TOWARDS BETTER DATA SHARING

If Wageningen's research data were stored using a common approach, the data would be found and used by other researchers much more often. In a new project, data scientists are looking for the optimal infrastructure for data sharing at WUR. 'We need input from all sides to find common ground.'

Text Ning Fan

'Research shows that scientists spend up to 40 per cent of their time on tasks such as finding the right data, checking its quality, moving it between systems and transforming it,' says Willem Jan Knibbe, the director of the Wageningen Data Competence Centre (WDCC). WDCC was set up to support developments in the field of big data and data science at WUR. 'Imagine how much time researchers could save if we had a common data sharing facility with guidelines for reproducibility, machine readability, data security and privacy policies.'

Currently, different WUR institutes and even different groups within the same institute use their own methods and tools for handling and storing research data. And new data programmes keep popping up. As a result, if data is collected by one group, others are not always aware of this data or easily able to find it.

The Wageningen Common Data Solutions (WCDS) programme aims to make research data easier to find and more accessible, interoperable and reusable. The programme started last September and is being funded by the Ministry of Agriculture for a period of two years. The programme explores the option of using generic open-source tools (iRODS and YODA) to connect data from all nine WUR research institutes in a common research data management infrastructure.

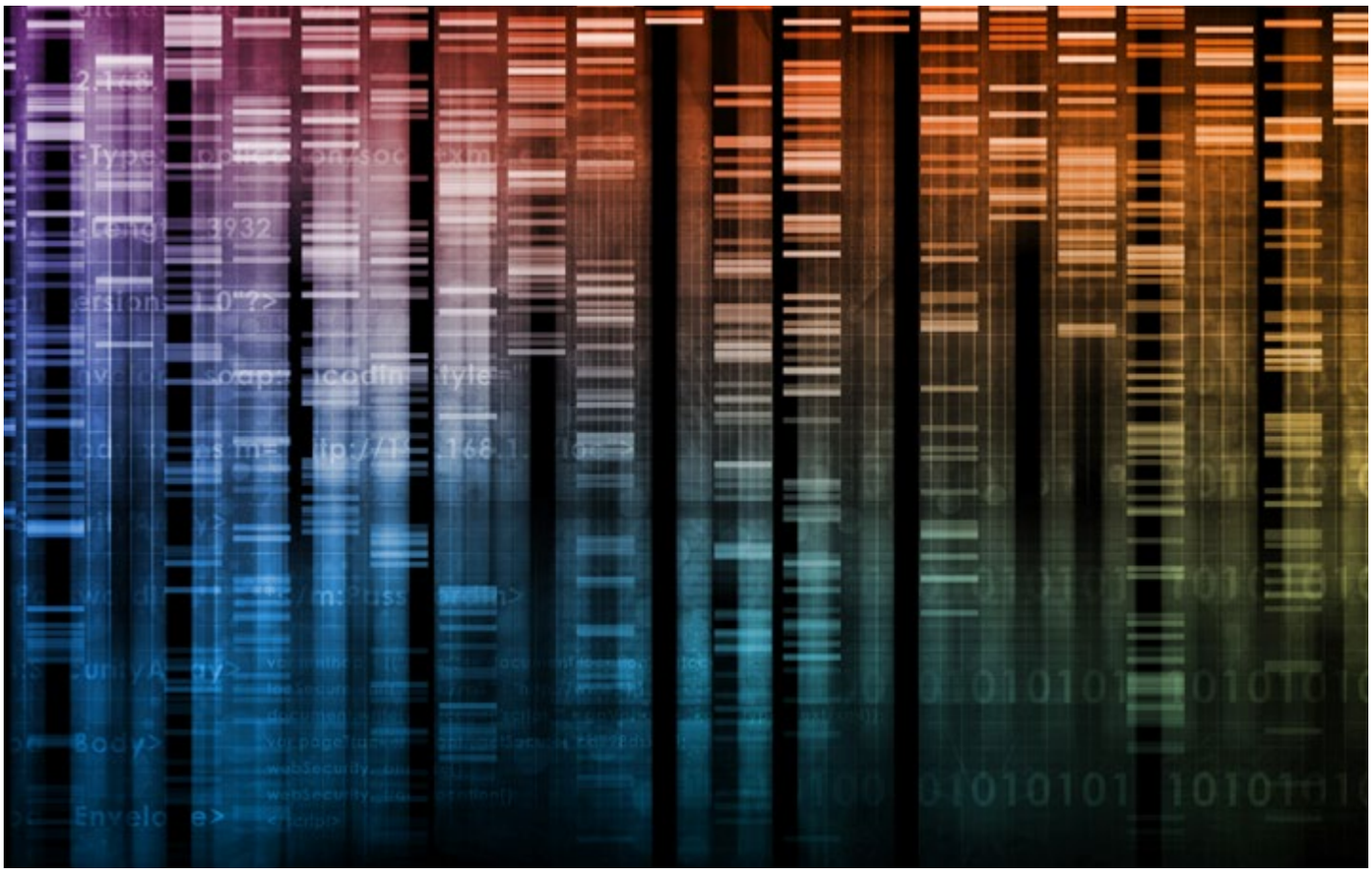
Common ground

'We need input from all sides to find common ground,' says Knibbe. 'As a starting point, we suggest using a combination of iRODS (advanced but complex) and YODA (basic but accessible) for fundamental data management tasks. By testing possible solutions across all nine institutes, we aim to find the right balance between central support and local expertise.'

In the ideal shared data solution Knibbe has in mind, users would not even realize they are using it. 'We should approach research data from various sources intuitively,' he explains. 'Imagine a situation in which an environmental research group uses soil data for a study, then stores the data in a WUR data hub following pre-defined rules in their data management plan. Plant science researchers can then reuse the data to improve crop resilience, while animal scientists can use it for research into improving the quality of livestock feed. Not only can this save researchers a lot of time, but it can also open up more research opportunities. A single dataset has the potential to unlock limitless insights, and that is precisely what WUR needs. Of course, right now we still have a long way to go to make this happen.'

Ronald Petie of Wageningen Livestock Research is one of the data researchers involved in WDCC. 'I am pleased that WUR is taking data handling and data sharing seriously,' he says. Petie and his team are developing a data handling workflow for epidemiological research data on bird flu outbreaks as part of the WCDS programme. 'We are currently

'A SINGLE DATASET HAS THE POTENTIAL TO UNLOCK LIMITLESS INSIGHTS'



'Imagine how much time researchers could save if we had a common data sharing facility.' ◆ Photo Shutterstock

using YODA to manage animal disease data. We will first see if this new system is suitable for managing bird flu data and explore its possibilities with other research groups. If it works, we could use the results to advise other scientists on the appropriate way to manage animal disease research data.'

Cost reduction

In addition to making it easier to find and share data, the costs of data storage also need to be reduced, points out Tim van Daalen, a horticultural data scientist at Wageningen Plant Research. In the WCDS programme, he and his team are working on finding an affordable way to store greenhouse data. 'Everyone agrees research data is valuable, but the price we pay per terabyte per year has a big influence on how much can be stored. We worked out several examples showing that new solutions like a tape archive reduce the yearly costs by more than 80 per cent. But can we data science researchers convince WUR to invest? I wonder what will happen when this project is finished. Data sharing infrastructure is the future; there is no plan B. But WUR is lagging behind in some aspects. The reality is that data storage is expensive and adding metadata is cumbersome. Saving everything would cost too much time and money, so we need specialist staff to determine which data to store and to store it in a way that lets it be reused. Most Dutch technical universities have appointed dedicated data stewards, someone who spends all their time dealing with data collected by researchers. But at WUR, these tasks are partly carried out by the researchers.'

When asked if finding a common solution for all research groups at WUR is feasible, Knibbe is positive: 'Definitely! I do see a need to secure commitments from both researchers and support staff. We need to build on our current initiatives step by step, finding sustainable models for financial viability, developing expertise, and building technical infrastructure and partnerships. I hope we can make convincing progress in the two-year WCDS programme so we can continue our efforts.' ■

1 February is Common Data Day, with the presentation of the status and plans for the 15 use cases from the nine institutes. Researchers, data stewards and information management staff will get an opportunity to share their thoughts. For more information about the event, please contact ning.fan@wur.nl.