Open Science Talk No. 42 (2022) Dataverse.no: a computer-generated transcript [1]

00:00:11 Per Pippin Aspaas
Open Science Talk, the podcast about open science. My name is Per Pippin Aspaas, and today I'm joined by two colleagues from the University library here at UiT the Arctic University of Norway, Philipp Conzett and Leif Longva. You are managing the Dataverse.no Archive. So first of all, Leif: where did it all start?

00:00:43 Leif Longva
It started back in the autumn of 2013. We were actually approached by some linguists who asked us for help to establish an archive that they needed to disseminate their research datasets in linguistics. And I guess it was based on these things that we've done on other types of open archives – meaning publications and master's theses, doctoral theses – establishing an archive for these kind of outputs. So they knew that we had some competence on open archives and so they approached us. And so that was the fall of 2013. The Linguistics Archive was established in 2014. And we built on that to establish what became Dataverse.no, eventually.

00:01:42 PPA
Yeah. So now when you talk about the linguistics archive, it's TROLLing, I guess – what today's called TROLLing. And what is that, Philipp, could you explain?

00:01:50 Phillip Conzett
Well, TROLLing stands for the Tromsø Repository of Language and Linguistics. And as Leif just said it's the main specific repository for research data from linguistics. It was launched in 2014, so in spring 2014, after we had this project that Leif was talking about. And I was happy just to start at that time when the library started to engage in research data management support services. So it has been an exciting time ever since then, with a lot of stuff happening all the time and new features coming up and new recommendations and more and more researchers, kind of, understanding the rationale behind open science, and so on. So yeah, so we're still running TROLLing from Tromsø and yeah, it's has been growing ever since.

00:02:52 PPA
Yeah, I guess we could have an own podcast just on TROLLing one day, but then it moved from a very discipline-specific archive, TROLLing, to what is now Dataverse.no – and how did that happen?

00:03:09 LL
Well, we firstly, we didn't think of Dataverse.no as an idea all, we were thinking of an open research data repository for UiT the Arctic University of Norway – for our institution. So that was what we did: to make the archive suitable for any subject areas – any research dataset from any subject area, and we spent a couple of years on that. In 2016 it was, that we launched the UiT Archive. And then when we did that, we were actually approached by some other institutions who asked us "could we use

[1] This is a computer-generated transcript of the podcast episode Open Science Talk No. 44 (2022): https://doi.org/10.7557/19.6773. The automated transcript has been proofread by Nikolas Tobias Freiberger and Per Pippin Aspaas. It is included here for the sake of Universal Design and improved discoverability by full-text search engines.

your repository in some way?" And that gave us the idea of … we found out that it wasn't that much work to be done to adjust the software so that we could make it a multi-institutional repository and then we named it Dataverse.no. And in the fall of 2017, we launched Dataverse.no.

00:04:24 PPA
Yeah, the software you just mentioned, Philipp – in name Dataverse. And could you tell us what is that?

00:04:33 PC
Yeah, so the Dataverse software is an open-source software designed for archiving and publishing research data. It was developed at Harvard University quite a long time ago, and it's still Harvard leading this open-source community for the Dataverse software and now a lot of other developers and research data management experts around the globe are contributing to develop this software further. And, as Leif said, from a technical perspective the software is quite flexible. So from a technical perspective, it was quite easy to adapt our institutional repository to become a national repository with multiple institutional collections. So I think the main work in establishing a repository is always to create the supporting services around the technical solution.

00:05:39 PPA
Yeah. The supporting services, one thing is the technical software – setting it up so that it works properly and there has been plenty of help from the IT department also – we should remember to mention, it's not strictly librarian-run service. But what kind of support services are then built around this software?

00:06:04 LL
Well, first and foremost, we have a curation. We are curating the datasets to make sure that the datasets are archived properly, and by properly we mean that it's archived so that anyone who finds the dataset is understanding what this dataset is holding – what kind of data and what kind of information is in the data. So we need to curate the datasets that the researchers submit to the archive. So the curation job is an important and a core service that we do, making the datasets as FAIR as possible – the FAIR Principles that we may talk more about.

00:06:57 PPA
Yeah, the Findable, Accessible, Interoperable and Reusable, the FAIR principles. They are sort of the basis, as I understand. But when you say curation: what happens if a researcher submits a dataset? Then I mean, if you submit to one of the online for-free services out there, there is no curation, basically, I guess, but they get something else from Dataverse.no. Could you explain that, Philipp?

00:07:27 PC
Yeah, it's basically the service that Leif just has mentioned. So when a researcher deposits a dataset and submits it for review, this means that – so, in Dataverse.no we have configured the system so that it's only possible for curators to publish datasets. So researchers, they create the dataset, a draft and submit it for review and then someone that's usually at the library or in the research administration at the Dataverse.no partner institutions – like UiT and University of Oslo, Bergen and so on – they will then have a look at the dataset and check if it complies with our deposit guidelines, basically. So for instance is there a ReadMe File, describing and documenting the dataset. Did they have, did they add appropriate keywords, for instance, in the metadata scheme and so on. So it's basically just a kind of quality check – not so much about the quality or the contents of the data, which would be a kind of peer review – it's more about the FAIR aspects of the dataset.

00:08:40 PPA
Yeah. So more like the technical aspects, are there good keywords? Is it sustainable in the long run, technically speaking?

00:08:47 PC
Yeah. Also file formats for instance. So if they just have added a tabular file as an Excel file, we will tell them to also convert it into a plain text file to make sure that the data can be preserved and reusable also in the long term.

00:09:02 PPA
I see. Are there other support services that we should mention? Leif, I know that you've been into PhD teaching, does this Dataverse.no have any role to play there?

00:09:17 LL
Oh, sure. The keyword “teaching” is central to the list of services that we offer. We offer designated courses for PhD candidates. But we also run open courses or webinars that anyone may just drop in to follow. These are thematic courses that follow each step on the way – from planning your research project all the way up to archiving your dataset. So sort of following this research data management lifecycle, all the steps included there are subject to a thematic course. And those courses are quite popular and people who follow these courses or webinars – we did both in class but also webinars especially during the COVID period we needed to develop these into webinars – and people who follow them are quite happy with them. The feedback is good. So this is a much needed service that we do to sort of train our researchers and graduate students on how to do these things.

00:10:47 PPA
Yeah, we could mention that there are many institutions across Norway participating in this. So it's not just UiT who runs this so there is teaching and curation going on all across Norway. How many institutions have joined so far, Philipp?

00:11:03 PC
Just a couple of weeks ago, Nofima Research Institute for Food Sciences in Norway, they joined us – as number 14, I guess. So it started with two institutions the UiT here in Tromsø and the southernmost university in Agder and then it increased. It has been increasing gradually and I just wanted to add to what Leif has been talking about, teaching and training and so on. So the courses Leif has been mentioning they are targeted, mostly at researchers at the different partner institutions. So as you said, Per, it's not only UiT, but most of the other Dataverse.no partner institutions, they offer these courses or webinars for their employees, or also, if they are open, anyone can join them – if it's a webinar, it's online for instance. But in addition to that, we also have webinars and training events for the Dataverse.no curators. So for the people that curate the datasets in the repository. So we run a couple, yeah – it depends, but we run them on a on a regular basis, each semester. We would talk about, you know, different topics: about what is important to look after when you curate the dataset like file format and metadata and so on.

00:12:30 PPA
So proper training of the curators themselves, I guess, is key here. We could also mention that it's five years ago this autumn semester that the first institution joined and Dataverse.no was created. So this is a sort of “birthday celebration podcast”. What about the kinds of data stored in Dataverse.no, Leif. Is it for any kind of research data or is it more designated for certain disciplines?

00:13:02 LL
You may say that it's for any kind of data, but there is one definite limitation: that these data are, or may be, openly available, it's for open research data. No person-data should be in there, no sensitivity regarding persons or regarding business sensitivity or national security or whatever. So those kinds of data are not suitable for Dataverse.no, we can have a lock on the files for a limited time, but the data in Dataverse.no are meant to be open, at least in a year or two. So a researcher may ask us to have them locked for a while, but not more than a year or two.

00:14:03 PPA
I see. So it's for open data and mostly not data involving people, because then you have regulations regarding sensitivity, etc. I know that Dataverse.no a couple of years ago, got something called a CoreTrustSeal attached to it. What is that, Philipp? And could you tell us why Dataverse.no has this seal – briefly?

00:14:30 PC
CoreTrustSeal is a certification framework and it allows repositories – or, mainly repositories, research data repositories – to demonstrate that they operate their service according to best practices. So you have to document that you follow best practice in different aspects, for instance in file format assessment, long term preservation – but also, you know, the duration or how you guide your researchers when they deposit data, about fulfilling requirements from different research communities, if there are some, for instance. So we basically applied for this certification to demonstrate for our researchers … So that it's, kind of, a convenient way for them to also demonstrate for their funders, for instance, or for a journal publisher that they comply with the quality requirements, or best practise recommendations, in the field.

00:15:37 PPA
Yeah. So you get sort of a stamp, it's a research dataset that has been stored in a proper archive, which is trustworthy.

00:15:46 PC
Yeah, and also for instance, the EU recommends the researchers to choose repositories that are certified because it makes it easy to choose a repository, so you can then, kind of, trust them – that they are operated in a responsible way.

00:16:07 PPA
You mentioned earlier, Leif, that a researcher submits a data set to the archive and then it goes not through peer review – to take the analogy with Journal Publishing – it goes through curation, and you are one of the curators. Is there some sort of author facing processing charge – like you have APCs, Article Processing Charges in the journal business? Do you have this kind of charges?

00:16:35 LL
Nice question. No, we don't. So for the researcher, there is no processing charge for archiving his or her dataset in Dataverse.no. So how is it financed? It's backed by our university, UiT, and then there is a fee for any partner partnering up with it – so that's the financial basis of the service. So no payment per research dataset.

00:17:14 PPA
Yeah. And there are of course – the reader, so to speak, of the dataset will have no charge either?

00:17:21 LL
Sure, sure. It's definitely Open Access, yes.

00:17:25 PPA
Yeah. So what we're touching upon here is the Norwegian word “dugnad”. It means working voluntarily together, you could say. So these people involved as curators, they are of course employed by their universities, but the universities – and the university colleges and research institutes that have joined so far – they join in a collaborative effort called dugnad, you could say. Philip, if there are listeners out there, not from Norway saying, “oh, this sounds like a good archive, this is the place where I want to store my datasets”. I guess they would need to be a linguist in order to submit because this is for Norwegian partner institutions, isn't it?

00:18:09 PC
Yeah, it's basically for any kind of data and from any discipline. Our main designated community, or targeted community is, kind of, the researchers from the long tail of data – from the long tail of research, sorry. Meaning that some researchers, they would have their domain-specific repositories that are more tailored towards their needs and for them, most often they would choose this kind of repositories if they are suited. But in many disciplines there are no such domain-specific services, so they would choose services like Dataverse.no and others.

00:18:53 PPA
Yeah, but if you were a linguist situated in, let's say, Berlin, then you could submit your dataset to TROLLing?

00:19:00 PC
Yeah, for instance.

00:19:02 PPA
Because you don't have to be part of a Norwegian institute to actually use TROLLing, but I guess for you to use Dataverse.no then you would need to be affiliated, at least, to some institution in Norway or have collaborators there.

00:19:17 PC
Yeah, it's a national repository and the our main target group are our researchers associated with the Norwegian research organisations, yes. And just to add to something Leif said about the how it is funded: it's as Leif said, there is a fee for partner institutions but the main contribution is the curation work that the partner institutions contribute with. So that, as you said, it's the people contributing backstage to the service.

00:19:51 PPA
Are there other services like this – I mean, in other countries, Philipp?

00:19:58 PC
Yeah, it's becoming more and more popular, the Dataverse software, and there is currently a number of national installations or repositories based on Dataverse, for instance in the Netherlands, and just recently there was one established in France, we have one coming up in Hungary and Denmark – it's very popular across the globe.

00:20:32 PPA
So if you're not from Norway and you're listening to this programme, maybe there is a Dataverse installation that you can use anyway. Well, is there anything else we should add here, at the end of the programme? Do you have any extra thoughts?

00:20:51 PC
Yeah. So we're trying to improve our service continuously – for instance, you know, improving our training webinars that we have for curators, and also we are working on, we have been working on a grant proposal for external funding from the Norwegian Research Council to upgrade the, kind of, make a major upgrade of the repository, of the whole infrastructure, also on the technical side and also on the competence side and the human side. So we are going to resubmit our proposal next year and hopefully this will also allow us to make a, kind of, major upgrade. So usually when we have our smaller improvements we have to do this, you know, besides our other commitments and so this kind of external funding would help us to, kind of, make a major improvement.

00:21:50 LL
I'd like to add that one important thing that we still need to work on is to make the researchers aware of the importance of research data management and research data archiving. Researchers are very occupied with publishing their papers in the “correct” journals, in the “good” journals. And of course they are. But they should also – we need to still work on making them aware of the importance of good data management and good data archiving because the datasets from a research project, those are what makes your research accountable and reproducible. If you want to, if someone wants to, look into what you did as a researcher and question your research findings, then they need access to the data and the description of what you did when you collected the data and all these things. So awareness around these things is something that we still need to work on.

00:23:11 PC
I totally agree. And that's also why it's important to have services like Dataverse, you know, that, kind of, support the researchers in this endeavour and this work with open science to make their research more transparent and more reproducible and accountable, as you said.

00:23:32 PPA
Well, at the very end of this podcast episode, I would like to congratulate you both with this service. It's now almost five years old to the day. How are you going to celebrate the birthday?

00:23:47 LL
With a cake!

00:23:49 PC
Yeah, that's true. So maybe, you know that there will be a conference on scholarly publishing at the end of this month and the start of December. And at the last day of this Munin conference, we will celebrate with, I hope, quite a large cake, yeah.

00:24:10 LL
Yeah. And then there will be a jubilee seminar, a small jubilee seminar, the following day.

00:24:21 PPA
Yeah. With that, I wish you a happy birthday when that, those days come and thank you so much for coming to the podcast.

00:24:30 PC
Thank you for having us here.

00:24:33 LL
Thank you.

00:24:37 PPA

This podcast is produced by the University Library at UiT the Arctic University of Norway. Thanks for listening.