



Exactus White Paper

The story of Exactus

One of the interesting areas of writing a report on electronic discovery processing software is it provides additional avenues into the weaknesses of the programs as a whole. This is something that is not possible via testing one of the programs. In 2006 we came out with a compare and contrast report of third party EDD processing tools on the market. If there was one specific area that caught my attention it was the accuracy of these programs at the loading stage. The same data set and methodology was used on all software only programs. The same held true for the software and hardware solutions. Why was it that each program had a different number for the loading of files? On some occasions a rerunning of the data took place and still the same results. Why? To answer this let us start at the beginning. There were three data sets that were run thru. The first was a set of loose files around 500 megabytes. Of the three data sets this is where there was the most consistency. Yet most of the programs had slightly different document counts. Why was one program able to load a few more files than another program? The answer dealt with embedded items. A few of the programs were able to extract embedded items thus having a higher document count than those that did not. Each program with the exception of one found the correct number of duplicate files. As it would turn out this would be the simplest data set that needed to be analyzed.

The next test was on four .PST files that were run together equalling around 1.5 gigs. Every program tested including the ones that did not make it in the report had a totally different document count for the loading of data. Little did I know at the time that this would begin a journey into the E-Mail accuracy issue. Further analysis was done on why a few programs were able to load hundreds of more documents than others. After peeling layer of layer off this difficult riddle a few answers came to light. Trying to complete an apples to apples comparison was nearly impossible because of how each program loaded documents and the methodology used for duplication. Not only were there different document counts at the loading stages but different counts for de-duplication. A few companies told me the reason why it could not load every document was it had to be a Microsoft issue and that Outlook was volatile. This simply was not good enough of an answer. After extensive research this turned out to be indeed a falsehood. The bottom line was a few of the products just could not load every single e-mail and attachment. One program that did not make the list loaded more documents than what was in the

.PST data set. I am still at a loss to explain that. Months later while running new data sets on a few programs more answers started to be realized. Not every program can go unlimited in e-mail nesting. For example say there is an e-mail with embedded e-mails of 5. In those 5 e-mails is another embedded e-mail with two embedded attachments. How deep can a program go? Can it reach those embedded attachments? Some programs can only go two to three levels deep. Another problem I saw was that I came across a data set that had a container file within a container file. The EDD programs I tested could not read the contents of the second container file. Nor could they read any attachments in a .PDF file.

The final test was a rather small set of four .NSF files. The results were all over the place. Two of the files did not contain any e-mails or attachments. Some of the programs could not even process these. Two programs could only process one .NSF file at a time and only if it had e-mails. One program that did not make the list crashed every time it tried to extract a large attachment from an e-mail. There was only one program that could process all files and have close to an accurate count. At the time of writing the report it was obvious that a majority of the programs just did not understand the ins and outs of Lotus Notes. To this day the same problem arises with migration tools converting from .NSF to .PST. Simply put they are not accurate. With new technology advances hopefully these programs and others in the future that come on the market will have a firm understanding of Lotus Notes.

To download a free copy of this outdated report that was released in December 2006. Please visit our website. <http://www.randallconsulting.net>

A good amount of time was spent in 2007 looking for a tool that can read everything in a .PST file. Was there a tool on the market to help with the e-mail accuracy issue? We tested every program on the market in relation to E-mail reading and reporting. None were able to give me what I wanted. A program that can tell me with close to 100% accuracy what was in a .PST file. These programs at best offer a quick peek with some nice reporting features. One of the better programs by a reputable computer forensics software company could not read all embedded items and was unable to read any attachments in a .PDF file. Randall Consulting created small data sets to run with these programs but again the majority of them only provided a quick peek. It was unable to read all attachments and filetypes.

In conversations with others in this industry on the E-Mail accuracy issue the responses are like the .NSF tests. They were all over the place. Some say that if a program can extract at least half of the embedded items in an e-mail file than it is a solid program. While others have said most firms understand that not every e-mail and attachment will be extracted. To this author that is completely unacceptable. A program needs to be close to one hundred percent accurate. What if one of the e-mails or attachments or embedded items was extremely relevant to a pending litigation matter? What then? Are we now talking about litigation against the company who processed that data?

In October of 2007 Randall Consulting again did even more testing on the E-Mail reporting tools on the market. These programs had the same results as most offered nothing more than a quick peek. We communicated to several of these companies about why embedded e-mails, attachments and objects were so important. A few of these companies never responded back. Some EDD processing programs

were able to read and load more documents than that of an E-Mail reporting tool that is still on the market. I asked several EDD processing software companies why it could not read and extract attachments and embedded items in .PDF files. Their responses were along the lines of that are just too difficult to figure out from a development point of view.

Up until the end of October 2007 Randall Consulting never had any intention to develop software. We wanted to stay an electronic discovery consulting and training company and one who stays neutral and gives an unbiased opinion on EDD software products. However, an organization had to get involved to figure out a way to help solve the E-Mail accuracy issue. The best way to go in my opinion was a reporting tool. We felt so strongly about this issue and the ramifications of EDD programs not being 100% accurate that yes it was now time to develop a software program. In December of 2007 which was a year after we released the processing report software development began on a .PST reporting tool. A program that could go unlimited nesting, read multiple container files, all embedded items and to even read attachments in a .PDF files.

In 2008 new sample .PST data sets were created to see what kind of accuracy there would be with electronic discovery processing programs. One of the data sets was more of a hodgepodge of issues that processing programs have. It had .PDF's with attachments and embedded attachments. It had deep nesting and a lot of embedded objects. It also had some Office 2007 files. Not every program was tested. But those that were the highest grade was an 84%. The majority of the programs did a better job with nesting and recognizing some embedded items than that of a year prior. They all failed in reading and extracting anything in a .PDF file. We understood that most .PST files would not be this complicated and would not have as many embedded items. But for a complicated data set the best was 84% accurate. If we remove the attachments and embedded items from .PDF's in the equation the accuracy goes up to around 92%.

The second half of 2008 data sets were created of .NSF files to test again in some of the EDD processing programs. Most were unable to extract all e-mails and attachments. A few reported that a .NSF file was corrupt because it did not have the default e-mail template. Of the few reporting tools still on the market that can handle .NSF files everyone received a failing grade internally. Thus going back to the two words I used frequently in 2007...a quick peek.

To this author there is an E-Mail accuracy issue. Having a program read and extract say half of all embedded items are unacceptable. Having a program go four to five levels deep in terms of nesting is also unacceptable. Programs not being able to read attachments or any type of embedded items in a .PDF file are again unacceptable. Maybe two or three years ago this was accepted in the market place. With electronic discovery taking over and the FRCP rules now in affect for well over 2 years accuracy is extremely important.

I have seen two programs even at the export level fail to export all documents that it was supposed to. Now have these programs gotten better? Absolutely! Are they close to 100% accurate especially for larger E-Mail stores? In the testing we have done the answer would have to be no. Not to beat a dead horse here but one never knows if the documents that could not be loaded are imperative to a litigation

matter. The problems becomes that a user loading data in a 3rd party program or even a proprietary one is taking the total document count loaded at face value.

Exactus

Exactus (Latin for accurate...precise...Exact) has been in development since December 2007. The goal again with this program was to create something that this industry needs. To create a .PST reading and reporting tool that can read and report unlimited nesting, unlimited container files, embedded e-mails, attachments and objects. To do something no other processing program that I am aware can do is to read all attachments and embedded items in a .PDF file. To have detailed reporting options, some basic search functionality and the ability to see the parent and child relationship of each e-mail.

Exactus is not competition to an electronic discovery processing program instead it works as a complement for a user to take more control over their data.

Embedded Files - Rich Text Files, Microsoft Word Documents, Outlook Messages and other file types may have files embedded within them. Information contained in these files is not made available by simply printing or converting these documents. Embedded files need to be detected and processed, and their link to the container file maintained. In some cases there may be many levels of files inside files that require tracking and conversion.

Exactus is a .PST reporting tool, just released from Randall Consulting. There are a few rules to follow to effectively use this software.

1> Exactus should be the only software running on a computer when processing a .PST file. Microsoft Outlook should be closed when running it.

2> **A copy of a .PST file should always be used. That copy should not be used to then run the .PST file thru an electronic discovery processing program. Why? Exactus opens up every attachment and sniffs for embedded items. Thus it changes some file system metadata. This is extremely important. If a law firm is using Exactus and makes a copy that copy should not be given to a vendor to process.******

3> A user may find some E-Mail reporting tools to be faster than Exactus. Consider this though this is the most accurate utility on the market. Speed and performance will continue to improve as new updates are released.

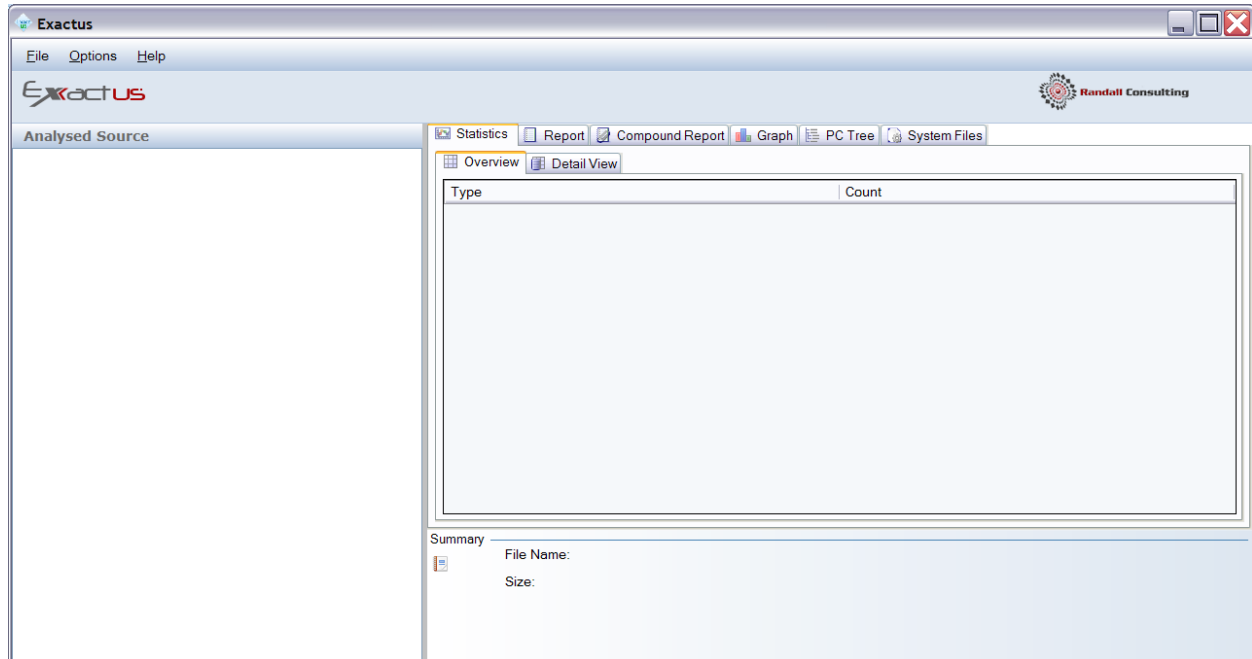
4> At the moment Exactus reads file types based on extensions. This will change in the next major release as extension reading will be replaced by that of file headers.

Upcoming Features

1> .NSF support should be added by the end of the year. As well as the ability to tell how many duplicates are in a .PST file.

2> Randall Consulting will also add support for loading multiple .PST files at once. For example if a custodian has more than one .PST file.

3> Interactive graphs.



Exactus opening screen.

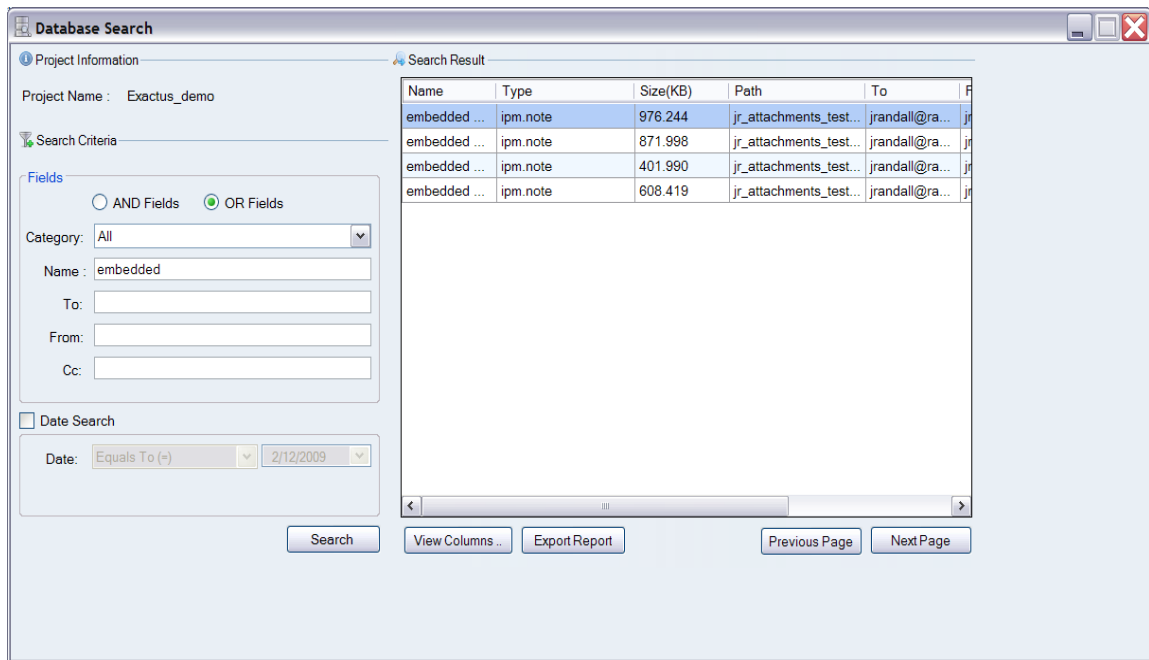
Why should a service provider use Exactus?

There are numerous reasons why all service providers should have Exactus in its EDD toolkit. I will go over a few here:

- 1> To be able to quickly and efficiently tell a client what is in a .PST E-Mail store. Such as total number of documents. This includes all attachments and embedded items. A service provider can also provide detailed reporting options such as the parent and child relationship of all e-mails, what attachment filetypes there are as well as what e-mails they come from. Also to be able to give a report on the compound document count of each e-mail. E-mail plus any attachments and embedded items. For example if an E-Mail had two attachments then the compound document count would be 3. To be able to tell their client if any system files are in a .PST. As well as all e-mail addresses within that .PST file.
- 2> The ability to use Exactus as a safeguard program in relation to their processing. A service provider would load a .PST file into Exactus to get a total document count. Then load a new copy of the same .PST file into their electronic discovery processing application whether it is a 3rd party product or proprietary. Do they match up? Does Exactus give a higher document count? Is

the processing software possibly missing some documents? This will then give a vendor an idea of what documents possibly need to be processed manually. The key here is to be close to 100% accurate as possible in processing all data given to you by a client.

- 3> To be able to use it as a data filtering reporting tool. Do you have a client giving you a date range and wanting to know how many e-mails come up? Or possibly asking for certain e-mail addresses in a .PST file? You can use Exactus to export out a report. For example here is a screenshot of the word **embedded** in all e-mails of this small .PST file.



- 4> Using Exactus as a means to test new software applications. Are you as a service provider looking at a different electronic discovery processing software? Or more than one? Run a few .PST files into Exactus and then do the same with the processing software. How accurate is the software you are testing?

What are the advantages for a law firm/Government Agency/ Corporation in using Exactus?

- 1> Having the ability to see what is in a .PST E-Mail store prior to sending it to a service provider. How accurate is a service provider in processing all documents in a .PST file? For example if running an E-Mail store thru Exactus and the document count is 1,200 documents which are all e-mails and attachments. Yet the document count you receive from your vendor is say 1,189

documents. Based on the reporting options available in Exactus you can now begin to analyze your service provider's performance. Were there any system files that were culled out? Was any de-duplication done? Or was it a situation that certain embedded documents could not be extracted? This is also a great way to test out a new service provider to see how well they perform from an accuracy point of view in terms of document count. The ability to run date searches in Exactus and see if your vendor finds the same number of e-mails and attachments.

- 2> If you do any processing in-house just like that from a service provider it is a great way to see how well your in-house processing tool is. To load a .PST into Exactus and get a document count and then see what the document count upon loading is for your processing program. Does it match?
- 3> As a way to verify to your client what is in a .PST store. With the numerous reporting options available in Exactus this is a way for a client to decide how it might be processed. Should it be filtered down? Are they looking for specific e-mail addresses or subjects? Should it be a native production?...etc
- 4> From a corporation point of view it gives you access to the .PST stores of your employees. Who are they e-mailing? What filetypes are being included as attachments that are sent out and coming in. Also as a way to verify if you do any processing in-house the accuracy of your product.

What are the key features in Exactus?

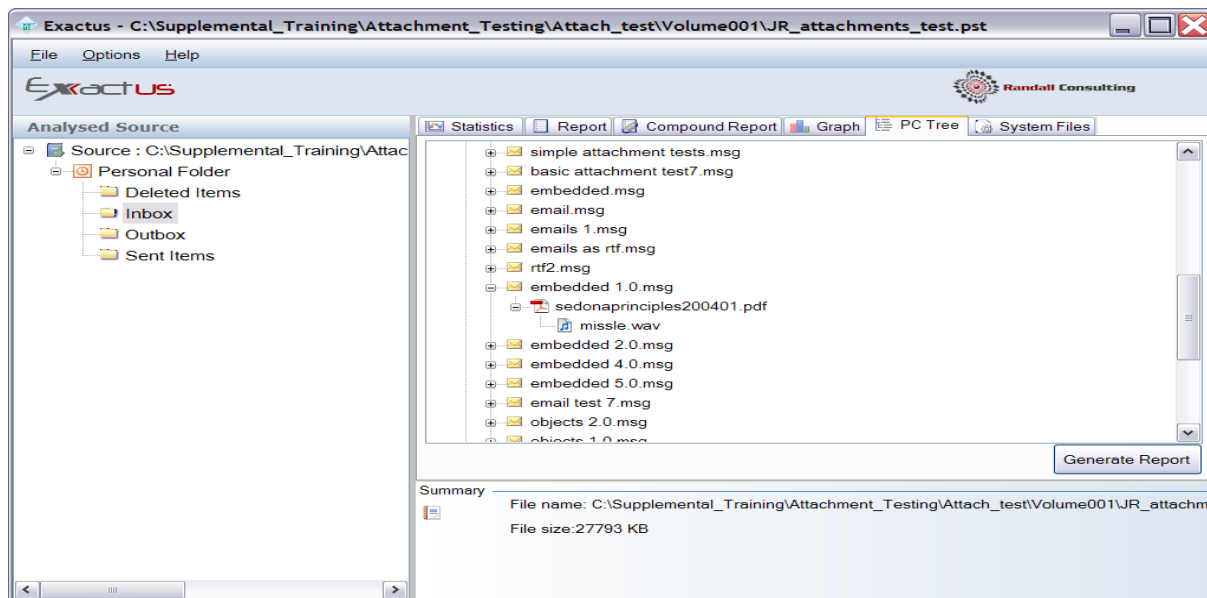
- Support Office Files (2005, 2007), PDF Files and External MSG Files.
- Support attachment drill down to N level
- Support various archive formats like zip, rar, tar, tar.gz.
- Support various outlook file formats. i.e. Notes, Calendar, Journal etc
- Supports Microsoft Outlook files i.e. EML files.
- Supports Unicode Characters in Folder and Message Level.
- Supports Parent Child View of Message and Attachments.
- Supports viewing and saving of specific type of Attachments in CSV and XML formats.
- Supports viewing and saving of Compound Documents
- Customized Graphical representation of message with regard to size and type count.
- Lists the documents available and saves it as a CSV file.
- Saves all unique e-mail addresses as a CSV file.
- Displays the full parent and child relationship for all embedded items and saves it as a CSV file.
- Reports Compound documents information.
- Reports all known system files(PST,DLL,EXE,OCX,BAT,COM,CMD,SYS,INI)
- Saves search result to CSV file.
- Option to drill N-Level deployed.
- Exception while extracting items from Password protected zip files are handled.

Convention Followed for Attachments:

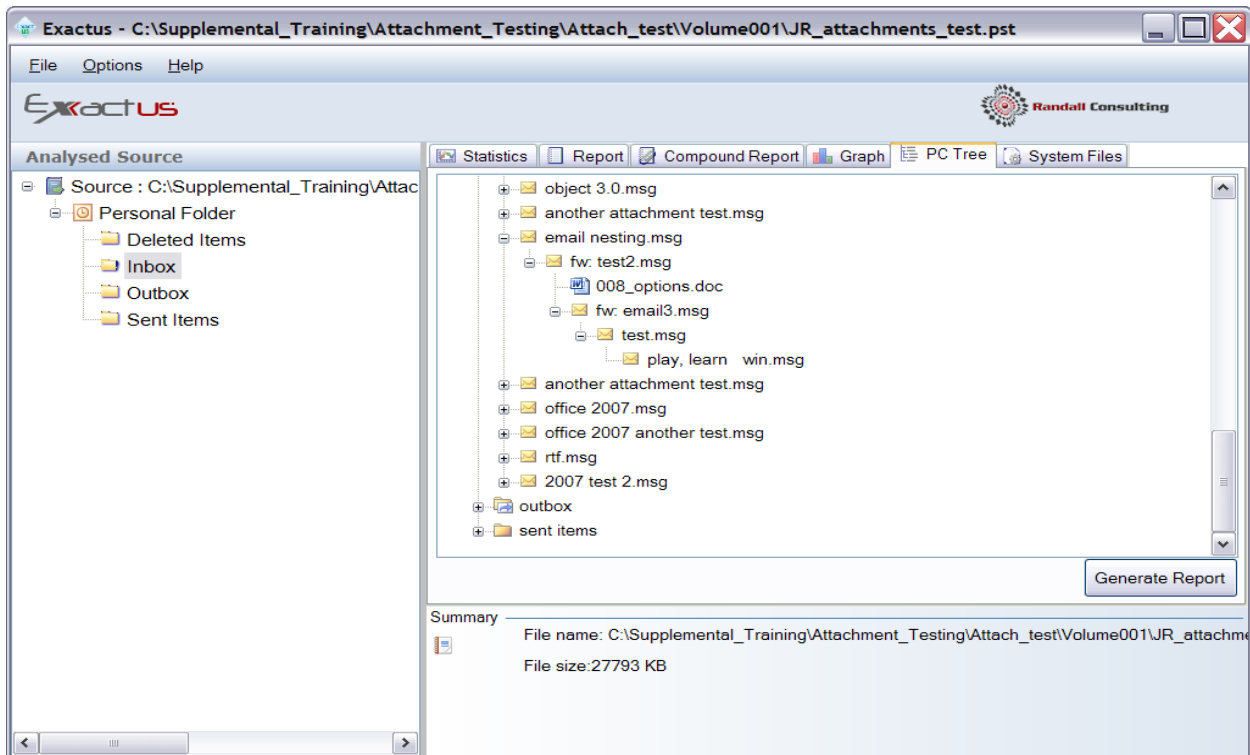
- Anything inside Office Files or PDF File are termed as Embedded Attachments
- Anything that is contained in the body of MSG file is termed as Embedded Objects.
- Every other thing is termed as General Attachments.
- Since for every attachment inside Office 2007 files, their snapshot image is also stored so Exactus extracts and lists them as well.

Parent and Child relationship tree

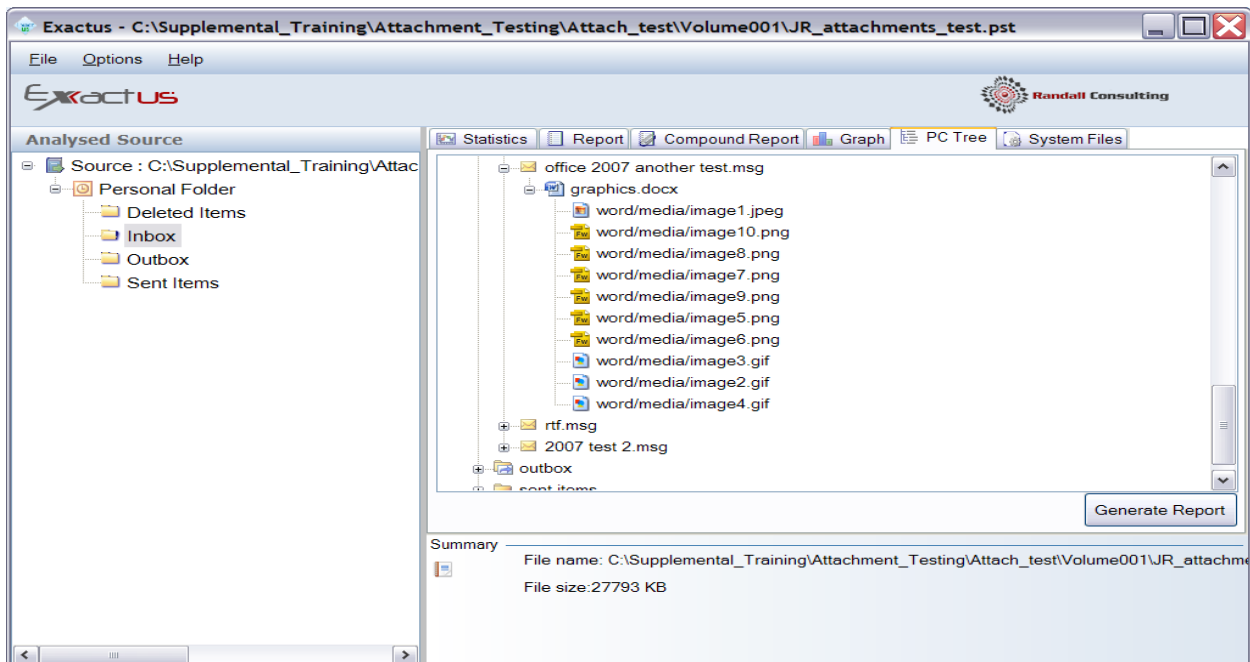
Another key feature in Exactus is the ability to see all e-mails along with their corresponding attachments and embedded items. One of the advantages of this program is that Exactus can read everything inside a .PDF file such as attachments and any embedded items. For example here is a screenshot from a small .PST file that I created. As you can see under Embedded 1.0 msg we have a PDF file with a .WAV as an attachment. Some processing tools would not be able to recognize.



Let us look at another e-mail called e-mail nesting. We have two embedded e-mails as attachments to start out with. In the first embedded e-mail titled test2 is a Microsoft word document. In the other embedded e-mail titled e-mail 3 there is another e-mail followed by another one. In this specific example we have 5 levels deep of nesting. The question again is can your electronic discovery processing program recognize all of these files?



What about Microsoft Office 2007 files? In this final example we have an e-mail titled office 2007 another test. There is a Microsoft Word 2007 file. Inside that are 10 assorted embedded object files. They are all graphic files. The question again is can your ediscovery processing program recognize all of these. Compound document count would be 12.



Yes, most .PST files would not be this complicated in terms of attachments in .PDF files and numerous embedded objects. When processing a .PST file one never knows what they might run into. Missing even a few documents could result in possible litigation down the line.

Questions to possibly consider with your electronic discovery processing application.

- How do you know that all e-mails and attachments are being properly loaded in your electronic discovery processing application?
- Have you done any auditing of a test file?
- Can your program handle embedded items?
- Does it give an accurate count of the documents it could not extract?
- Can your processing tool read and extract attachments within Adobe 8 files?
- Can it read container files within container files?
- How deep can it nest?

Randall Consulting also offer a 15 day fully functional trial version of Exactus. To continue using it after that requires that a license be purchased.

It can be downloaded here:

<http://store-.randallconsulting.net/>

This white paper was written by John Randall. Any comments please e-mail jrandall@randallconsulting.net

Do you have a 3rd party processing program or a proprietry one? Exactus can be integrated via OEM opportunities.

