California Recreational Fisheries Survey (CRFS) Infrastructure for Analysis and Reporting Tools

FY 2016 Proposal

Connie Ryan

Created: 09/28/2015

1. Overview

1.1. Sponsor

Steve Williams and Ed Hibsch

1.2. Focus Group

Information Management

1.3. Background

CRFS Data Collection:
The marine waters off California's 1,100 miles of coastline are home to a diverse array of fish species. Many of these species are targeted by recreational anglers who take more than four million fishing trips annually in California's marine waters. The California Department of Fish and Wildlife's (CDFW) California Recreational Fisheries Survey (CRFS) collects data on these trips and the resulting catch using eight types of field surveys and a telephone survey. Each year CRFS staff complete about 6,500 assignments, interview over 60,000 angling parties in the field, identify over 200,000 fish to species and measure about 120,000 fish. In addition, about 26,000 licensed anglers are interviewed via a telephone survey. CDFW uses these data to estimate catch and effort by month, mode, six geographic regions, water area and eight trip types.

CRFS Data System:
CDFW maintains the CRFS data and estimates in a fully relational Microsoft SQL Server Enterprise database which is an effective tool for collecting and storing the CRFS data. In addition to supporting the collection of and estimation from the survey data, the database also aids in making monthly assignment draws, importing data from a telephone survey, and providing data to assist the CRFS staff in tracking assignments and performing data quality checks. Currently, there are approximately one hundred data tables in the CRFS database. In addition, there are over forty tables that provide an audit history by automatically generating a backup record whenever data is added, edited, or deleted in the CRFS primary data tables.

While numerous aspects of a relational model contribute to accurate and reliable data, a relational database can be difficult to use for analysis and reporting. The traditional approach for improving analysis and reporting capabilities has been to copy the data out of the relational model and into a read-only data warehouse, which uses a star-schema model or a tabular model instead of a relational model.

Pacific RecFIN:
Pacific Recreational Fisheries Information Network (Pacific RecFIN) is the repository for Marine Recreational Fisheries Statistics Survey (MRFSS) data and estimates for the west coast and for the state surveys that replaced MRFSS. California, Oregon and Washington all collect recreational fishing data using different survey methods and each state stores their data in different database formats. Pacific RecFIN is responsible for compiling these disparate data sets into one uniform format and making the data and estimates available to the public for regional analyses. Pacific RecFIN is currently moving from a SAS-based database to Microsoft SQL Server database and plans to develop a data repository for these data that provides advanced analytic capabilities.

The relational database models used by CDFW and Pacific RecFIN will be different, due to their different purposes. CDFW maintains data associated with making draws, completing assignments and producing estimates; Pacific RecFIN compiles data from disparate data stores and provides a unified repository. In addition, each state will have their own reporting tools, with decisions about those tools being made based on a wide range of data sets – not just the recreational fishery data.

CRFS Data Dictionary:
The data dictionary that will be created during the 2015 MRIP IT project "California Recreational Fisheries Survey (CRFS) Data Access" will be used extensively in this project. The elements of the data dictionary will be embedded into the transformed data to provide user-friendly field names, to create online help documentation, and to provide documentation for the current transformation processes.

Fisheries Data Exchange Standards:
The 2015 American Fisheries Society Annual Meeting had a session titled "Managing Data to Meet Shrinking Budgets and Growing Needs". Included in that session were topics such as "Fisheries Data Exchange Standards are Important to AFS" and "Developing a National Fisheries Data Exchange Standard." What was interesting about this session was how the discussions were not about what type of database is used for storage, or about what type of tool is used for reporting, but instead, about how to standardize the names, descriptions and methods for sharing of data. Although, much of the work has been with inland fisheries, there were discussions about marine fisheries exchange formats. The data exchange among the western states and Pacific RecFIN represents an opportunity to explore the feasibility of implementing a standardized exchange similar to what is being proposed with the National Fisheries Data Exchange Standard (NFDX). At a minimum, we must be able to provide a cross-walk between the state's relational data and the Pacific RecFIN data warehouse (probably in a cube format) repository.

California Needs Not Addressed by Pacific RecFIN Data Migration Project:

CDFW transmits data and preliminary estimates to RecFIN monthly and periodically provides finalized data and estimates. Currently the monthly transmissions consist of 21 “flat files” compatible with the current Pacific RecFIN format. At some point, CDFW will need to provide the data in formats useable by the new Pacific RecFIN relational database and/or in formats consistent with their new data repository. In addition, CRFS will need to produce automated validation scripts that confirm the data has been transmitted correctly.

In addition to maintaining the catch and effort data and estimations, the CRFS database must handle the sample draw process, the assignment tracking process, and the unique aspects of each of the estimation processes. There are numerous tables that support all of these processes. Pacific RecFIN does not maintain the survey level data from the sample draw, assignment tracking and individual assignments. These data are important for use in resource damage assessments, in various models and analyses conducted in support of resource management, in budgeting, in proper weighting of samples, and in estimating variance. Should estimation procedures change at a later date, these underlying survey data will be needed. This project will give scientists and managers tools for using these data and enhance understanding to the CRFS data collection program.

Pacific RecFIN maintains estimates at the mode level (e.g., private and rental boat) to enhance regional comparisons. CRFS maintains the estimates at the level of the survey type (e.g., survey of primary private and rental boat sites, angler telephone directory telephone survey) and in the original estimation domains. These survey-level estimates have proven critical in regulatory analyses, stock assessments and in understanding local fisheries. The data repositories and tools developed by this project will enhance access and usability of these estimates.

1.4. Project Description

This project will create the infrastructure required for the integration of the California Recreational Fisheries Survey (CRFS) data into analysis and reporting tools (also known as business intelligence and analytic applications), and into a standardized exchange format for transmission to the Pacific Recreational Fisheries Information Network (Pacific RecFIN) and other governmental agencies. A number of powerful commercial business intelligence applications are available that, with the proper infrastructure, allow users to quickly summarize data, conduct multidimensional analyses, and quickly navigate around the data to find trends and to “drill down” through the layers of data and estimates. In addition, this project will also allow California Department of Fish and Wildlife (CDFW) to automate, document, and validate the submission of CRFS data to be incorporated in the new Pacific RecFIN’s relational model or repository.

The proposed project will create a library of transformations required to move data from a relational model into a star-schema or tabular model, and into data exchange formats. These transformations will result in data structures that will comprise the reporting and exchange repositories for CDFW. The CDFW staff will likely use commercially available business intelligence and analytic applications such as MS Power Pivot and MS Power Query to access its data repositories. Note, that the data model incorporated into Power Pivot (in-memory, tabular model), is capable of reading XML formatted data, which is the most commonly used data exchange format.

Whether based on a star schema model, a tabular model or an xml-data-exchange model, the function of the read-only data repository is to provide data users with tools to easily and quickly summarize data and conduct analyses. The benefits of a read-only data repository for analysis, reporting and validation as compared to a relational database include:

(a) Easy to understand and use: The structure is easy for end users and applications to understand and navigate the data.
(b) Aggregation speed: The query response time is fast since the structure allows for rapid aggregations (e.g., counts, sums, averages).
(c) Multidimensional analyses: The types of analyses that can be conducted depend on the “front end” tool, but the structure allows for easy filtering and grouping of data, “drilling down” to investigate the underlying source data or aggregations and for trend analyses.
(d) Built-in referential integrity: Only those records with the correct key values will be available for specified analyses.
(e) Security: End users can be given varying levels of access to data so that confidential data is kept secure.

The library of transformations created by this project should be useful to Pacific RecFIN and possibly other agencies. Examples of the types of transformations and typical analyses that the system can support include:

(a) The movement of data from one model (system) to another,
(b) The draw for each mode, district and month and all of the supporting analyses for the draw (e.g., prior years’ pressure at each site),
(c) Assignment tracking,
(d) Data validation and verification (e.g., identifying length and weight outliers, over limits, species geographic outliers),
(e) Providing data extracts in support of fishery management decisions.

To enhance the ability of others to use or adapt the transformations, the final report will contain a description of each transformation in the library. This description will include: a list of data elements from the relational model, an explanation of the transformation (including the code), and a list of the resulting tables and elements. In addition, for each transformation we will describe the purpose, and if possible give an example for the type of analysis that would be performed using the transformed data.

This project aligns with the MRIP Information Management focus areas of Data Access and Analytical Tools. The project will facilitate the creation of a data repository that allows fishery managers, Pacific Fishery Management Council teams and researchers to use commercially available analytical tools to better understand recreational fisheries and to better understand the CRFS estimates and underlying data. It will assist the users in finding trends, spotting patterns and quickly "slicing and dicing" the data in ways that are useful for improving understanding. The project will also improve access, usability and validation of the California's recreational fisheries data by creating a data exchange standard.

1.5. Public Description

1.6. Objectives

(1) Improve transmission of California data and estimates to Pacific RecFIN by developing a data exchange standard in coordination with Pacific RecFIN and creating automated validations for the data transmission.
(2) Facilitate the use and understanding of California's recreational fisheries data by creating an infrastructure that will allow data users to use various analysis tools to quickly summarize and compare data, to conduct multidimensional analyses and to identify patterns and trends.
(3) Create a library of transformations for use by CDFW, Pacific RecFIN and possibly other state agencies, that presents fisheries data in a way that is conducive to various analytical investigations.

1.7. References

2. Methodology

2.1. Methodology

Throughout the project CDFW will work closely with Pacific RecFIN in developing the data exchange standard, and in creating the automated validations for the data transmissions.

The initial tasks of the project will include:

(a) Define a data exchange format for the new Pacific RecFIN repository.
(b) Identify a set of analyses that are needed to validate the data being sent to Pacific RecFIN.
(b) Identify analyses of interest to data users (e.g., fishery managers, decision-makers, stock assessors).
(c) Determine whether a star schema or tabular model best meets the requirements of the data users.

Once the data repository models are identified, Extract-Transform-Load (ETL) scripts and packages in Microsoft Integration Services will be created. These transformations prepare the data for multidimensional cubes, in-memory-tabular models and data exchange formats. Also, data validation and data comparison scripts could be created to help verify the exchange of data.

The following major tasks will be undertaken to create infrastructure for analysis and reporting tools:

(a) Identify and design the dimension and fact tables or the tabular model structure appropriate for the designated analyses.
(b) Develop the transformations required for each to move the data from the transactional, relational database into the new dimensional or tabular database.
(c) Document and test the transformations.

An Application Software Specialist (consultant) will be hired for nine months to work directly with the CRFS Programmer/Analyst in the CDFW Data and Technology Division (DTD) in Sacramento, California. The current CRFS Programmer/Analyst will guide and monitor all work by the consultant on a daily basis. The final decision on all schema, tabular structures and elements of the transformations will be made by CDFW-DTD staff, and the consultant will transmit schema, tabular structures and elements to DTD staff bi-weekly for their approval.

2.2. Region

Pacific

2.3. Geographic Coverage

California

2.4. Temporal Coverage

Not a data collection project

2.5. Frequency

Not a data collection project

2.6. Unit of Analysis

Not a data collection project

2.7. Collection Mode

Not a data collection project

3. Communication

3.1. Internal Communication

COORDINATION WITH PACIFIC RECFIN:
The Pacific RecFIN programmer/analyst will be an integral member of the project team. The Team Leader (the CRFS programmer/analyst) will work closely with the Pacific RecFIN programmer/analyst to ensure there is no duplication of effort between this project and the Pacific RecFIN data migration project and to share information and examples of transformations.

COMMUNICATION AND PROJECT TRACKING:
The Team Leader and Application Software Specialist (who will be hired specifically for this project) will work in the same office and will communicate daily, and the Team Leader will track progress weekly. Progress and blockages will be addressed in the CDFW CRFS data team weekly check-in calls.

The entire team for the CRFS Infrastructure for Analysis and Report Tools Project will have periodic project calls. The Team Leader and Application Software Specialist will contact other team members for input as needed via e-mail and telephone calls.

SHARING AND DISTRIBUTING INFORMATION AND PRODUCTS:
The primary means of distribution will be through e-mail. Files that are too large to share via e-mail will be placed on an ftp site.

3.2. External Communication

A monthly report will be submitted to the MRIP Operations Team using the MRIP reporting system. During the project, regular updates will be provided to the Pacific RecFIN Technical Committee and the project's sponsor. A final report will be posted on the MRIP reporting system and distributed to the Pacific RecFIN Technical Committee.

4. Assumptions/Constraints

4.1. New Data Collection

N

4.2. Is funding needed for this project?

Y

4.3. Funding Vehicle

Pacific RecFIN Grant awarded to Pacific States Marine Fisheries Commission

4.4. Data Resources

CRFS database/ Microsoft SQL Server and comprehensive data dictionary will be fully available to the Application Software Specialist who is hired with the funds for this project.

4.5. Other Resources

CDFW will provide staff to lead and assist the Application Software Specialist, and will provide the required office space. The computer purchased as part of the 2015 MRIP grant would be used by the person hired for this project.

4.6. Regulations

No new regulations required.

4.7. Other

(1) Assume that we will be able to find an Application Software Specialist (programmer) with the requisite skills willing to work on a short-term project (about nine months) in Sacramento, California.

(2) Assume funding will be available no later than August 1, 2016.

5. Final Deliverables

5.1. Additional Reports

5.2. New Data Set(s)

5.3. New System(s)

6. Project Leadership

6.1. Project Leader and Members

First Name	Last Name	Title	Role	Organization	Email	Phone 1	Phone 2
Ed	Hibsch	RecFIN Programmer /Analyst	Team Member	Pacific States Marine Fisheries Commission	ehibsch@psmfc.org	503-595-3100	
Kevin	Hitchcock	Research Analyst II	Team Member	California Department of Fish and Wildlife	kevin.hitchcock@wildlife.ca.gov	707-576-2865	
Jeanne	Rimpo	CRFS Programmer /Analyst	Team Leader	California Department of Fish and Wildlife	jeanne.rimpo@wildlife.ca.gov	916-327-8767	
Connie	Ryan	Senior Environmental Scientist	Team Member	California Department of Fish and Wildlife	connie.ryan@wildlife.ca.gov	650-631-2536	
Ashok	Sadrozinski	Environmental Scientist	Team Member	California Department of Fish and Wildlife	ashok.sadrozinski@wildlife.ca.gov	650-631-2535	
TBD	TBD	Application Software Specialist hired with MRIP funds for this project	Team Member	Pacific States Marine Fisheries Commission			

7. Project Estimates

7.1. Project Schedule

Task #	Schedule Description	Prerequisite	Schedule Start Date	Schedule Finish Date	Milestone
5	design dimension and fact tables or tabular model structure, develop transformations, document and	1,2,3,4	01/02/2017	07/14/2017	Y

Task #	Schedule Description	Prerequisite	Schedule Start Date	Schedule Finish Date	Milestone
	test transformations				
1	Develop position description, identify interview team, and develop interview questions		07/01/2016	08/31/2016	
2	Advertise position for Application Software Specialist, interview candidates, make selection	1	09/01/2016	09/30/2016	
3	Hire Application Software Specialist and begin project	1,2	10/03/2016	10/21/2016	
4	define data exchange format, identify analyses, determine the model that best meets the data users (e.g., star schema, tabular)	1,2,3	10/24/2016	12/30/2016	Y
6	Write and submit final report	1,2,3,4,5	07/17/2017	09/29/2017	Y

7.2. Cost Estimates

Cost Name	Cost Description	Cost Amount	Date Needed
overhead	overhead at 11.82%	$8362.00	08/01/2016
salary	9 months salary for Application Software Specialist	$60516.00	08/01/2016
benefits	9 months of benefits for Application Software Specialist	$10232.00	08/01/2016
TOTAL COST		$79110.00	

8. Risk

8.1. Project Risk

Risk Description	Risk Impact	Risk Probability	Risk Mitigation Approach
Funds for project not available by August 1, 2016	Would not start the project until funds were available and the project timeline would need to be adjusted.	Medium	Delay the project start data and adjust the timeline accordingly.

Risk Description	Risk Impact	Risk Probability	Risk Mitigation Approach
A project team member is unavailable for a significant period of time (e.g., a month).	The impact would depend on the expertise of the team member. In the worst case scenario, the project might be delayed.	Low	If the project member is the only person with the needed expertise, then the time line would be adjusted. If another team member has the expertise, then that person would work on the tasks. If someone outside the team has the needed expertise and is available, then add that person to the team.
Delay in the completion of the data dictionary for the CRFS database [work supported by 2015 MRIP Grant for CRFS Data Access Project].	Could delay the start of the project.	Medium	Start working on task #4 (define data exchange format, identify analyses, determine the model that best meets the data users).
Not able to hire an Application Software Specialist with sufficient expertise.	The project could be delayed until an applicant with the required skills is found, or it could limit the amount of work that would be completed.	Low	If an Application Software Specialist with sufficient expertise cannote be hired, we would train a less qualified individual.

9. Supporting Documents