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

Grant Agreement Number 773330

Deliverable D7.3
Title: Data Management Plan

Due date of deliverable: 31/10/2018
Actual submission date: 31/10/2018

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	17/10/2018	Table of contents	Fearghal O'Donncha
Version 2	23/10/2018	First draft	Roberto Pastres
Version 3	28/10/2018	Second draft	Fearghal O'Donncha
Version 4	31/10/2018	Final version	Roberto Pastres

Recommended Citation

O'Donncha F., Pastres R., 2018. Data Management Plan. Deliverable 7.3. GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 773330. 13 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

Introduction

This document fulfils deliverable 7.3 by providing information on the data management policy for GAIN project. This 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"*.

Guidance from the H2020 Programme indicates that a data management plan should be submitted early in the project and subsequently revised as the project matures and more information becomes available. Thus, it is not expected that a first version of this plan will provide complete detail on all aspects of data management. Accordingly, updates to this document are expected in line with interim project reviews.

The remainder of the document is structured following the Guidelines on FAIR Data Management in Horizon 2020 (European Commission, 2016). Section 1 provides a summary of data collected, generated and re-used on the project. Section 2 discusses making data findable, accessible, interoperable and re-usable (FAIR). Section 3 describes the allocation of resources for making data FAIR. Section 4 considers data security. Section 5 notes ethical aspects related to data management. Section 6 concludes the report.

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

Data summary

The GAIN Consortium includes the 20 partners listed in Table 1 and an International Partner, namely NOAA, National Ocean and Atmospheric Administration (US), which cooperates: as Third Party of the Coordinator, UNIVE: non-EU partners are marked in bold.

Table 1. GAIN Consortium

Participant Nº (leadership role)	Participant legal name	Country	Type
1 (Coordinator; WP5; WP7)	Universita Ca' Foscari Venezia (UNIVE)	Italy	RTD
2 (WP3)	The University of Stirling (UoS)	UK	RTD
3 (WP1)	Alfred-Wegener-Institut Helmholtz- Zentrum für Polar- und Meeresforschung (AWI)	Germany	RTD
4	IBM Ireland Limited (IBM)	Ireland	CORP[1]
5 (WP2)	Agencia Estatal Consejo Superior de Investigaciones Cientificas (CSIC)	Spain	RTD
6 (WP4)	Longline Environment Limited (LLE)	Ireland	SME
7 (WP6)	Sparos Lda (SPAROS)	Portugal	SME
8	Salten Havbrukspark (SHP)	Norway	SME
9	Wageningen University (WU)	Netherlands	RTD
10	Johann Heinrich von Thuenen-Institut, Bundesforschungsinstitut Fuer Laendliche Raeume, Wald Und Fischerei (TI)	Germany	RTD
11	Agrifood and Biosciences Institute (AFBI)	UK	RTD
12	Zachodniopomorski Uniwersytet Technologiczny W Szczecinie (ZUT)	Poland	RTD
13	Asociacion Nacional de Fabricantes de Conservas de Pescados y Mariscos-Centro Tecnico Nacional de Conservacion de Productos de la Pesca (ANFACO)	Spain	NPO[2]
14	Multivector AS (MV)	Norway	SME
15	Gildeskal Forskningsstasjon AS (GIFAS)	Norway	SME
16	Lebeche (LEBCH)	Spain	CORP[1]
17	Sagremarisco-Viveiros de Marisco Lda (SGM)	Portugal	SME
18	Fondazione Edmund Mach (FEM)	Italy	NPO[2]
19	Dalhousie University (DAL)	Canada	RTD
20	South China Sea Fisheries Research Institute (SCSFRI)	China	RTD

GAIN is structured in 7 Work Packages, plus an Ethics Work Package, which was added by the EC during the negotiation (see Fig. 1). WP leaders are indicated in Table 1. The main objects of each WP are listed below.

WP1 - Production and Environment: will develop novel sustainable feeds and tools for enhancing aquaculture sustainable management of aquafarm based on Big Data analytics.

WP2 - Secondary products: will develop new co-products, in order to enhance circularity, sustainability and profitability of aquaculture supply chains.

WP3 - Policy and markets. will analyse the state-of-the-art of EU and national legislations with respect to the valorisation and marketing of innovative GAIN products and co-products and provide suggestions to policy makers.

WP4 - Eco-intensification: will develop new approaches and tools for assessing the level of eco-intensification of GAIN innovative solutions, in comparison with standard practices.

WP5 - Professional development: will deliver both on-line and in presence courses, in order to facilitate the adoption of GAIN innovative solutions by aquafarm operators.

WP6 - Dissemination, Exploitation, Communication: will maximize GAIN impact, by careful matching communication&dissemination tools to targeted audiences and developing platforms for exploiting GAIN results beyond its life time.

WP7 - Coordination: will ensure the timely delivery of all GAIN contractual items.

[1] Corporation (Not SME)

[2] Non-profit Organisation

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

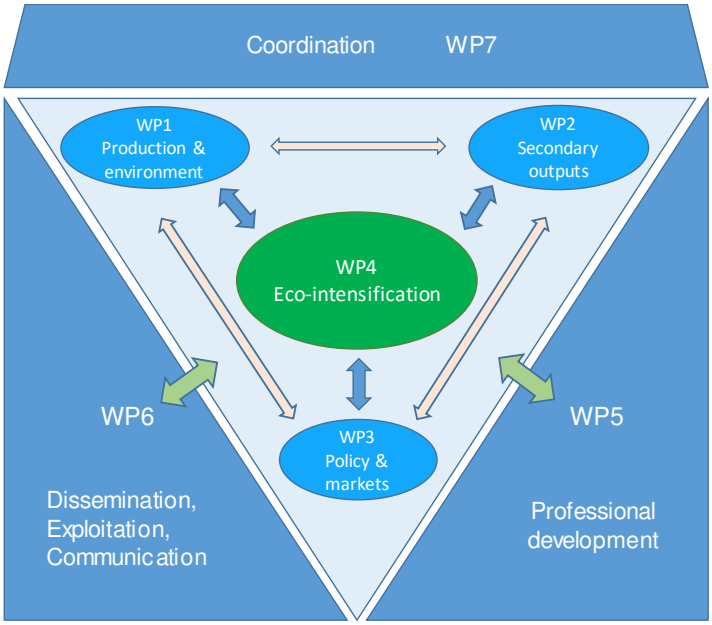

Fig. 1. GAIN structure

GAIN collects, generates and re-uses data of different types and from different sources to accomplish the project objectives. In this section we summarise: the purpose of the data collection/generation, data type-and formats, origin of the data and expected volumes.

Sensor data collected at pilot sites

GAIN will collect large volumes of sensor data at the pilot farm sites. These sites include:

- a salmon farm at GIFAS production site of Rossøya, Norway;
- a salmon farm belonging to Cooke Aquaculture, which joined the GAIN Consortium as committed end-user.
- a salmon farm in Nova Scotia, Canada
- LEBCH seabass and seabream farms located near Murcia, Spain
- a rainbow trout farms, i.e. Troticoltura Leonardi, located in Trentino Alto-Adige, Northern Italy, which joined the GAIN Consortium as committed end-user.
- SGM mussel farm, located in the Algarve, Portugal.
- a set of mussel farms located in Northern Ireland, UK, which are currently been monitored by AFBI.

Data collected will include:

- time series of environmental variables, e.g. water temperature, dissolved oxygen, chlorophyll a concentration, salinity, current, etc.
- Acoustic monitoring data providing information on biomass, movement patterns, etc.
- Imagery data from in-situ camera and drone imagery.

The collection of this data is necessary to enable 'precision aquaculture' type analytics and better inform on farm condition, environmental status and time evolution. This data originates at the pilot farm sites participating in the project and is not a re-use of existing data. The data will be critical to the Information Management System developed during the project and of significant scientific value to the academic partners in the project. The total size of data generated is expected to be in the range of Petabytes and varies according to the type. For example, the acoustic monitoring system returns very large datasets generating 2D spatial

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

maps at five second frequency, while the time series datasets return hourly data at five pilot sites producing about 43,000 data points per year for each variable collected.

Operational data collected at pilot sites

We will collect a range of operational datasets at the pilot sites. Data collected will include:

- 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).

The collection of this data is necessary to enable 'precision aquaculture' type analytics and understand how external factors influence farm productivity and effects. This data originates at the pilot farm sites participating in the project and is not a re-use of existing data. The data will be critical to developing data-driven modelling related to feature extraction, predictive analytics and decision support. Further, the data is of significant scientific value to the academic partners in the project. Data type will encompass both unstructured data (e.g. reporting) and structured data (e.g. biomass data) and expected volumes will be in the Gigabytes per farm.

Data related to feed design

We will collect a range of data to guide feed design also re-using data collected at pilot farm sites (described above). Data collected will include:

- Microalgae species and strains, based on literature and specific experiments carried out in GAIN.
- Environmental variables affecting microalgae growth.
- Concentrations of Zn and Selenium in microalgae and culture medium.
- Economic data on microalgae production at pilot scale and industrial scale.

The collection of this data is necessary to explore the viability of using algae as feed components. Further data will be collected related to the selected feed formulations including:

- 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.
- Data related to feed trials testing the performance and FCR of the selected formulation.

These datasets are critical to developing novel feed formulation in WP1.

Data concerning feed pellets will be determined during specific tests at its premises by SPAROS, a GAIN partner with strong expertise on feed formulation and manufacturing. Data related to feed originates from a set of feed trials on four selected species, namely Atlantic salmon, rainbow trout, seabream, and turbot as a niche species, in at least 8 feeding trials performed by SPAROS (seabream), AWI (turbot), FEM (trout) and salmon (GIFAS). Key Performance Indicators determined in these trial are summarized in Table 2.

The expected volume of data related to feed design is in the range of 100,000 data points.

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

Table 2. Key Performance Indicators determined in GAIN feed trials.

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[3]	***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***
	Mucosal function	***Skin, gills***	***Histology[4]***	***All***
	Plasma lysosyme	***Plasma***	***Enzymatic activity***	***All***
	Bactericidal activity	***Plasma***		***All***
	Biomarkers[5]	***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

We will collect a range of data related to aquaculture by-products. Data collected from finfish products will include:

- 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.),
- Data concerning the yield and composition of marine peptones, protein hydrolysates, oils, minerals, collagen and gelatines obtained from Enzymatic Hydrolysis of by-products.

Data collected will be critical to allow an analysis of potential valorisation, and an assessment of increase of the edible proportion and waste reduction. It will be used to determine environmental and economic benefits of redirecting by-product fractions to innovative eco-intensification. Data will be collected from laboratory analysis during the project and is not a re-use of existing data.

Data collected from shellfish by-products will include

- Data concerning the efficiency of shell-based materials for water purification in RAS, biofilters in Aquaponics and Phosphorus removal from land-based fish farm effluents.
- Data about the efficiency of fillers for the cement industry based on shells.

Data collected will be required to achieve project objectives related to the valorisation of shellfish by-product. It will be used to assess the viability of bivalve shell in several applications, namely as a biofilter in land-based aquaculture systems and as substitute in construction industry. Data will be collected from laboratory and prototype testing during the

[3] Application of predictive biomarker models developed in the FP7 ARRAINA-project
[4] Non-invasive microbiopsies to map density and volume of mucosal cells (see http://www.quantidoc.com/)
[5] 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

project and is not a re-use of existing data. 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.

GAIN model data

A range of model data will be generated during the GAIN project. 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 intends to make use of several geospatial datasets during the project. Specifically, we will use the following:

- Weather 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 will make use of 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.

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

FAIR data

This section details how GAIN will make data ***Findable, Accessible, Interoperable, and Reusable*** (FAIR). Each topic is addressed in turn.

Making data findable, including provision for metadata

Following H2020 guidelines, two types of datasets are considered here:

1. **The 'underlying data' –** data and metadata related to scientific publication generated during the project
2. **Any other data –** for instance curated data not directly attributable to a publication, or raw data

GAIN plans to make project data findable by associating open data sets with scientific publications. Publications may include traditional research papers as well as "data only" articles which present and summarize data, but do not offer in-depth analysis or draw conclusions. A scientific paper about a dataset is an ideal forum to provide metadata, discuss naming conventions, point to standards, and list relevant keywords. 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 also.

Making data openly accessible

Data will be made openly available by uploading to an appropriate repository. Data which is uploaded to a repository will use an interoperable format with the exact format dependent on the type and volume of data together with domain conventions and standards. Example data formats will be 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.
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, 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 3: 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;	?

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
	however we will explore making open partial sets of the data	
Data related to aquaculture by-products	Data may be commercially exploitable but we will explore making open partial sets of the data	?
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 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

Interoperability of GAIN data will be accomplished through standardised data formats (and we aim to promote standardisation of data formats through GAIN) 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).

Data re-use

Re-use of data will be 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

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

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

Allocation of resources

The project budget allocates resources to make GAIN data FAIR. Tasks in the project which are related to open data include

- Task 1.2 related to developing novel feed components
- Task 1.3 Instrumentation of commercial aquaculture facilities
- Task 1.4 Development of a real-time Information Management System
- Task 3.2 Assessment of production and consumption data and implications for policy

Once data has been made FAIR during the course of the project and placed in a repository, further costs are not anticipated for the project team. Benefits of long term preservation include historical benchmarking and comparison activities.

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 will be taken to facilitate data recovery, provide adequate data storage and enable transfer of sensitive data as required by the project.

Responsibility for data recovery falls to individual software and hardware components of the GAIN system. Much of the data considered for open access will move through the cloud service platform for real-time management of aquaculture data. The data recovery strategy for cloud service platform will rely on the recovery infrastructure of the cloud environment where it is hosted. Box provides a secure storage environment for data in the cloud. 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.

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 project Deliverables 8.1 and 8.2.

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

Conclusions

This deliverable describes the data management plan for the GAIN project and our participation in the Open Data pilot. The range of data being collected, generated and used as part of the project are described; data typology and volume are presented together with the utility both towards the GAIN project and for the wider scientific community. A detailed description of open-access procedures are provided together with an assessment of the datasets that can be made open either partially or in totality. A final decision on providing any particular dataset as open data will be documented in a future revision of this report.

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.

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