**Title**: **The Harper Farm Database: a platform and dashboard to showcase best practice for all Farm Data stakeholders**

**Applicants**:Eric Siqueiros, Ed Harris, Karl Behrendt

# **Underlying rationale**:

Data democratization is the process of removing all the bottlenecks/silos that restrict access to the data, providing an easy way to organize, understand and access it. Data democratization can help to expedite the decision-making process and identify new opportunities by getting the right data to the people who need it when they need it, for example, teaching, research, and on farm decision making. Moreover, it will allow the release of information from the existing data silos and turning it into a borderless ecosystem of information (a data lake). A data lake is a central location that holds a large amount of data in its native, raw format (Laskowski, 2016). This eliminates problems with data silos offering downstream users a single place to look for all sources of data.

Efforts have been carried out to mind-map the Harper Farm data, which have showed that it is diverse, largely disconnected, heterogeneous and disparate. Further to this understanding of what data is available on the farm and what is needed for all the potential different end-users is as yet largely undefined. Importantly, data collection and centralisation needs to be efficient, sustainable and autonomous where possible. This is supported by the recently completed survey (by questionnaire and interview) of Harper Adams Farm Data producers, resources, and users. Key stakeholders were identified and surveyed about current Farm Data assets, their Farm Data needs, whether existing Farm Data assets meet their needs, and their perception of how Farm Data availability and utility can be improved.

The Farm Data stakeholders are using and generating is diverse and ranged from common crop yield and livestock productivity records, to more specialized data such as genetic information and GIS land use assets. Several stakeholders cited the heterogeneous nature or lack of records relevant to carbon storage and land use best practice.

The main recommendations from this work include:

* A need for a Farm Database that is searchable, complete and secure
* Data design should be driven by end user information needs and easy to use
* Responsibility for data capture and design requires specialized skills and should be allocated
* A formal role (within a defined entity/institute)
* A Farm Data capture exercise for a subset of Farm Data assets (e.g. database and

dashboard for yield, inputs and meta-data for specific fields) to acquire feedback from users as stakeholders of Harper Farm Data.

This project will build on the recommendation of the capture exercise for a subset of data and the development of a streaming data pipeline. This is in direct alignment with the objectives of the Data and System Boundaries group’s, which are to organise the collection and inputting of all data into AgreCalc and Sandy, develop a dashboard to make farm data available to all, and define the systems boundaries of the farm Net-Zero. This proposal also directly addresses the priority area of the SSFF Project Funding call on using farm data to enhance decision making.

Laskowski, N. (2016). Data lake governance: A big data do or die. URL: http://searchcio. techtarget.com/feature/Data-lake-governance-A-big-data-do-or-die

# **Specific objectives:**

1. Conduct Farm Data capture best practice exercise for a subset of Farm Data assets based on the previous work conducted by the Data and System Boundaries Group.
2. To develop a data pipeline for efficient inputs and outputs for the selected Farm Data
3. Design and develop an example dashboard to find, visualize, and explore data.
4. Deliver Farm Data workshops to showcase the Farm Data Dashboard and acquire feedback from users as key stakeholders

# **Programme and methodology:**

The methodology will be to capture a subset of the Harper Farm Data (such as from the dairy, livestock, & crop enterprises). A data pipeline will then be implemented, which will start with a data capture process and end with a platform to find, visualize and explore Farm Data in a dashboard environment. The goal is to streamline data input for Harper Farm staff, and also enable academics and students to easily access Farm Data for teaching and research.  This work will assist in identifying the processes required to centralise and automate data collection from the other data streams, as well as assist in developing the potential integration of Harper Farm data with other data systems such the Trinity AgTech’s Sandy platform, John Deere’s data platform, and others, so as to support scalability, data volumes, analytics and potential end-users.

# **Costing and justification of resources: total: £ 7700**



# **Anticipated outputs:**

We will demonstrate a data resource based on selected a Harper Farm Data. This will be used as proof of concept and roadmap to integrate further Farm Data streams into the ‘data lake’ where we will demonstrate analytics queries, analysis and visualisation in a data dashboard, and data exploration and subsetting for teaching and research purposes as well as other Farm Data operations. The process will be documented and presented in conferences as a general model for farm data and information regarding commercial and academic opportunities in agriculture settings. Furthermore, it will support the development of a data management system considering technical and practical challenges for Farm Data platforms. Moreover it will allow skill/knowledge development of stakeholders. The whole process will be documented with the aim to be published in a relevant journal.

Sep 2022 Selection of the data stream

Oct – Dec 2022 Data pipeline implementation

Jan – Mar 2023 Development of Farm Data Dashboard

# **Calls for external funding:**

This work will be used to motivate an external funding bid and will fit within the remit of several funding schemes for data management e.g. Innovate and Farmers Innovation Programme.