

Breakthrough in global data exchange on plant breeding



Longread

June 11, 2020

Access to data is crucial for the huge growth in artificial intelligence (AI) applications. No data means no applications. There has recently been a breakthrough - two in fact - that may open up for use the stores of experimental data from companies, government bodies and research institutes that have to date largely been closed off. This paves the way to efficient and transparent research into better crops that are, for example, more resistant to climate change, or require fewer pesticides. Read about how WUR contributes to Open Science.

How do you make data available for research on a global scale? What arrangements do you make in order to write up research, i.e. experimental data, in such a way that anyone in the world can make use of it? This is precisely what plant breeding researcher Richard Finkers has been working on for years, along with a large group of international researchers.

Data driven science

Richard Finkers started out as a plant breeder, but he currently focusses primarily on bioinformatics and big data, working on 'data-driven science' as he describes it. Finkers: "Plant breeders aim to improve plants through breeding. These breeding experiments generate an increasing amount of data. We want to incorporate all this information into the breeding process. But it is virtually impossible for an individual to have an overview; the quantity of data can no longer be processed. So, we are automating as much as possible. If you want to extract information from a computer or present it to others, unambiguous terms are essential for exchanging the data. For this reason, you have to come to mutual arrangements regarding: 'What are we talking about?' This is an initial step towards exchangeability."

MIAPPE allows data to be exchanged

A breakthrough has now been achieved in this field with the publication of a data standard called MIAPPE. The abbreviation stands for Minimum Information About a Plant Phenotyping Experiment. [MIAPPE](#) is in fact a set of agreements on the description and documentation of experiments in the field of plant phenotyping (*a phenotype is the external manifestation of the plant, ed*).

In MIAPPE, we have reached agreement on how observations are best recorded, so that everyone knows what you are talking about. That's the only way to make sure the information can be used by everyone.

Richard Finkers

As Finkers explains: "This data standard sets out what we write up in order to understand an experiment; the data and metadata of the data. In a potato field trial, for example, we write up the weight of the tubers in kg. And if it is measured using a different system, you include this, so that the reading can be converted. It should be seen as a kind of language that we agree on for this type of experiment. In MIAPPE, we have reached agreement on how observations are best recorded, so that everyone knows what you are talking about. That's the only way to make sure the information can be used by everyone."

Converting units in particular generates ambiguity and misunderstanding, often obstructing the large-scale and efficient use of data. Many organisations, especially within the EU, perceive the benefits of these agreements and have collaborated on MIAPPE. The first version was published in 2015, with an updated version appearing in January 2020. Work on MIAPPE is being funded by the [Elixir](#) infrastructure. Finkers relates that fellow researchers are constantly working to improve this standard, for example those in the [EU phenotyping network](#).

Not only for plant breeders

Willem Jan Kribbe, [Wageningen Data Competence Center \(WDCC\)](#) director, sees the use of a data standard like MIAPPE as a significant breakthrough, not only for plant breeders. "In Wageningen, we are conducting research into an extremely complex domain with many interconnected links," he says. "Data provide us with the means to comprehend this domain increasingly effectively. Each part of this domain has its own idiosyncrasies, in the field of data collection as well. If we succeed in improving the mutual accessibility of these data, this then opens up fantastic possibilities for computer-based research. MIAPPE makes clear that it is possible to work towards exchangeability with major players that are globally active. It's a magnificent example of how we are on the road to Open Science at Wageningen."

Breeding API (BrAPI)

Once the data sets have been standardised and made exchangeable, for example with the aid of MIAPPE, the next step in the process follows. How do you make data exchangeable using computer software? For this, an Application Programming Interface (API) is often used. An [international group of researchers](#) has developed a set of agreements for exchanging data on plant breeding; the so-called Breeding API, or BrAPI. [BrAPI](#) is a technical description of how plant breeding researchers automatically exchange data (on phenotype and genotype) between each other's computers. The plant breeders, computer scientists, biometricians etc. worked to improve this language of exchange during twice yearly hackathons. BrAPI was presented to the world in 2019 in [a publication](#).



Finkers summarises ideal processing of the data drawn from a field experiment as follows: "We record the data from the experiment in documents in the way we have agreed in MIAPPE. We then make use of BrAPI for the exchange. As a result of these international agreements, you not only have unity at a global level and are able to exchange data, but you can also consider practical applications. If you set to work by this method, you can go into the field in Africa and input data into a database with your smartphone via BrAPI (for example by using the app [Field Book](#)). [Several apps](#) have been developed in this way."

What does exchangeability achieve?

Finkers believes that global data exchange generates added value for everyone. "Information collected around the world facilitates improved choices in plant breeding. Data availability also makes a difference in the number of additional experiments that are needed. It is also often a commitment on the part of researchers with respect to their funders. A data management plan is a requirement for all the research that is funded for example by the NWO, so that the information from the study can be reused. If you are working with public/private money, your data do have to be exchangeable. Transparency and exchange are increasingly a must."

What could be better than being able to reuse the data in another study, without having to reinvent the wheel every time?

Richard Finkers

He notes that, as with all new developments, there are *early adapters* who see the advantages in the additional effort required; a middle group that does not as yet perceive the benefits; and a group that experiences a barrier. The last fears that they are merely giving away data without getting anything in return, Finkers says. He has a number of examples for them (see [article with cases](#)) where this could work well. He also recommends that they visit the [BrAPI website](#) to see how the exchange of data can assist plant breeders and researchers.

Data reuse

You may query whether reusing data makes sense. But a plant breeding company in fact does little else. They are constantly building on the material and knowledge that they have acquired in the past. Continuity is less obvious in research, Finkers says. Research is frequently funded ad hoc by the government and partners, usually in projects that are wound up after a couple of years. "What could be better than being able to reuse the data in another study, without having to reinvent the wheel every time? In particular with constantly improving methods, as in machine learning, this opens up all sorts of perspectives in promising research.

Until now this has not been that simple. If you start using old data, you spend three quarters of the time working out what was done. Once you have found out, you still have to shape the data in order to work with it. Agreements in this respect are able to save a lot of time; as a researcher you can simply elaborate on trials that have previously been conducted. Personally, I see benefits especially in yield trials in relation to stability: which plant generates higher yield and also does so reliably over an extended period? It is difficult for us to do experiments of this sort. They demand a lot of observations, multiple soil types, many years and a wide range of different conditions. You need a large amount of data to gain an understanding of which cultivars are stable under all conditions.

For me as a plant breeder, being able to use data from for example starch processor AVEBE is hugely significant. This company possesses enormous amounts of cultivation data from stakeholders over long periods, linked to starch percentages and quality per potato variety."

Biometris and precision agriculture

Finkers offers as another example that of WUR colleague Maikel Verouden, a researcher at the Biometris Business Unit. In processing genotype and phenotype data, Verouden makes use of R - software for statistical calculations and graphical presentation. Along with colleagues at Biometris, Verouden has developed a statistical genetic pipeline – inter alia for the Integrated Breeding Platform (another partner in BrAPI) – for genomic selection and prediction of the phenotype based on the genotype of new material.

The genotype and phenotype data needed for this pipeline may derive from databases that are compatible with the BrAPI standard. For this, Verouden has – along with Reinhard Simon, who works at Plant Breeding – developed the BrAPI R software package to provide the link between R and BrAPI-compatible databases. Using it, he is able to extract data for further analysis quite simply from databases. Finkers sees this as a good example of universal and simple exchange via BrAPI.

He also believes there are opportunities in the use of data from precision agriculture, a field that WUR's Corne Kempenaar is working on. In this type of farming, the farmers record a great deal of data from their operations, on occasion down to small details. For example, the yield of a crop is already measured when it is harvested in the field.

By the time that researchers and companies share data efficiently and transparently across the world, we will be five or even 10 years further down the line. So, we're taking the long-term view, but I'm pleased that WUR is able to make its small contribution to this development.

Richard Finkers

Finkers: "Precision agriculture is generating multiple data streams that I would like to be able to use in the future. Virtual plant breeding experiments could be set up using this data."

Long-term view

Finkers is confident that an increasing number of people will perceive the benefits of efficient data sharing. Finkers and his co-researchers in EU projects are now working in cooperation with the Wageningen Data Competence Center (WDCC) on workflows where they use and improve systems like MIAPPE and BrAPI - all with a view to improved crops.

Finkers: "By the time that researchers and companies share data efficiently and transparently across the world, we will be five or even 10 years further down the line. So, we're taking the long-term view, but I'm pleased that WUR is able to make its small contribution to this development."