NAAR HET BETER DELEN VAN DATA

Als alle Wageningse onderzoeksdata op dezelfde manier worden opgeslagen, dan kunnen veel meer collega's de data vinden en gebruiken. In een nieuw project gaan datawetenschappers op zoek naar de beste infrastructuur voor het delen van gegevens binnen WUR. 'We hebben van alle kanten input nodig om raakvlakken te vinden.' Tekst Ning Fan

'Onderzoek wijst uit dat wetenschappers tot veertig procent van hun tijd kwijt zijn aan het vinden van de juiste gegevens en het checken van de kwaliteit daarvan, het verplaatsen van data van het ene systeem naar het andere en het transformeren van de data', zegt Willem Jan Knibbe, directeur van het Wageningen Data Competence Centre (WDCC). Het WDCC is opgericht om ontwikkelingen te ondersteunen op het gebied van (big) data bij WUR. 'Stel je voor hoeveel tijd onderzoekers kunnen besparen als er een gemeenschappelijk systeem is om data te delen met regels voor hergebruik, leesbaarheid door machines, beveiliging van gegevens en privacybeleid.'

Op dit moment gebruiken verschillende WUR-instituten en zelfs verschillende groepen binnen hetzelfde instituut hun eigen methoden voor het verwerken en opslaan van gegevens. Daarnaast schieten nieuwe dataprogramma's als paddenstoelen uit de grond. Daardoor zijn data die bij de ene groep vergaard zijn, niet altijd bekend of gemakkelijk te vinden door anderen.

Het afgelopen september opgerichte Wageningen Common Data Solutions-programma (WCDS), dat gedurende twee jaar wordt gefinancierd door het ministerie van LNV, heeft als doel onderzoeksgegevens vindbaarder, toegankelijker, beter te combineren en beter herbruikbaar te maken.

Het programma kijkt naar algemene open-source tools (iRODS en Yoda) waarmee gegevens van alle negen WUR-onderzoeksinstituten verbonden kunnen worden aan een gezamenlijke infrastructuur voor datamanagement.

Raakvlakken

'We hebben van alle kanten input nodig om raakvlakken te vinden', zegt Knibbe. 'Om te beginnen stellen we een combinatie van — het geavanceerde maar complexe — iRODS en het eenvoudige maar toegankelijke — Yoda voor om basis datamanagementtaken te verrichten. Door mogelijke oplossingen te testen in alle negen instituten, hopen we de juiste balans te vinden tussen gecentraliseerde ondersteuning en lokale expertise.'

De ideale oplossing voor gegevensuitwisseling zou voor Knibbe een systeem zijn waarvan de gebruikers niet in de gaten hebben dát ze het gebruiken. 'We moeten intuïtief kunnen omgaan met onderzoeksdata uit verschillende bronnen', legt hij uit. 'Stel je voor: nadat een onderzoeksgroep van omgevingswetenschappen bodemgegevens heeft gebruikt voor hun onderzoek, worden de data opgeslagen in een WUR-datahub. Dat gebeurt volgens vastgestelde regels van het data-managementplan. Onderzoekers van plantwetenschappen kunnen de data vervolgens gebruiken voor onderzoek naar de weerbaarheid van planten en onderzoekers van dierwetenschappen kunnen de gegevens gebruiken om diervoeding te verbeteren. Dat scheelt onderzoekers niet alleen tijd, maar er ontstaan zo ook meer onderzoeksmogelijkheden. Een enkele dataset heeft dan het potentieel oneindig veel inzichten te ontgrendelen.Uiteraard hebben we nog een hele weg te gaan voor dit werkelijkheid wordt.'

Ronald Petie van Wageningen Livestock Research is een van de deelnemende onderzoekers aan het WDCC. 'Ik ben blij

'MET EEN ENKELE DATASET KAN POTENTIEEL ONEINDIG INZICHT ONTSLOTEN WORDEN'

‘Stel je voor hoeveel tijd onderzoekers kunnen besparen als er een gemeenschappelijk systeem is om data te delen.’ ◆ Foto Shutterstock

dat WUR het beheren en delen van gegevens serieus neemt’, zegt hij. Petie en zijn team ontwikkelen een workflow voor verwerking van epidemiologische onderzoeksgegevens over uitbraken van vogelgriep als onderdeel van het WCDS-programma. ‘Op dit moment gebruiken we Yoda om data over dierziekten te managen. We zullen eerst met verschillende onderzoeksgroepen kijken of het nieuwe systeem geschikt is voor het verwerken van data over vogelgriep. Als dat werkt, kunnen we de uitkomsten gebruiken om collega’s te adviseren over hoe om te gaan met gegevens over dierziekten.’

Kostenbesparing

Naast het verbeteren van de vindbaarheid en deelbaarheid van data, moeten ook de kosten van dataopslag worden verlaagd, zegt Tim van Daalen, Tuinbouw Informatiewetenschapper bij Wageningen Plant Research. Hij en zijn team werken in het kader van WDCC aan het vinden van een betaalbare manier om gegevens van kassen op te slaan. ‘Iedereen is het erover eens dat onderzoeksdata waardevol zijn en de prijs die we per jaar per terabyte betalen, bepaalt hoeveel data we kunnen opslaan. We hebben verschillende voorbeelden uitgewerkt waaruit blijkt dat oplossingen zoals een tape-archief de jaarlijkse kosten met zo’n tachtig procent terugdringen. Maar kunnen datawetenschappers WUR ervan overtuigen te investeren? Ik vraag me af wat er gebeurt wanneer dit project ten einde loopt. Data-infrastructuur voor gegevensuitwisseling is de toekomst. Toch loopt WUR in sommige opzichten achter. Dataopslag is duur en het toevoegen van metadata ingewikkeld. Alles bewaren zou te veel tijd en geld kosten, dus er is behoefte aan specialisten die bepalen wat bewaard moet worden en die ervoor zorgen dat data hergebruikt kunnen worden. Bij de meeste andere Nederlandse universiteiten is er een data steward aangesteld die zich uitsluitend bezighoudt met onderzoeksgegevens. Maar bij WUR worden deze taken deels door de onderzoekers uitgevoerd.’

Op de vraag of het haalbaar is een gezamenlijke oplossing te vinden voor alle onderzoeksgroepen binnen WUR reageert Knibbe positief: ‘Zeker. Ik zie wel dat het noodzakelijk is dat zowel onderzoekers als ondersteunende medewerkers hun medewerking verlenen. We moeten stapsgewijs voortborduren op onze huidige initiatieven en duurzame modellen vinden voor financiële haalbaarheid, het ontwikkelen van de nodige expertise, technische infrastructuur en samenwerkingsverbanden. Ik hoop dat we erin slagen met het tweejarige WCDS-programma een grote stap zetten om ons werk voort te zetten.’ ■

Op 1 februari is de Common Data Day waar alle gebruikers van de negen instituten hun plannen en stand van zaken kunnen delen. En onderzoekers, data stewards en informatiemanagers kunnen er met elkaar van gedachten wisselen. Voor meer informatie over het evenement mail naar: ning.fan@wur.nl