Translation and Transliteration of Contact Information PDP Working Group Meeting
TRANSCRIPTION
Thursday 12 June at 1300 UTC

Note: The following is the output of transcribing from an audio recording of Translation and transliteration of Contact Information DT on the Thursday 12 June 2014 at 1300 UTC. Although the transcription is largely accurate, in some cases it is incomplete or inaccurate due to inaudible passages or transcription errors. It is posted as an aid to understanding the proceedings at the meeting, but should not be treated as an authoritative record.
The audio is also available at: http://audio.icann.org/gnso/gnso-transliteration-contact-20140612-en.mp3

Attendees:
Pitinan Kooarmornpatana - GAC
Petter Rindforth – IPC
Jennifer Chung - RySG
Chris Dillon – NCSG
Rudi Vansnick – NPOC
Ubolthip Sethakaset – Individual
Peter Dernbach – IPC
Jim Galvin – SSAC
Mae Suchayapim Siriwat – GAC
Wanawit Ahkuputra – GAC
Peter Green (Zhang Zuan) - NCUC
Marc Blanchet – guest speaker
Percy Ephraim Kenyanito – NCUC

Apologies:
Justine Chew

ICANN staff:
Julie Hedlund
Amy Bivins
Lars Hoffmann
Nathalie Peregrine

Coordinator: And your call is now being recorded.

Nathalie Peregrine: Thank you very much (Indrid). Good morning, good afternoon, good evening everybody and welcome to the Translation and Transliteration PDP Working Group call on 12 June 2014.

On the call today we have Marc Blanchet, Peter Green, Petter Rindforth, Rudi Vansnick, Wanawit Ahkuputra, Chris Dillon, (Hubortit Dakaset), Pitinan Kooarmornpatana, Jennifer Chung, Peter Dernbach and May Suchayapim Siriwat.

We have received no apologies for today's call. And from staff we have Julie Hedlund, Lars Hoffman and myself, Nathalie Peregrine.

I'd like to remind you all to please state your names before speaking for transcription purposes. Thank you very much and over to you, Chris.

Chris Dillon: Thank you very much indeed. And let's start with the compulsory Number 3 on the agenda and that's just the statements of interest. I have to ask whether there has been any change in statements of interest since we last met?

Seeing and hearing nothing that means we can continue. And that takes us into responses from the SOs and ACs. We haven't had any new responses recently and so I think we can leave that.

Now later on in the call there will be a briefing on the study to evaluate available solutions for the submission and display of internationalized contact data but that won't be I think until half past or somewhere around than. So we might as well just work through the other parts of the agenda before they join us.

And so that brings us to the - we've got that work plan on that - oh yes okay that's really good to have that on the screen. And I think, you know, we've got a little bit of time to think about this and really with two major questions I would say.

So one of them is, I mean, my general feeling is that we have slipped, you know, that the timeline has slipped slightly but I think we need to be saying is there's still realistic? So are we likely to finish in December this year? I think that's certainly one question I think we should consider. I wonder if anybody has any comments on that?

Seeing no comments I would like to suggest that although we perhaps have had some delay that we should still try and keep to that deadline. Now specifically one of the issues is that although we wanted to receive comments from the Registrar Stakeholder Group they haven't actually got back to us. And certainly I am, I mean, I am concerned that that could cause a delay at some point. Petter, would you like to got something there?

Petter Rindforth: Yeah, sorry, I just checking out the agenda during the summer of July and August and just a suggestion if it's possible I see that there is a lot of meetings noted as continue review of comments received. Maybe we could at least shorten that one or two meetings to get on time again and try to be more efficient on the meetings when we review the comments.

And perhaps even if we have weekly meetings, and I should not say that's because I haven't been so active in between the meetings but it may be also good to try to supply comments and discussions online by email or so between the meetings so that we can prepare or summarize when we are on the phone calls.

Chris Dillon: Thank you very much for that. That's a very, you know, that's more or less exactly what I feel about this. So, you know, there are certainly quite a few meetings there. Historically we've canceled the audit meeting now and then so to some extent I think that may continue but that's really why - one of the reasons why I'm thinking that although there has, you know, perhaps there is a small delay now there may be additional delay conceivably. But I think we probably still are on schedule. Let us hope so.

Now, leaving that aside for a moment and, the other big question is what people sometimes refer to as the dog that didn't bark. So I'm sort of - I mean I've already mentioned that we've got one stakeholder group that ideally we would have liked to have had comments from.

But I'm sort of just thinking are the things we're missing here - is this actually complete and that's - I think that's quite an important thing because it's quite easy to talk about the things that are raised in specific documents but every now and then it's just quite good to stand back and think is there anything we are missing here. I don't know whether anybody spotted anything but I thought I would like to, you know, just you bring that up and actually not just about the work plan, that's quite a general issue.

Well okay, seeing nothing and I think perhaps let me encourage you to click that generally and as I say not just about the work plan. And let me just ask whether there are any other comments about it or I think this is almost certainly something we will actively visit during the London meeting. But just before we let it go is there any other aspect that anybody would like to pick up?

Okay, let's leave this one until London. And as I say I'm pretty sure that we will revisit it at length during that meeting. Okay so that then brings us down into any other business in fact because we're going to come back to the responses at the end. So any other business is going to be the - well there are also some actions and we can work through those but any other business.

At some point we need to talk about, yeah, Julie is actually saying in the chat, and the other thing we do need to talk about is the slide update so we'll, if it's possible to load those that would be really good.

But I think perhaps we also need to talk about London - the London meeting generally. I'll just put that on the table for a moment. Whether there's any, I mean, obviously there's been some discussion about some aspects of the London meeting but it's just whether anybody wants to raise anything that we haven't already spoken about.

Okay well in that case I will just - yeah, if you have the slides loaded up. It's a horribly long URL back I will paste it into the chat room now. Okay, I hope that works. Yeah, yes, that is fine. Just wait for a moment so that you can open that either in another tab in your browser or I think sometimes preferably in another browser altogether because then it just makes it easy - easier to look at the chat room and to look at the presentation at the same time.

Okay, I'm fairly sure I have spotted something in this which I didn't spot when I first read it. But I'm just wondering what the easiest way of doing this is going to be. So, well, yeah, I'll talk through the slides; I think that's probably easiest.

So we have a slide on why, you know, why is that PDP important. So this is the continued internationalization of the Domain Name System. Need to allow for standardized query as internationalized registration data; ongoing reforms of the gTLD directory services makes the need to establish GNSO policy.

Now what I will say this when giving this slide, when I get it to public audiences, so not to technical audiences but to a public audience I would be also stressing that, you know, a lot of information which is currently available either in English or in ASCII, may not be available in the future. So I will stress that actually as I - as I do this slide.

Then if we have a look at the next one, we've got just something on the history of the working group and then we've, you know, the input request.

Then feasibility study on transformation of contact information. And then, you know, the fact that we are systematically monitoring other efforts in this area, you know, that we know about.

Now this is the date which I think got forgotten so I meant to stress that we keep everything in our wiki because we do, I mean, all the documents are there, you know, everything we've created, everything we've received it's systematically there so I think that's an additional point that's worth making if not on this slide but on one of the slides.

Then if we perhaps go down into next steps, we've got the - okay the fact that we are reviewing the feedback using the tool and that we're providing updates which is actually what we will be doing or what I will be doing as, you know, a psychic this presentation.

And the Julie is asking in the chat could we add it on next steps? Yes, and certainly include the link to the wiki because the reason I stress it is that I think some groups are possibly not as systematic as we've been and so people just looking at what the groups are doing feel quite overwhelmed because they feel that they have to really go into a lot of detail where we have systematically, you know, go through calls, go through transcripts, heaven knows what; we're as we have been very very systematic and, you know, as far as I know there isn't anything that isn't in there.

So actually it does mean that we are more welcoming for people just, you know, who haven't been involved so far and are just interested in what we're doing but yes, I think that would work nicely. And we've been did a sort of rather colorful slide I think and that is then - and then what follows is the rather more detailed information. And we've got the repetition of the main charter questions, links to the charter.

Oh, here's the wiki. Yeah, in fact I'd just forgotten that there was a link here. Maybe we just get away if I stress absolutely everything is in there, maybe that's all we need to do. Okay.

And then we've got the issue report there. Then we've got background. Yeah, okay, which is actually quite technical, request and content return by DNRDS like Whois, coded using ASCII. No standards for Whois implementations to support other character sets, yeah, okay this is really what I was thinking.

And it is anticipated that a new domain name registration data service may not be ASCII-centric. Yeah, okay. Some of this I was saying earlier. Then we've got definitions of translation and transliteration. Slightly wondering whether we should put something on transcription on this slide.

That could be another one because I think in practice it's, you know, it is a word that's used but, you know, it is phonetic based so that might be - that might be a question that we could ask later on actually about, you know, to what extent transcription could play a role. I think we've largely talked about translation and transliteration.

Oh yeah, okay, Julie is making the point that because we are quoting that issues report we can't really do that which is probably not a big problem but it's still a good question.

Okay, and then more stuff on the background here which is the definition of contact information. Yeah, and Julie is just typing that we just need to be prepared for that if it comes up. So, yeah, take a few notes during the next presentation and what contact information includes.

And then specification of data elements by the RAA but no necessity for those to be transformed. Oh, and that's actually the end of the presentation. Okay. Right, again, any dogs that didn't bark, is there anything that should be in that presentation that we haven't put there?

I think I really honestly can't think of anything but I think we're very close to a final version there. And, you know, it is the case also that, you know, I think sometimes when giving a presentation it's actually quite good to hold some small thing back because otherwise you've got to be able to talk about it.

I was very keen to give a sort of a general background being because one of the presentations I gave in Singapore I really had the impression that there was an audience that didn't really understand the presentation because there wasn't background information but I think we've fixed that problem now.

Jim, would you like to say something about that?

Jim Galvin: Thank you. Just in the category of yes - this is Jim Galvin for the transcript. In the category of whether there's something else to add to the presentation or not I'm wondering if since we mentioned the gTLD directory services group is there, you know, some value in pointing out that there's additional work going on in the Internationalized Registration Data Expert Working Group too? And just at least indicating that, you know, you're where of that work and what's going on.

Chris Dillon: Yes, that's a good improvement. I mean, you know, you hope anybody requiring further information we'll go to the wiki but, yes, you're right, I mean, it's good to put the major things in here. But, yes, yeah I think that's a good improvement.

Okay, now so if we just - okay yes so that's - yeah, so Julie is just replying to that in the chat so that should be fine. Okay, anything else about this? Jim, I think you've probably got your hand up but if you would like to say something that's also fine. Yeah, okay.

Right, now, what else have we got on here. I think - now one being we could do but I'm just wondering whether we've got enough time for this is that I know that Jim has some questions about some of the parts of the review tool. I don't know whether, Jim, you would be interested in asking those now.

I mean the only slight issue is that we might need to get the review tool up on the screen to be able to do that. But, Jim, would you like to do that now?

Jim Galvin: I have to be honest, I confess I'm really not sure. I don't remember what questions I had or...

((Crosstalk))

Chris Dillon: Oh well, the good thing is that I like an elephant on things like this. I think it's Number 39. So if we go all the way down there I think the Number 39 and it's actually something from the Thai GAC representative but I seem to remember - I have a note that you weren't very sure about it looking at that...

Jim Galvin: Yes. Thank you. So I remembered. The question that I was asking at the time was more in the context of sometimes when you get a comment back you understand the comment because, you know, you know the community and so you kind of expected them to say, you know, what it was that they said.

And I was a little bit struck by the fact that I didn't at least expect, you know, the Thai GAC representative to say what they said here. And my question was really, do we have any additional context to go with their position? and, you know, is there - what else is in the background here that would help us to understand why it is that they have this particular position?

Chris Dillon: Thank you very much. I don't know - I think we've got at least one Thai representative on the call, Pitinan, and somebody else who has a name which might be Thai. I wonder if either - now Pitinan is writing something so let's just wait a moment. Okay yes, Pitinan is definitely doing something in the chat.

Pitinan Kooarmornpatana: Hello?

Chris Dillon: Oh.

Pitinan Kooarmornpatana: Hi.

Chris Dillon: Hello, Pitinan. Yes, we can hear you.

Pitinan Kooarmornpatana: Hi. I'm Pitinan for the transcript. Sorry, Jim, could you probably ask again specifically what's make you wonder about their position.

Jim Galvin: So, yes. This is Jim Galvin for the transcript. I was struck by the suggestion from the Thai representative that the registrant should bear the cost of converting to a common language. And that stood out for me because at least so far, you know, conventional preference has been in the direction of that the registrant should only have to, you know, work within the language that they are most skilled at and that they know and have.

And, you know, Thailand is also, you know, one of those countries with a language which is pretty far removed from US ASCII as compared to other languages. And so it struck me that they would choose that kind of suggestion. And I just wondered if there was some additional information that might be helpful in this discussion to understanding why it is they would suggest something like that. I mean...

Pitinan Kooarmornpatana: Right, right.

((Crosstalk))

Pitinan Kooarmornpatana: Okay.

Jim Galvin: Thank you.

Pitinan Kooarmornpatana: Thanks, Jim. I would say this position is - the background of this is, given that they should have the common, you know, the local (unintelligible) to validate and do the (translation) of the contact information which is - that is one of the slide in the wiki that we propose in the (unintelligible) earlier - a few months ago that we see that is the government role to facilitate this kind of infrastructure to happen in each country.

The idea is given that we have the single point to validate the contact address (unintelligible) validation and transform to from Thai to English then it kind of makes sense that the registrar would be the one to do the validation, and to registrant would be the one who bear the cost for translation which should happen only once.

This will make much more sense if there is one - not one but there is the accredit contact information (unintelligible) for each country. Does that answer your question?

Jim Galvin: So thank you, that's helpful. It answers part of the questions that I'm asking. Let me make sure I understand what you are saying. You started with the premise that if there is to be a common language then you were going down the path that the registrant should have the burden of doing the translation or transliteration of the contact information.

So let me expand my question a bit. First let me ask if did I summarize that correctly?

Pitinan Kooarmornpatana: Yes.

Jim Galvin: Okay.

((Crosstalk))

Pitinan Kooarmornpatana: But if this will - given that you have to do only once. Right, because you don't change your name that often so the first time you register to any domain name registrar you already done the translation. And that English information will be held in - well we will call it contact information, this for Thailand we are setting it up actually.

So this contact information system is the place - is the center that you can register your English name, English information and let the registrar or any other entities to come and check. So for us the registrant will translate only once.

Jim Galvin: So this is essentially supporting rather strongly the proposal from the Directory Services Expert Working Group. They have a rather complete validation ecosystem, if you will, that they are proposing which in particular conceptually has, you know, a database of validated contact information and whatever other elements need to be there.

So if I understand correctly you're suggesting that, assuming that such a system is to come into place, then this kind of translation and transliteration only needs to happen once. And if it's in this convenient database...

Pitinan Kooarmornpatana: Right.

Jim Galvin: ...then, you know, that one time cost is most conveniently or most practically something that the registrant should be responsible for. Correct?

Pitinan Kooarmornpatana: Right, right. Correct. Correct.

Jim Galvin: Okay thank you. That helps.

Pitinan Kooarmornpatana: Okay thank you.

Chris Dillon: Thank you very much indeed. Jim, just before we move on was that the only question you have about the document or was there anything else?

Jim Galvin: Well, thank you for pointing out to me that it was Item 39 because then I was reminded immediately of the question. And, no, nothing else jumps out at me here that I remember as a question that I had so hopefully that's it. If it comes back to me I'll jump in. Thank you.

Chris Dillon: Thank you very much indeed. I've been remembering the Number 39 for all of meetings and it's good to get that sorted out. And thank you very much Pitinan as well.

Okay, well we're now going to go back to Item 4. And we have the presentation on the studies so we're all very much looking forward to that.

Julie Hedlund: And, Chris, this is Julie. I don't know if Steve Chan has been able to join, he was having some network difficulties. And then comment Steve, are you on the call? Nathalie, do you know if he's joined?

Then Steve has asked if Marc Blanchet could do the slides. And I'd be happy to run them for you, Mark. Would you be able to do that, give the presentation?

Marc Blanchet: Sure. Do you hear me?

Julie Hedlund: Hear you loud and clear.

Chris Dillon: Indeed. Thank you very much.

Marc Blanchet: So shall I start?

Chris Dillon: Yes, please.

Marc Blanchet: Okay. Thank you. Maybe go back to (unintelligible)? So we were commissioned to study the available solutions and to evaluate be available solutions for submission and displaying of internationalized contact the data. The study team can be reached out the email address there. The study team comprises of a few of my colleagues. And we were more on the technical side of the study. And Sarmad Hussein was most on the linguistic part of the study.

Next slide. Yeah so it's internationalized registration data has been, you know, discussed in various places. The study we were trying to do was to document some current practices and transformation possibilities that could be used for the various working groups within ICANN such as yours.

We look at the practices of handling the registration data. We did a survey of registries and registrars. We looked at various electronic merchant services, the protocols, specifications to see if there is missing pieces.

And we actually exercised tools - some tools that are available both commercial and - either commercial as a shrink-wrapped software or commercial of proprietary as an EPI or public domain. And we actually created many test cases in different languages and exercised those tools to see what's the actual output and provide the accuracy of the outlet.

Next slide. I want to make sure - okay. So...

Woman: Sorry about that. I don't know why this is acting so strange. You did say next slide, is that correct?

Marc Blanchet: Yes. I'm at Slide 3.

Woman: Excellent.

Marc Blanchet: As you know there is roughly two categories of data. One is the contact data and one is transactional. We obviously look at the contact data which is more related to the local language instead of the user registrant is using, so a person, name, address, city, state, country.

Next slide, Slide 4. I think I'll skip this because everybody is talking about this so we in the report that has been sent for public review like a week ago I would suggest that you look at it. We provided some (unintelligible) and sometimes we talk about transformation in the sense of any kind of transformation, either translation, transliteration or transcription of any kind. That's where we use a job and transformation, where in some other cases we specifically say something about one method, either translation or transliteration for example.

Next slide. So we kind of tried to define a level of transformation that actually, you know, is within the study, kind of background information used because obviously depending on your needs, your being the person who is looking for the data, then you may require, you know, some kind of accurate transformation which means, you know, it really needs to be valid, you know, legally and all that stuff, matching to legal environments. And consistent meaning that it will always be the same way. While it may not be accurate, it is consistent transformation and that is more like, you know, the basic transformation I would say.

Next slide, Slide 6. There is ways that you could I'm not sure how to say it in English but convert or divert, or I don't know, pertaining to restoration from languages, all languages. That has also been looked at. So for example from French to Thai or from Chinese to Arabic, and United Nations we looked at some studies of United Nations and the - what they suggested or recommended was to use manuscript as fiber for - to go from one language to another. However, the study found that it's, you know, on theory, on paper it looks just fine but in practice it's very, very difficult to get this function really working well.

Next slide, Slide 7. So we surveyed a few e-merchants. Obviously very large, we spent a whole year surveying these, but can you look - move to Slide 8 please? Yes. So the merchants that we included were in different languages and scripts, sometimes global, sometimes more local, but roughly the yellow or the data to be put into local languages are split. We usually verify the type of data in a very limited extent.

They essentially just accept the user input, you know. The problem is for the user to make sure that he puts the right information. That entered them in for merchants as opposed to our active markets where they don't support dominant script or language use, for example, you know, e-merchant, U.S. e-merchant that is selling to Arabic people, you know, support Arabic entry of data while, you know, not their primary market.

So next slide, 9. We also did a survey of registries and registrars, a separate survey. So we were trying to see what's the current practice in the field. We got obviously, you know, 12 registries responded, so not a bad sample but obviously not all the registries. So the registrars were actually - the survey was actually responded by only two registrars. Since then we made some significant effort to get this, you know, get all the other contacts and people to respond but.

So the conclusion should be, you know, there is good data there but we need to be careful in, you know, assuming what happens in the overall market. So large registries -- just a minute -- registries, large gTLDs and ccTLDs covering multiple languages and scripts to Arabic and Cyrillic, Japanese, German, French and English, so that is actually a fairly good sample. The registrar survey only one very large registrar, very well known, so this by itself is actually very good. The other one is a more a local one. So, you know, that gives us some good idea but obviously not a full blown statistical survey.

Next slide, 10. So - and you have almost - in the report you have all the survey questions kind of summarized but you have a lot of details in the survey. So, you know, please look at it for additional information. So essentially one of the key results I think it could be said is that none of all the registrars and registries that we got information from them actually transformed the contact information. So whenever there's multiple language data collected then it's always provided by the registrant. So I think that's the main thing. And the other thing is none actually verified any data that they received.

Next slide, 11. We surveyed relevant protocols, obviously as you know is we're supporting SDN. EPP supports UTFA and coding for transmitting and receiving data obviously but without a language specification and doesn't have the - we're talking here about the standard track RFCs and does not record mutable linguistic versions of the same data. (Unintelligible) however can encode language information and can handle multiple versions in parallel.

So if we're talking about EPP here then if people are actually - so later in the report at the end of the findings if people want to look into going into dramatic transformation of contact data then one of the two augments the accuracy of the transformation the more language - actually the strings are tied with language details and there's more than just a language tag. The more there's data about the string itself, the origin of the strength, the better chances for better transformation. Therefore this has impact on the whole chain of getting the data from the user up to, you know, to registry store and then to the display through (unintelligible) or other means. So keep this in mind.

Next slide, 12. So I'm not sure - okay yes. So okay thanks. So this slide, transformation, actually talks about the testing we did. So we did -- and you'll see in the next slide the details -- but roughly we have created test data including individual and names, including family given names, organization names, addresses, abbreviations, you know, all kind of stuff to prevent - state names, country names and for short forms in multiple languages and scripts such as (unintelligible), Arabic and in some cases multiple languages in the script, and Cyrillic.

Next slide, 13. So this is the testing data that we took and they were provided by either - they were all provided by people that know well the language, are either language experts or, you know, native speakers. And one of the things that we did was actually after transformation of exercising the different tools that we used then the same person provided the test data, actually verified by the size, the results of the transformation transformed data and then was able to qualify if the output was accurate or not and the level of accuracy of the output. So for example therefore the whole process was handled by linguistic persons from the source to the transformed data and the matching between the two.

So next slide. So, you know, you could see that from the previous slide that, you know, there's a good set of data but again it's certainly not a full blown, you know, linguistic study I would say. However - and the interesting part also is that the focus from the registration data we didn’t try to, you know, look at, you know, generic transformation of, you know, languages and scripts and texts in general, which has pros and cons because there's for addresses and names and stuff, you don't have a lot of context so it's sometimes more difficult to translate or transliterate.

We provided a few measures too that I would like to talk about and you will see the results later, is accuracy which is essentially a binary measure of if the transform data compared to the, you know, what we call goal reference which is essentially the right way to transform it from a linguistic perspective. So when the tool provided the exact same transformed data as the linguistic expert would provide then it's 100%, you know.

The thing that we did was actually that the expert, the linguistic expert, that was looking at the transformed data was still able - the exact match was not only - was not necessarily on the code point level but on the linguistic level. Therefore two variations of the same - two valid variations of the same transformed data would have 100% accuracy. So that's the accuracy measure.

The second is Levenshtein distance. So this is the number - it's a well know measure in the linguistic environment for transformation of linguistic data. Levenshtein distance is a non-binary and it's the number of edits, insertion deletion and substitution between two strings.

So here's an example in Russian where you see that for example if the transformation - the appropriate transformation of the word in Cyrillic Russian and transliteration is V-E-L-Y-O-V; however, though this is the accurate transformation by a linguistic expert but the transformation tool will provide V-I-E-L-'-O-V, then the Levenshtein distance is two, meaning that there's two changes which is delete the I and substitute the apostrophe with a Y to go back to the gold reference, the right transformed data. So for this transformation, the result is a Levenshtein distance of two.

So the lower the Levenshtein distance is the better it is because it has less number of transformation. However on the accuracy side, the higher it is, the better it is. So the maximum distance of Levenshtein is obviously the length of the longer string. So the best is zero.

So in the study because the words have a different number of characters, you cannot use the actual value of the Levenshtein distance because it actually depends on the length of the word. Therefore we normalize from zero to 100%. One hundred percent in the study - 100% Levenshtein distance means that the whole string is completely different, right. Whatever the length of the string, 100% means, you know, it's all different. There's no correlation between the expected transformed data and the string that you get from the tool. If you think about 50% of Levenshtein distance that means that 50%

normalized of this - 50% of the string has been changed. It is changed from the expected output.

Slide 15. So this is a list I think there's a few slides of various translation tools that we looked at. I think 15 - go to Slide 16. And so either general or specific transliteration or transformation tools so - and obviously there are more. So we contacted most of them trying to get access to their software or their API by some means. We - some of them offered us to actually submit our data to them and they will provide the answer and then we refused because we were not able to control exactly what was happening, not that we didn’t trust them but - so we only took the tools that were able to emphasize by ourselves and look at the results. For the purpose of, you know, careful, you know, legal or other means we didn't disclose the, in the study, the tools that we used because it may have some implications. So they are essentially anonymized.

Next slide, 18. So here - so you will see a few slides of transformation that was a summary of transformation. You have all the details in the study. So the first one on Page 18 is the Levenshtein distance normalizing percentage across all tools that we tested. So for example if you look at the first row that says for name, so for the name category for at hand, the average of all the tools that we exercised had a Levenshtein distance of 26%, which means that roughly a quarter, one every four characters is wrong or changed from the expected output. And then you have all the numbers for all the different languages and all the categories.

If you look at the last column, which is the average across tools and across languages, then you see that it's, you know, between 29 to 50%. So 50% meaning that half of this string is essentially wrong from the transformation of the data compared to the expected transformation made by the linguistic expert.

And remember that the linguistic expert actually reviewed with the linguistic guys of the output of this transformed data by the tools, which means the tool actually provided a different but okay transformed data then the Levenshtein data was zero, meaning that there was some judgment from and leeway from the linguistic expert to make sure that the score was not bad just because there was a variation in, you know, one way to encode a character but was also valid. So if we were doing really straight comparisons it would be worse, but so this is giving, you know, the best chance for the tools to provide good results.

Next slide, Page 19. So this is an example - okay across languages, so here's some example of the type of - so this kind of show that depending on the type of tool, the type of transformation, sorry, then you may have better or worse results. And it actually depends also on the type of data, input data. If it's a name or country name or address, the results are different if it's transliterated or translated. So there is a lot of variations depending on the actual, you know, the set, the matrix of considerations.

So either the category of the data, the method, if it's transliteration or translation, the type of language, because depending on the language or script, there are scripts that are really are transformed badly more, for example transliteration of Arabic is not because the language itself has not that many vowels, and therefore when translated to Latin it actually loses a lot of things. We actually have a colleague who is an Arabic native and he was not able - that person was not able to read the transformed data. It was nonsense to that person by all the tools just because all the metrics that was used by the transformation of the data, transliteration of the data by the tools were all, you know, make no sense for the actual user.

So back to the results, so accuracy, you know -- sorry, Page 19 -- so accuracy of - so remember that accuracy is kind of binary measure if it actually provides, you know, one-to-one relationship. As you can see for all kind of contact information for all across languages, you see a very different accuracy average of accuracy which is around 60% for translation compared to 16% which is very low. Remember accuracy, you know, 100% accuracy is the best.

So as you see overall, translation is more accurate, you know, typically. However the lever - well in other results is the average of Levenshtein distance on the transliteration part in all languages and all contact information altogether average transliteration gives 52% average of Levenshtein distance, which means that on average for all categories, all languages, there is half of the strings that are - essentially half of the string is wrong. And in translation it's still almost 40%.

Next slide. Obviously this is a very summarized table. A good part of summary is you get a feeling; however, you will see in the report that depending on as I said the category of data or the language or the method, you will see a large variation of percentages. So that means that you may be more likely using not I would say lucky, you could have better results with one set of conditions and very bad results with another set of conditions.

Some findings of our study. The following information is needed for transformation: the current language and script method of obtaining trend data. So back to what I said before which is the more information you have from where the source comes from then the likelihood to have better transformed data. And moreover, for the transformed data, especially if you want to do reversibility, which we haven't talked too much but you have a lot of data on reversibility in the study, reversibility meaning that you want to come back to the original string. So you transform to or you transliterate for example to a Latin form then you go back to the source string in the native script, then if the reversibility is - there's some data and stuff about this but don't skip.

So reversibility when you want to do the reverse side of the transformation then however the transform data to be reversed back to the native, the additional information needs to be recorded, such as the source language, the type of transformation, all this information if it's kept with the transformed data that would help to go back to the reverse or the native version of the string. So again - and obviously as you see, all this is essentially not to our knowledge, you know, taking into account any registry operation at the moment and in protocols and anywhere else.

Page 21. Yes, one - and this is kind of the first sentence here says is written as one tool may not work for all contact information. Roughly speaking I would rephrase that as one tool does not work for all contact information. We haven't, you know, you've seen the list of tools. There's obviously more than what we listed but the one that we tested and none were able to provide very good, accurate with the low Levenshtein distance for all kinds of data.

For example there's tools on the market that are related to addresses. They do verification, they do a pretty good job in addresses themselves, but then for names, organization names are really bad or other stuff. So there is no one that does very well at least to our knowledge. Transliteration is usable for script which works with consonants and vowels but do not work with script where consonants or vowels are (unintelligible). An example of this is Arabic as I said before.

Ad hoc transformation using translation systems usually give average output. It's not really predictable. Therefore the action if you go back to the different levels of transformation that we discovered at the beginning, then translation doesn't fit well in some of those. Translation actually more readable independent of the scripts of language there (unintelligible) from an end user perspective but there's a very limited set of language barriers that have automatic translation systems.

And any new translation system from a language barrier is very challenging. So for example, if you go from Russian to English for translation, you may find, you know, good tools, but if you go from Russian to I'd say Arabic, then it may be very bad because the two - the language barrier is not well implemented in tools. And then if you go through a third hybrid system, it's not, you know, demonstrably better.

Page 22. Consistent, as you see the difference here is that translation gives a better for human perspective but transliteration gives more consistent transformation. But then the output may not be usable from a user point of view. Accurate transformation is not - is essentially not possible through automated processes. It requires a very manual report and, you know, registrant verification of someone who has an understanding of a native language.

Sorting through randomization interest - is an interesting possibility to provide like local language to local language, but then you're actually doubling the issue of the accuracy that we've seen between - with the tool itself. That is even more challenging.

Next slide. Or I think that's the last one, right? Yes. So again...

Chris Dillon: Thank you very much, Marc.

Marc Blanchet: I can probably - let me - give me a sec, I can probably find - so...

Man: Hi, Marc, this is (unintelligible). I think we probably need to wrap up here.

Chris Dillon: Thank you for that, Marc. It was a very, very interesting presentation and the report is also very interesting. I think there are probably several questions. I have at least a couple myself. But we are over time and so we'll need to do that either on the mailing list or in London I think at this stage if that's all right, unless there's something very burning.

Marc Blanchet: I put the information in the chat room, the URL of the study, so.

Chris Dillon: Yes. Thank you very much for that and I will start a discussion on the mailing list picking up one or two things soon after the call. Again many thanks and thank you all of you for attending today's call and very much looking forward to seeing many of you in London in a couple of weeks' time.

Julie Hedlund: And, Chris, this is Julie. Just to remind everyone, there is no call next Thursday. The next meeting will be in London.

Chris Dillon: Thank you very much. Goodbye then.

Julie Hedlund: Thank you, everyone. Bye.

Man: Bye.

Man: Thank you very much indeed. We'll now stop the recording.

END