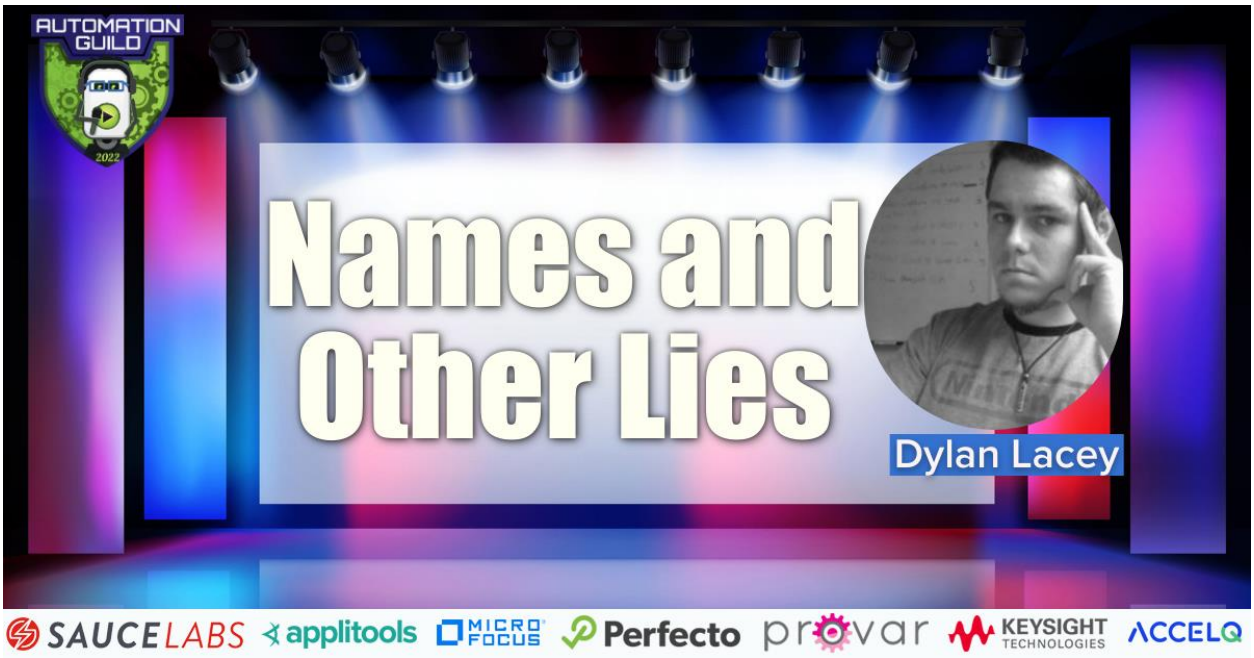

===

[00:00:00]

[00:00:00] G'day folks, I want to tell you a story about names and other lies. It's applicable worldwide, but mostly. It's a story about Japan.

[00:00:13] In 2016, I'd been working at Sauce Labs for three years. Every year Sauce gives us a retention bonus. And the three-year bonus was a travel package anywhere I wanted. Well, I've wanted to go somewhere ever since. I badly attempted to learn the language in high school to Japan. And I loved it. We went back in 2017 and in 2018 and in 2019, and in 2020, we didn't go and we didn't go in 20 21 either.

[00:00:45] So instead I did what every reasonable adult does and went back to college part-time and got a linguistics degree. Well, that's still in progress, but as part of the degree, I wanted to write to my friend, Hajime. Now [00:01:00] Hajime lives in Tokyo. He's a conversation partner of mine, and he really loves postcards.

[00:01:04] So I decided to send him one from my hometown of Brisbane. And just in the process of writing that postcard, I got thinking about names and addresses. So I did some research and what I discovered about them was really surprising. And I want to talk about how that relates to testing. But first I'm going to have to talk about what testing means because testing means getting things, right.

[00:01:28] We want to make sure that we spend the right amount of money with the right risk profile and get the right functionality for it. Now, usually when we're testing, we're using acceptance criteria,

[00:01:41] See, here's the thing about acceptance criteria. Acceptance criteria often suck. They're only a proxy for being right. Now when what we're trying to do is say, implement a tax algorithm. Well, that's easy. We can look up the algorithm, make sure that our acceptance criteria says do these things. [00:02:00] But when we're talking about names and addresses and stuff, we're relying on what everybody knows about names and addresses, which isn't always right.

[00:02:09] So instead of acceptance, I want to focus on what I'm calling the diverse demographic principles. These are ways that we can look at demographic data and make sure we're using it in a way that most helps for people. Now, I want to start easy with names. There's a great quote about names, which is that anything someone tells you is their name is by definition, an appropriate identifier for them.

[00:02:38] What that means is that if somebody tells you their name, you can use that that's their name. They'll respond to it. There's no official way of getting their name except legal names, which we'll get into later. So names can be anything pretty much. We already know that names can change [00:03:00] during marriages and divorces.

[00:03:02] They might change because of a change of religion to Islam, or when you're confirmed as catholic. Well, they might just change it. No set time. Like in Thailand where people change names to avert bad luck. We also probably know that names aren't unique in 2007 Beijing news reported that there were nearly 300,000 people named Zhang Wei, but names also aren't unique to the people who have them.

[00:03:26] Lady Gaga is known by lady Gaga by way more people than her real name, but it's probably notwhat's on her licence. So this leads us to the first of the diverse demographic principles. Don't use personal ids as system IDs.

[00:03:42] If you need to have a unique ID in a system, either using an external identifier that will never change, which is much harder than you think, or just make a random one up. And that way users will be able to change demographic details, like name, address, email, whatever, whenever they need to. So [00:04:00] those are the easy ones. Let's take a look at spelling.

[00:04:03] Japan uses multiple character systems. Here is one of them. Hiragana is syllabic which means that each of those characters can be read exactly as it's pronounced.

[00:04:13] Another character system is called Kanji. Kanji is logographic, meaning it's little pictures. That means specific sounds. For instance, the Kanji on the left means calm and the kanji on the right. It means Ji they called Kanji because Kan is the Japanese word for China. And these characters were imported from China.

[00:04:34] Here's an example of the difference. This is the verb to receive. Uketamawa is the pronunciation pronunciation of the Kanji on the left. And ru is the pronunciation of the Kanji on the right so far so good.

[00:04:49] Because the characters were imported from China. They had Chinese pronunciations. These are called theon yomi, well, the sound reading this characters on yomi. But because China used [00:05:00] to have a habit of dramatic civil war, the balance of power would change. And the dialect that was spoken prominently would as well, which meant that how characters were pronounced would change and Japan thought this was great and just imported all of the new sounds.

[00:05:16] So instead of having on yummy, we have on Yomis, so this character can also be pronounced as ka but it gets. See, the Japanese already had words for these characters and they weren't going to just throw out their whole language in order to use them. So instead characters also have a kun yomi meaning reading.

[00:05:39] This is a Morphographic reading. It means the character. Doesn't tell you how to say it, but instead tells you what it means. So for instance, this character means motto, but in some words, it doesn't see as it's used in different ways. And as it's used in different dialects of Japanese, it attaches to [00:06:00] different meanings.

[00:06:00] So it could be under or below or inadequate or difficult. These are made up examples, but here are some that aren't made up. This character can be pronounced as shita moto sageru kudasu oriru and several others. How do you tell them apart? Use context. What doesn't have any context? Names! This name is spelled that way, but pronounced either as tokairin or Shoji, do you know which one to use?

[00:06:34] No, you don't. You either have to guess or write down which variation it is. It works the other way to. See Hajime means beginning. There's a lot of kanji for beginning, which is why Hajime has nearly a dozen way to spell their names.

[00:06:51] So that's fun. Now, these characters, obviously aren't going to fit in ASCII, so you're going to have to use Unicode, but the [00:07:00] problem is the Unicode isn't good enough. You see unicode stores about 24,000 of the Kanji, but there are about 35,000 knji some of which are historic and aren't used anymore, but you'll never know which ones they are.

[00:07:15] There are also languages whose character sets don't fit into Unicode. Like the Aymara people of south America who have over a million speakers and have to transliterate everything through English, which is pretty common and a whole linguistic student rant that I'm not going to go into.

[00:07:30] What about data sanitation? Unfortunately, real names. Aren't sanitary, and it's very cultural specific. If you're in the U S and you tell one of your UK friends. Fannie and Willie

getting together with Randy. They'll probably start sniggering. If you use the name, Nick in France or Cara in an Arabic speaking country, you sound like you're swearing Gary in Japanese.

[00:07:53] Sounds like you have the runs. And Chloe sounds like toilet in German. Okay. Well [00:08:00] what about capitals? Unfortunately, capitals don't make sense either. There are no capital in Japanese and I don't think there are in Chinese either, but don't quote me on that because there are a lot languages. The problem is sometimes you do need capitals.

[00:08:15] For instance, the name's Mackenzie and Mackenzie, a very different to people named Mackenzie or Mackenzie. We can probably strip out symbols though. Of course not! Not symbols are totally valid in names. Let's look at Mackenzie or McKenzie, for example, what about. Well, numbers of valid. I think her majesty would be very annoyed if you weren't going to let her use her name, but maybe you're thinking, well, my system deals with business people, not royalty.

[00:08:46] So I'm never going to have to deal with numbers and names. Well, what about William Henry Gates? The third, the founder of Microsoft. All right. Arity and order now [00:09:00] Arity is the number of parameters. So I'm misusing it a bit to mean how many parts of a name there are, for instance, middle names. Hajime, doesn't have a middle name, most Japanese don't, but people from Asia may decide to give themselves a middle name in English.

[00:09:15] When in an English speaking country, in order to make it easier for the other English speaking people to use their name again, a name is anything that they tell you you can use. Well, what about the other extent? What's the maximum middle name length there is. Here's a perfectly valid name with a lot of middle names that I'm not going to try to pronounce because I'm terrible at Spanish.

[00:09:38] You probably know him better as Pablo Picasso.

[00:09:43] Well, what about family name by the way family name is preferable to last name because ordering of names can be different. Family name is fairly understandable across cultures. Getting back to it. Everybody only has one family namethat, right.

[00:09:58] Well, no [00:10:00] Picasso has two. Ruiz is his mother's family name and Picasso with his fathers. Can you pick one to use?? .

[00:10:07] Absolutely not. You don't know which of those family names is more valid. The best option is to use both of them or none, because as it turns out, family name is well,

[00:10:21] totally. You don't have to have a family name and you don't have to be some peasant from a village in the middle of nowhere, either. You could be, for instance,

[00:10:31] Suharno the former president of Indonesia. You might think that's a one-off, but you'd be wrong. According to Sukarno also former president, I mean, this leads to our second

diverse demographic principle. Be as permissive as you can use the most amount of data that you can get your hands on in order to accommodate the most number of people.

[00:10:55] Now your marketing team might say, but we need all of this detail so we can [00:11:00] send nicely worded letters and ask people gendered buy our stuff. Oh, and if we don't know how to address people properly, how will we categorize them in our giant database full of information that we probably shouldn't have about people?

[00:11:11] Well, I think that that is in general, an absolutely terrible reason. People are most comfortable with companies when companies treat them the way they want to be treated. And that includes using their name in the way they want it used, not telling them that it's invalid. Gendering names is particularly fraught. Most countries don't restrict what gender name you can give a child. And then there are any sex names like Casey or Ashley or Dylan.

[00:11:39] You might have a legal need to store names, but you can still be smart about it. What you should do is ask the user's preferences and then test those instead.

[00:11:48] Ask them for their legal name and then ask them what you should call them and then never use their legal name for anything except things where you're legally required to. Cool. We have names [00:12:00] down. What else is horrifying and confusing?

[00:12:02] Addresses. Addresses change dramatically between countries. Let me show you an example from Japan.

[00:12:11] Here's a Japanese address. It's probably a bit confusing and it being in Kanji definitely won't help. So, I'll fix that last part.

[00:12:21] No, it's still confusing. See let's break it down line by line. First, we start with a postcode and then a prefecture. The prefecture name ends in to for Tokyo do for Hokkaido, foo for Kyoto and Osaka and Ken for all the others. For reasons! Next you have the municipality.

[00:12:40] This will either be a , city, which ends in Xi or a county which ends in gun. For one of the smaller prefectures big cities divide straightaway into wards or ku , so Chiyoda-ku here, means the Chiyoda ward of Tokyo. Small cities then divide into Cho or machi which means town or [00:13:00] village, big cities can optionally divide into these, but they don't have to add cho or machi to the end.

[00:13:06] Here in Tokyo, our town name is Maranouchi . That then divides into the block or chome, which are numbered here, we're at ni-chome 2nd block. So here's the thing addresses can be inconsistent within countries. Let's look at Kyoto. Here's an address in Kyoto prefecture Otokuni county.

[00:13:29] It's in Ooyamazaki-cho. Ooyamazaki town in the Shimoueno district, but instead of dividing into chome some parts of Kyoto divide into aza Kitahosoike-aza and then optionally

into smaller child, aza called ko-aza like one aza here. The Daihatsu plant is big enough that it takes up the entire area. So it doesn't divide further.

[00:13:51] Oh. And some cities do different things. Again.Ibaraki, for instance, jumped straight from Ibaraki to the chome level So [00:14:00] we know how to get to what block we want. Now we want to get to a specific building. We start by getting either the city block number, the new style, or the lot number from the land registry. And if it was ever subdivided, it then needs to have the subdivision number, and then we have the building number and optionally the building name in this case, the Tokyo Chuo Yubin-Kyoku, the central post office. So, here's a handy little cheat sheet. You have the prefecture, to/do/fu/ken the city county shi/gun. The ward ku the town village machi slash cho, the district chome/aza, the ko-az or block and sub block, the building or building name, but you don't use a sub block if you aren't using a block and you don't use a ko-aza, if you're not using an aza and aza and ko-aza tend to only be used in Kyoto and Ibaraki doesn't bother with the ward the machi the cho at all.

[00:14:54] Sorry, I got a little carried away. Oh, also, if you don't have a lowlotnumber, it's [00:15:00] just called , which means unregistered land. And you go straight to the building number. And I guess hope?

[00:15:07] Also, when people are telling you this, they might compress everything from the block level down.

[00:15:12] So instead of being two chome seven ban two go, they'll just say seven dash two dash seven, all use the possessive particle and use two, no seven, no, two. Oh, say it in Japanese ni no, nana, no ni you might be expecting some more traps now. You're right. See, buildings can be numbered in construction order.

[00:15:36] And since Japan can have 400 year old shops next to one year old apartments, that means block numbers are completely insane. Also, blocks are irregular. They're logical, not. You might have a block that's actually five city blocks long and shaped like a Tetris piece because it's all part of one shopping street.

[00:15:56] And so that all goes together. Oh, and in Niigata instead of [00:16:00] using divisions, based on city blocks, they use traditional divisions numbered with a system based on the Chinese heavenly stems.

[00:16:07] I have no idea what that means. I didn't understand. I moved on. I hope I just never need to send mail to Niigata. Another thing you might've noticed is that street names aren't required in addresses, which is very convenient because street names aren't required in Japan. Major streets might have names, but they don't have to, except in Kyoto.

[00:16:31] See Kyoto has a bunch of cho with the same name in the same. Which is confusing.

[00:16:36] Since Kyoto is on a grid, they use a citizens' addressing system.

[00:16:39] Here's how it works. This is the address of Kyoto Tower. One address for it is Karasuma, shichijio sagaru. It's on Karasuma street. Find the shichijio intersection and go south. But since Karasuma intersects with other streets, you can also give the building the address, Karasuma, [00:17:00] Shiokoji agaru. Find the Shiokoj iintersection and go north.But they might want to be very clear. So they might say Karasuma dori, shichijio sagaru. Where dori specifically indicates that Karasuma is the street and it's the street the building is on. Which means that buildings can have multiple addresses, including their official address.

[00:17:23] Oh. And the government and the tax department and the postal service will all use them as valid addresses. Which leads us to diverse demographic principle four: validity is local. You can't validate a piece of data unless it's local to your jurisdiction.

[00:17:39] If the jurisdiction the data belongs to says it's valid, it is. It's incredibly arrogant to say, well, our system, isn't going to accept your addresses because they are wrong. Who says they're wrong. If an Australian system says to a Japanese. Your address is silly and we don't accept it. We're basically saying you don't live at a proper place, which is a [00:18:00] weird and unpleasant thing for your system to say to somebody. But, it turns out that addresses are a mess. Worldwide See, you know, building numbers and street numbers.

[00:18:12] They have no constraints, none whatsoever.

[00:18:15] See, buildings don't have to have numbers like parliament house in Canberra. Numbers can be names like at plan 1944 in the Netherlands and numbers don't have to be odd or even on any specific side, like the Boulevard, Theophile Sueur in Montreal. Sorry about the pronunciation. Numbers aren't necessarily unique to streets; 50 Ammanford road in Tycroes and Llandybie, are both in the same area.

[00:18:39] And not very far from each other. Numbers don't have to be positive integers. Zero Edgemont road is a real address. Minus one. Priory road is a real address that has to spell out their number because no system they've encountered will allow it. And my university uses numbers that are negative to indicate a building is below ground level, according to the entrance of the campus.[00:19:00] Buildings aren't necessarily either numbered, or lettered for instance, there's a building in Florida that has both a 39 and alpha 39.

[00:19:08] Building numbers can also be names. There's a different building at T E N post office square from the numeral 10 post office square in Boston. There are fractional street numbers, like 43rd and a half street in Pittsburgh, which can also be spelled 43 RD, one slash two or 43 dot five or 43 dot five RD.

[00:19:28] You can't necessarily ignore leading zeros, apartment 0, 0 1 and apartment one at one-on-one Elma street in Plato Alto in California, are different buildings and streets themselves have no constraints. The type of street isn't always last french streets have it first the type of street isn't always first or last 14, Brisbane technology park is a street.

[00:19:50] Street doesn't necessarily mean street, especially in the town of Street in Somerset, in the UK. Street names aren't unique. There are 16 streets [00:20:00] containing the name high street in London. Seven of them are literally just called high street. And if they don't have the same name, they're not even necessarily in different areas.

[00:20:10] Bosch Holt in Germany has five streets called up dwell in the same area. They even all connect with each other in different weird ways that make it clear; it's not just one street.

[00:20:21] Oh, and here's a bunch we can disprove at once. Streets have a single name. They have a maximum length. They're only present in a single city. They're only present in a single state.

[00:20:31] See, the same street can be present in many places.

[00:20:35] Meet the M1! It looks like this it's Australia's national highway and it goes the entire way around the country. Some of its names are: the Pacific highway, the Princes Highway, the Bruce Highway, the Monash Freeway, the Brand Highway, the Victoria Highway, the Stuart Highway, the Carpentaria highway, the Savanna Way, the Kennedy Highway, the Pacific Highway, the Warringah Freeway [00:21:00] and othres, but they're all the M1.

[00:21:02] Also streets that go to many places. Oh, that should say aren't they aren't contiguous. Here's an example. Meet the M1. Again! Because it also looks like this. This is Tasmania. It's also part of Australia.

[00:21:17] In Tasmania. The M1 is also called the Brooker Highway, the Midland Highway and the Bass Highway. Towns can also be present in many places. In fact, their name can even be nested. For instance, there's a city called Singapore.

[00:21:31] It's in a state called Singapore in the country of Singapore. In particular, Springfield needs to calm down. There are 86 Springfields worldwide, as far as that. 7 in Australia WA is the only state that doesn't have one. There are 18 Canada 2 in Ireland and 14 in the UK. The US is winning because there's 52 there, which is at least one per state. Some states have three.

[00:21:56] So this leads us to the final diverse demographic principle: [00:22:00] rely on expert help. Your team has no way of getting this right by themselves. So don't try! Use an open source library or a vendor. And then instead of testing that the library of a vendor works properly, which is frankly their job, just make sure that you're calling out to it correctly and handling the results.

[00:22:16] Experts are there to make these things better. Open source systems in particular can accept help from residents around the world. Making sure that systems are robust and support their own use cases.

[00:22:26] The diverse demographic principles are super helpful for making sure that your system and your tests are as friendly, welcoming, and open as possible.

[00:22:35] One last piece of Japanese. Arigato Gozaimasu. Thank you very much. I've had a blast putting this talk together. If you liked it, please follow me over on Twitter and you can download resources and the slides at bit dot lee slash names and other lies.

[00:22:51] Thanks very much.