MetadataMiner v1.0

User Guide

Ian P. Duffy

[ipduffy@yahoo.com](mailto:ipduffy@yahoo.com)

1 Overview 2

2 Getting Started 3

3 Searching, Exporting, and Reporting 9

3.1 Searching 9

3.2 Exporting 11

3.3 Reporting 12

# Overview

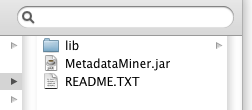
MetadataMiner is a Java-based application that allows a forensic analyst to quickly parse a directory containing Microsoft Office documents (including Office XML documents created by Microsoft Office 2007) to parse and extract document metadata that is saved into these documents by the Microsoft Office applications. MetadataMiner parses and extracts this data and saves it into a SQL database for easy searching, facilitating E-Discovery based on the document metadata. This metadata can provide valuable insight into documents that have been authored, printed or edited by a particular individual, identify what versions of an Office application were used to create a document, and can be used to create and augment a timeline of document creation, editing, and printing, which, coupled with analysis of document author/editor metadata can provide valuable insight into the activities of an individual at any given time.

My inspiration for writing this application was a forensic case that involved a large number of office documents on a suspect’s hard drive. I wanted a quick way to eliminate duplicate documents and identify all of the documents that were created or edited by the suspect and then create a timeline based on those documents so that I could see what the suspect was doing at any given time. There were no good tools to do this available commercially at the time. Thus I began in earnest writing this program. I hope that you find it useful.

# Getting Started

The first thing you will need to do to get started with MetadataMiner is install a Java Runtime Environment, if you don’t already have one. MetadataMiner was developed on a Mac OS X system, so if you are using Mac OS X you should be good to go with the native JRE included with Mac OS X. MetadataMiner has also been tested with Sun/Oracle Java Runtime Environment (version 6, update 26) available at <http://www.oracle.com/technetwork/java/index.html>. If you prefer to use another JRE, your results may vary; if you run into any issues with them, please hit me up at the email above and I will do my best to try to rectify the issue.

Once you’ve got your JRE installed and working, you can extract the MetadataMiner zip file to a local working directory. You should have a file named MetadataMiner.jar and a directory named “lib” that contains a bunch of other libraries that MetadataMiner uses for various functions. In Windows and Mac OS X, you can run the program simply by double-clicking on the MetadataMiner.jar file.



In Linux, you will have to open a command shell and go to the directory where you extracted the MetadataMiner files. Type the following command to execute MetadataMiner:

java –jar MetadataMiner.jar

(NOTE: this command-line will also work at the command prompt in Windows and in Terminal on Mac OS X.)

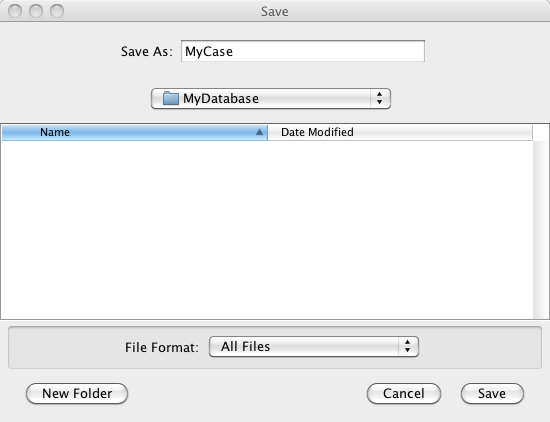
Once you have MetadataMiner up and running, you should see the main window of the application:



There are only a few things you can do at this point, as shown by the buttons:

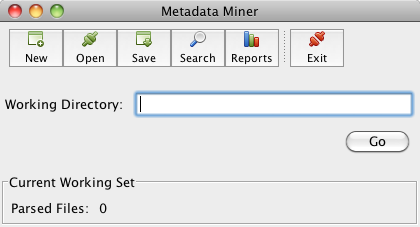
1. Create a New MetadataMiner database
2. Open an existing MetadataMiner database
3. Exit the application

The other buttons are disabled until you have either created or opened a MetadataMiner database. To get started, assuming you have not yet created a MetadataMiner database, click on the “New” button, and in the file save dialog that appears, navigate to a folder where you would like to store the MetadataMiner database files and type a name for your MetadataMiner database files:

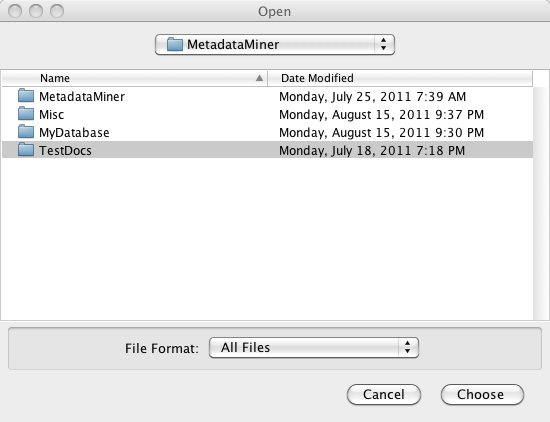


In the folder that you selected, MetadataMiner will create several files. These are related to the backend database (HSQLDB) and you don’t have to worry about them. If you would like to read more about HSQLDB, you can find more information about it at [www.hsqldb.org](http://www.hsqldb.org). You can certainly write your own utilities and scripts to interact with the database on your own.

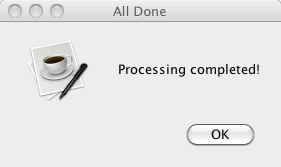
Once the working database is created, MetadataMiner will create all the tables that it needs and you should be set up to begin parsing and analyzing documents. Your other buttons will become active and MetadataMiner will tell you that there are currently no documents in your database:



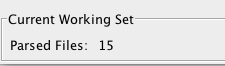
At this point you are ready to start parsing Microsoft Office documents. To do so, select the “Working Directory” text box. A file selection dialog should open, prompting you to select the directory where you have saved the Microsoft Office documents that you wish to analyze. NOTE: MetadataMiner will only allow you to select a directory, not individual files. It is designed to recursively parse an entire directory, reading in the metadata from each file in that directory and any subdirectories that it discovers. Select the folder/directory where you have saved/exported all the documents from your case:



When you have selected the working directory, all you have to do is hit the “Go” button, and MetadataMiner will begin searching for Microsoft Office documents in the directory you have specified, and any subdirectories therein. Depending on how many documents you have exported from your case, this may take some time. Once MetadataMiner is done parsing the documents in the Working Directory, it will notify you that it is done processing:



MetadataMiner will show the number of unique documents that it processed into the database. This number does not include duplicate files.



Note about duplicate files: MetadataMiner will calculate and MD5 hash of each file that it analyzes. It stores this MD5 hash in the database. If it encounters a file that it has already parsed into the database (based on this MD5 hash), it will not re-parse the file. Rather, it will make a note of the fact that it found a duplicate file, recording the file name and path into the database.

Once you have parsed in the documents that you are interested in analyzing, you can begin executing searches, generating reports, and exporting the data to Microsoft Excel for timeline analysis.

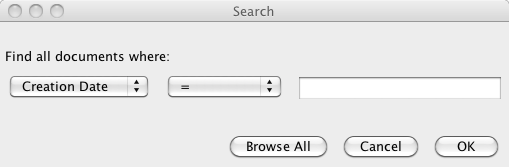
# Searching, Exporting, and Reporting

The true power of MetadataMiner becomes apparent when you use it to start analyzing the data that it has extracted from the documents. You can begin running searches on the data based on any of the different fields that MetadataMiner extracts from the Microsoft Office documents. These fields include the following:

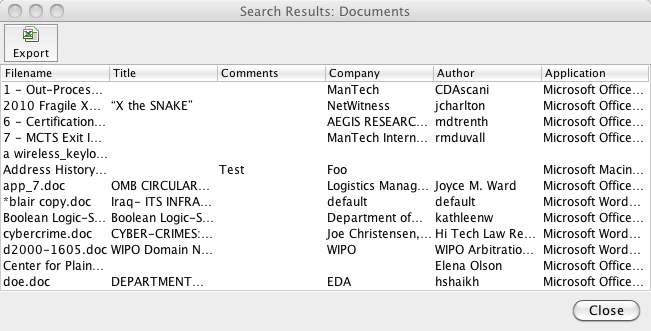
|  |  |  |  |
| --- | --- | --- | --- |
| Author | Filename | File Editing Time | Last Print Date |
| Editor | File Extension | Application\* | Word count |
| Title | MD5 hash | Creation Date | Character count |
| Comments | File revision | Last Save Date |  |

## Searching

The search dialog allows you to specify a particular metadata field and allows you to tailor your query to zero in on the exact subset of documents that match your parameters:



Alternatively, you can click on the “Browse All” button and you can view a list of all of the documents that MetadataMiner has parsed into the database. When MetadataMiner is done retrieving the documents that you are interested in, it will display basic information about those documents in a table:

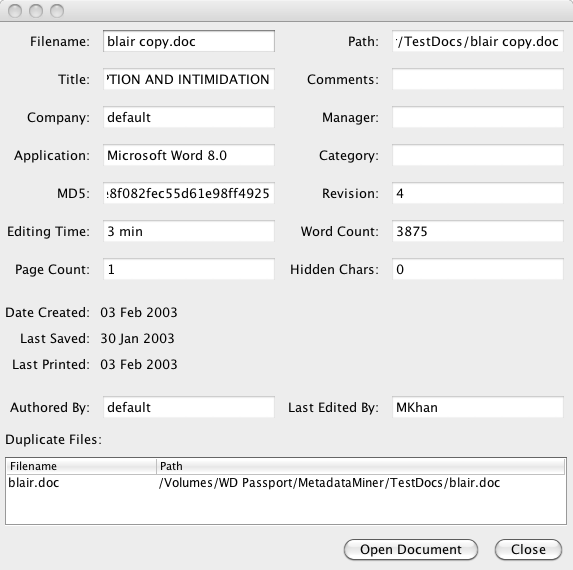


In order to view more detailed information about a particular document and its metadata, you can double-click on any row in the table and MetadataMiner will display a dialog box with all of the parsed metadata for that particular document. Alternatively, if you would like to open/view a document, you can right-click on a row and select “Open Document” in order to view this document in your native application that you use to view Office documents.

NOTE: It is potentially dangerous to open Microsoft Office documents from a suspect’s computer. I have encountered several documents that contain malware. Open at your own risk!

NOTE 2: A document whose filename begins with an asterisk indicates that duplicate copies of that document were encountered when MetadataMiner parsed in the documents in the working directory. You can view more information about duplicate files, to include filenames and paths for the duplicate files, by double-clicking on the row to display the detailed information about that document.

When you double-click on a row, detailed information about that particular document (and any duplicate copies of that document that were encountered) is displayed:

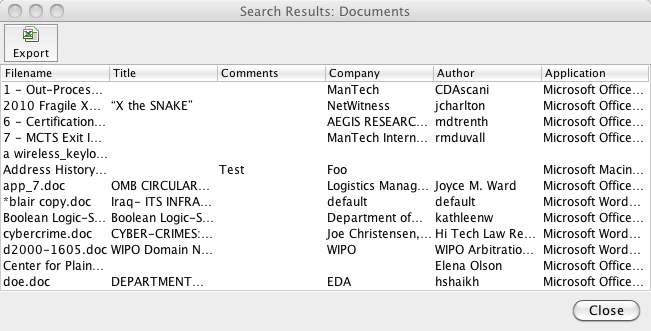


NOTE: Any fields that are partially obscured can be viewed in their entirety by hovering the mouse over the field. A tooltip will display the full contents of the text box.

From this point, a user can choose to open the document to view its contents or close the dialog box, which will return you to the list of documents.

## Exporting

From the list of documents, you also have the option to export a Microsoft Excel spreadsheet containing all of the metadata for the documents in the window (This could be all documents in the database if you chose to browse all documents, or only a subset of the documents if you searched for a particular value in one of the metadata fields). To export the list of documents to Microsoft Excel, click on the “Export” button on the top of the document list:



Select the location to save your Excel spreadsheet. Click Save. You’re done! You can now search, sort, and browse all of the metadata for all of the documents you’ve selected using the powerful tools included in Microsoft Excel.

## Reporting

MetadataMiner includes some canned reports that can be used to quickly identify interesting leads. These include the following:

1. Top Authors (authors who have created the greatest number of documents)
2. Top Editors (authors who have edited the greatest number of documents)
3. All Documents by Author (for each author, a list of all the documents they’ve written)
4. All Documents by Application (a list of each document that was created by a particular application / version – i.e. Microsoft Word 8.0)
5. Most Heavily Edited Documents by Editing Time (a list of documents, sorted from highest to lowest by the amount of time that they have been edited)
6. Most Heavily Edited Documents by Revision Number (a list of documents, sorted from highest to lowest by the revision number)
7. Recently Printed Documents (a list of documents sorted chronologically by most recently printed)
8. Recently Edited Documents (a list of documents, sorted chronologically by most recently saved)

These reports will be displayed in a separate window, where the report can be saved in various formats (PDF, DOCX, HTML, XML, CSV, and various others):

