4.2 Central Database Management

John P. Bolte, Douglas H. Ernst, and Duncan Lowes
Department of Bioresource Engineering
Oregon State University
Corvallis, OR USA

Background

The PD/A CRSP Central Database is a centralized data storage and retrieval system for PD/A CRSP research and for other aquaculture research programs with compatible objectives and standardized methodology. The Database, started in 1983, currently contains over eighty aquaculture production studies and represents the world's largest inventory of standardized aquaculture data. Through its World Wide Web site, the Database is available to aquaculture researchers, educators, outreach and extension agents, and producers worldwide.

Two fundamental objectives for the original development of the Database were to 1) create a mechanism for analysis of variance among geographically dispersed aquaculture research sites, in addition to analyses within single ponds and among ponds at a single location, and 2) support development of predictive models for aquaculture pond processes (Egna et al., 1987). Ten years later, these objectives remain central to the purpose of the Database. Further discussion on the purpose and methods of the Database can be found in Batterson et al. (1991) and Ernst et al. (1997).

Objectives

The Database was relocated to Oregon State University in May 1996. Since then, objectives in the development and management of the Database have focused on procedures to promote error-free and timely data submissions from PD/A CRSP research projects and on the infrastructure and mechanisms necessary to support Database publication on the World Wide Web. These objectives include those originally stated in the Database Proposal (Eighth Work Plan) as well as priority issues that have emerged over the last year. These objectives were to:

1. Reconstruct the Database under a rigorous, relational framework.
2. Enter missing, overdue data for Workplans 1 through 7.
3. Add experiment protocol information to all existing datasets.
4. Develop a tracking mechanism for current PD/A CRSP research projects.
5. Publish a Database Manual for data submission requirements and methods.
6. Update and expand the PD/A CRSP Handbook of Analytical Methods.
7. Publish the Database on the World Wide Web.
8. Establish context-sensitive linkage between the Database and the Program Management Office Web Sites.
9. Enhance awareness of the Database in the greater aquaculture community and create additional opportunities for its use.

Rationale

Standardized methods for aquaculture research and standardized databases for aquaculture information are fundamental requirements for the continued advancement of aquaculture science and engineering. The PD/A CRSP is a world leader in both standardized methods and database publication of aquaculture research.

However, as of December 1996, the Database served mainly as a data repository for PD/A CRSP research projects with relatively few requests for data use (about 30 total from 1983 to 1996). This lack of use by the aquaculture community was likely due to a combination of factors, including lack of awareness, difficulties in database access, and lack of the necessary database infrastructure to facilitate the search and extraction of specific datasets. An additional problem with the Database was a continuing accumulation of overdue data submissions from PD/A CRSP research projects.

The objectives enumerated above directly address these concerns. The people who will benefit from this work, i.e. existing and potential users of the PD/A CRSP Database, include aquaculture researchers, educators, outreach and extension agents, and producers worldwide.

Objective 1: Database Reconstruction

Tasks. Reconstruct the Database under a relational organization framework. Find and remove all erroneous data. Implement procedures to prevent future entry of erroneous data.

Accomplishments. The new, hierarchical organization of the Database mirrors that of the PD/A CRSP program, including the various research facilities involved, the experiments performed at each of these facilities, and the treatment protocols and replicate data comprising each of these experiments (see diagram on p. 38). Five levels of hierarchical organization are used: 1) global, 2) facility, 3) experiment, 4) experiment treatment, and 5) treatment replicate. The relational-database software used to manage the database enforces these data relationships.

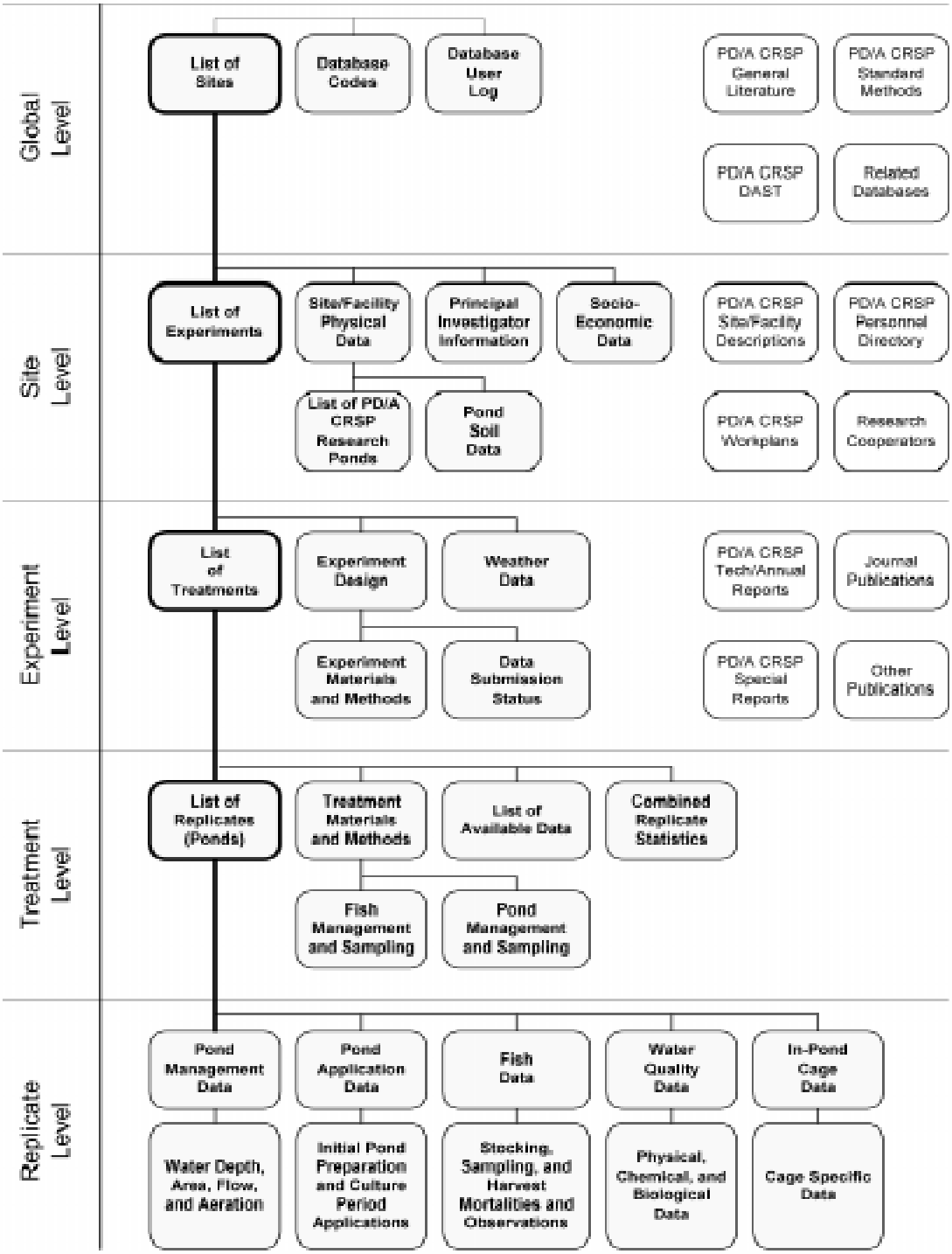
Considerable reorganization and editing of the database was required to remove erroneous data, implement relational data structures, and support efficient mechanisms for data access and publication. Some of these changes unfortunately required adjustments to data-table formats used by researchers. However, these changes were essential to the development of a database architecture that supported data-entry error checking and data-extraction user queries and graphical formatting.

The degree of work required under this objective, about three months total, was not anticipated in the Database Proposal (Eighth Work Plan). It was incorrectly assumed that the Database was already in a workable, relational, organization structure.

Specific accomplishments include:

- The Database has been reorganized from 720 files, maintained in dbf format using FoxPro database software, to one file (60 MB) maintained under Microsoft Access® database software.

THE NEW, HIERARCHICAL ORGANIZATION OF THE PD/A CRSP DATABASE MIRRORS THAT OF THE PROGRAM.



- Primary-key based data indexing and automated relational data organization has been implemented.
- Partial and fully duplicate records have been removed.
- Inconsistent pond names have been corrected.
- Redundant and undefined names for fish stocks and pond application materials have been corrected.
- Reorganization of time and depth of water samples to primary-key based indexing has been accomplished
- Removal of erroneous data values exceeding reasonable range values has been completed.

Objective 2: Overdue Data Submissions

Tasks. Enter missing, overdue data for Workplans 1 through 7. First, rely on Principal Investigators to come forward with overdue data until Dec. 1997. Then, generate a list of overdue data submissions, organized by Principal Investigators responsible, and pursue directed inquiries.

Accomplishments. Historically, submission of data from PD/A CRSP research projects to the Database has not been adequately enforced. As a result, for Workplans 1-7, approximately one-third of the total studies that should be in the database is missing. At the 1997 Annual Meeting, the Database Manager provided a list of experiments that were in the Database and asked that overdue data through the Seventh Work Plan (Sept. 1, 1993 to Aug. 31, 1995) be submitted. However, since May 1996, the Database Manager has received only one data submission from a PD/A CRSP research project.

To be completed. The next step to be completed is to generate a list of all overdue experiments, identified by Work Plan and Experiment Title and organized by Research Site and Principal Investigator. This list will be published and kept current at the Database Web Site, where it will be accessible to all past and current Principal Investigators. With this list as a reference, correspondence with Principal Investigators (copy Program Management Office) will be initiated regarding specific datasets due, their required content, and their anticipated timelines of completion.

Objective 3: Experiment Information for Existing Datasets

Tasks. Add experiment protocol information to all existing experiment data already submitted to the Database. Require this information for all future data submissions.

Accomplishments. For all PD/A CRSP studies submitted to the Database through May 1996 (80 total), data in the Database consisted of replicate sampling data only and lacked additional information regarding research protocols and experiment treatment specifications. With no information regarding fish pond management, there was a corresponding absence of a fish production-methodology context from which a database user could identify and extract specific datasets. The PD/A CRSP Workplans, Technical Reports, and Annual Reports were of limited use for defining treatment specifications in the

Database, especially after Work Plan 3, given their superset relationship to the Database subset, re-mixing of experiment treatments between proposals and reports, and lack of linking references.

To the greatest degree allowed, experiment treatment specifications have been gleaned from the Database itself by compiling fish stocking, pond application, and water management data into overall treatment values. These specifications are organized by Experiment ID (combines Work Plan, Site Code, and Experiment Number) and by the specific experimental replicates (Facility Ponds) assigned to that treatment.

All experiment and treatment information is now required at the time of data submission. As itemized in the Database Manual, this information includes research objectives, experimental design, sampling protocols, additional materials and methods not described in the PD/A CRSP Handbook of Analytical Methods, explanation of departures from planned protocols, and significant problems encountered in the course of the study.

The degree of work required under this objective was not anticipated in the Database Proposal (Eighth Work Plan).

To be completed. The next step to be completed is to circulate these specifications to individual Principal Investigators for review, correction, and additional information.

Objective 4: Data-Submission Tracking

Tasks. Develop a tracking mechanism for current PD/A CRSP research projects regarding required content and due-dates of data submissions.

Accomplishments. For the Eighth and subsequent Work Plans and greater, requirements for data submission to the Database are defined in the individual research sub-contracts (Article VII, Reporting Requirements) and in the Database Manual. All data collected under a study must be submitted within three months (90 days) of the scheduled end of an experiment. It is the responsibility of the Database Manager to circulate data submission reminders and overdue-notifications to the Program Management Office and Principal Investigators.

To effectively carry out this task, a Work Plan summary table is required that lists all research studies of each Work Plan organized by research site. Also included would be experiment start and end dates and a list of, or reference to, all data that are to be collected and submitted for each experiment. Given mutual benefits and complementary responsibilities, discussions with the Program Management Office have shown that this table would best be accomplished as a combined effort between the Database Manager and the Program Management Office.

To be completed. The next step to be completed is to build a template for this table and post it at the Database Web Site. The Program Management Office will enter Work Plan information into the table and keep it updated. The Database Manager will mark check-off boxes when data is correctly submitted and the Program Management Office will do the

same for other project deliverables. This list will be readily available to Principal Investigators, the Program Management Office, and the Database Manager at the Database Web Site.

Objective 5: Database Manual for Data Submissions

Tasks. Develop a Database Manual to provide a single source for all data submission requirements and procedures. Distribute manual to Principal Investigators.

Accomplishments. Procedures that must be considered by Principal Investigators to properly submit data to the Database include data-submission timeline requirements, data-submission mechanisms, personnel and publication referencing, site and facility specification, data organization, format, and use of templates, experiment and experiment-treatment specification, and description of departures from Work Plan protocols and standard methods. All of this information has been made available in the Database Manual, available in printed form from the Database Manager or in electronic form at the Database Web Site. The first version was completed January 1997 (Ernst, 1997).

To be completed. An updated version of the Database Manual is a high priority and will be completed by Sept. 1, 1997.

Objective 6: PD/A CRSP Handbook of Analytical Methods

Tasks. Update and expand the PD/A CRSP Handbook of Analytical Methods (Piedrahita et al., 1991). Establish direct, one-to-one linkages between the Handbook sampling variables and the Database data fields by use of common data (variable) names. Put the Handbook into the Database and make it available to PD/A CRSP research personnel and data users.

Accomplishments. At the 1997 Annual Meeting, the Materials and Methods Technical Subcommittee delegated responsibilities for method revisions (e.g. weather, soil, water, and fish sampling) and method additions (e.g. facility and experiment specifications and socio-economic data). The Database Manager was assigned the task of receiving and collating these updated methods, using the existing version of the Handbook as a starting point.

To be completed. An electronic form of the PD/A CRSP Handbook of Analytical Methods has been located and now needs to be added to the Database. As inputs from the Materials and Methods Technical Subcommittee come forward, methods in the Handbook will be updated.

The existing PD/A CRSP Handbook of Analytical Methods (Piedrahita et al., 1991) contains copyrighted materials used directly from external sources. To honor copyright agreements between these sources and the PD/A CRSP, copyrighted material will continue to be made available to PD/A CRSP researchers only. For public domain publication of the Handbook at the Database Web Site, copyrighted sections will be replaced with references. This public domain version of the Handbook will be useful to data users, as contextual information for specific studies, and to aquaculture research projects outside of the PD/A CRSP that wish to submit data under the required standardized methodology.

Objective 7: Database Publication on the World Wide Web

Tasks. Provide immediate, minimal cost, worldwide access to the Database by establishing an Internet site for the Database and publishing its data on the World Wide Web.

Accomplishments. The Database currently resides on a Windows-NT server and is maintained using relational database software (Access, Microsoft). A server application (Cold Fusion, Allaire) is used to support client-server database access and database publication via the Internet (World Wide Web). A number of Web forms have been developed to support tabular data retrieval. A programming language (Java, Sun Microsystems) is used to embed time-series and water-depth based plots in Web pages for graphical data retrieval. Further discussion of computer software technology required to support the Database Web Site may be found in the Database Proposal (Eighth Work Plan). Location of the Database on an OSU College of Engineering Network Server, at the Bioresource Engineering Deptartment, provides permanent housing of the Database, centralized, secured storage with automated backup procedures, and Internet accessibility. Internet accessibility to the Database via the World Wide Web is available at http://www.biosys.orst.edu/crspdb.

Data may be searched, extracted, reviewed, and/or downloaded according to user-specified geographical location, inclusive calendar years, fish species and stocks, and fish production methods. In this respect, the experimental treatment protocols by which the research was accomplished correspond to equivalent fish-culture management scenarios that can be considered by data users. Data available from the Database include 1) site weather, 2) pond/site soil composition, 3) pond mapping and water management, 4) pond application materials, rates, and compositions, 5) fish numbers, weights, and lengths at stocking, during culture period, and at harvest, 6) water quality variables, and 7) natural biological-productivity variables.

To be completed. Data currently available at the Database Web Site are raw, replicate sampling data in tabular format only (e.g. water temperature and fish weight time-series data). Under development is presentation of data in graphical formats and in a variety of additional forms, including 1) calculated data (e.g., fish biomass productivity and feed conversion efficiency), 2) statistical summaries (e.g., treatment means, variances, and analysis of variance), and 3) fitted model parameters (e.g. fish growth and water quality models). In this respect, experimental treatments may be viewed and analyzed as individual entities, grouped according to their original experiments, or recombined to create new experiments.

Also to be completed is to provide data-users with additional experiment information specific to an extracted dataset, including research personnel citations (analogous to printed publications), physical descriptions of research facilities, and references to related publications (see Objective 8).

Objective 8: Linkage with the Program Management Office Web Site

Tasks. Provide direct, context-sensitive linkage between the Database and the Program Management Office Web Sites. Support Database access to research site descriptions, research personnel citation information, publication references, and publication texts.

Accomplishments. A need for Database linkage to additional PD/A CRSP information was recognized in order to provide necessary contextual support for specific, extracted datasets. This information includes:

- Principal investigators responsible for specific experimental datasets, to support data citations and referrals, analogous to printed publications.
- Site and facility descriptions and physical data, to be used in conjunction with experiment-treatment specifications.
- References to PD/A CRSP literature (Workplans, Technical Reports, Annual Reports, etc.) and other external publications, to be used to augment experiment treatment descriptions, provide research objectives and context, and provide experimental results and discussions.

Discussions with the Program Management Office have shown that they already maintain information on research personnel, research facility descriptions, and research publications (references and texts). Thus, this information will continue to be maintained by the Program Management Office, while actual housing of this information at the Database or Program Management Office Web Sites is under discussion. Context-sensitive linkage will be supported by simply using the specific location (research site) and time (calendar years) of an extracted dataset.

To be completed. Further discussion with the Program Management Office is required to determine data housing locations, specific mechanisms of data linkage and access, and division of responsibilities between the Program Management Office and the Database Manager. An overall plan that is emerging is to use the Database to house all elemental data (numbers and text strings) including facility specifications, publication references, and personnel information. The Program Management Office Web Site would be used to house complete document texts. As done historically, the Program Management Office will be responsible for maintaining personnel, facility, and publication information.

Objective 9: Database Promotion

Tasks. Enhance awareness of the Database in the greater aquaculture community and create additional opportunities for its use. Make the Database available through other public databases. Actively promote its use through aquaculture conferences and publications.

Accomplishments

- Database Web Site Publication. As presented above, publication of the Database on the World Wide Web will greatly enhance its visibility and availability to the aquaculture community worldwide.

- International Symposium on Tilapia in Aquaculture (ISTA IV). A paper entitled "PD/A CRSP Central Database: A Standardized Information Resource for Pond Aquaculture" (Douglas H. Ernst, John P. Bolte, Duncan Lowes, and Shree S. Nath) has been prepared for ISTA IV, for presentation at its upcoming meeting (Nov. 1997, Orlando, FL) and publication in its proceedings.

- AquaNIC. A link to the Database Web Site is provided at AquaNIC, an Internet-based aquacultural information service maintained at Purdue University.

- Oregon Department of Fish and Wildlife (ODFW). A seminar by Doug Ernst and John Bolte was given to ten people from the ODFW (May 23, 1997), with one component addressing procedures and methods of the Database and its Web Site interface. Applications similar to the Database for use in ODFW salmon hatchery management were discussed. The relationship of the ODFW Central Office to its multiple hatchery sites and needs for standardized data storage and access are similar to the data publication needs and relationship of the Program Management Office and its multiple research sites.

To be completed

- ICLARM FishBase. Summary data compiled from the Database will be included in FishBase (Froese and Pauly, 1996), starting with FishBase 1998. FishBase provides a wealth of fish biology, fish taxonomy, fisheries, and aquaculture information compiled from a wide range of sources.

- Consortium of International Earth Science Information Networks (CIESIN). To consider fish production information available in the Database in conjunction with additional, geographical information available for a given research site, data from the Database may be used in conjunction with geographical data from other sources. One approach to this task, that will be available by Dec. 1997, is to access summary data compiled from the Database via CIESIN. CIESIN maintains a worldwide, geographically based, environmental information database that is publicly available at the CIESIN Web Site (access via the Database Web Site).

- World Aquaculture Society (WAS). A technical session entitled "Use of Computer Tools for Aquaculture Planning, Design, and Management" is being organized for the next WAS annual meeting (1998, Las Vegas, NV) by Shree Nath and Doug Ernst. Use of standardized aquaculture research methods and databases for computer tool development, as represented by PD/A CRSP research and Database, respectively, will be an important component of this session.

Literature Cited

Batterson, T., Berkman, H., Hopkins, K., Piedrahita, R., and Pompa T., 1991. Final Report on Database Management. Pond Dynamics/Aquaculture CRSP, OSU, Corvallis, OR USA. 51 pp.

Egna, H.S., Brown, N., and Leslie, M. (Eds.), 1987. Pond Dynamics/Aquaculture Collaborative Research Data Reports, Vol. 1: General Reference. Pond Dynamics/Aquaculture Collaborative Research Support Program, Oregon State University, Corvallis, OR USA. 84 pp.

Ernst, D.H., 1997. Pond Dynamics/Aquaculture Collaborative Research Database Manual. Pond Dynamics/Aquaculture Collaborative Research Support Program, Oregon State University, Corvallis, OR USA. 30 pp.

Froese, R., and Pauly, D. (Eds.), 1996. FishBase 96: Concepts, Design, and Data Sources. ICLARM, Manila, Philippines. 179 pp.

Piedrahita, R.H., Boyd C., and Szyper, J., 1991. Handbook of Analytical Methods. Pond Dynamics/Aquaculture Collaborative Research Support Program, OSU, Corvallis, OR USA. 150 pp.