A relational database for California salmon data

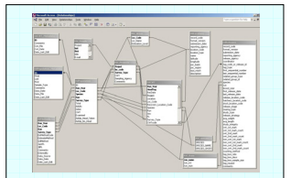
Jason Azat

Anadromous Resources Assessment, Fisheries Branch, California Department of Fish and Game

Abstract

Two persistent problems facing California salmon researchers are difficulty sharing study data, and data with little documentation. A centralized relational database (RDB) can help solve these problems. I describe an RDB tailored to anadromous species projects such as hatchery releases, screw trapping, mark recapture surveys, and Coded Wire Tag recoveries.

An RDB is a set of relationships among a set of tables. These referencing relationships make it possible to store extensive documentation down to the record level, while maintaining fast and simple data interactions.



* RDB relationships Graphic Design Interface.

Introduction

Tehama Colusa Fish Facility	
Coleman National Fish Hatchery	
Sacramento River:	RBDD, Carcass, Redds, Anglers
Clear Creek:	Carcass, Snorkel
Cow Creek:	Video
Bear Creek:	Video
Cottonwood Creek:	Snorkel, Video
Battle Creek:	Ladder, Snorkel, Video
Antelope Creek:	Snorkel, Video, Redds
Mill Creek:	Carcass, Video, Redds
Deer Creek:	Carcass, Snorkel, Redds
Big Chico Creek:	Snorkel
Butte Creek:	Carcass, Snorkel
Feather River:	Carcass, Anglers
Feather River Hatchery	
Yuba River:	Carcass
American River:	Carcass, Anglers
Nimbus Fish Hatchery	
Mokelumne River:	Carcass, Redds, Anglers
Mokelumne River Fish Hatchery	
San Joaquin River:	Carcass
Cosumnes River:	Carcass, Redds
Stanislaus River:	Carcass
Tuolumne River:	Carcass
Merced River:	Carcass
Merced River Fish Facility	

An example of the many areas and survey methods throughout the California Central Valley.

Problem

* Difficult to compile all data for analysis.
* Difficult to track multiple estimate methods for the same time/areas.
* Difficult to document all study methods.
* Difficult to maintain version control.

Solution

* An RDB stores all data in central location for analysis.
* An RDB tracks all estimate types for all time/areas.
* An RDB documents study methods of each estimate.
* Data products programmatically display RDB source data.
  * Because these data products do not modify or store data:
    * Most current data.
    * Most reliable data complete with documentation.
* SQL easier to learn, faster to write, easier to debug, faster to run than conventional implementations such as Matlab, Visual Basic.

SQL

```
Select year, location, count(*)
From Survey
Where year = 2009
Group By location
```

Matlab

```
surveys=[[list of surveys]]
[m,n]=size(surveys)
for I=1:n
  bad=0
  good=0
  str=['c:\data\salmon\'num2str(surveys(I)),'.txt']
  ans=load(str)
  good=length(ans(ans(:,2009)=='hatchery',:))
  bad=length(ans(ans(:,2009)=='river',:))
  result(I,1)=surveys(I)
  result(I,2)=length(ans)
  result(I,3)=good
  result(I,4)=bad
end
```

Visual Basic

```
Dim intResult As Integer
Dim intRatCount As Integer
Dim intSirCount As Integer

Sheets("surveys").Select
Range("C2").Select
Do While ActiveCell.Value <> ""
    If ActiveCell.Value = 2009 Then
        If ActiveCell.Offset(0, -1).Value = "hatchery" Then
            intRatCount = intRatCount + 1
        ElseIf ActiveCell.Offset(0, -1).Value = "river" Then
            intSirCount = intSirCount + 1
        End If
    End If
ActiveCell.Offset(1, 0).Select
Loop
intResult = MsgBox("2009 Hatchery Count = " & intRatCount & Chr(13) & "2009 River Count = " & intSirCount & "")
```

* SQL is easier to learn, faster to write, and easier to debug.
* Comparison of SQL, Matlab, and Visual Basic code implementations, all producing identical results.

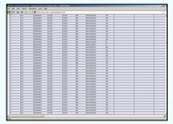
Methods

What was achieved:

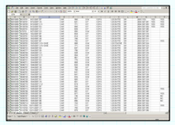
Data compiled

* Population estimate documentation provided by California Department of Fish and Game.

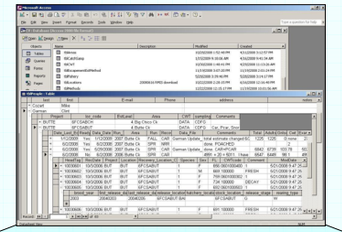
* Historic California salmon population data imported from GrandTab flat file storage.



* California salmon release data imported from Record Mark Information System database.



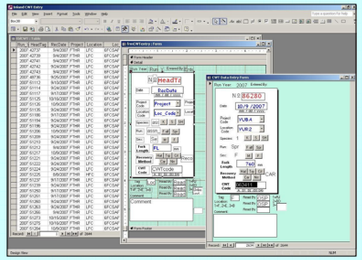
* Coded wire tag recovery data imported from six California coded wire tag reading labs.



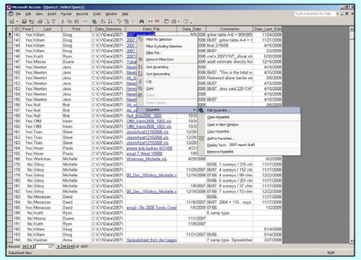
* Population estimate documentation provided by cooperating agencies.

* Source Data collected from many Contributors.
* Database design requirements defined.
* Database Structure designed to store data, document data, and meet output needs.
* Relational Database built in MS Access, California Fish and Game's standard database application.

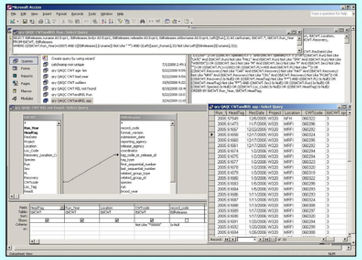
Database infrastructure built



* Data entry forms implemented using MS Access.



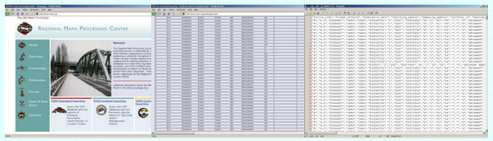
* Electronic source documentation library begun.



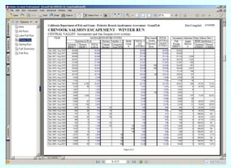
* QAQC queries implemented using Structured Query Language and Graphic User Interface.

Results

Data Products



* New Coded wire tag release and recovery data sent to RMIS database, available for public download.
* http://www.rmpc.org/



* Population data published on Calfish website, available for public download.
* http://www.calfish.org/IndependentDatasets/CDFGFisheriesBranch/tabid/157/Default.aspx

Discussion

What is proposed

* Centralized coded wire tag lab.
* Online data entry.
* Dynamic data reporting and analysis.

Necessary resources

* Construct coded wire tag lab.
* Purchase, install, design, program, implement server.
* Hire supervisor, biologists, database administrator, web programmer, statistician, and lab technicians.

References

ATLAS (A Toroidal LHC ApparatuS) Experiment. 2006. Information systems and data acquisition for ATLAS. CERN (European Organization for Nuclear Research), Available: http://sim.fc.ul.pt/en/images/e/e7/FCT-ATLAS-2006.ppt

Bradley D, Mascaro M, and Santhakumar S. 2005. A relational database for trial-based behavioral experiments. J Neuroscience Methods. 141:75-82

Cotter, G. 2006. Biological Informatics: Challenges of data integration. USGS (U.S. Geological Survey), Available: www.nbii.gov/images/uploaded/8496_1179747727020_The_Challenge_of_Data_Integration.pdf.

FGDC (Federal Geographic Data Committee). 1999. Content standard for digital geospatial metadata part 1: biological data profile. USGS (U.S. Geological Survey) Report FGDC-STD-001.1-1999. Reston, VA.

Irvine, L., Kleczkowski A., Lane A., Pitchford J., Caffrey D., and Chamberlain P. 2006. An integrated data resource for modelling the soil ecosystem. Applied Soil Ecology 33:208–219.

Jacob, C., Kant, A., Benson, B., Newton, R., and K.D. McMahon. 2005. Biological databases for linking large microbial and environmental datasets. Proc. 9th World Multiconference on Systematics, Cybernetics and Informatics Vol VII: Informatics Systems Development II. Orlando, FL.

Rice, M., Gladstone, W., and Weir, M. 2004. Relational databases: A transparent framework for encouraging biology students to think informatically. Cell Biology Education 3:241–252

Stitt, S. 1999. The federal geographic data committee's biological data profile. USGS (U.S. Geological Survey), Available: http://www.nbii.gov/images/uploaded/151871_1166715705793_NBII_bdp_graphical.doc.

Sucaet, Y., Van Hemert, J., Tucker, B., and Bartholomay, L. 2008. A web-based relational database for monitoring and analyzing mosquito population dynamics. J. Med. Entomol. 45(4):775-784

White, D., Wolf, D., Porter, D., Sanger, D., Riekerk, G., DiDonato, G., Holland, A., and Dabney, D. 2008. Development of a data management framework in support of southeastern tidal creek research. Environ. Monit. Assess. 150:323-331

Acknowledgements

The following play an important part in developing the salmon data systems presented here: Dr. Robert Titus, Melodie Palmer-Zwahlen, Joseph Duran, Brenda Erwin, Don Pearson, Alice Low, Stan Allen, Douglas Killam, Connie Shannon, Robert Kano.

The following dedicated scientists and hatchery personnel provided the data to populate the database: Bob Anderson, Steve Boyd, Matt Brown, Mike Brown, Kari Burr, Clint Garman, Michelle Gilroy, Kyle Hartwigsen, Colleen Harvey-Arrison, Mike Healey, Paula Hover, Anna Kastner, Trevor Kennedy, Duane Massa, Tracy McReynolds, David Menasian, Jess Newton, Kevin Niemela, Bob Null, Kevin Offill, Colin Purdy, Bill Smith, Steve Tsao, Robert Vincik, Michelle Workman.