Going Dutch with your data!

In this month's interview, the roving EUDAT reporter caught up with Ton Smeele and Joyce Nijkamp after they gave a presentation on the Dutch collaborative project, U2Connect, at the EUDAT User Forum that was held in Prague on 23-24 April. Ton and Joyce are both employed by central Information and Communications Technology (ICT) departments in the Netherlands: Ton is a data management specialist at Utrecht University, and Joyce is an enterprise architect at the University of Amsterdam.

Firstly, Ton and Joyce, what is U2Connect? And what is it doing?
U2Connect is a collaborative project being undertaken by eight of the Dutch universities in conjunction with the Netherlands Institute of Ecology and SURFsara (the Dutch national high-performance computing and e-Science support centre). Our aim with the project is to help Dutch researchers to manage and use their data more easily and thus make it simpler for them to collaborate and share their data. To achieve this, we plan to enhance current university data infrastructures to include connections with EUDAT services and, in particular, we are focussing on enabling researchers to transparently work with data, irrespective of where it is stored. By "transparent" we mean that it will not make any difference to the researchers whether the data is stored locally or at another research institute. Currently our data infrastructures service roughly 60% of the academic researchers in the Netherlands.

So, in essence, what you are doing is connecting the data infrastructures at the different Dutch research institutions with EUDAT's services. How do you see that as benefiting researchers in the Netherlands?
In their present state, the EUDAT services already provide significant benefits to our researchers and, by integrating these services more tightly into our national infrastructure, we will make it even easier for the researchers to use the services and share data with each other. EUDAT refers to this type of access as "joining" EUDAT (rather than simply "using" EUDAT's services).
In addition, as things stand, it can be somewhat difficult for researchers who want to collaborate with each other and work on the same data during the course of their research. If they want to avoid having to centralize all the data analysis activities or want to avoid having to duplicate all those terabytes of research data, there is no easy way to do that at present. U2Connect is therefore exploring the option of using iRODS technology for collaborating on distributed data (that is, data stored at different locations) during research. Since iRODS is already used by EUDAT as the underlying technology for the B2SAFE service, we consider it to be a convenient choice, as well as providing a functionally rich platform.
The main thing to remember here is that it is essential for researchers nowadays to be able to collaborate easily and share data between different organisations (in the Netherlands or further afield) while maintaining the integrity of the data.

Well, it definitely sounds useful to have better data integration between the various research centres in the Netherlands so it is easier to access and share data from different locations. I notice you mentioned that the institutions currently participating in the U2Connect project cover about 60% of the Dutch academic researchers. What about the other 40%? Are there longer term plans to integrate the data infrastructures of all the Dutch universities and research institutions with U2Connect (and thus with EUDAT)?
U2Connect is an action-oriented initiative. We wanted to have a sufficient number of the Dutch researcher institutions involved from the start of the project to build momentum, and then aim for a snowball effect with more and more institutions joining as we moved forward. In actual fact, more Dutch institutions may be joining the project in the forthcoming months. Additionally, the project results will be incorporated into a blueprint so that other institutes and research groups can join at later stages and benefit from the experience of those who adopted the project early on.

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

So, here we are seeing a national level initiative of Dutch research institutions that may even end up incorporating all the research centres in the Netherlands, and you have jointly chosen EUDAT as one of your primary research data service providers. Why choose to collaborate with EUDAT in particular?
What better alternatives are there? The needs of our researchers cross the borders of academic disciplines and also countries, and therefore effective data management is likewise a cross-border challenge, both academically and geographically. To manage such diverse data successfully requires levels of coordination and investment like those now being provided by the European Commission (EC). We are excited that the national data centres across Europe have joined forces with the research communities to respond to this challenge.

Ok, so now we understand what you're doing and why (and it definitely sounds like a good idea!), would you give us some details about how you're going about this process? Personally I sometimes get a bit frustrated reading material about EU projects as there seems to be a lot about solving grand challenges and exploiting synergies, but what I'd like to know is what people are actually doing to make all these amazing things happen in real life, or at least in "e-life".
Indeed programs like EUDAT can be dauntingly complex and involve many cycles of development. So we are taking this project step by step, starting with pilot studies to develop our expertise and build trust between the initial institutions involved with the project, and then we will be scaling up to a address a broader audience.
First we are assembling one pilot group per university whose members will be helped to gain a comprehensive understanding of the depth and reach of the EUDAT services through training and hands-on experience. The pilot groups typically include IT research support staff and library specialists, and are linked to a real life research project that acts as a use case. The total dataset per group is expected to be from 1–20 TB in size.
Next the pilot teams will test drive the EUDAT services and iRODS technology locally and observe how they are used by researchers. This will enable us to detect areas where the services are beneficial as is, or where they need to be complemented or extended locally to meet the needs of the researchers. We will provide EUDAT with our findings from these observations. In parallel with observing the test drives, we will hold workshops to shape a computing platform (which will be based on iRODS grid software) that will enable researchers based at different universities to collaborate and share data. The idea behind this platform is that, while part of the research data may be stored at another university, from the perspective of each individual researcher, all the data will be made available as if it were stored locally and in one directory. University research support staff and EUDAT specialists will participate in these workshops to ensure that the platform will be able to interoperate with EUDAT.
Finally we will share our experiences and let people know what we've done, so our work can be used on a wider basis. We will package an out-of-the-box blueprint implementation of the U2Connect data infrastructure and make it available to the research field at large (possibly via the EUDAT website) to help other countries or collectives of research centres that are looking to connect with EUDAT's Collaborative Data Infrastructure. We are also looking forward to presenting an overview of the U2Connect process at a forthcoming EUDAT conference, so that other research collectives that are at earlier stages of the data integration process can benefit from our experiences.

Wow! That was great - you've actually given a really concrete overview there so we can see how this will come about. And I seem to recall that you mentioned that you were expecting more Dutch institutes to join the project soon too. It is quite exciting to see basically a whole country of researchers having their data connected to EUDAT's data services.
The U2Connect initiative is linked to the simultaneous development of university research data management policies in the Netherlands. The Dutch universities recognize that they have a responsibility to help their researchers be able to properly manage and share research data. Open Access to research data is already being widely encouraged in the Netherlands, and, in fact, it is expected to be a prerequisite in the near future when it comes to being awarded grants at Dutch institutions. We expect that the results from our project will also contribute significantly to those goals.

It is quite interesting that the Netherlands is approaching the management of research data, and providing Open Access to such data, on a national basis. You haven't limited yourselves to working

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

on data management within individual research communities and disciplines. It seems like you are taking a far-sighted view that should ultimately encourage more cross-disciplinary utilisation of research data. So, please tell us a bit more about the Dutch approach where you are developing a blueprint for a common data infrastructure implementation that connects universities to EUDAT and supports collaboration on research data between researchers of different universities and disciplines? How do you see your work on a national level tying in with work done within specific research communities?

To claim that U2Connect addresses these topics on a national basis would highly overstate our efforts and also fail to take into account the excellent work that is being done by other initiatives and communities in relation to this highly complex global issue. Existing research data infrastructures are typically based on centrally stored data and that is indeed fine for a lot of purposes. However, as typical data sets continue to grow in size, and as more and more research crosses disciplinary boundaries, we may need to adapt our approach and extend towards other architectures that make data available on demand, irrespective of where it is based. U2Connect is one step on the way towards such a global architecture. Variations in data format pose another huge challenge. Within U2Connect we plan to make a start on tackling this by agreeing on a minimum amount of mandatory descriptive data (known as “metadata”) that would, as a minimum, encompass information about the provenance of the stored data. When everyone starts consistently providing this basic level of metadata, it will provide a much healthier basis for managing different types of data.

Thanks very much, Joyce and Ton, for really giving us a concrete feel for how U2Connect is teaming up with EUDAT to offer the Dutch research community cross-disciplinary research data services that are resilient and integrated. We look forward to hearing about your progress at the 3rd EUDAT Conference, which coincidentally will be held in Amsterdam in September this year.

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.