Using a Malware Ontology to Make Progress Towards a Science of Cybersecurity Transcript

Part 1: Why Ontologies Are Critical

Julia Allen: Welcome to CERT's Podcast Series: Security for Business Leaders. The CERT Program is part of the Software Engineering Institute, a federally-funded research and development center at Carnegie Mellon University in Pittsburgh, Pennsylvania. You can find out more about us at cert.org. Show notes for today's conversation are available at our podcast website.

My name is Julia Allen. I'm a principal researcher at CERT, working on operational resilience and measurement. Today I'm very pleased to welcome David Mundie. David is one of my colleagues and a member of CERT's Enterprise Threat and Vulnerability Analysis Team. I think you'll find today's conversation pretty interesting. It's a little bit of a departure from our normal operational topics.

Today, David and I will be discussing the need for controlled vocabularies, taxonomies, and ontologies -- all to the end pursuit of making some substantive progress towards a science of cybersecurity, as opposed to an art form. And David and I will be specifically discussing malicious code, also referred to as malware analysis, based on a report that he and his team have recently published, called "The MAL: A Malware Analysis Lexicon," to try and give you a little bit of an example of what we're talking about.

So with no further ado, welcome David; really good to have you on the podcast series.

David Mundie: Well thank you Julia. This is a great time to be an ontologist and I'm happy to have the opportunity to talk about some of the recent developments in this field.

Julia Allen: Yes, I suppose not too many people might know what an ontologist is. So I look forward to you illuminating that area of expertise for us. So let's set the stage a little bit. And when we talk about a science of cybersecurity -- usually the words science and cybersecurity aren't juxtaposed. So that in and of itself I think is a great thought to pursue. But when you talk about a science of cybersecurity, what exactly do you mean?

David Mundie: Well of course Julia a lot of things go into making up a science. But I think that one of the most basic elements was highlighted by a MITRE report that came out back in 2010 and that report was actually entitled "The Science of Cyber-Security." That report came about because the DoD had asked the authors to take a look at the theory and practice of cybersecurity and to figure out what would really be needed to make cybersecurity into a science, the assumption being that it is not currently a science.

And the very first conclusion of the report was that the most important development would be the creation of - - and I quote-- "a common language and a set of basic concepts about which the security community can develop a shared understanding." In other words, they were saying that what we need is an ontology of security. And in my view, all the other aspects of science -- the statistics, the hypothesis testing, etc. -- all of that can only be built on top of that shared understanding that the report highlighted.

Julia Allen: Great. So if I'm following you, you're talking about having a language or terminology or a common understanding. I can certainly appreciate why that's foundational to any science.

But to expand on that a little bit, could you say more about ontologies in general and what you've observed as their importance today?

David Mundie: Sure. First off, we need to make clear that we're not talking about the philosophical notion of ontology, which has a long history going back all the way to the Greeks. And that's the study of what is, the study of the concepts of being and becoming and so forth. What we're talking about is the modern, computer science notation of an ontology. And that is simply a hierarchical collection of standardized terms, a controlled vocabulary, along with the relationships among those terms.

And this whole idea began to crystalize around the year 2000 when DARPA (Defense Advanced Research Projects Agency) came out with an initiative to apply some very basic artificial intelligence techniques to the World Wide Web. The theory was that the World Wide Web was this enormous knowledge representation that could really benefit from having some artificial intelligence applied to it.

And that idea merged with the idea of a semantic web that came out of the World Wide Web Consortium from the inventor of the World Wide Web, Tim Berners-Lee to be precise. And that produced a web ontology language called OWL. And it's taken awhile -- or maybe I'm just an impatient person -- but it's taken awhile for this idea to get traction. But it seems to me it is finally taking off and that's part of why I'm so excited about all this.

Just to take a couple of examples: The 11th edition of the International Classification of Diseases, which is the standard for describing medical conditions -- and it has 68,000 diseases and their interrelationships -- and all of that is now being produced as an OWL ontology. And another nice example of this is what Google calls their semantic network that has 570 million objects in it and 18 billion facts. And I could go on. There are many other examples. But I just have a feeling that this is something that, something whose time has come.

Julia Allen: So this whole idea of having a structured way of capturing, codifying, and describing a language for a particular domain or discipline, that's really what is behind this emergence or this increasing interest in a construct like OWL. Is that correct?

David Mundie: Yes, that's right. And this goes hand in hand with a lot of developments in logic. As many of your listeners will understand, traditionally it's been very difficult to avoid things like deadlock to make sure that the reasoning terminates in a reasonable amount of time.

And just in the last ten years, more and more logics that improve on the state of the art in terms of that reasoning -- which is what makes it possible to apply these reasoning techniques to the particular language, OWL, at a very large scale. That's why we can talk about doing this in a big data context.

Part 2: An Ontology for Malware Analysis and Competency Frameworks

Julia Allen: Great. So, this is great in terms of background but I suspect our listeners might be itching to get to the meat of our conversation because this is a security podcast series, right?

David Mundie: Yes.

Julia Allen: So I know that you've specifically applied this very rich background and understanding to the malware or malicious code lexicon that you've recently been working on

and attempted to fit it into this picture of ontologies and taxonomies. So can you say a little bit about how that particular lexicon fits into this overall picture?

David Mundie: Sure. Obviously we have fewer than 300 lexical items, 300 items in our ontology. So it's certainly not on the same scale as the International Classification of Diseases. But the underlying goal of being able to reason about a formal representation of a knowledge domain -- that is the same.

And I'd just like to say that we started with the controlled vocabulary, the lexicon, which was published in the tech note. But since then we've moved on to actually create an OWL ontology out of the lexicon, using Protégé, which is Stanford's ontology creation tool. And on top of that we've started building what is called an ontology- based competency framework.

In the last few years there've been a number of initiatives to use ontologies as the foundation for managing workforce competencies. And this makes a lot of sense to us because knowing about the pieces of the domain and their interrelationships makes things like reasoning about job descriptions or task analyses or figuring out what the training needs are and many other aspects of human resource management -- the ontologies make all of that a lot easier. And we think that's particularly true in malware analysis, where there are a number of very difficult human resource questions.

Julia Allen: Until you and I started preparing for this podcast, I hadn't really thought about -- I mean, I can see the benefit of having a foundational language for a domain or a body of knowledge like malicious code.

But I never thought about its applicability to the knowledge, skills, and abilities of the people that were actually going to go try to find malware and analyze malware and correct malware and come up with doing root cause analysis for malware that actually not applies just to the domain but the people that are actually performing in that domain. That's what you're saying here, right?

David Mundie: Yes, exactly. And if it's any consolation to you, we have thought about doing competency models for malware analysis for a number of years now. But it is only quite recently that I discovered this notion of an ontology-based competency framework.

And our current plan is actually to have a small internal R&D project to pursue that idea and to more fully develop the ontology-based competency framework that we have started.

Julia Allen: Great. So let's go a level deeper, if you would be so kind, and talk about how you actually went about developing your malware analysis lexicon. I think that's a fascinating story of how you get from something that's vague and not well specified, where there's a lot of ambiguity, to something that's much more precise.

So could you take us on that journey?

David Mundie: Oh Julia, that's so wonderful. There are so few people out there that find lexicography exciting. I'm glad to hear that you find it an interesting topic.

Julia Allen: Well, we'll see if it captures our listeners' interest. But I do find it a fascinating topic. So please go ahead.

David Mundie: Well from a lexicographic standpoint, I think the most interesting thing we did was to gain access to ten years' worth of email from the CERT Malware Analysis Team. And that was a great corpus to start from. But the problem was that it was about 90,000 terms. And most of them were days of the week and what people had for lunch and so forth. And so the issue that we grappled with was how to reduce those 90,000 terms to just the ones that had to do with malware analysis.

And I'll tell you, that was quite an adventure. The first step that we took was to use some scripts that were developed internally for other purposes that looked for any term that occurred in Moby Dick or the Bible -- on the grounds that Captain Ahab was pretty unlikely to have known about malicious software. So that if it's a term that occurred in Moby Dick, it's probably not a term for some sort of malware technique.

Julia Allen: And David, if I may interject, is that a pretty common approach? Because you want to get rid of -- if you will, you're talking about signal to noise -- so you want to get rid of the noise, the background data, the terms that really don't contribute to your taxonomy or ontology of interest. So is using Moby Dick and the Bible and other sources like that -- pardon my ignorance -- but is that a pretty common practice?

David Mundie: Well as far as I know, we're the only ones who use Moby Dick and I think that was just a tactical decision. But the general principle of filtering out known uninteresting terms, yes, that's pretty widely spread. And I'll just throw in here that there's a lot of research being done today in how to automatically mine older forms of knowledge representation -- legacy systems -- in order to extract an ontology out of them.

And yes, as part of doing that, filtering out is absolutely essential because otherwise there'll be just too much noise for the amount of signal you get.

Julia Allen: Okay. So once you went through that first pass, what happened next?

David Mundie: Well, we didn't get the number down as much as we had hoped to because we still had about 70,000 terms in the lexicon after we ran all the scripts on it. And the realization that came to us was that well it might be that there was a very important malware term that only occurred once in those ten years of emails but the probability of that happening was pretty low.

And so what we did was to strip out all the terms that occurred only four times or less. And that got the lexicon down to just about 5000 items. And my colleagues and I sat down and manually went through all of them, and finally came up with 40 terms out of the 70,000 that were clearly malware analysis terms.

So that was a lot of filtering for 40 terms. In addition to that corpus, we also looked in textbooks on malware analysis and came up with about 200 terms and on various internet sites we found about 25 more. And so we ended up with about 270 - 275 lexical entries.

Julia Allen: Okay. And then was there a final step that you went through?

David Mundie: Yes, we thought a lot about what to do with the dictionary once it was finished. And we published the tech note, of course. But then we also encoded them in a structured dictionary that's based on the IETF's -- the Internet Engineering Task Force -- standard for dictionary servers.

We're working on an actual e-pub file; an electronic version of it as an e-pub. And just for good measure we threw in a couple of other vocabularies that we happened to have kicking around; like the (SEI's) CMMI vocabulary, the vocabulary from the (CERT) Resilience Management Model, our insider threat vocabulary. It's a fairly large lexicon of information security terms.

Julia Allen: Okay so that's great. So you were able to fold in some other CERT bodies of knowledge, just to see how they would round out some of the language and structure that you wanted to capture, correct?

David Mundie: Yes. Call me a romantic, call me an idealist, but I have this vision of everybody at the SEI and beyond -- well it's the vision of the MITRE report; a common vocabulary and common understanding about all these terms. And I think that the more comprehensive that can be, the better off we are.

Part 3: Additional Security Ontologies in Development

Julia Allen: Fantastic. Well I know that this is just the beginning of a very exciting journey for you and your research colleagues. So before I let you go, could you say a little bit about other related ontological development efforts to see if we can push this whole science of cybersecurity forward? Are you applying this same method and type of thinking to other domains?

David Mundie: Yes I am. And I am very proud of the fact that nobody looks at me funny when I talk about an ontology anymore. A couple of years ago that was not so much the case. But I'm not the only one now that is talking about ontologies.

We're actually working on a variety of them inside CERT. We have one for insider threat detection; for insider threat modeling; for coordination center information sharing; for incident response. And they're all in varying degrees of completeness. And we're certainly looking forward to continuing to work on them in the upcoming months and eventually to merge them, again, into one large ontology.

Julia Allen: Excellent. Well David, do you have some resources for anybody whose interest we have piqued on this very foundational topic for our field -- some places where they can get some additional information?

David Mundie: Yes. The method that we are following -- and we sort of invented this method for developing ontologies -- was published last year in the Proceedings of the First Annual Workshop on Taxonomies and Ontologies for Security. And that also, by the way, includes an informal ontology for incident management.

And we have submitted a paper, which we hope will appear in that same Workshop on Taxonomies and Ontologies for Security for 2013 and that paper will be about our ontology-based competency framework. And the malware vocabulary itself is described in an SEI technical note. As you mentioned it was -- the title of that is: "The MAL: A Malware Analysis Lexicon." And that's available on the SEI and the CERT websites.

Julia Allen: And you also mentioned this idea of dictionaries and OWL and other electronic or digital representations of these ontologies. Is there any currently available -- and maybe it's a future thing -- but is there any place where folks can go if they wanted to check out an electronic or digital structure for this type of information?

David Mundie: Yes. The W3C -- World Wide Web Consortium -- is the owner and publisher of the OWL standard itself, which is by now pretty much the de facto standard for doing computer ontologies.

So I would recommend looking at the OWL standard itself. It's not all that complicated. Another very good site is the Protégé site at Stanford -- Protégé being the ontology development tool that, at least in the open source world, is the standard for doing that.

Julia Allen: Fantastic. Well David, I can't thank you enough for your time, for your expertise, for the excellent preparation for our conversation today. I really appreciate it and thank you so much.

David Mundie: It's been my pleasure.