The Importance of Data Quality

featuring Dave Zubrow interviewed by Shane McGraw

Shane McGraw: Welcome to the SEI Podcast Series, a production of the Carnegie Mellon Software Engineering Institute. The SEI is a federally funded research and development center at Carnegie Mellon University in Pittsburgh, Pennsylvania. To learn more about the SEI, please our website at sei.cmu.edu. A transcript of today's podcast is posted on the SEI website. My name is Shane McGraw, and today I am pleased to introduce you to Dave Zubrow, manager of the Software Engineering Measurement and Analysis Initiative at the SEI. His research focuses on applications of measurement and analysis techniques to process improvement, project management, and product development. In today's podcast, Dave and I will be discussing the importance of data quality. Welcome, Dave.

Dave Zubrow: Thank you, Shane, I'm glad to be here.

Shane: Okay. So, first question for you. According to a recent report, released by Gartner, the average organization loses 8.2 million dollars annually because of poor data quality. So let's start off by having you tell us about the impact of poor data quality in organizations that you've worked with.

Dave: Well, Shane, the impact can be tremendous; and the work of many researchers in this area on data quality has primarily focused on outward-facing data, and so, you think about your billing issues. You know, mistakes in billing; and, in fact, I had an example happen to me yesterday that I can share with you about a bill from a physician's office. And I was questioning it with my insurance company, and what happened when I called the insurance company about the claim, is they immediately said there must be a mistake on this. It was 250 dollars. But, it was just a telling incident of the way in which data quality can affect our everyday lives. Right? And if I hadn't have called, and found out that it had been a coding error in the type of procedure that had been done, I would have just ended up paying that cost. So you can see it in billing. Think about when you go to a restaurant. There are mistakes there.

Dave: You can think about mistakes, perhaps, on filings with the IRS. Since it's April here, when we're recording this, that's probably a topic that's on a lot of peoples' minds. Now, those become very visible, and there's a huge industry that's grown up around data quality and the notion that data warehousing, because, as they say, garbage in, garbage out. And so you want to make sure that the data are good. An example we encountered with this was with one of our research projects, when we were trying to better understand a defect analysis and defect detection process. And we went in and started looking at the data, and cleaning the data up before we were going to conduct the kind of analysis, you know, that was part of our research. And one of the things we discovered is that from the perspective that we were looking, the data didn't seem to make any sense. And yet, this data had been in the organization for many, many years. And one of the things that happens is when you look at data from a transaction viewpoint, what do I need to do to close this inspection? To resolve this defect? Right? There may be integrity there, but when we go through to say "Let's analyze your overall defect removal process," now we're doing these aggregates, and you start to see patterns emerge that are inconsistent. Errors will become apparent that aren't apparent in the small, if you will, at the transaction level. This kind of data quality issue is one that, in my view, and from our research, has been relatively hidden, if you will. Because the apparent financial consequences are not as visible.

Shane: Right.

Dave: People don't see it. You don't have customers calling you angry, because of this kind of thing. So, this got us thinking, that there's probably a need for some work in this particular area. And the other confluence of events is as more organizations attempt to reach or strive to reach high maturity, they're analyzing and using their data more and more and more. And they're building models on it, they're using it to forecast the future of projects, or to make decisions regarding investments and process improvement. Well, we want that foundation, that data that they're using to drive those decisions, to be as good as possible. And this is another area where we're finding that there's a lot of error. And not only is there a lot of error, but the real problem is nobody's looking for it. And so, there's a presumption that once the data have been captured and recorded, that they're accurate, that they're of high quality. And that is an assumption that needs to be checked.

Shane: So is this something that you can get in front of executives and explain? Are they open to hearing this information, or is your research needed to back it up?

Dave: Well that's a wonderful question, because unless they've experienced the problem directly, it's one of those ones that tends to be sort of hidden away.

Shane: Right.

Dave: And so you have to either be able to go in and do an analysis, a small pilot study within their organization and show them the results, or as we're doing work in this area and building up more and more examples, we take that along, too. But, you know, it's sort of, "If it's not broken, don't fix it." And if their perception—and again, perception is the keyword here—if their perception is that it is not broken, that they don't have a problem in this area, then they don't often see the need to invest. Some *will* say, "I would like to know where we stand in that regard," and that's wonderful because then you have that opportunity to go in and do the research, and have the engagement to produce some results, and say "Hey, you know, you really are looking pretty good., you know, and my hat's off to you." Or, "Here's where you might run into some trouble, because of the data quality problem." And one thing I want to point out, and this led us into some other work in this area, is that understanding data quality is only one part. I already mentioned that, really the reason you have concern about data quality is the use of the data for making decisions that can significantly impact the organization. You're making investments. You're making decisions about customers, you're making decisions about new products to pursue, perhaps, or process improvement investments. That's one part, and one expression of it. That's where it all comes home to roost, so to speak. But, the quality of the data are influenced by the way in which they're collected, too, so the planning and the analysis of what's leading to issues with data quality has to ultimately trace back to "What are your data collection methods?" So, are they automated? Or, if they're manual, what kind of training, what kind of forms, what kind of support.

Shane: And does that differ from organization to organization a lot, or is it just every organization makes their own decisions?

Dave: Yeah. Oh, absolutely. Absolutely. And the degree to which they enforce what they say they're going to do, and to what extent is that actually practiced, will vary from organization to organization.

Shane: Okay.

Dave: And so you might even think, if you want to know how much effort it takes to perform a project, well then you have to be recording people's effort in terms of the projects that they're working on. Right? And furthermore, then you also want to understand to what extent is there any kind of monitoring or auditing of that reporting, if it's done, self-reports, and all individuals ever see is that it's used, you know, to pay me for my 40 hours per week, and there's no feedback at all about the degree to which they're accurately charging against their different accounts. Then, what you get is basically 40 hours a week charged to whatever's most convenient to them at the time.

Shane: Or I would imagine if something takes 60 hours to complete, that work is unaccounted for.

Dave: That's actually a real important different kind of data quality. And here's where that becomes pernicious, is in the realm of project estimating. So, what happens in that case, to follow up on your example, Shane, is we use the historical data as to what we believe was the effort it took to create that.

Shane: Right.

Dave: And we use that to estimate the next one. Now what have we done? We've built into that estimate the fact that you're going to give me one and a half times your effort to complete that project. And that becomes very difficult when we start bidding to customers, based on our historical data which underreports the amount of effort it actually took.

And that's another kind of data quality problem. But the issue of data quality creeps in in many, many guises. And you can think about it as maybe a typo. Okay. Sometimes that's what the most common thing. You slip a digit somewhere. And that's what people think of. But there can be data that appears to be accurate, okay, it sort of masquerading as accurate that's an error. And so, should it have been, you know, if it was 60 hours, and you billed that in a week, maybe that would be flagged because it's overtime in some systems? They'd have a check for that. There might be a pay differential, and so that might garner some attention. But what if it was the difference between 20 and 30 hours?

Shane: Right.

Dave: That may not cause anybody to look at it, in which case it's still a significant difference.

Shane: Absolutely.

Dave: But, it appears to be accurate. And those kind of subtle errors are actually the focus of some research that we did this past year.

Shane: Excellent. We'll go with the next question for you here, Dave. A number of software applications have been introduced in recent years to help organizations address data quality issues. Can you tell us a little bit about your research and how it complements these applications?

Dave: Right. So one of the things that we looked at was the functionality that's provided by some of these tools. And they can do a lot of business-rule-type checking. They can be configured to do that, and they'll also run some very basic integrity checks, and so when I say that, checking for missing data, checking data type. You know. If this is supposed to be a numeric field, is it**?** Are you putting alpha characters in that field or not? If the range is supposed

Dave: to be from zero to 100, you shouldn't be seeing negative numbers in there, and you shouldn't be seeing 101 in there. Okay. So there's some very basic types of integrity checking

that can be done. Summing certain fields should sum to 100. You can enforce that. You can enforce "This number should never be greater than that number." So there's those kinds of tools. And they can automate that, and it's better yet when that kind of checking is automated as part of the data entry process, so you stop the erroneous data at the point of entry, and make whoever is entering it or whatever systems it's coming from actually go back and correct that value before it gets in. It's always going to cost you more to fix it after the fact, and it's very hard to go back to the point of origin and correct data once it's entered the system.

Now, our specific research last year was to investigate the use of some statistical techniques, primarily associated with outlier detection. And the idea was that perhaps these techniques would provide value beyond sort of the basic business-rule-checking. So it's really looking at trying to find more subtle kinds of errors, based on patterns. And also to find errors that may require a historical distribution, you know, as a foundation. So, beyond the business rule kind of configuration. So we conducted that research; we had some promising results around several statistical techniques. Unfortunately, the data set that we were using had some limitations to it, and so the ability to gauge just how much beyond the margin or how much better, what the value add would be beyond the business rules, we were not able to actually measure that or assess that particular value. But we were able to say, "Here's a set of techniques that were demonstrably better than another set." And that's looking at them from the viewpoint of their statistical performance. Now, from a viewpoint of an organization implementing these, they would also be concerned about the degree to which these techniques can be automated, the ease of implementing them, and the costs for doing so. Because one of the things that's also important is the automation. And this comes back around to the tools. Because what companies are trying to do, if you will, is take humans out of the loop.

Shane: Right.

Dave: Okay. Reduce that labor effort and try to put as much onto the automation. As my story at the beginning of this interview illustrated about my encounter with the insurance company yesterday, the human in the loop immediately saw a discrepancy in the billing codes for all the procedures listed on the form. And said, oh, was this a routine procedure? And I said, yes. And she said, oh, well, you know. Then, if that's true, then you should not have been charged this amount.

Shane: Right.

Dave: Okay. So, it's that kind of intelligence that you'd like to start to see built in to some of these systems. Certainly it can be done, but how common that is, to what extent that the tools allow you to actually configure all of that information, and then, that gets deployed.

Shane: And is the cost justified? Catching just a couple errors right up front?

Dave: Depending on how many errors there are in the system, and how severe, what their potential consequence of those errors are.

Shane: Great.

Dave: Then that would also figure in to the, you know, the cost effectiveness or cost justification for deploying and configuring these kinds of tools. And especially when you think today, some of these systems are interfacing with inputs from many, many different sources, and then you're trying to configure, you know, have a configuration for each individual type of source that you're pulling from.

Shane: Okay. So where can listeners go if they want to learn more about your work in data quality? Where would you point them to?

Dave: So, our website would be a wonderful starting source, and as mentioned at the beginning of this podcast, it's www.sei.cmu.edu, and in addition to that, you would add forward slash SEMA.

Shane: S-E-M-A.

Dave: On there, we'll have a link for data quality, and associated with that page, then, you would find out not only the work that I've described, but some reports talking about the measurement and analysis infrastructure diagnostic. And this is a report and an idea that talks about sort of data in terms of its life cycle. From the time it's being planned and generated, to when it's turned into information that's being used by a decision maker. And it's important to take that life cycle perspective when thinking about data quality, because there are many things that will influence the quality of the data, but the data by themselves only sort of represent potential energy or potential value. And it's when you transform them into information and get decision makers to use it, that it becomes kinetic, that there's action that's taken and there's value realized by the organization.

Shane: Absolutely. Dave, thank you very much for joining us today. Great information. Some of the reports that Dave was talking about, one was titled Issues and Opportunities for Improving the Quality and Use of Data in the Department of Defense, and the other is Can You Trust Your Data? Establishing the Need for a Measurement and Analysis Infrastructure Diagnostic. And they're at that URL Dave read off a moment before. This podcast is also available on the SEI website at sei.cmu.edu/podcasts, and on Carnegie Mellon University's iTunes U site. As always, if you have any questions, please don't hesitate to email us at info@sei.cmu.edu. Thank you.

Dave: Thank you.