(U) Ask Raul: Getting the Most from Metadata

FROM: Raul
A DNI Analyst
Run Date: 04/28/2005

Dear Raul,

(C) I am swamped with metadata and things that look like metadata. Can you think of something important and easy that I should be able to use from all the metadata we generate which would provide a quick and easy payoff?

Betsy

Dear Betsy,

(C) I feel your pain. To get directly to your question, let me recommend the fields found in OLE (Object Linking and Embedding) property sets for Microsoft Office products. What the heck are you talking about?! Let me explain.

(U//FOUO) Ages ago (I'm looking at an article from 1994 on OLE 2 and I know I have a book at home from the 80's on OLE), Microsoft began using OLE to make their products, especially documents, spiffy. To save you a bunch of details, the basic idea behind OLE was that each OLE document would take the form of a file system and if you know what you are doing, you can read it. Here is an example:

sample.xls:OK :Workbook :
Root = Start / Count = 006c / 1
SBD = Start / Count = fffffffe / 0
Size of BBD = 1
BBD = 006b
OLE DIRECTORY
ID Name Type Start Size Prev Next Dir
0:Root Entry ROOT 0000 0 -1 -1 2
1:Workbook FILE 0000 46112 -1 -1 -1
2:eSummaryInformation FILE 005b 4096 1 3 -1
3:eDocumentSummaryInformation FILE 0063 4096 -1 -1 -1

This is what a plain old Excel spreadsheet looks like, OLE-wise. This structure is such that it makes it fairly straightforward to repair some damaged documents. Don't get me going here! Just thinking about it has me fit to be tied.

(S) The Root Entry tells us where everything is. The spreadsheet is actually the Workbook section but of particular interest here are the eSummaryInformation and eDocumentSummaryInformation property sets. Inside these property sets is information you can use to track your target such as the company name, author and last author. Here is an example of a dump of a property set from a document we collected:

OS: Win32 (Windows) Version 5.1
Summary Information Property Set

SERIES:
(U) Ask Raul - Answers to DNI Questions

Codepage is 1256 Arabic
Title: No
Subject: Empty. No value.
Author: Masoud
Keywords: Empty. No value.
Comments: Empty. No value.
Template: Normal.dot
LastAuthor: Masoud
RevNumber: 8
AppName: Microsoft Word 9.0
EditTime: 20 Hrs 28 Min
Create_DTM: April 11, 2005 1246 hrs
LastSave_DTM: April 11, 2005 1453 hrs
PageCount: 1
WordCount: 3450
CharCount: 19670
Security: 0
Summary 1

OS: Win32 (Windows) Version 5.1
Document Summary Information Property Set
Codepage is 1256 Arabic
Company: IRISL
LineCount: 163
ParCount: 39
Invalid or Unknown Document Summary Property 17 value is 24156
Invalid or Unknown Document Summary Property 23 value is 592544
Scale: False
LinksDirty: False
Invalid or Unknown Document Summary Property 19 value is False
Invalid or Unknown Document Summary Property 22 value is False
DocParts: ... Error. No further processing.
Doc_Summary 1

Notice the Author's name: Masoud, Last Author: Masoud and Company: IRISL. If you are still awake, you can probably guess that this info could also be used to quickly link identical documents from different sources together or as a means to validate the integrity of documents as well as a whole host of other things.

(C) The two most commonly collected file types you will see these property sets in are Word and Excel. They aren't the only file types, but certainly the most common.

(S) Now the nice thing about this is that often, very often, our targets fill in these values, or the machine does it for them, and they never get changed. Are things getting clearer? If Joe Badguy isn't paying attention, and he usually isn't, and you get an OLE document from him, you can then target what he has in his property sets or you can go look in Pinwale for additional documents from him without having to know his email address, domain or anything else. You can also become a DNI wizard and guess at what might be in the property sets and have some fairly good luck. In the example above, IRISL is the publicly known abbreviation for the foreign company in question. We could have guessed at this without having to have the first collection to discover it. You can hunt, not fish.

(S) Now here's the nitty-gritty. The property sets use a specific

format which is at least 15 years old now. Using the example above, the Company name is actually stored as follows in the file:

0x1e 0x00 0x00 0x00 0x06 0x00 0x00 0x00 I R I S L 0x00

Here is what it means:

0x00 0x00 0x00 0x1e String (8-bit), null terminated
0x06 0x00 0x00 0x00 String length + null
String IRISL0x00

(The first two values are reversed because they are stored in the file in Little Endian format.)

If we wanted, we could slap the above hex string into a dictionary and suddenly we'd be hauling in Word and Excel documents from IRISL without having to know who was sending it or to whom. Additionally, we could take this term and convert it to base64 and task that also. Like this:

AAAeAAAABgAAAElSSVNMAC8v
LwAAHgAAAAYAAABJUklTTAAvLy8v
Ly8AAB4AAAAGAAAASVJJU0wALy8A

Although this isn't perfect (again, don't get me going), it will work just fine and keeps you from having to worry about what equipment is or isn't at what site. It's that simple. The Unicode version of the string type is not much different. Plus, lucky us, what we do for the dictionary is easily adapted for Pinwale and vice versa.

(S) Now, imagine this: You're sitting here digging through the mess you've made with your traffic and lucky you, you find a Word document with a value set for the company or author in the property set. You then take that string and search through Pinwale and discover that there were more documents lying there waiting for you that you otherwise never would have found. You didn't have the email addresses or other hard selectors but you had that Word document and the property sets. Additionally, you now harvest several new email addresses and other bits of EEI from the sessions you found that help you to find even more goodies. Pretty spiffy!

(S) Needless to say, if we harvested this data and used it properly, it would also be a big help for DNI chaining. Let's say Joe Badguy, whom we know about, sends a message with an OLE document in it to Bill Badguy, whom we don't know about and unlucky us, we don't collect this session. Three months later, Bill forwards that file to Tom Badguy, again whom we don't know about but because he is on a link we're watching, and because we had the value in the property set tasked, we haul in the session and suddenly we know about Bill and Tom and that they are associated with Joe. Not bad! Some of us have actually been doing this. It works.

(C) There you go. In case you are wondering, corporately we've had the ability to extract the data from property sets for about five years. Also, other data types have similar bits of information that can be exploited. Is that metadata being extracted? It pays to know how your data works.

(U//FOUO) Enjoy! DNI is a lot easier than you think. Or at least it can be.

Raul
