Scientific Symposium FAIR Data Sciences for Green Life Sciences
12 December 2018 | Wageningen | the Netherlands

Conference Proceedings | DOI: https://doi.org/10.18174/FAIRdata2018.16282

Using Data Lake Stack in Animal Sciences

Schokker, D.[1], I.N. Athanasiadis[1], B. Visser[1], R.F. Veerkamp[1], C. Kamphuis[1]

[1] Wageningen University & Research, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands
Corresponding author's e-mail: dirkjan.schokker@wur.nl

Big Data is a theme that receives a lot of attention, and is often characterised as managing and analysing large datasets to reveal new valuable patterns. In the livestock domain, big data is also becoming more common and is being anchored into the mind-set of researchers, due to, for example, sensors generating large amounts of data which are stored and analysed to generate management information for the farmers. The data of such sensors can be of any given nature, i.e. structured, semi-structured, or unstructured. With these increasingly availability of large amounts of data with varying nature, there is the challenge how to store, combine, and analyse these data efficiently. With this study, we explored the possibility of using a data lake stack for storing and analysing sensor data, using an animal experiment as use case.

A data lake is characterised by three key attributes: collect everything, dive in anywhere, and flexible access. Within Breed4Food, a public-private partnership, we have selected a case study involving an experiment in which the gait score of 200 turkeys is determined. This gait scoring is traditionally performed by a trained person. In the experiment, a variety of sensor data types were recorded, including data from inertial measurement units (IMUs), 3D-video camera, force plate, and weights to explore the usefulness of these sensors in automating the visual gait scoring. The resulting sensor output, i.e. raw data, were stored using a data lake stack. Our technology stack uses Apache Spark with Jupyter front-end and delivered using Docker. An important aspect we encountered was the meta data quality. Unique identifiers were present for the individual animals and were embedded in a searchable resource. Nevertheless, the identifiers per sensor differed for individual animals, meaning more data structuring and linking were necessary before conducting any data analyses. Accessibility and interoperability of the data was poor, as well, mainly because data were stored using proprietary software, which is not compliant with FAIR principles. Because data involved were proprietary, licensing was another point of friction that impedes data reusability.

In conclusion, we managed to set up a data lake stack and load animal experimental data of the case study. The data lake allows to easily scale up, when more data are collected in the future. Although the data is not yet completely FAIR, a first step is made by identifying the caveats. Future research will also have to demonstrate whether the process of analysing data using a data lake outperforms the traditional approach.