This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 773330

Deliverable report for

GAIN

Green Aquaculture Intensification in Europe

Grant Agreement Number 773330

Deliverable D7.5

Title: GAIN Open Data Archive

Due date of deliverable: 31/08/2021

Actual submission date: 31/10/2021

Lead beneficiary: IBM

Authors: Fearghal O'Donncha,

WP 7 – Coordination

Task 7.2 – Data Management Plan

Dissemination Level:		
PP	Restricted to other programme participants (including the Commission Services)	
PU	**Public**	**Y**
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

Document log

Version	Date	Comments	Author(s)
Version 1	01/10/2021	TOC + first draft	Fearghal O'Donncha
Version 2	28/10/2021	Second draft	Fearghal O'Donncha
Version	29/10/2021	Final version	Fearghal O'Donncha

Recommended Citation

O'Donncha F., 2021. GAIN Open Data Archive. Deliverable 7.5. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 11 pp.

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

GLOSSARY OF ACRONYMS

Acronym	Definition
DMP	Data Management Plan
ECMWF	European Centre for Medium-Range Weather Forecasts
FAIR	Findable, Accessible, Interoperable and Re-usable (data)
LSTM	Long short-term memory
RNN	Recurrent Neural Network

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

Table of Contents

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

Executive Summary

This deliverable describes GAIN contributions to the Open Research Data Pilot. This document builds on the data management plan developed and executed in Deliverables 7.3 and 7.4. The focus of this deliverable is how data, code, and innovation developed within GAIN were archived and managed for exploitation and value add by the scientific community and industry. A core guiding principle at all stages was FAIR: making data findable, accessible, interoperable, and re-usable

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

1. Introduction

GAIN participates to the Open Data pilot (DG, 2017; EC, 2021) and therefore strived to share scientific data by making publicly available. Scientific reproducibility and replication were key objectives of the project. Where feasible, we made datasets used for scientific investigation (e.g., publication, white papers, etc.) publicly available. Further, we assessed all data (sensors, experiments, code, innovation) generated during the project – results that were deemed of value to the scientific community were again made publicly available.

This deliverable outlines the data management plan (DMP) adopted for GAIN, in line with the H2020 guidelines for data management plan creation (Commission, 2016). It describes the repository used for data archiving, the infrastructures set in place to allow partners easily distribute, store and archive datasets and protocols on data documentation and meta-description.

The rest of the document is structured as follow: Section 2 presents the process for archiving data considering different datasets, different communication methodologies, and different degrees of commercial sensitivities. Section 3 briefly draws conclusions from the open data pilot and how the data generated during the project may be of use for future work.

2. Archiving of GAIN data

Data archiving within GAIN was composed of four pillars: the first three considered open data, while the fourth considered proprietary or commercial data that could not currently be made publicly available.

The first pillar considered an open data repository. We used the Zenodo repository (https://zenodo.org/) for all datasets generated during the project. The close connection between Zenodo and Horizon 2020 projects made it a natural and easy fit. We created a Zenodo community for the GAIN project curated by IBM. This links directly to the OpenAire page for the GAIN project and provides a concise summary of publications emanating from the project. All GAIN datasets are readily available by searching for the ‘gain_2020’ community in Zenodo. Further, any dataset uploaded to Zenodo is assigned a unique DOI (unless one already exists), which makes referencing (e.g., in publications or technical reports) very convenient.

The second pillar considered code and related assets developed within the GAIN project. Naturally, much of the code developed within the project contained sensitive IP and could not be publicly released. However, wherever possible we committed to open-sourcing code to further impact and allow extensions of our work by the scientific community. We relied on GitHub for all code dissemination due to its ubiquity and its excellent code versioning.

The third pillar considered scientific publications. Naturally, we committed to open access of all GAIN publications. Achievement of this relied on gold and green access protocols but we also committed to making these publications findable and replicable to all extents possible. We strived to make code and datasets connected to publications publicly available (where

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

possible).

The final pillar considered data that could not be made publicly available, at present. While commercial sensitivities precluded making some data publicly available, it is hoped that with time data may be less sensitive but still of significant scientific value. The focus from this perspective was on ensuring data was stored in a consistent format with appropriate meta descriptors to allow long-term storage and retrieval of data.

2.1. Zenodo open data archive

Deliverable 7.4 summarised data collected during the GAIN project (O'Donncha & Pastres, 2019). GAIN collected, generated, and re-used data of different types and from different sources. While data collection continued during the second half of the project the format and context of data did not change. Deliverable 7.3 (Pastres & O'Donncha, 2018) describes the data management plan outlined in early stages of the project and general overview of the types of data expected. While this was updated in Deliverable 7.4 (O'Donncha & Pastres, 2019), we did not observe any deviations from that outlined in the first stage of the project.

The Zenodo archive contains publications, presentations, and datasets connected to the GAIN project. GAIN partners started populating the archive, uploading data collected at three pilot sites, as described below. Details on the monitoring strategy and instrumentation of pilot sites are given in GAIN deliverable D1.1

1. **Data set collected by Sagremarisco at Sagres pilot site (Portugal)** https://zenodo.org/record/5874010
   The data were collected at a mussel farm in Sagres, using Hobo Sensors, a moored bouy and remote sensing data. The dataset includes the following variables: phytoplankton counts, spectrophotometric chlorophyll as well as HPLC phytoplankton pigments.
2. **Data set collected by ZUT at the Nowe Czarnowo pilot site (Poland)** https://zenodo.org/record/5863000
   Data were collected at a land-based common carp farm in Nowe Czarnowo, Poland. The following variables were recorded: water temperature, dissolved oxygen, and pH.
3. **Data set collected by GIFAS at the Rossoya pilot site (Norway)** https://zenodo.org/record/5870744
   The data were collected at one of the Atlantic salmon farming sites managed by GIFAS. The dataset includes: 1) time series of environmental variables, namely water temperature, salinity, and dissolved oxygen 2) spatial time series of hydroacoustic datasets, concerning fish size distribution and biomass. A detailed description oft he dataset is given in "Data Driven Insight Into Fish Behaviour and Their Use for Precision Aquaculture". Included are tools to analyse the data in Julia programming language. More details on the code and analysis is included in our talk at JuliaCon 2021.
4. **Data collected by FEM and UNIVE at the Preore pilot site (Italy)**

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

https://zenodo.org/record/5907365
The data were collected by FEM and UNIVE at a rainbow trout farm located in Preore, Italy. The dataset includes time series of the following environmental and animal variables: i) water temperature, dissolved oxygen, nitrates, conducibility; ii) fish weight distribution. Time series of dissolved oxygen concentration and fish weight forecast obtained using the model developed in GAIN WP1 and embedded in AQUARADAR, see GAIN deliverable D6.6, are also available (Royer et al., 2021).

All external presentations and publications were stored in this repository to provide a central point of access to core innovation from the GAIN project. Two situations were identified as exemptions from requiring relevant data be uploaded to the Zenodo repository:

- Commercial sensitivity: if partners were unable to release data due to commercial sensitivity or specific IP requirements
- Alternative dissemination routes: If partners identified an alternative dissemination route for their data that justifiably provided more valuable impact. This in particular focused on GAIN initiatives to open-source code developed within the project, and to make experiments repeatable by also including the associated datasets. In these cases, we wished to avoid repeating data archival in Zenodo and to guide traffic towards our code repositories.

2.2. Code and innovation

The GAIN project developed innovative technological and experimental datasets. In this section, we present the approach adopted to make our technological developments available to the wider scientific community.

Code – and in particular code connected to scientific publications – were made publicly available where possible. From this regard, we also considered how to make the experiments repeatable by addressing both 1) making data available and 2) how to make the code reusable, including for non-expert users or those not from a data science background.

The GAIN project developed an extensive library of machine learning models to forecast environmental and aquaculture conditions. A core part of this considered the scalable deployment and management of machine learning models. We also considered the development of innovative models to better represent environmental conditions and aquaculture operations. Prominent examples here include a geospatial LSTM (Goodfellow et al., 2016) model to represent the spatiotemporal dependencies in aquaculture farms (Hu et al., 2021), and a machine learning approach to learn relationships between environmental conditions and salmon response in cages (O'Donncha et al., 2021).

The former presents a novel LSTM framework specifically adapted to datasets with a distinct spatial and temporal dependencies. LSTM are part of the family of Recurrent Neural Networks (RNN), and has achieved extremely high performance levels in Deep Learning (Goodfellow et al., 2016). However, the situation in aquaculture is more complex since there are very distinct dependencies in both space and time. Within GAIN, we developed and implemented a novel bidirectional LSTM model that learns in two directions – in space between different sensors and across the time dimension. We applied the model to data from the DAL farm in Canada

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

and to publicly available ADCP data from Norway. To support further usage and extension of the approach, we open-sourced the code here together with subsets of the data for reuse. This can be a valuable contribution towards increased usage of deep learning techniques in aquaculture and environmental science. Indeed, GAIN are participating in a series of workshops related to "AI for Earth Systems Predictability" in November to share their experience of applying machine learning to ocean industry.

The latter paper (O'Donncha et al., 2021), focused on data driven insight into fish behaviour on farms and how machine learning could help decompose the complex dependencies. In this case we developed a data wrangling and machine learning framework and applied to our three pilot study sites cultivating salmon (Norway, Scotland, Canada). We open-sourced the code used to develop the paper results here together with subset of the data. We note the complexity of making public data from commercial farms. Farm operators did not agree to releasing data at two of the three sites. Instead – with agreement from GIFAS pilot site owners – we released the data from that farm and included in the GitHub repo. This allows users to replicate the framework at one of the three sites and more importantly understand the key concepts to allow apply the approach on their own datasets.

Another key consideration of open-sourcing code is to ensure that users can understand the code and replicate. While every effort was made on documentation and code readability, it does assume a basic understanding of data science and machine learning concepts. Of course, aquaculture stakeholders are often experts in aquaculture rather than data science. To help bridge this gap, we connected our open research activities with our professional development initiatives in WP5. Deliverable 5.1 reports on GAIN online courses related to aquaculture professional development. Our course on precision aquaculture is hosted on OpenLearn and includes a detailed overview of developing data driven management processes for aquaculture farms.

The course consists of two main components:

- Material in terms of text and video describing what is precision aquaculture and how it can be developed
- A series of code-based tutorials that covers introduction to data science, fundamentals of machine learning, applied data science, and a demonstration example on the real-world data from the GIFAS farm.

We believe this approach consists of a valuable contribution to the materialisation of precision aquaculture. The code tutorial can be readily accessed (through our guided OpenLearn course or independently) by anyone, and by using a cloud-hosted environment, users avoid the complexity of instantiating their own local environment. Addressing skills development and open research simultaneously can be a pathway to increased uptake of precision farming concepts in the industry.

2.3. Scientific publications and presentations

Throughout the GAIN lifetime, project partners published prodigiously. This arguably reached its zenith at the European Aquaculture Society 2021 meeting where GAIN chaired three sessions and contributed more than 20 submissions. Naturally, we made all publications

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

available through gold or green open access. We also made every effort to:

- extend the reach of our scientific publications through targeted blog posts and social media activities
- enhance the replicability of scientific publications by making data publicly available and ensuring clear documentation of sources of external data.

Zenodo was our main outlet to share data as well as presentations and technical reports created within the project. These allow for a high degree of visibility of our research and allow interested people to reach out to principal investigators for more details on the specific of experiments and potential data sharing agreements.

2.4. Data that could not be made publicly available

Finally, there is the question of archiving data that could not be shared. Naturally when working with real world data from farms, feed companies, and associated stakeholders, commercial sensitivity and IP are paramount. We experienced reluctance from data owners to publicly disseminate data for these reasons. GAIN partners engaged with industry partners to better understand how data sharing in the aquaculture industry could be enhanced (e.g. Bela Buck, WP1 leader chaired a workshop by EATIP, EMODNet, and Copernicus on "Marine Data to support aquaculture in the North Atlantic"). Further, we actively engaged with the AquaCloud initiative to better understand how frameworks and data sharing agreements developed within the Norwegian AquaCloud partnership could be extended to the European aquaculture industry.

A key concept is that while stakeholders might be unwilling to share data currently, that is subject to change. Hence, a key objective of GAIN was to ensure that data that could not be made public, be archived in a secure manner that is also accessible and self-describing for future stakeholders. We relied on DB2 and box.com to host GAIN data. Structured data on farms, environmental data, etc. were uploaded to DB2 and stored with a logical contextual layer to inform on data descriptors. These will be retained after the project. For unstructured datasets such as reports, documents, presentations, etc., we used box.com cloud storage. This allows efficient sharing within the project consortium and allows partners to identify material that they wish to retain after the lifetime of the project.

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330

3. Conclusions

This deliverable describes GAIN contributions to the Open Research Data Pilot. Building on the data management plan described in Deliverables 7.3 and 7.4 (O'Donncha & Pastres, 2019; Pastres & O'Donncha, 2018), we ensured a framework that enabled data sharing. The structures adopted ensured data was "FAIR", while the multi-pronged approach to data sharing considered raw data, code, and innovation. We believe that this approach is a key facilitator towards:

- Reproducibility and replication of scientific research
- Extension of scientific results and uptake by the academic and industrial community
- Progression of the industry towards increased data sharing initiatives.

References:

Commission, E. (2016). *Guidelines on FAIR Data Management in Horizon 2020*. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

DG. (2017). *European Commission Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

EC. (2021). *Open Research Data Pilot of the European Commission*. https://www.openaire.eu/what-is-the-open-research-data-pilot

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press. https://books.google.co.uk/books?hl=en&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&dq=goodfellow+deep+learning&ots=MMR2imlINV&sig=ygji347eXNZRvEcWQsbWlcNuNS8

Hu, Y., O'Donncha, F., & Palmes, P. (2021). A spatio-temporal LSTM model to forecast across multiple temporal and spatial scales. *AAAI 2021*.

O'Donncha, F., & Pastres, R. (2019). *Update of the Data Management Plan. Deliverable 7.4 GAIN - Green Aquaculture INtensification in Europe*.

O'Donncha, F., Stockwell, C. L., Planellas, S. R., Micallef, G., Palmes, P., Webb, C., Filgueira, R., & Grant, J. (2021). Data Driven Insight Into Fish Behaviour and Their Use for Precision Aquaculture. *Frontiers in Animal Science*, *2*. https://doi.org/10.3389/FANIM.2021.695054/PDF

Pastres, R., & O'Donncha, F. (2018). *Data Management Plan. Deliverable 7.3 GAIN - Green Aquaculture INtensification in Europe*.

Royer E., F. Faccenda, R. Pastres. (2021) *Estimating oxygen consumption of rainbow trout (Oncorhynchus mykiss) in a raceway: A Precision Fish Farming approach. Aquacultural Engineering 92 (2021) 102141*

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 773330