Open Science Talk No. 58 (2022): 10 Years of TROLLing : a computer-generated transcript [1]

00:00:08 Per Pippin Aspaas

Open Science Talk, the podcast about open science. My name is Per Pippin Aspaas. Today, my two guests are Lukas Sönning and Laura A. Janda. First to you, Lukas: you are with us online from Germany. Welcome to the podcast.

00:00:27 Lukas Sönning

Thanks for having me.

00:00:29 PPA

And in the studio in Tromsø, Laura, welcome to you as well.

00:00:33 Laura A. Janda

Hi. Thanks a lot.

00:00:36 PPA

So we're here to talk about something called TROLLing, the Tromsø Repository for Language and Linguistics. Is that right?

00:00:44 LAJ

Uhum.

00:00:45 PPA

It was actually your idea, Laura, I've been told. 10 years ago exactly now, TROLLing was launched, and you're sort of the founding mother, if that's allowed to say. Could you tell us why TROLLing came into being?

00:00:59 LAJ

Yeah, so I was already at that time using quite a bit of statistical data in my linguistic research, and even though I try very hard to keep things tidy in my own computer, sometimes it can be hard to find things again, and I thought it would be great if somebody more professional was curating and taking care of that data and making sure that it was professionally managed and also that it would be available not just to myself but also to others. And yeah – so that was, kind of, the idea.

[1] This is a computer-generated transcript of the podcast episode Open Science Talk No. 58 (2024): https://doi.org/10.7557/19.7631. The automated transcript has been proofread by Per Pippin Aspaas. It is included here for the sake of Universal Design and improved discoverability by full-text search engines.

00:01:37 PPA

Yes. And then you came to some colleagues at the library here at UiT. We should have mentioned by now that you are a professor of Russian linguistics. And so you came to the library and asked for their help. And what was the response?

00:01:51 LAJ

Yeah, remarkably, the library was very happy to take on this rather large project. And we were really grateful for that because of course, a library has a much more professional expertise in the archiving and managing of data and information than I would have as a lowly professor.

00:02:14 PPA

Yeah. And over to you, Lukas. You work with English linguistics, is that correct?

00:02:21 LS

Yeah, that's right. I'm currently a postdoc at the University of Bamberg in the English linguistics department.

00:02:30 PPA

And how did you even hear word about TROLLing?

00:02:35 LS

At a 2018 conference – ISLE 5, I think it was – we hosted a workshop on implications of the replication crisis for linguistics, and in the course of preparing this workshop we had a look at different open science channels that are available to linguists, and that's how I became aware of TROLLing. And I remember very well watching the YouTube clip that recorded Laura, which was a very nice entry into the TROLLing archive. So that's how I became aware of it. And I only started using it a year or so later, so that was in the course of this research for that workshop.

00:03:33 PPA

But since then you have been a quite frequent user of TROLLing. I could see, by now, 13 published datasets. And that's in a short span of time. Why do you use TROLLing for your datasets?

00:03:50 LS

I think, by now I'm making an effort to make every academic contribution, or publication, into a bundle, which does not just consist of the the paper, but also of the data and code. And I think, after the first few submissions to TROLLing I really began to appreciate the excellent service and feedback and advice that the team provides and that the data curators offer. So I've also become a bit of a fan, I think, of TROLLing and I think – yeah, that's probably the main point.

00:04:41 PPA

Thanks for that. It's called the Tromsø Repository, but it's free for researchers all over the world. Did you give it much thought when you decided to baptise it 'Tromsø Repository', Laura? Can you remember?

00:04:59 LAJ

Yes, I remember quite a lot because we we got a certain amount of pushback too, because of course TROLLing is also a word for, sort of, negative activities that take place on the Internet, let us say. But, I don't know, I really like the name anyhow. And I think that – I don't think that that association has really gotten in the way, but it's completely straightforward – the name describes exactly what it is, when you look at the spelled out name and the short name is easy to remember. So I don't know, although there were a few people who complained in the very beginning, after that it kind of died away and people got used to the fact that we were going to take back this word and give it a positive spin.

00:05:56 PPA

Can you relate to that, Lukas?

00:06:00 LS

Yes, I can. And of course, as an external scholar, I also have to say that this is quite remarkable – that, people from other universities, other countries can make use of the same – of your resources, actually. So I think that's one of the great parts of TROLLing, that it's open to external scholars. And yeah, from day one, however, I've had kind of a bad conscience for using resources at another university for improving my research. So I think that's something – at least for me, because I'm quite an active user of TROLLing – I would like to contribute to it in other ways as well, not just take advantage of the repository.

00:07:00 PPA

How can you contribute to it in other ways? That made me curious.

00:07:05 LS

So, about half a year ago, I was invited to be part of the Scientific Advisory Board and we had a meeting – in January, I think it was. And there were some some thoughts circulated as to how TROLLing could manage an increasing number of submissions, and so on. And I think for me personally, one option would be to contribute as a kind of a 'reviewer' or a 'preliminary data curator' – to do some of the work, the repetitive work that's part of curating a data set. so that would be one concrete way in which I can picture myself making a contribution, I think.

00:08:08 PPA

It's interesting that we now use words that are similar to those that are used in journal publishing, for instance. It's like submitting a data set is comparable to submitting an article manuscript to a journal. Then it goes through peer review by other scholars, and then if it's approved and improved upon by the author – usually there are some rounds with the editor and so on – and then, finally, it gets published. In the data world, there are actually two options for researchers. Either you go this long route through a kind of peer review, where the peer reviewer is usually set in a library – like TROLLing, these, the curators are working in the library, so they are the 'reviewers', you could say, and the editors of the TROLLing 'journal' in quotation marks. But you can also go another and simpler route. It's just submitting it somewhere, uploading it somewhere on the Internet without no curation at all. Of course that is hugely faster. It's extremely fast. Why do both of you – I should mention that Laura didn't just start TROLLing, she has been using it ever since, more than 20 datasets by now, so, I mean, you continue to use it. Why do you go that extra mile through the TROLLing curation process?

00:09:37 LAJ

I think I do that because it's very important to me to hold to the highest standards and best practises for our field. Because if we go that extra mile then our datasets – we feel very assured that they can be reused. And they – I know they are being reused – and that way, we contribute to the entire field and we contribute to helping people to – how can we say – we contribute to helping people to raise the level of professionalism and expertise. So yeah, I think it's, I think it's extremely important to do. If you're going to do something, do it well, don't do it halfway.

00:10:37 PPA

Exactly. Just to pick on one thing you said, data are being reduced, you say, how do you – as the author, so to speak of those data sets – how do you get notified that they are being reused? Do you get quotation citations? Do they send you an e-mail or how does it work?

00:10:54 LAJ

I don't get any citations from TROLLing – I mean, there's no way that TROLLing would actually know either. I guess they would see if something was downloaded, but we actually don't get those. I was once an external examiner for a dissertation – this was just, it was just before the pandemic – for a dissertation in Leiden. And when I was reading that dissertation prior to going for the dissertation defence, I discovered that a large part of that dissertation had taken my data and my statistical code and reused it on another data set. So in other words, the PhD defendant had used my data and my code in order to carry out his own dissertation research. And I thought that was really exciting.

00:11:54 PPA

So you felt that was exciting, others would would say: this is a robbery, it's your data, it's your code. But that's not how you feel about it?

00:12:00 LAJ

Oh no, no. I mean, he didn't steal it, you know, obviously he cited it and everything, but he took what I had done and took it in a slightly different direction and gathered his own data but it made it possible for him to do the research that ultimately earned him a PhD. I think that's wonderful.

00:12:24 PPA

How about you, Lukas? Do you have any dream scenario about how your data should be reused by others?

00:12:34 LS

Well, first of all, I'm usually quite happy if someone's interested at all in my research. So, I think the fears that some people have, I don't share them. I've only been contacted once by someone on a TROLLing post and I actually learned something from that exchange because it turned out that the way I had obtained data from a corpus would have been possible in a much easier way. So there was another option that I wasn't aware of. To the other person, the interesting bit was that there was a very, very minor difference in the output of those two data retrieval strategies. But for me it was something I learned, so that was already one nice scenario that unfolded. At the moment, for instance, I'm doing mostly methodological research, and I'm grateful if – I'm always grateful if I find data from other people that I can use – like, real, actual data to test things. And that often leads to an interesting exchange. So, for instance, at the moment I'm collaborating with Raquel Romasanta from the University of Santiago. And we're presenting a paper at ICAME together in a few weeks, and this is based on her data, and kind of connecting to methodological work I'm doing at the moment. So yeah, not my data, but another nice scenario, I think.

00:14:21 PPA

Yeah. Excellent. How about you – that 'extra mile' and this question about the curation process, which of course takes up some of your time, it must surely – there are things to fix on your datasets and that takes up some of your time. How do you feel about this? Why is it worth it?

00:14:41 LS

Yeah. I'm also. I'm actively advocating TROLLing and I'm advertising it to other people. And I always say that the first data set publication is the hardest and after that it becomes more and more of a routine. I mean, the nice thing also of TROLLing – about TROLLing – is that the documentation of a data set is quite standardised. So there's a fixed and recommended structure of what's called a 'readme file', where you describe your data and methods and other things in detail. And writing this document is the most work when preparing a TROLLing post. But this is an example where the second or third 'readme file' you write, you know what goes where and so on. And I think, the moment I opened my first curation report – this is the document you get after the first round of review – I was overwhelmed by the thought and care that went into quality control from the data curators at TROLLing and – I mean, in that instance I realised that even though I had to invest quite a bit of work into that very first TROLLing post, a lot of work was done in return in giving me feedback and suggesting improvements and so on. So yes, I mean, I also feel that we should be aiming for the highest standards and I also see, however, that the quality of my output and only the data and the documentation I'm providing has improved a lot from the feedback – or based on the feedback – I've received from the data curators. So I've really grown, I feel, from my history with TROLLing.

00:16:57 PPA

Yes. And in another sense, TROLLing as a database is growing – it's now 173 datasets at the time of recording. So, 173 datasets for 10 years. Lots of curation work has been going on. So the question for the future, then, Laura: where do you see TROLLing in 10 years from now?

00:17:23 LAJ

OK. Yeah, 10 years from now, I would hope that there are more people like Lucas who will become active users of TROLLing. Also, I hope that TROLLing will become more of a standard best practise for everyone, that journals will require this more and more. I mean, I think this is a growing trend and also that this will be required as part of the process of applying for grants, in order to make sure that all of the data is publicly available, and also the code. I think that TROLLing can be a major tool for people who are learning how to use statistical methods in linguistics. I myself – I'm not teaching that course right now, this semester, but – many years I've taught the course in the use of statistical methods for linguists at the university here, and I've reused all of my own data sets in that course, so that the students can see how to set those up. So yeah, I think this will be – will become more like standard practise for the field.

00:18:48 PPA

You, Lukas?

00:18:50 LS

I hope that in the future, TROLLing will also find a way of obtaining the necessary resources to cater external people like myself. Because I think if the goal is for TROLLing to grow and the quality is supposed to remain at the same high level, there's lots of curation work which will have to be done, and I wish that in a few years time at the latest, maybe I can play my part in the review and curation work, if TROLLing decides to 'outsource' parts of the work.

00:19:44 PPA

That bodes very well for the future of TROLLing. And so, at the end of this episode – I always ask, do you have anything else that you would like to add before we stop recording?

00:19:58 LAJ

I was interested that Lucas mentioned this video that we made about TROLLing. It was one of the first things that we did when TROLLing was launched – we made a kind of humorous video, and in that video there's somebody who – a graduate student – who comes and knocks on my door and I'm like the – I'm the difficult old professor who never – I'm playing that role – who never archived anything and can't provide the data that goes along with the articles that I've published and isn't willing to help this poor, this poor dissertator, the poor student. I slammed the door in his face and then he cries out in the hallway: 'What about my dissertation?' And anyhow, I wanted to tell Lukas that what happened was that we actually – we rehearsed that the day before. And the person who plays the dissertator, the PhD student, he actually was my PhD student. So we rehearsed this the day before, and then the next day we came back with the camera crew and actually, we actually did the video. But the second day, some of my colleagues down the hall, they said: 'oh, it's just for, you know, this is just a recording, this is just a joke, they said, we thought Laura was abusing her graduate student'.

00:21:39 LS

Yeah, that's a nice story.

00:21:42 PPA

OK, so you you got away from that reputation.

00:21:46 LAJ

Yeah, yeah.

00:21:53 PPA

Yes. So at the very end of this episode, anything you would like to add, Lukas?

00:22:02 LS

Yes, I think I would like people who consider using TROLLing to know that the staff at TROLLing – the data curators – are very, very friendly and also very patient dealing with your submission. So that's been a completely positive experience. So I think I would like people to know this.

00:22:28 PPA

Open Science Talk is produced by the University Library of UiT The Arctic University of Norway. Thanks for listening.