Alaska Hatchery Research Program **Technical Document:**[1] 10

Title: AHRP Data Flow **Version:** 1.0
Authors: Timothy R. Frawley, Eric Lardizabal, and Scott Johnson
Date: August 9, 2016

Abstract

The Alaska Hatchery Research Program is designed to answer questions regarding concerns that hatchery fish released by private non-profit corporations in Prince William Sound (pink and chum salmon) and in Southeast Alaska (chum salmon) may have a detrimental impact on the productivity and sustainability of natural stocks. The study that was designed to answer these questions requires that data collected by a contractor and by various Alaska Department of Fish and Game laboratories be combined to test hypotheses. Collecting, storing, correcting, archiving, retrieving, and ensuring quality control and reliability of these data are critical to the success of this project. This Technical Document describes the flow of data between the contractor who collects the specimens and records the sampling event data and the various Alaska Department of Fish and Game offices and laboratories responsible for producing results from processing the specimens collected and to explain how each step of the process works and describe why it is necessary.

Background of AHRP

Extensive ocean-ranching salmon aquaculture is practiced in Alaska by private non-profit corporations (PNP) to enhance common property fisheries. Most of the approximately 1.7B juvenile salmon that PNP hatcheries release annually are pink salmon in Prince William Sound (PWS) and chum salmon in Southeast Alaska (SEAK; Vercessi 2015). The large scale of these hatchery programs has raised concerns among some that hatchery fish may have a detrimental impact on the productivity and sustainability of natural stocks. Others maintain that the potential for positive effects exists. To address these concerns Alaska Department of Fish and Game (ADF&G) convened a Science Panel for the Alaska Hatchery Research Program (AHRP) whose members have broad experience in salmon enhancement, management, and natural and hatchery fish interactions. The AHRP was tasked with answering three priority questions:

I. *What is the genetic stock structure of pink and chum salmon in each region (PWS and SEAK)*?

[1] This document serves as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and other members of the Science Panel of the Alaska Hatchery Research Program. As such, these documents serve diverse ad hoc information purposes and may contain basic, uninterpreted data. The contents of this document have not been subjected to review and should not be cited or distributed without the permission of the authors or the Commercial Fisheries Division.

II. *What is the extent and annual variability in straying of hatchery pink salmon in PWS and chum salmon in PWS and SEAK*?

III. *What is the impact on fitness (productivity) of natural pink and chum salmon stocks due to straying of hatchery pink and chum salmon*?

Introduction

Specimens collected for this project include tissue samples, otoliths and scales. Where tissue and otolith specimens are paired with result data, they are placed into the same well (cell) of a uniquely numbered 48 deep-well plate (tray). When only otoliths are collected, they are placed into cells of uniquely numbered 96 shallow-well trays. Scales are collected and stored separately on gum cards and carefully aligned with the tray and cells containing tissue and otolith samples using a mobile based application created and maintained by the Sitka Sound Science Center (SSSC). This Tray/Cell combination creates a unique relationship wherein sample and result data may be combined into a single dataset for analysis. Because data generated by the Alaska Hatchery Research Program exist in multiple systems with differing data structures and business rules, maintenance of these relationships between databases is critical to the success of data integration.

To answer the priority questions, sample-specific data collected by the contractor and ADF&G laboratories must be stored and organized into a system that meets the following objectives:

1) The ability to upload data from all contributing parties
2) A repository for the most up-to-date data
3) The ability to correct data as errors are discovered
4) Archive versions of records that were edited
5) The ability to retrieve paired data from multiple parties for analysis
6) Reliability including automated backup

The data flow diagrams and narratives herein document the flow of data from the contractor to each ADF&G system and finally into the Division of Commercial Fisheries (DCF) Data Warehouse where it is linked into OceanAK, the DFC's statewide reporting solution.

This document describes the flow of data between the SSSC, who collects the specimens and records the sampling event data, and the various Alaska Department of Fish and Game offices and laboratories responsible for producing results from processing the specimens collected and explains how each step of the process works and why it is necessary. This document also serves as a technical reference for future maintenance.

AHRP data flow process: Component Integration

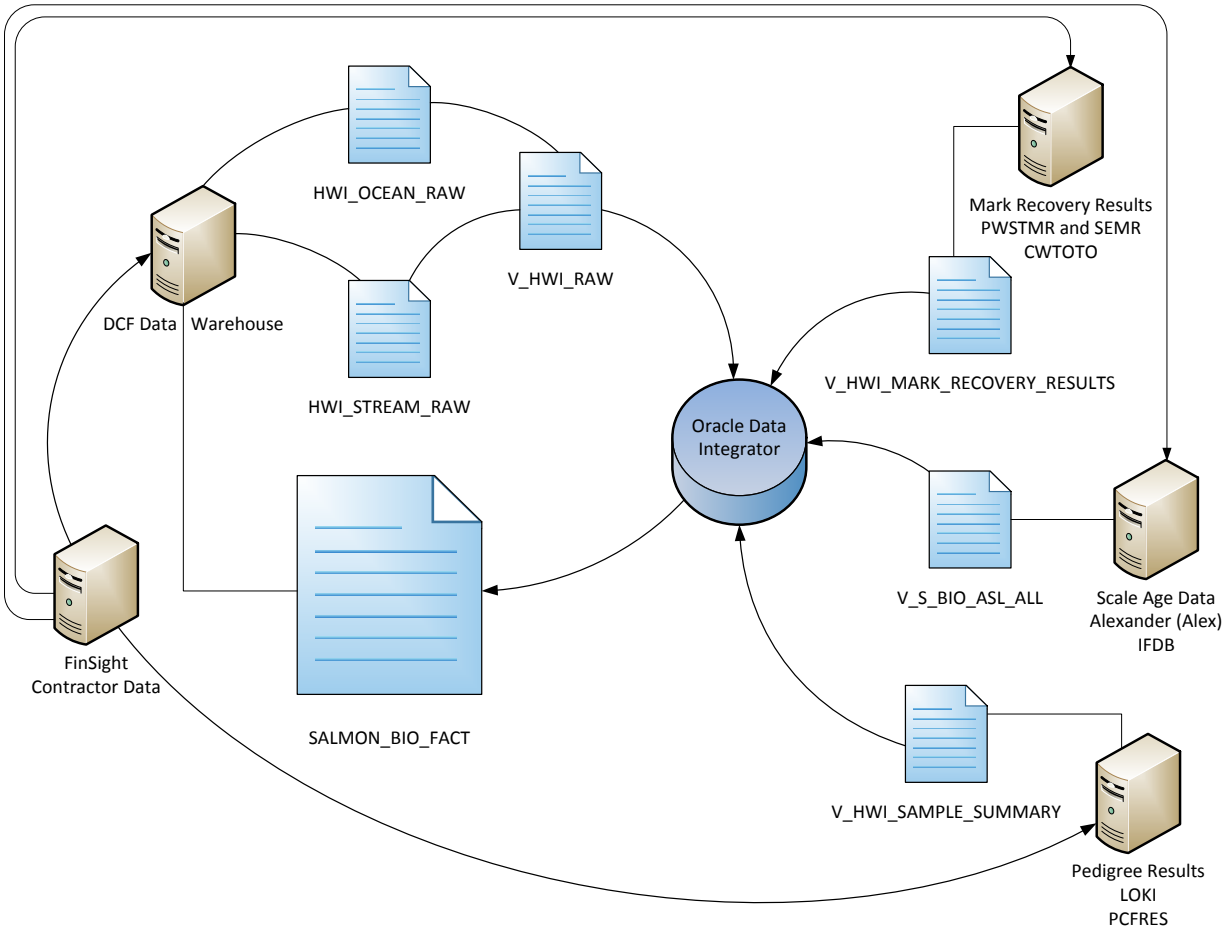

Figure 1. Diagram of the data flow among components of the Alaska Hatchery Research Program. Each server icon represents a database containing data related to this project. Blue page icons represent database objects where data resides in the attached server icon. The circle icon represents the Oracle Data Integrator processes that combine all data into SALMON_BIO_FACT, the ultimate destination for all the data for this project. See text for descriptions of objects.

The following is a description of the objects named in Figure 1. Match the object in Figure 1with a bold heading below and follow the lines indicating where objects reside. Arrows indicate the direction of data flow.

FinSight:

The contractor collects data using tablet devices. These data are uploaded into the FinSight database. ADF&G downloads data manually using the contractor's reporting engine (http://keta.finsight-ak.com/HatcheryWild/), emailed Excel or comma separated value (CSV) files the contractor creates, or via web service.

- **FinSight to Pedigree Results arrow:** The Gene Conservation Lab (GCL) loads Stream Sampling data from the FinSight database into an Oracle table called HWI_SAMPLING (not shown in the diagram above). The Oracle View, V_HWI_SAMPLE_SUMMARY, formats Sample data into the Salmon BIO Data Repository format, filters for Pedigree trays only. It will also include Pedigree result data once the best way to represent those results has been determined.
- **FinSight to DCF Data Warehouse arrow:** The Mark, Tag, and Age Laboratory (MTA Lab) loads Stream Sampling data from the FinSight database into an Oracle table called HWI_STREAM_RAW and Ocean Sampling data from the FinSight database into an Oracle table called HWI_OCEAN_RAW. See Oracle Data Integrator (ODI) description below for more details.
- **FinSight to CWTOTO and IFDB arrows:** Contractor collected Ocean and Stream Sample data are transferred to the MTA Lab, Cordova Otolith Lab, and the Southeast Alaska (Region 1) office via emailed Excel or CSV files, and hand written tray labels. This information is data entered and/or imported into each system. More information is provided in the description of each database icon below.

DCF Data Warehouse:

ADF&G, DCF Data Repository. This warehouse is a repository of data related to the AHRP, as well as other DCF projects.

HWI_OCEAN_RAW and HWI_STREAM_RAW:

DCF Data Warehouse Oracle tables containing all raw Ocean and Stream Sampling data downloaded from the contractor's FinSight database.

V_HWI_RAW:

DCF Data Warehouse Oracle View that combines and formats Ocean and Stream Raw data. These data are the basis for all contractor provided sample meta data that exist in SALMON_BIO_FACT. Data are filtered by maximum "last modified date" value.

Oracle Data Integrator (ODI)

ODI is a process by which data are combined from multiple datasets for importation to another location. This process combines Ocean and Stream Sample data from V_HWI_RAW with Pedigree Sample data and results from V_HWI_SAMPLE_SUMMARY, Otolith Mark Recovery results in V_HWI_MARK_RECOVERY_RESULTS and Scale Age data from V_S_BIO_ASL_ALL, populating it in SALMON_BIO_FACT. Data are related using Sample Year, Tray and Cell Number values.

Contractor Raw Data: Contractor raw data are downloaded once per week from the FinSight database, updating the HWI_STREAM_RAW and HWI_OCEAN_RAW Oracle tables. Rows in either table are replaced where the downloaded Last Modified Date is greater than the existing Last Modified Date by Year, Tray, and Cell number. Old records and deleted records are moved to HWI_STREAM_RAW_ARCHIVE and HWI_OCEAN_RAW_ARCHIVE respectively (ARCHIVE tables not shown). Archived records are never deleted or overwritten with newer updates.

Pedigree Sample Data: V_HWI_RAW provides all of the Tray/Cells by Year as collected and reported by the contractor. Sampling meta data, such as sample date, stream code, lat./long., etc. are retrieved from V_HWI_SAMPLE_SUMMARY and only from V_HWI_RAW when any given Tray/Cell combination does not exist therein.

All other: V_HWI_RAW is the source of all sample meta data for Ocean and Stream Otolith Mark Recovery and Scale Age data. Result data such as MARK_ID, MARK_PRESENT, MARK_STATUS_CODE are populated solely from V_HWI_MARK_RECOVERY_RESULTS and CARD_NUMBER, FISH_NUMBER, FW_AGE, SW_AGE and AGE_ERROR_CODE are populated solely from V_S_BIO_ASL_ALL. It is important to note that CARD_NUMBER and FISH_NUMBER are populated from V_S_BIO_ALL.

SALMON_BIO_FACT:

ADF&G, DCF Data Warehouse, Statewide Salmon Biological Data Repository Oracle table. This table currently contains salmon age, sex, length (ASL) data for Alaska as well as for this project. To access information related specifically to AHRP, the user must filter for project data using “BATCH_NUMBER like ‘HWI%’” or “BATCH_NUMBER in (‘HWI-OCEAN’,’HWI-OTOLITH’,’HWI-PEDIGREE’)”.

V_HWI_MARK_RECOVERY_RESULTS:

MTA Lab, Coded Wire Tag and Otolith Recovery database (CWTOTO) Oracle view that combines Cordova Otolith Lab and MTA Lab otolith marked recovery data related to this project. Data are formatted based on the Salmon Bio Repository specification.

Mark Recovery Results, PWSTMR and SEMR, CWTOTO:
This icon represents the MTA Lab's Oracle database. Ocean, Stream and Pedigree Sampling data for this project are received from the contractor on labels attached to Otolith trays and/or a spreadsheet or hardcopy inventory of otolith trays. The Otolith tray label by itself provides enough information for data entry; however, an electronic tray inventory is preferable. A Tray Inventory is necessary for data entry of Pedigree Sample data as the deep well plate labels do not provide enough information for data entry. Tray data entry occurs manually due to the nature of the tray data creation process in each otolith lab's data entry software and the business rules that manage it. After tray data entry, technicians run an Access-based report which duplicates the Tray Inventory spreadsheet provided by the contractor for side by side comparison. Any discrepancies are resolved that this point. As otoliths are processed, Otolith Mark Recovery data appears in V_HWI_MARK_RECOVERY_RESULTS.

V_S_BIO_ASL:
Region 1 Integrated Fisheries Database (IFDB), Salmon Age, Sex, Length Oracle view. Data includes all salmon ASL data collected by Region 1 and the Scale Age data related to this project by filtering the view where "BATCH_NUMBER like 'wildchum%.csv'". Data are formatted based on the Salmon Bio Repository specification.

Scale Age Data, Alexander (Alex), IFDB:
This icon represents the Region 1 Integrated Fisheries Database (IFDB). A small portion of Pedigree Stream Sampling data has been paired with the collection of salmon scales that are examined to estimate the age of the specimen. Initial sample data are received from the contractor in the form of a CSV file (wildchumXXXX.csv, where XXXX represents the sample year) containing enough sample information for importation into the IFDB system so that the scales may be "read" by scale aging technicians. Age data appear in V_S_BIO_ASL_ALL as each specimen is read.

V_HWI_SAMPLE_SUMMARY:
GCL LOKI database Oracle view includes contractor provided Pedigree Stream Sample and Pedigree Result data. Data are formatted based on the Salmon Bio Repository specification.

Pedigree Results, LOKI, PCFRES:
This icon represents the GCL's Oracle database, Production Commercial Fisheries Resource (PCFRES), aka LOKI. Stream sampling data are downloaded in Excel format from the Finsight web portal and uploaded to an Oracle table HWI_SAMPLING on the database instance PCFRES using Oracle SQL Developer. Using a custom Adobe AIR program named LOKI, technicians use the stream sample data to organize the tissue samples by location and create collections. The collections in LOKI store individual sample data that include sample Tray ID, Well ID, and an internal individual number assignment. Once the collections are created, the individual sample data and collection location data are joined to the stream sample data in a view called

V_HWI_SAMPLE_SUMMARY. Sample Tray IDs and locations are then integrated into SALMON_BIO_FACT. The framework to handle DNA extraction and genotyping results is now in place. The project lead retrieves otolith data from the warehouse and links them to genotype data via Tray IDs so pedigree analysis can begin. The parental results will be stored in an Oracle table in LOKI that will be integrated into the data warehouse through V_HWI_SAMPLE_SUMMARY.

AHRP data flow process: Contractor to DCF data warehouse details

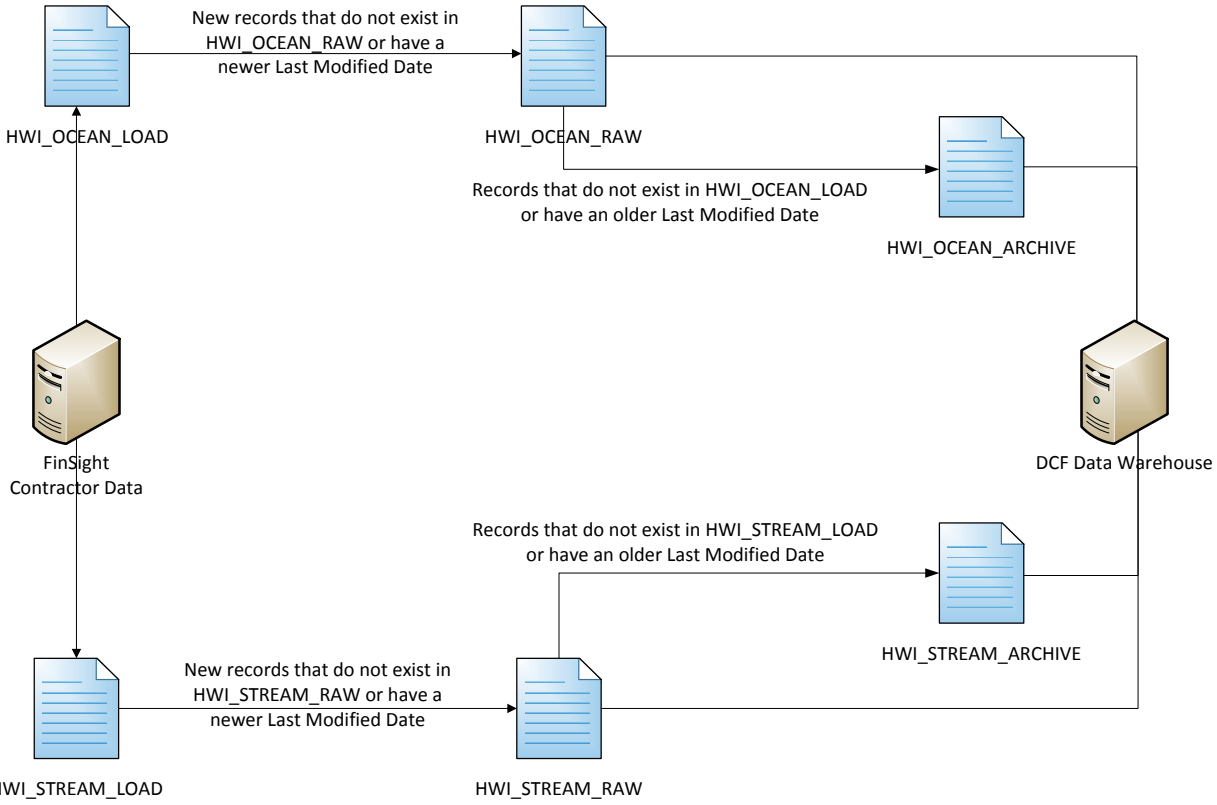

Figure 2. Diagram illustrating the flow of data flow from the contractor's database "FinSight" to the Division of Commercial Fisheries (DCF) Data Warehouse. Each server icon represents a database containing data related to this project. Blue page icons represent database objects where data resides in the attached server icon.

The following is a description of the objects shown in Figure 2. Match the object named in Figure 2. Lines indicate where objects reside and arrows indicate the direction of data flow. Contractor collected data reside in the FinSight database and are available for download from the contractor's reporting engine. These data need to be transferred into the DCF Data Warehouse

for integration with Pedigree, Otolith Mark Recovery and Scale Age data. Figure 2 depicts the concept behind this data transfer.

In order to maintain the most current set of Contractor data, the MTA Lab has developed an application to download CSV report files from the FinSight reporting engine that are stored locally on the automation server's (dfgjnudcf-adm1.dfg.alaska.local) file system. These files are then parsed into memory and loaded into HWI_OCEAN_LOAD and HWI_STREAM_LOAD respectively. Records in the LOAD tables (HWI_OCEAN_LOAD and HWI_STREAM_LOAD) are compared to the existing RAW data tables (HWI_OCEAN_RAW and HW_STREAM_RAW) by Year, Tray and Cell values. The ARCHIVE tables (HWI_OCEAN_ARCHIVE and HWI_STREAM_ARCHIVE) are used to house older versions or deleted records of HWI Raw data. Records that exist in RAW but not in LOAD are inserted into the ARCHIVE table, then deleted from the RAW table. Records in LOAD that exist in RAW with a newer Last Modified Date are inserted into ARCHIVE from RAW, then deleted from RAW, and finally the new record from LOAD is inserted into RAW. The LOAD tables are replaced every time the custom application successfully downloads a new CSV data file. Archived records are never deleted or overwritten by newer versions. A control table, HWI_DATA_CONTROL (Appendix Table 1) is used to specify the Year and report type to be downloaded when the custom application executes. Data Direction value IN is used when specifying the Year and report type to be downloaded from the contractor by the custom application. Data Direction value OUT is used in conjunction with reporting in OceanAK or for reports downloadable by the contractor.

Questions for the AHRP Science Panel

Are the objectives for handling AHRP data appropriate?

Will these processes meet the objectives?

AHRP Science Panel Review and Comments

This technical document has been reviewed.

This document covers the data flow process for this research. There were no comments from the AHRG.

This document is acceptable to the AHRG.

References

Vercessi, L. 2015. Alaska Salmon Fisheries Enhancement Program 2014 Annual Report. Alaska Department of Fish and Game, Anchorage. http://www.sf.adfg.state.ak.us/FedAidPDFs/FMR15-15.pdf

Appendices

Appendix Table 1.

Used to specify the Year and report type to be downloaded when the custom application executes. Data Direction value IN is used when specifying the Year and report type to be downloaded from the contractor by the custom application. Data Direction value OUT is used in conjunction with reporting in OceanAK or for reports downloadable by the contractor.

HWI_DATA_CONTROL		
SAMPLE_YEAR	**COLLECTION_TYPE**	**DATA_DIRECTION**
2013	OCEAN	IN
2013	STREAM	IN
2014	OCEAN	IN
2014	STREAM	IN
2013	HWI-OCEAN	OUT
2013	HWI-PEDIGREE	OUT
2013	HWI-OTOLITH	OUT