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 D7.4
Title: Update of the Data Management Plan

Due date of deliverable: 31/10/2019
Actual submission date: 31/10/2019

Lead beneficiary: UNIVE
Authors: Fearghal O'Donncha, Roberto Pastres
WP: 7

Dissemination Level:		
PU	**Public**	**Y**

Document log

Version	Date	Comments	Author(s)
Version 1	24/10/2018	First draft	Roberto Pastres
Version 2	31/10/2018	Final version	Roberto Pastres, Fearghal O'Donncha

Recommended Citation

O'Donncha F., Pastres R., 2018. Update of the Data Management Plan. Deliverable 7.4. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 15 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)
FCR	Feed Conversion Ratio
RAS	Recirculating Aquaculture System

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

A data management plan (DMP) is required for projects participating in the Horizon 2020 data pilot. The objectives of the DMP is to detail *"what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved"*.

GAIN DMP was structured in accordance with the Guidelines on FAIR Data Management in Horizon 2020 (European Commission, 2016): the plan is presented in (O'Donncha F. & Pastres R., 2018) which was submitted at an early stage of the project, Month 6, as required by GAIN Description of Action.

At the end of the first reporting period, the present document reviews GAIN DMP, in order to identify gaps in its early version and amend the plan, if required. The document includes:

1) a summary of data has been collected, generated and re-used during the first reporting period;

2) an assessment of the strategy implemented for making data findable, accessible, interoperable and re-usable (FAIR)

3) an assessment of the data security measures adopted, including

The main conclusions are:

1) the data collected during the first reporting period fell into the data typologies identified in D7.3 and the strategy and protocols for archiving them proved effective: therefore, there is no need to change them;

2) data processing and dissemination of data/results will take place mainly during Month 18- Month 42: based on the limited experience pertaining the first reporting period, the strategy for making data findable, accessible, interoperable and re-usable (FAIR) seems adequate;

3) thus far, no leaking of personal data and sensitive data has been detected, thus confirming that the security and data protection measures adopted in GAIN proved to be effective.

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

Summary of data collected during the first reporting period

As foreseen in (O'Donncha F. & Pastres R., 2018), during the first reporting period GAIN collected, generated and re-used data of different types and from different sources. In this section we summarise the purpose of the data collection/generation and provide an overview of the data collected thus far, in order to identify deviations from the data type- and formats and archiving protocols planned at an early stage of the project and summarized in (O'Donncha F. & Pastres R., 2018).

Sensor data collected at pilot sites

GAIN collected large volumes of sensor data at the pilot farm sites. The data collection is ongoing and will end at M36 (April 2021). The 10 GAIN Pilot sites are described in detail in (Service, M. et al., 2019): their main features are summarized in Table 1.

Table 1 - Summary of Pilot Sites

Pilot Site	Country	Species	Type of Aquaculture	Partner
1) Dundrum Bay	Northern Ireland - UK	Oysters (*Magellana gigas*)	Shellfish Aquaculture	AFBI
2) Belfast Lough	Northern Ireland - UK	Mussels (*Mytilus edulis*)	Shellfish Aquaculture	AFBI
3) Sagres	Portugal	Mussels (*Mytilus galloprovincialis*)	Shellfish Aquaculture	SGM
4) Rossøya Nord	Norway	Salmon (*Salmo salar*)	Sea cages	GIFAS
5) Carness Bay	Scotland	Salmon (*Salmo salar*)	Sea cages	UoS
6) McNutt's Island, Shelburne	Canada	Salmon (*Salmo salar*)	Sea cages	DAL
7) El Gorguel, Cartagena	Spain	Seabass (*Dicentrarchus labrax*)	Sea cages	LEBCH
8) Preore, Troticoltura Leonardi	Italy	Rainbow trout (*Oncorhyncus mykiss*)	Land-based raceways	UNIVE
9) FES NOWE Czarnowo	Poland	Carp (*Cyprinius carp*)	Land-based RAS/pond	ZUT
10) Fenzhou Village	China	Shrimp (*Litopenaeus vannmei*, *Macrobrachium roserbergii*)	Land-based pond	SCSFRI

The variables collected at each site are summarized in Table 2 and 3: data collection frequencies, f_d, are summarized as follows:

- H : $f_d \geq 1$/hour
- D : 1/day $\leq f_d <$ 1/hour
- W: 1/week $\leq f_d <$ 1/day
- F: 1/2 weeks $\leq f_d <$ 1/week
- M: 1/month $\leq f_d <$ 1/week
- S: fd $\leq$ 1/month and/or as a response to observed anomalies.

  In accordance with the theoretical framework of Precision Fish Farming, these

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

variables are classified as environmental and animal variables.

Table 2. Summary of environmental variables monitored on GAIN pilot sites

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Site 10
Water Temperature	H	H	H	H	H	H	H	H	H/D	D
Salinity	H	H	H	H	D			H		D
Dissolved Oxygen	H	H	H	H	H	H	H	H	H/D	D
pH								D	H/D	D
ORP								D		W
Turbidity	H	H			D					
Chlorophyll a	H	H	H				H			W
Tryptophan										
N-NH4		F						D	W	D
N-NO3		F						H		D
P-SRP		F								
Si-SiO2		F								
POM		F								W
POC										W
TSS		F								W

Table 3. Summary of animal variables monitored on GAIN pilot sites

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Site 10
Size (weight/length) distribution	F	F	M	H	S	S	S	H	W/M	W
Biomass	S	S	S	H	S	S	S	W	W/M	S
Relative biomass distribution in cages				H	H	H				
Feeding activity					H	H				
Fish speed and location						H				
Parasites				M	S	S	S	S		
Welfare indicators				M	M	M				
Mortality	S	S	S	D	W	W	D	D	D	S

The data are fed to the Information Management System, which is being developed in Task 1.4 “Development of a real-time Information Management System” As described in detail in (O’Donncha F., et al., 2019), a protocol for standardizing the data flow from pilot sites to the platform was set and implemented, in order to ensure proper data archiving and interoperability.

The total size of data generated is in accordance with the estimate provided in

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

(O'Donncha F. & Pastres R., 2018), i.e. in the range of Petabytes, as some sensors returns very large datasets: for example, the acoustic system CageEye, implemented in the Canadian and Scottish pilot sites, is generating spatial maps of biomass relative density at hourly intervals. Figure 1 presents an example of CageEye data coming from one of our sites to illustrate the nature of the data. This is complemented by other acoustic sensors (such as ABM) generating large datasets and data that are extracted from external sources (e.g. satellite, weather, numerical model generated).

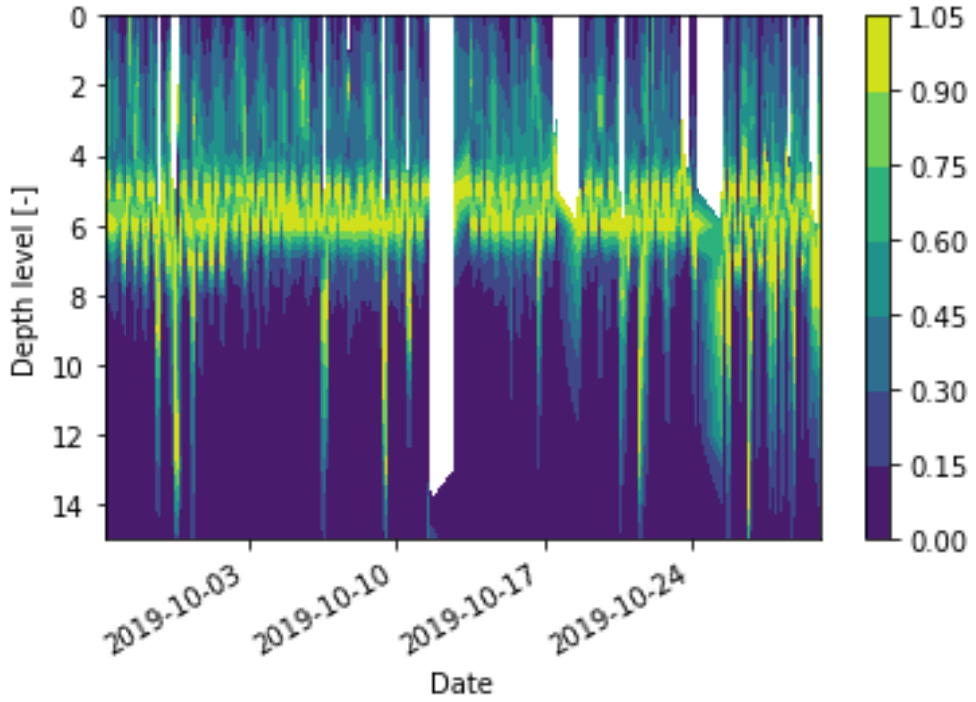

Figure 1: Estimates on vertical fish distribution returned by the CageEye sensor for a representative sample time

Operational data collected at pilot sites

Besides sensor data, additional data and qualitative information concerning husbandry practices were collected, including:

- Production data concerning GAIN pilot sites. (e.g. total biomass, yield, feed conversion ratio).
- Data concerning husbandry practices, (e.g. feeding time, feed rations, feed compositions, parasite presence).
- Data concerning fish welfare (e.g. behaviour, sea louse counts, mortalities, fish growth and condition, fish health parameters).

These data are being processed for developing data-driven and mechanistic models implementing Precision Fish Farming and Precision Shellfish Farming, with the aim of optimising the use of resources and reducing costs. As foreseen in (O'Donncha F. & Pastres R., 2018), data type encompasses both unstructured data (e.g. reporting) and structured data (e.g. biomass data): expected volumes will be in the Gigabytes range per farm.

As an example, at the Rossøya Nord pilot site in Norway, a range of datasets related to both reporting metrics sampled manually by staff, and automated estimates of fish size and distribution generated by the Acoustic Biomass Monitor (ABM). The latter generates an estimate of biomass distribution every five minutes corresponding to about 10Gb of data per month. On the other hand, a range of operational metrics are sampled manually generating

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

much smaller data but in a more unstructured format, Examples of these type of measurements coming from Rossøya Nord are:

- 'Welfare-Score_Condition',
- 'Welfare-Score_Deformity',
- 'Welfare-Score_EyeHealth',
- 'Welfare-Score_FinDamage',
- 'Lice-Average_chalimus',
- 'Lice-Average_matureFemale',
- 'Lice-Average_pre-adult',
- 'Mortality',

Data related to feed design

During the first reporting period, the collection of data concerning feed design, manufacturing with innovative ingredients and testing was planned in detail.
Task 1.1 concerns the identification of microalgae strains which could bioaccumulate Zinc (Zn) and Selenium (Se). These trace elements are very important for fish growth and welfare: small amounts of Zn and Se enriched microalgae in feed could therefore enhance feed performances. Therefore, in Task 1.1 the following data are being collected.

- Variables affecting microalgae growth, e.g. light intensity, water temperature, concentration of macronutrients in the culture medium.
- Concentrations of Zinc and Selenium in microalgae and culture medium.
- Economic data on microalgae production at pilot scale and industrial scale.

The first feed trials are being carried out and preliminary results about feed Key Performance Indicators, see Table 4, are available, as described in GAIN Deliverable D1.4. Data concerning feed pellet features e.g. mixing homogeneity, pellet durability index, pellet hardness, physical water stability, sinking velocity, fat absorption/leaking, starch gelatinization and nutrient leaching.

No marked deviation from the GAIN DoA and D7.3 can be expected at this stage.

Table 4. Key Performance Indicators for assessing novel feeds.

Category	Key performance Indicators	Tissues/samples	Analysis	Species
Performance	Feed intake	Whole body		All
	Weight gain	Whole body		All
	Condition factor	Whole body		All
	Biomarkers[1]	Liver	PCR array	Seabream, salmon
Resource efficiency	Feed conversion (FCR)	Whole body		All
	Digestibility	Feed, Faeces	Macro and micronutrients	All
	Retention efficiency	Whole body	Macro and micronutrients	All
Health & welfare	Mortality			All
	Enteritis	Intestine	Histology	All
	Parasitic infestation	Skin Intestine	*Lepeophtheirus salmonis* *Enteromixum leei*	Salmon Seabream

[1] Application of predictive biomarker models developed in the FP7 ARRAINA-project

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

Category	Key performance Indicators	Tissues/samples	Analysis	Species
	Mucosal function	Skin, gills	Histology[2]	All
	Plasma lysosyme	Plasma	Enzymatic activity	All
	Bactericidal activity	Plasma		All
	Biomarkers[3]	Head kidney	PCR array	Seabream, salmon
Quality	Dressout loss			All
	Fillet yield	Fillet		All
	Pigment	Fillet	Visual and/or chemical	Salmon, trout
	Texture	Fillet	Shear force	All
	Taste	Fillet	Organoleptic (chefs)	All

Data related to aquaculture by-products

The enhancement of circular economy in the aquaculture sector is one of the pillar of GAIN approach to ecological intensification.
Data collected from finfish products included:

- Data concerning potential by-products composition, in terms of specific flesh yields, fatty acid and different fractions (e.g. fillet, head, trimming, viscera, skin, etc.), as detailed in (Malcorps W, et al., 2019).
- Data concerning the yield and composition of marine peptones, protein hydrolysates, oils, minerals, collagen and gelatines obtained from Enzymatic Hydrolysis of by-products, as detailed in (Vazquez J.A., et al., 2019).
- Data concerning yield and safety of by-products from the innovative processes for mortality disposal and RAS wastewater treatment.

These data are being collected from laboratory analysis.

Data collected from shellfish by-products included:

- Data concerning the efficiency of shell-based biofilters for water purification in RAS and Aquaponics and for Phosphorus removal from land-based fish farm effluents.

Data will be used to assess the viability of bivalve shell in several applications, in land-based aquaculture systems and as substitute in construction industry. We expect to generate tens of thousands of data points assessing each valorisation application.

Production and consumption data

To achieve project objectives related to production and consumption of seafood and implications for policy, the GAIN project will collect data related to the amount of seafood consumed for different product/species in different countries. Data will be collected based on information from large retailers, interviews with key operators, questionnaires, ... The data will be collated from a variety of sources and will hence be partly a re-use of data but will generate a much more complete representation of production and consumption data. The data is necessary to inform policy recommendations ensuring the most comprehensive data. We expect the data volume to be relatively small resulting in a few hundred data points per Country.

[2] Non-invasive microbiopsies to map density and volume of mucosal cells (see http://www.quantidoc.com/)
[3] Application of predictive biomarker models developed in the FP7 ARRAINA-project

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

GAIN model data

A range of model data has already been generated during the first reporting period. These include mechanistic models related to fish growth and data-driven models for 1) prediction, 2) feature extraction and 3) decision support platform. These data will be generated during the project and will not be a re-use of existing data. The data will be critical to the precision aquaculture ambition of the project that aims to leverage analytics to better inform farm operations. We expect the data volumes to be in the range of 100s of Gigabytes for individual farms

Existing Geospatial data

GAIN is using of several geospatial datasets during the project, such as

- Meteorological data (e.g. wind speeds, air temperature)
- Ocean model data (e.g. ocean currents, wave heights)
- Satellite data (e.g. sea-surface temperature, Chlorphylla)

GAIN is using these datasets to inform on environmental condition at pilot farm sites. The project intends to re-use these datasets rather than generate and a number of potential sources have been identified (e.g. MODIS satellite data, ECMWF ocean model data). The data will be used as part of the precision aquaculture component in WP1 - Production and Environment.

FAIR data

The steps for making data ***Findable, Accessible, Interoperable, and Reusable*** (FAIR) are described in detail in (O'Donncha F. & Pastres R., 2018). In summary:

Findable

GAIN plans to make project data findable by associating open data sets with scientific publications. Publication venues which utilize Digital Object Identifiers will be favoured for GAIN and all attempts will be made to associate unique DOI with research data (the facility of being able to associate DOI to data is one of the reasons we favoured Zenodo). We have posted preprint of scientific publications in our public repository to ensure availability to the scientific community.

Accessible

Collected data which are non-personal and non-confidential are, at present, made available to project partners through the project hub. Data which is uploaded to this repository are available in an interoperable format, depending on the type and volume of data as well as on domain conventions and standards. Example data formats are: text files of comma separated values and CF convention NetCDF files. We will use the Zenodo repository (zenodo.org) provided by CERN for open datasets generated during the project. Datasets made publicly available will not require restricted access.

We created a Zenodo community for the GAIN project curated by IBM (https://zenodo.org/communities/gain_h2020/). To protect the integrity of the repository, we enforced a number of conditions of use, namely:

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

Only GAIN partners are allowed to upload new data to the community.

Institutions uploading data for this community are responsible to ensure that:

1) they do not upload any sensitive personal data

2) they do not upload regulated data, i.e. medical data, defence/judicial data, export regulated data or any export sensitive data

The Zenodo repository links directly to the OpenAire page for the GAIN project, provides a concise summary of publications emanating from the project and all GAIN datasets will readily accessible by searching for the 'GAIN community in Zenodo. Further, any dataset uploaded to Zenodo is assigned a unique DOI, which makes referencing (e.g. in publications) very convenient.

Table summarises the datasets being generated in GAIN and open-access status.

Table 5: Perspectives on open-access for GAIN datasets

Dataset	**Perspective on open-access**	**Plan to open**
Sensor data collected at pilot sites	Data related to environmental data collected at farm sites is scientifically useful but often commercially sensitive	Ongoing communication with farm owners to separate commercially sensitive data from data that can be released
Data on farm operations	Data related to farm operations are often commercially sensitive	NO
Data related to feed design	Data related to feed design may be commercially exploitable; however, we will explore making open partial sets of the data	Results of trials will be published and data concerning KPI listed in Table 4 will be made available.
Data related to aquaculture by-products	Data may be commercially exploitable, but we will explore making open partial sets of the data	Results concerning optimal enzymatic hydrolysis conditions, and main parameters leading to cost effective drying of RAS wastewater and mortalities will be published in peer reviewed papers.
Production and consumption data	Some production and consumption data may be commercially sensitive (retailer data), but we will explore making open, possibly with suitable anonymisation and aggregation	As stated in GAIN Deliverable 8.1, confidential information on production and consumption will be collected only upon informed consent of the data owners. Informed consent forms will clearly state that the data will be used only after anonymisation and

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

Dataset	Perspective on open-access	Plan to open
		aggregation
GAIN model data	Data related to environmental and operational insights on farm production are often commercially sensitive	Ongoing communication with farm owners to separate commercially sensitive data from data that can be released.
Existing geospatial data	These are re-used data which are either already openly available or which the project does not have the right to publish.	NO

Interoperability of data

As foreseen in (O'Donncha F. & Pastres R., 2018), interoperability of GAIN data has been accomplished through standardised data and appropriate meta-description and documentation. We plan to accompany open data with a scientific article or technical white paper to promote reuse. GAIN aims to promote semantic mapping and relevant ontologies for aquaculture farms to promote interoperability and increase access to individuals from outside domains (particularly from data science fields where the focus is much more on the data itself rather than domain characteristics). We have made public the code to upload and download data to our cloud service to ensure that the approaches we use are transparent and visible to others (note that for security reasons, user credentials are required to actually upload or download data).

Data re-use

Re-use of data has been encouraged through clear licensing, prompt dissemination, archiving, and quality assurance. At this writing, license for GAIN data has not yet been selected. We note the preference expressed in the OpenAIRE Licensing Study (Dietrich et. al. 2014) for version 4.0 of the Creative Commons Licenses. Dissemination of GAIN data will promptly follow publication of related scientific papers. We plan to disseminate data once papers have been accepted for publication. Data published from GAIN will be disseminated through Zenodo facilitating re-use of the data by third parties for an indefinite period after the project is over. Data published from the project will undergo a quality assurance process managed by the individuals responsible for the scientific publication

Data Security

The security of data collected and generated during GAIN is important to the success of the project and essential to providing subsequent open access to selected project data. In this regard, care has been taken to facilitate data recovery, provide adequate data storage, through the project hub, and enable transfer of sensitive data. Partners, however are responsible for transferring the data to the project hub and Information Management System for real-time management of aquaculture data. The data recovery strategy for this

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

cloud service platform relies on the recovery infrastructure of the cloud environment where it is hosted: see (O'Donncha F., et al., 2019) for more details. Where it is necessary to transfer sensitive data, care will be taken to use encrypted connections. Taking these measures in relevant activities contributes to data security on the GAIN project. Thus far, the operational protocols adopted in GAIN has ensured that no data leaking has occurred.

Ethical aspects

The ethics review for the GAIN project raised the issues of:

- informed consent for the collection, storage, and protection of personal data.
- In case personal data being transferred from/to a non-EU country or international organisation, confirmation that this complies with national and EU legislation, together with the necessary authorisations.

These issues are treated in (Pastres R., & Licata C., 2018).

Conclusions

This deliverable reviews the data management plan for the GAIN project presented in (O'Donncha F. & Pastres R., 2018), in order to identify order to identify gaps in its early version and amend the plan, if required.

The main conclusions are:

1) the data collected during the first reporting period fell into the data typologies identified in (O'Donncha F. & Pastres R., 2018) and the strategy and protocols for archiving them proved effective: therefore, there is no need to change them;

2) data processing and dissemination of data/results will take place mainly during Month 18- Month 42: based on the limited experience pertaining the first reporting period, the strategy for making data findable, accessible, interoperable and re-usable (FAIR) seems adequate;

3) thus far, no leaking of personal data and sensitive data has been detected, thus confirming that the security and data protection measures adopted in GAIN proved to be effective.

References and useful links

European Commission (2016), Directorate-General for Research, H2020 programme guidelines on FAIR data management in Horizon 2020, 2016 Jul.

Dietrich, Nils et al. (2014) Study on licensing of publications and research data: Summary of findings. https://www.openaire.eu/openaire-licensing-study.

Malcorps W, Newton R, Little D., 2019. Report on value chain mapping for key species/systems, with SWOT analysis of key informants. Deliverable 3.3. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 57 pp.

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

O'Donncha F., Gormally, R., Akhriev, A., Palmes, P., 2019. Information Management System: first release. Deliverable 1.5. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 25 pp.

O'Donncha F., Pastres R., 2018. Data Management Plan. Deliverable 7.3. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 77330. 13 pp.

Pastres R., Licata C., 2018. H Requirement N° 1. - Deliverable 8.1. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 21 pp.

Pastres R., Licata C., 2018. POPD Requirement N° 2 - Deliverable 8.2. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 13 pp.

Service, M.; Grant, J., Icely, J.D., Lopez, M., Moore, H., Micallef, G., Panicz, R., Pastres, R., Rey Planellas, S , Zhu, C., 2019. Report on Instrumentation of GAIN pilot sites. Deliverable 1.1. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 77330. 59 pp.

Vazquez J.A., Pérez-Martín, R.I., Méndez D., Sotelo, C.G., 2019. Fish by-products from aquaculture and fish discards as feed ingredients. Deliverable 2.3. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 38 pp.

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

List of Tables

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