Ten Lessons Learned: Data Warehouse Development Project, California Department of Fish and Game

Crilly Butler, Jr.
California Department of Fish and Game

Despite an abundance of data in numerous internal and external systems, the California Department of Fish and Game struggled over the years to assemble this information into coherent and meaningful representations of its core business. A highly successful data warehouse project has largely resolved this problem, yet the development process had to face and overcome numerous obstacles over the span of several years before all the important lessons were learned. The project was hindered by all the usual culprits: heterogeneous data systems, lack of consistent data standards, limited funds, lack of internal expertise, competing political priorities, and rigid organizational boundaries. Unexpected hurdles included the sudden and serious illness of a key contractor, internal reorganization of business units, and a statewide economic downturn. Each hurdle was met and surmounted in both creative and practical ways, yet widely available within most organizations.

"Water, water everywhere, nor any drop to drink…"

Famous lines by Samuel Taylor Coleridge from *The Rime of the Ancient Mariner*. Could not the same thing be said about most organizations where, despite an abundance of data, staff often go thirsty for information? The California Department of Fish and Game (DFG) has struggled for years to generate comprehensive reports which accurately portray important aspects of its core business. Historically, such reports have had to be cobbled together manually from several data stores – a process that can occupy staff for days or weeks at a time. The recent development and implementation of a large data warehouse with a number of targeted business areas has successfully addressed this problem, yet the sailing was neither smooth nor uneventful. The project sat moored in the harbor for years, got off to several false starts, then sailed around in circles before finally getting its bearings and reaching port. During the project course, DFG was faced with numerous obstacles, both internal and external, yet managed to surmount each with a combination of creative and practical solutions. Perhaps these 10 lessons learned can help others chart a less-perilous course towards the ultimate goal of putting information in the hands of those who need it.

The DFG manages California's diverse fish, wildlife, and plant resources, and the habitats upon which they depend, for their ecological values and for their use and enjoyment by the public. In support of this mission, the department's information technology hardware and network infrastructure have been regularly upgraded to provide most employees with high-speed access to their local data assets and to the outside world via the Internet. Several large, production database applications automate procurement and payments, allocate and adjust the budget, track the sale of hunting and fishing licenses, manage contracts, store information about properties and assets, oversee commercial fishing activities, monitor habitat restoration, and more. Like many other state agencies, DFG also relies on external data processing systems for everything from recording accounting transactions to managing its human resources. These external systems, supplied by control agencies and hosted at large data centers, provide relatively inflexible access to data – generally limited to lookup screens on terminal emulators, voluminous fanfold reports, or downloads of huge fixed-format data files. Direct communication between these external and internal systems is rare, creating the ubiquitous *islands of information* that hinder knowledge management in most organizations.

The need for a centralized, consolidated, departmental data store, optimized for online analytical processing and easily accessible by all employees, was recognized by business analysts at DFG for years. However, implementing such a solution is complex, both with regard to the technology involved and, more importantly, from a business process perspective since it involves the close cooperation of numerous organizational entities, requires dedicated resources throughout the department, and, most critically, depends on active sponsorship at the highest levels – in other words, by people with a lot of other oars in the water.

Initial attempts at data warehouse development stalled at the conceptual level. Though there was widespread support for the idea even among executive management; as a priority, there was little wind in the project's sails. Meetings were held, preliminary analysis was done, and some logical data models were developed but the effort eventually ran out of steam. Part of the problem was a lack of internal expertise in data warehouse development – after all, the department had never attempted it before. This would not have prevented forward progress in and of itself, but when resources were at a premium, and the choice was between *doing the business* or *improving the business*, day-to-day operations generally took precedence. Nevertheless, occasional discussions continued, and when some unexpected funding became available, the decision was made to outsource a project to someone with data warehouse expertise. A contractor was hired, the existing documentation was dusted off, and the initial project was launched.

1. Start Small to Achieve Early Success

The department's internal Budget Management System (BMS), developed as a client/server application some years previously, was woefully lacking in reporting capability, and each request for a new report had to be passed to the programmers for development and deployment. Even tiny reports that would be used only once had to make their way through this same narrow channel.

Deploying a modest data warehouse, both to streamline report development and, most importantly, to put it into the hands of the business users themselves, would be a significant *proof of concept*. Though its scope was small – only a single branch's data was involved – the benefit was immediate.

2. Optimize the Design

Budget branch staff were recruited to participate in the development of the

warehouse, along with the department's own analysts. After detailed discussions, the contractor produced a denormalized, business-centric *star schema* (see Figure 1). This is a standard data warehouse design with one or more *facts* tables comprising the hub of each star, surrounded by various *dimension* tables that allow the level of granularity of the *facts* to be drilled into or rolled up along relevant vectors. For example, budget *allotment facts* could be viewed along different *time dimension* slices so that allotment adjustments, as they evolved from day to day or week to week, could be visualized. Or, alternatively, budget allotment *facts*, and their specific impacts on various programs or funds, could be either summarized or detailed throughout the entire program or funding hierarchies. Or, lastly, budget allotment *facts*, as they are distributed throughout the department, could be viewed with respect to any level of the organizational hierarchy.

This is quite different from the fully normalized relational architecture adopted by most online transaction processing (OLTP) systems. The *star schema* pays no consideration to data redundancy since quick response time is the highest priority. Because the schema is denormalized, accessing and filtering huge data sets is extremely fast since table joins are kept to a minimum and data access paths are fully pre-planned in accordance with the current business needs. Robust indexing has no impact on performance during *insert*, *delete*, or *update* transactions since only *select* transactions by end-users are permitted. This also eliminates the kinds of insert, delete, and update anomalies inherent within denormalized designs and ordinarily resolved by a normalized relational architecture.

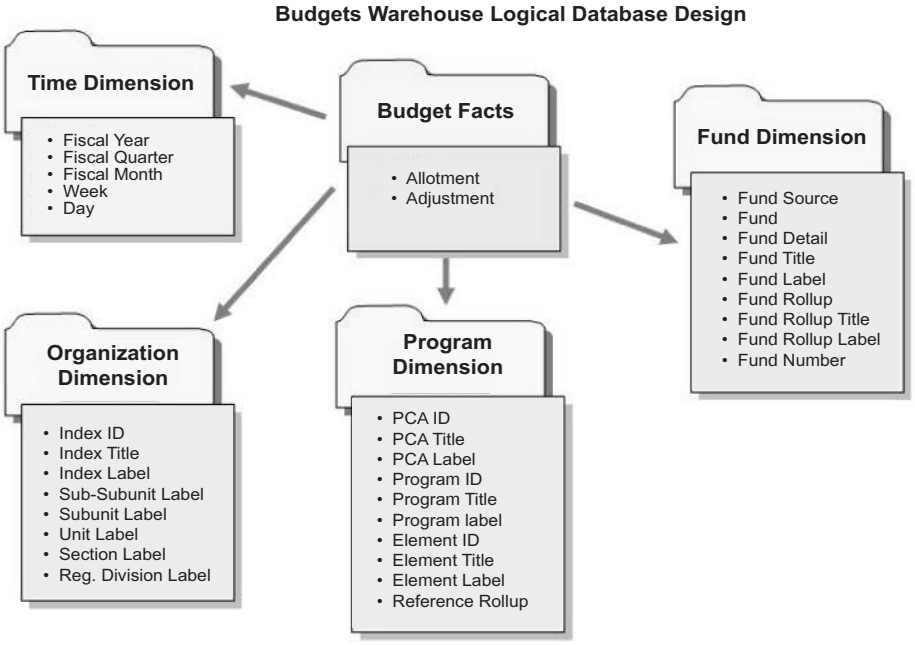


Figure 1: *Budgets Warehouse Logical Database Design*

3. Implement in Increments

The information technology branch (ITB) staff granted the contractors read-only access to the production client/server database so that the OLTP data could be imported into the warehouse using an export/import tool bundled with the department's relational database management system (RBDMS). These kinds of tools were generally able to accept as a data source nearly any form of exported file, either in text or binary format, including comma-separated values, fixed-length/fixed offset data, extended binary coded decimal interchange code, etc. They could then reformat the data on the fly and import it into the relational database tables in fairly flexible ways.

Once the data was loaded, a set of preliminary reports was generated to permit the knowledge experts to vet the warehouse data against the source data. Problems and discrepancies were identified and reconciled in an iterative process spanning several weeks. Some were the result of misunderstandings about the data and data relationships, and others were due to mistakes that were made during imports and exports.

Finally, when the warehouse data and the production data were seen to consistently match, a nightly, automated replication of the data into the warehouse was scheduled, tested, and confirmed. Over time, the budget branch staff began to use the warehouse with increasing confidence, and for the first time since the inception of the BMS, its data was now available to everyone in the branch in an organized, intuitive and easy-to-access manner.

4. Seize Opportunities

Due to this preliminary success, a second, more ambitious project was planned. The new warehouse would allow for the accurate and timely tracking of expenditures and encumbrances against allotments, permitting regional and divisional staff to know precisely where they stood, at any time, in managing their year-to-date budget. Unfortunately, the currents driving this new effort soon shifted and the project was essentially scuttled. The primary contractor had become gravely ill, the initial funding was nearly exhausted, and internal reorganization led to the remaining resources being redirected elsewhere. More importantly, the enhanced scope of this new warehouse project was obviously going to require collaboration among several branches within the department, such that managers within each branch would have to commit significant staff time to the development effort and would have to make themselves available for meetings and interviews. In light of everything else that was happening in the department, this kind of commitment was simply not forthcoming.

Fortunately, this same departmental reorganization produced a new business unit with staff who had the rare combination of both strong technical skills and business expertise – perfect for taking charge of a data warehouse development project. Additionally, the state's economy was suffering through a downturn, and the department's executive management was being required to submit numerous fiscal reports to both the legislature and the governor to present plans for budget and staffing cutbacks. These reports were difficult and onerous to create, given the scattered and heterogeneous nature of the department's data. During good economic times, state departments can think in terms of dollars, but during the kind of lean times DFG was facing it had to scramble for every penny. The need for a fiscal data warehouse, in which budget allotments could be compared directly with detailed and accurate year-to-date expenditures and encumbrances, sliced and diced through any cross-section of the department's internal structure, became an imperative, and the payoff also directly benefited all the business units required to participate. The

Assistant Deputy Director for Administration said *make it happen*. Thus, the project acquired an admiral, and the ship finally picked up some steam.

The recently formed Business Analysis Unit (BAU) was thus asked to create a fiscal data warehouse that combined allotment data from the internal BMS with financial transactions (expenditures and encumbrances) from CalStars, California's statewide accounting system, hosted by the Department of Finance (DOF). No additional resources could be made available, and there was no budget for training, outsourcing, or hiring of consultants. Still, a great deal had been learned during the initial project that could be effectively leveraged for this one.

5. Involve End-Users Throughout

The importance of good working relationships between all the stakeholders was clear from the earlier development effort, but we learned that it was not enough to simply hold interviews, gather requirements, and then build a system that meets them. Small but noticeable problems with the initial project were recognized to be the result of insufficient involvement by business unit staff at every stage. Input and feedback from those who would ultimately use the warehouse were seen as necessary throughout the entire development process, leading not only to the satisfaction of basic requirements, but just as importantly to the satisfaction of the users themselves. Consequently, management from each unit was asked to recruit their most appropriate knowledge experts to actually join the development team.

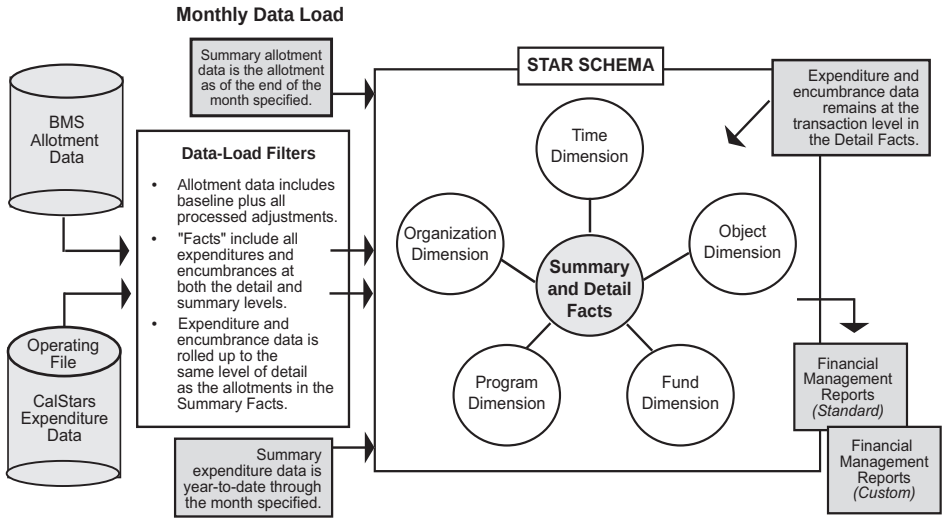
Discussions began in earnest between key staff from the department's budget branch, accounting branch, and BAU staff. ITB was also engaged, even though the actual development work would be done by the BAU since a fair amount of database administration would certainly be required in addition to planning for and accommodating the increased demands on servers, storage, and the network as a whole. Executive management was also consulted regularly to make sure that their reporting requirements could be fully met by the proposed design.

New *facts* and *dimension* tables were sketched out and held up before the business users for their input and suggestions. Again, BAU staff worked closely with those who would be using the warehouse until consensus was reached with regard to the essential facts to be focused on, and the critical *dimensions* along which the data needed to be manipulated.

DOF, host of the CalStars accounting system, made available two large binary data files at the close of each fiscal month – one containing year-to-date totals for transactions posted against all funds as of the close of that month and one containing all the individual transactions posted that month. Correspondingly, the requirement became obvious for two basic sets of *facts* – allotments, expenditures, encumbrances, and account balances from both a *year-to-date* perspective (cumulative amounts), and a *current month* perspective (individual transactions). The first set, *summary facts*, would enable managers to see exactly what they had been allotted to spend, and exactly what has already been spent or encumbered, year-to-date, from any account. The second set, *detail facts*, would allow managers to examine each expenditure individually, looking for erroneous charges or overcharges or to explain exactly why the *year-to-date* totals looked the way they did. Several discreet dimensions would enable these dollar values to be analyzed along the axes of time, organizational entity, program, fund, or object (expenditure category).

Budget allotment data would then be exported from the department's internal RDBMS and merged with the accounting data from CalStars. Since the accounting data was stored at a lower level of detail than the budget allotments, accounting transaction dollars had to be rolled up to this higher level in the *summary facts* table to make comparisons valid (see Figure 2).

Figure 2: *Allotment-Expenditure Warehouse High-Level Design*



6. Make It Business Friendly

Once the logical warehouse design was agreed upon and in close consultation with the business users, the physical database was developed in such a way that all the *facts* and *dimension* tables and data elements were assigned names based on their common usage within the department. We discovered one characteristic of a well-designed data warehouse is that everything in it should be immediately recognizable and intuitively meaningful to those who will use it. In addition, by honoring user input in this way, a certain *pride of ownership* among the business unit staff develops that is generally lacking in many software development efforts, not to mention the feeling of respect that this inherently communicates to staff.

Another important feature was to build in flexibility so different users could visualize the data in ways that are meaningful to them. For example, those who were familiar with coded values, such as A1772, should continue to be able to do so, generating concise reports that gave them exactly what they needed. At the same time, managers – more familiar with complete titles – should be able to see *San Joaquin River Salmon Telemetry* instead. Similarly, those dealing with fiscal year data and those basing their reports on calendar year data are easily accommodated.

Lastly, *facts* and *dimension* tables should be designed so that they can be prejoined in only a single way, eliminating the need for users to have to understand the underlying data structures or be required to make technical decisions when choosing which data to display in their reports.

7. Add Value Wherever Possible

Because data from two different systems was to be merged together in the warehouse, it soon became obvious that we

had to reconcile the different representations of the same data. This led to a concurrent, long overdue, and ultimately successful effort to create some department-wide data standardization. This was an unintended benefit, since the need for such reconciliation only became apparent during these first attempts to combine data from the two systems. Data standardization had been talked about within the department for some time and seen as a *desirable goal*, but no one had ever been given the responsibility for *data administration* to provide this oversight or spearhead this effort. Since there was no way to successfully merge the two data sets without reconciling the differences, the BAU was assigned *de facto* responsibility to become the data administrator for the department. Since disparities were significant but not huge, the effort had only minimal impact on the overall development timeline, yet it provided important value to the department, both in the current system and for future development efforts.

Conversations to understand and reconcile data variations involved the owners or stewards of each set of data plus the BAU analysts. Simply drawing attention to these disparities was an eye-opener to many who had no idea that other units were representing identical information quite differently. Where identical codes with disparate meanings were being used by different organizational entities, a consensus was reached as to which version would be kept and which modified, cascading those modifications out to all other affected systems. Where the same codes were used to represent the same data, but where the titles varied somewhat, the more complete or descriptive title was chosen as the standard. Abbreviations, viewed as a holdover from legacy systems and their limitations, were generally eliminated as unnecessary in the belief that full names produce fewer misunderstandings and less confusion.

8. Provide a Simple Implementation

Each new fiscal month's data would be loaded into the warehouse by the BAU just after the close of the fiscal month. A link to the Web-based user interface was then deployed to the appropriate section of the department's intranet home page by the ITB's Web Services Administrator – a simple HTML change pointing to the correct server and directory. The data warehouse interface itself was provided by a Web service that comes bundled with the department's RDBMS. This tool, being Web-based, can be made easily accessible to any employee able to connect to the department's intranet and with an appropriate database account, without the need for time-consuming and difficult-to-maintain software installations. First-time users were required to download a browser plug-in that enables the browser to communicate with the database, but this installation is triggered automatically during the first login session and never needs to be repeated.

Within six months, the warehouse had been designed, developed, and deployed. Not only was the data now available on the desktop to everyone who needed it, but because the lengthy and tedious distribution of paper reports throughout the state was no longer necessary, the new data were available weeks earlier than ever before. Budgeting and procurement decisions could now be based on and supported by detailed and accurate information. The warehouse development project had evolved into an unqualified success.

9. Maintain Momentum

There is a saying that quickly became applicable to this project: *supply breeds demand*. No sooner was the fiscal data warehouse available, than employees throughout the department began clamoring for more. The business users quickly understood the power of the data warehouse concept and began wanting the same kinds of access to other areas of the department's data.

If a fiscal data warehouse could be developed so quickly and easily, why not develop one for each of the following:

- Covering all departmental contracts.
- Developing labor cost accounting that brings together personnel data from the state Controller's Office with time sheet hours and charges from the CalStars accounting system.
- Reporting revenue.
- Reconciling staff positions from Human Resources with position funding from the BMS.
- Tracking and reporting on procurements and invoice payments.
- Monitoring vehicle usage.

In other words, staff had glimpsed an alternative to isolated and difficult-to-access islands of information and wanted consolidated, business-centric data available to them immediately, regardless of how many distinct and heterogeneous physical systems on which the production data actually resides.

10. Support Your Users

As demand escalated, the introductory information covered in the initial round of presentations was no longer sufficient to meet the complex needs of business users. They wanted greater access, training, documentation, and improved performance. So a warehouse primer was created, with step-by-step instructions for accessing the warehouse, running canned reports, modifying existing reports, creating new reports from scratch, sharing reports with other users, and exporting data to spreadsheets for additional manipulation. More comprehensive two-day trainings were scheduled throughout the state to provide business-unit-specific instruction and business-unit-specific reports. New dedicated servers, optimized for an online analytical processing environment, were procured to decrease wait-time and improve the user experience. The business-user response could not have been more positive or appreciative.

Ten Lessons Learned

Despite encountering many difficulties, enduring numerous delays, and proceeding with only limited resources and a small budget, the DFG has evolved a successful data warehouse development methodology centered around 10 lessons learned. In summary, they are the following:

1. Start small to achieve early success.
2. Optimize the design.
3. Implement in increments.
4. Seize opportunities.
5. Involve end-users throughout.
6. Make it *business friendly*.
7. Add value wherever possible.
8. Provide a simple implementation.
9. Maintain momentum.
10. Support your users.

Once the decision was made to move forward, the likelihood of success was maximized by constraining the scope of the initial project to a small and workable scale. Producing the first, modest product and demonstrating its immediate and important benefits, gave impetus to future efforts to expand our data warehouse development.

Starting with a design optimized for data warehousing maximized the quality and value of the initial product while laying the foundation upon which future efforts could be built, layer by layer, in sensible and incremental efforts.

While obstacles were encountered, and resources ebbed and flowed, opportunities did present themselves, and it was obvious that each, no matter how small or fleeting, had to be seized and exploited

CROSSTALK
The Journal of Defense Software Engineering

Get Your Free Subscription

Fill out and send us this form.

517 SMXS/MDEA
6022 FIR AVE
BLDG 1238
HILL AFB, UT 84056-5820
FAX: (801) 777-8069 DSN: 777-8069
PHONE: (801) 775-5555 DSN: 775-5555

Or request online at www.stsc.hill.af.mil

NAME:_____

RANK/GRADE:_____

POSITION/TITLE:_____

ORGANIZATION:_____

ADDRESS:_____

_____

BASE/CITY:_____

STATE:_____ZIP:_____

PHONE:(____)_____

FAX:(____)_____

E-MAIL:_____

CHECK BOX(ES) TO REQUEST BACK ISSUES:

JUNE2005 ☐ REALITY COMPUTING
JULY2005 ☐ CONFIG. MGT. AND TEST
AUG2005 ☐ SYS: FIELDG. CAPABILITIES
SEPT2005 ☐ TOP 5 PROJECTS
OCT2005 ☐ SOFTWARE SECURITY
NOV2005 ☐ DESIGN
DEC2005 ☐ TOTAL CREATION OF SW
JAN2006 ☐ COMMUNICATION
FEB2006 ☐ NEW TWIST ON TECHNOLOGY
MAR2006 ☐ PSP/TSP
APR2006 ☐ CMMI
MAY2006 ☐ TRANSFORMING
JUNE2006 ☐ WHY PROJECTS FAIL
JULY2006 ☐ NET-CENTRICITY
AUG2006 ☐ ADA 2005
SEPT2006 ☐ SOFTWARE ASSURANCE

TO REQUEST BACK ISSUES ON TOPICS NOT LISTED ABOVE, PLEASE CONTACT <STSC.CUSTOMERSERVICE@HILL.AF.MIL>.

if development were to produce the maximal benefit.

The department's most valuable asset, however, consists of its hard-working and dedicated employees and their knowledge and understanding of both the business and the data. This became the project's greatest resource. Project developers learned, from the inception, to involve knowledge workers and stakeholders at every stage and to work closely with them to ensure a usable and effective product. At the same time, active executive sponsorship at the highest levels was found to make the ultimate difference between moving forward or floundering. Though this support initially seemed to hinge on the scale of the crisis, resulting from not having access to the data that this project would provide, the long-term benefits soon became clear and executive managers became strong supporters of ongoing efforts.

It was also important to produce a resource that was *business-centric* rather than *information-technology* centric. This maximized usability, acceptance, and value to those who would need to depend on it for gathering and making sense of the department's huge stores of data, and it also served to demystify the data and data relationships.

We learned that mining data also uncovered heretofore hidden data anomalies and discrepancies within the organization. Rather than reacting to this with surprise or distress, it was seen as a valuable opportunity to launch efforts to better manage and administer the department's data assets, improving its consistency and standardization across diverse business areas.

The *simple is better* axiom was also confirmed when it comes to implementing these projects. Our thin-client, Web-based interface, immediately accessible by anyone on the network, was low in cost, simple to maintain, and quick to deploy. Perhaps not as robust as the client/server version of the same tool, this was more than compensated for by its many virtues.

Early successes can quickly be forgotten if momentum is not maintained. Fortunately, the groundswell within the department for more was immediate and energetic. By responding quickly with helpful documentation, on-site training, ongoing phone and e-mail support, and the continual creation of new and varied custom reports, forward progress rarely faltered.

Conclusion

DFG has now developed seven discreet business areas in its comprehensive data warehouse, each addressing one or more of the needs described earlier. Some are in full production, others are in pilot rollout, and two are still in development. A new and ambitious project to assimilate all the human resource branch data into a new business area is in the planning stages. At this point, the creation of new business areas in the data warehouse has become almost routine. Scores of canned reports have been written, each carefully named so that its content is clear and unambiguous. Business unit staff are requesting help building custom reports almost daily and many, within a short time, have become adept at producing their own reports without assistance. The time and costs associated with printing, copying, and distributing hard copies throughout the department have been dramatically reduced as users now simply log on to the warehouse, highlight a report, and run it.

The need for making data available to the people who require it is clear and ever-present. In many cases, vital facts and figures that used to take days or weeks to collect and organize are only a mouse-click away. Now that the ship has arrived with its cargo of information, the department's employees are *thirsty* no longer – at least for today.◆

About the Author

Crilly Butler, Jr. is a senior information systems analyst with the California Department of Fish and Game. A licensed psychotherapist, he switched careers when he realized that a computer would never call him at 2:00 a.m. threatening to jump. Butler specializes in database and data warehouse design and development, along with information systems and business analysis, leveraging his past training in interpersonal communication to *build bridges*. A previous software project he led was accepted into the Smithsonian's permanent research collection in 2000.

California Department of Fish and Game
1416 9th ST
RM 1248B
Sacramento, CA 95814
Phone: (916) 651-7844
Fax: (916) 651-8763
E-mail: cbutler@dfg.ca.gov