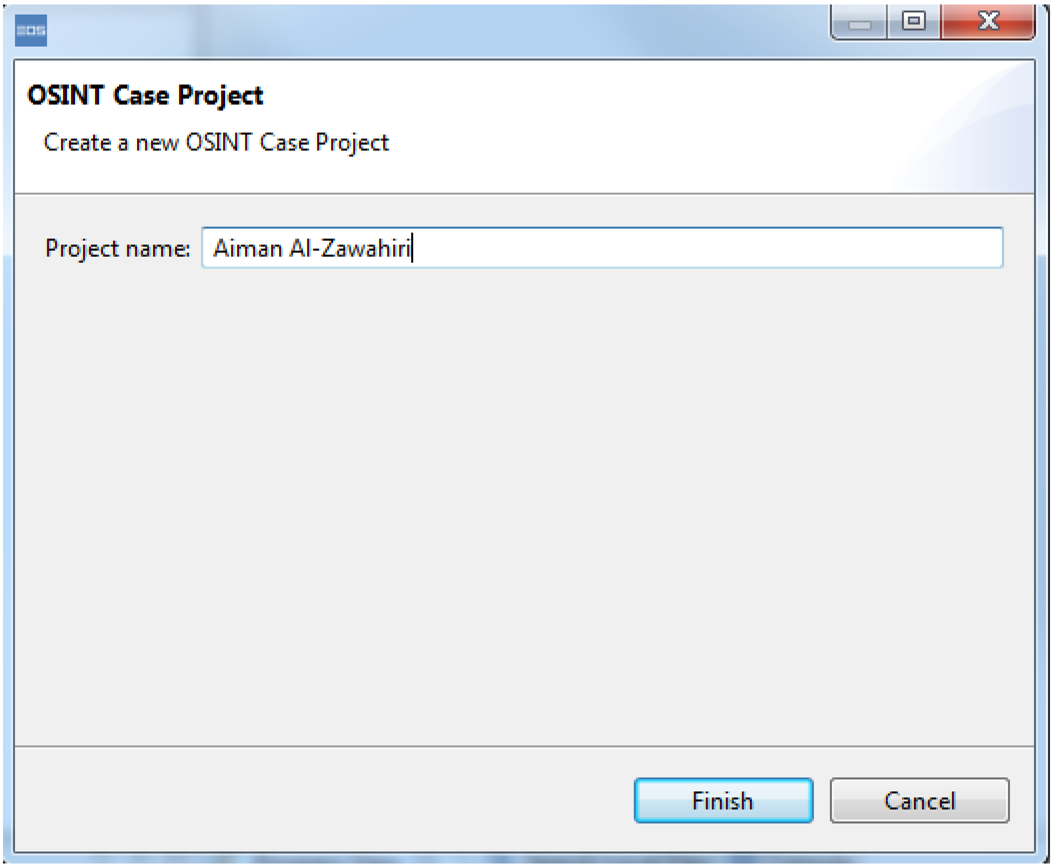
A case project is a type of top level folder in your workspaces which contains data about a single investigation. It consists of a predefined structure of sub-folders to store various files of the investigation.

To create a case project, perform the following steps:

- In the main menu click **File > New Wizards > Case Project**. A project creation dialog is opened.



- Enter the name of the project in the **Project name** field.



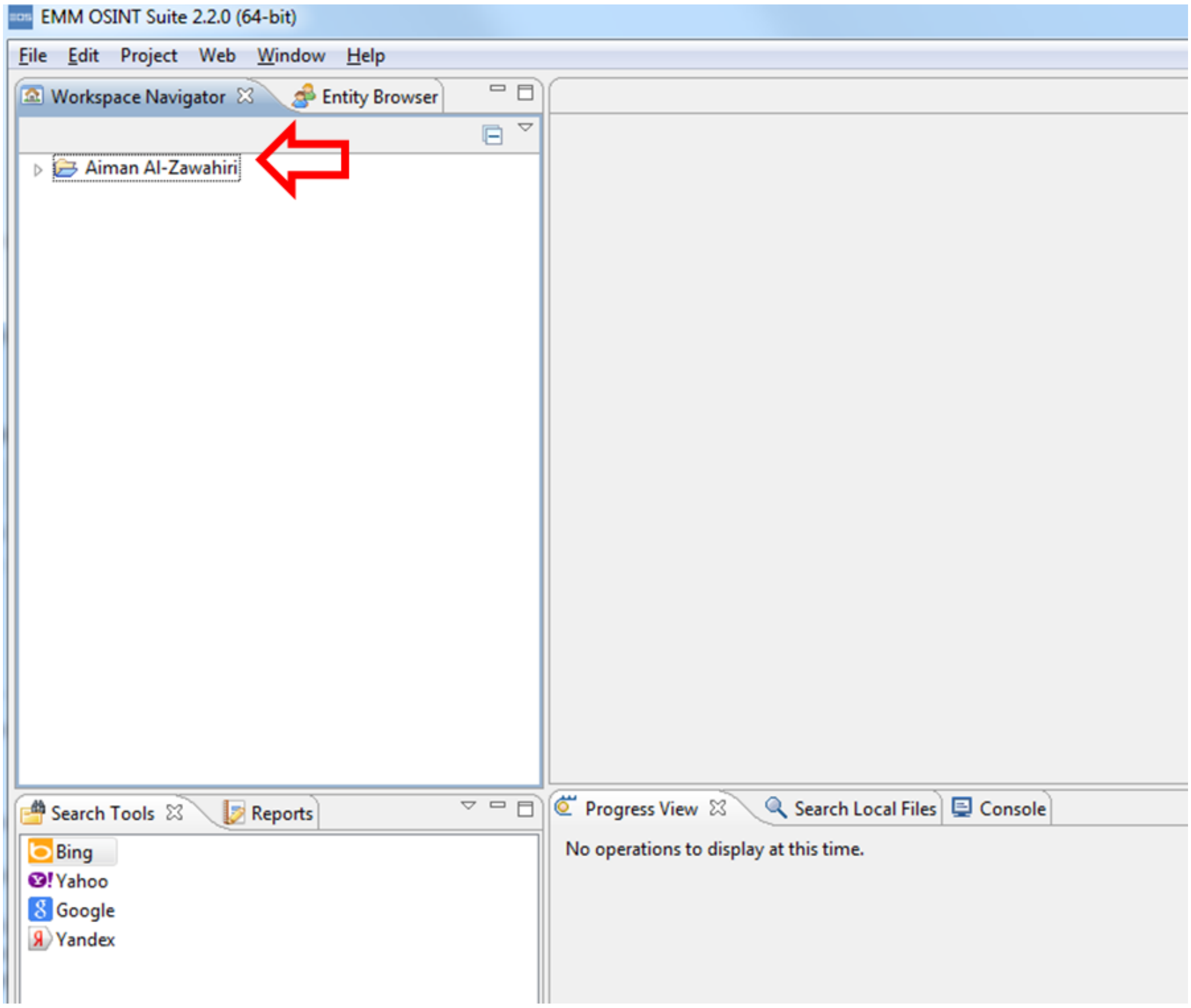
OSINT Case Project

Create a new OSINT Case Project

Project name: Aiman Al-Zawahiri

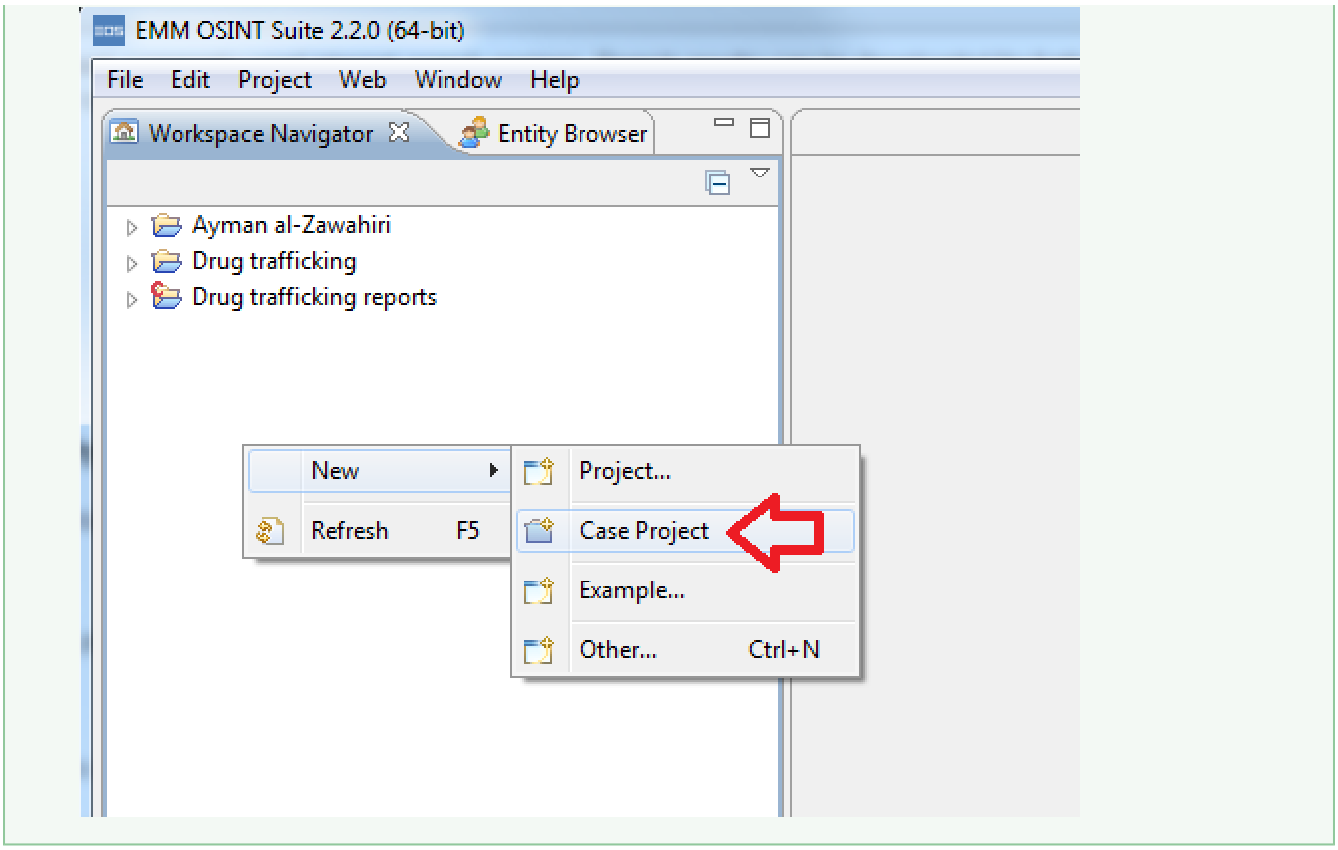
- Click **Finish** to create the project in the workspace.

After the project creation dialog has closed the new case project appears in the **Workspace Navigator** view:



No operations to display at this time.

✓ Instead of opening the creation dialog from the main menu (step 1) you can also right-click in the **Workspace Navigator** view and click **New > Case Project**

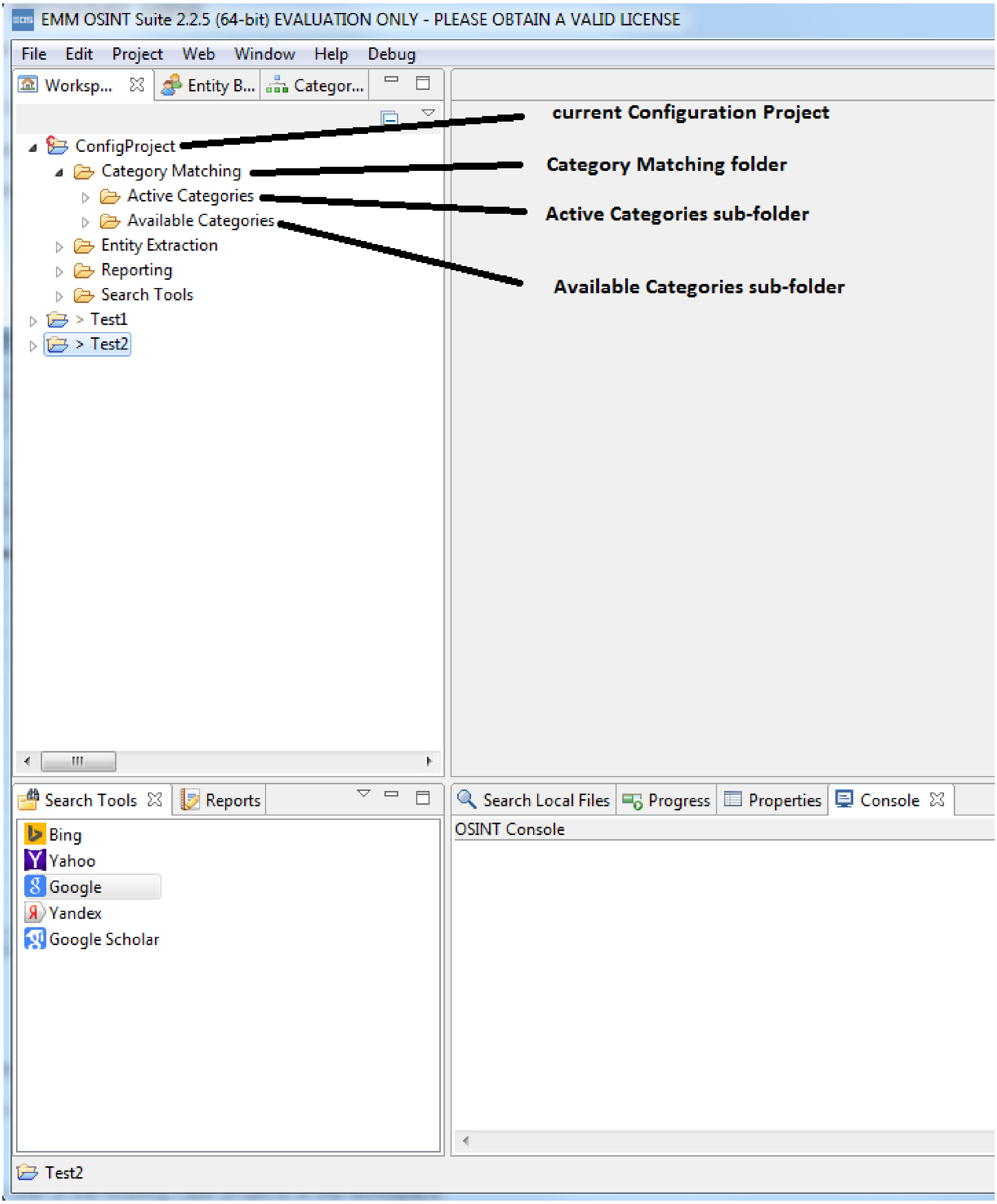


Creating a Category Definition File

Prerequisite: Creating a Configuration Project

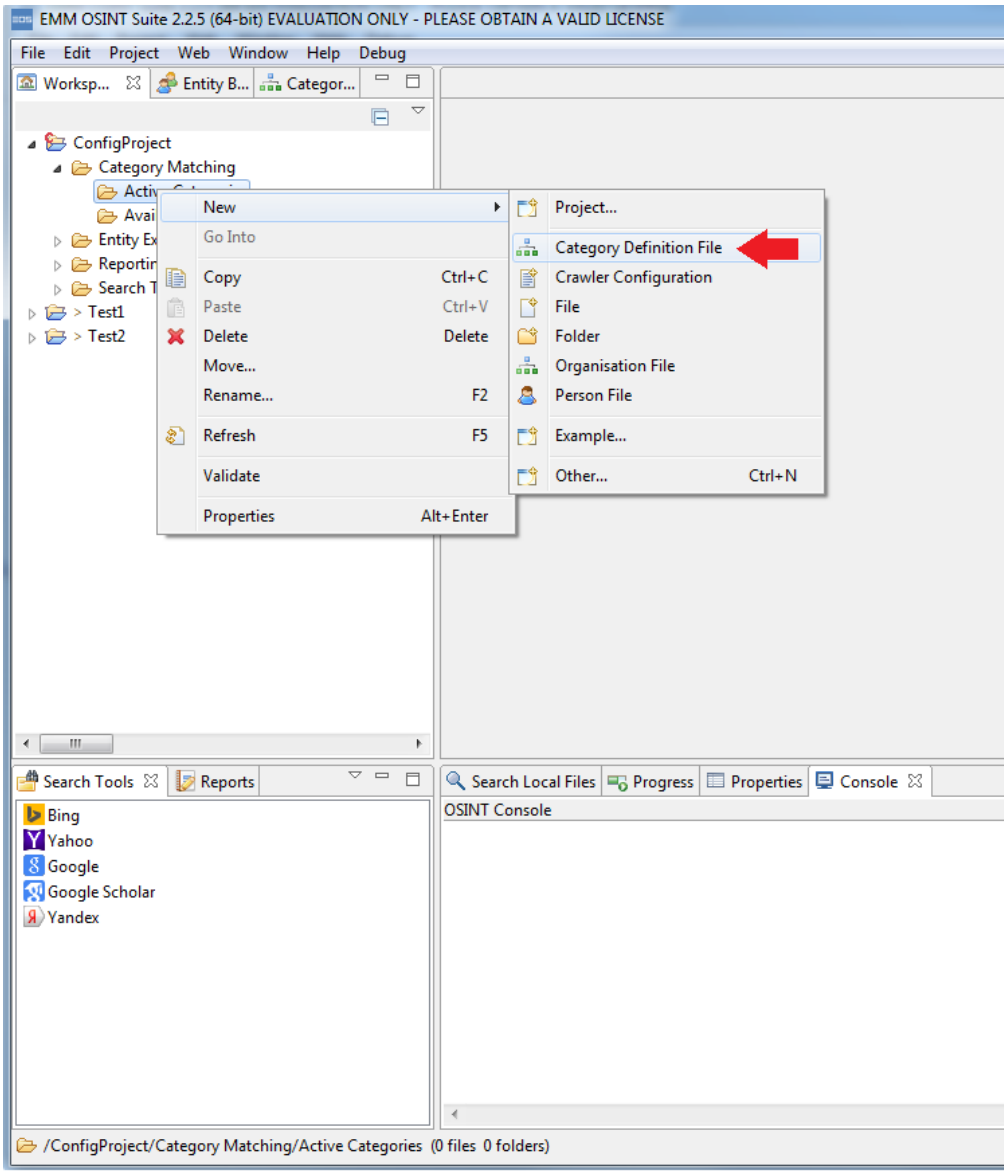
A **Category Definition File** is a file type used to define a category or alert within the OSINT Suite. This type of file should be created under the **Category Matching folder** that can be found within a Configuration Project. Once a Configuration Project is created, two different sub-folders appear within the Category Matching folder:

- **Active Categories** sub-folder: contains the active categories that can be currently selected in the category selector shown in the Category Browser view.
- **Available Categories** sub-folder: contains other category definition files that can be used later as active categories. The categories defined under this sub-folder are not available in the category selector of the Category Browser view.

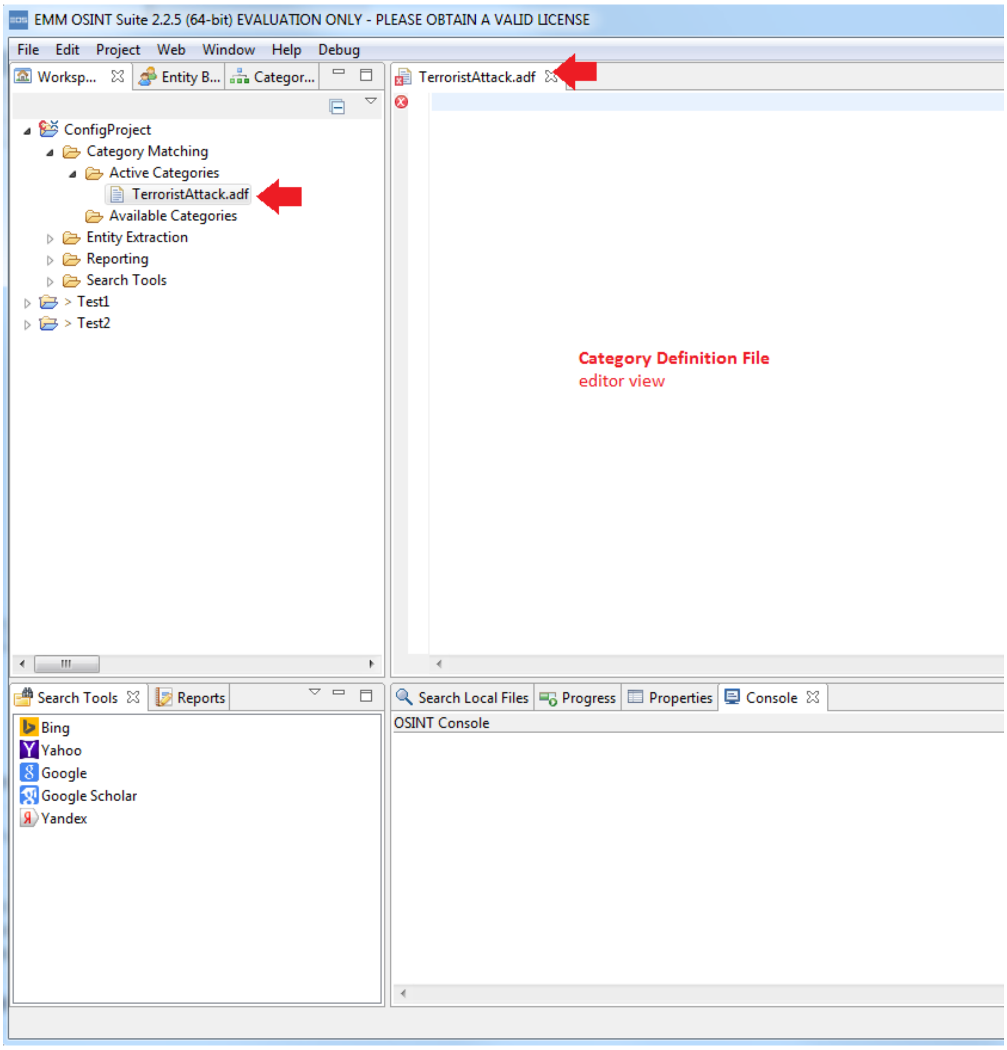

EMM OSINT Suite 2.2.5 (64-bit) EVALUATION ONLY - PLEASE OBTAIN A VALID LICENSE
File Edit Project Web Window Help Debug
Worksp... Entity B... Categor...
ConfigProject
Category Matching
Active Categories
Available Categories
Entity Extraction
Reporting
Search Tools
> Test1
> Test2
current Configuration Project
Category Matching folder
Active Categories sub-folder
Available Categories sub-folder
Search Tools Reports
Bing
Yahoo
Google
Yandex
Google Scholar
Search Local Files Progress Properties Console
OSINT Console
Test2

To create a **Category Definition File** do the following:

- **Right click** on one of the sub-folders mentioned above (**Active Categories** or **Available Categories**) and click on **New > Category Definition File**. Choose a file name and click on Finish.



- The **Category Definition File editor** is automatically opened showing the empty content of the new file. The file extension used for Category Definition files is **adf** (alert definition file), as shown in the workspace navigator:



> The OSINT Suite also allows using category definition files generated by the official EMM Alert Editor (http://emm.jrc.it/AlertEditor/). Files generated by the EMM Alert Editor are basically XML files that conform the XML Schema Definition (XSD) provided by the EMM team to define categories or alerts. The alert XSD is available at http://emm.jrc.org/alert.xsd

To know how to define categories or alerts within a specific category definition file see Defining a Category.

Creating a Configuration Project

A configuration project is a type of top level folder in your workspace which contains configuration data used by various tools of the system. It consists of a pre-defined folder structure which contains the various configuration files for different tools.

To create a onfiguration project, perform the following steps:

- In the main menu click **File > New Wizards > Config Project**. A project creation dialog is opened.
- Enter the name of the project in the **Project name** field, preferably name it "*Configuration*"
- Click **Finish**. The configuration project is created in the workspace.

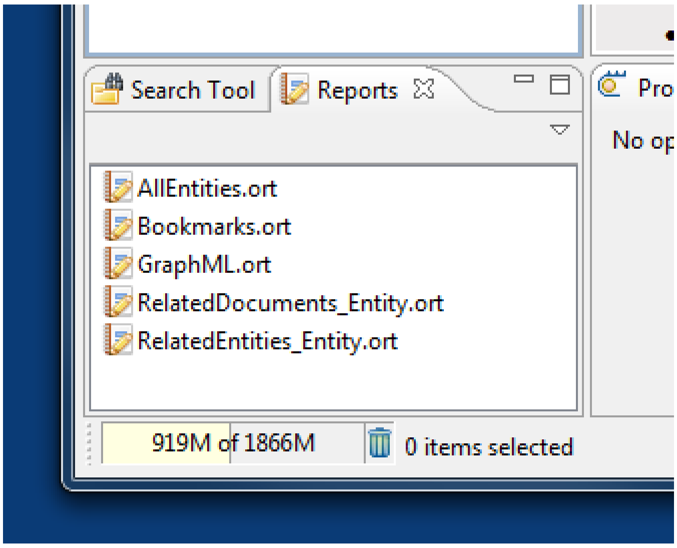
The configuration project appears as a top level folder in the workspace. Its icon is marked with a small red "c" to show that this it is a configuration project.

> **Tip**
> A configuration project contains templates and configuration information which is needed by different modules of the system. These modules use the settings and templates to process data in Case Projects. Even though, it is possible to create more than one configuration project (only one is active at any given time as defined in the Preferences), it is recommend to create only **one** c onfiguration project to keep things simple.

Creating a Custom Report

Reports are a way to export analysis data from a Case Project into flat files. These files can be HTML documents or machine readable formats such as tsv files (tab separated value - to be imported into MS Excel).

If you open the application, you see in the lower left corner a **Reports** view which may be hidden behind the **Search Tool** view:



This view shows the currently available reports in the system. For example, if you double-click on *AllEntities.ort* report, the system creates a HTML report in the current project which shows a list of all found entities.

However, some reports need to be customised for your project in order to perform as desired or you need to create a custom report from scratch. This tutorial gives an introduction on how to create a custom report.

The following tasks are described

- Setting up an example case project
- Creating a configuration project
- Modifying the report "RelatedDocuments_Entity.ort" to list documents related to "Barack Obama"
- Modifying the report and its output template to list documents related to "Barack Obama" and similar named entities

Setting up an example Case Project

To have some data for this tutorial we set up an example project about "Barack Obama". If you have already a Case Project where you need to change the reports, please skip this step.

Procedure:

- Create a Case Project named "Barack Obama"
- Search on Google about "Barack Obama"
- Download the bookmarks we gathered from the Google search result pages
- Run the entity extraction to generate analysis data

See also: Quick Start Guide for descriptions of the individual tasks.

Creating a Configuration Project

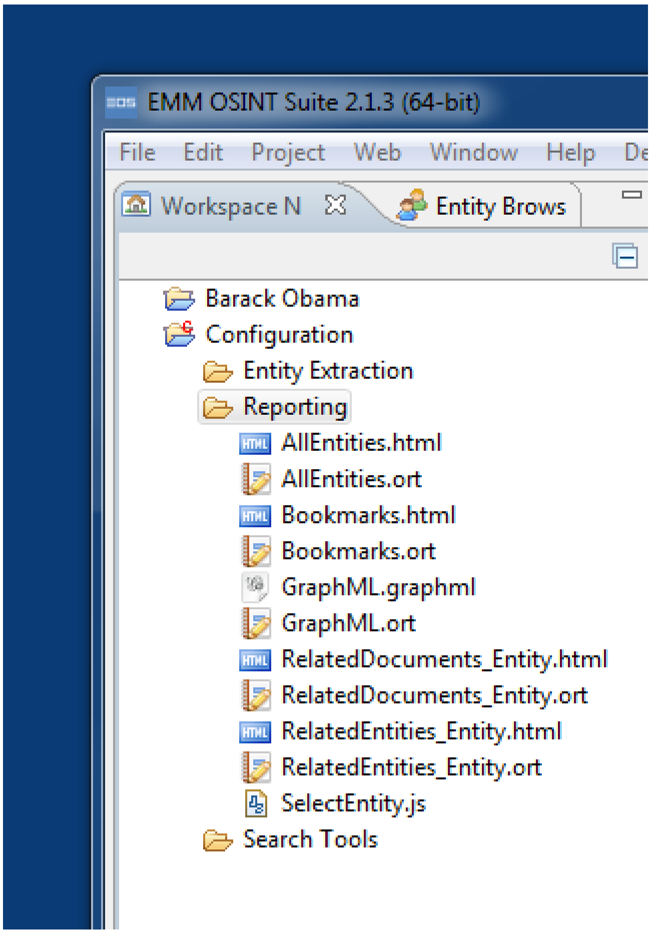
The first step is to create a configuration project, which allows you to modify the existing report templates (.ort stands for OSINT Report Template) or create new ones.

In order to create a configuration project do the following:

- In the main menu click on File > New Wizards > Configuration Project
- Fill in "Configuration" as project name and click on "Finish"

The system now creates a new project which is marked with a small red "c" to show that this is not a normal Case Project but rather contains configuration data and templates.

The following sreenshot shows this configuration project with the "Reporting" folder expanded:



EMM OSINT Suite 2.1.3 (64-bit)
File Edit Project Web Window Help De
Workspace N
Entity Brows

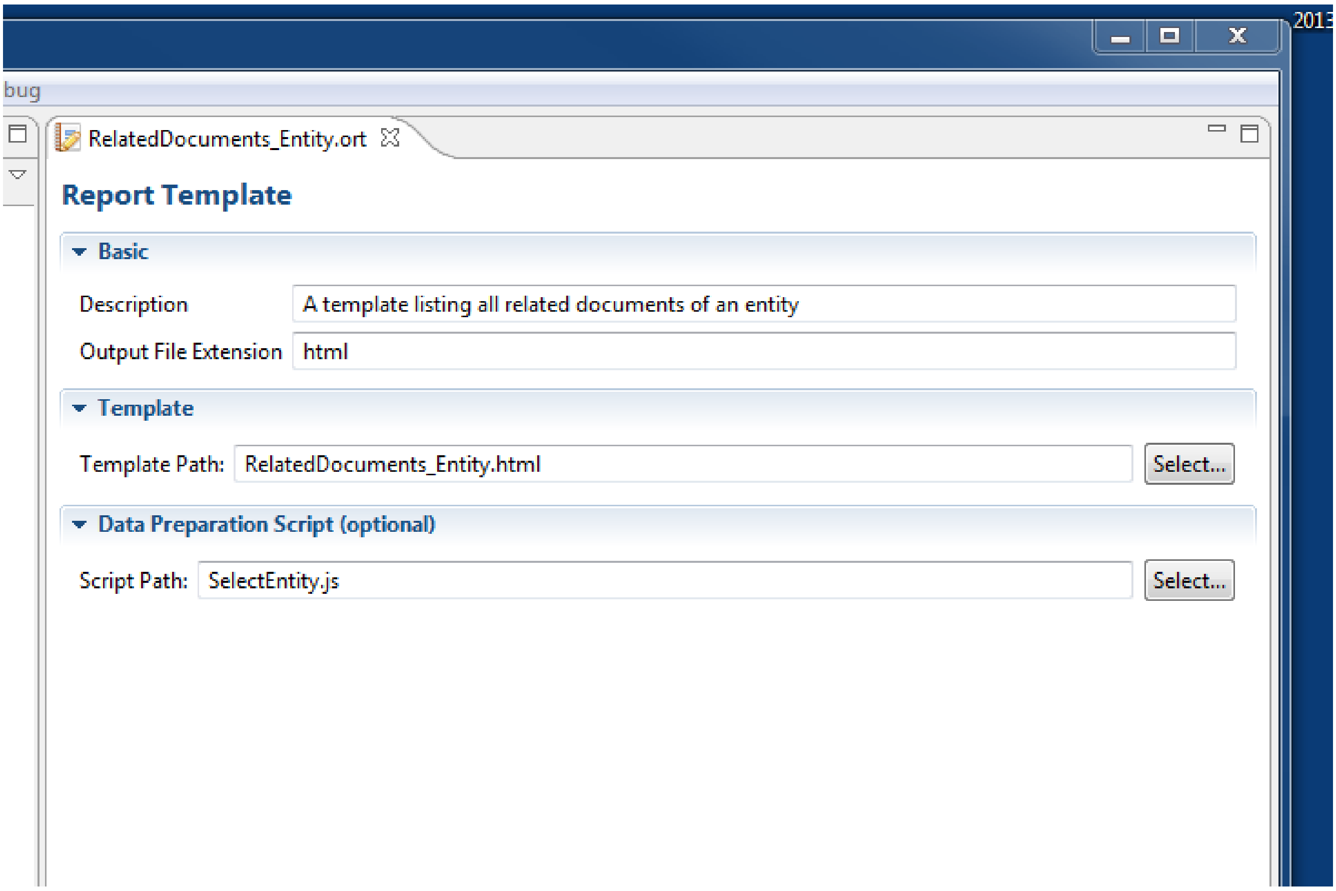
Barack Obama
- Configuration
  - Entity Extraction
  - Reporting
    - AllEntities.html
    - AllEntities.ort
    - Bookmarks.html
    - Bookmarks.ort
    - GraphML.graphml
    - GraphML.ort
    - RelatedDocuments_Entity.html
    - RelatedDocuments_Entity.ort
    - RelatedEntities_Entity.html
    - RelatedEntities_Entity.ort
    - SelectEntity.js
  - Search Tools

> *A configuration project contains templates and configurations which apply to all case projects in the system. It is possible to create more than one configuration project in this case only one is active. (You can choose the active one from the Preferences). However, in practice we recommend to create only one configuration project.*

Open a report template for editing

Inside the "Reporting" folder you see different types of files. There are files with the *.ort ending which are the report template files for individual reports. A report template contains definitions for the system to generate a report. These available report template files are also accessible from the Reports View in the user interface. The other files are used from within the *.ort files and define output templates (for example to create HTML output) and data selection scripts used by the system to create the final report.

The system provides a report template editor to edit the definitions of the template. We open the "RelatedDocuments_Entity.ort" file by performing a double click on it. It opens in an editor window:



bug

RelatedDocuments_Entity.ort

Report Template

Basic

Description: A template listing all related documents of an entity

Output File Extension: html

Template

Template Path: RelatedDocuments_Entity.html

Data Preparation Script (optional)

Script Path: SelectEntity.js

The fields defines the needed template file and optional data preparation script to compile a report:

Field	Description
Description	Gives a short descriptions what kind of report is generated
Output File Extension	Defines which file extension the generated report will have. This can be for example html if the report creates a html file or csv for a file readable by MS Excel.
Template Path	The template which is used to generate the report. During report generation the template is filled with the analysis data from the selected case project.
Script Path	An option javascript file which can be used to pre-process the data before filling it into the template. For example, a list of entities could be sorted or filtered before it is used in the template. Note: The scripting will change in a future version of the software, use it as little as possible.

The default report template creates a report showing all documents related to "Franz Marc" (which is the example name we use in the quick start guide). We want to change this and generate a report which shows all documents related to "Barack Obama".

Create a custom report based on an existing example

The first step to create a custom report for "Barack Obama" is to copy the existing RelatedDocuments_Entity.ort file to a new file:

- Close the editor window showing RelatedDocuments_Entity.ort
- Right click on "RelatedDocuments_Entity.ort" in the Workspace Navigator View and select "Copy"
- Right click on the "Reporting" folder and select "Paste"
- The system opens a dialog, we change the name of the copied file to "RelatedDocuments_BarackObama.ort"

Now, as second step we do the same for the javascript file belonging to the report template:

- Right click on "SelectEntity.js" in the Workspace Navigator View and select "Copy"
- Right click on the "Reporting" folder and select "Paste"
- The system opens a dialog, we change the name of the copied file to "SelectBarackObama.js"

The javascript is used to select the needed data for the report. It defines that our main entity is "Barack Obama". (In the next major version of the software this will be done by using a data selection dialog.)

As a third step, we edit the new report template file (double click on "RelatedDocuments_BarackObama.ort" and insert the following data:

Field	Input
Description	Generates a list of documents where Barack Obama is found
Output File Extension	html
Template Path	RelatedDocuments_Entity.html
Script Path	SelectBarackObama.js

We save the report template file (Main Menu > File > Save).

As a final step, we modify the javascript file "SelectBarackObama.js" to select Barack Obama as main entity for our report:

- Double click on "SelectBarackObama.js"

Now, the javascript editor opens and we change the script to the following:

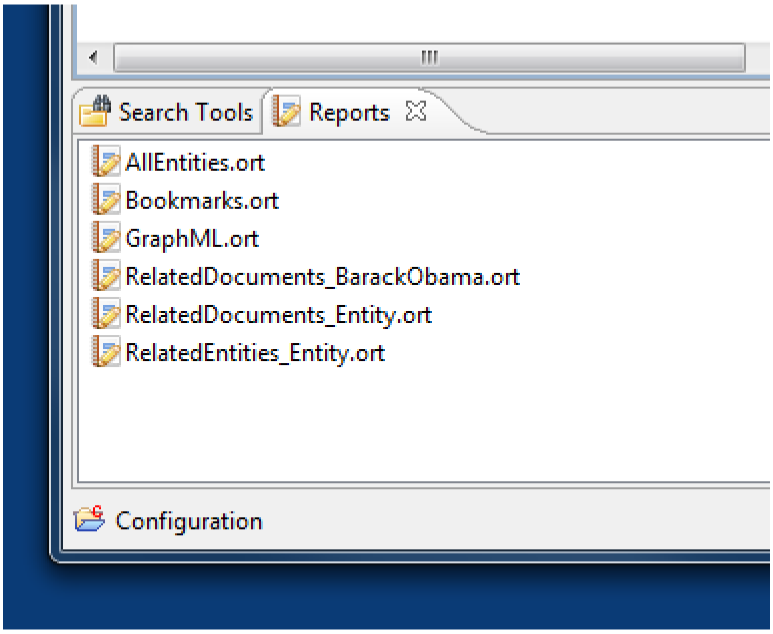
Select Entity

```
/**
 * JavaScript to select an entity and store it into context of
* report template.
*/
var selectedEntity = project.getEntityByName("Barack Obama");
templateContext.put("selectedEntity", selectedEntity);
```

This selects the entity named "Barack Obama" from the selected project and stores it into the template context. During generation the system replaces the $selectedEntity variable in the html template "RelatedDocuments_Entity.html" with the entity for Barack Obama.

After saving the javascript file, we can test it by doing a right mouse click on "RelatedDocuments_BarackObama.ort" and selecting "Generate Report".

The new custom report also shows up in the Reports View:


Search Tools
Reports
AllEntities.ort
Bookmarks.ort
GraphML.ort
RelatedDocuments_BarackObama.ort
RelatedDocuments_Entity.ort
RelatedEntities_Entity.ort
Configuration

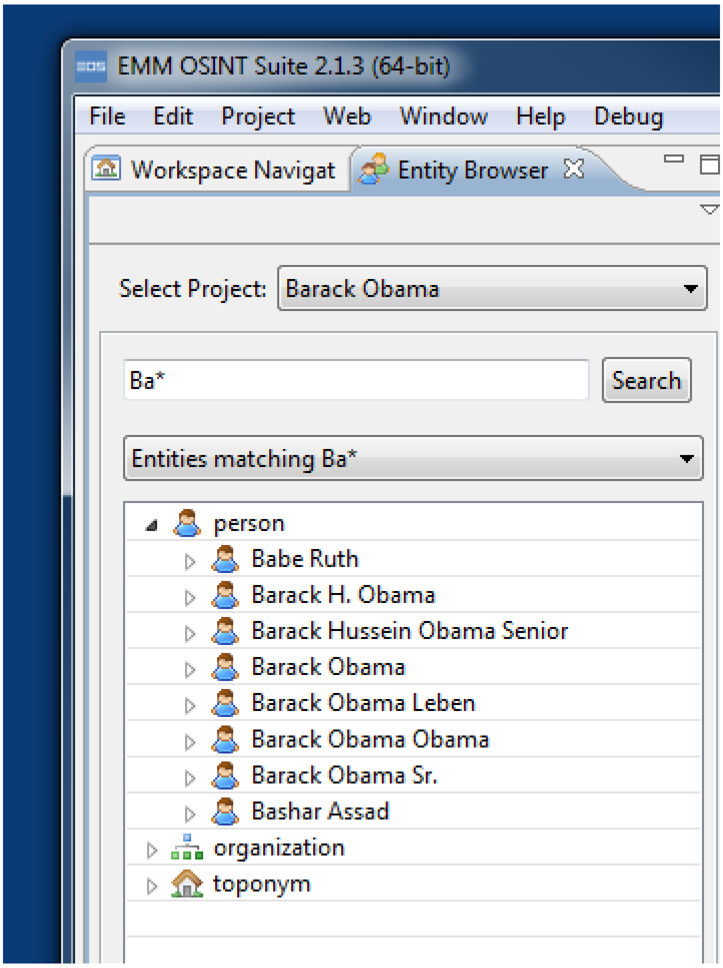
> ***Note:*** *This complicated procedure to edit the javascript file manually to select data will go away in the next major version of the software. Instead a selection dialog will be used (similiar to the one to select the source project) to select the entity (or other data) for the report*

(Advanced topic) Improve the report to include similar entities

Our custom report uses the "RelatedDocuments_Entity.html" output template to embed the project's entity data into a HTML page. Now we want to improve our report with the following goal:

- List all documents related to the entity named "Barack Obama" and *similar named entities*

The entity extraction engine has a normalisation step which tries to match similar names and combine these name variants to belong to a single entity. This way we avoid to have too many entities which represent the same person. However, there is a limit for the system to decide which names are similar enough. If we look at the entity browser view of our application we see that there are quite a few entities listed which represent "Barack Obama":


EMM OSINT Suite 2.1.3 (64-bit)
File Edit Project Web Window Help Debug
Workspace Navigat
Entity Browser
Search

Select Project: Barack Obama

Ba*

Entities matching Ba*

- person
  - Babe Ruth
  - Barack H. Obama
  - Barack Hussein Obama Senior
  - Barack Obama
  - Barack Obama Leben
  - Barack Obama Obama
  - Barack Obama Sr.
  - Bashar Assad
- organization
- toponym

We now want to improve our report to include also documents which contain one of the different entities above. (Please note, that by editing the name variant database of the software, we can avoid these different entities about the same person. How to do this will be covered in a different tutorial).

In order to do so, we need to do the following:

- Change the data selection script "SelectBarackObama.js" to obtain a list of entities with names starting with "Bar"
- Change the output template "RelatedDocuments_Entity.html" in a way that it lists documents of a list of entities not of a single one

Select a list of entities representing Barack Obama

The javascript file "SelectBarackObama.js" defines which data is avaiable to the report generator to include it in the final report. In a first step we modify it to include all entities with names starting with the letters "Bara". (We defined these four letters by looking at the entity browser view.)

We open the "SelectBarackObama.js" javascript from the configuration project by performing a double click on it:

Select List of Entities

```
/**
* JavaScript to select an entity and store it into context of report template.
*/
var selectedEntity = project.getEntityByName("Barack Obama");
templateContext.put("selectedEntity", selectedEntity);
```

We see that the ***selectedEntity*** variable is defined in a first step by calling the function property ***getEntityByName*** of the ***project*** object. This object is a predefined object which is the entry point into the analysis data of the source project for the report. The source project in turn is selected during report generation from a selection dialog.

Now, if we want to define a list of entities with names starting with "Bara". We look into the online documentation to find a suitable function of the ***project*** object which provides this list:

- From the main menu open Help > Help Contents > OSINT Suite User Guide > Reference > Reporting > Data Objects
- Review the table showing the object properties of the *project* object

The function property ***project.getEntitiesByNamePattern(namePattern)*** seems to be suitable for our purposes. It needs a regular expression pattern as parameter to match against the available entities in the system.

Now, we adapt the javascript code as follows:

Select Similar Entities

```
var selectedEntities = project.getEntitiesByNamePattern("Bara.*");
templateContext.put("selectedEntities", selectedEntities);
```

> Note: in order to match "Bara" as the start of an entity name we provided the pattern "Bara.*" which is a regular expression pattern. Soon, we will provide a tutorial to show you how to write these patterns to match text.

After selection of the entities, we store them in the templateContext which defines a set of data available to the report output template.

Adapt the output template to show documents relating to a list of entities

After changing the data selection script, we need to adapt the output template "RelatedDocuments_Entity.html" to show all documents related to a list of entities and not only to a single entity.

We do the following:

- Copy the RelatedDocuments_Entity.html to a new file:
  - Right click on "RelatedDocuments_Entity.html" in the Workspace Navigator View and select "Copy"
  - Right click on the "Reporting" folder and select "Paste"
  - The system opens a dialog, we change the name of the copied file to "RelatedDocuments_Entities.html"
- Adapt the report template file to use "RelatedDocuments_Entities.html" as output template instead of "RelatedDocuments_Entity.html"

Now, we open the new "RelatedDocuments_Entities.html" file in a text editor (right mouse click > Open With > Text Editor) and edit it.

Since we use internally Apache Velocity as templating engine (see the User Guide), the output template consists mainly of html code with variables starting with $ and directives starting with #:

HTML Template

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<title>Related Documents of Entity $selectedEntity.Name</title>
</head>
<body>
<h2>List of related documents of $selectedEntity.Name:</h2>
<ul>
#foreach( $doc in $selectedEntity.RelatedDocuments )
<li><a href="file://$doc.FilePath">$doc.Title</a></li>
#end
</ul>
</body>
</html>
```

The current output template has a *foreach* loop directives which loops of the list of all documents related to the *selectedEntity* (this is the variable we have defined in the javascript).

Now, in order to show all documents of a list of *selectedEntities* (see javascript above), we need to add another foreach loop directive to first loop over all entities and then internally to loop over all related documents. The resulting code of the output template looks like this (not the new #foreach directive):

HTML Template

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<title>Related Documents of a list of entities</title>
</head>
<body>
#foreach ($selectedEntity in $selectedEntities)
<h2>List of related documents of $selectedEntity.Name:</h2>
<ul>
#foreach( $doc in $selectedEntity.RelatedDocuments )
<li><a href="file://$doc.FilePath">$doc.Title</a></li>
#end
</ul>
<br/>
#end
</body>
</html>
```

After saving the new output template, we can test our new report by selecting it from the config project with the right mouse and choosing "Generate Report".

Please find all files of this tutorial attached to this page. You can simply unzip them to disk, then select from the main menu File > Import > Documents > File System and import them to your configuration project.

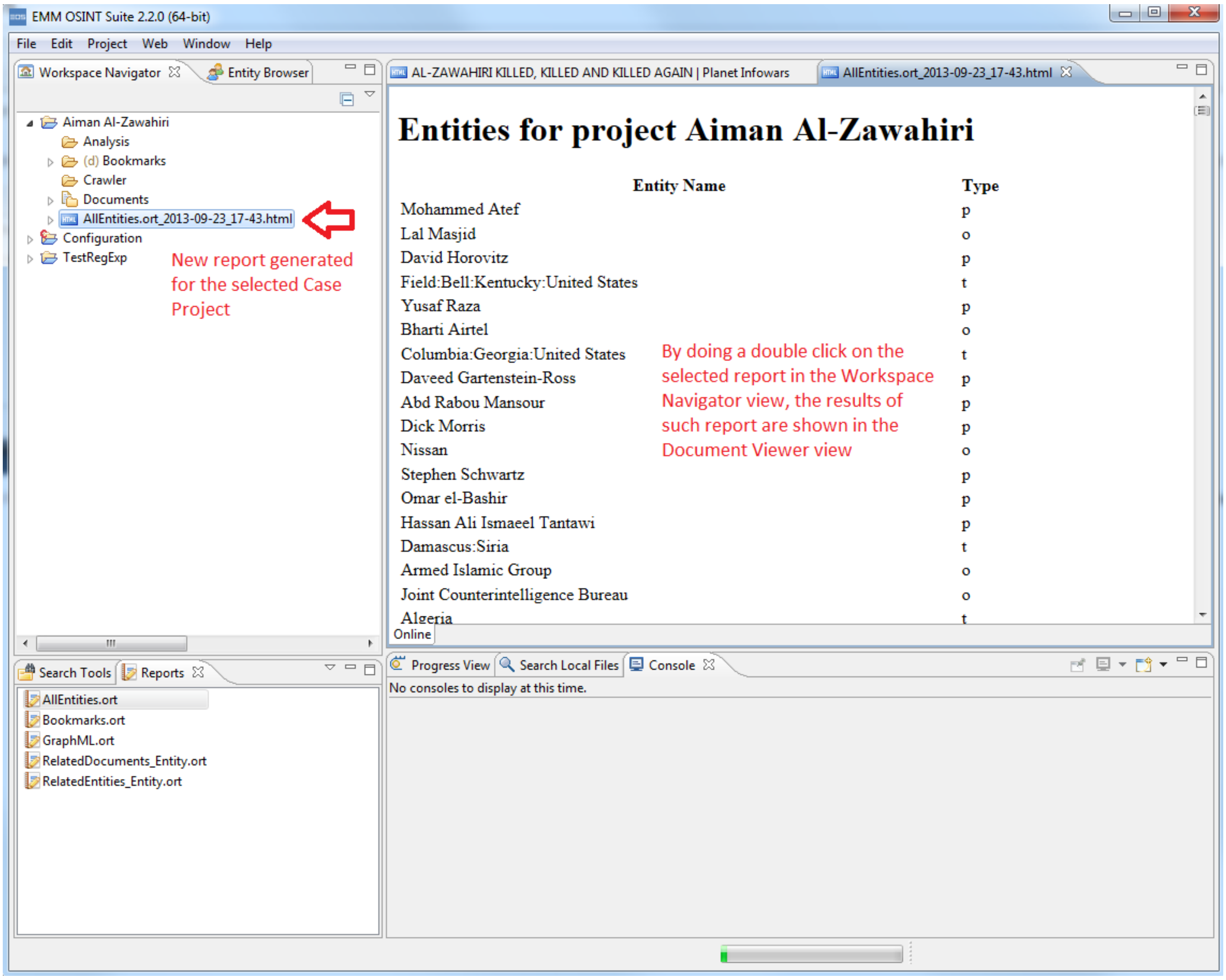
Generating a Report

To generate a report, perform the following steps:

- **Make sure the entity extraction has run**
- Double-click on one of the predefined report templates shown in the **Report Tools** view in the lower-left side of the application window. This view shows the currently available reports in the system. A dialog appears to select the source case project for the report.

? Unknown Attachment

- Once we have selected the desired Case Project from the list of available projects, the system creates the report and shows the progress in the **Progress View**. After the report has been generated it will be available from the **Workspace Navigator** view.


EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help
Workspace Navigator
Entity Browser
Aiman Al-Zawahiri
Analysis
(d) Bookmarks
Crawler
Documents
AllEntities.ort_2013-09-23_17-43.html
Configuration
TestRegExp
AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN | Planet Infowars
AllEntities.ort_2013-09-23_17-43.html
Online
Search Tools
Reports
AllEntities.ort
Bookmarks.ort
GraphML.ort
RelatedDocuments_Entity.ort
RelatedEntities_Entity.ort
Progress View
Search Local Files
Console

Entities for project Aiman Al-Zawahiri

Entity Name	Type
Mohammed Atef	p
Lal Masjid	o
David Horovitz	p
Field:Bell:Kentucky:United States	t
Yusaf Raza	p
Bharti Airtel	o
Columbia:Georgia:United States	t
Daveed Gartenstein-Ross	p
Abd Rabou Mansour	p
Dick Morris	p
Nissan	o
Stephen Schwartz	p
Omar el-Bashir	p
Hassan Ali Ismaeel Tantawi	p
Damascus:Siria	t
Armed Islamic Group	o
Joint Counterintelligence Bureau	o
Algeria	t

No consoles to display at this time.

New report generated for the selected Case Project

By doing a double click on the selected report in the Workspace Navigator view, the results of such report are shown in the Document Viewer view

In the image above, the results for the predefined report "AllEntities" are shown in the Document Viewer view.

Defining a Category

- Category names
- Patterns section
  - Words threshold and keywords weight
- Combinations section
  - Proximity
- Wildcards
- Uppercase/Lowercase definitions
- Useful tricks
- Examples

Prerequisite: Creating a Category Definition File

EMM OSINT Suite provides a **Domain-Specific Language (DSL)** to define a category or alert by using the Category Definition File editor view. A Category Definition File (adf file) consists of two different sections mainly:

- **Patterns section**, which is a simple keyword-weight list with a defined threshold (also optional).
- **Combinations section**, which is a list of keyword combinations and optionally a proximity attribute.

The keywords in the patterns section determine whether a text is classified to fit a category or not.

The basic structure of a Category Definition File is the following:

```
define alert <ID>
  (
   (words threshold <INT>)?
     define patterns
    <ID> <INT> (, <ID> <INT>)*
     end patterns
  )?
  (
   define combination
    (proximity <INT>)?
    ( begin or
     <ID> (, <ID>)*
    end or )+
    ( begin not
     <ID> (, <ID>)*
    end not )?
    end combination
  )*
 end alert
```

Category names

Regarding the names to be used as **category names**, they should be unique in the system. This means that a name can only be used for one category exactly. It should contain only non-accented alphanumerical characters. The name is case sensitive but case should nevertheless NOT be used to distinguish between categories, i.e. category 'myTest' and 'mytest' should not be used at the same time.

Patterns section

The first way to define an alert is to use a **list of keywords with associated weights**. This simple keyword method is preferred for performance reasons and it is very effective if you are looking for precise, unambiguous terms or names (e.g. *Gazprom Media*, *brucellosis*, *Michael Phelps*). If a precise term consists of two or more words, you can use the wildcard “+”. For instance, if you are interested in “yellow fever” you do not want all the documents containing the word “yellow” and “fever” individually anywhere in the document, but you want them to appear together and in the same order, so you can use the “+” symbol (“yellow+fever”). “+” effectively skips the white space between the 2 words. Note that “+” also skips punctuation marks, so “yellow, fever” would be valid.

An important feature of the Category Matcher is that it is **multilingual**. If the users want to categorize articles in various languages, they have to define their alerts also in different languages. The system will accept terms in any language, and an alert may have any numbers of keywords.

Words threshold and keywords weight

The value of the words threshold for each alert and the weight for each pattern can be chosen by the user and both are optional. If the user wants to define one, they have to be **integer values**. Eventually, the system keeps track of the total weight of the individual patterns, and only if the threshold set by the user has been reached, the document will be categorized to the category. The words threshold is ONLY used within the current definition. The value has no particular meaning other than to check against the total of the values of the patterns found in the text. The word weight list can also be used with weights less than the threshold value.

The system already takes into account the same pattern for a maximum of 8 occurrences. It assigns a decreasing weight based on the following multipliers: 1.0, 1.0, .6, .4, .4, .2, .2, .2. So, if a pattern matches multiple times, the system will automatically reduce the value of the weight and one pattern will never score more than 4 times its full value.

Combinations section

The second way to define an alert is to create one or more **combinations of lists of keywords**. A combination section is composed of one or more "OR" lists of patterns (***or*** **sub-sections)** and optionally one "NOT" list of patterns (***not*** **sub-section)**. When a combination section is defined by the user, at least one of the keywords belonging to each *or* sub-sections must be found in the document to assign such category to it. Obviously, if any of the keywords defined within the *not* sub-section is found, then the document would be automatically discarded although some *or* combinations were found. Therefore, the "NOT" list of patterns means "unwanted words".

One should use a combination for a broader term or concept (e.g. *imported disease*, *release of toxic substances*, *equal rights* or such combinations as *Russian peacekeeping mission in Caucasus*, *Russian Georgian Conflict*). As explained above, each "OR" list of patterns would express a certain concept and a document would be considered for the alert if every concept is found in the document but rejected if the “NOT” concept is found.

Proximity

The “proximity” value is optional and it can be very handy. It will define a word context size within which the combination terms have to occur. If this value is not defined by the user, then the Category Matcher will use 10 as default value.

Wildcards

The Category Matcher allows using several wildcard characters:

- **% (percent)** for 0, 1 or more characters. E.g. origin% would match original, originality, originally, originate, originating, originator, origination... This wildcard can be very useful with inflecting/fusional languages like Russian for example.

- **_ (underscore)** for exactly one character, it does not denote a blank. E.g. p_t would match pot, put, pat…., “organi_ation” would match both “organization” and “organisation”.

- **Set**: [abc] in a pattern definition means that the system will match either an ‘a’, a ‘b’ or a ‘c’ in that position. E.g. c[aou]t would match ‘cat’, ‘cot’ and ‘cut’.

- It is possible to introduce prefixes in the following way: **@prefix]**. Please note that you have to introduce a prefix only once in the whole system. E.g. together with words like bug, bunk, but, claim you can introduce @de] and the system will automatically get debug, debunk, debut, declaim etc.…This symbol should be used with caution as it will affect all other alert definitions.

- The **“+” (white space)** sign can be used to build or unite term strings. E.g. Olympic+games, News+Brief, dmitry+medvedev would match dmitry (white space) medvedev.

Using these wild cards can be very helpful to build common patterns for multiple languages because they can substitute accented characters. E.g. ent_rotox% would match enterotoxine (de), ente’rotoxin (fr), enterotoxín (sk) etc.

It is possible to use word-initial wild cards (both _ and %), but these should be used as little as possible because they are computationally heavy. If you only want to cover one word-initial letter, you should use the _ (underscore) instead of %(percent). If there are only two or three variants of your chosen word/patter, it would be much better to put them in explicitly instead of using a wildcard.

Uppercase/Lowercase definitions

A pattern definition should normally be in lowercase, but can contain upper case characters. In that case the pattern will only match text that has an uppercase character in the same position. Forcing uppercase can be used for acronyms that would otherwise cause problems. This means that a lowercase character in a patter matches both lower and uppercase in the incoming text, but an UPPERCASE character only matches uppercase, so the pattern e.g. “abc” would match ABC, ABc, Abc, aBC,AbC etc but the pattern “Abc” would only match Abc, ABC, ABc etc all with the uppercase “A”.

Useful tricks

- **Negative weights** can be used. Negative weights can be useful if a search word is homographic with some other unrelated word or with a person name, or if a search word has many meanings. E.g. if you are interested in finding texts mentioning “tsunami”-sea storm, you could put several words with a negative score of let’s say -999.
  - rock-band -999 (there is a famous Indian rock band “Tsunami”)
  - Arashi+Tsunami -999 (a Japanese voice actor)
  - Satoshi+Tsunami -999 (a Japanese football player)
  - deodorant -999 (“Tsunami” fragrance by Axe)
  - politics -999 (“Tsunami” term used to describe an overwhelming victory by a political party)
  - Another example: if you are interested in Michael Jackson the Canadian actor and not the musician, you could put words like pop music, songwriter, dancer etc. with a negative weight.

- **To use weights and a threshold** (e.g. 50), so that some words can trigger the alert on their own (weight = 50) while other words need to occur cumulatively (several times) before reaching the threshold (e.g. weight = 20). For example:

```
define alert Biotechnology
 words threshold 50
 define patterns
  genetic_ 40, cancer 20, genomics 50, antibodies 40, biotechnology 50
 end patterns
end alert
```

- **Be careful with abbreviations**. E.g. a very simple (at first sight) abbreviation “ABC” can stand for the following: Latin Alphabet, American Broadcasting Company, Australian Broadcasting Company, Associated British Company, Appalachian Brewing company, Atlanta Bread company, Agricultural Bank of China, ABC (programming language), abc conjecture, All Lesotho Convention, ABC (island in Alaska) etc. Please keep in mind that a simple word (or an abbreviation) in English can mean a completely different thing in Italian, German, Russian, Bulgarian… E.g. a Portuguese word for “vomiting” is “emese”, and “Emese” is a very common first feminine name in Hungary. Another typical example is the work ‘mais’ which means the agricultural crop in many languages, but in French means ‘but’. The French version is written with an accented ‘i’. In order to avoid these conflicts it is sometimes useful to define multiple combinations with the word lists in the combination reflecting the various languages or language groups. A combination is unlikely to trigger the category if one or-list consists of English words and the other of French words.

Examples

Here you can find some examples of using the DSL to define Category Definition Files:

```
define alert TerroristAttack
 words threshold 20
 define patterns
  stratégiai+bombázás% 10, film% -999, car+bomb% 10, bomb%+detonat% 10,
  attentato+suicide 10, camion+bomba 10
 end patterns
 define combination
  proximity 5
  begin or
   terroris%, bioterrorism%, attack%, attentat%
  end or
  begin or
   attacco+chimico, attacco+tossico, allarm_, sostanz_+tossic%, sostanz_+chimic%
  end or
 end combination
 define combination
  proximity 15
  begin or
   ETA%, IRA%, Al-Kaida%, szélsoséges
  end or
  begin or
   bomba%, terror%, tömeg+gyilkosság%, gyilkosság%, merényl%, mészárl%, csoport%
  end or
  begin not
   könyv%, gól, film%
  end not
 end combination
end alert
```

```
define alert WaterConflict
 define combination
  begin or
   water, eau
  end or
  begin or
   conflict, conflict%
  end or
  begin not
   book%, film%, movie%, game%, song%, mostra+fotografica
  end not
 end combination
end alert
```

```
define alert NaturalDisasters
 define combination
  begin or
   tsunami, volkanik, erozyon, volkan%+pat%, volkan%+kül%, sel%+felaket%, heyelan%
  end or
  begin or
   ferit_, ferida%, vittim_, crash%, explosion%
  end or
 end combination
end alert
```

By doing **Ctrl + Space**, the adf editor view shows the next command feasible to be used at the current position of the cursor

Defining the maximum number of search result links

See also: Performing an Internet Search

The search result link extraction function extracts links from the result pages of search engines. It tries to extract the result links on the first and following pages up to a maximum number of links. You can define this maximum link count as follows:

1. In the main menu click **Window > Preferences**. The **Preferences** dialog opens.
2. In the topic tree on the left side, expand *OSINT Preferences* and click *Link Extraction*. The **Link Extraction** preference page is shown on the right side.
3. In the **Link Extraction** preference page enter a number in the **Maximum link count** field, then click **OK**.

Downloading of Bookmarks

Prerequisites: Creating a Case Project and Performing an Internet Search

See also: Setting HTTP Proxy Information

Bookmark files contain URLs pointing to some resources on the web. Theses resources are mostly web pages or file resources (such as .pdf files). To analyse theses resources locally, they need to be downloaded to the Case Project in the workspace. Once all the search result links (bookmarks) have been downloaded under the *Bookmarks* folder, the next step is to download the result pages as text files:

- Right click on the desired *bookmark* folder in the **Workspace Navigator view** and click on **Download Bookmarks.** The **Progress View** shows the progress of the download. The downloaded web pages or files (for example PDF files) are stored in the predefined *Documents* folder of the case project.

Unknown Attachment

- After downloading the search results, a new folder will appear under the ***Documents*** folder with the same name as the selected *Bookmarks* folder. It will contain a text file with the raw content for every link stored in the *Bookmarks* folder. After the download has finished, the system automatically extracts the raw text from the web pages or files. The system can extract raw text from a variety of file formats (such as PDF, MS Office formats and others). Also, the system detects the language of the text.

Unknown Attachment

After the download has finished the system automatically starts to **extract the text from the resources**.

The download process is performed in parallel, but in some cases the process has to wait for URLs to respond which are pointing to slow servers.

The files in the *Documents* folder are marked with an **> to show the entity extraction has not yet been performed**. Some files may be marked with a red indicator showing that the text extraction has failed. In most cases the file is either an unsupported file type or does not contain enough relevant text.

The downloaded result pages or result files are given a name based on the result URL they were downloaded from. If a title can be extracted from the extracted text of the file this title is shown as an overlay in the **Workspace Navigator** view. If the text extraction failed to find a title, the original file name is shown instead.

Reviewing result pages and files

Double-click one of the downloaded files (HTML files have a small “HTML” icon in front of them) to open the file in the **Document editor view**. The **Document editor view** shows the extracted text. Since the entity extraction has not yet run, just the plain extracted text without any mark-up for entities is shown.

Unknown Attachment

Encoding a File in UTF-8

The software uses UTF-8 (Unicode) encoded files to support multiple languages and scripts. If you need to encode a file from the standard encoding on your PC (such as Windows-1252) to UTF-8, please do the following:

- Download Notepad++ which is a free text editor and install it to your system
- Load your file in Notepadd++
- Change encoding of your file to utf-8
- Save utf-8 encoded file

Loading a File in Notepad++

- Start Notepad++
- Click on **File > Open...**
- Select the file to open and click on **Open**

Changing encoding to UTF-8

- Click on **Encoding > Convert to UTF-8 without BOM**

Saving UTF-8 encoded File

- Click on **File > Save** to save the converted file to disk

Filtering Search Engine Result Bookmarks

Sometimes the Internet Search might result in a number of Bookmarks from sites which contain no relevant results. For example, if searching for a person often the Linkedin.com profile pages shows up containing other persons who have no releationship with the main person of the profile.

The software allows to filter Bookmarks out by matching the URL to custom patterns.

To define filter patterns do the following to open the Link Extraction Preferences:

- Click on **Window > Preferences** to open the **Preferences Dialog**
- Expand the **OSINT Preferences** branch in the left hand panel
- Click on **OSINT Preferences > Link Extraction** to show the preferences for search engine link extraction

Now in the **Filter out URL Patterns** table custom patterns can be defined:

- Click **New...** to open the **New Pattern dialog**
- Enter a pattern which matches URLs to be filtered out from the results
- Click **OK**

Filter Pattern
The patterns use regular expression syntax to match against search result URLs

Example

You want to filter out all result Bookmarks which point to linkedin.com pages. Use the following pattern:

```
.*linkedin\.com.*
```

Importing Bookmarks from web browsers

Prerequisite: Creating a Case Project

The EMM-OSINT Suite also supports importing *bookmarks* from different web browsers such as Internet Explorer, Firefox or Google Chrome into the *Bookmarks* folder of the Case Project. Therefore, the first step will be to export bookmarks from our web browser into a HTML file. Follow the next links in order to find out how to export bookmarks from your web browser:

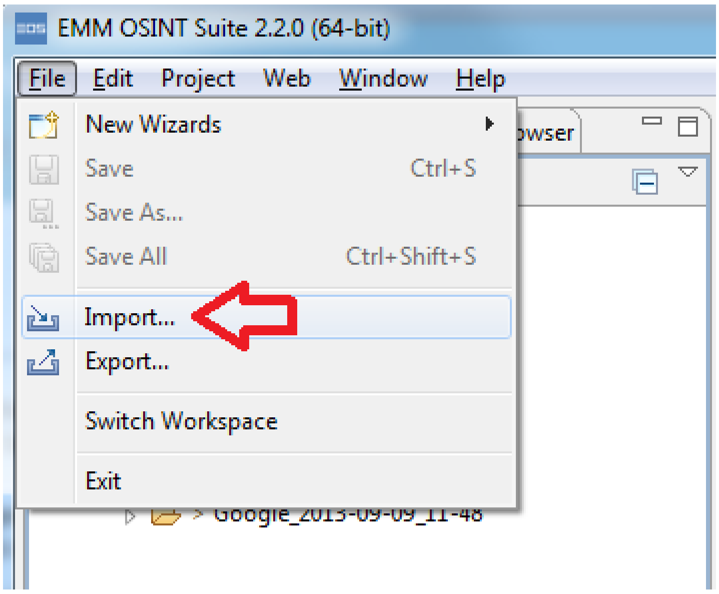
Find out how to
Export bookmarks from Mozilla Firefox

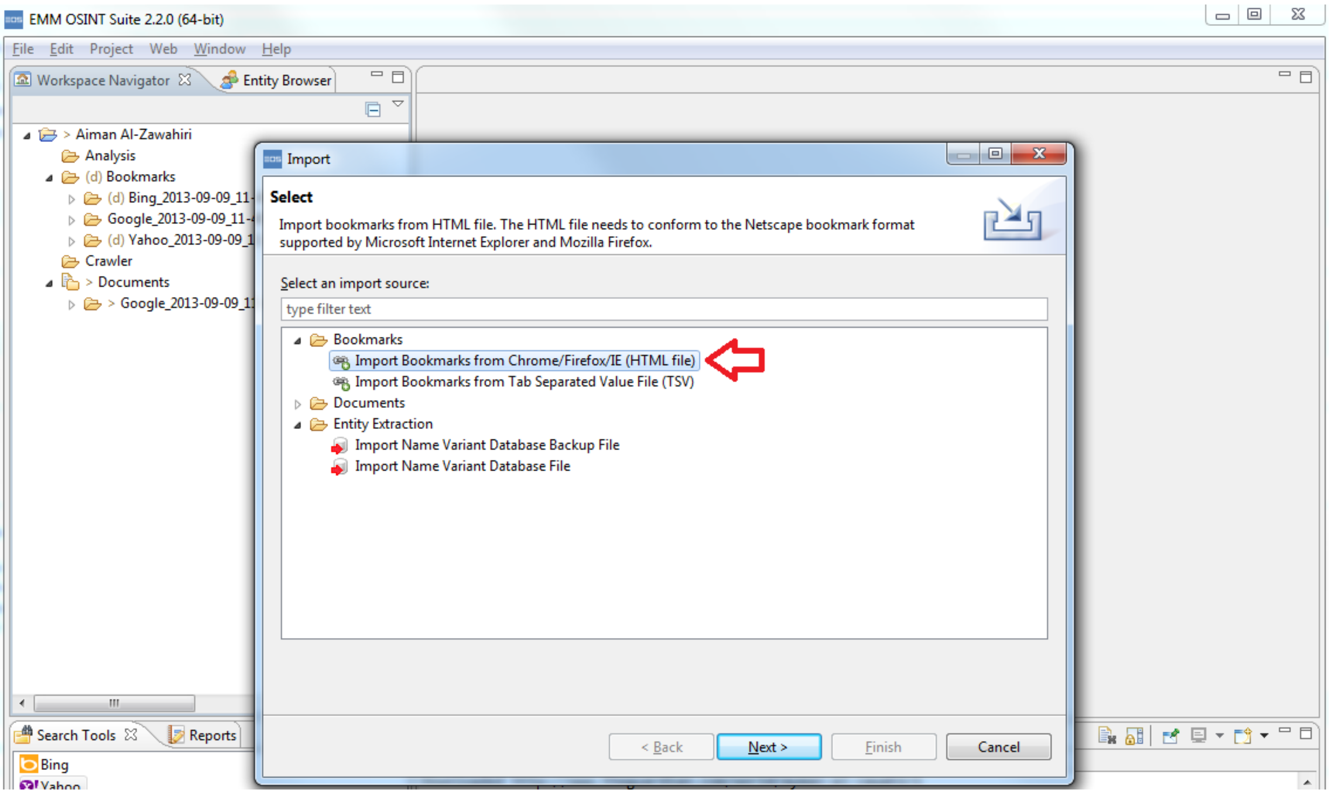
Export bookmarks from Google Chrome

Export bookmarks from Internet Explorer

Once the bookmarks have been exported from the web browser into a file (usually it is based on an ancient HTML format), to import it into EMM-OSINT Suite do the following:

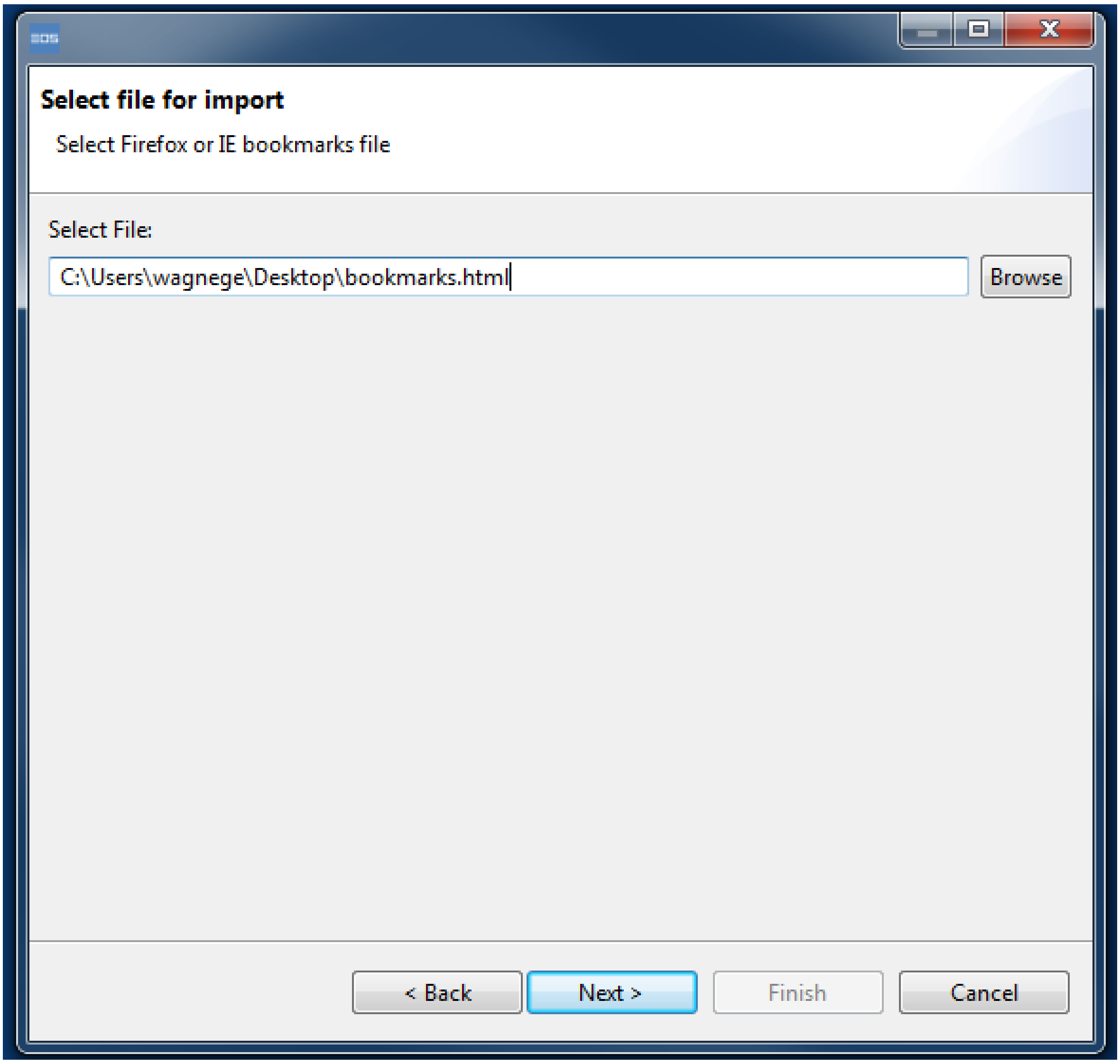
- Click in the main menu **File > Import** and from the list of available import sources select **Bookmarks > Import Bookmarks from Chrome/Firefox/IE (HTML file)** and click **Next**. The system shows the Import dialog





Import bookmarks from HTML file. The HTML file needs to conform to the Netscape bookmark format supported by Microsoft Internet Explorer and Mozilla Firefox.

- Click **Browse** and select the file exported previously from the web browser. Click **Next** to proceed and the import wizard shows the page to define the import location in your Workspace.

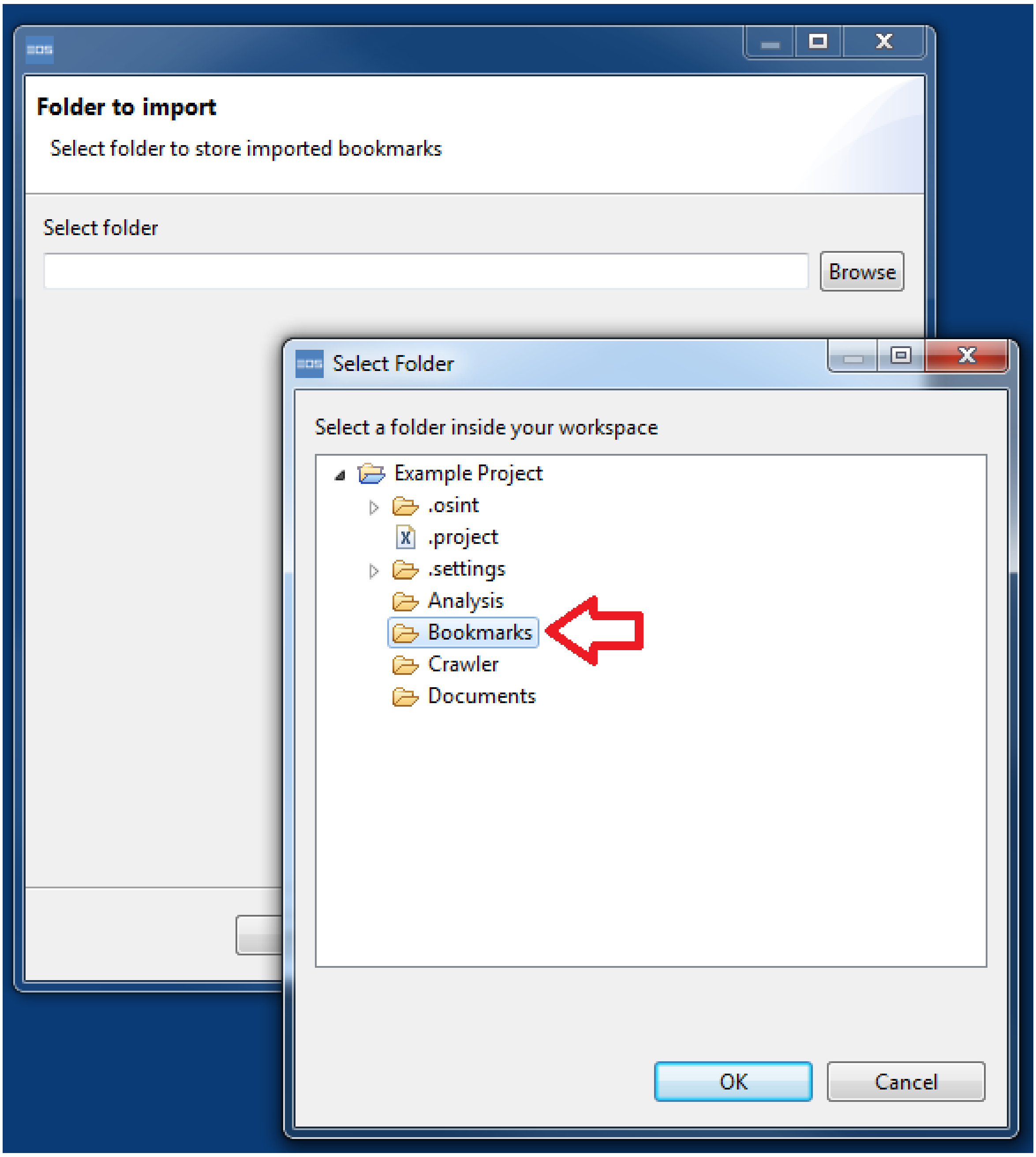


Select file for import

Select Firefox or IE bookmarks file

Select File: C:\Users\wagnege\Desktop\bookmarks.html

- Click **Browse** to select a folder in your Workspace. A file selection dialog opens. **Select the *Bookmarks* folder or a sub-folder** and click **OK**. The file selection dialog closes



Folder to import

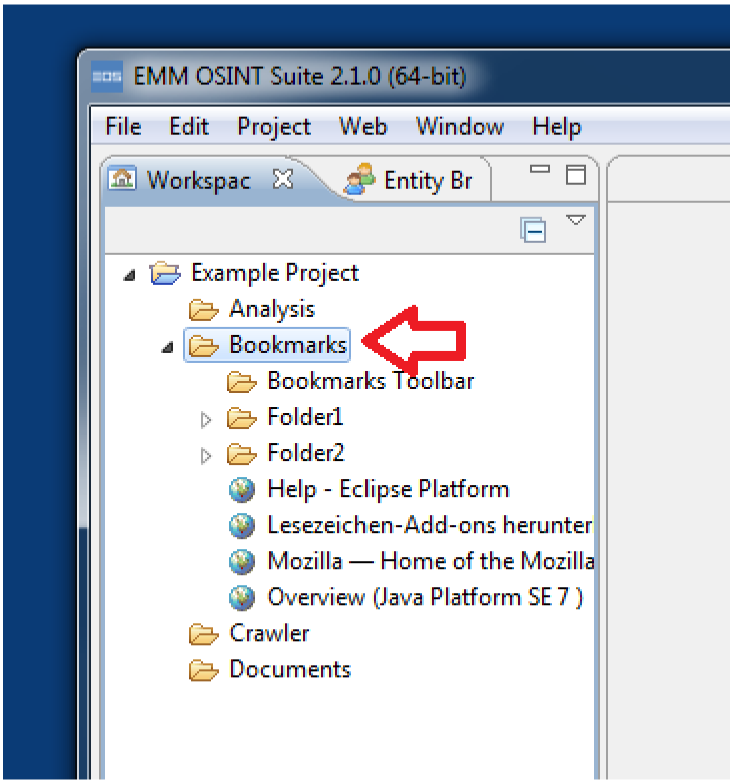
Select folder to store imported bookmarks

Select folder

Select Folder

Select a folder inside your workspace

- Click **Finish** to perform the bookmarks import. The system imports the bookmarks from the HTML file and prints a status message for each imported bookmark in the **Console** view. After the import has finished successfully, the imported bookmarks appear in your Workspace

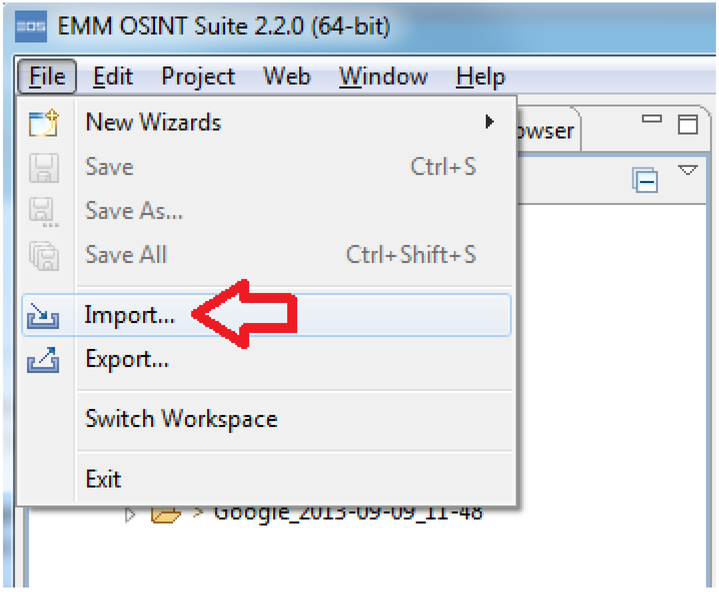


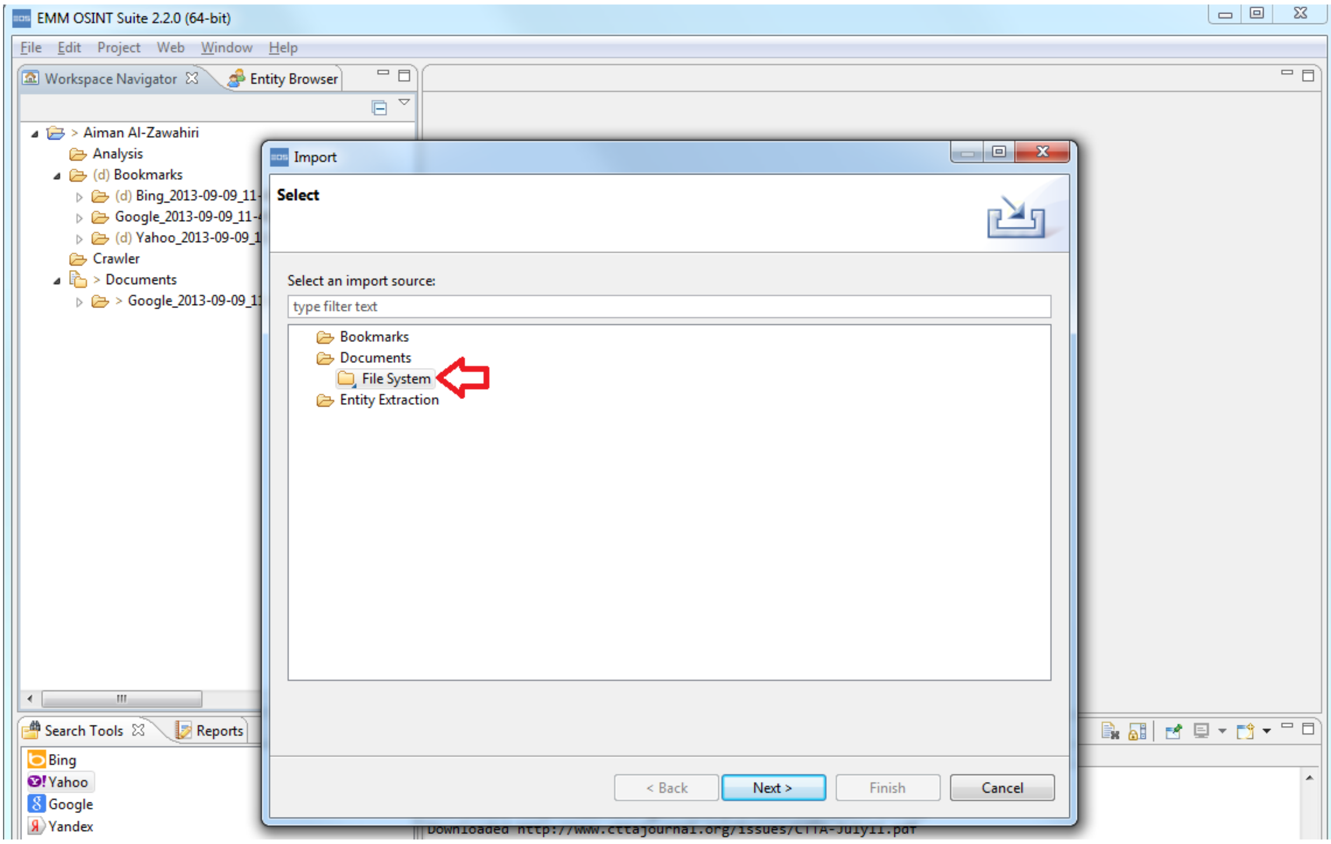
Importing Documents from Local Disk

Prerequisite: Creating a Case Project

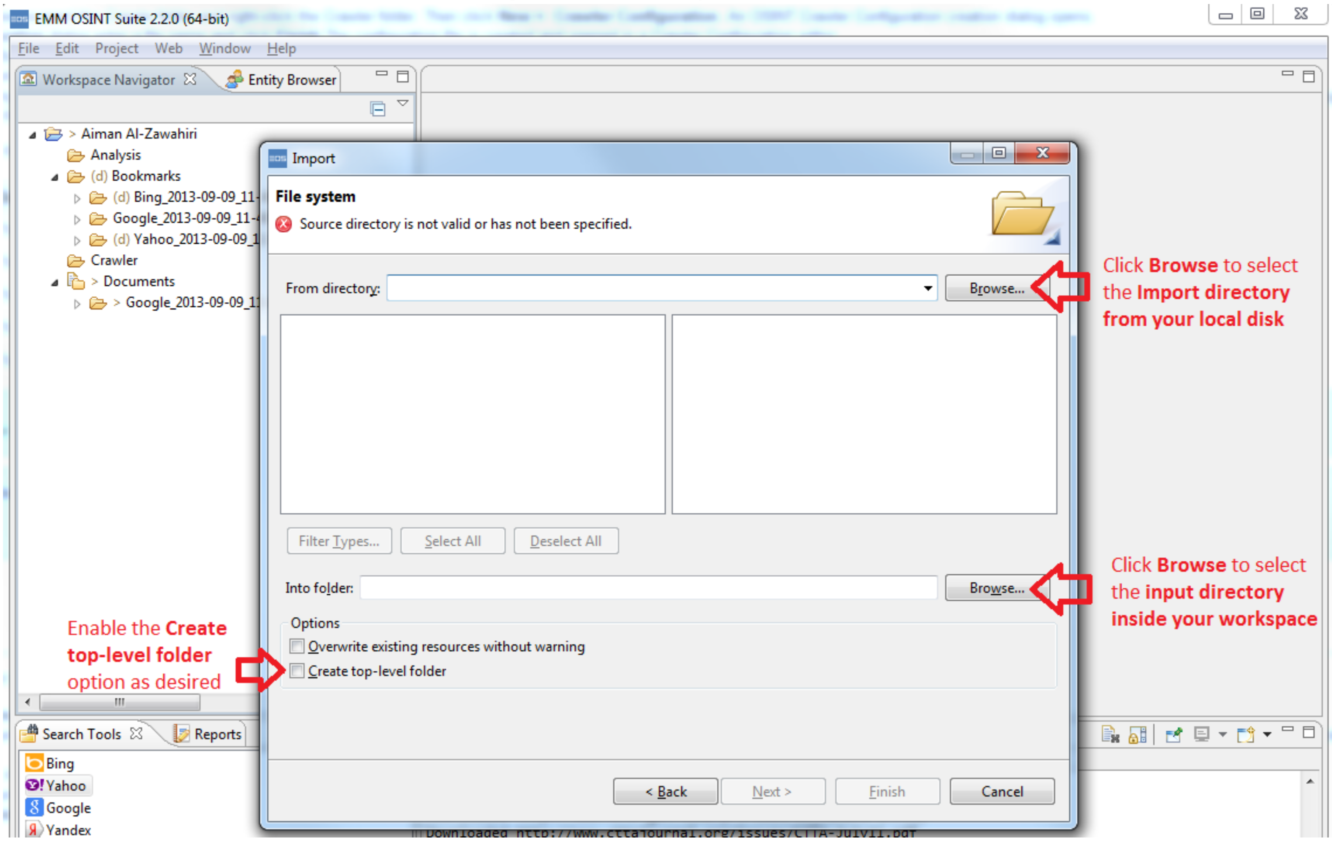
The EMM-OSINT Suite allows importing files from local disk into a Case Project for a further analysis. To import files perform the following procedure:

- Click in the main menu **File > Import** and from the list of available import sources select **Documents > File System** and click **Next**. The system shows the Import dialog





- Click **Browse** to select the import directory in the **From directory** field. The system shows an **Import from directory** file dialog. Fro m the file system list select the import base directory and click **OK**. The system closes the dialog and returns to the **Import** dialog. C lick on a folder or file to select it for import (optional: click **Filter Types** to filter the import for specific file types). Click **Browse** to select the import directory **inside your Workspace** (note**:** make sure to select a Case Project). Enable the **Create top-level folder** option.



Source directory is not valid or has not been specified.

□ Overwrite existing resources without warning

□ Create top-level folder

Click **Browse** to select the **Import directory from your local disk**

Click **Browse** to select the **input directory inside your workspace**

Enable the **Create top-level folder** option as desired

- Click **Finish** to start the import

Files must be imported into the ***Documents*** folder of a Case Project (or in a subfolder of the Documents folder) in order to be analysed later using the Entity Extraction

Installing on Linux (Ubuntu)

Note: This document describes the installation under Ubuntu Linux. Other Linux variants may vary.

Prerequisite: Download the application archive from our download page. After download you should have an application archive named emm-osint-suite-*<version-number>.linux.gtk.x86_64.tar.gz* (for the 64-bit version) in your download folder.

Unpacking and Installing EMM OSINT Suite

Extract the application archive to a location of your choice. According to the File System Hierachy Standard (FHS) the application should be placed in */opt/emm-osint-suite-<version-number>*

Add executable permissions

If the software does not start by clicking on the osint executable in the installation directory or gives an error message, please add executable permission to two files as follows (the directory names may change for later versions - commands for version 2.3.3 shown):

Add executable permissions

```
cd /opt/emm-osint-suite-2.3.3
sudo chmod 755 osint
sudo chmod 755 features/it.jrc.osint.jre.linux.gtk.x86_64.feature_1.8.45.201507271600/jre/bin/java
```

Related Topics:

- Running on Linux
- Requesting Support

Installing on Windows

Prerequisite: Download the application archive from our download page. After download you should have an application archive named *osint_<version-number>_x86_64.zip* (for the 64-bit version) in your download folder.

Note: For a successful installation local administrative rights may be needed. Contact your PC support for help if the installation fails.

The application archive is a compressed ZIP archive which can be decompressed using standard tools of MS Windows (or using tools such as 7zip or Winzip).In order to install the application on a MS Windows PC, perform the following steps:

1. Double-click the application archive to uncompress the application archive in the download folder. The uncompress utiltiy should create an application folder which is named *osint_<version-number>_x86_64*.
2. Open the **Windows File Explorer** and navigate to the download folder
3. Move the application folder *osint_<version-number>_x86_64* from the download folder to the standard Windows program folder: *c:\Program Files\osint_<version-number>_x86_64*

> You can install the application also in a non-standard location. Since the version number is included in the name of the application folder, multiple version may co-exist on your system.

Related Topics:

- Requesting Support
- Running on Windows

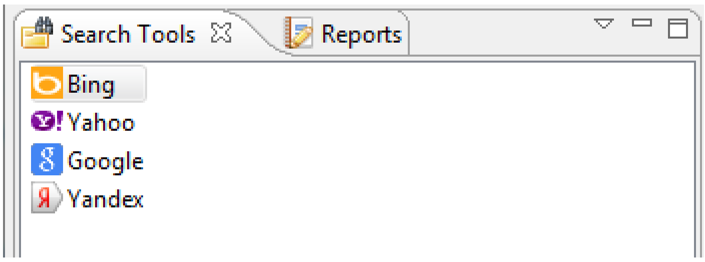
Performing an Internet Search

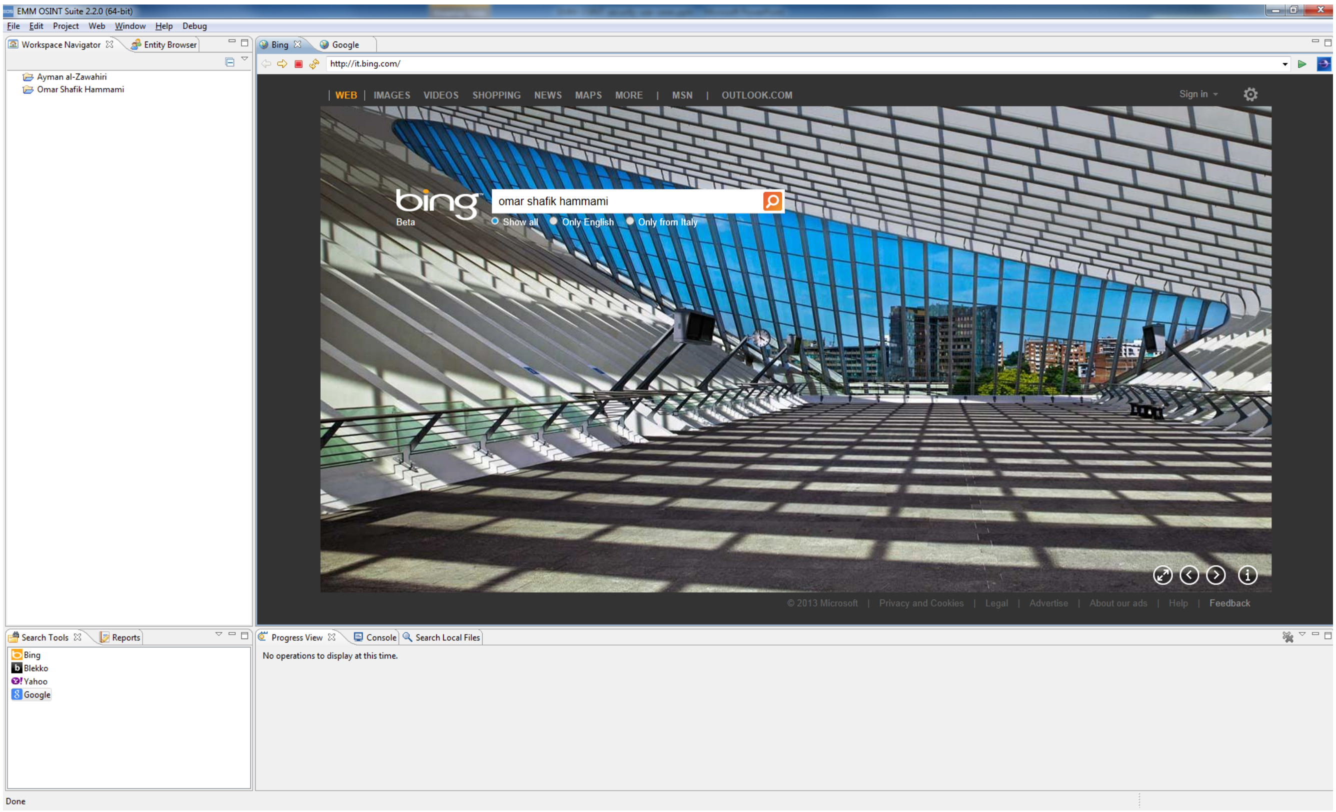
Prerequisite: Creating a Case Project

To find relevant information on the Internet the application allows you to search using the major Internet search engines and then use the search results for further research.

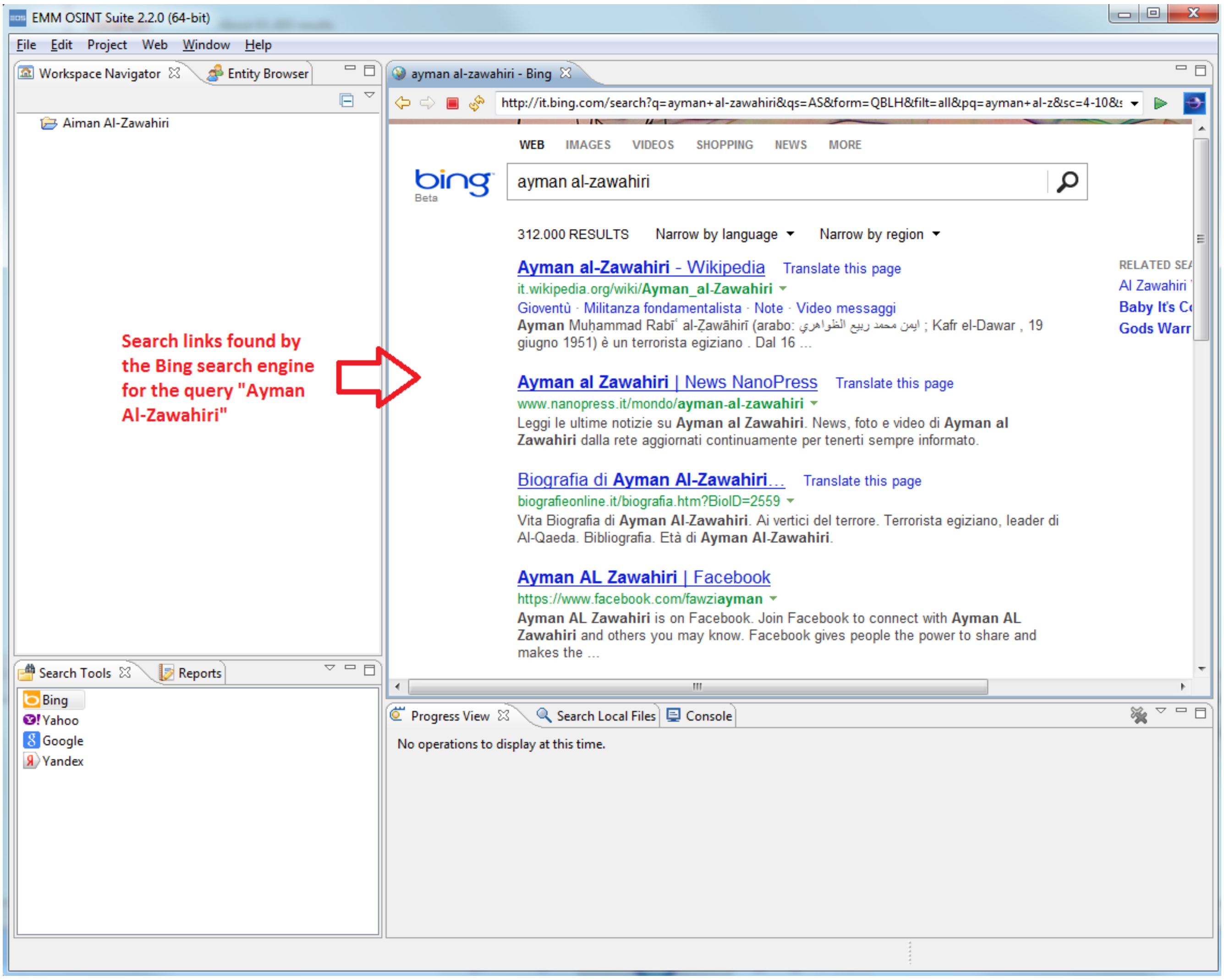
To perform an Internet search, do the following:

- Double-click on one of the available search engines in the **Search Tools** view in the lower-left corner. For this example choose Microsoft's search engine Bing. The start page of such search engine will be opened in a new Browser editor window.




EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help Debug
Workspace Navigator
Entity Browser
Ayman al-Zawahiri
Omar Shafik Hammami
Bing
Google
http://it.bing.com/
WEB | IMAGES VIDEOS SHOPPING NEWS MAPS MORE | MSN | OUTLOOK.COM
Sign in
bing
Beta
omar shafik hammami
Show all
Only English
Only from Italy
© 2013 Microsoft | Privacy and Cookies | Legal | Advertise | About our ads | Help | Feedback
Search Tools
Reports
Bing
Blekko
Yahoo
Google
Progress View
Console
Search Local Files
No operations to display at this time.
Done

- **Perform your search** as usual using the online search engine. Then, the result links for such search will appear on the **Browser editor window**:



Search links found by the Bing search engine for the query "Ayman Al-Zawahiri"

312.000 RESULTS Narrow by language Narrow by region

Ayman al-Zawahiri - Wikipedia Translate this page
it.wikipedia.org/wiki/Ayman_al-Zawahiri
Gioventù · Militanza fondamentalista · Note · Video messaggi
Ayman Muḥammad Rabīʿ al-Zawāhirī (arabo: ايمن محمد ربيع الظواهري ; Kafr el-Dawar , 19 giugno 1951) è un terrorista egiziano . Dal 16 ...

Ayman al Zawahiri | News NanoPress Translate this page
www.nanopress.it/mondo/ayman-al-zawahiri
Leggi le ultime notizie su Ayman al Zawahiri. News, foto e video di Ayman al Zawahiri dalla rete aggiornati continuamente per tenerti sempre informato.

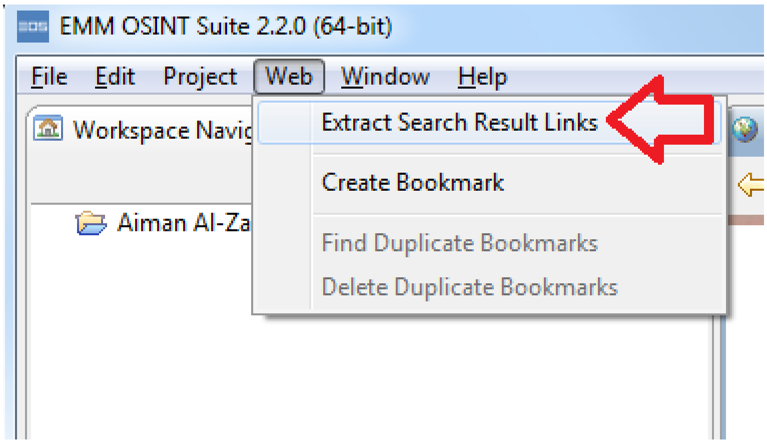
Biografia di Ayman Al-Zawahiri... Translate this page
biografieonline.it/biografia.htm?BioID=2559
Vita Biografia di Ayman Al-Zawahiri. Ai vertici del terrore. Terrorista egiziano, leader di Al-Qaeda. Bibliografia. Età di Ayman Al-Zawahiri.

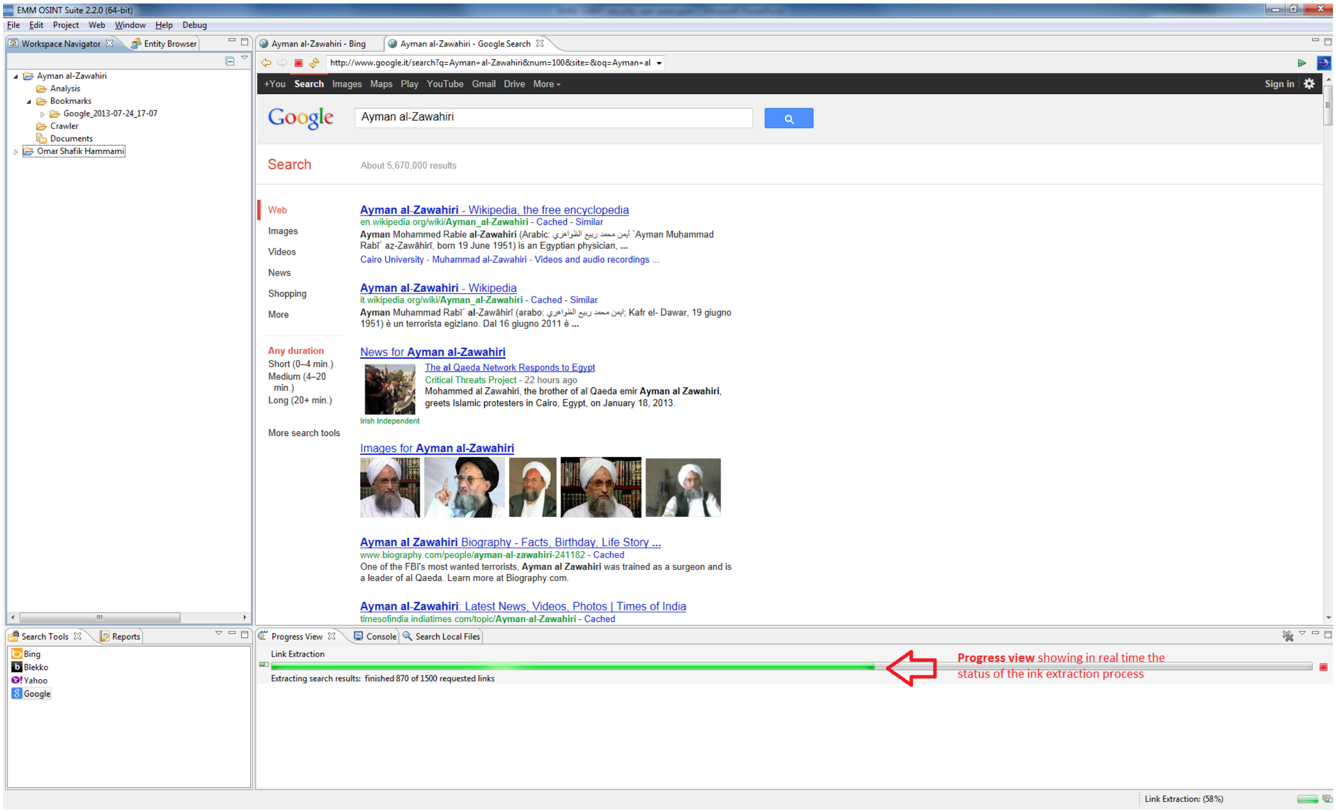
Ayman AL Zawahiri | Facebook
https://www.facebook.com/fawziayman
Ayman AL Zawahiri is on Facebook. Join Facebook to connect with Ayman AL Zawahiri and others you may know. Facebook gives people the power to share and makes the ...

No operations to display at this time.

Since you are using the normal web interface of the selected search engine, you can use all available options and refine your search to be as relevant as possible. As soon as you are satisfied with the shown results, you can proceed to the next step

- Click on **Web > Extract Search Result Links** to extract the search results from the result page of the search engine and store them in your Case Project. The link extraction process starts and the system shows a warning that it will take control of the Browser editor in order to extract the links. They will be stored as ***bookmarks***. If more than one case project exists, the system opens the Project Selection dialog to choose in which case project we want to store the search links (bookmarks). The software extracts ("*scrapes*") the search result links from the first and subsequent pages of the search engine. After it has extracted the maximum number of links (by default 100 links), it returns to the first result page. The progress of the scraping is shown in the **Progress View**. You can stop the link extraction process any time by clicking on the red cancel icon in the **Progress View**





Ayman al-Zawahiri - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Ayman_al-Zawahiri - Cached - Similar
Ayman Mohammed Rabie al-Zawahiri (Arabic: أيمن محمد ربيع الظواهري ʾAyman Muḥammad Rabīʿ aẓ-Ẓawāhirī, born 19 June 1951) is an Egyptian physician, ...
Cairo University - Muhammad al-Zawahiri - Videos and audio recordings ...

Ayman al-Zawahiri - Wikipedia
it.wikipedia.org/wiki/Ayman_al-Zawahiri - Cached - Similar
Ayman Muḥammad Rabīʿ al-Ẓawāhirī (arabo: أيمن محمد ربيع الظواهري; Kafr el- Dawar, 19 giugno 1951) è un terrorista egiziano. Dal 16 giugno 2011 è ...

News for Ayman al-Zawahiri
The al Qaeda Network Responds to Egypt
Critical Threats Project - 22 hours ago
Mohammed al Zawahiri, the brother of al Qaeda emir Ayman al Zawahiri, greets Islamic protesters in Cairo, Egypt, on January 18, 2013.
Irish Independent

Images for Ayman al-Zawahiri

Ayman al Zawahiri Biography - Facts, Birthday, Life Story ...
www.biography.com/people/ayman-al-zawahiri-241182 - Cached
One of the FBI's most wanted terrorists, Ayman al Zawahiri was trained as a surgeon and is a leader of al Qaeda. Learn more at Biography.com.

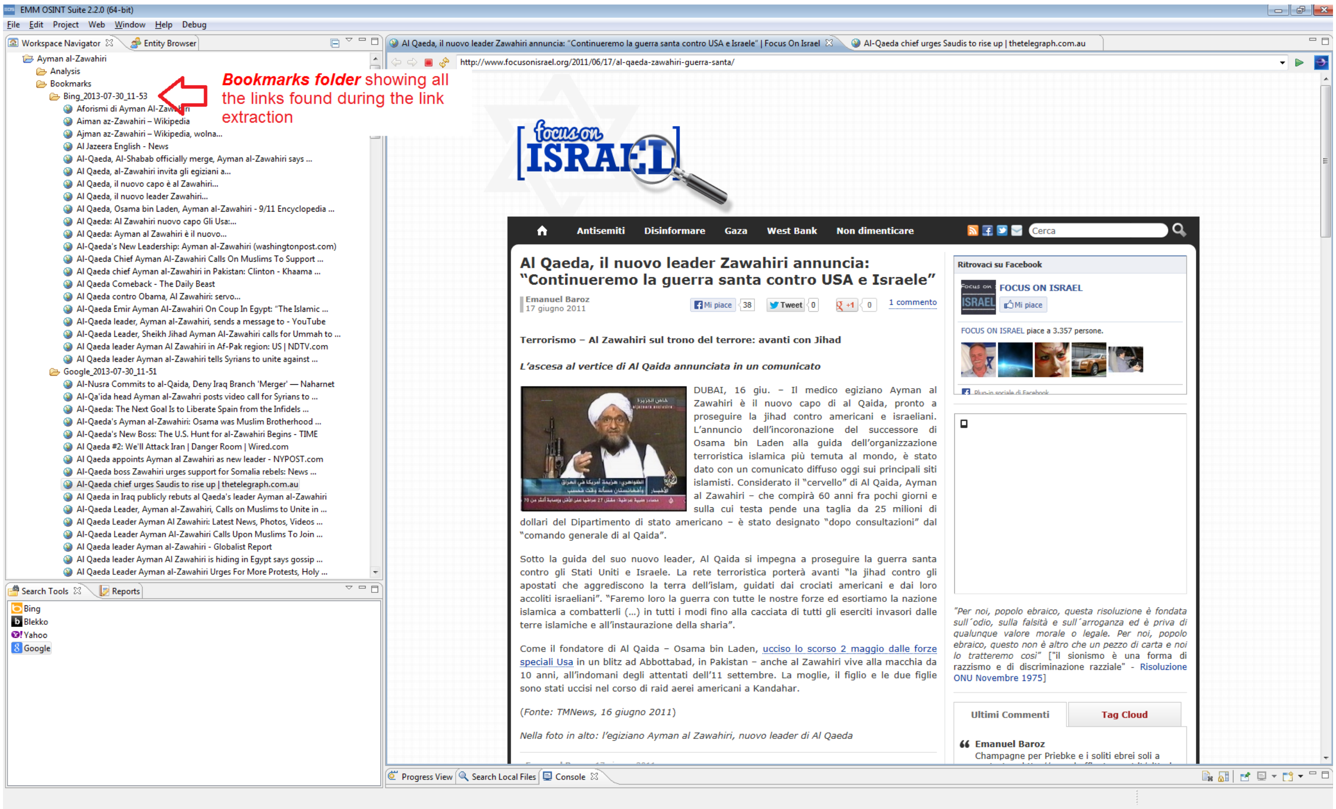
Ayman al-Zawahiri: Latest News, Videos, Photos | Times of India
timesofindia.indiatimes.com/topic/Ayman-al-Zawahiri - Cached

Link Extraction
Extracting search results: finished 870 of 1500 requested links

Progress view showing in real time the status of the ink extraction process

Link Extraction: (58%)

- After the link extraction has finished, it creates a folder with Bookmarks files in the ***Bookmarks* folder** of the defined target Case Project**.** The extracted result links are stored *as .obm* files. A *.obm* file which we also call a "***bookmark file***" contains the actual result link plus meta information such as the search engine, the search query and a time stamp. Such meta information can be found by clicking on the bookmark file and then clicking on the **Properties view**.


EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help Debug
Workspace Navigator
Entity Browser
Al Qaeda, il nuovo leader Zawahiri annuncia: "Continueremo la guerra santa contro USA e Israele" | Focus On Israel
Al-Qaeda chief urges Saudis to rise up | thetelegraph.com.au
http://www.focusonisrael.org/2011/06/17/al-qaeda-zawahiri-guerra-santa/
Search Tools
Reports
Bing
Blekko
Yahoo
Google
Progress View
Search Local Files
Console
focus on ISRAEL
Antisemiti
Disinformare
Gaza
West Bank
Non dimenticare
Cerca
Mi piace 38
Tweet 0
+1 0
1 commento
Ritrovaci su Facebook
FOCUS ON ISRAEL
Mi piace
FOCUS ON ISRAEL piace a 3.357 persone.
Plug-in sociale di Facebook
Ultimi Commenti
Tag Cloud

Bookmarks folder showing all the links found during the link extraction

Ayman al-Zawahiri
Analysis
Bookmarks
Bing_2013-07-30_11-53
Aforismi di Ayman Al-Zawahiri
Aiman az-Zawahiri – Wikipedia
Ajman az-Zawahiri – Wikipedia, wolna...
Al Jazeera English - News
Al-Qaeda, Al-Shabab officially merge, Ayman al-Zawahiri says ...
Al Qaeda, al-Zawahiri invita gli egiziani a...
Al Qaeda, il nuovo capo è al Zawahiri...
Al Qaeda, il nuovo leader Zawahiri...
Al Qaeda, Osama bin Laden, Ayman al-Zawahiri - 9/11 Encyclopedia ...
Al Qaeda: Al Zawahiri nuovo capo Gli Usa:...
Al Qaeda: Ayman al Zawahiri è il nuovo...
Al-Qaeda's New Leadership: Ayman al-Zawahiri (washingtonpost.com)
Al-Qaeda Chief Ayman Al-Zawahiri Calls On Muslims To Support ...
Al Qaeda chief Ayman al-Zawahiri in Pakistan: Clinton - Khaama ...
Al Qaeda Comeback - The Daily Beast
Al Qaeda contro Obama, Al Zawahiri: servo...
Al-Qaeda Emir Ayman Al-Zawahiri On Coup In Egypt: "The Islamic ...
Al-Qaeda leader, Ayman al-Zawahiri, sends a message to - YouTube
Al-Qaeda Leader, Sheikh Jihad Ayman Al-Zawahiri calls for Ummah to ...
Al Qaeda leader Ayman Al Zawahiri in Af-Pak region: US | NDTV.com
Al Qaeda leader Ayman al-Zawahiri tells Syrians to unite against ...
Google_2013-07-30_11-51
Al-Nusra Commits to al-Qaida, Deny Iraq Branch 'Merger' — Naharnet
Al-Qa'ida head Ayman al-Zawahri posts video call for Syrians to ...
Al-Qaeda: The Next Goal Is to Liberate Spain from the Infidels ...
Al-Qaeda's Ayman al-Zawahiri: Osama was Muslim Brotherhood ...
Al-Qaeda's New Boss: The U.S. Hunt for al-Zawahiri Begins - TIME
Al Qaeda #2: We'll Attack Iran | Danger Room | Wired.com
Al Qaeda appoints Ayman al Zawahiri as new leader - NYPOST.com
Al-Qaeda boss Zawahiri urges support for Somalia rebels: News ...
Al-Qaeda chief urges Saudis to rise up | thetelegraph.com.au
Al Qaeda in Iraq publicly rebuts al Qaeda's leader Ayman al-Zawahiri
Al-Qaeda Leader, Ayman al-Zawahiri, Calls on Muslims to Unite in ...
Al Qaeda Leader Ayman Al Zawahiri: Latest News, Photos, Videos ...
Al-Qaeda Leader Ayman Al-Zawahiri Calls Upon Muslims To Join ...
Al Qaeda leader Ayman al-Zawahiri - Globalist Report
Al Qaeda leader Ayman Al Zawahiri is hiding in Egypt says gossip ...
Al Qaeda Leader Ayman al-Zawahiri Urges For More Protests, Holy ...

Al Qaeda, il nuovo leader Zawahiri annuncia: "Continueremo la guerra santa contro USA e Israele"

Emanuel Baroz
17 giugno 2011

Terrorismo – Al Zawahiri sul trono del terrore: avanti con Jihad

L'ascesa al vertice di Al Qaida annunciata in un comunicato

DUBAI, 16 giu. – Il medico egiziano Ayman al Zawahiri è il nuovo capo di al Qaida, pronto a proseguire la jihad contro americani e israeliani. L'annuncio dell'incoronazione del successore di Osama bin Laden alla guida dell'organizzazione terroristica islamica più temuta al mondo, è stato dato con un comunicato diffuso oggi sui principali siti islamisti. Considerato il "cervello" di Al Qaida, Ayman al Zawahiri – che compirà 60 anni fra pochi giorni e sulla cui testa pende una taglia da 25 milioni di dollari del Dipartimento di stato americano – è stato designato "dopo consultazioni" dal "comando generale di al Qaida".

Sotto la guida del suo nuovo leader, Al Qaida si impegna a proseguire la guerra santa contro gli Stati Uniti e Israele. La rete terroristica porterà avanti "la jihad contro gli apostati che aggrediscono la terra dell'islam, guidati dai crociati americani e dai loro accoliti israeliani". "Faremo loro la guerra con tutte le nostre forze ed esortiamo la nazione islamica a combatterli (...) in tutti i modi fino alla cacciata di tutti gli eserciti invasori dalle terre islamiche e all'instaurazione della sharia".

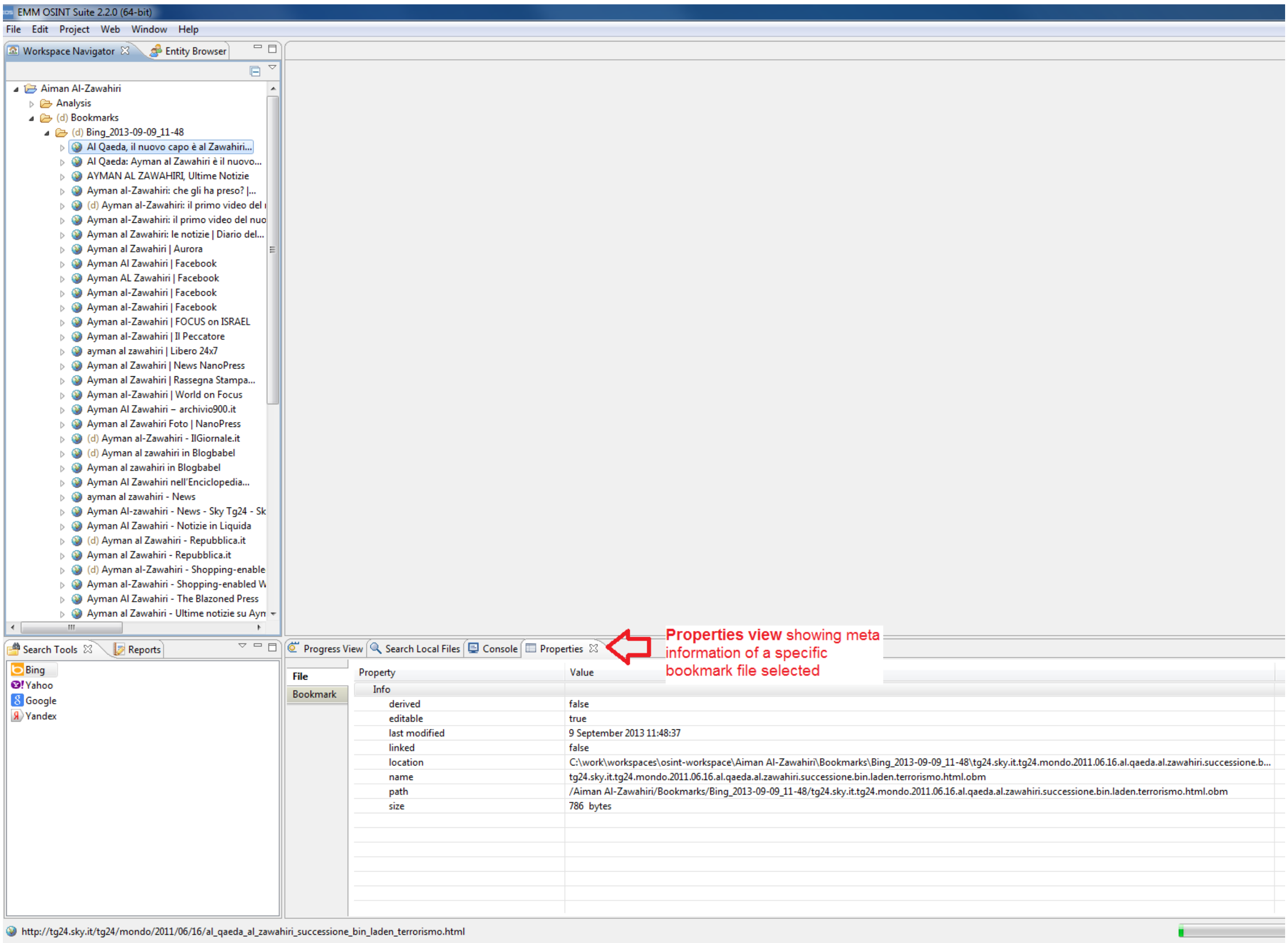
Come il fondatore di Al Qaida – Osama bin Laden, ucciso lo scorso 2 maggio dalle forze speciali Usa in un blitz ad Abbottabad, in Pakistan – anche al Zawahiri vive alla macchia da 10 anni, all'indomani degli attentati dell'11 settembre. La moglie, il figlio e le due figlie sono stati uccisi nel corso di raid aerei americani a Kandahar.

(Fonte: TMNews, 16 giugno 2011)

Nella foto in alto: l'egiziano Ayman al Zawahiri, nuovo leader di Al Qaeda

"Per noi, popolo ebraico, questa risoluzione è fondata sull´odio, sulla falsità e sull´arroganza ed è priva di qualunque valore morale o legale. Per noi, popolo ebraico, questo non è altro che un pezzo di carta e noi lo tratteremo così" ["il sionismo è una forma di razzismo e di discriminazione razziale" - Risoluzione ONU Novembre 1975]

Emanuel Baroz
Champagne per Priebke e i soliti ebrei soli a



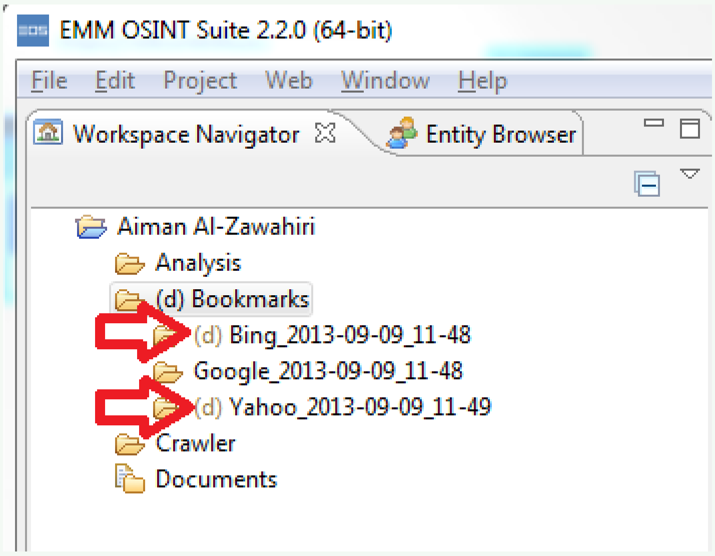
Aiman Al-Zawahiri
Analysis
(d) Bookmarks
(d) Bing_2013-09-09_11-48
Al Qaeda, il nuovo capo è al Zawahiri...
Al Qaeda: Ayman al Zawahiri è il nuovo...
AYMAN AL ZAWAHIRI, Ultime Notizie
Ayman al-Zawahiri: che gli ha preso? |...
(d) Ayman al-Zawahiri: il primo video del
Ayman al-Zawahiri: il primo video del nuo
Ayman al Zawahiri: le notizie | Diario del...
Ayman al Zawahiri | Aurora
Ayman Al Zawahiri | Facebook
Ayman AL Zawahiri | Facebook
Ayman al-Zawahiri | Facebook
Ayman al-Zawahiri | Facebook
Ayman al-Zawahiri | FOCUS on ISRAEL
Ayman al-Zawahiri | Il Peccatore
ayman al zawahiri | Libero 24x7
Ayman al Zawahiri | News NanoPress
Ayman al Zawahiri | Rassegna Stampa...
Ayman al-Zawahiri | World on Focus
Ayman Al Zawahiri – archivio900.it
Ayman al Zawahiri Foto | NanoPress
(d) Ayman al-Zawahiri - IlGiornale.it
(d) Ayman al zawahiri in Blogbabel
Ayman al zawahiri in Blogbabel
Ayman Al Zawahiri nell'Enciclopedia...
ayman al zawahiri - News
Ayman Al-zawahiri - News - Sky Tg24 - Sk
Ayman Al Zawahiri - Notizie in Liquida
(d) Ayman al Zawahiri - Repubblica.it
Ayman al Zawahiri - Repubblica.it
(d) Ayman al-Zawahiri - Shopping-enable
Ayman al-Zawahiri - Shopping-enabled W
Ayman Al Zawahiri - The Blazoned Press
Ayman al Zawahiri - Ultime notizie su Ayn

Properties view showing meta information of a specific bookmark file selected

Property	Value
Info	
derived	false
editable	true
last modified	9 September 2013 11:48:37
linked	false
location	C:\work\workspaces\osint-workspace\Aiman Al-Zawahiri\Bookmarks\Bing_2013-09-09_11-48\tg24.sky.it.tg24.mondo.2011.06.16.al.qaeda.al.zawahiri.successione.b...
name	tg24.sky.it.tg24.mondo.2011.06.16.al.qaeda.al.zawahiri.successione.bin.laden.terrorismo.html.obm
path	/Aiman Al-Zawahiri/Bookmarks/Bing_2013-09-09_11-48/tg24.sky.it.tg24.mondo.2011.06.16.al.qaeda.al.zawahiri.successione.bin.laden.terrorismo.html.obm
size	786 bytes

http://tg24.sky.it/tg24/mondo/2011/06/16/al_qaeda_al_zawahiri_successione_bin_laden_terrorismo.html

If several searches are carried out on the same Case Project, a sign ***(d)*** might appear on the left side of the *Bookmarks* folders. It means that there are some duplicate links inside that folder regarding all the links downloaded for the same Case Project. To discard them, click on **Web > Delete Duplicate Bookmarks**.


EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help
Workspace Navigator
Entity Browser
Aiman Al-Zawahiri
Analysis
(d) Bookmarks
(d) Bing_2013-09-09_11-48
Google_2013-09-09_11-48
(d) Yahoo_2013-09-09_11-49
Crawler
Documents

Reviewing the bookmark files

In order to review the result links faster without opening a new browser window each time, you can select from the main menu **Window > Reuse Browser** . If this menu item is enabled the current browser editor view is reused and loads the link of the bookmark file with a simple click on the file in the **Workspace Navigator** view.

If you do a double-click on a bookmark file you can still force to open a *new* browser editor. If you want to delete irrelevant bookmark files, you can simply right-click the file and click **Delete**.

The system automatically detects bookmarks pointing to the same URL (so-called duplicates). These duplicates are marked with a *(d)* prefix in the **Workspace Navigator** view. For each duplicate bookmark there exists another bookmark pointing to the same URL but with an older time stamp. In order to delete all duplicates under a certain folder: right-click on the folder in the **Workspace Navigator** view and click **Organize Bookmarks > Delete Duplicates**. The duplicate detection is handy if you gather search results from multiple search engines and merge the results before you download the result pages to your case project.

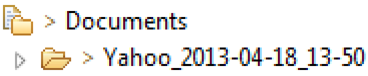
See also: Using the duplicate bookmark detection and Delete duplicate bookmarks.

Performing the Entity Extraction

The goal of the Entity Extraction process is to find locations in the text which contain *entity information*. In other words, it tries to find occurrences of person names, locations, organizations, VAT numbers, etc.

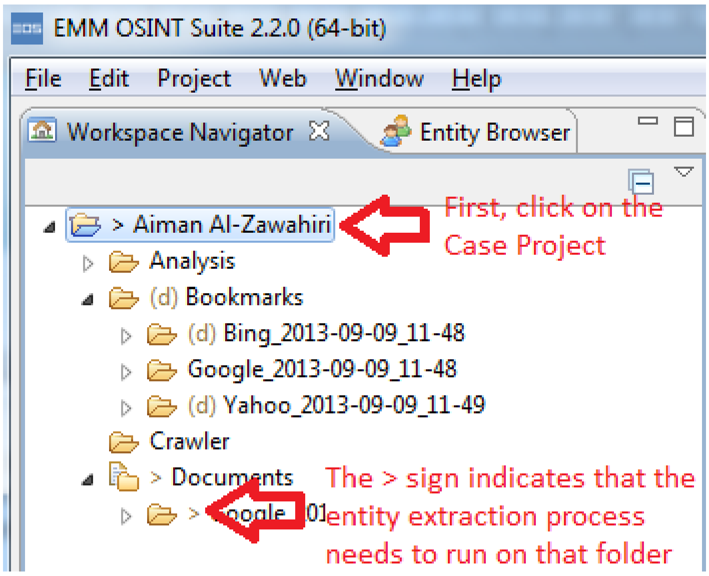
The entity extraction finds entities in the set of documents located in the ***Documents*** folder of a Case Project. The ***Documents*** folder is a *special predefined* folder which contains all input documents for the entity extraction. This means that **download bookmarks or import documents from local disk into the *Document* folder is a prerequisite for running the entity extraction process**.

The ***Documents*** folder is marked with a special icon: . Whenever the *Documents* folder shows a greater sign in front of its name, the entity extraction needs to run:



You can run the entity extraction as follows:

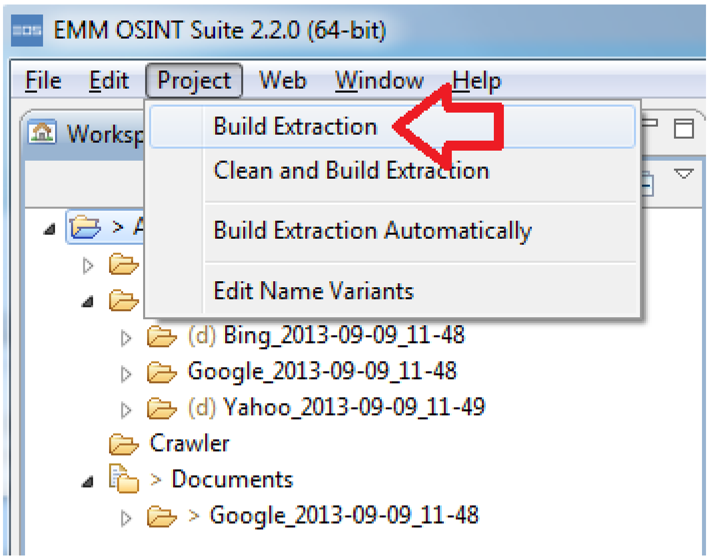
- In the **Workspace Navigator** view click on the case project containing the *Documents* folder which is not up-to-date.



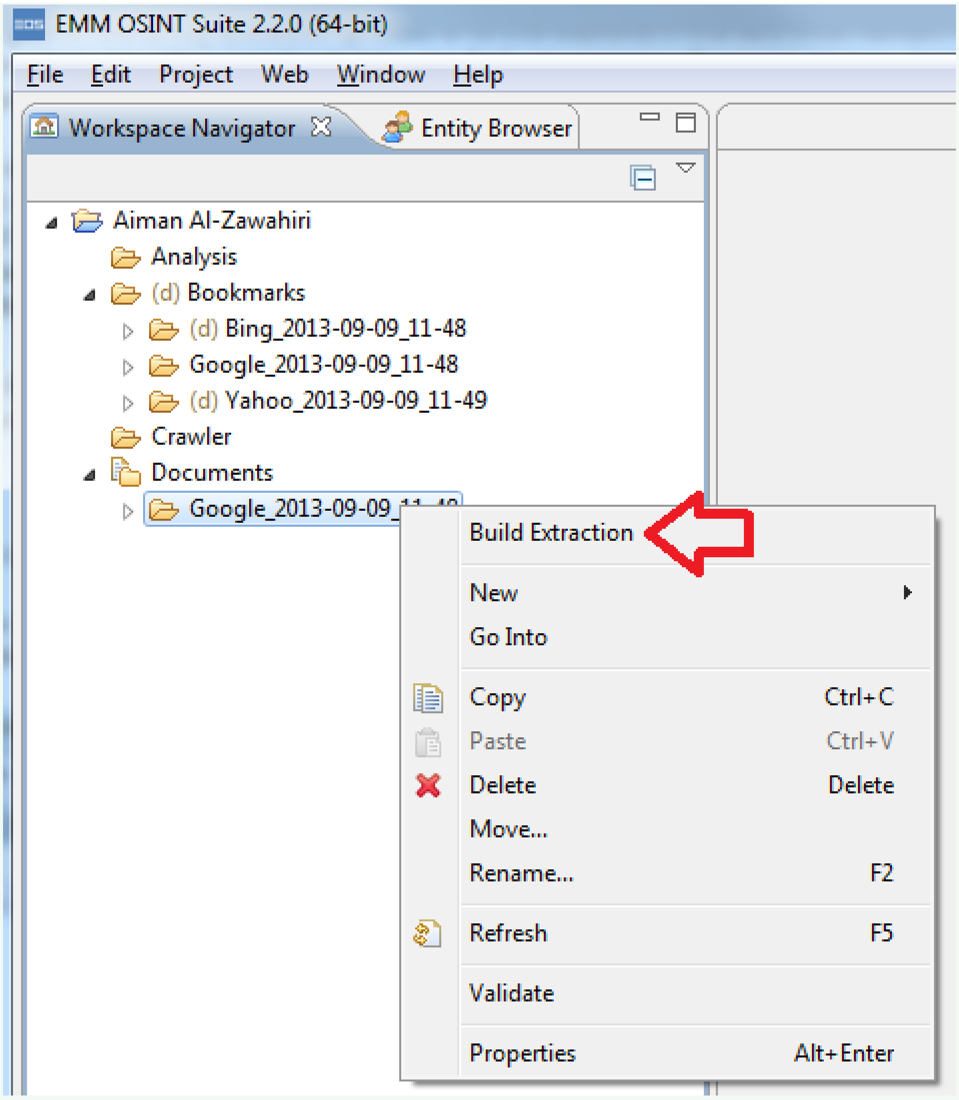
First, click on the Case Project

The > sign indicates that the entity extraction process needs to run on that folder

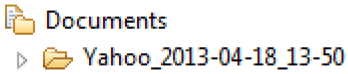
- In the main menu click on **Project > Build Extraction**. The entity extraction starts to run, you can monitor the progress in the **Progress View**


EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help
Build Extraction
Clean and Build Extraction
Build Extraction Automatically
Edit Name Variants
Worksp
> A
(d) Bing_2013-09-09_11-48
Google_2013-09-09_11-48
(d) Yahoo_2013-09-09_11-49
Crawler
> Documents
> Google_2013-09-09_11-48

Alternatively, you can use the context menu to run the Entity Extraction. Right-click on the *Documents* folder in the **Workspace Navigator** view, then click **Build Extraction**. The entity extraction starts to run.



As soon as the entity extraction has finished, the *Documents* folder is no longer marked with the **>** sign:



Requesting Support

If you experience any problems with the software or would like to suggest a new feature, please contact us.

Submitting an Issue by Email

Write an email to mailto:xxxxxxx.xxxxxx@xx.xxxxxx.xx and include as many details as possible in your email.

- Attach the log file to your email which allows us to find the error cause more quickly:
    1. Open you current workspace directory using the **Windows File Explorer** and open the sub-directory *.metadata/.plugins/it.jrc.osint.logging*
    2. The sub-directory contains the log file named *osint.log* (possibly along with some older archived log files)
    3. Zip the *osint.log* file and attach it to your email

Note: Normally, the log file should not contain any confidential data. However, you may want to review it before sending it to us.

Running on Linux

Prerequisite: Installing on Linux (Ubuntu)

Once the EMM OSINT Suite has been installed on your system, running the application requires you top define a workspace folder on a local disk drive to keep your user data.

⚠ **Important**
The workspace folder needs to be placed on a local disk drive. The application ***will fail*** if the workspace is on a network drive.

By default the application places the user data in the default user home directory */home/<user-id>/osint-workspace*.

To run the application do the following:

- Open a Terminal window
- Navigate to the installation location /opt/osint_<version>_linux.x86_64 (for the 64-bit version)

```
cd /opt/osint_2.3.1_linux.x86_64
```

- Starting the application from a terminal window as follows

```
env UBUNTU_MENUPROXY=0 ./osint&
```

> **Ubuntu Unity Global Menu Workaround**
> Ubuntu's Unity desktop environment display the menu of an application in the global menu bar. This causes problems with the EMM OSINT Suite (and other programs). To disable the global menu please set the environment variable UBUNTU_MENUPROXY as shown above.

Problems starting the application

If the application does not start, make sure that the following files are executable:

/opt/osint_<version-number>_linux_x86_64/osint
/opt/osint_<version-number>_linux_x86_64/jre/bin/java

Under Ubuntu you can set the execute permissions with the following commands:

```
sudo chmod ugo+x /opt/osint_2.3.1_linux.x86_64/osint
sudo chmod ugo+x /opt/osint_2.3.1_linux.x86_64/jre/bin/java
```

Running on Windows

Prerequisite: Installing on Windows

Once the EMM OSINT Suite has been installed on your system, running the application requires you to **define a workspace folder on local disk** to keep your user data.

> **Important**
> The workspace folder needs to be placed on a local disk drive. The application ***will fail*** if the workspace is on a network drive.

By default the application places the user data in the default **Windows User Directory** under *c:\Users\<User-Name>\osint-workspace*.

To run the application do the following:

1. Open the installation location in Windows File Explorer. By default the location is *c:\Program Files\osint_<version-number>_x86_64*.
2. Double-click on the ***osint.exe*** executable to start the application. The application shows a splash screen, then shows the **Workspace Selection** dialog.
3. In the **Workspace Selection** dialog either
    - Click **OK** to choose the default workspace location
    - or Click **Browse** to define a custom workspace location. The application shows a file dialog to choose the custom workspace location.
4. In the Workspace Selection dialog you can also enable the **Remember Workspace Location** check box. If enabled, the application will not show the Workspace Selection dialog during next startup but simply re-use the selected location.

> You can switch to a different workspace location from within the running application by clicking on **File > Switch Workspace** in the main menu. This option is needed if you enabled **Remember Workspace Location** previously in the **Workspace Selection** dialog and want to change to a different workspace location later on.

Startup Fails - What now?

[Windows 32-bit only]

If starting the application fails and you see a message "failed to create Java Virtual Machine", you can try the following:

- Edit *osint.ini* in the installation directory
- Change the line "-Xmx924m" which defines the maximum amount of memory for the application to a lower number (but not much lower, maybe something like "-Xmx850m")
- Save *osint.ini*
- Double-click on *osint.exe* to start again

Repeat if this procedure if necessary.

The reason is that the underlying Java VM needs a contiguous memora area. If there are a lot of other applications open and drivers loaded the system is not able to allocated the needed memory area.

(The 64-bit version does not suffer from this limitations)

Related Topics:

- Requesting Support

Setting HTTP Proxy Information

The application tries to automatically detect the system-wide proxy settings. However, if the proxy requires authentication this authentication information is not provided by the operating system and must be set manually.

Complete the following steps:

- Detecting the needed proxy settings
- Opening the Network Connections preference panel
- Providing Proxy Settings for different protocols

Detecting the Needed Proxy Settings

If the needed proxy settings are not known, the *Verify Network Connection* feature can be used to detect proxy information on the system:

- Click on **Help > Verify Network Connection**

The system prints the found proxy information in the **Console View**.

Use the printed information to set the proxy information in the **Network Connections Preference** Panel.

> **Proxy Information Must be set Manually**
> Currently, the Verify Network Connection command does not automatically set the needed proxy information. Please set it manually as described below.

Opening the Network Connections Preference Panel

- Click on **Window > Preferences**
- Click to expand the **General** node and select **Network Connections**

Providing Proxy Settings for different protocols

- Click the **Active Provider** drop down and change to **Manual**
- Select in turn the HTTP and HTTPS protocol schemes from the **Proxy entries** list and click **Edit...** to change settings.

The system opens the **Edit Proxy Entry** dialog.

- Fill in the **Host** and **Port** fields with the proxy host and port.
- If the edited proxy schema requires authentication, enable the **Requires Authentication:** option.
    - Fill in **User** and **Password** for authentication
- Click **OK** to save the schema entry

> **SOCKS Scheme not supported**
> If the **Proxy entries** list contains a SOCKS protocol entry, clear it by selecting it and clicking on **Clear**. Some user reported problems if a SOCKS entry was present.

In the **Network Connections** preferences dialog click **OK** to save the new preference settings.

> **Changing Active Provider**
> Chaning the **Active Provider** back to **Native**, requires a manual clean of all entries in the **Proxy entries** list before closing the dialog. Otherwise the changes may not be properly propagated in the system.

Using Local Search

The OSINT Suite allows performing local searches within the documents previously downloaded or imported into OSINT, i.e. those found under the Documents folder. It is important to note that local searches must be always performed after the text extraction process ends.

The local search in the OSINT Suite provides a rich query language through which the user can perform wildcard searches or boolean queries. A query can be broken up into terms and operators. There are two types of terms: **single terms** and **phrases**. A single term is a single word such as "house" or "car". A phrase is a group of words surrounded by double quotes such as "white house". Multiple terms can be combined together with boolean operators to form a more complex query.

Wildcard searches

The local search supports single and multiple character wildcard searches within single terms (not within phrase queries):

- To perform a **single character wildcard** search use the "**?**" symbol. The single character wildcard search looks for terms that match that with the single character replaced. For example, to search for "text" or "test" you can use the search: te?t
- To perform a **multiple character wildcard** search use the "*****" symbol. Multiple character wildcard searches looks for 0 or more characters. For example, to search for test, tests or tester, you can use the search: test*. You can also use the wildcard searches in the middle of a term: te*t . Note that you cannot use a * or ? symbol as the first character of a search.

Boolean searches

Boolean operators allow terms to be combined through logic operators. These are the operators that our local search supports:

- AND
- +
- OR
- NOT or !
- -

<u>Note</u>**:** boolean operators **must be ALL CAPS**. Some examples of using boolean searches:

1. To search for documents that contain either "white house" or just "house" use one of the following queries (note: you can use double quotas for searching an exact phrase):

   - `"white house" house`
   - `"white house" OR house`

2. To search for documents that contain "white house" and "black house" use the query:

   - `“white house” AND “black house”`

3. The "+" or required operator requires that the term after the "+" symbol exist somewhere in the document. Thus, to search for documents that must contain "house" and may contain "white" use the query:

   - `+house white`

4. The NOT operator excludes documents that contain the term after NOT. This is equivalent to a difference using sets. The symbol ! can be used in place of the word NOT. To search for documents that contain "white house" but not "black house" use the query:

   - `"white house" NOT "black house"`

   Note: The NOT operator cannot be used with just one term. For example, the following search will return no results:

   - `NOT "white house"`

5. The "-" or prohibit operator excludes documents that contain the term after the "-" symbol. To search for documents that contain "white house" but not "black house" use the query:

   - `"white house" -"black house"`

Grouping

Our tool also supports using parentheses to group clauses to form sub queries. This can be very useful if you want to control the boolean logic for a query. To search for either "white" or "black" and "house" use the query:

- `(white OR black) AND house`

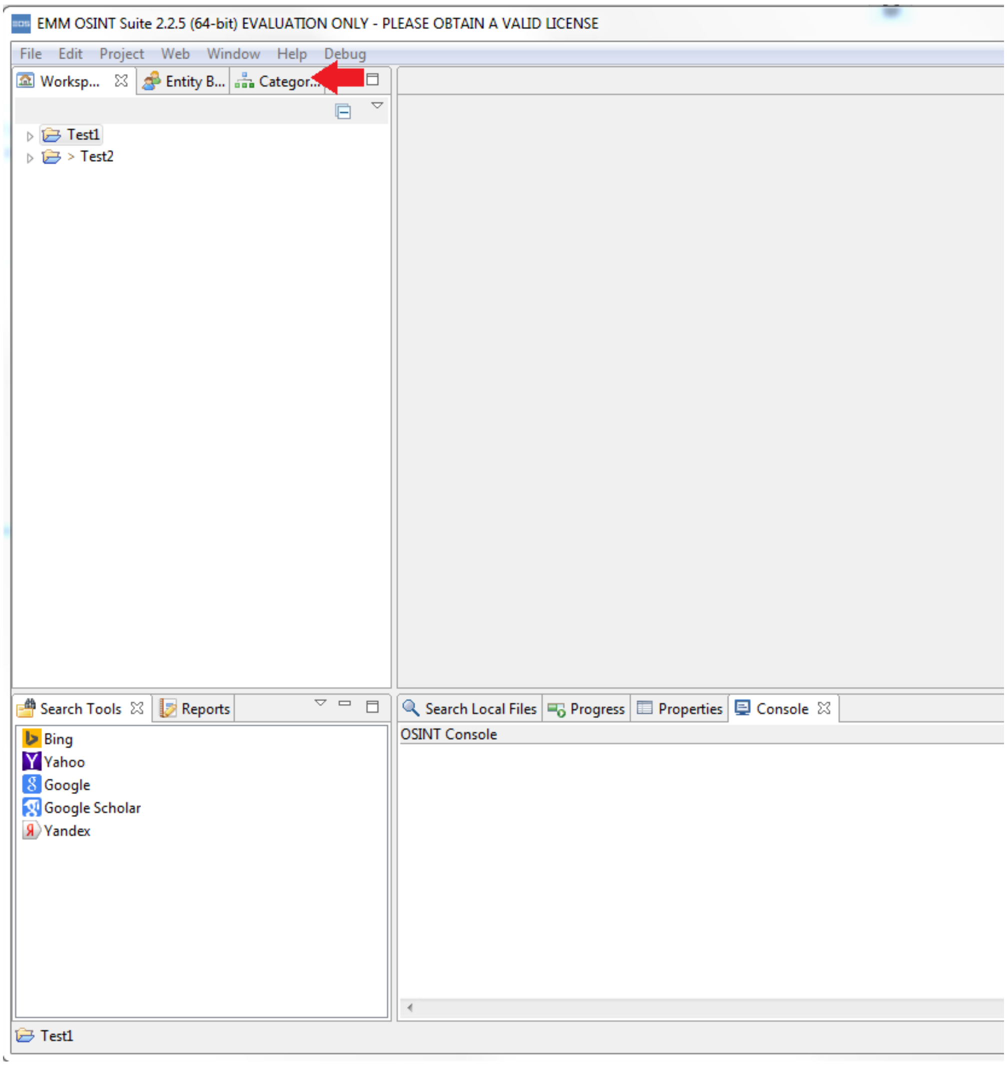
Using the Category Browser view

Prerequisites:

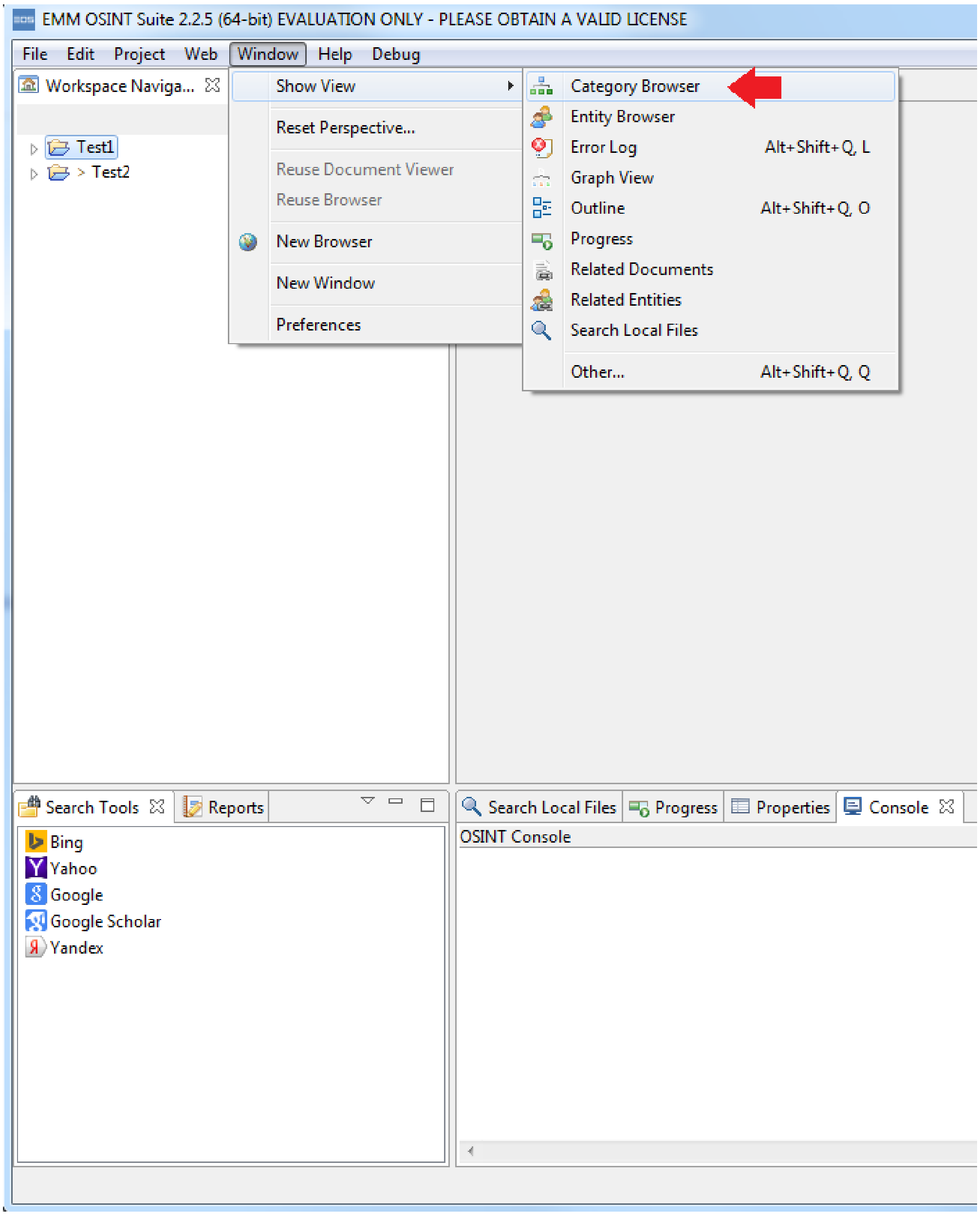
- Creating a Configuration Project
- Creating a Category Definition File

The **Category Browser view** is used to find which documents belong to any category previously predefined in the OSINT Suite. A **category** or **alert** can be seen as a file in which several keywords or even combinations of keywords are defined by the user. Therefore, before using the Category Browser view, a Configuration Project should be created and then one **Category Definition File** should be generated at least in order to check how the Category Browser view works.

The **Category Browser view** is usually opened behind the Entity Browser view:

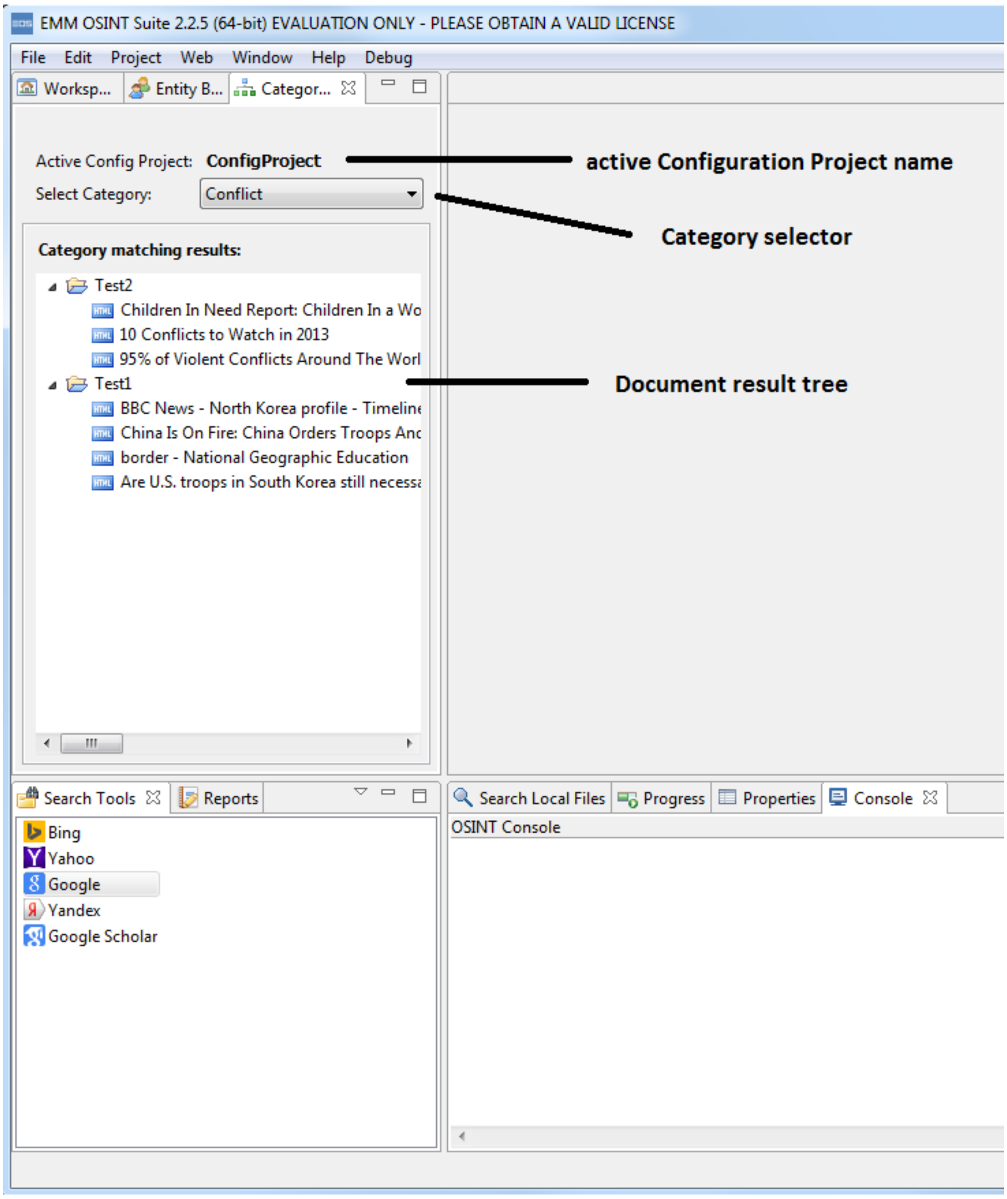


If it is not shown behind the Entity Browser tab once OSINT starts, then it can be activated by clicking on **Window > Show View > Category Browser**. To bring it to front, click its title bar.



ⓘ After activating the Category Browser view, a message indicating active Configuration Project was not found may be shown if there is no an active Configuration Project in the workspace.

The view consists of the following elements:



Active Config Project: **ConfigProject**

Select Category: Conflict

Category matching results:

- Test2
  - Children In Need Report: Children In a Wo
  - 10 Conflicts to Watch in 2013
  - 95% of Violent Conflicts Around The Worl
- Test1
  - BBC News - North Korea profile - Timeline
  - China Is On Fire: China Orders Troops Anc
  - border - National Geographic Education
  - Are U.S. troops in South Korea still necess

active Configuration Project name

Category selector

Document result tree

- **Active Configuration Project name**: it shows the name of the current Configuration Project activated in the workspace.
- **Category selector**: it is used to select one of the existing active categories predefined within the Configuration Project currently active.
- **Document result tree**: this is a tree viewer which shows the downloaded documents that match with the selected category. They are shown ordered by Case Project.

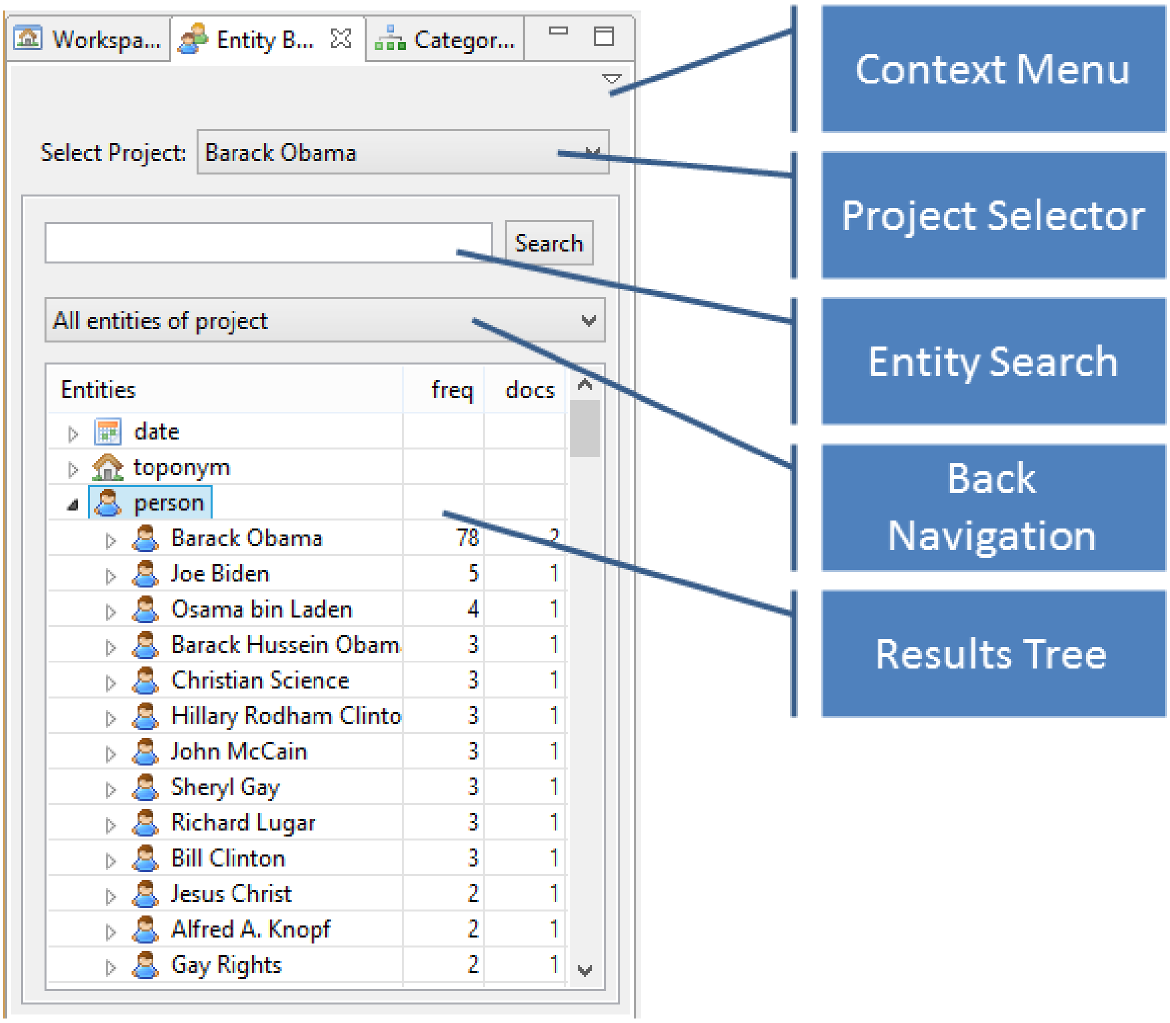
Using the Entity Browser view

Prerequisite: Performing the Entity Extraction

The **Entity Browser** view can be used to browse the entity information extracted from the set of documents. By default the **Entity Browser** v iew is opened behind the **Workspace Navigator** view. To bring it to front, click its title bar.

Elements of the Entity Browser View

The view consists of the following elements:



Select Project: Barack Obama

All entities of project

Entities	freq	docs
date		
toponym		
person		
Barack Obama	78	2
Joe Biden	5	1
Osama bin Laden	4	1
Barack Hussein Obam	3	1
Christian Science	3	1
Hillary Rodham Clinto	3	1
John McCain	3	1
Sheryl Gay	3	1
Richard Lugar	3	1
Bill Clinton	3	1
Jesus Christ	2	1
Alfred A. Knopf	2	1
Gay Rights	2	1

Context Menu:

The context menu can be accessed by clicking the small triangle next to the title tab. The context menu contains the following actions:

- Refresh - forces the system to reload the data in the Entity Browser
- Filter Entity Types... - Allows to filter out certain entity types
- Sort Entity By
    - Alphabetic - orders entities in the Results Tree in alphabetic order
    - Frequency - orders entities in the Results Tree by number of occurrences (frequency)
    - Related Docs - orders entities in the Results Tree by number of related documents (documents which contain the entity)

Project Selector:

The project selector is used to select one of the existing case projects in the workspace.

Entity Search field:

The search field is used to search for specific entities. It supports * as a wildcard for any characters. For example, the search term "Barac*" will find any entity starting with "Barac". Enter a search term and click Search. The results will be shown in the result tree.

Back Navigation:

This drop-down element shows a list of previously defined search operations. Select one to go back to previous search results.

Entity Result Tree:

This is a tree-like viewer which shows the found entities ordered by type.

The viewer has two additional columns:

- Freq - Frequency of Occurrence, the number of occurrences of an entity across all documents
- Docs - Related Documents, the number of documents which contain the entity

By default the columns show the frequency and related documents of an entity across all documents in the case project. If a search for a specific entity or entities related to a specific entity is active (see the Back Navigation drop down) the data in relation to the current search is shown.

Entity Context Menu:

Right-click on an entity to show the Entity Context Menu. The context menu contains the following actions:

- Add to Graph - adds the selected entity to the Graph view
- Show related entities - shows the related entities of the entity in the Browser view. This action is added to the Back Navigation. This action switches the data in the frequency and docs column in relation to the selected entity.

Using the Entity Browser view

The view shows all found entities in a result tree ordered by entity type.

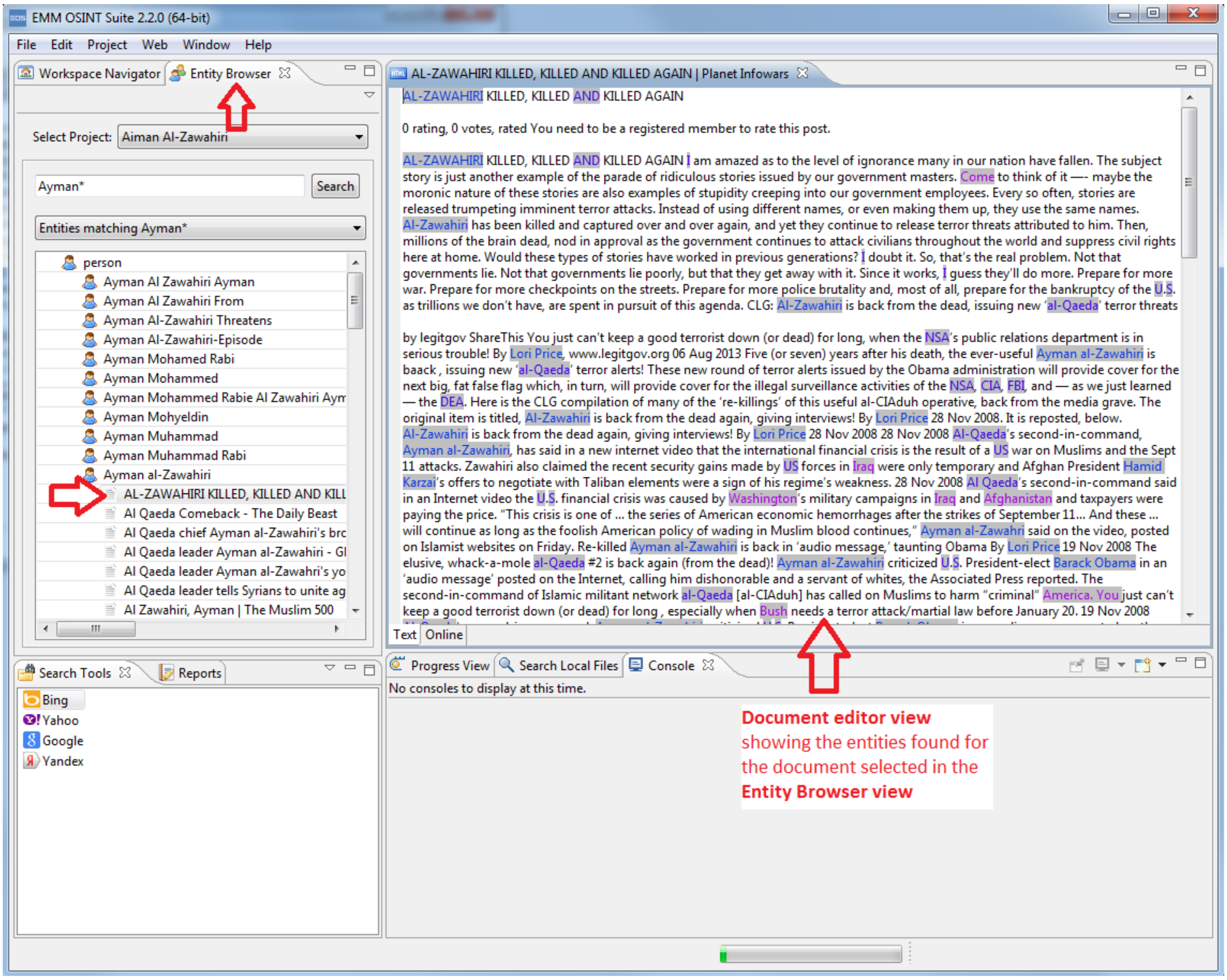
Searching for a specifc Entity

In order to search for a specifc entity, do the following:

1. Enter the name of the entity or the start of the name in the Entity Search Field
2. Click Search or hit the enter key

The search field supports the use of wild card symbols to search for parts of names, for example “Franz*” searches for all entities which start with “Franz”. The wildcard patterns support *?* to replace a single character and *** to replace one or multiple characters. By default the system adds the * wildcard pattern to the end of a term to match all entities starting with this term. If you want to search for an exact term, enclose the term in double quotes.

In the result tree either all entities or the entities matching the search term are shown. Expand the result tree to the next level to view the documents where the different entities are found. Double-click on a document icon in the result tree to open the document in the **Document editor view**.



Select Project: Aiman Al-Zawahiri

Ayman*

Entities matching Ayman*

person
Ayman Al Zawahiri Ayman
Ayman Al Zawahiri From
Ayman Al-Zawahiri Threatens
Ayman Al-Zawahiri-Episode
Ayman Mohamed Rabi
Ayman Mohammed
Ayman Mohammed Rabie Al Zawahiri Aym
Ayman Mohyeldin
Ayman Muhammad
Ayman Muhammad Rabi
Ayman al-Zawahiri
AL-ZAWAHIRI KILLED, KILLED AND KILL
Al Qaeda Comeback - The Daily Beast
Al Qaeda chief Ayman al-Zawahiri's brc
Al Qaeda leader Ayman al-Zawahiri - Gl
Al Qaeda leader Ayman al-Zawahri's yo
Al Qaeda leader tells Syrians to unite ag
Al Zawahiri, Ayman | The Muslim 500

AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN

0 rating, 0 votes, rated You need to be a registered member to rate this post.

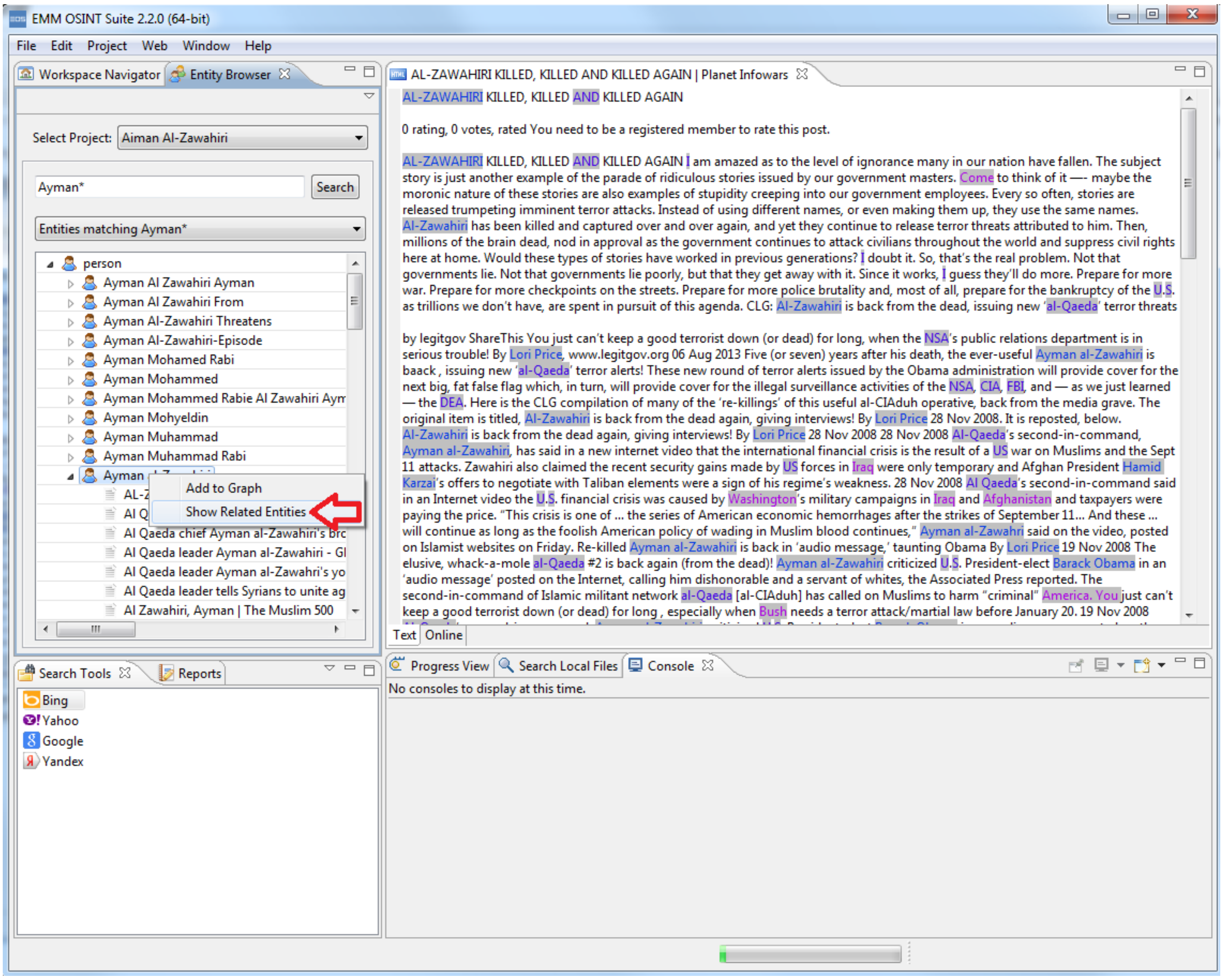
AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN I am amazed as to the level of ignorance many in our nation have fallen. The subject story is just another example of the parade of ridiculous stories issued by our government masters. Come to think of it —- maybe the moronic nature of these stories are also examples of stupidity creeping into our government employees. Every so often, stories are released trumpeting imminent terror attacks. Instead of using different names, or even making them up, they use the same names. Al-Zawahiri has been killed and captured over and over again, and yet they continue to release terror threats attributed to him. Then, millions of the brain dead, nod in approval as the government continues to attack civilians throughout the world and suppress civil rights here at home. Would these types of stories have worked in previous generations? I doubt it. So, that's the real problem. Not that governments lie. Not that governments lie poorly, but that they get away with it. Since it works, I guess they'll do more. Prepare for more war. Prepare for more checkpoints on the streets. Prepare for more police brutality and, most of all, prepare for the bankruptcy of the U.S. as trillions we don't have, are spent in pursuit of this agenda. CLG: Al-Zawahiri is back from the dead, issuing new 'al-Qaeda' terror threats

by legitgov ShareThis You just can't keep a good terrorist down (or dead) for long, when the NSA's public relations department is in serious trouble! By Lori Price, www.legitgov.org 06 Aug 2013 Five (or seven) years after his death, the ever-useful Ayman al-Zawahiri is baack , issuing new 'al-Qaeda' terror alerts! These new round of terror alerts issued by the Obama administration will provide cover for the next big, fat false flag which, in turn, will provide cover for the illegal surveillance activities of the NSA, CIA, FBI, and — as we just learned — the DEA. Here is the CLG compilation of many of the 're-killings' of this useful al-CIAduh operative, back from the media grave. The original item is titled, Al-Zawahiri is back from the dead again, giving interviews! By Lori Price 28 Nov 2008. It is reposted, below. Al-Zawahiri is back from the dead again, giving interviews! By Lori Price 28 Nov 2008 28 Nov 2008 Al-Qaeda's second-in-command, Ayman al-Zawahiri, has said in a new internet video that the international financial crisis is the result of a US war on Muslims and the Sept 11 attacks. Zawahiri also claimed the recent security gains made by US forces in Iraq were only temporary and Afghan President Hamid Karzai's offers to negotiate with Taliban elements were a sign of his regime's weakness. 28 Nov 2008 Al Qaeda's second-in-command said in an Internet video the U.S. financial crisis was caused by Washington's military campaigns in Iraq and Afghanistan and taxpayers were paying the price. "This crisis is one of ... the series of American economic hemorrhages after the strikes of September 11... And these ... will continue as long as the foolish American policy of wading in Muslim blood continues," Ayman al-Zawahri said on the video, posted on Islamist websites on Friday. Re-killed Ayman al-Zawahiri is back in 'audio message,' taunting Obama By Lori Price 19 Nov 2008 The elusive, whack-a-mole al-Qaeda #2 is back again (from the dead)! Ayman al-Zawahiri criticized U.S. President-elect Barack Obama in an 'audio message' posted on the Internet, calling him dishonorable and a servant of whites, the Associated Press reported. The second-in-command of Islamic militant network al-Qaeda [al-CIAduh] has called on Muslims to harm "criminal" America. You just can't keep a good terrorist down (or dead) for long , especially when Bush needs a terror attack/martial law before January 20. 19 Nov 2008

No consoles to display at this time.

Document editor view showing the entities found for the document selected in the **Entity Browser view**

Right-click on an entity, and click **Show related entities** to show all related entities in the **Entity Browser** view. In order to go back, select a previous set of entities from the drop down menu which contains a list of previous actions.



AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN

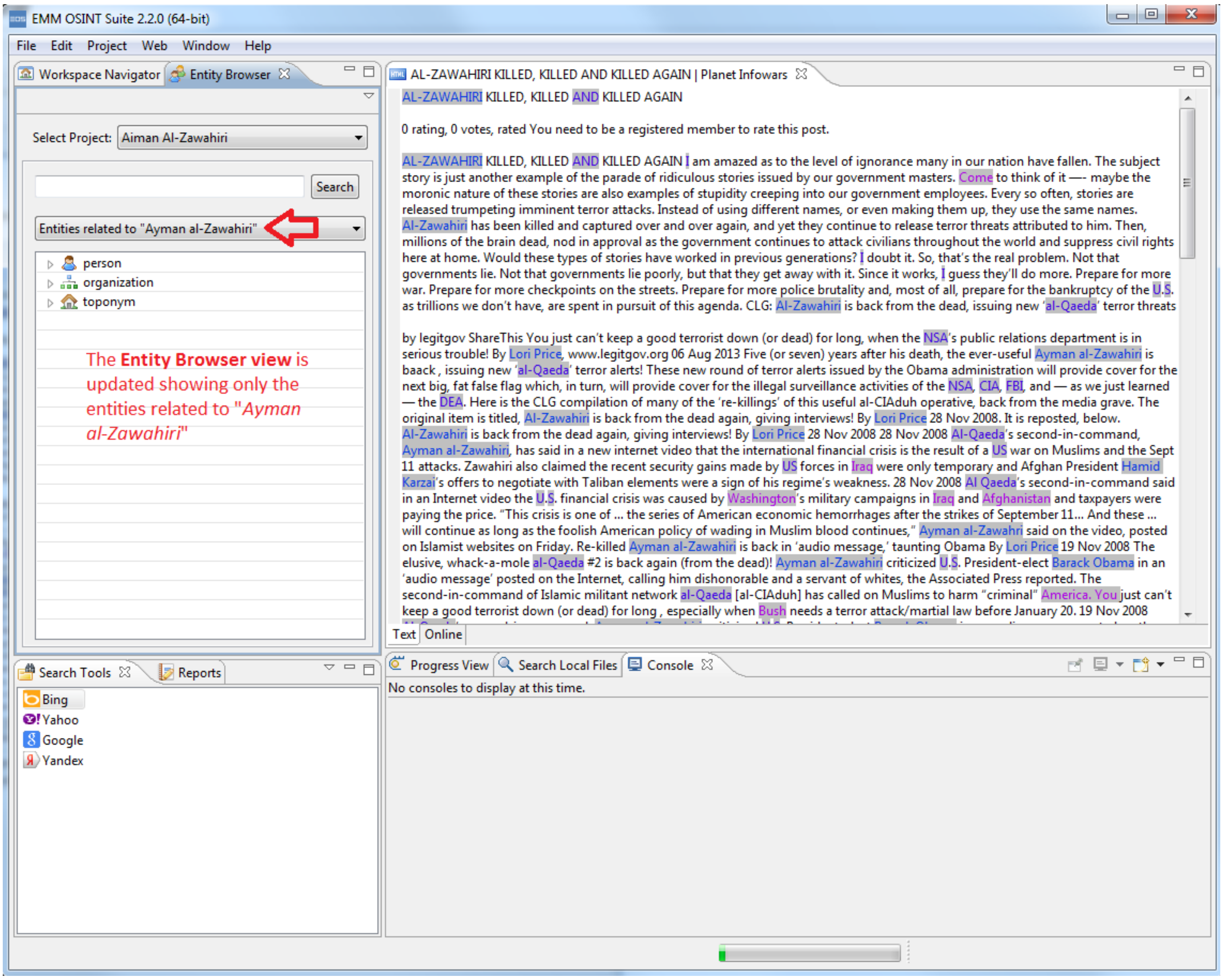
0 rating, 0 votes, rated You need to be a registered member to rate this post.

AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN I am amazed as to the level of ignorance many in our nation have fallen. The subject story is just another example of the parade of ridiculous stories issued by our government masters. Come to think of it —- maybe the moronic nature of these stories are also examples of stupidity creeping into our government employees. Every so often, stories are released trumpeting imminent terror attacks. Instead of using different names, or even making them up, they use the same names. Al-Zawahiri has been killed and captured over and over again, and yet they continue to release terror threats attributed to him. Then, millions of the brain dead, nod in approval as the government continues to attack civilians throughout the world and suppress civil rights here at home. Would these types of stories have worked in previous generations? I doubt it. So, that's the real problem. Not that governments lie. Not that governments lie poorly, but that they get away with it. Since it works, I guess they'll do more. Prepare for more war. Prepare for more checkpoints on the streets. Prepare for more police brutality and, most of all, prepare for the bankruptcy of the U.S. as trillions we don't have, are spent in pursuit of this agenda. CLG: Al-Zawahiri is back from the dead, issuing new 'al-Qaeda' terror threats

by legitgov ShareThis You just can't keep a good terrorist down (or dead) for long, when the NSA's public relations department is in serious trouble! By Lori Price, www.legitgov.org 06 Aug 2013 Five (or seven) years after his death, the ever-useful Ayman al-Zawahiri is baack , issuing new 'al-Qaeda' terror alerts! These new round of terror alerts issued by the Obama administration will provide cover for the next big, fat false flag which, in turn, will provide cover for the illegal surveillance activities of the NSA, CIA, FBI, and — as we just learned — the DEA. Here is the CLG compilation of many of the 're-killings' of this useful al-CIAduh operative, back from the media grave. The original item is titled, Al-Zawahiri is back from the dead again, giving interviews! By Lori Price 28 Nov 2008. It is reposted, below. Al-Zawahiri is back from the dead again, giving interviews! By Lori Price 28 Nov 2008 28 Nov 2008 Al-Qaeda's second-in-command, Ayman al-Zawahiri, has said in a new internet video that the international financial crisis is the result of a US war on Muslims and the Sept 11 attacks. Zawahiri also claimed the recent security gains made by US forces in Iraq were only temporary and Afghan President Hamid Karzai's offers to negotiate with Taliban elements were a sign of his regime's weakness. 28 Nov 2008 Al Qaeda's second-in-command said in an Internet video the U.S. financial crisis was caused by Washington's military campaigns in Iraq and Afghanistan and taxpayers were paying the price. "This crisis is one of ... the series of American economic hemorrhages after the strikes of September 11... And these ... will continue as long as the foolish American policy of wading in Muslim blood continues," Ayman al-Zawahri said on the video, posted on Islamist websites on Friday. Re-killed Ayman al-Zawahiri is back in 'audio message,' taunting Obama By Lori Price 19 Nov 2008 The elusive, whack-a-mole al-Qaeda #2 is back again (from the dead)! Ayman al-Zawahiri criticized U.S. President-elect Barack Obama in an 'audio message' posted on the Internet, calling him dishonorable and a servant of whites, the Associated Press reported. The second-in-command of Islamic militant network al-Qaeda [al-CIAduh] has called on Muslims to harm "criminal" America. You just can't keep a good terrorist down (or dead) for long , especially when Bush needs a terror attack/martial law before January 20. 19 Nov 2008

No consoles to display at this time.

Now the Entity Browser view is updated showing only the related entities to the previously selected one


EMM OSINT Suite 2.2.0 (64-bit)
File Edit Project Web Window Help
Workspace Navigator
Entity Browser
Select Project: Aiman Al-Zawahiri
Search
Entities related to "Ayman al-Zawahiri"
person
organization
toponym
Search Tools
Reports
Bing
Yahoo
Google
Yandex
AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN | Planet Infowars
Text
Online
Progress View
Search Local Files
Console

The **Entity Browser view** is updated showing only the entities related to "*Ayman al-Zawahiri*"

AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN

0 rating, 0 votes, rated You need to be a registered member to rate this post.

AL-ZAWAHIRI KILLED, KILLED AND KILLED AGAIN I am amazed as to the level of ignorance many in our nation have fallen. The subject story is just another example of the parade of ridiculous stories issued by our government masters. Come to think of it —- maybe the moronic nature of these stories are also examples of stupidity creeping into our government employees. Every so often, stories are released trumpeting imminent terror attacks. Instead of using different names, or even making them up, they use the same names. Al-Zawahiri has been killed and captured over and over again, and yet they continue to release terror threats attributed to him. Then, millions of the brain dead, nod in approval as the government continues to attack civilians throughout the world and suppress civil rights here at home. Would these types of stories have worked in previous generations? I doubt it. So, that's the real problem. Not that governments lie. Not that governments lie poorly, but that they get away with it. Since it works, I guess they'll do more. Prepare for more war. Prepare for more checkpoints on the streets. Prepare for more police brutality and, most of all, prepare for the bankruptcy of the U.S. as trillions we don't have, are spent in pursuit of this agenda. CLG: Al-Zawahiri is back from the dead, issuing new 'al-Qaeda' terror threats

by legitgov ShareThis You just can't keep a good terrorist down (or dead) for long, when the NSA's public relations department is in serious trouble! By Lori Price, www.legitgov.org 06 Aug 2013 Five (or seven) years after his death, the ever-useful Ayman al-Zawahiri is baack , issuing new 'al-Qaeda' terror alerts! These new round of terror alerts issued by the Obama administration will provide cover for the next big, fat false flag which, in turn, will provide cover for the illegal surveillance activities of the NSA, CIA, FBI, and — as we just learned — the DEA. Here is the CLG compilation of many of the 're-killings' of this useful al-CIAduh operative, back from the media grave. The original item is titled, Al-Zawahiri is back from the dead again, giving interviews! By Lori Price 28 Nov 2008. It is reposted, below. Al-Zawahiri is back from the dead again, giving interviews! By Lori Price 28 Nov 2008 28 Nov 2008 Al-Qaeda's second-in-command, Ayman al-Zawahiri, has said in a new internet video that the international financial crisis is the result of a US war on Muslims and the Sept 11 attacks. Zawahiri also claimed the recent security gains made by US forces in Iraq were only temporary and Afghan President Hamid Karzai's offers to negotiate with Taliban elements were a sign of his regime's weakness. 28 Nov 2008 Al Qaeda's second-in-command said in an Internet video the U.S. financial crisis was caused by Washington's military campaigns in Iraq and Afghanistan and taxpayers were paying the price. "This crisis is one of ... the series of American economic hemorrhages after the strikes of September 11... And these ... will continue as long as the foolish American policy of wading in Muslim blood continues," Ayman al-Zawahri said on the video, posted on Islamist websites on Friday. Re-killed Ayman al-Zawahiri is back in 'audio message,' taunting Obama By Lori Price 19 Nov 2008 The elusive, whack-a-mole al-Qaeda #2 is back again (from the dead)! Ayman al-Zawahiri criticized U.S. President-elect Barack Obama in an 'audio message' posted on the Internet, calling him dishonorable and a servant of whites, the Associated Press reported. The second-in-command of Islamic militant network al-Qaeda [al-CIAduh] has called on Muslims to harm "criminal" America. You just can't keep a good terrorist down (or dead) for long , especially when Bush needs a terror attack/martial law before January 20. 19 Nov 2008

No consoles to display at this time.

Using the Graph view

The **Graph** view is automatically shown if you add an entity to it. It shows the co-occurrence relationship between entities in the case project. Right-click on a single or multiple selected entities in the **Entity Browser** view and click **Add to Graph** to add it the the **Graph** view. Whenever an entity is added to the **Graph** view, the system tries to find existing co-occurrence relationships with the entities already shown in the graph.

Note: The search for co-occurrences may block the user interface for a short while, this is a known-issue and will be fixed in one of the future versions of the software

Double-click on the title tab of the **Graph** view to maximize it. A click on the small triangle to the right of the view's title tab reveals the view's context menu. To refresh the **Graph** view click on the triangle and then click **Refresh**.

Double-click on the edge connecting two entities in the graph to show a dialog with the documents establishing the relationship between the two entities. From the connection dialog you can select one or many documents to open them in the editor area.

? Unknown Attachment

Installing on Mac OS X using VirtualBox

Prerequisite: Recent Apple Mac computer with a current OS X version and at least 8GB of RAM and 10GB free disk space (less RAM could work but may be slow). You need to to have administrative rights to install software.

We provide a virtual image with a Ubuntu 14.04LTS installation to be able to run the EMM OSINT Suite software on Mac OS X.

Installing VirtualBox

- Download VirtualBox from its web site https://www.virtualbox.org/
- Choose "VirtualBox for OS X hosts"
- Install VirtualBox from DMG image using installer

Downloading an Image File

- Download a preconfigured zipped image file for VirtualBox here: vbox_ubuntu-14.04.3-x86_64_osint.zip
- Copy it to a folder on your Mac (for example create Documents/virtual-box and drop it in)
- Unzip it, it creates a directory with some files

Launching the Image

- Open VirtualBox
- Select Machine > Add... and select the .vdi file in the unzipped directory from the step above
- Start the image by clicking on the green arrow in VirtualBox

Logging in to the Image

- Use the user name "osint" and the password "osint" to log on
- Start the EMM OSINT Suite by opening the folder on the desktop and double click on the "osint" executable

Frequently Asked Questions

General

What is the EMM OSINT Suite?

Licensing

How can I license the EMM OSINT Suite?

Technical Questions

How can I recover data from a corrupt workspace?

Result Link Extraction works, but the download of the resulting Bookmarks fails

Shall I download the 32-bit or 64-bit version?

The application startup fails due to "Companion shared library not found"

What are the System Requirements for OSINT Suite?

How can I license the EMM OSINT Suite?

The EMM OSINT Suite is available free of charge for public authorities and institutions. However, depending on where you work, the licensing process may vary. In order to obtain the software please review the following cases:

I work for a public authority or institution in one of the member states of the European Union

Please send us a signed license agreement. The agreement should be signed at least by a department manager (head of unit) or above and can then be used for the complete department or authority.

I work for the European Commission

Please send an email to xxxxxxx.xxxxxx@xx.xxxxxx.xx to obtain access to the software.

I work for an European Institution (Council, Parliament, Agency)

Please send us a signed license agreement. The agreement should be signed at least by a head of unit (or above) and can then be used for the complete unit (or institution).

I work for an International Institution

Please contact xxxxxxx.xxxxxx@xx.xxxxxx.xx we need to ask permission from our hierarchy. After your request has been reviewed we will contact you explaining the next steps (e.g. license agreement, etc.).

I work for a public authority or institution outside of the European Union

Please contact xxxxxxx.xxxxxx@xx.xxxxxx.xx we need to ask permission from our hierarchy. After your request has been reviewed we will contact you explaining the next steps (e.g. license agreement, etc.).

Where can I download the License Agreement

The latest model of the license is available from here:

File	Modified
EMM OSINTSuite license agreement.pdf	Sep 24, 2013 by Gerhard Wagner

Drag and drop to upload or browse for files

Where can I send the signed license to?

Please send the signed license via postal service to

European Commission, Joint Research Centre
To the attention of: Gerhard Wagner, TP 440
Via Enrico Fermi 2749
21027 Ispra
Italy

I have obtained a valid license, where can I download the software?

Please refer to the download section of this site.

How can I recover data from a corrupt workspace?

The workspace is a folder on disk which stores all your user data. The user data consists of all the files you have either downloaded from the Internet or imported from local disk plus some meta data the application creates when extracting information. Due to bugs in the software it is possible that the meta data is corrupted and the application cannot process it anymore. Since the application never changes the imported files, it is possible to recover most of the work and run the extraction again to re-create the correct meta data.

To recover data from a corrupted workspace, please perform the following steps:

1. Closing all EMM OSINT Suite application instances
2. Creating a new workspace
3. Copy your user files over

Closing all EMM OSINT Suite application instances

- Close all instances of EMM OSINT Suite
- Make sure a crashed osint process is no longer running in the background.
  - On Windows you can access the Task Manager by using ctrl-alt-del, look for osint.exe in the Processes tab and click End Process.
  - On Linux use top and kill to halt the process

Creating a new workspace

- Running a new instance of EMM OSINT Suite (Windows) (Linux)
- Choose a new workspace folder (e.g. "osint-workspace-new")
- Recreate the projects from the old workspace (see Creating a Case Project or Creating a Configuration Project)

Copying Case Project data from the old workspace to the new workspace

While the EMM OSINT Suite instance is running with the new workspace, you can copy the following user data simply using the **Windows File Explorer** or the command line on Linux from the old project directory to the new project directory:

- Bookmark files
- Crawler Configuration files
- files from the Documents folder
- modified and addef files from a custom Configuration Project

You **must not** copy files from the .metadata or .osint directory while an instance of the application is running.

Copying a modified Name Variant Database file

If you have modified the Name Variant Database (see Importing or updating the name variant database file) you can also copy the name variant database file across:

- Closing all instances of the EMM OSINT Suite
- Navigate to <old-workspace>/.metadata/.plugins/it.jrc.osint.extract/

- Copy entity_20.h2.db to the corresponding location in the new workspace

Result Link Extraction works, but the download of the resulting Bookmarks fails

The extraction of links from the result page of a search engine works, but I cannot download the resulting Bookmarks, what is the matter?

The software most likely failed to detect the correct proxy settings. The Browser View relies on an embedded system browser and therefore uses the system-wide settings. However, the operating system does not provide proxy authentication information so the automatic detection of proxy settings may have failed.

Please refer to **Setting HTTP Proxy Information** to find out how to detect and set the proxy information manually.

Shall I download the 32-bit or 64-bit version?

Choosing 32-bit or 64-bit

Currently we provide versions in 32- and 64-bit flavour for Microsoft Windows and Linux. If you run Windows XP please use the 32-bit version.

If you run Windows 7 or better, you can do the following to check whether you have already a 64-bit system:

- Open The Windows Explorer (Right mouse click on Start button)
- Check if the directory "c:\Program Files (x86)" exists on your system.
  - If yes, download the 64-bit version
  - If no, download the 32-bit version

Note: The 32-bit version runs also on 64-bit systems. However, if supported by your system, prefer the 64-bit version.

The application startup fails due to "Companion shared library not found"

The application startup may fail with the error message "Companion shared library not found". A possible reason is that the companion shared library (which is a DLL under Windows) may have been taken away by an anti-virus scanner running on your system.

Check if the companion shared library can still be found in your installation directory:

- Open the installation directory (e.g. *osint_2.2.3_win32.x86_64*) in the **Windows File Explorer**
- Open the sub directory *plugins*
- Check if there is a sub directory starting with *org.eclipse.equinox.launcher.win32.win32.x86_64<version-id>*
- Check if the sub directory contains a file *eclipse_<build-id>.dll* (the current build-id is 1503 but may change for newer releases)

If the shared library is not there, please check your anti-virus scanner if it has "quarantined" this DLL.

Running without the eclipse launcher and the companion shared library

If due to restrictions on your PC the companion shared library cannot be used, the software can also be run from the command line:

- Open the Windows Command Prompt by selecting it from the Start Menu (usually it can be found under Accessories)
- In the Command Prompt change to the installation directory of the application (for example if it is installed in the program folder)

```
cd "c:\Program Files\osint_2.2.3_win32.x86_64"
```

- Start the application from the command line as follows:

```
jre\bin\java.exe -Xmx3072m -jar
plugins\org.eclipse.equinox.launcher_1.3.0.v20130327-1440.jar -data @noDefault
```

What are the System Requirements for OSINT Suite?

Microsoft Windows

The system requires Microsoft Windows 7 or better (older version such as Microsoft Windows XP and Microsoft Vista will probably also work but are no longer tested). The software runs on normal 32-bit installations, but 64-bit versions of Microsoft Windows 7 (or better) are strongly preferred.

The information extraction modules require a lot of memory and CPU power. The faster the CPU is the better. System memory should not be less than 4GB.

Mac OS X

We always test against the latest Apple desktop operating system (as of 01/2014 OS X Mavericks). The system should have a fast CPU and at least 4GB of RAM. Please check the Release Notes which restrictions for OS X may apply.

Linux

The software is tested to run on Ubuntu desktop 13.04. Both 32-bit and 64-bit versions are available. However, we stronlgy advise you to run 64-bit. Fast CPU required and minimal 4 GB of RAM.

What is the EMM OSINT Suite?

The EMM OSINT Suite is a desktop software package which provides you with Entity-centric search support. Based on EMM technology you can extract information from documents downloaded from the web or imported from local disk A good over of the functions can be found in the product flyer.

Glossary

Boolean Logic

Boolean logic is named after the mathematician George Boole. It is a form of algebra in which all values are reduced to TRUE or FALSE. We can use the Boolean operators to form search queries and to filter results of a search more effectively.

Boolean Operators

There are three different operators: AND, OR and NOT.

AND Operator

The and AND operator requires that all terms connected by the operator appear in the search results. For example, if someone searches for ***Barack AND Michelle***, only results will appear where both terms are present. Google uses an implicit AND opoerator, that means a search query ***Barack Michelle*** equals ***Barack AND Michelle*****.**

OR Operator

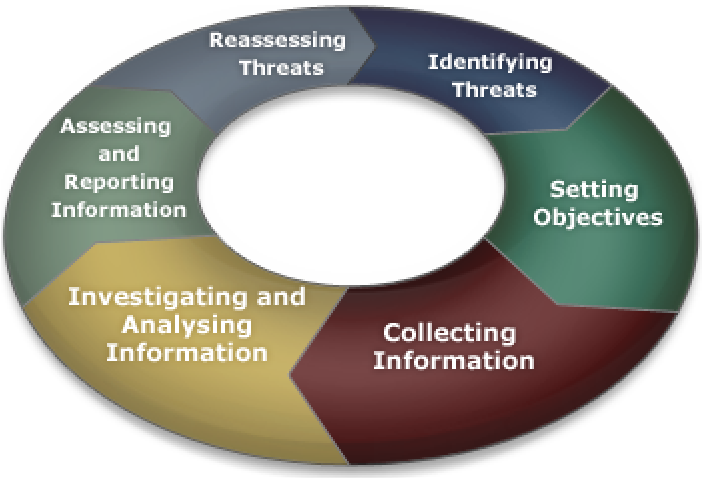
The OR operator is used to connect terms in a search query. The search engine results list found pages containing either of the two or more connected terms. For example, ***Barack OR Michelle*** will find all pages where either ***Barack*** is mentioned or ***Michelle*** is mentioned or both are mentioned. This example will yield a hugh number of results, since "Michelle" is a pretty common name.

NOT Operator

The NOT operator is used to exclude pages with certain terms from a search result. For example ***Barack AND NOT Michelle*** will yield all pages containing the name ***Barack*** but which do not contain ***Michelle***. Note: Google uses a dash "-" for the NOT operator and has an implicit AND. Therefore, ***Barack AND NOT Michelle*** will be written as ***Barack -Michelle*** in Google's query language.

Intelligence Cycle

Law enforcement authorities, security and intelligence services rely on some core processes to derive intelligence from input data. The classical intelligence cycle forms a first framework to detect the consecutive stages from finding and acquiring raw data to deriving intelligence in a determined way.



(Security Intelligence Cycle of the New Zealand Security Intelligence Service, http://www.security.govt.nz/our-work/our-methods/)

The Intelligency Cycle contains the following steps:

- Identifiying Threats - this is the inital step which starts the cycle
- Setting Objectives - in this step the goal is to specify which questions need to be answered to be able to assess the threat
- Collecting Information - this step covers the activity to harvest data from a wide variety of sources (OSINT - publically available sources)
- Investigating and Analysing Information - this step contains automatic and manual information extraction and processing, e.g. find the name of a specific person in a large collection of documents
- Assessing and Reporting Information - this step describes how the gathered information is put into reports.
- Reassessing Threats - with the gained knowledge, a threat is reassessed and depending on the outcome, the cycle may start again.

Even initially developed for intelligence services, the process framework can be equally used for classical law enforcement investigations.

Internet Protocols

At the heart of the Internet is a set of rules defining how computers can exchange messages. In computer science a set communication rules is commonly called a protocol. The messages are well-defined and each message has an exacte meaning to provoke a particular response of the receiver.

The Internet Protocol Family consists of various protocols which describe different aspects of network communication. Commonly, these protocols are put in a layered reference model to categorize the protocols. According to Andew S. Tanenbaum (see [1], chapter 1.4.3) the reference model for the Internet Protocol contains five layers:

Layer	Description	Example Protocol
Application	Contains programs that make use of the network,	HTTP (Web), SMTP (Mail), RTP, DNS
Transport	Provides protocols describing delivery abstractions, such as reliable transport of datagram.	TCP
Network	Deals with connecting different networks and the routing of datagrams within the networks	IP, ICMP
Link	How to send finite length messages between directly connected computers	Ethernet, WiFi
Physical	Describes how to transmit dfata as electrical signals	DSL

In the following we will describe only the most important protocols, for an in-depth description of all protocols, see [1].

Internet Protocol (IP)

The central protocol to make the Internet work is the Internet Protocol (IP). It is responsable to address hosts (computers attached to the Internet) and for routing data packets from a source computer to a destination computer.

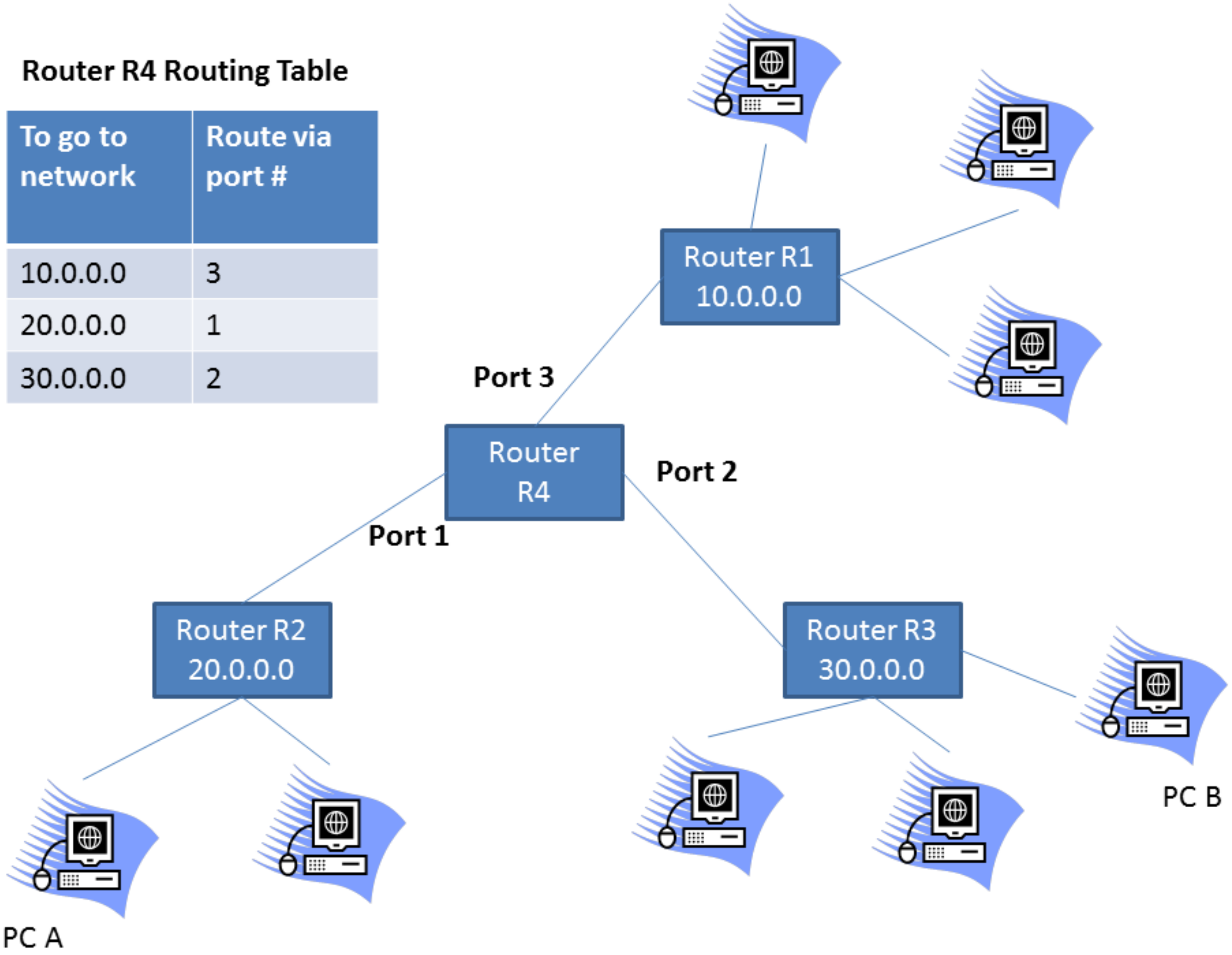
Every device attached to the Internet must have a unique address. In the current dominant protocol of the Internet, Internet Protocol Version 4 (IPv4), the addresses are made up of four sets of numbers separeted by periods (e.g. 192.168.2.1.). An IP address is similar to a postal address, it addresses a unique destination in the network. However, postal addresses are usually fix, whereas IP addresses may be assigned to a device only for the time the device is connected to the network. Network components, such as routers which are permanently connected to the network have usually permanently assigned address (so-called static IP addresses).

IP Datagrams

A datagram (also commonly called a data packet) is a fundamental unit of data that is sent between a source and a destination in the Internet. A datagram is made up of an IP header and the payload with the actual data (e.g. the content retrieved from a web site). The header contains the source IP address, the destination IP address and other meta-data needed to deliver the packet.

Routing Protocol

Routing is the process of finding the destination computer (host) for a packet which is being forwarded in the network. Since the Internet is a network of networks, there may exist multiple paths from a source IP address to a destination IP address (like for a postal package there are many roads leading from one city to another). The task of the router is to forward a data packet towards is destination. For this purpose the router uses routing tables to determine where a packet is going and how to send it.



Router R4 Routing Table

To go to network	Route via port #
10.0.0.0	3
20.0.0.0	1
30.0.0.0	2

In the above example, a PC A wants to send some data to PC B. The IP network software on PC A encapsulates all data in IP datagrams (IP packets). These packets are then send to the Router R2 which connects PC A to the Internet (in this example made up of three sub networks). The datagrams contain the source IP address of PC A inside the network with addresses starting with 20.0.0.0 and the target IP address of PC B which is 30.0.0.1:

Header	
Source	20.0.0.1
Destination	30.0.0.1
Payload	
Hello There!	

The router R2 which connects PC A to the Internet forwards all datagrams with an address other than 20.x.x.x to router R4. Now, router R4 needs to decider where to forward the incoming datagram. For this reason router R4 consults its routing table. The routing table defines that all datagrams with a destination address of 30.0.0.0 need to be forwarded via Port 2 to router R3. Router R4 forwards the datagram to router R3. Router R3 knows all connected PCs and forwards the datagram to the final destination 30.0.0.1 which is PC B.

Application Protocols

Application layer protocols describe how applications using the network can communicate with each other.

Hypertext Transfer Protocol (HTTP)

The Hypertext Transfer Protocol (HTTP) is the fundamental protocol enabling the World Wide Web. It describes how to transfer hypertext documents between a client and a server in a request-response manner.

A web server program is providing a web site made up of resources such as HTML documents and media files (images, movies, etc.) to clients on the WWW. A web browser is a typical client which requests resources from the server. For this reason the web browser sends a HTTP Get request to the web server. The web server responds either with the requested resource (for example a HTML document) or a status code describing an error condition. Resources on a server are identified using Uniform Resource Identifies (URIs) or, more spcifically for HTTP, using Uniform Resource Locators (URLs)

Simple Mail Transport Protocol (SMTP)

The Simple Mail Transport Protocol is the Internet standard protocol to transmit elecontric mail. Usually SMTP is used to exchange message between mail servers. Most email client programs use it only to send email, and use other protocols, such as the Post Office Protocol (POP) or the Internet Message Access Protocol (IMAP) to retrieve email messages from a mail server.

References and Further Reading

- [1] Andrew S. Tanenbaum, Computer Networks 5th Edition, Pearson 2011, ISBN 0-13-255317-1

- The Internet Protocol Suite, Wikipedia.org
- The Internet Protocol, Wikipedia.org
- Internet Engineering Task Force, IPv4
- Router Terminology, Answers.com
- The World Wide Web Consortium (W3C)

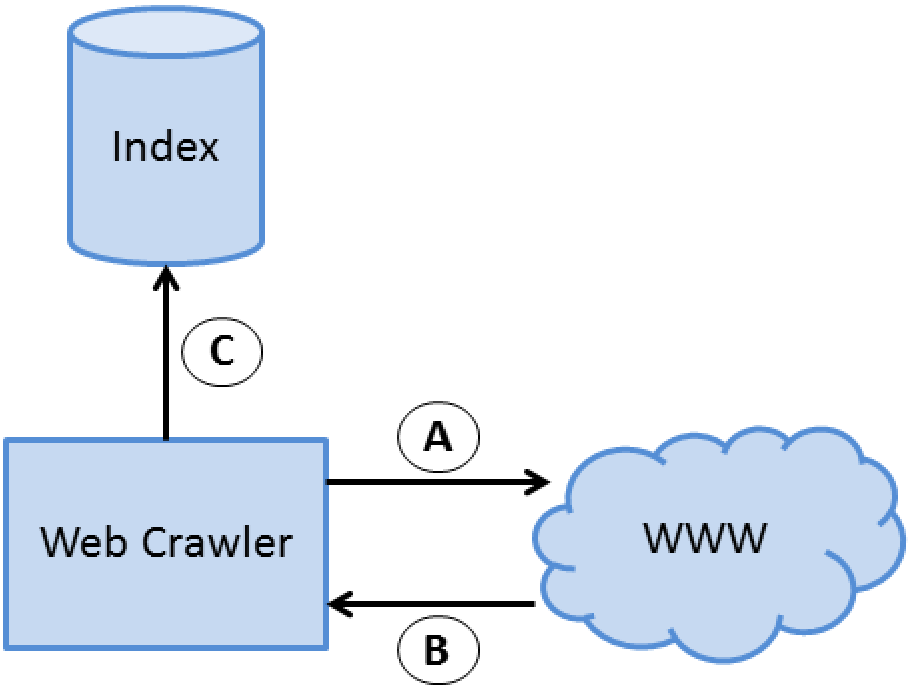
Internet Search Engine

Commercial grade web search engines, such as Google, Bing or others consists of thousands of servers in distributed data centres. However, the present themselves to the user as a simple web page. To understand how a search engine works, we look at two core processes. The first process is how the search engine continuosly collects crawls the web to build its search index. The search index is a data structure which allows to find web pages for specific keywords.

Crawling the Web

In order to build its search index, a search engines continuously downloads web pages and makes them searchable for keywords by storing them into an index. The links of web pages are then extracted and are downloaded next. In this incremental fashion a complete copy of all public web sites is downloaded into the index of the search engine. Given the scale of the World Wide Web today this can only be done by employing tens of thousands of servers in concurrent fashion.

The basic concept is shown in the next diagram:

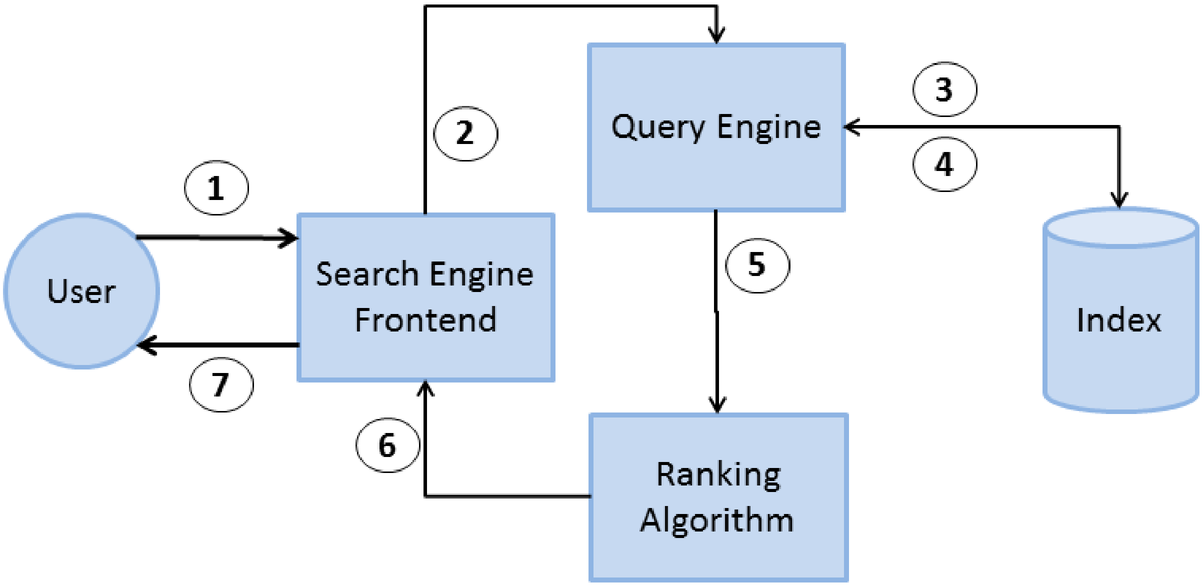


The Web Crawler components performs the following steps:

- A: It starts with a set web page addresses (URLs)
- B: the web pages behind the URLs are downloaded
- C: the crawler puts the web pages into its index and extracts URLs embedded in the web pages
- The process starts at step A again using the URLs extracted from the web pages

Querying a Search Engine

Now we look at the process how a user queries a search engine.



The search engine use the Index component which is created during the web crawl operation.

1. A search engine user types in a search query into the web form presented by the Search Engine Frontend (web server)
2. The Frontend component translates this query into internal retrieval commands. Furthermore, it takes into account additional knowledge about the user such as the user's search history or typical search queries used by a large number of users. The retrieval commands are forwarded to the Query Engine component.
3. The Query Engine is responsible to retrieve results for the query from the Index. The Query Engine and the index are massively distributed systems so many servers will be queried in parallel to produce results.
4. The Index component returns results to the Query Engine.
5. The results are forward to component providing a ranking algorithm. The results are ranked according to different criteria calculating a score for each search result. Typical static criteria include the number of occurrences of a search term in a page. These static criteria are combined with dynamic criteria such as user related information (search history, etc.).
6. The ranked results are forward to the Search Engine Frontend.
7. The Front End component renders the search result pages and returns it to the user. The result page is in most cases enriched with data from other systems such as advertising servers.

Open Source Intelligence (OSINT)

Open Source Intelligence (OSINT) is a term from the world of intelligence services. However, in more general terms it refers to the collection, processing and analysis of information from publicly available sources. The goal is to create actionable information supporting operations or decision making processes.

Typical public sources include (see also [1]):

- Mass media products, such as newspapers, magazines, radio, televsion
- Web-based information, such as homepages
- Web-based user generated content, such as blogs, social-networking sites, video sharing sites, wikis
- Government data, such as official reports, budget information, press information
- Professional and academic literature, e.g. conference proceedings, research papers, professional associations

For the purposes of law-enforcment the web based information becomes more and more prevalent to support investigations into fraud and criminal activities.

References

[1] Wikipedia Entry Open Source Intelligence. Retrieved September 10, 2013, from http://en.wikipedia.org/wiki/Open-source_intelligence

Concepts

- Entity Extraction
- Reporting
- Text Extraction

Category Matching

Category Matching is the process of classifying or grouping documents according to a field of interest (which we call a *category*). EMM OSINT Suite allows to classify documents on disk which have been previously downloaded or imported.

The software provides an editor to define categories. Each category definition consist of a combination of keywords. With the help of these keyword combinations the software can categorize a set of documents.

Entity Extraction

The goal of the Entity Extraction is to find locations in the text which contain *entity information*. In other words it tries to find occurrences of person names, locations, VAT numbers, etc..

The overall process is split into sub modules which run in a pipeline like fashion.
The system performs the following extraction steps:

Matching Module	Matched entities
Name Variant Matching	Matches name variants from the name variant database
Geo Matching	Matches geo locations (countries, regions, cities)
Regular Expression Matching	• Matches buit-in types such as vat number, email address, url, ip address, credit card numbers, date, phone number, zip code, personal id • Custom user-defined types based on regular expressions
Entity Guessing	Guesses further entities according to built-in rules
Entity Normalisation	Combines similar name variants to a single entity profile, provides unique ids to entities accross the document set

Name Variant Matching

The *Name Variant Matching* module simply matches entries from the *Name Variant Database* to the document texts. The found matches are then marked as entities with the type and id from the database.

Name Variant Database

The *Name Variant Database* contains entities of various types (e.g. Person, Organisation, etc.). It is amended each time the *Entity Normalisation* process finds a new entity.

Note: The initial *Name Variant Database* is created automatically from the EMM NewsBrief system. Therefore, the quality of the entries may vary.

Each possible spelling of an entity is called a *Name Variant* (or short “a variant”). Since a person entity can have many different spellings of its name, the variants are clustered in a so called *Name Variant Profile* (or short: “profile”). The name of a profile is taken from one of its variants. We call this variant the canonical variant.

For example, the profile for “Franz Beckenbauer” (a former German soccer player) contains a variety of variants which can also contain misspellings of his name.

- Franz Beckenbauer (canonical variant)
- Franz Beckenabuer
- Franz Beckenbaur
-
-

The profile is named “Franz Beckenbauer” after the canonical variant. Each profile has a unique id in the system. Therefore, all variants found belonging to the same profile will get the same profile id (and represent the same Entity in the system). The *Name Variant Database* is automatically amended by the entity normalisation module which tries to find variants belonging to the same profile. In addition, the profiles and variants in the database can be edited manually.

Geo Matching

The Geo Matching module matches the text against a database of location names (Countries, Regions, Cities, etc.).

Regular Expression Matching

The Regular Expression Matching module matches built-in entity types and user defined custom entity types. The matching is based on Regular Expressions.

Entity Guessing

Entity Normalisation

Regular Expressions

A Regular Expression (often abbreviated regex or regexp) is a sequence of characters which forms a search pattern used to match strings.

Each character in a regular expression is either a meta-character with a special meaning or a regular character with its literal meaning.

Example Regular Expression:

```
.*linkedin\.com.*
```

This consist of the following characters:

- meta-characters **.*** which means "zero or more occurences of any character"
- regular characters **linkedin**
- escaped meta-character **\\.** means the literal value of . (a dot) should be matched
- regular characters com
- meta-characters .* which means "zero or more occurrences of any character"

The example regular expression matches any URL of linkedin.com.

Resources

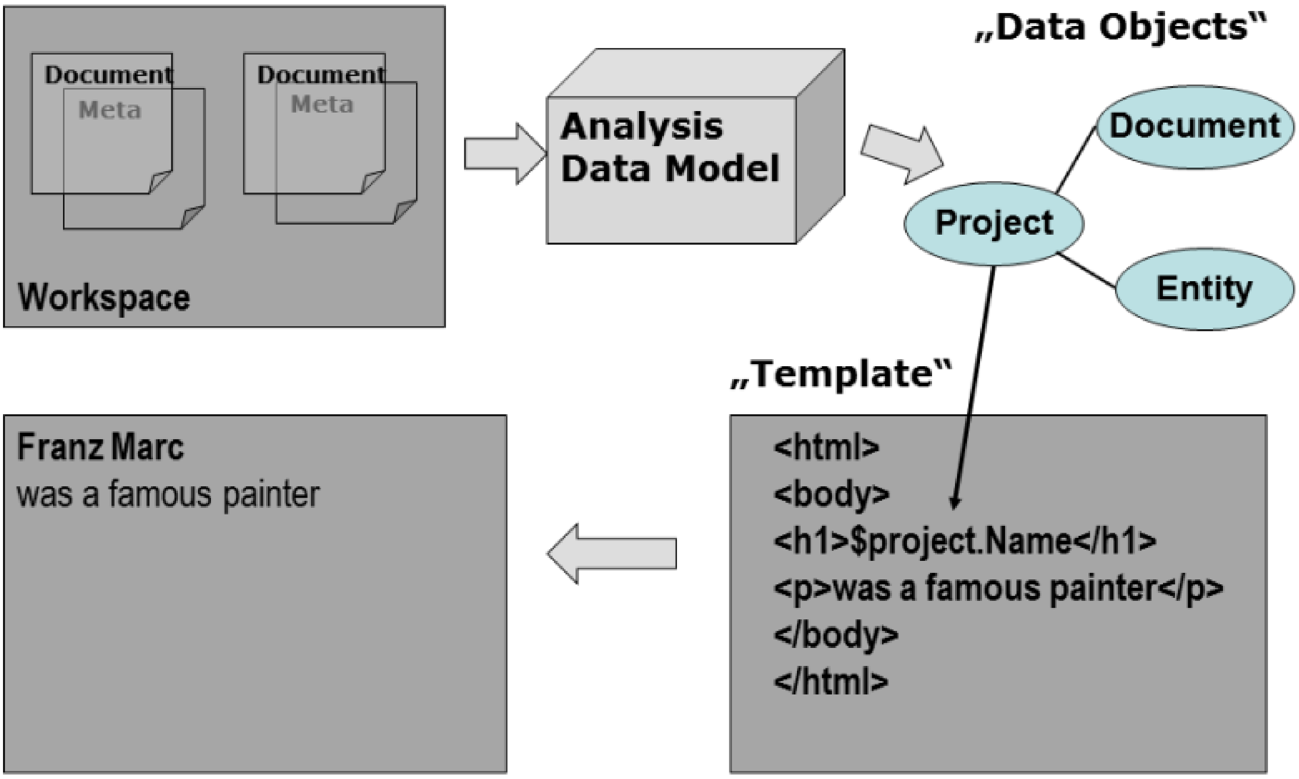
- Wikipedia: Regular Expression
- Online Testing of Regular Expressions regexr.com

Reporting

A report is a way to export the analysed data (entities and relationships) of a case project. Using different templates output file with different formats (Text, HTML, CSV, XML) can be created. The reporting mechanism is made up of three components:

1. Templates
2. Data Objects
3. Scripts

If you generate a report the following happens:



The entity data which is stored in the meta data of the individual documents is loaded into an aggregated Analysis Data Model object. This object makes the data available via an API represented by Data Objects. A template file can embed the data objects using variable names (e.g. $project.name). If the report is generated the variables are replaced by the actual data (e.g. the project name "Franz Marc" replaces $project.name in the template).

The templates used to create a report are stored internally and can be modified by creating a configuration project.

Data Objects

Project Object

The project object is the main entry point to access the data (documents and entities and relationships) of a project.

Object Property	Description	see also
$project.Name	name of the project	
$project.Documents	retrieves a list of all documents of the project	Document Object
$project.Entities	retrieves a list of all found entities of the project	Entity Object
$project.getDocumentById(id)	retrieves a Document object by its unique id	Document Object
$project.getEntityById(id)	retrieves a single Entity object by its unique id	Entity Object
$project.getEntityByName(name)	retrieves a single Entity object by its name	Entity Object
$project.getEntitiesByNamePattern(namePattern)	retrieves a list of Entity objects matching a regular expression pattern	Entity Object
$project.getEntitiesByEntityType(typeId)	retrieves a list of Entity objects matching the entity type id (2 letter code)	Entity Object
$project.DocumentEntityRelations	retrieves a list of DocumentEntityRelation objects	DocumentEntityRelation Object
$project.EntityEntityRelations	retrieves a list of EntityEntityRelation objects	EntityEntityRelation Object
$project.Bookmarks	retrieves a list of all Bookmark objects in the project	Bookmark Object

Note: The algorithm to calculate the EntityEntityRelations list is currently quite slow and takes for large data sets a considerably amount of time. This is a known issue and will be improved in a future version.

Entity Object

The entity object represents a single entity found in a document of the project.

Object Property	Description	see also
$entity.Id	the unique Id of the entity	
$entity.Name	the name of the entity	
$entity.RelatedDocuments	a list of Document objects in which the entity occurs	Document Object
$entity.RelatedEntities	a list of Entity objects which occur in the same document	Entity Object
$entity.Type	the type acronym of the entity (for example “p” for Person)	
$entity.TypeName	the type name (for example “Person”)	

Document Object

The document object represents a document of the selected project.

Object Property	Description	see also
$document.Id	the unique id of the document	
$document.Title	the title of the document	
$document.Link	the source link of the document if it was retrieved from the web	
$document.RelatedEntities	a list of Entity objects which occur in the document	Entity Object
$document.FilePath	the file path to the original document on local disk	
$document.Text	the extracted text of the document	
$document.DocumentEntityRelations	a list of DocumentEntityRelations involving this document	DocumentEntityRelation Object

DocumentEntityRelation Object

The documentEntityRelation object represents the relation between a document and an entity.

Object Property	Description	see also
$documentEntityRelation.Document	returns the document of this relation	Document Object
$documentEntityRelation.Entity	returns the entity of this relation	Entity Object
$documentEntityRelation.TextPositions	returns a list of TextPosition objects describing where the entity was detected in the document	TextPosition Object

TextPosition Object

The textPosition object represents a position in a document where an entity was detected (see **$documentEntityRelation.TextPositions**).

Object Property	Description	see also
$textPosition.Pos	the character position from the beginning of the document	
$textPosition.Length	the length of the found entity	

Text Extraction

Text Extraction is the process of extracting raw text from multiple input file formats.The *Text Extraction* module of EMM OSINT Suite is based on the open source project Apache Tika.

Currently, the module supports the following input file formats:

- Plain text (no text extraction needed)
- XML
- HTML
- PDF
- Microsoft Office 97 formats (doc, xls, ppt)
- Microsoft Office Open XML (2007) (docx, xlsx, pptx, thmx)
- Open Office Text, Presentation, Spreadsheet(odt, odp, ods)

In addition to extract the text, the language of the text is identified and stored as meta data.

Training

Welcome to the EMM-OSINT Suite Training Sessions!

The EMM Open Source Intelligence Suite is a desktop software application which helps to find, acquire and analyse data from the Internet and local sources. It provides automatic means to gather intelligence from open available sources by removing the need to search manually through vast data sets.EMM OSINT Suite comprises a set of powerful tools to support the main processes of intelligence gathering from open sources. Documents can be acquired from the public internet as well as from local sources. The core of the software is the entity extraction module which matches text locations against pre-defined patterns for different type of entities, such as person, organisation and place names, credit card numbers, VAT identifiers, URLs, etc.. User defined patterns can be added to find investigation specific entity types, such as number plates or tax identifiers. The analysis views allow to make sense of the data.

Here you will find material for the training sessions introducing different modules of the software:

1. Installing EMM-OSINT Suite

The EMM-OSINT Suite can be installed on different platforms:

- Microsoft Windows
- Linux (Ubuntu)
- MacOSX

Linux (Ubuntu)

Installing EMM-OSINT Suite on Linux (Ubuntu)

Installing on Linux.

Microsoft Windows

Installing EMM-OSINT Suite on Microsoft Windows

Installing on Windows.

MacOSX

Installing EMM-OSINT Suite on MacOSX

Installing on Mac OS X using VirtualBox .

2. Getting Started

Quick Start Guide.

3. Module Sessions

- C1 - Data Acquisition
  - C1 - Lab Exercises
- C2 - Entity Extraction
  - C2 - Lab Exercises
- C3 - Entity Extraction Advanced
  - C3 - Lab Exercises
- C4 - Reporting & Data Export
  - C4 - Lab Exercises
- C5 - Category Matching

C1 - Data Acquisition

EMM-OSINT Suite provides a browser based search interface to the internet search engines. Search results can be downloaded for further local processing. In addition to using search engines, targeted websites can be crawled using the embedded web crawler. The crawler follows the link structure of a website and downloads relevant pages to local disk. A file import wizard complements the acquisition tools. It allows importing locally stored documents for further analysis. For further processing the plain text is extracted from a variety of document formats, such as HTML, PDF and Microsoft Office.

The Data Acquisition session is composed of the following sections:

- C1.1 Searching the Web
    - Performing a Search
    - Managing Bookmarks
- C1.2 Crawling a targeted Web Site
- C1.3 Importing Files

Before starting, **a Case Project must be created** as prerequisite (see creating a Case Project).

C1.1 Searching the Web

Performing a Search

The search module allows to gather result links as bookmarks from internet search engines

- Performing an Internet Search

In order to **customize the web search**, you can:

- Define the maximum number of search result links to be extracted

Managing Bookmarks

- Filtering out search links (bookmarks) before extracting them
- Using the Duplicate Bookmark Detection

C1.2 Crawling a targeted Web Site

The crawler allows to collect data from specific web sites by following the site structure

- Crawling a Web Site

C1.3 Importing Files

Data can be imported from local disk into a Case Project. One way is to import documents (plain text files, PDF files, Microsoft Office files) for later analysis. Another way is to import bookmarks from web browsers:

- Importing documents from local disk
- Importing bookmarks from web browsers

C1 - Lab Exercises

C1.1 Searching the Web with OSINT

Search for "Al Qaeda":

- Search on Google
- Search on Bing

- Extract 50 result links each and store as bookmark files
- Remove duplicate bookmarks
- Review and remove invalid bookmarks
- Download web pages of bookmarks to Documents Folder

Also refer to:

- Creating effective search queries

C1.2 Crawl a Targeted Web Site

- Crawl a news web site depth 1
- Review and remove invalid bookmarks
- Download crawled pages to Documents Folder

C1.3 (Advanced) Creating a custom a Search Provider

- Creating a configuration project

- Create a custom search provider for
  - https://ixquick.com

C2 - Entity Extraction

The Entity Extraction finds locations in the text which contain "entity information", such as person names, locations, organizations, VAT numbers, etc.

The entity extraction searches a set of documents located in the ***Documents*** folder of a Case Project. The ***Documents*** folder is a *special predefined* folder which contains all input documents for the entity extraction.

The Entity Extraction session is composed of the following sections:

- C2.1 Performing the Entity Extraction
  - Using the Entity Browser
  - Using the Graph view
- C2.2 Editing the Name Variant Database
- C2.3 Adding Custom Entity Types

C2.1 Performing the Entity Extraction

- Performing the Entity Extraction

Using the Entity Browser

Once the Entity Extraction has finished, you can review the entities which were found.

- Using the Entity Browser.

Using the Graph view

The system provides a basic view to show relationships in a graphical way.

- Using the Graph View

C2.2 Editing the Name Variant Database

The Name Variant Database is used to match variants of an entity. (Such as Barack Obama, President Obama, etc.)

- Editing the current Name Variant Database
- Importing a Name Variant File

C2.3 Adding Custom Entity Types

- Introduction to Regular Expressions (ppt)
- (External) Regular Expression Tutorial http://regexone.com/
- (External) BRICS Syntax http://www.brics.dk/automaton/doc/index.html

See Adding a Custom Entity Type for more information.

C2 - Lab Exercises

C2.1 Performing the Entity Extraction

Run the entity extraction on a set of documents.

- Creating a Case Project
- Searching the web for a person or organisation of your choice
  - Use at least two search engines
- Consolidate bookmarks
  - Review extracted search bookmarks
  - Remove duplicate bookmarks
- Download web pages
- Run the entity extraction
- Review the results

C2.2 Editing the Name Variant Database

Open the Entity Browser and look for entities with similar names

Use the editor to consolidate two entity variants into a single profile

C2.3 Creating a Custom Entity Type based on a fixed pattern

Create a regular expression pattern to extract container numbers (BIC code) from a set of documents.

You can use http://www.regexr.com/ to develop a pattern according to the wikipedia definition of container numbers.

A first working pattern for container numbers:

```
[A-Z]{3}[UJZR][0-9]{7}
```

` Now you have to create a custom entity pattern file in <config-project>/Custom Entities/Active Entities named bic.xml`

with the following content:

bic.xml

```
<?xml version="1.0" ?>
<expressions>
<!--
OPTIONAL: the declaration tag can be empty
This declaration section defines additional scripts used for
validation. The scripts are written in Groovy, a scripting language
which is a superset of Java. See http://groovy.codehaus.org/ for more information.
-->
<declaration><![CDATA[

 /**
 * This predefined init method is called only once, during initialisation of the
 * custom entity module. It should be used to load resources for validation.
 */
  public void init() {
   //this function can be used to load resources from
   //the "Custom Entities/Resources" directory in the active configuration project.
   //The path of this directory can be accessed from the predefined resourcespath
variable.
   //The loaded resources should be stored to the context variable in order to be
accessible
   //from other scripts which are called for each text match.
  }

  /**
  * This is an example of a validation function, which can be accessed as
  * global.validate(term) to further validate a matched pattern.
  */
  public boolean validate(String term) {
   return true;
  }

 ]]>
</declaration>
<!--
MANDATORY: the type tag must define a two letter id and a description
of the custom pattern.
-->
<type id="bc" description="container bic"/>
<!--
MANDATORY: At least one expression definition must exist.
```

```
-->
 <expression>
  <!--
  MANDATORY: regex contains the regular expression to match the text.
  By default it uses the restricted "Brics syntax" for performance reasons.
  See http://www.brics.dk/automaton/doc/index.html?dk/brics/automaton/RegExp.html
for
  more information.

  Alternatively, it can have the attribute mode="groups" to fall back on Java
regular
  expression syntax which is much slower but allows to match groups.
  If the mode attribute is set to "groups", the matched groups can be accessed
  from the variable groups (array, 1-based, 0 contains the full match).
  -->
    <regex><![CDATA[[A-Z]{3}[UJZR][0-9]{7}]]></regex>
    <!--
    MANDATORY: a unique name of this regex pattern. A custom type can be matched
    using multiple patterns, but each pattern must have a unique name. (Otherwise
    we cannot show error messages which pattern has failed, etc..)
    -->
    <name>container-bic</name>
    <description>a custom pattern</description>
    <!--
    OPTIONAL: validate section can be empty or simply contain "return true;"
    -->
    <validate><![CDATA[
     //Example: calls the optional validate function to check
     //if the matched term (stored in the term variable) is actually validate
     return global.validate(term);
     ]]>
 </validate>
 <!--
 MANDATORY output: The type must be returned under the "type" key as defined in the
id attribute of the type tag
 -->
    <output key="type" value="bc"/>
    <!--
    MANDATORY output: The matched pattern must be returned under the "name" key
    -->
    <output key="name">
     return term;
    </output>
  </expression>
```

```
</expressions>
```

C2.4 (Advanced) Creating a Custom Entity expression with multiple patterns

Create a custom entity type with patterns which match the ISIN number of at least three European countries

File	Modified
companies-cac40_utf8.txt	Sep 17, 2013 by Gerhard Wagner

Drag and drop to upload or browse for files

C3 - Entity Extraction Advanced

In this session, some advanced tasks related to the Entity Extraction process are described:

- C3.1 Customizing the Name Variant Database
- C3.2 Customizing Regular Expressions for Entities

C3.1 Customizing the Name Variant Database

See Name Variant Database for understanding the main concept.

See Editing the current Name Variant Database in order to modify or add new entities to the current Name Variant Database in OSINT.

C3.2 Customizing Regular Expressions for Entities

See Adding a Custom Entity Type.

C3 - Lab Exercises

C3.1 Import a Custom Name Variant Database

You want to search for a specific person.

- Create a Name Variant file with an entity profile and a number of variants for the person name
- Import the Name Variant file into the Name Variant Database

Search for documents, review found entities

C3.2 Export the Name Variant Database and Import an improved one

- Export the name variant database (from C3.1)
- Review the exported file and consolidate multiple profiles into one
- Import the improved file into the Name Variant Database

C3.3 (Advanced) Creating a Custom Entity Type to match a list of company names

Create a custom entity type which loads a list of person names from a text file.

C4 - Reporting & Data Export

C4 Reporting & Data Export

It is possible to export data from the application. Currently, the following types of data can be exported:

- Bookmarks (i.e. URLs scraped from search engines)
- Documents (downloaded from the web)

This session covers the following topics:

C4.1 Exporting Bookmarks

Bookmarks are files in the workspace containing a web URL and some meta data. The meta data contains the following data:

Key	Description
URL	web address the bookmark points to
Search Engine	search engine the URL was extracted from (e.g. Google)
Search Query	search query used to find the URL
Title	title as shown by the results page of the search engine
Timestamp	date and time the bookmark was created

The system allows to export bookmarks to three target formats:

Target format	Description
Netscape bookmarks file	An old html based file format to export a hierarchy of bookmarks. This format can be read by Firefox, Microsoft Internet Explorer and Google Chrome. No meta data is exported, only the URL and the title.
Plain text	A plain text file containing a list of all URLs
Tab separated value file	A plain text file containing all bookmarks and meta data in columns with tabulator

C4.2 Exporting Documents

Documents can be exported to a proprietary XML format. Either all documents are exported into multiple XML documents or a single large XML document can be created.

The format of the export is as follows:

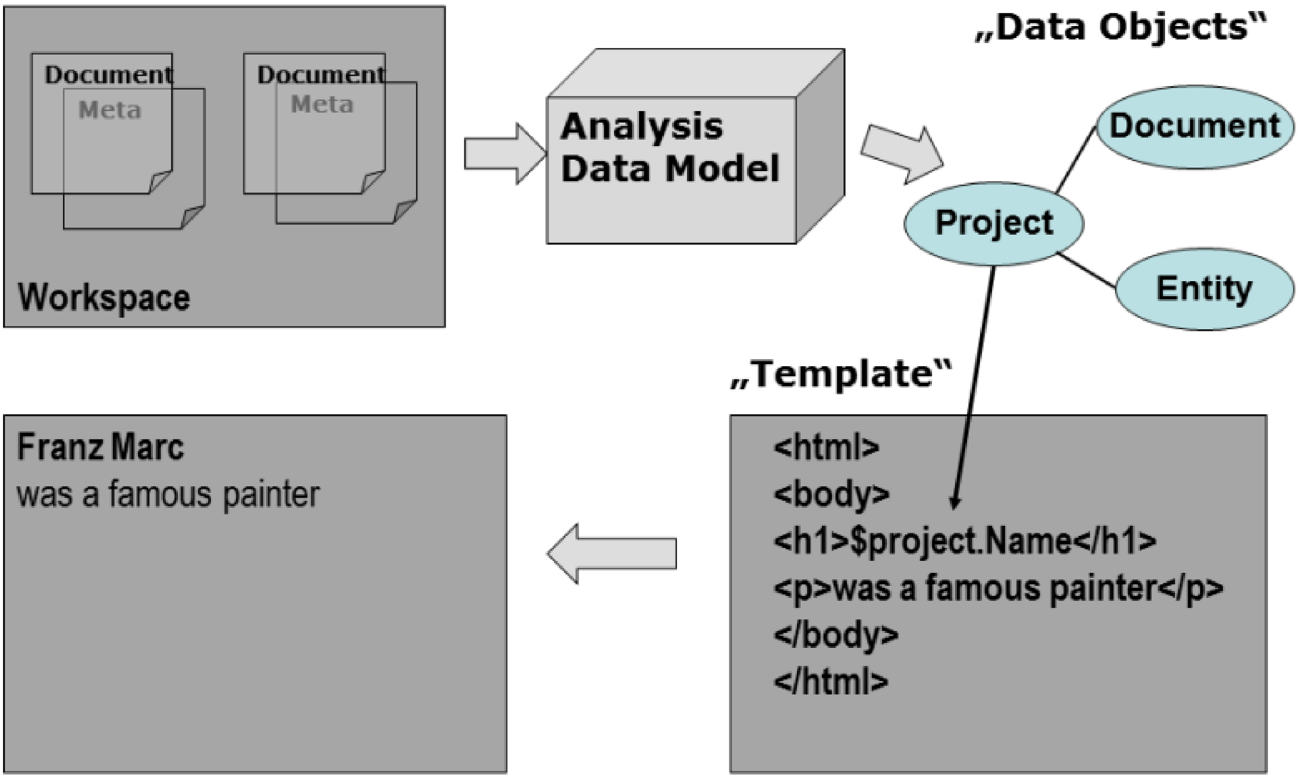
Tag	Description
doc	encloses a complete document
title	title of the document
text	extracted plain text of the document
desc	a short description of the content
url	the URL of the original document
lang	the detected language

C4.4 Reporting

A report is a way to export the analysed data (entities and relationships) of a case project. Using different templates output file with different formats (Text, HTML, CSV, XML) can be created. The reporting mechanism is made up of three components:

1. Templates
2. Data Objects
3. Scripts

If you generate a report the following happens:



The entity data which is stored in the meta data of the individual documents is loaded into an aggregated Analysis Data Model object. This object makes the data available via an API represented by Data Objects. A template file can embed the data objects using variable names (e.g. $project.name). If the report is generated the variables are replaced by the actual data (e.g. the project name "Franz Marc" replaces $project.name in the template).

The templates used to create a report are stored internally and can be modified by creating a configuration project.

C4.4 Use existing reports

The system ships with a set of predefined reports. Which are accessible from the Reports view. Please refer to Generating a Report.

C4.5 Creating a custom report

Please refer to the tutorial Creating a Custom Report.

C4 - Lab Exercises

C4.1 Exporting and Importing Bookmarks into Firefox / IE / Chrome

Export bookmarks from your project and import the Netscape Bookmark file into the browser of your choice

C4.2 Exporting Bookmarks and editing them with MS Excel

Export bookmarks to a CSV file

Open the CSV file in MS Excel, edit it.

Import the amended CSV file into your case project.

C4.3 Use a predefined template to generate a report

Select an existing template to generate a report

C4.4 Improve a predefined template to generate a nicer, more complete report

Create a configuration project

Copy a reporting template

Edit and improve the copy

Use the new template to generate a report

C4.4 (Advanced) Create a custom CSV report

Create a report template to generate a CSV list of entities (with all attributes)

Import the CSV file into MS Excel

Note: Look in the documentation section

C5 - Category Matching

A set of documents can be categorised by using the Category Matching module. Each category is defined by a set of keywords (or combinations of keywords). All documents in the system are matched against the active category definitions.

- DSL Overview and Introduction (ppt)
- Creating a Configuration Project
- Creating a Category Definition File
- Defining a Category
- Matching Documents against Categories