A New Approach to the Management of Environmental Information

by

Blair A. King
B.Sc., Queen's University, 1989

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Chemistry and School of Environmental Studies

We accept this dissertation as conforming
to the required standard

Supervisors: Dr. T.M. Fyles
Dr. P.R. West

Abstract

Environmental science is a growing field that draws data from a broad range of disciplines. These data represent the intellectual and financial efforts of countless individuals and institutions and are invaluable for continued research on the environment. This thesis details three case studies that center on providing users with improved access to environmental data and suggest an information model. Users will be better served by environmental information systems that provide detail on the strengths and limitations of data in archives, and that give direct access to individual measurements accompanied by metadata. Metadata provides the required, essential summary of the applicability of data.

The first case study describes the creation of a prototype metadata system CODIS (the Continental and Oceanographic Information System). It examines the creation of an effective database organization for a multidisciplinary information system and the generation of conventions and techniques to assemble and structure multidisciplinary data. These conventions included the requirement for input using previously prepared lists and the development of parallel data structures between disciplines to facilitate data entry and searching. This improved database organization was demonstrated to decrease the time needed for data entry while reducing error rates in the entered data.

Data in CODIS are appraised for reliability using discipline-specific protocols. The protocols are based on a dichotomous, decision tree format accompanied by detailed guidelines. The output from the appraisal process is a non-hierarchical assessment based on a five-point scale and comments from appraisers. These products inform users about the reliability of the included data. The protocols were examined for repeatability and replication between appraisals. The outputs from the appraisal processes were demonstrated to be comparable to peer review.

Contextual evaluation, developed in the second case study, provides insight into the potential applicability of data in databases. The NCIS (National Contaminants Information System) study examines the development of a system to create contextual metadata to be stored with archival data. Contextual evaluation is carried out by examining and documenting each step in the experimental process. This study entailed developing a set of protocols for the assessment, and creating educational tools to ensure their effective implementation. NCIS groups datasets as either experiments or surveys, with only experiments being evaluated for context. It was necessary to develop a unified organizational scheme to classify diverse research and monitoring activities into defined categories. The process was reviewed and a refined version is currently in use across Canada in the implementation of NCIS. The case study highlighted difficulties associated with the division into experiments and surveys.

The third case study examines the censoring of data, a practice that involves reporting values as unknown or undetected when their existence is known. This study of the British Columbia, Ministry of Environment's Environmental Management System (EMS) examines the limitations placed on secondary users and metadata systems by storing censored data in archives. It includes a survey of current practices in environmental analytical laboratories and investigates the statistical tools used to remediate censored data. The case study concludes that censoring of data severely limits the secondary use of otherwise high-quality data.

A gap-analysis of the studied systems leads to a set of recommendations and responsibilities that highlight the critical insights derived from the case studies and emphasize shared responsibility by all partners in the data-to-decision process. The thesis then presents a three-tiered conceptual model for a general environmental information system. In order to facilitate this task three new information elements are proposed and defined: datasets, infosets and metasets. It is anticipated that this work may serve to influence the direction of environmental data management practices by providing a model for future environmental information systems.

Table of Contents

List of Tables

List of Figures

List of Excerpts

List of Acronyms

ACSCEI	American Chemical Society Committee on Environmental Improvement
ADCAP	Arctic Data Compilation and Appraisal Program
APHA	American Public Health Association
ARRP	Aquatic Resources Research Project
ASTM	American Society for Testing Materials
CODIS	Continental and Oceanographic Data Information System
DFO	Department of Fisheries and Oceans Canada
DS_ID	Dataset Identification
DSIDXREF	Dataset to Bibliography Cross-reference Table
EC	Environment Canada
EDI	Electronic Data Interchange
EMS	British Columbia, Environmental Monitoring System
EPA	United States Environmental Protection Agency
EPP	Environmental Protection Program
EQUIS	Environmental Quality Information System
FREDI	Fraser River Estuary Directory Information
FSDB	Forest Science Data Bank
GIS	Geographical Information System
IOS	Department of fisheries and Oceans, Institute of Ocean Sciences
LABMAN	Laboratory Management System
LOD	Limit of Detection
LOQ	Limit of Quantification
MDL	Method Detection Limit
MEDS	Marine Environmental Data Service
MELP	British Columbia, Ministry of Environment, Lands, and Parks
NCIS	National Contaminants Information System
NUSAP	Numeral, Unit, Spread, and Assessment Pedigree system
ODIS	Oceanographic Data Information System
QA	Quality Assurance
QC	Quality Control
RSCAMS	Royal Society Analytical Methods Committee
SEAM	System for Environmental Assessment and Management
SFU	Simon Fraser University
SPARCODE	Specific Parameter Analytical Route Code
UBC	University of British Columbia
UVic	University of Victoria
VEC	Valued Ecosystem Component
WESCAP	West Coast Data Compilation and Appraisal Program
WQN	Water Quality Network

Acknowledgements

I would like to express my thanks to Dr. Tom Fyles and Dr. Paul West for their help, patience, faith and guidance throughout my time at UVic. I would like to acknowledge my many collaborators from the Environmental Information Research Group, Simon Fraser University, Environment Canada, the Department of Fisheries and Oceans and the British Columbia Ministry of Environment, lands and Parks. Special thanks to the many friends who made me welcome and taught me to enjoy the academic experience especially Dr. Dave, Todd, Scoots, Dave R., Mark, LL, Simon, Scott F., Rob, Tony and Sandra S. Financial assistance in the form of awards from the British Columbia Ministry of Environment, Lands and Parks; the Family of Edward Bassett; and the University of Victoria was very much appreciated. Finally I am grateful to my family for their continued love and support throughout the long course of my education.

Dedication

This work is dedicated to the two people
who taught me that anything is possible
with hard work, dedication and love,
thanks Mom and Dad.

Chapter 1 Introduction

1.0 Introduction

This thesis is about the management of environmental data. The Canadian environmental data management strategy has been characterized as localized and uncoordinated, with data collection, management, access and preservation being driven by the needs of individual disciplines, institutions and projects (Canadian Global Change Program, 1996). Examples exist of major irreplaceable collections of data being lost in all disciplines and in all sectors in Canada (Canadian Global Change Program, 1996). These data represented the intellectual and financial efforts of countless individuals and institutions and would have been invaluable in continued research on environmental variables. Data are one of the few assets in an institution that increase in value. Samples taken at one time cannot be re-taken, and new techniques or hypotheses are continually being developed, thus permitting new interpretations of documented historical data (Clay, 1997). Improved access to this irreplaceable resource is essential to support future environmental research, monitoring and decision-making. Delays in developing systems to effectively preserve these data and information have very serious consequences. Researchers retire and information can become lost or irretrievable due to poor filing systems, incomplete documentation of files, or technological obsolescence. Besides the obvious desire of institutions and individuals not to see their work lost or forgotten, making historical data and information available can have numerous positive consequences including reducing the need to reproduce work or carry out new sampling when pre-existing results can be used in their stead.

Figure 1.1 displays the current state of the environmental data climate in Canada. In the center is the real world, which serves as the source of measurements intended to serve to understand the system. These measurements are carried out for numerous reasons and may be stored in a variety of locations and on a variety of media. Users interested in accessing this data seldom have access to all the useful measurements that have been carried out. Instead, they are generally limited to the data in published reports or certain

select archives. Consequently a great deal of useful information is not considered in decision-making.

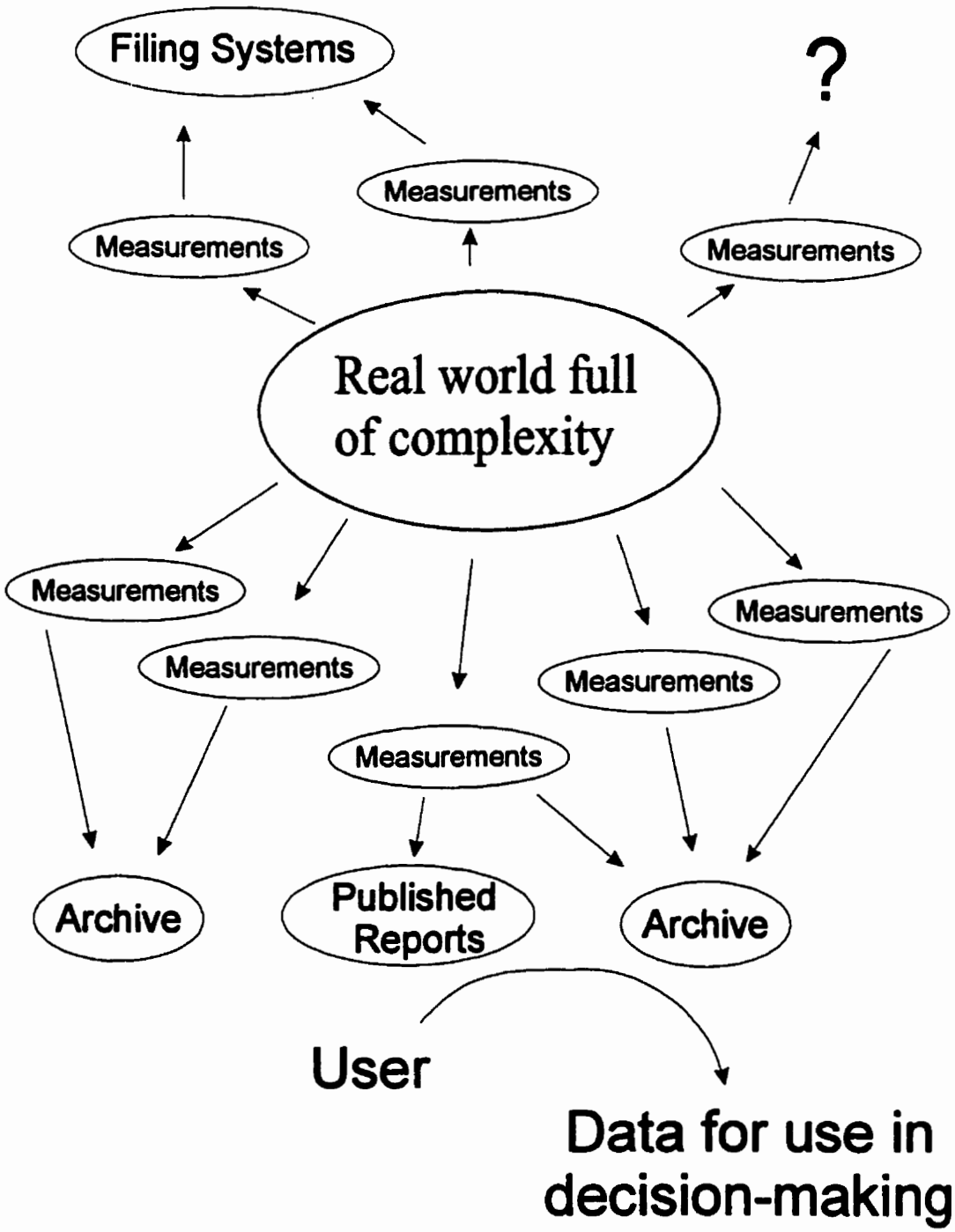


Figure 1.1 The Canadian Data Management Situation

The current focus of the environmental data management field has been on developing structured archival databases and geographical information systems (GIS) containing primary data (National Research Council-USA, 1995). In Canada, organizations like the Federal Department of Fisheries and Oceans (DFO) and the British Columbia Ministry of Environment, Lands and Parks (MELP) are dedicating significant amounts of time and effort in the creation of databases such as the National Contaminants Information System (NCIS) and the BC Environmental Monitoring System (EMS) (AXYS, 1994; and LGS, 1995). These systems are designed to preserve the data accumulated by these institutions for subsequent reuse (LGS, 1995). The challenge in designing such systems lies in making the primary data available in a form that allows for effective re-use. In particular, it lies in designing systems that promote the use of data by secondary users.

Secondary users are data consumers. They make use of the data in information systems but are not directly involved in the initial process that produced the data. Consequently, they are often limited in their understanding of the data accessed and rely on the information systems to provide them with reliable, applicable results. In essence, secondary users are seeking information.

Data and information are different entities (Samli, 1996). When data are properly gathered, organized, processed, analyzed, and delivered they become information (Samli, 1996). Roots (1992) emphasized that the main barriers to production and dissemination of information that can contribute to effective environmental knowledge were those that affect the reliability, adequacy, accessibility, and understandability of environmental knowledge. These researchers both emphasize the same point, that data must be associated with some additional elements in order to be useful as information. Identifying these additional elements and designing systems that incorporate them in order to promote the effective re-use of scientific information about the environment, are the chief goals of this research.

Chechile (1991) has described an idealized decision-making process as consisting of six steps:

1) Identify the problem and define the goal
2) Identify alternatives including the status quo
3) Gather and analyze information about alternatives, probabilities, implementation plan, risks and benefits
4) Apply a decision tool, e.g., systems model, decision tree or linear programming
5) Make the decision
6) Implement the decision

The quality of any decision derived from this process is dependent on the quality of each step, if any of these six steps is poorly executed, then the result will be a flawed decision process (Chechile, 1991). The research in this work concentrates on the information management needs of the third and fourth steps, as these are most directly related to the results of research scientists. The implication of this data-to-decision model is that different individuals will be involved in the process at different stages. These individuals vary greatly in expertise and may be involved in the overall decision-making for only one or a few of the steps. They include persons in a wide variety of situations at all levels of organizations, ranging from elected officials, agency representatives, department heads, and bureau chiefs, to program managers, field supervisors, and technicians (Holcomb Research Institute, 1976). Different decision-makers are often sensitive to different issues or inputs and have differing priorities in the temporal (day-to-day operations or long-range policy making) and spatial scope (the amount of land or number of people affected) of their decisions.

While many decision-makers are experts in their fields, none can be experts in all fields. When faced with multidisciplinary data they seldom have complete certainty of the data's disciplinary-based uncertainties. This can result in environmental management decisions that do not consider disciplinary-specific uncertainties in their calculus (Reckhow, 1994). Assessing the quality and general availability of data and reporting that information in a manner that allows for effective discipline-specific and cross-disciplinary analyses is critical.

Data are the most valuable assets that most organizations possess (Clay, 1997). It has been estimated that 50% of the costs of any study will be directly related to data collection and up to a third of the entire research budget of an institution will be required for editing, documenting and archiving that data (Clay, 1997). Any process that can encourage the reuse of pre-existing data will, thus, provide an added return in both a scientific and institutional sense. Samli (1996), in his work on the use of data in marketing, suggests that there are at least five criteria for good data: reliability, validity, sensitivity, relevance, and versatility.

> **Reliability** means that the data were produced in such a way that if the study were to be replicated using the same techniques; the same results would be obtained. That means that the data is not loaded with random errors that make them undependable.
>
> **Validity** indicates that the data show what they are supposed to show. In other words, the research instrument has measured what it was supposed to measure.
>
> **Sensitivity** implies that the data indicate small changes and variations in the phenomenon that is being represented (or measured) by the data. When the data lack sensitivity, research will not yield significant results and the efforts will be wasted.
>
> **Relevance** means that the problem to be solved or the decision to be made is practical and important. The data that are gathered will be able to accomplish what they were supposed to do, meaning that the proper data were collected.
>
> **Versatility** includes robustness. In other words, the data can be used for various statistical analyses. Measuring the phenomenon for various interpretations is made possible if the data are versatile (Samli, 1996).

Samli (1996) notes how critical it is that secondary users of data, like decision-makers, be able to decide on the quality of the data. He points out that although the researcher must generate good data, in the final analysis, secondary users are responsible for determining if the data being accessed will be reliable enough for their use. If the quality is not acceptable, the data can never become information or be used effectively (Samli, 1996). Bolin (1994) insists that researchers must take more responsibility. He states that it is essential for scientists to recognize and communicate clearly, and as objectively as

possible, the limitations of their information provided. What Bolin recognizes is that secondary users are often unable to determine the reliability of data. Secondary users can be faced with situations where insufficient annotative information is available to determine the reliability of data. Alternatively, some users simply lack the expertise to carry out an analysis of the reliability of the accessed data.

Advances in computer technology have provided numerous tools to aid accessing data. The increasing power and decreasing cost of computerization have resulted in the creation of larger and more complex databases, which are able to store more data about more phenomena than was ever thought possible. The variety of data included in these new databases results in a number of potential difficulties, which can hinder secondary users and thus decrease the value of the data stored in these systems. The stored data often vary greatly in spatial and temporal scales. This results in the need for systems of increasing complexity in both size and design (Stafford, Brunt and Michener, 1994). The data collected into these systems will also be derived from a multitude of disciplines; each with its own specialized analytical and discipline-specific language requirements (Stafford, Brunt and Michener, 1994). Secondary users unfamiliar with the requirements and jargon of other disciplines will be ill equipped to search for applicable data and, if that data are identified, will be unable to address any uncertainties regarding the data's reliability.

With the advent of improved technology the volume and availability of data has increased tremendously, but the community of users who can make use of the information derived from that data has decreased. This decrease is the result of the heightened sophistication of these new systems, which require more sophisticated users and increasingly refined technologies (Roots, 1992). The expanding volume and rate of data acquisition and transmission has resulted in the requirement for increasingly sophisticated means of dealing with it (Roots, 1992). The management of data, which was formerly the purview of archivists and some, few scientists, has become central to important economic, environmental, intellectual, and social questions (Canadian Global Change Program, 1996). The flow of scientific data from a very large variety of sources is increasing yet the

development of systems and agreements necessary to make the best and most cost-effective use of these data lag behind (Canadian Global Change Program, 1996). Consequently, issues relating to data preservation and accessibility are receiving increased attention from the broad scientific community (Michener et al., 1997).

Environmental science is particularly sensitive to technologies that increase the availability of data. In the evaluation of an environmental problem one might be expected to examine physical, chemical, biological, technological, economic, philosophical, ethical, legal, and political factors (Chechile, 1991). Omission of any of these factors is likely to oversimplify the problem and render the decision process incomplete and unrealistic (Chechile, 1991). The data used in environmental research have historically been collected through small-scale studies involving one or a few investigators in a single discipline and funded for relatively short periods (Stafford, Brunt and Michener, 1994). Consequently, available data on the environment are usually unique to a particular sector and are collected to satisfy particular operational requirements (Manning, 1992). This has increased the difficulty in reporting of changes in the environment and developing synergistic information (Manning, 1992). The effective management of this growing, multidisciplinary, data stream is an underlying challenge of the environmental field. Assembling and processing data from a broad range of basic sciences for application in addressing environmental problems is one of the key functions of environmental science (Caldwell, 1990).

There is general agreement that current database systems are inadequate for managing large heterogeneous sets of scientific data. Gosz (1994) indicated that in order to increase the value of datasets for future work, databases should document the many conditions associated with the original measurements. As Gosz (1994) pointed out, data becomes more valuable for subsequent studies if the appropriate ancillary data are archived. Ward, Power and Ketelaar (1996) analyzed the computational and information management needs of geoscientists and identified key shortcomings in current geoscientific data analysis practices. They suggested that the key concepts of a proposed system

architecture would include the management of data, data analysis operators, and experiments; the maintenance of supporting data for each of these components; and interoperability among diverse data source and application software packages (Ward, Power and Ketelaar, 1996).

In a similar study, Brown (1994) analyzed the information requirements for ecology. He suggested that ecologists must confront numerous challenges in their efforts to address environmental questions including: incorporating information from new data sources and other disciplines; standardizing and controlling the quality of data; and integrating, synthesizing and modeling knowledge about ecological systems. Brown (1994) pointed out that the variation in the quality of data makes the need for standards for data collection, management and analysis critical. He suggested that all data does not need to achieve the same standards of accuracy and precision, requirements vary with the problem being addressed. Instead, Brown (1994) considered that the quality of data was critical. It had to be known to be accurate in order to ensure that it was sufficient for the application. This requires attention to documentation and standardization at all stages of data processing, from initial collection through management to final analysis (Brown, 1994).

As noted above, data are a valuable asset, however, improvements in environmental information management systems have multiplied the opportunities for data to move from one user to another, eventually escaping the bounds of intended use (Chrisman, 1994). This necessitates the association of supplementary information to accompany the escaping data and provide a context for their secondary use. Consequently, environmental information systems must ensure that data are only available when accompanied with that supplementary information. As Stafford, Brunt and Michener (1994) noted, this will greatly increase the complexity of any system designed to contain this data.

The requirement that data only be available if accompanied by supplementary or contextual information will not prevent secondary users from accessing individual

measurements. It will merely guarantee that the individual measurements are accompanied by sufficient details to ensure that the data are used appropriately. This protects the secondary user from inadvertently misapplying the data, while reassuring the primary data producers that their work will not be misused. Only when both parties are satisfied can an effective environmental information system be developed. An effective system serves both the primary and secondary users. A system is effective for primary data producers when they feel comfortable entering their data and are certain that the data will be both securely stored and protected from accidental misuse. Secondary data users require systems that allow them to understand the strengths and limitations of the data accessed while remaining confident in their knowledge that all potentially useful data has been identified.

There is a demonstrated need for a methodology that represents a new approach to the growing data problems in the environmental fields. This new approach must acknowledge that as technologies progress, more data will be collected by more agencies about more phenomena. The traditional approach of simply increasing the size of data repositories will not address this problem. Data archives serve their purpose, but as these archives increase in size and complexity the need arises for tools to communicate the contents of these archives in an efficient manner for use by decision-makers. These tools must reflect the fact that the environmental field is interdisciplinary, as is the expertise of potential users. It must also acknowledge the differing needs of primary and secondary users.

As summarized above, numerous workers in the field have presented a similar group of requirements for new environmental data management systems (Brown, 1994; Stafford, Brunt and Michener, 1994; Gosz, 1994; and Ward, Power and Ketelaar, 1996). All emphasized data with supporting data elements, which appraise reliability and describe the context of data and data collection. Collectively, these supporting data are the "metadata". This thesis asserts that properly defined and controlled metadata will encompass the additional elements that convert data to information. Consequently, in this work, information can be defined as data plus its associated metadata. This thesis examines the implications of metadata for environmental data management. One aim of

this research is to demonstrate that value added to the data, through their link to associated metadata, enhances the applicability and usability of the original data for subsequent reuse, and in particular, for decision-making in the environmental field. By developing an effective methodology to create, store and disseminate data and its associated metadata, the major concerns of both data producers and data users can be addressed.

1.1 Goals of the Research

This thesis will present a new approach to the management of environmental data that effectively translates the uncertainty associated with environmental data in a transparent and objective manner for use in environmental information systems. This process will accommodate the complexity of real-world situations, include natural variability and uncertainty and acknowledge the multidisciplinary nature and differing needs of the receiving audience. This new approach, based on metadata, involves the creation of new information tools and systems that will facilitate access to the archival records while adding value to the data by appending indicators of reliability and context to individual records. The information model developed in this thesis involves the creation of datasets as displayed in Figure 1.2.

These datasets act as an organizing tool to preserve the relationships between measurements in archives, while also serving as the basic information unit in a new generation of information systems called inventories (Figure 1.3). Establishing the baseline metadata requirements of datasets and developing a process by which they are applied are two of the major goals of this work. Then a process is needed to associate critical contextual information with datasets. Consequently, an additional goal of this work will be to develop a procedure to associate contextual information with datasets in environmental information systems.

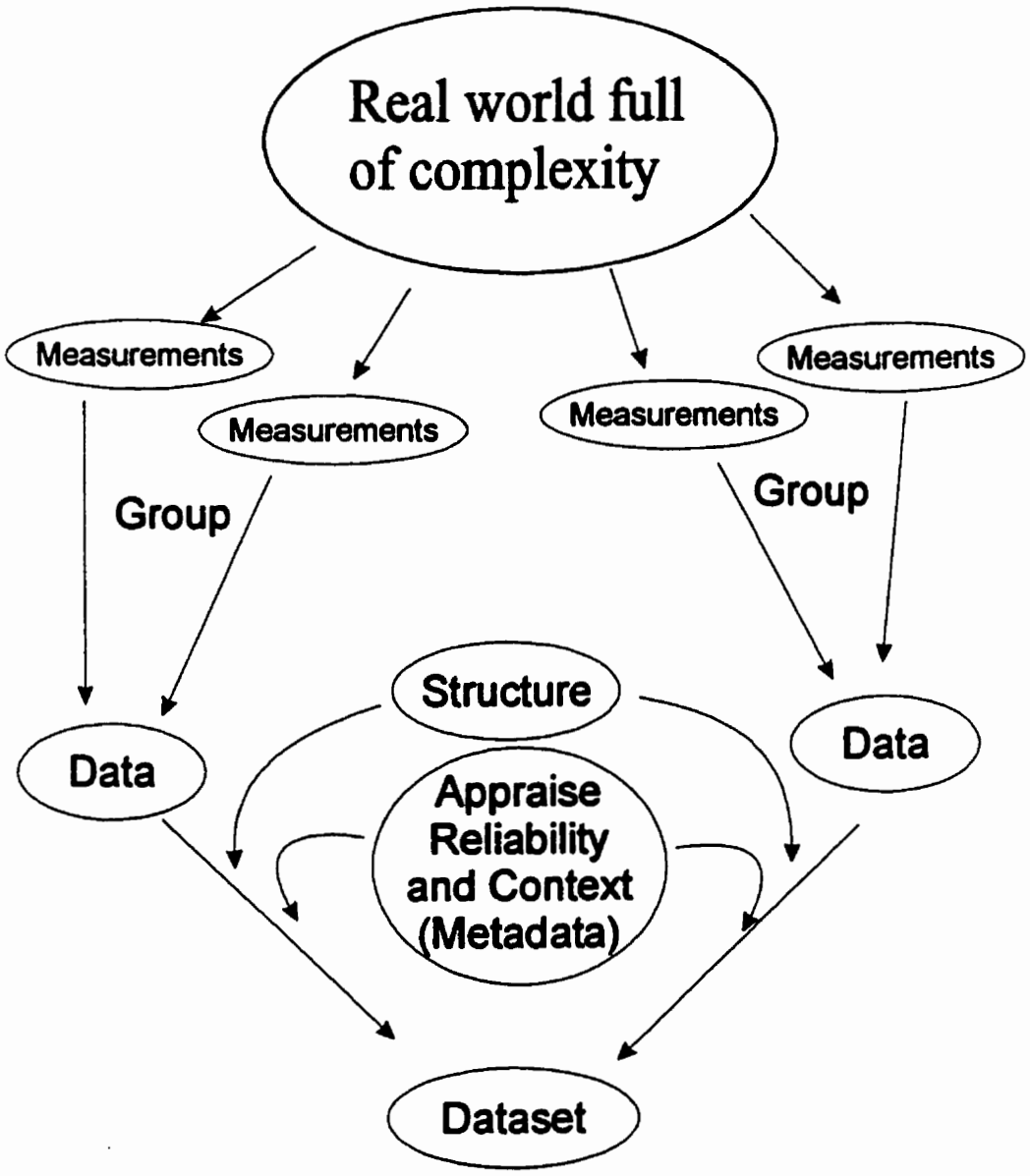


Figure 1.2 Information Model for the Creation of Datasets

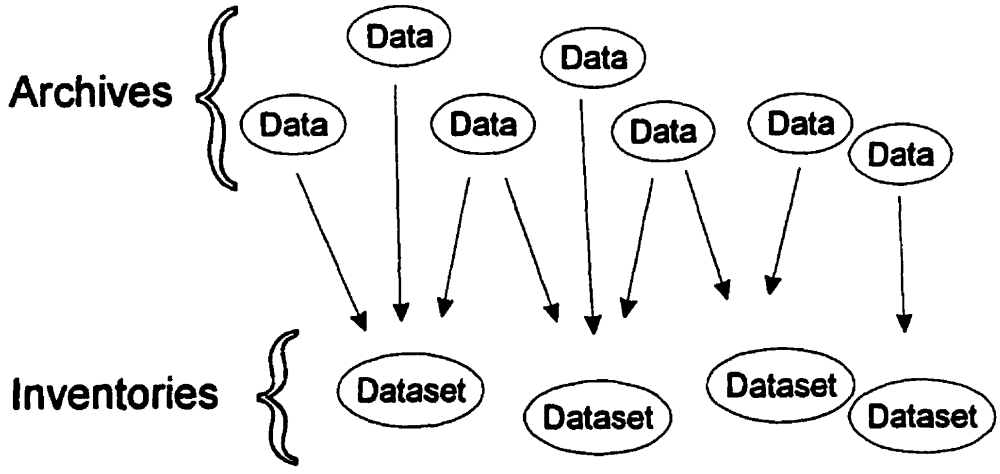


Figure 1.3 The Relationship between Data and Inventories

These three goals are met through the completion of a number of component objectives:

- Determine the basic requirements for storing multidisciplinary data.
- Identify the basic requirements of metadata for multidisciplinary data.
- Establish the baseline metadata elements needed for differing types of environmental information systems.
- Develop of a set of structuring and appraisal tools to apply metadata to data in an objective and reproducible manner.
- Elaborate a methodology to evaluate the contextual basis of data and report that information.
- Apply these structuring, appraisal and contextual tools in real systems in order to test their efficacy and assess how they respond to natural uncertainty and variability.
- Evaluate and review the process and incorporate improvements

A general model can then be developed based on these recommendations and this analysis. This model will provide a theoretical and practical foundation and structure for an environmental information system that provides an effective basis for decision-making.

1.3 Methodology-A Case Study Approach

This thesis is an account of a program of interactive research. That program began by identifying the strengths of the natural sciences. The research then incorporated concepts and tools from various sources including reliability ratings, standardized protocols and

data structures. Each of these concepts and tools was refined to become compatible with an overall methodology. The evolution was carried out through case studies that developed and implemented these concepts in real systems. Through this activity, weaknesses were analyzed and omissions identified. Subsequent systems were then developed and the process repeated. At each stage, input was sought from experts and reviewers and incorporated into subsequent development.

This iterative research project will be presented as a series of three case studies. The case studies represent independent research activities that shared critical characteristics. Each case study examined some aspect of the process of storing data, derived from measurements of environmental variables, in information systems, in order to facilitate effective decision-making. In addressing the case studies a number of tools were developed to transform the goals of the study into the architecture needed for a general model of an refined overall process to improve the use of data in environmental decision-making. This model will be described in Chapter 6.

The first case study described the creation of a prototype metadata system called the Continental and Oceanographic Information System (CODIS). Creating CODIS required developing an intellectual framework for metadata. In order to apply this framework, it was necessary to design a number of data structuring and reliability appraisal tools. CODIS provided an opportunity to test and critique these tools and improve their efficiency. The outcomes of this case study were an intellectual framework for metadata systems and a set of protocols to structure data and appraise their reliability.

The second case study examined the creation of a system to appraise experimental activities to be incorporated into an environmental information system being developed by DFO called the National Contaminants Information System (NCIS). The creation and application of this appraisal process refined many of the tools developed for CODIS and required the development of additional approaches. The outcome of this project was a functional system to evaluate the context of experimental events, which is being used to

input new data into the NCIS. In addition, this case study provided an improved understanding of the data and information needs of environmental decision-making and insight into the limitations of many environmental information systems in current use.

The third case study explored the use of truncated data derived from analytical laboratories in MELP's new Environmental Monitoring System (EMS). This research examined the data requirements of archives and investigated how changes in the data transmission or reporting affect the ability of researchers to make use of that data for alternative tasks.

The work in these case studies made it possible to identify gaps in current environmental information systems. This examination provided critical insight from which a number of recommendations and responsibilities for environmental information systems and their users could be derived. The gap analysis, recommendations and responsibilities together suggested a conceptual model for an ideal environmental information system that met all the requirements discussed. This ideal model is presented in Chapter 6.

Chapter 2 Definitions

In the design of environmental information systems, controlled terms and definitions are critical. The following terms will dominate the discussion of the case studies.

2.1 Datasets

In order to facilitate the long-term, computerized storage of scientific data it is necessary to break down standard reports and publications to "datasets" which can be readily input into the storage systems. The McGraw-Hill Dictionary of Scientific and Technical Terms (4th ed.) defined a dataset as a named collection of similar and related data records, recorded upon some computer-readable medium. The Concise Oxford Dictionary of Current English defined "data" as known facts or things used as a basis for inference or reckoning. It defined "set" as a number of things grouped together according to a system of classification or conceived as forming a whole. From these two definitions, it is clear that the term "dataset" must preserve the sense of expectation of internal consistency.

Since 1979, the Arctic and West Coast Data Compilation and Appraisal Programs (ADCAP/WESCAP) of the Institute of Ocean Sciences (IOS) of Fisheries and Oceans Canada (DFO) have produced catalogues for all types of physical, chemical, and biological oceanographic data. The compilations attempt to examine all data regardless of their source and status. Twenty-two catalogues have been published to date in the Canadian Data Report of Hydrography and Ocean Sciences No. 5 and 37 series, as volumes of the Arctic (ADCAP) and West Coast (WESCAP) Data Cataloguing and Appraisal Programs, respectively. The catalogues developed for ADACP/WESCAP assemble groups of measurements together into entities, which they called data sets. The developers of ADCAP/WESCAP did not define "data sets" but did stipulate:

> Each data set comprises sampling or chemical measurements taken during a single cruise, or during a sampling excursion usually by a single agency. It is assumed, then, that data within a given data set have been collected uniformly and should be internally consistent insofar as sampling methodology is concerned.

From this definition, it is evident that the term "dataset" must also preserve the sense of consistency derived from a single source. From these various sources Fyles et al. (1993a) defined a "dataset" as:

> a collection of measurements unified by one or more of the following characteristics: chemical species, biological species, physical matrix, geographical locations, or sampling methodology. The measurements must be treated uniformly, ideally by a single agent or agency and should be internally consistent with respect to sampling methodology. The measurements within the dataset need not always be of the same type.

In addition Fyles et al. (1993b) stipulated that the derivation of individual datasets from a data source (or sources) must strive to maintain the expectations of internal consistency of the original workers. This definition will be used in this research.

2.2 Metadata

Metadata is "data about data" or more completely, "data about the content, quality, condition and other characteristics of data" (Federal Geographic Data Committee, 1994). A commonly recognized example of metadata is the Library of Congress system used to organize library holdings using call numbers. Books are ordered on shelves using a call number system based on content and characteristics of the book (i.e. subject, genre, author, and publication date). A user seeking books on a subject need only identify the appropriate call number in order to locate the correct section of the library where all the books covering that subject should be stored.

The metadata concept has a rich history in the social sciences (Zhao, 1991) while in computer science metadata and its use have become an important issue of investigation for the last two decades (Al-Zobaidie and Grimson, 1988). The most prominent current use of metadata has been in the geospatial field, specifically in reference to geographical information systems (GIS). The standardization of information used in federally funded geospatial data systems in the United States began in 1995 when federal agencies were instructed to develop and use a "standard" to document new geospatial data and to provide these metadata through a National Geospatial Clearinghouse (Federal

Geographic Data Committee, 1994). The term metadata, however, should not be restricted to geographical data. As Hsu et al. (1991) put it, the scope must be extended from simply representing data systems to including knowledge resources as well. For the purposes of this work, the Michener et al. (1997) definition of metadata will be used:

> all information that is necessary and sufficient to enable long-term secondary use (reuse) of data sets by the original investigator(s), as well as use by other scientists who were not directly involved in the original research efforts (Michener et al., 1997).

This definition responds directly and completely to the who? what? where? when? how? and why? questions posed at the outset by any user confronting a new piece of information.

Metadata is a product of data that can be used without referring to the original data itself. Computer systems based on metadata can be used to search for the existence of data without referring to archival systems containing the raw data just as libraries can be searched for one tome without reading every book. Metadata can also provide insights not readily available from the primary data themselves. Since the scale is larger, metadata offer the potential to examine large-scale trends, which are missing in the smaller scale of individual studies. The metadata offer direct access to cross-, inter- and multidisciplinary analyses of regional monitoring and management significance. Even on its simplest level, the metadata provide a useful resource for the communication of specialist information to non-specialist audiences and non-expert users.

2.3 Reliability Indicators

The appraisal of measurements and observation is widely practised. In the biomedical field, meta-analyses of the results of several similar studies are a common approach to evaluation of the efficacy of various procedures (Mann, 1990). In order to combine data from individual studies, an appraisal of each study is an essential prerequisite. The classification schemes are usually not particularly subtle, using categories of "good", "reasonable", "poor", and "bad" as one example (van Beresteyn et al., 1986). Similarly, meta-analyses in forestry (McCune and Menges, 1986), in ecology (Gurevitch et al.,

1992), in institutional analysis (Roos et al., 1989), or in agricultural economics (Fletcher and Phipps, 1991), all confront the same issues with descriptive scales to express the degree of reliability of the data. In an organizational climate where defined experimental protocols have been developed, the reliability of information collected can be expressed in the degree to which the "right" methods were used. This circumstance occurs in multi-center biomedical studies with rigorous clinical protocols and in water quality programs with significant investment in protocol development. One example of the latter is the Puget Sound Water Quality Authority (Puget Sound Estuary Program, 1991).

The ADCAP/WESCAP data appraisal effort made use of a common five-level scheme, or reliability rating to express the potential reliability of data (Cornford et al., 1982). This system has been adapted and refined for use in this project. A breakdown of the five ratings is provided in Table 2.1. While hierarchical in appearance, this scheme is meant to establish the intercomparability of data. Hence "2" rated data is not necessarily less valuable (worse) than "4" rated data, provided it is applied with knowledge of its limitations. "4" rated data has both demonstrated internal consistency between measurements and has been standardised with some external standard while "3" rated data shows only internal consistency, without benchmarking by an external standard.

An alternative approach to the appraisal and classification of scientific information is the NUSAP system described by Funtowicz and Ravetz (1991, and Costanza et al. 1992). NUSAP stands for Numeral, Unit, Spread, and Pedigree, and was designed to describe the reliability of parameters such as the mean temperature rise due to a particular global warming model. A NUSAP notation for a value would be given as: a numeral value (3), a unit value (°C), a spread (± 50%), and a pedigree grade (0.5). The spread is just the statistical uncertainty in the result derived by conventional statistical techniques. The pedigree expresses the limits of a scientific field in which the process knowledge was generated. It serves as an assessment of the strength of the scientific result.

Table 2.1 ADCAP/WESCAP Rating Scheme from Fyles, King and West (1993b)

Rating	Data Reliability
0	Data are found to have errors. The data source contains obvious discrepancies.
1	Data are suspect because of recognized weaknesses which compromise the internal consistency of the data. Patterns or trends within the data are probably not real.
2	Insufficient information is provided to assess the reliability of the dataset. Trends in the data may, or may not be, real.
3	Data are internally consistent. Patterns or trends within the data can be used with relative confidence. Comparisons with other datasets may be difficult or unachievable.
4	Data are internally consistent and are sufficiently standardized to permit comparison with other datasets of this rating.

The NUSAP grade describes the assessment of the model according to matrix shown in Table 2.2. Cornford and Blanton (1993) use similar prose classifications to describe the degree of certainty in process knowledge. Within NUSAP, high scores imply a sound theoretical framework based on substantive experimental validations, and enjoying a wide degree of consensus support. Information from such a source is likely to have high predictive value and could be used with confidence in a variety of contexts. Lower scores imply a weaker theoretical framework, more anecdotal experimental work, or less consensus in the scientific community. The predictive capacity would also be lower, and the uncertainty in the information would be correctly communicated to the public policy forum in the lower pedigree score. In the example above (3 °C ±50% [0.5]) the pedigree of the model was assumed to be {2,2,2} indicating a computational model using indirect estimates, from one of several competing models.

The pedigree grade is the average of the scores normalized on the scale 0-1. The NUSAP grade expresses the uncertainty in a form which is amenable to an "arithmetic of uncertainty" (Costanza et al., 1992). More importantly, it provides a suggestive index rather than a defined mathematical quantity.

Table 2.2 Numerical Pedigree Matrix (Costanza et al., 1992)

Score	Theoretical, Quality of model	Experimental, quality of data	Social, Degree of consensus
4	Established theory - many validation tests - causal mechanisms understood	Experimental data - statistically valid samples - controlled experiments	Total - all but fringe
3	Theoretical model - few validation tests - causal mechanisms hypothesized	Historical/field data - some direct measurements - uncontrolled experiments	High - all but dedicated disputants
2	Computational model - engineering approximations - causal mechanisms approximated	Calculated data - indirect measurements - handbook estimates	Medium - competing schools or methodologies
1	Statistical processing - simple correlations - no causal mechanisms	Educated guesses - very indirect approximations - "rule-of-thumb" estimates	Low - embryonic field - speculative and/or exploratory
0	Definitions/assertions	Pure "guesses"	None

The numerical rating of data reliability as used by DFO or the NUSAP scheme, are both effective ways to communicate scientific uncertainties to non-experts. They reflect a consensus approach to the doing and reporting of science. Unfortunately, any appraisal intended to classify data also infers personal judgement, presumably by an expert, but nonetheless potentially imprecise and subjective. To enforce objectivity and ensure confidence in the assessments, appraisal processes require well-described protocols for the analysis of primary data.

Chapter 3 CODIS Case Study

3.0 Introduction

CODIS (the Continental and Oceanographic Data Information System) is a geo-referenced data information and retrieval system based upon metadata. CODIS was developed as a functional prototype upon which to test theories of information management using metadata. CODIS also serves as a stand-alone management tool. The research activity that eventually became the CODIS project pre-dates the beginning of the research program described in this thesis. CODIS, however was an integral part of the formulation of the approach to environmental information management used in this work. It served as the central research activity of the early years of this project and provided a vehicle to test and refine many of the principles that are fundamental to the completion of the model presented in Chapter 6.

The CODIS case study is an examination of the process that began with the decision to create a metadata system, it evolved into a systematic methodology to apply metadata and appraise datasets. The methodology developed in the creation of CODIS was subsequently refined in the development of the Department of Fisheries and Oceans Canada (DFO) National Contaminants Information System (NCIS) and tested against other models including the British Columbia, Ministry of Environment's Environmental Management System (EMS). Case studies of these other systems are the subject of subsequent chapters. The goal of this case study was to carry out a critical analysis of how metadata could be created and organized for use in an information system. It included examining the lessons learned in that process, which were incorporated in a general model for information management (Chapter 6). Specifically, the CODIS case study examined the process by which the CODIS metadata system was designed, how metadata was assigned to datasets and practical applications of the model. The metadata creation process involved developing methodologies to structure multidisciplinary data, building a multidisciplinary information system and appraising individual datasets for their

quality. It involved developing a general structure for multidisciplinary data and the creation of a methodology to appraise scientific data.

3.1 CODIS History

CODIS began as a sub-component of the Aquatic Resources Research Project: Environmental Risk Assessment and Management (ARRP) in April 1991 (Farrell, 1993). ARRP was a multi-faceted research project centered at Simon Fraser University (SFU) (Farrell, 1993). It combined the expertise of 50 researchers drawn from the Geography, Environmental Toxicology, Biology, Zoology, Chemistry, Resource Management and Statistics departments at SFU, UVic and the University of British Columbia (UBC). The focus of the project was on the partitioning of toxic compounds in the biota, waters and sediments of the Fraser River and on the linkage between scientific data and resource management. The research program involved five mutually supported sub-components aimed at contributing to a design for an integrated strategy for improved ecosystem management (Farrell, 1993). Sub-component IIIA, at UVic, involved creating a sustainable, functional database of organic data in the Fraser River estuary. This database was intended to support other sub-components by facilitating liaisons between datasets and data users. It eventually expanded to become CODIS.

The crucial feature of the sub-component IIIA database was its need to serve as a cross-disciplinary link that would allow the interdisciplinary team to identify critical elements of multidisciplinary data for use in their own discipline-specific research. The ultimate goal of this project, within ARRP, was multifaceted and included: providing reliable datasets for an environmental modeling sub-component; identifying critical data gaps and focusing on critical criteria for modeling purposes; providing a close link to policy and decision-making groups within ARRP; and serving as a powerful monitoring and planning tool, in a pro-active support role for the social science components of the project.

CODIS, from the outset, was intended to serve as more than a limited tool for use by ARRP in the Fraser River. Early development involved cooperation with the Data Assessment group (DA) at the Institute of Ocean Sciences (IOS) in Sidney B.C. The DA group, in association with the Native and Regulatory Affairs Division and the Freshwater Institute, had been involved in a process to review the sufficiency and suitability of available scientific data collected in the Arctic and West Coast of Canada (Ratynski, and de March, 1988; Birch et al. 1983). ADCAP/WESCAP was designed to collect and publish this data in the Canadian Data Report of Hydrography and Ocean Sciences Series No. 5 (ADCAP) and No. 37 (WESCAP). The cooperative research venture was intended to combine the development of CODIS, at UVic, with the efforts of the DA group. The aim was to create a system that would incorporate both the Fraser River and ADCAP/WESCAP data into a single system.

When the CODIS project was initiated in 1991, the DA group had already published 22 ADCAP and three WESCAP catalogues. These catalogues covered a diverse range of disciplines from ocean chemistry to marine zoobenthos. In order to simplify the task of publishing the ADCAP/WESCAP catalogues, some of their data was collected into computer files (Wainwright, 1991). After their publication in paper format, an effort was made to create a computerized catalogue, which would contain some of the information from the catalogues. This system, called the Oceanographic Data Information System (ODIS), was designed to support efficient computer access to the ADCAP/WESCAP information and the tides and currents data being stored at IOS (Wainwright, 1991). ODIS was developed in Oracle with custom FORTRAN procedures that provided map display and "query from map" capabilities and resided in the MicroVax system at IOS (Waiwright, 1992).

ODIS served as the starting point for CODIS and so initial work of CODIS involved rationalizing the ODIS data structure formalized by Wainwright in his "PC ODIS Data Dictionary" (1992). The ODIS data dictionary presented six disciplines: physics, chemistry and biology (which consisting of four sub-disciplines: fish, marine mammals,

plankton and benthos) (Wainwright, 1991). Since the data was collected for publication and not for the creation of software, each ODIS discipline had its own distinct data structure. Few structural details were held in common between disciplines. In effect, ODIS was six separate databases connected through a single software shell for use as a single system (Figure 3.1).

The CODIS software was initially envisioned as a PC tool that would combine the ODIS data with Fraser River organic contaminants data. As mentioned above, the ODIS model consisted of three, discipline-based systems that operated under a common software shell. If a dataset contained data from more than one discipline, separate files would be created in each discipline to which data might belong. The disciplines were linked through sampling locations, source documents and people. The lack of structure in the applicable files meant that none could be used for searching purposes. Figure 3.1 clearly demonstrates that the different disciplines were supposed to work in parallel, as part of a large combined system. The common thread between each was supposed to be the Dataset Identification (DS_ID) field. The DS_ID field was a method designed to uniquely identify every dataset in the system (Wainwright, 1991). Significant overlaps in DS_IDs existed between disciplines. This compromised the functionality of DS_IDs in ODIS.

The design presented in Figure 3.1 was never fully implemented (Smiley, B., Pers. comm.). Instead, each discipline worked independently. Shared files were not actually shared and each discipline had its own unique structure. A typical data structure (Ocean Chemistry) is displayed in Figure 3.2.

CODIS was originally intended to expand on the ODIS model by adding a new discipline: organic contaminants in the Fraser River Basin (called Continental Chemistry), and transferring the entire product from a mainframe environment to one capable of being used on a personal computer. Early in the development of CODIS it became apparent that seven different data structures made for an exceedingly complex programming task.

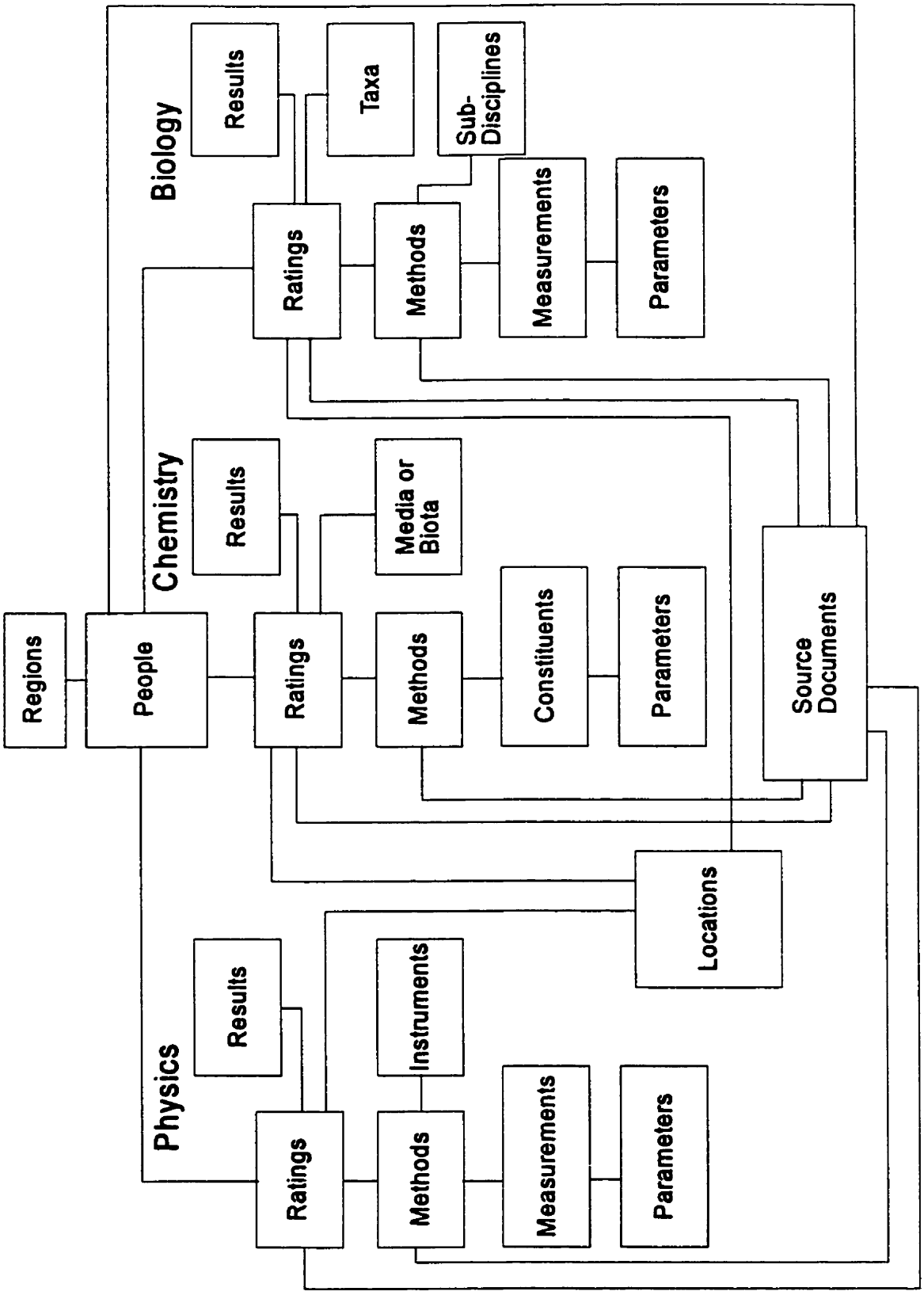


Figure 3.1 ODIS Overall Data Structure from Wainwright (1992)

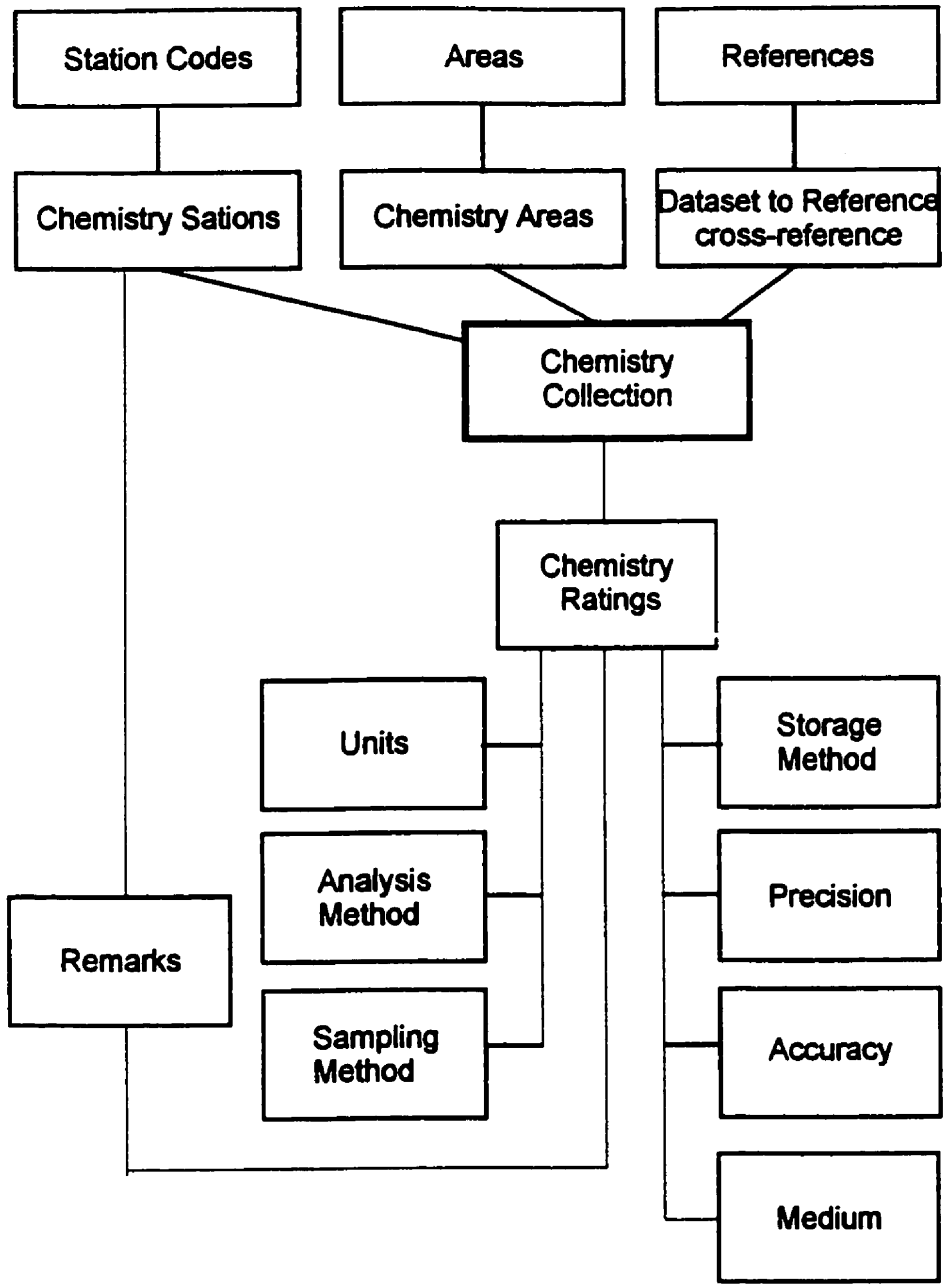


Figure 3.2 ODIS Ocean Chemistry Structure from Wainwright, (1992)

It was, therefore, decided that CODIS 1.0 be designed to formalize a data structure across disciplines and then to create a system that demonstrated effective functionality in a single discipline (Continental Chemistry) (Fyles et al., 1993a).

A critical feature of CODIS was the ability to provide effective geo-referencing of its data. In both ODIS and CODIS this was done through the use of proprietary software called QUIKMap. QUIKMap is a desktop mapping and database management program developed by Environmental Sciences Limited (ESL) in Sidney B.C. (ESL, 1988). QUIKMap has a number of unique features, which greatly aided the creation and development of CODIS. QUIKMap separates map overlays. The maps and the data are treated as separate entities. Consequently, a single map can be used to display multiple sets of data and a single set of data can be plotted on several different maps of different scales and projections (ESL, 1988). This feature facilitates the creation of databases, independent of the mapping program, but provides for the use of maps in the assembly of data. In CODIS, this feature allowed developers to derive latitude and longitude values for data using the "point and click" features provided in QUIKMap.

CODIS 1.0 was completed in 1993. Subsequently, a new version was proposed to expand the number of disciplines covered, to upgrade the software platform, and to expand functionality to all the disciplines covered. The result was CODIS 2.0 released in 1997. One of the major additions in the creation of CODIS version 2.0 was the incorporation of a new catalogue of benthic invertebrates in the Fraser River Basin (Continental Benthos). The data structure for the Continental Benthos catalogue was created at the University of Victoria. Experts at Simon Fraser University (SFU) carried out the cataloguing and inputting task. Details of this process are available in Johansen and Reis (1994).

CODIS version 1.0 was a DOS application. For CODIS version 2.0 the platform was shifted to Windows and from proprietary software packages to MSAccess. CODIS 2.0 runs under MSAccess version 7.0 for Windows95 and WindowsNT and uses QUIKMap for mapping functions (CODIS User's Manual 1997). CODIS 2.0 achieved all the original design goals and contained metadata for eight disciplines covering the Canadian Arctic, the British Columbia West Coast, and the Fraser River Basin. Within the regions and disciplines defined, the coverage was believed to be comprehensive. The metadata range

from the early 1800's to 1996, from isotope ratios to whale behaviour, from established accuracy and precision to established errors (CODIS 2.0 Users Manual, 1997).

CODIS 2.0 had a number of features common to many types of databases. Metadata could be searched; the results rapidly browsed and printed using standard reports; and all data could be mapped. User's search files could be restored or deleted and the metadata files maintained using the software. New metadata catalogues could be created using the software. CODIS 2.0 also had a number of features rarely seen in databases. The documentation was extensive and could be manipulated separately from CODIS. The metadata was accessible to all, and users were encouraged to explore the metadata using the tools of MSAccess to develop customized queries unique to each user's needs. Every aspect of the database was open and accessible to users. CODIS 2.0 is currently freely available on the World Wide Web for download and use.

A number of researchers were involved in the CODIS project. My responsibilities included a) developing the initial data structures and structuring tools, b) creating the data entry look-up lists, c) transforming the ADCAP/WESCAP data for inclusion into the system, d) developing the decision tree methodology, e) creating the Continental Chemistry decision trees and guidelines, f) appraising the Continental Chemistry data, g) testing the appraisal system, and h) producing the initial drafts of all reports. Dr. Fyles and I worked jointly in the a) development of the final data structures, b) QA/QC analysis of the Continental Chemistry and Continental Benthos data files and appraisal systems, and c) supervision of the Continental Benthos cataloging task. Other researchers assisted in locating much of the Continental Chemistry data, while data entry and software development were contracted out.

3.1.1 CODIS Design Goals

CODIS was envisioned as an accessible, multidisciplinary research tool that would speed up the retrieval of background material at the start of any monitoring program or research project in order to avoid costly duplication of data collection activities. By using a simple PC platform the objective was to provide effective environmental data management tools to a wider range of user. By using CODIS before initiating a testing program, a user was expected to be able to identify data produced from pre-existing, original research or testing. Suitable data would be located using the CODIS search features including the option of both textual and map-based searching. Output from CODIS would include both the option of on-screen viewing of both maps and text, and detailed printed reports. A first-time user approaching CODIS would expect to encounter an intuitive system that used standard scrolling lists to identify data of interest. This user would be provided with sufficient details from the metadata to determine if original data merited retrieval from their archived location, which was identified through bibliographic data supplied in the included standard reports.

CODIS was also envisioned as a tool for data cataloguers. In order to satisfy the needs of cataloguers, CODIS required tools to break down larger data sources into datasets and report the reliability of the data that made up those datasets. This involved developing a methodology to objectively appraise datasets and report the outcome of those appraisals. Once created, such datasets had to be accurately and efficiently entered into the system. Unnecessary duplication of effort had to be eliminated. By simplifying the input task cataloguers would be more likely to use the system. This required eliminating as much textual input as possible and limiting typists to the use of previously prepared lists for input. This would reduce the number of keystrokes required to input data, which would subsequently decrease the possibility of input errors. Cataloguers needed tools to confirm their input and customize the process for their particular needs.

While providing useful tools for users and cataloguers, the primary goal of CODIS was to serve as a practical platform to test information management theories. CODIS was designed as a multidisciplinary index to primary datasets rather than an archive of raw data. It included metadata that spanned the scientific disciplines represented: chemistry, benthos, fish, marine mammals and ocean physics. The datasets incorporated into CODIS were associated with indicators of their reliability that would provide users with a measure of their potential utility. This methodology developed for appraising reliability had to be both robust and objective in order to be acceptable to the scientific community. The process of inputting metadata into CODIS had to be simple, accurate and fully documented.

3.2 CODIS Structure Development

Data can vary tremendously in style and format. In order to be of future use, data must be organized (National Research Council, 1995). Structuring of data serves to standardize documentation while describing all pertinent aspects of data collection (Stafford 1994). As such, structure is an essential component of an archival database. Structure arises from applying controlled terms or vocabulary to the data. The application of controlled language is a well accepted tool in database creation and is used primarily to limit the number of alternatives that need be searched in order to identify applicable data (AXYS, 1994). As a simple example, consider the use of the Chemical Abstract Service (CAS) registry number to identify chemical compounds. Due to the complexity of nomenclature in chemistry, a single chemical may be known by a common name, a trade name, and by a formal chemical structure. The CAS registry number insures that, regardless of how the compound is named, a search of the CAS registry number will identify all occurrences of that compound in their database.

3.2.1 Data Input Lists

Data input into CODIS was carried out using previously prepared lists called look-up lists. In order to input data a cataloguer was required to go a list and pick an appropriate, pre-existing alternative. All the potential alternatives for a particular field, such as chemical contaminants, were stored in a file (in this case called "Chemical Contaminants"). Each contaminant had an associated number in the list. In order to enter data, the code number was entered in the appropriate field. This resulted in the appropriate contaminant appearing both in the file and in any subsequent report. The reason for using such lists for data input was to preserve the relational database structure, simplify searching, and to speed up the system. Like every trade-off, this process had its advantages and disadvantages. The advantages for this system were three-fold: since data was input using lists the typist required a greatly reduced number of keystrokes to enter data; numerical input meant that if the typist slipped and entered a number not on the list the computer would not accept the character thus avoiding many errors; and using look-up lists with defined terms eliminated the use of "other" as an input. In similar cataloguing efforts elsewhere the word "other" was an accepted term, and in some portions of the ADCAP/WESCAP catalogues approximately 30% of all the data was entered as "other" (Fyles et al., 1993a). The disadvantage of using previously prepared lists in input was that before a dataset could be entered in the dataset it had to appear on the list. Any value not on a look-up list at time of data entry could not be admitted to the system by the typist.

The preparation of the lists for data input required the creation of an overall data structure. Formulating this overall data structure is the initial task of any cataloguing activity that follows the approach used in CODIS. A critical feature of CODIS-like systems was that all disciplines covered should share a similar parallel structure. Each discipline had its own discipline-specific data fields while sharing common non-disciplinary system-level files with all other disciplines in the system. The parallel structure served as a template for data input and look-up list creation and was both flexible and robust. The sharing of system-level files served to decrease unnecessary duplication and improved the ability to handle multidisciplinary datasets. Two types of look-up lists were needed to

input data into CODIS: general shared lists and discipline-specific lists. The general lists were shared by all disciplines while each discipline-specific list was unique to its own discipline.

The need to develop an overall structure and look-up lists before data input placed the majority of system design effort at the start of the cataloging process. In doing so it reaped rewards of decreased input time and increased quality of data that more than made up for the initial investment. An initial task in developing CODIS was to translate the critical elements of the data source in question using a controlled vocabulary with fixed, unique definitions (West, Fyles and King, 1993). Structural qualifiers included source, location (in time and space) and data specific elements such as species and collection method. The formal process that insured the one-to-one relationship between structure elements and the controlled vocabulary was maintained through detailed protocols that defined their relationship. All metadata files were governed by protocols. In some disciplines this involved grouping of measurement techniques (thermocouples and thermometers), or media (migratory fish, to include both salmon and trout). Controlling terms available for data input limited the number of alternatives available, this reduced the need for experts to input the data.

The creation of lists for entering data entailed developing unique definitions for all terms. Since all these terms were unique, any search of these files using these terms would be "exact". Only the controlled terms could be used in an exact search so a user was guaranteed to get a result if one occurred in the file, and no result if the file truly contained no occurrence of the search term (West, Fyles and King, 1993). In contrast, searches of text strings (lists of characters) have a number of potential disadvantages. Text string searches could be slower, case sensitive ("Basin" versus "basin"), and might yield unwanted results ("cat" would also find "category"). If the textual information was only loosely defined or incomplete then these searches also ran the risk of being "inexact" and could miss occurrences in the database. The distinction between "exact" and "inexact" searches was a consequence of the metadata concept. Exact searches were those which

dealt directly with the metadata while inexact searches focussed on aspects specific to a given data file.

3.2.2 CODIS Structural Features

The structuring approach used in CODIS began by assembling data into datasets and was governed by a set of rules and guidelines (called conventions hereafter)(Fyles et al., 1993a). The entire list of these conventions was presented in Fyles et al. (1993a). The convention for collecting datasets required that the division of a large report into datasets strive to maintain the expectations of internal consistency of the original workers. It stated that this subdivision must also take into account some general realities.

> 1) When subdividing a large report into simpler datasets one should strive to maximize the size of the datasets.
> 2) When subdividing reports, the new datasets should be easy to derive from the original report.
> 3) Datasets should have uniform quality rankings. A large dataset could be fragmented to preserve quality "3" or "4" data together with "0" or "1" quality data in a separate subdivision.

A critical aspect of structure when dealing with multidisciplinary data was the process of identifying the appropriate discipline to which the dataset belonged (Fyles et al., 1993a). Disciplines were a conceptual tool that provided system designers with ability to refine their work based on the understanding, limitations and requirement of the specific branch of science (discipline). In CODIS, disciplines shared a parallel data structure but had elements that were unique to the discipline. As an example, individuals studying contaminant loads in soil samples might need a list of contaminants while marine biologists might need a taxonomic key of large marine mammals. The translation of text and ideas into code suitable for input into a system could vary in complexity depending on the discipline.

The use of disciplines in system development provided advantages for designers. One such advantage was that the existence of disciplines allowed for discipline-specific

language, methodologies and techniques. This limited the requirements of input tools and meant that appropriate lists could be targeted for each discipline. Parallel data structures across disciplines allowed for structured, formalized, discipline-specific protocols for the evaluation of the reliability of the data (West, Fyles and King, 1993). The parallel structure implicitly recognized that no two disciplines could be treated using the same protocols. A consequence of this division was that the same data source may have differing levels of detail in different disciplines but all the data would be accessible via either discipline.

CODIS was a multidisciplinary tool that could also effectively handle interdisciplinary data (the difference between the two is a fundamental one in nature). Environmental research has both an interdisciplinary and a multidisciplinary character. This results in activities that can be both multidisciplinary and interdisciplinary. Consider the publication of the cruise report for a science vessel. Over the course of the cruise, measurements might be taken of ocean currents and metal concentrations in ocean water. This research would be multidisciplinary, with the ocean current data being catalogued in one discipline (Ocean Physics) while the chemical data being catalogued in another (Ocean Chemistry). The structure developed for CODIS ensured that the two datasets preserved that relationship through the shared file structure to be discussed later. Interdisciplinary data involves data that crosses between disciplines or includes fundamental aspects of differing disciplines. Consider as an example a study detailing the concentration of a chlorinated compound in benthic invertebrates and its effects on population. The cataloguer might classify the dataset as belonging in the discipline of benthic invertebrates; while a chemist might classify it as a chemical dataset involving benthic organisms. Both of these allocations are appropriate. The structure developed for CODIS left the allocation of discipline under control of the cataloguer. The process relied on cataloguer judgement and not protocols, but cataloguing activity affected it. An effective multidisciplinary system would produce useful metadata in both cases. How a dataset was initially classified would not affect the eventual usefulness of the derived metadata.

Geographical data presented additional complications. Location may be variably defined. A botanist may need to be able to find a single, endangered plant in a small plot, while a marine biologist may only require a general map quadrant in the ocean. Using multi-tiered location descriptions, data could be associated within an appropriate level of reliability. Within CODIS, locations could be assigned by exact latitude and longitude, by a named area, or by a general location or region (Fyles et al., 1993a). As an example, consider Roberts Bank in the Strait of Georgia off Vancouver. Roberts Bank covers an area of several square kilometers in the Georgia Strait, at the mouth of the Fraser River. It can be located by name (Roberts Bank), by specific latitude and longitude (49° 123° W), by area (Fraser River Estuary and/or Georgia Strait) and by region (Fraser Basin and/or Strait of Georgia and Adjoining Waters). Areas and regions were not exclusive and as such locations like Roberts Bank could be included in more than one region or area.

3.2.3 CODIS Software Structure Considerations

The goal of the CODIS data architecture (detailed in the CODIS data dictionary) was to elucidate a framework to support an interdisciplinary information system that could be used to bridge the gaps between disciplines. This could only be accomplished if all data types could be identified and dealt with in a similar manner. Given the recognition that a standardized data structure was needed, an effort was made to investigate the commonalties of the initial data groups in the system (chemistry, physics, marine mammals, marine fish and physical oceanography) in order to identify the common features to all disciplines. The outcome of this process is displayed in Figure 3.3.

As a point of departure, it was recognized that a data source, in order to be of use, must be capable of being described by a bibliographic citation of some type (author, title, source of the information, and publication type). It was also recognized that all data sources (except laboratory studies, which were not included in the CODIS database) have information about measurements at locations; the data are inherently "geo-referenced".

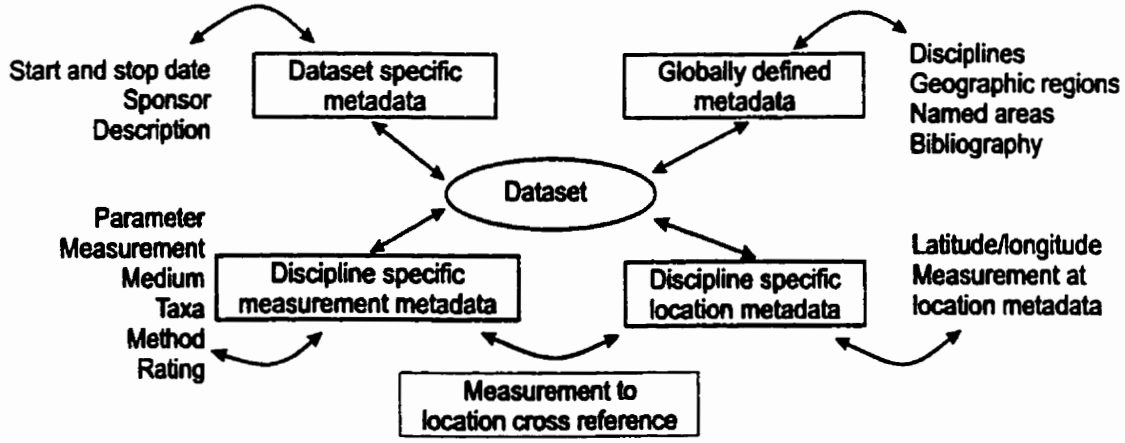


Figure 3.3 CODIS interpretation of dataset components.

Since geo-referenced data collection must involve expenditure of time, effort, and resources, CODIS assumed that there was information about the project that gave rise to the data (start and stop date, sponsor, and description of the project).

There were further assumptions about the measurements and locations components of the data structure. With respect to location, a hierarchy of detail was assumed. At the largest scale were regions; followed by named areas; and the precise latitude and longitude of sampling locations. The main assumption at this stage was that regions were geographically defined within CODIS, while areas followed place name usage. Areas could be subject to definition by external sources such as a standard gazetteer, while regions were larger and needed a controlled definition possibly unique to CODIS.

The regions used in CODIS arose from the ADCAP/WESCAP process in which coverage was partly defined by this type of geographical definition. Not all regions needed the same surface area: data-dense regions could be smaller than data-lean regions so that roughly the same amount of metadata would be associated with each region. The definition of a region also needed to accommodate conventional perceptions of the name given to the region. This led to some overlap at the boundaries of adjacent regions. Locations near such a boundary could be associated with two or more regions.

Discipline-specific metadata had two components: location metadata, and measurement metadata. Even in a multi-discipline dataset, it was extremely unlikely that exactly the same measurement at location metadata (sampling location, time, and depth) would be the same for all measurements across the disciplines. Consequently, all facets of the measurement process were handled on a discipline specific basis. The CODIS software was expected to re-synthesise the multi-discipline dataset from the discipline-specific components stored individually.

Figure 3.3 displays these minimum structural requirements. Each dataset was expected to have a start and stop date; a sponsor; and an overall description that applied to the dataset as a whole. Of these, the most unique features were the dates, and these were used as the starting point in the process of defining new CODIS datasets.

Figure 3.4 displays how this theoretical interpretation was implemented in the CODIS system. The pivotal table in Figure 3.4 is the Dataset Identification Table (DS_ID), which contained the dataset-specific information. The regions, areas, and bibliographic information tables were linked to DS_ID via cross-reference tables, which supported the many-to-many relationships in the data. The dataset to bibliography cross-reference table (DSIDXREF) also accommodated the required sorting of bibliographic information by discipline.

The programming task required that a unique identifier be used to associate datasets. In the ODIS system this unique identifier was the Dataset Identification number (DS_ID). Cataloguers assigned the DS_ID values which served as a universal pointer that guaranteed that information related to a dataset was always associated with that dataset. The historical ADCAP/WESCAP data had some duplicate usage of DS_ID, so uniqueness could not be assumed for this field. In CODIS 1.0/2.0 the DS_ID was supplanted by a unique identifier called the UNIQUE_ID. In all cases, the UNIQUE_ID was accompanied by a DS_ID, and in

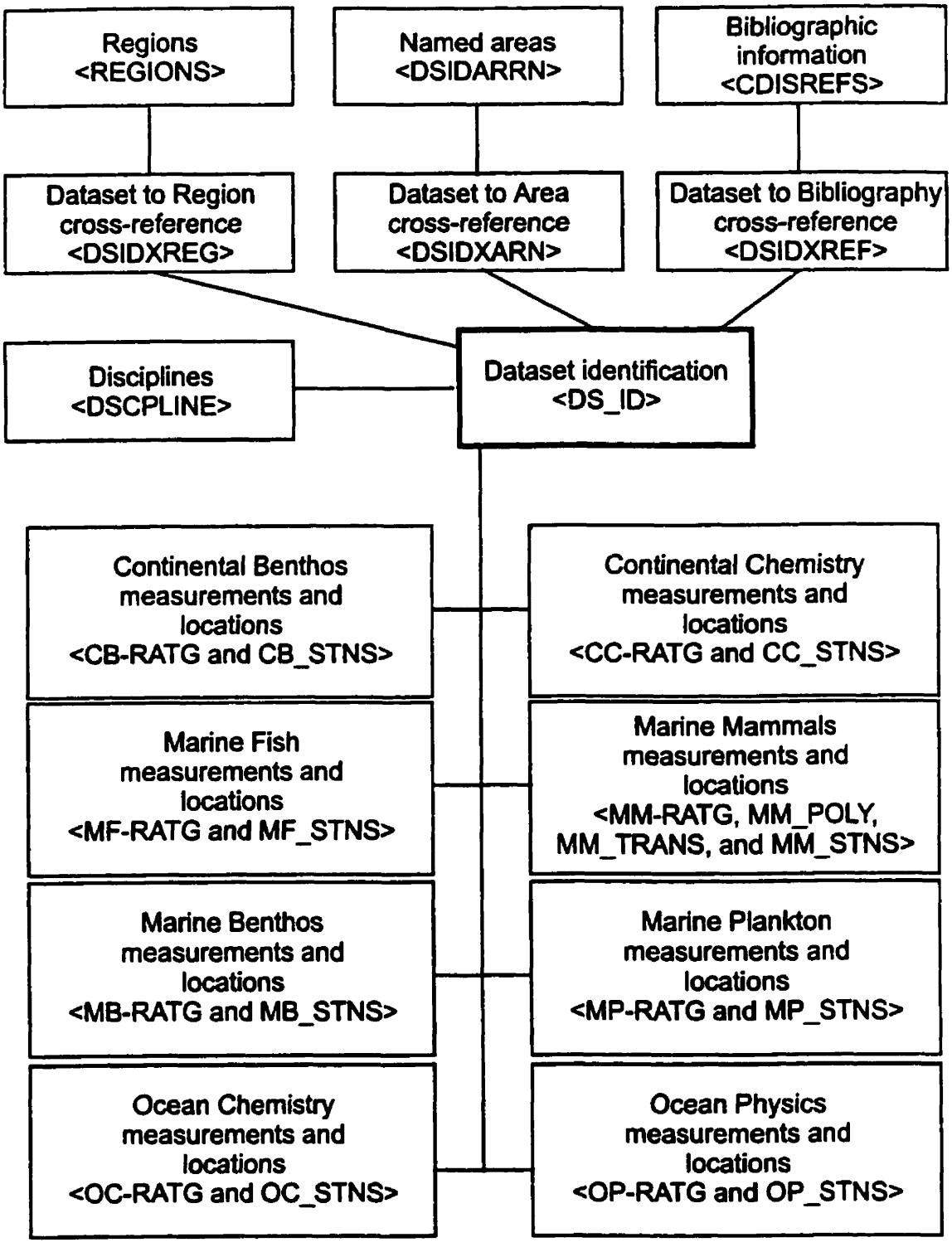


Figure 3.4 Overall Organisation of Tables in CODIS

some circumstances, the two could be used interchangeably. However, the UNIQUE_ID values were machine generated and could be guaranteed to be unique in all relationships. The DS_ID was useful for reporting and provided the linkage to the historical data, so it was preserved.

Associated with each dataset were globally defined metadata. These were broad features of the metadata, which were used in different combinations by all datasets in CODIS. These included the list of disciplines, the list of regions, the list of named areas, and the bibliography component of CODIS. A single dataset may involve multiple disciplines typically at a common set of locations (regions, areas), but the results might have been published on a discipline-specific basis. Consequently, the bibliography needed to incorporate a discipline selection. Maintaining a common bibliographic component was judged to be more effective than trying to maintain multiple discipline-specific bibliographies. When a report was generated the globally defined metadata would use the discipline selection to identify which references to include and in what order. Thus, a chemistry report that included previously published benthic observations would appear as the primary reference in a chemistry report but as a secondary reference when reporting the benthic data.

The hierarchy of measurements was more complex. At the highest level were disciplines. These were defined conventionally with the understanding that some duplication must arise at discipline boundaries. Within a discipline were groups of similar measurements called parameters, which divided all the potential measurements into 8-12 large groups. In the chemistry disciplines these parameters were large groups of chemical compounds that shared chemical characteristics. Most life sciences disciplines had similar parameters based on common criteria of diet or morphology and these differed from parameters in the physical sciences disciplines. Measurements were made with respect to some organism or medium. In the life sciences disciplines, the taxonomic classification of the organism was a critical element of the metadata. In the physical sciences, the comparable element of

metadata was the medium of the measurement. In all disciplines the method used to make the measurement and the data appraisal rating were included in the metadata.

The lower part of Figure 3.4 implies the relationships with the discipline-specific tables. The figure uses two letter codes to represent disciplines. These shortened codes were a requirement of disk operating software at the time, which only allowed for table names of eight characters or less. The full list of codes was: CB, Continental Benthos; CC, Continental Chemistry; MF, Marine Fish; MM, Marine Mammals; MB, Marine Benthos; MP, Marine Plankton; OC, Ocean Chemistry; OP, Ocean Physics. In order to simplify the discussion an additional code (XX) is used in the remainder of this work. XX is used in cases where all disciplines share a common table name.

Each of the eight disciplines had two key tables linked to the DS_ID table. One table (XX_RATG) contained and directed relationships to the discipline-specific measurement metadata. The other played the same role for the discipline-specific location metadata. The lower part of Figure 3.4 also implies that the eight disciplines in CODIS shared common data structures. The differences between the life sciences and physical sciences have been noted above. The extent of these differences made directly parallel structures in all disciplines impossible. Nonetheless, there existed basic similarities shared by all disciplines. These common features are illustrated in Figure 3.5.

All disciplines had location metadata (called stations in CODIS) stored in a table named XX_STNS, and measurement metadata (called ratings information) stored in table XX_RATG. The key field in the XX_STNS file was a unique identifier for individual stations (XXSTNS_ID); the parallel field in the XX_RATG file is the XXRATG_ID. These two keys supported the relationships of the discipline specific information to the UNIQUE_ID.

The locations metadata was largely contained in the XX_STNS Table. The measurement metadata was partly contained in the XX_RATG Table, and partly related to tables for the

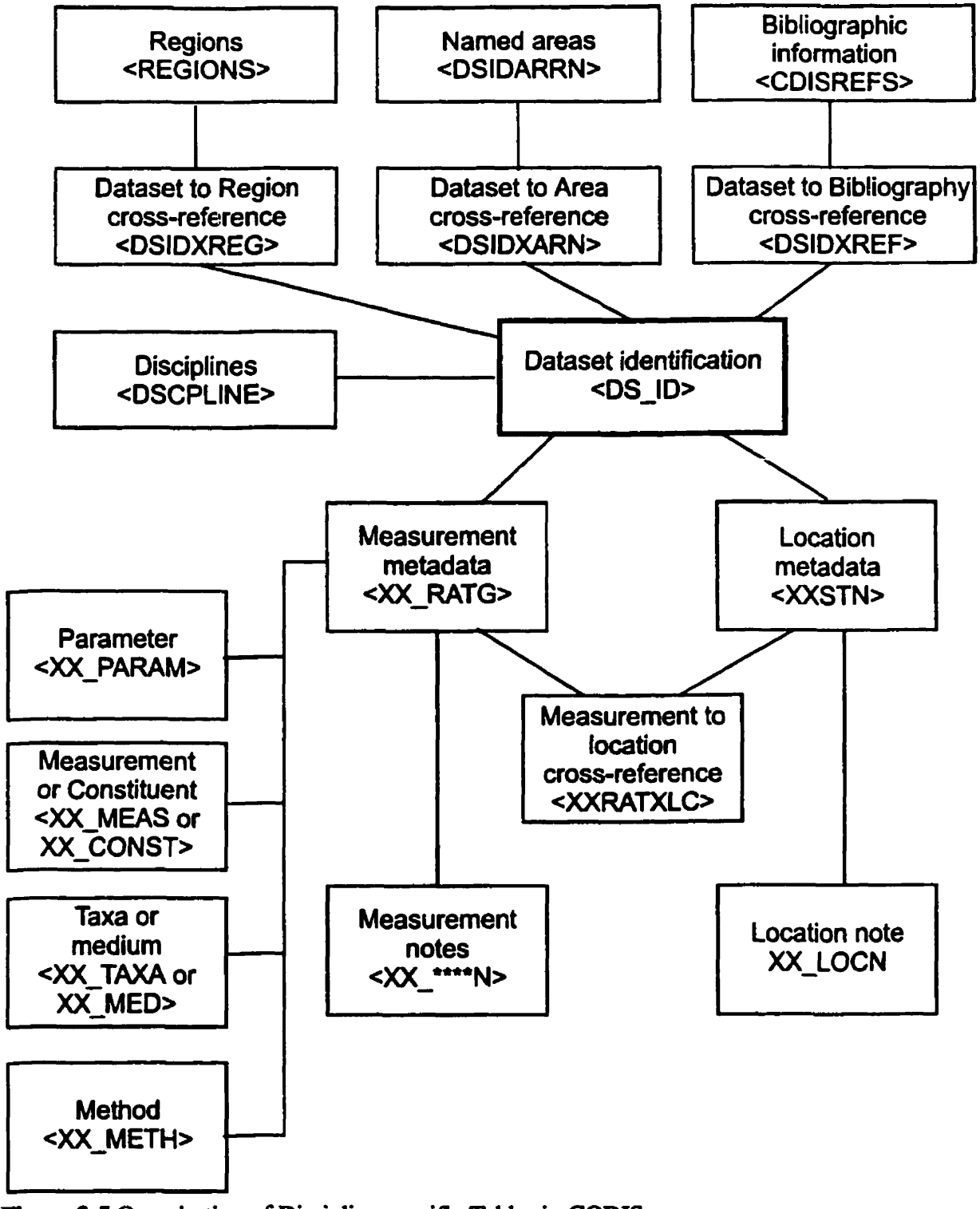


Figure 3.5 Organisation of Discipline-specific Tables in CODIS.

discipline, parameter, measurement or constituent, taxa or medium, and method. These comprised the metadata for the dataset, and were the basis for the searches CODIS

supported. There were several additional tables associated with the locations and measurement metadata tables that provided additional information, which could be reported. These additional items were stored in "notes" tables. The notes tables were intended for reporting but could be subjected to text searches if required.

Discipline specific tables (XXRATXLC) supported the relationship between individual measurements and individual locations (measurement-to-location relationship). These linked one-or-more unique measurements (XXRATG_ID) to one-or-more unique stations (XXSTN_ID). In certain disciplines, some historical data had this information embedded into the measurement metadata (Ocean Physics), some had it explicitly represented (Benthos and Plankton), and some ignored it entirely. A similar situation existed for the method metadata tables in all disciplines. Historically, some cataloguers attempted to systematise this aspect of the metadata, while others regarded it as only information to report, akin to the notes tables. The CODIS structure required that this should be controlled metadata, and incorporated it, even where the available information was poorly systematised.

In order to create the CODIS tables it became necessary to reconstruct the original ODIS data files. This activity consisted of taking the original files and manually translating them for use by the new data system. This process, while time consuming, was especially useful for model development as it highlighted a number of significant data problems that had to be faced in order to create a functional multidisciplinary information system. The following sections discuss some of the problems, discovered through this and an earlier process, and considers them in light of the overall goal of designing an effective tool for environmental information management.

3.2.4 Data Structure Development

The most significant issue faced by the chemistry cataloguers was the development of a complete constituents list. A constituent was a CODIS term for a chemical compound,

mixture of compounds, or chemical property that had been measured (Fyles et al., 1993a). As a starting point, there existed a constituent file from the PC-ODIS Ocean Chemistry discipline. This list was derived from a data dump of all the chemical files included in the system. It was essentially unstructured and was related to a parameter file which broke the compounds down into nine parameters, these being:

1) Hydrocarbons;
2) Organochlorines;
3) Metals;
4) Pigments;
5) N-, P-, Si-based Nutrients;
6) Dissolved gases;
7) Isotopes and Isotopic Ratios;
8) C-H-N-P; and
9) Other (Wainwright, 1991).

While each parameter list contained some compounds, 30% of the overall list was held in the "Other" category. Given the lack of structure in the list and the existence of the "other" both as a category and a constituent, there were numerous duplicates and the system was essentially unusable for the input of new datasets.

Each constituent was subsequently examined and uniquely related to the CODIS parameters. Each constituent was ultimately defined by its Chemical Abstract Service (CAS) registry number (pure compounds or defined mixtures) or by a protocol definition where no CAS registry number could be found. Compounds having a variety of common names, synonyms, acronyms or trade names were combined under a single constituent name. In order to simplify the process, an additional level of division was incorporated into the process (the parameter group), which will be discussed in more detail later. The new CODIS definitions for parameter and parameter group were:

> **Parameter** A CODIS term for large groups of chemical constituents. The parameters recognized by CODIS are: "Hydrocarbons", "Chlorinated organic compounds", "Inorganic", "Natural Products", "N,P, and Si nutrients", "Gases", "Isotopes and isotope ratios", "C,H,N,O,P,S and X compounds", and "Properties".
>
> **Parameter group** An organizational level between parameters and constituents used by CODIS when necessary to divide the total list of constituents into manageable

groups using definitions based on structural or functional similarities (Fyles et al., 1993a).

The parameters and their associated groups are presented in Table 3.1. Each of these parameter groups and constituents was strictly defined in Fyles et al. (1993a).

The constituent list was one of the early products of this research program. It underwent scrutiny at the University of Victoria and was then reviewed and expanded by researchers at the Federal Departments of the Environment and Fisheries as well as the provincial Ministry of Environment. A revised version of the list was subsequently presented for review at a workshop held at the Institute of Ocean Sciences in Sidney B.C. of May 18-20th, 1993. The aim of the workshop was to develop protocols for a new National Contaminants Information System (NCIS). It was attended by 18 DFO, Environment Canada and private sector experts from across Canada. At the workshop several different models were presented for categorizing chemical constituents these included the CODIS scheme as well as the ADCAP/WESCAP system mentioned previously and two additional systems: The Environment Canada-ENVIRODAT scheme and the Puget Sound Ambient Monitoring Program PSAMP file (Blyth, Thomas and Gormican, 1993). After a discussion the Workshop chose to use the CODIS list.

The systemization process used to create the constituent list was subsequently carried out for the list of media. While the constituents list had an obvious pre-existing structure, no obvious structure existed for potential media in which an organic contaminant might be found. The creation of this list, therefore, was carried out in an *ad hoc* manner. A list was made of the most common media in which organic contaminants were measured. As new data were collected for the database, new categories were identified, comprehensively defined, and if necessary, added to the list. The outcome of this process was a comprehensive media list.

When CODIS was expanded to create a Continental Benthos discipline a number of data issues had to be addressed. The ODIS Marine Benthos data structure was unworkable.

Table 3.1 Chemical Parameters and Groups in CODIS 1.0

Parameter	Parameter Group Name
Hydrocarbons	Aliphatic Hydrocarbons
	Aromatic Hydrocarbons
	Polycyclic Aromatic Hydrocarbons
	Hydrocarbon Materials
Chlorinated Organic Compounds	Chlorinated Pesticides
	PCB's
	Chlorinated Compounds
	Dioxins and Furans
	Chlorinated Phenolic Compounds
Inorganic	Metals
	Non-Metals
	Inorganic Compounds
Natural Products	Natural Products
N,P and Si Nutrients	N,P and Si Nutrients
Gases	Gases
Isotopes	Isotopes
C,H,N,O,P,S and X Compounds	Phenolic Compounds
	Phthalates
	Resin and Fatty Acids/Esters
	Organophosphates
	Carbamates
	Triazines
	Herbicides and Fungicides
	Anti-Sapstains
	C,H,N,O,P,S and X Compounds
Properties	Properties

It included a number of files which were intended to serve as a cross-references between records (Wainwright, 1991). The existence of these files made the ODIS look-up lists unsalvageable for use in the creation of the CODIS Continental Benthos catalogue. As a result, the Continental Benthos cataloguers were required to create look-up lists in an *ad hoc* manner.

A serious difficulty arose in the development of the parameter-measurement relationships. In CODIS 1.0, all parameter-constituent (the equivalent relationship in chemistry) relationships were strictly defined. This requirement was a hallmark of the CODIS approach. This strict control was missing in the benthic discipline and resulted in difficulties in determining why a measurement was associated with one parameter and not another. As an example, the measurement "density and identification" was associated with the parameter "density" while "identification and enumeration" was associated with "identification". Moreover, some datasets included "density", "enumeration" and "identification" as three separate measurements (Fyles and King, 1994). An examination of the original data appeared to indicate that the difference involved the association of data with a specific location, versus composite station information (Fyles and King, 1994). This depth of insight would not be available to the first-time or naïve user and would pose serious problems for users interested in identification information, as several parameters are involved (Fyles and King, 1994).

Another issue arose in the creation of note fields. As mentioned previously, CODIS used note tables in order to provide commentary regarding specific datasets. The ODIS note files were unstructured which resulted in huge lists that included entries that varied in length from one or two characters up to 762 characters (the maximum note-file length in the ODIS database). This variety in sizes made the creation of standardized reports difficult. Large note fields also detracted from the quality of the data holdings as cataloguers occasionally chose to include information in a commentary rather than including it in the appropriate fields in the database. In the CODIS approach, note fields

were contained in look-up lists that could only be added by the data manager and not by individual cataloguers. This use of look-up lists sharply decreased the freedom of expression in the note fields.

3.2.5 Data Structure Analysis

An important issue in developing look-up lists for input into a data system was the "coarseness" of the groupings used. Coarseness has a profound effect on the usefulness and user-friendliness of a system. Users who carry out unsuccessful searches become disaffected with the system, while users who receive multiple, useless, "hits" become frustrated. Both problems decrease the likelihood that the user will use the system again. Cataloguers, on the other hand, want clear divisions for cataloguing and resent systems that require unnecessary, time-consuming, scrolling when inputting data. Both groups resent non-intuitive groupings and structuring systems that "lose" their data.

The aim in developing CODIS was to structure the data in intuitive groupings of similar data-density. As an example, consider the breakdown of the major locations (called regions). As mentioned previously in the discussion on locations, not all regions were alike. In British Columbia there exist a number of distinct geographic regions that are well recognized by most researchers. In addition, much more work on organic contaminants had been carried out in the population rich southwestern corner of the province (the lower Fraser River area) than in the much less densely populated central portion of the province. CODIS acknowledged these two considerations by breaking the data-rich southern areas into a number of smaller data-dense regions while retaining the geographic integrity of the larger data-poor regions. The outcome of this process is illustrated in Table 3.2.

While the data-density was by no means even, it was broken down in a manner that was intuitive for the user. Greater balance was evident in the distribution of the chemical

parameters and constituents. Table 3.3 indicates the distribution of the Fraser Basin organic chemistry by parameter.

Table 3.2 CODIS 1.0 Dataset Density by Region

Station Regions	Number of Datasets
Lower Fraser River	393
Middle Fraser River	73
Upper Fraser River	27
Thompson Sub-Basin	62
Nechako Sub-Basin	19
Fraser River Estuary	265
Georgia Strait Offshore of Fraser River	115

Table 3.3 Fraser Basin Datasets by Parameter

Parameter	Number of Datasets
Hydrocarbons	131
Chlorinated Organic Compounds	201
Inorganic	25
Natural Products	133
N, P and Si Nutrients	0
Gases	0
Isotopes and Isotope Ratios	0
C,H,N,O,P,s and X Compounds	293
Properties	3

The entries under the Inorganic parameter refer to organic carbon and total organic carbon. Both these constituents were placed in the Inorganic parameter to place them in proximity to total carbon, total inorganic carbon, graphite, and the other elemental forms. The entries under the Properties parameter reflect a single constituent (volatile residue) that did not fit in any of the other groupings.

A difficulty arose, however, when it came to entering constituents into the system. The original constituents list in CODIS was very long (containing 812 constituents) as were some of the parameter lists (Chlorinated Organic Compounds had 176 constituents). This made data input frustrating as cataloguers were forced to scroll through pages of constituents in order to track down the appropriate one. As a result, it was necessary to create a sub-grouping called a "parameter group" which was used for input and is presented in Table 3.1. While the difficulty with data coarseness in the constituent list was recognized early in the data input stage, the same was not so for the media list. Table 3.4 displays the density of datasets in CODIS 1.0 by medium. It indicates a potential problem with the protocols for media. Although each medium had a precise definition, it is obvious that there were too many choices of media for fish. As a result, users of CODIS had to be cautioned to use a broad search of several related fish media together in order to ensure that they did not miss interesting data.

In summary, coarseness of data groupings is an issue for any process that assembles data. If the data screens are too large then groups become overfilled and searches become impractical as was the case with the constituents list. Too fine a screen results in underpopulated fields as was seen in the media list. Experience creating CODIS demonstrated that of the two choices the smaller screen was preferable. It was significantly easier to lump together several smaller groups than it was to break a larger group into smaller substituents. In the case of the media list, it would be relatively simple to assemble the 14 fish tissue sub-groups into three or four larger groupings. The act of breaking down the constituents list, however, was both time-consuming and difficult.

The understanding developed in CODIS on data structuring served to improve subsequent systems. The DFO-NCIS workshop discussed later, benefited from the lessons learned in the development of the CODIS media list. As a result, a carefully excised media list was created for use in NCIS.

Table 3.4 Fraser Basin Datasets by Medium

Medium	#	Medium	#
River/Lake Water	132	Fish	19
Estuarine Water	20	Fish-Muscle	8
Sea Water	14	Fish-Liver	14
Sea/Est./River Water	2	Fish-Tissue	5
Interstitial Water	1	Fish-Gill	1
Run Off/Leacheate	38	Fish-Bile	0
Waste Water	68	Migratory Fish	10
Well Water	158	Migratory Fish-Muscle	6
Liquid	3	Migratory Fish-Liver	6
Sediments	65	Migratory Fish-Tissue	3
Bottom Sediments	2	Semi-Migratory Fish	7
Intertidal Sediments	2	Semi-Migratory Fish-Muscle	2
Beach Sediments	0	Semi-Migratory Fish-Liver	1
Dredged Sediments	1	Semi-Migratory Fish-Tissue	1
Subtidal Sediments	1	Non-Migratory Fish	8
Soil	4	Non-Migratory Fish-Muscle	3
Solids	3	Non-Migratory Fish-Liver	1
Biota	14	Non-Migratory Fish-Tissue	1
Benthos	8	Birds	2
Plants	0	Aquatic Birds	5
Amphibians	0	Non-Aquatic Birds	0
Reptiles	0	Mammals	0
Aquatic Mammals	0	Non-Aquatic Mammals	0

3.2.6 Initial Quality Assurance/Quality Control of CODIS Continental Chemistry Files

The goal of the Quality Assurance/Quality Control (QA/QC) program in CODIS was to ensure that the data was both correctly abstracted from the original source and that data entry was correct. The key question was - how accurately did the CODIS generated report reflect the original data source? Accumulation of errors could occur throughout the process: during transcription from the data sources to the standardized forms used for input, during data quality appraisal and decisions about divisions of data sources into unique datasets, as well as during data entry.

Once the Fraser Basin data was completely entered, it was possible to explore the errors incorporated in the CODIS files. This involved three steps:

1) selection of a random subset of the DS_ID file and generation of dataset ID, ratings and station reports for each of the selected dataset IDs.
2) field-by-field comparison of the information in the CODIS reports with the information in the original data source documentation, and with the information on the standardized forms used for data input.
3) tabulation of the errors detected using a standard method to count the errors.

The first step in the QA/QC process was the random choice of twenty datasets. These datasets were chosen by ranking all 578 datasets plus subsets in increasing numerical/alphabetical order and then using a standard pseudo-random number generator (QBasic, default seed) to choose twenty random numbers between 1 and 578. A further 20 dataset IDs were identified for subsequent use at a later stage using the next 20 numbers generated.

Errors were detected by comparison of the original data source to the printed report, using a field-by-field comparison, as well as a full analysis of the standardized data entry forms. This step also checked the data quality appraisal step and the division of the data source into datasets. Errors were assigned to various categories depending on which type of fields in the final report were affected by the error, and the total number of fields in the

report affected by the error. The main categories were report header, ratings fields, station fields, and data product fields. The following division was used:

Report header fields = 9:
Start date, Stop date, dataset ID, Interim dataset ID, system manager note, CONT_CHEM flag, reference code, status of document, collection area

Ratings fields = 17:
overall quality rating, medium sampled, parameter, constituent, # of stations, # of samples, five quality ratings and five comments, remarks

Data product fields = 7:
units of measurement, detection limit, number of samples > detection limit, minimum, maximum, mean, median

Station and QuikMap fields = 17:
station key, station ID, station code, start date at station, stop date at station, remark, sampling depth, maximum depth, sampling time, number of samples at station, latitude, longitude, QM data type, QM symbol code, QM symbol colour, QM symbol thickness, label

Total fields: 50
Six fields were commonly not used (mean, median, station code, sample depth, maximum depth, sampling time) and three other fields usually took default values (QM data type, QM symbol thickness, QM symbol colour)

One factor that greatly aided the QC process was the relationship of typist to dataset provided by the Unique Identifier function. This allowed two systematic errors particular to typists and document types to be uncovered. In the first twenty reports, the most common error was clearly linked to one of the typists and a particular report type. On further review, it was discovered that whenever that particular typist came across a constituent not currently on the constituent list that typist would simply omit to include that constituent and all its line items. Once noted, this problem was easily solved by identifying all the datasets entered by that individual and confirming the numbers of line items in each dataset with reference to the original data entry forms. The result was that all further omissions of that type were eliminated from the file as a whole. A summary of the error classification is presented in Table 3.5

Table 3.5 Classification of Errors in the Continental Chemistry Catalogue

Error type	Maximum number of affected fields	Error assigned to.
Typo in a text field	1	Header
Typo in a code field	1	Field as defined above
Typo or omission in a data field	Varies: counted on the ratings report	Data
Omit a medium/constituent	24	17 to Ratings + 5 to Data
Error in compilation or data rating	Varies: counted on the ratings report	Ratings
Duplicate a medium/constituent	24	17 to Ratings

The second systematic error was the result of a misunderstanding by one typist, in a set of 138 datasets. In this group there was a systematic technique used to indicate the occurrence of a test where no pesticides were detected. This resulted in one misnamed constituent in the dataset that was repeated 138 times. This systematic error was identified in the first QA/QC step and once identified, was corrected.

This first QA/QC step did not eliminate all the errors, but the number of remaining errors was small. A second group of nineteen DS_IDs was examined to improve the sampling statistics. While the statistical significance of 39 datasets from a population of 576 appeared to be adequate, the number of fields checked was inadequate. In this case, while 6.77% of the datasets had been checked only 4.2% of the fields in the CC_RAT.DBF [the ratings files were stored in Dbase IV format] file had been checked. Whilst the average dataset had about 8 line entries in the CC_RAT.DBF file, there was a small group of much larger datasets that had not been properly sampled by the random selection process used. Clearly, there were a few big "nuggets in the sand" of smaller datasets. A group of the 25 largest datasets had an average of 42 constituent/medium lines and contained 22% of the rating fields in the CC_RAT.DBF file. Using a random number generator, 3 of the largest 25

datasets were identified and checked as previously, and the combined 42 datasets were then used to determine the overall error rate.

The overall error rate in the Fraser River Continental Chemistry files was estimated to be about 1.3%. The overall figure masks the very low error rates in the stations and header fields (0 and 0.03% respectively). The most significant error was a ratings error for the precision of a particular study (4 assigned in place of 2), but this had no consequence for the overall rating (2 due to undocumented sampling and storage). This was a single error at the entry form stage, but it was replicated 31 times for the 31 constituent/medium lines in the CC_RAT.DBF file.

Other abundant errors were omissions of means reported in the original data source, and to a lesser degree the omission or confusion of the maximum detected. These, and the remaining data product errors, were relatively unimportant, as they did not influence the quality of the metadata. Confusion over which constituent code to use might have been an historical result of the evolution of the constituent codes list as the cataloguing was proceeding.

The total number of metadata errors directly attributable to the cataloguing task was only 5 in an estimated total of several thousand cataloguing decisions. No errors were detected in station location information, no errors having to do with incorrect medium or parameter were detected, and no errors relating to overall data quality were detected. Thus, a user of CODIS Continental Chemistry would be certain to locate the information sought, but would run a risk of error of about 1-2% in using the summary data provided on the CODIS report in place of the data of the original data source.

3.2.7 Initial Error Analysis of the CODIS Continental Benthos Files

There were 168 DS_IDs assigned to the Continental Benthos catalogue. Using the same technique as described above a set of 18 was selected for analysis. An examination of the overall file structure indicated that unlike the Continental Chemistry catalogue, in the Continental Benthos catalogue no datasets involved more than 18 different measurements (2.4% of the CB_RATG file). While many datasets involved large number of taxa, the 18 DS_IDs showed averages that were comparable to the global averages of the overall data files.

Each DS_ID was in turn viewed using the "View Datasets" function. There were six types of entry positions for the data entry tasks:

1) Dataset Identification	8 fields
2) Regions	1 or more fields
3) References	1 or more fields
4) Locations	13 fields plus 15 QUIKMap fields per station
5) Ratings	10 fields per measurement per assigned station
6) Taxa	1 field per taxon per assigned station

The QUIKMap fields were set by the program to default values. Once the program had assigned a DS_ID, the relationships with all the other forms were maintained. Thus, mixing between DS_IDs in principle could not occur (and in fact none were detected in the Benthos catalogue). The data structure permitted measurements to be "assigned" to station locations rather than at the full dataset level. This provided added detail at the station level. Moreover, the taxa could be assigned to each measurement, allowing the taxa at a station to be uniquely known to the user. This complicated the QA/QC analysis as the number of fields was determined by the data, not by the file structures. The entry forms for each of the datasets selected were viewed in turn: accounting for all the assigned stations and taxa, a total of 217 forms were examined, and compared with the primary documents and the rating information provided by the appraisers. Table 3.6 summarizes the outcome of the process.

Table 3.6 QA/QC results for Continental Benthos data

Field Name	Errors	Number of Fields Examined	Percentage of errors
Identification	5	180	2.8%
Regions	1	26	3.85%
References	1	18	5.5%
Locations	1	572	0.17%
Ratings	4	780	0.51%
Taxa	0	1529	0

In addition to the errors in Table 3.6, there were a number of systematic issues that might be considered as errors. For example: many of the primary documents included depth of sampling, but the appropriate location field was virtually never used. Clearly, the cataloguers decided to ignore this type of information. Similarly, the vast majority of ratings forms examined did not have a "measurement method" assigned. In many cases the measurement was obvious from the nature of the study, but a default method might have been assigned in these cases. Both of these systematic errors are ignored in the analysis.

Of the 3015 fields plus taxa examined, only 12 errors were detected for an overall error rate of 0.39%. There were no errors detected in taxa, even though the data entry of taxonomic information was very tedious. The errors in the ratings fields were all of the same type – the wrong parameter was assigned to the field measurement. This was a result of the copy-append function of the data entry process. In principle, this error could be avoided using software, as the assignment of measurement to a parameter was strictly controlled. Consequently, this field could be automatically generated. The remaining errors dealt with missing information type and were probably the result of legitimate entry mistakes.

As with the Continental Chemistry data, this error analysis indicated the great strength of using controlled lists for data entry. The highest error rates were in "note" fields where the data entry person had to type directly into a field. The low error rates were associated with rigidly fixed protocols.

3.2.8 Comparison of Error Rates with Other Systems

The previous sections have detailed the approach taken to structuring data in CODIS. They have demonstrated that the approach could be applied to a functional system. What remained unanswered, however, was whether this new approach added value beyond conventional databases. The results of the case study indicated that the approach to structuring data used in CODIS improved on traditional systems in a number of areas. The discussion below details these improvements. Concrete examples are provided of how the approach improved on aspects of comparable systems. The most compelling proof, however, lay in the fact that given several competing data structuring methodologies presented for their new NCIS system, the DFO experts chose the approach used in CODIS.

The use of look-up lists produced significant increases in input speed over comparable text fields. Since data cataloguers entered code numbers in lieu of text fields the number of keystrokes needed to input each individual dataset was greatly decreased. A single, two-keystroke input replaced entering a 250-character reference. Since this reference had to be keyed into the look-up list to begin, the benefits of this system were primarily seen when a number of datasets shared a common set of references. In the CODIS Continental Chemistry file, one reference was associated with 138 datasets while the majority of datasets had shared references. Data input using look-up lists increased the input speed but had an even more significant effect on input quality since the system exasperated errors. When an incorrect code was input for a measurement, an entirely different

measurement (associated with the actual code punched in) would be displayed on the screen. This evident error could be quickly identified and corrected.

Compare the system used in CODIS with alternative systems used elsewhere. In the Forest Science Data Bank (FSDB), developed for the Qualitative Sciences Group of the University of Oregon, data quality control was ensured through dual entry, with data being entered twice by different key operators (Stafford, 1993). Mismatches between the two entries were flagged and the original fields checked for accuracy (Stafford, 1993). Once the data were entered, a program compared the data elements to the formats in the metadata control file. There was no estimate given for transcription error from documents as that was the responsibility of the researchers who provided the data (Stafford, 1993). Clay (1997) estimated that every time data are transcribed for information storage, up to 5% errors are introduced. Another methodology suggested to limit errors in a data system was used by the Louisiana Department of Environmental Quality, Office of Water Resources in their Water Quality Network (WQN) (Hindrichs, 1998). In WQN, data was entered daily. At the end of the month all the data were printed out and manually checked against each station's field/lab sheet to identify and correct errors (Hindrichs, 1998). Any errors were then referred to an Environmental Quality Specialist for cross-checking and correction (Hindrichs, 1998). Users of CODIS avoided these time consuming tasks. In neither system presented, was an overall error rate provided with the documentation. The implication was that the final error rate was zero.

By reducing the number of keystrokes required for data entry and eliminating the need for duplicate entry procedures, the approach used in CODIS increased data input speed and accuracy when compared to similar systems in use in other jurisdictions. The reporting of overall data reliability in the system by CODIS appears to be a unique feature of this system. While others systems carry out very detailed QA/QC of their data, the author was unable to find any reported reliability figures for data in these or any similar systems.

3.2.9 Analysis of Overall Data Structure in CODIS

The analyses of the individual disciplines in CODIS established the effectiveness of the structuring task for the purpose of cataloguing. Through the input of chemical and biological datasets it was been demonstrated that the structure could effectively handle discipline–specific, multidisciplinary and interdisciplinary data and that all three types of datasets could be entered into the system and produced useful metadata. Once the system was shown to be effective for cataloguing, the next task was to demonstrate that the CODIS data structure effectively facilitated the task of searching for appropriate datasets. This process could be carried out through direct testing of CODIS and through virtual experiments. These experiments consisted of considering theoretical and real experiments carried out in environmental matrices in order to determine how effectively the CODIS structure could identify the data required by these projects.

The first virtual test involved a scenario based on the current data holdings in CODIS. In that case the question posed considered an analysis of the interaction of chlorinated organic contaminants in water and sediment with benthic organisms. Bio-accumulation, and effects on benthic populations was established in the 1970s from observations at a variety of locations (Foehrenbach, 1971; Duke, 1970; Tagatz, 1982). A researcher seeking to carry out direct research in this field would be looking for datasets involving both organic contaminants and benthic measurements. A search in the CODIS Benthos catalogue revealed that only two datasets involved concurrent measurements of organic contaminants and benthic measurements: a survey report from 1972/73 (Albright, 1975), and a monitoring report from 1989 (Swain, 1989). These reports would be the only publications available from a traditional search of the literature.

Using the structure elements developed for CODIS, additional datasets could be identified. The multidisciplinary nature of the CODIS structure allows a user to search for concurrent datasets, datasets containing either benthic or chemical data in close proximity to each other. Such a search resulted in the identification of 133 station locations derived from 53 datasets

from the Chemistry catalogue, which contained sediment and water levels of organochlorines. A similar search of the Benthic catalogue identified 25 benthos datasets covering 41 direct station locations. A station-by-station inspection of the benthos stations for chemistry stations in close proximity (< 1 km), and close in time (< 2 years), revealed 31 of the 41 benthos stations could be supported by the chemistry data. The overlaps were distributed widely over the map area, and occurred in roughly 5-yearly intervals (1972, 1977, 1983/84, 1989). Thus, using traditional research methods, insufficient data would be available to investigate the problem but through the use of metadata significantly more information could be uncovered.

This example illustrates the potential of metadata to use an imperfect and fragmented historical record to provide a composite, multidisciplinary picture. Clearly, if organisms were not analyzed in the past, no amount of modern manipulation could create data. Thus, the data to examine bio-accumulation was simply not available apart from the two datasets with concurrent benthic and chemistry data. However, the analysis of the metadata suggested that the historical data would support an analysis of ecosystem effects on benthic speciation, population, and community structure in the Fraser Basin estuary. The primary data still needed to be analyzed in detail, using statistical tools appropriate for the diversity of the data represented, but the metadata greatly facilitated this task. The example identified the specific locations to be pooled, and ensured that the primary data would support the trend-analysis required. The metadata also provided bibliographic references. Alternative searches of the metadata would provide even greater refinement, such as pooling by a specific measurement type, by a specific group of species, or within restricted time intervals.

Another method of determining whether the CODIS data structure could accommodate environmental data was carried out by examining how the structure handled published reports over a much broader range of science than was seen in the CODIS data. Consequently, a test was conceived that involved taking articles from the scientific literature and determining whether the CODIS data structure could satisfy their basic information needs. In this test seven issues of the journal *Ecology* (1998-1999) were obtained. Each issue contained between 24 and 30 individual articles. With the first issue

the initial ten articles were chosen and tested, in the following six issues three of the first ten articles were randomly chosen resulting in a total of 28 test articles.

The test consisted of examining each article and identifying the parameter, measurement, taxa and other relevant metadata elements. This information was then compared to the data structure to determine if any critical metadata was being excluded or missed. The aim of the test was to determine whether the data structure of CODIS would be able to identify any appropriate preliminary data in the literature. The outcome of this experiment was the determination that in each case the structure could effectively identify useful pre-exiting data. In order to carry out these tests a number of assumptions had to be made. Several of the articles included data from disciplines not currently covered in CODIS. A test was considered successful if the information content of a paper could be transferable to the general structure of CODIS, assuming appropriate new disciplines were completed. As an example, consider Harrison's (1999) paper on the local diversity of herbs in a patchy landscape. In order to be included in CODIS this paper would require the creation of a new discipline "Botany". This new discipline would require a taxon list that included the 50+ species of herbs noted in the paper. It would require the creation of a parameter called "diversity" with a number of specific measurements including "diversity on patches", "diversity on continuous areas", "diversity and soil calcium", and "diversity at elevation".

Of particular interest to users of CODIS was a tool developed for the first test: the proximity search. A proximity search is a direct search of metadata and involves identifying whether two or more activities were carried out within a given area of time and/or space. This can be done because the metadata includes ranges of times and locations. Thus, a proximity search has the potential to identify where significant overlaps exist in time and space between useful datasets. This feature is seen in many other non-metadata systems but the nature of metadata systems (being comprehensive in time and space) accentuate the advantages provided. The ability to carry out proximity searches provides a powerful tool for the analysis of environmental variables particularly when data

is associated with accurate location information, as in advanced GIS systems. In summary, the CODIS structure provided direct for direct searches of the literature and in addition allowed for secondary searches based on proximity in time and space.

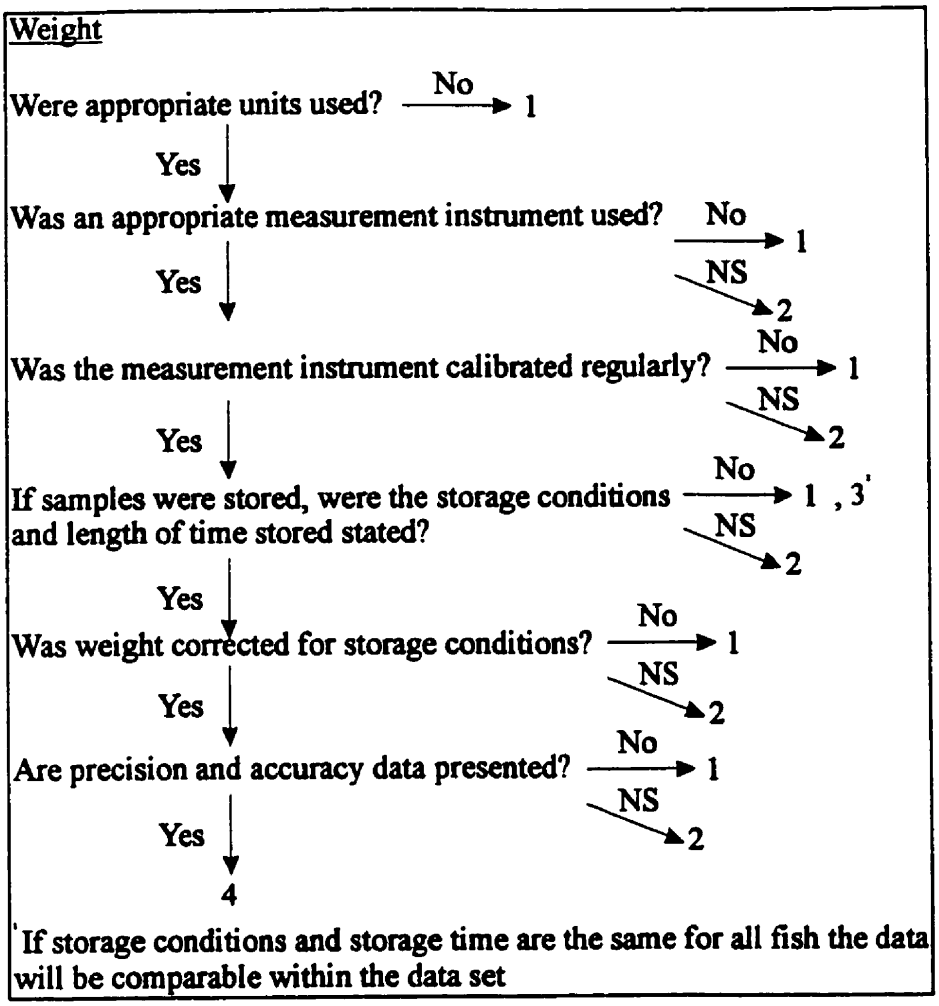
3.3 Decision Trees for Data Reliability Appraisal

3.3.1 Introduction

Data quality is a central issue for all users of data, irrespective of their individual goals. Data quality has a direct and controlling influence on the confidence level of conclusions drawn from the data and on the kinds of questions that can be appropriately posed. Data quality assessment is an area of continuing research. In environmental chemistry, data quality assessment is widely discussed in the context of quality assurance/quality control (QA/QC; see Mesley et al., 1991, for a review) and traditionally, the focus has been on QA/QC of analytical laboratory techniques. Some recent work has expanded this to include the entire process of sampling, storage, analysis, and data dissemination (Clark and Whitfield 1993). With respect to complete datasets, the traditional mode of quality assurance in the chemical field has been the use of peer review. Peer review is a system of decision-making by referees, editors, and research program directors in evaluating the quality of scientific research (Cicchetti, 1991). Hodgson (1997) notes that peer review is "so much part of the fabric of scholarly inquiry that it is often taken for granted". Moreover, peer review is a time consuming task. In the context of the Fraser Basin and the tremendous amount of data being collected for inclusion, the likelihood of identifying a sufficient number of reviewers was virtually nil. Even assuming peer review was an option, the appraisal of data quality infers personal judgment, presumably by an expert. This has the potential to be imprecise and subjective. What was required for this system was a method to encourage objectivity and ensure confidence in the assessments through a transparent, objective process. As Cornford et al. (1982) noted: the data quality appraisal process requires well-described protocols for the analysis of primary data.

3.3.2 ADCAP/WESCAP Methodology

In the ADCAP/WESCAP process, "data rating charts" were used to appraise datasets (Ratynski et al., 1988). These charts consisted of a series of questions that identified key characteristics of a dataset and resulted in the generation of a data rating (Ratynski et al., 1988). The charts, however, depended heavily on the expert knowledge of the appraiser (Ratynski et. al, 1988). The aspect of the data rating charts that was of particular interest to this research was its similarity to traditional decision trees. In a classic decision tree, a dataset is classified by sequentially subdividing it according to the decision framework defined by the tree. A class label (in this case a data rating) is assigned to each observation according to the leaf node into which the observation falls (Friedl and Brodley, 1997). A decision tree format has significant intuitive appeal because the classification structure is explicit and therefore easily interpretable (Freidl and Brodley, 1997). The ADCAP/WESCAP data rating charts met the requirements of a decision tree in that each classification procedure recursively partitioned a data set on the basis of a set of tests at each branch (or node) in the tree (Friedl and Brodley, 1997). A typical example of the ADCAP/WESCAP data rating charts is displayed in Excerpt 3.1 below. Note that the structure was based primarily on a root with three splits at each question: one that continued to the next split for yes, and two that gave ratings for no and not specified. While rating factors existed to aid in determining the appropriate answer for each question they consisted of recommendations for the appraiser and were not comprehensive, nor were they referenced. Excerpt 3.2 gives an example of rating factors for the weight chart presented in Excerpt 3.1. As is evident from the contents of Excerpt 3.2 the appraiser was expected to use professional judgement when doing appraisals.



Excerpt 3.1 Data Rating Chart for Marine Fish from Ratynski, March and Smiley (1988).

3.3.3 CODIS Appraisal System Design Principles

The appraisal system developed at UVic and used in CODIS was aimed to improve on the ADCAP/WESCAP methodology by deriving an objective data rating for a dataset in a manner that was reliable and repeatable.

Weight

TYPES OF UNITS USED: As with length, the size of the unit has to be appropriate for the weight of the fish.

CALIBRATION OF MEASURING INSTRUMENTS: Scales often go out of calibration and must be recalibrated at regular intervals. It should be stated that this procedure was carried out and at what intervals it was done.

SPECIFICATION OF STORAGE CONDITIONS: As with length, storage conditions of specimens prior to weighing may affect weight. If weights are taken after storage or preservation, the treatments should be described in detail. For best results, the effects of storage and preservation techniques should be determined and reported.

To receive a rating of 4 weight data must include type of scale used, estimates of the precision and accuracy of measurements, information about calibration of scales, information about storage of samples, and the effects of storage on weight.

Excerpt 3.2 Rating Factors for Fish Weight Measurement from Ratynski, March and Smiley (1988).

The design principles of the system were threefold:

1. Objectivity: the system had to both be, and appear to be, objective.
2. Simplicity: the system had to be sufficiently simple to allow appraisers to efficiently appraise numerous datasets in a limited period of time; and
3. Flexibility: the system had to be flexible enough to allow for advances in science and changes in procedures.

It was evident that the ADCAP/WESCAP data rating charts had encouraging aspects but needed redesigning in order to encourage objectivity and limit the possibility of appraiser bias.

The development of decision trees was carried out in parallel with the development of the CODIS structuring rules. It was the interchange of ideas and concepts from the structuring work that eventually resulted in the development of the CODIS decision trees. Much like in the structuring task, the ADCAP/WESCAP data rating charts served as an effective starting point. In CODIS, these charts were refined through adherence to the following principles:

1. The inclusion of a strictly dichotomous or binary structure,
2. The inclusion of guidelines which gave descriptive guidance to appraisers while providing for changes in methodologies with improvements in technologies or techniques, and
3. Peer-review and workshop testing of the decision trees to ensure a "consensus". This was expected to limit the subjectivity inherent in the ADCAP/WESCAP charts.

The approach used in CODIS to rate data used a series of structured questions designed to yield a unique reliability rating for each aspect of the measurement or experimental process. It included binary nodes with yes/no logic where a no resulted in the generation of a value. The dichotomous nature of the CODIS decision trees was designed to increase the simplicity of the system. As soon as a value was generated, the appraisal in that section was complete. The questions were taken in turn, with a "yes" response equivalent to "continue to next question". An example of a typical CODIS decision to appraise the sampling part of a measurement process in trace organic chemical analysis tree is presented in Excerpt 3.3.

Organic Contaminant Sampling Decision Tree	
1) Was collection documented? (see guideline 1)	no→2
2) Were collection apparatus and materials suitable? (see guideline 2)	no→0
3) Were all utensils and containers suitably cleaned? (see guideline 3)	no→0
4) Was cross-contamination avoided? (see guideline 4)	no→0
5) For benthic samples: Was suitable sampler used with disclosed mesh size?	no→1
6) For fish samples: Was trap method and fish type indicated?	no→1
7) For dioxins, furans and related analytes: were containers pre-washed with sample?	
8) For volatile analytes: was head space left in sample?	yes→1
9) For water column samples: was the functioning of the sampler established?	no→1
10) Exit with collection rating 4	

Excerpt 3.3 Decision Tree for Organic Contaminant Sampling (from Fyles et al., 1993)

The tree in the excerpt includes guidelines to guide the appraisal and produces an outcome consistent with the ADCAP/WESCAP rating scheme. As is evident in Excerpt 3.3, some questions did not apply to all samples and could be ignored when not applicable.

The critical advance on the rating charts was the inclusion of guidelines. They were designed to provide guidance for appraisals and not as an attempt to prescribe a correct method. The guidelines served as a source of information about the critical aspects of the measurement under evaluation. Guideline 3 from Excerpt 3.3 is shown below in Excerpt 3.4. The guidelines provided the appraiser with critical information in order to allow for an objective determination as to whether the dataset met the standard accepted requirements of the field or not. The guidelines, which were backed by an extensive bibliography, were derived from consensus protocols and appraised methods.

3.3.4 Using Decision Trees to Appraise Datasets

The decision trees provided a methodology to assess the reliability of each step in a measurement process. What was needed for CODIS was a process to appraise entire datasets. The system developed did this by developing protocols to derive the quality rating of the data in the dataset (Fyles et al., 1993). In the case of organic chemistry the evaluation of data quality was broken down into five independent categories: collection, storage, analysis, accuracy and precision. These categories were chosen to assess the confidence level in the full history of the sample from collection to final reporting. Each category was important, and relatively independent. Ambiguities in any of these categories would be sufficient to diminish the overall reliability of the data, despite excellent performance in the other categories. Each of these categories had its own set of decision trees which included an overall flow of questions to be answered, and a section of guidelines to assist in answering the questions. Each decision tree yielded a single reliability rating value, using the 0-4 scale discussed previously.

Guideline 3: Suitable cleaning procedure for collection materials:

Puget Sound Protocols for cleaning utensils

Utensils should be washed with detergent, rinsed with tap water then distilled water and dried at > 105 °C . Solvents can be used as well: these include acetone and high purity dichloromethane but utensils must be oven dried after the washing procedure. Sample bottles should be washed with detergent, rinsed with tap and distilled water then rinsed with acetone or high-purity dichloromethane and oven dried. Glass bottles should have their lids protected with PTFE to avoid contamination.

Environment Canada Method for cleaning of utensils, sample bottles and lid liners.
After washing with detergent and rinsing should be with tap water, distilled water or high purity deionized water. Then the following cleaning procedures should be followed:

Semi-volatile and Non-volatile Compounds:

Several rinses with acetone to remove any water followed by several rinses with (pesticide grade) hexane or methanol or petroleum ether to remove organics. Dichloromethane (pesticide grade) is used only for highly contaminated or dried on material. The utensils are then oven baked at 325°C for 12 hours or at 350°C for 6 hours or at 450°C for 4 hours. Solvent washing is recommended. All equipment can be solvent washed in the field to prevent cross-contamination of samples between different sites. Solvent washing also allows for consistency between field and lab cleaning.

Volatile Compounds

Utensils should be oven dried at > 105°C without solvent washing or rinsed several times with acetone or hexane and the solvent evaporated in an oven at 125°C.

Excerpt 3.4 Guidelines for Suitable cleaning procedure for collection materials (Fyles et al., 1993a)

The five ratings could then be compared, and the overall reliability of the measurement of the constituent/medium combination assigned. The overall rating was based on the "weakest-link" principle (Cornford et al., 1982). It stated that overall data can be no more reliable than the least reliable step in the entire measurement or experimental process (Fyles et al., 1993a). The rationale for each of the decisions made in the appraisal process

was documented to give a record of why a particular measurement gave the overall rating assigned (Fyles et al., 1996).

Figure 3.6 presents a flow chart of the process developed to appraise chemical measurements. The process commenced with the identification of a group of chemical measurements to be appraised. In the first phase, a list consisting of controlled terms for chemical constituents and media would be created as discussed previously. For each combination of constituent and medium, the details of the sampling, storage, precision, accuracy and analysis of the sample were appraised. Each of the five aspects of the appraisal involved subjecting the measurement process to an appropriate decision tree comparable to the one displayed in Excerpt 3.3.

Each decision tree would yield a single reliability rating value and associated documentation. The process would be repeated for each of the constituent/medium combinations in the dataset to create a set of overall reliability ratings. If all values in the set were the same, then the group would be considered to be fully appraised. If the overall ratings in the set differed, then the measurements assigned to the group would be redefined to achieve a uniform data rating. The overall reliability rating (the final result of the data appraisal) and the initial creation of a dataset were in a feedback relationship. If all the measurements in the initially assigned group could not be appraised on an equal basis then the group would need to be redefined, and the new group re-appraised (Fyles et al., 1996). As one example, consider a suite of sediment samples analyzed for volatile chlorinated compounds and PCBs. At the outset, this could be considered a single dataset. If the data appraisal process, however, determined that storage methodology was inadequate for the volatile chlorinated compounds, but adequate for the PCBs (storage period of 6 months before analysis), then the dataset would be split to reflect the differences between the two sets of measurements.

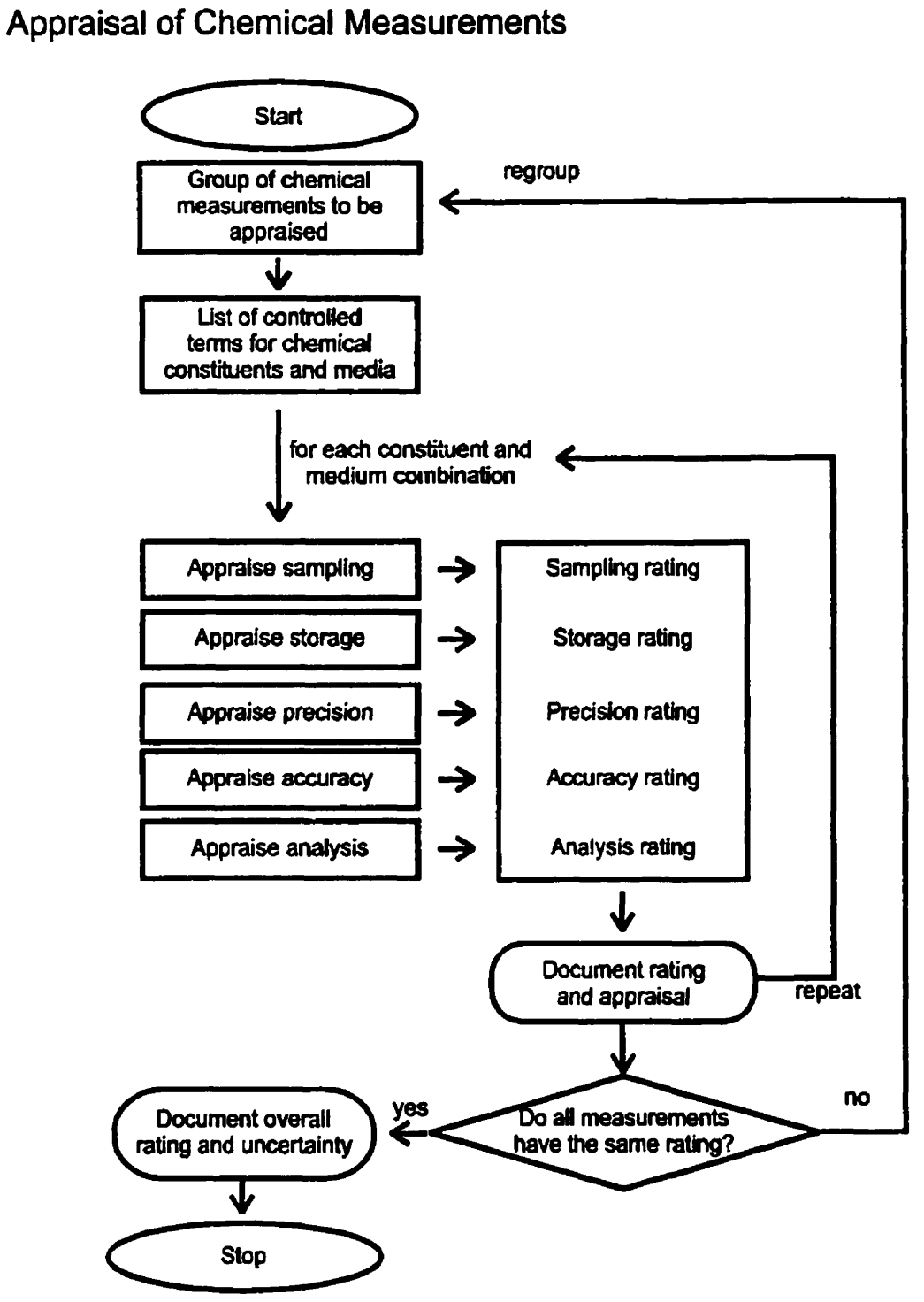


Figure 3.6 Appraisal Process for Chemistry from Fyles et al. (1996)

3.3.5 Decision Trees for Continental Benthos

The use of parallel data structures provided for the development discipline-specific appraisal protocols. In the case of the Fraser River Benthic catalogue, this meant a process designed specifically for the case of benthic invertebrates. It was determined that four decision trees could be used for the appraisal of benthic datasets: collection, storage, analysis and QA/QC. A sample decision tree for collection is shown in Excerpt 3.5 (Johansen and Reis, 1994). Note the similarity with the continental chemistry decision tree presented earlier. A dichotomous structure is presented, structured questions are posed and using the guidelines an outcome is derived.

COLLECTION:

1. Is there sufficient documentation of collection methods and substrate such that repeatability by other qualified investigators is possible? (See guideline 1.)

N⇨2

2. Is equipment suitable for substrate being sampled, the organism(s) of interest and the purpose of the study? (See guideline 2.)

N⇨0

3.Are the procedures suitable for the equipment, the conditions under which the equipment is being used and the purpose of the study? (See guideline 3.)

N⇨1

Exit with a rating of 4.

Excerpt 3.5 Decision Tree for Collection of Benthic Samples from Johansen and Reis (1994)

In designing the benthic appraisal process the authors used a slightly different approach as they noted in their final report:

> It is not the intent of the authors to produce a handbook of benthic study methods and as such the users of this document are directed to selected references on the subject. Those individuals assigned the task of rating data sets will be expected to

> be educated in this discipline and use their experience in conjunction with the protocols and appropriate references to assess the data. The authors have attempted to create a rating system that is of broad scope so it can be used for a wide range of study types. The questions in the decision trees have been developed to maintain objectivity when possible, given the many types of data sets expected (Johansen and Reis, 1994).

An examination of the guidelines developed for the benthic decision trees is presented in Excerpt 3.6. It presents an approach to guidelines that is less descriptive than those seen in Excerpt 3.4, relying instead on users referring back to the original references.

3.3.6 Critical Analysis of the Decision Tree Design and Functionality

The decision tree format and trees described in this work have been in use since 1993 and have undergone countless revisions, improvements and reviews. Extensive testing was carried out and the format used in the two new catalogues was incorporated into CODIS: Continental Chemistry and Continental Benthos. As part of the development of CODIS a detailed QA/QC evaluation of the decision tree methodology was carried out. The outcome of the evaluation follows.

The Continental Chemistry catalogue contained 578 unique datasets concerning 4663 chemical constituent/sample medium combinations for samples from at least 1394 sampling stations in the Fraser River Basin; the majority included multiple samples, over spans of time ranging from a few minutes to decades (Fyles et al., 1993a). Using the initial decision trees developed for CODIS 1.0 all these datasets were appraised. As a part of the QA/QC process approximately 10% of the rated datasets were re-analyzed and re-appraised. Of the 210 ratings checked, there was disagreement on only one, giving an estimated accuracy rating for the overall appraisal process of 99.5% (Fyles et al., 1993a).

Guideline 3

The various types of equipment available for benthic invertebrate sampling require specific techniques for their proper use, depending on the type and size of the equipment and the sampling conditions. Grabs of similar designs can be found in a variety of weights and sizes, and the procedures for their use will depend on certain parameters. Heavier and larger grabs may require winch and/or hoist equipment, whereas smaller, lighter grabs may be hand operated. Heavier grabs are more suitable for use in adverse weather conditions and for deep sampling than are lighter grabs. If inspection of grab samples upon retrieval shows evidence of washout or leakage, samples should be rejected and criteria used for their rejection should be documented.

Certain aspects of the use of a stream net sampler should be taken into consideration, including:

-avoiding disturbance of substrate upstream of the sampler
-avoiding clogs in nets which may cause backwash
-placing nets in areas of sufficient water velocity and appropriate depths

When using stream-net samplers, acceptability criteria should be related to the consistency of the sampling procedures, i.e. samples should be collected using identical methods.

When artificial substrates are used, recommended colonization time is six to eight weeks and exposure times should be consistent within a study. Extreme care should be taken during retrieval of artificial substrates to minimize loss of organisms. Ideally, a fine mesh net should enclose apparatus during retrieval.

Because of the small surface area sampled with cores, measures must be taken to compensate for the problems associated with patchy distributions of fauna. A sufficient number of replicates must be taken to ensure samples are representative of the study site.

Equipment of all types should be rinsed after each use to avoid contamination of future samples.

Selected References

Beak *et al.*, 1973; Eleftheriou and Holme, 1984; Gibbons *et al.*, 1993; Klemm *et al.*, 1990; Lewis and Stoner, 1981; Merritt *et al.*, 1984; Peckarsky, 1984; Plafkin *et al.*, 1989; Tetra Tech, Inc., 1989; Wainwright *et al.*, 1987.

Excerpt 3.6 Guidelines for Benthic Collection Decision Tree from Johansen and Reis (1994)

A similar process was carried out in the Fraser Benthics catalogue. The Fraser Benthic catalogue encompassed 168 datasets (Johansen and Reis, 1994). The average Continental Benthos dataset involved 3 different measurements, at 6 different locations, involving 15 different taxa (Fyles and King, 1994). To check for consistency between cataloguers using the trees, a random sample of datasets making up 10% of the total datasets were re-rated by different cataloguers. There was a 94.4% agreement between cataloguers using the decision trees (Johansen and Reis, 1994).

The outcomes of these evaluations indicated that the systems were indeed reliable. However, these evaluations did have a potential flaw. The same individuals who created the decision trees carried out the appraisals. While no individuals were given the opportunity to evaluate their own appraisals, the potential for bias existed. In addition, since their creators were the ones using the trees, no opportunity existed to determine whether the trees met another design criteria: ease of use. As a result, further testing of the trees was warranted.

In 1994 and 1996 workshops were held in which the decision tree methodology was evaluated. The 1994 workshop, held at the Institute of Ocean Sciences (IOS) on June 13th and 14th, and moderated by Dr. Dave Thomas of AXYS Environmental Consultants and Dr. Tom Fyles of the University of Victoria. The role of this workshop was to evaluate the decision trees developed to appraise trace organic and inorganic chemical analysis. This workshop occurred over two days. By the end of the second day the group agreed that by using the decision trees, consensus on appraisal outcomes could be obtained. The decision tree methodology was initially met with scepticism. This scepticism was dispelled by early into the first day, by which time the entire group had accepted both the concept and the overall process. The majority of the workshop time was spent discussing the contents of the guidelines, which were edited in detail. The only significant change involved reorganizing the order of the trees to facilitate the activity of appraising trace analysis methodology.

The 1996 workshop was held at the University of Victoria. It was attended by 15 senior scientists and data managers from DFO and the United States Environmental Protection Agency (EPA) and was intended to evaluate the protocols developed to appraise Contaminant Survey and Experimental events for NCIS. The attendees at the second workshop differed from those who attended the first with the exception of the University of Victoria researchers who ran both workshops and IOS researchers involved in the project. The aim of the second workshop was to evaluate the overall NCIS protocols (which will be discussed in detail in the NCIS case study); this included using decision trees to evaluate datasets. The workshop lasted two days with much of the time being spent introducing the delegates to appraisal protocols and decision trees. The attendees varied in expertise from senior administrators to laboratory technicians with the common thread being that all had worked on, or had interest in the development of the NCIS system. At the outset the delegates were provided with training materials in the form of tutorials which were intended to introduce them to the protocols and give them the opportunity to view the appraisal of a number of datasets. The first tutorial involved appraising chemical contaminant measurements and used modified CODIS chemistry decision trees.

The first session of the workshop consisted of a brief introduction followed by a run-through of the tutorial material. This was followed by a detailed evaluation of the first tutorial (using the chemistry decision trees) and an opportunity to carry out appraisals. In order to facilitate the learning outcomes the overall group was broken into four sub-groups each of which was given the same datasets to appraise. After a half-hour to carry out the appraisals, the four groups compared their outcomes. All four groups produced the same rating outcomes for the datasets although each group had its own unique notes associated with the rated datasets. The entire process from introducing the users to the decision trees, to the successful appraisal of chemical datasets was carried out in one morning.

At the outset of the workshop, the group expressed strong scepticism about the workability of using decision trees to appraise datasets. This scepticism was dispelled in one morning. This outcome indicated that the decision tree methodology had potential and in an environment of small controlled groups and consensus, could be used to provide reliable outcomes. Given the small sample number, this data is primarily of anecdotal interest. However, with regards to the decision trees, it is important to note that following the workshop a number of the attendees were then able to return to their own districts and teach other users how to appraise datasets using decision trees.

3.3.7 Evaluation of Continental Benthos Decision Trees

In order to carry out a more comprehensive examination of the decision tree methodology another set of tests was carried out in 1998. The first evaluation involved benthic datasets and consisted of two stages. The first stage was a retrospective analysis. This involved re-evaluating benthic datasets and comparing the results with those obtained by the original cataloguers in their 1993-94 appraisals. The second stage involved training an individual in the use of the benthic decision trees and then having that individual evaluate 10 datasets that had previously been evaluated in the first stage. Thus, the first appraiser, at least one historical cataloguer and the new appraiser would have appraised these 10 datasets.

In the first evaluation, 50 random datasets were chosen from the benthic catalogue and the processing numbers used by the original data cataloguers to organize their records. The first four datasets were set aside for use as a training set while 17 additional datasets (10% of the total) were used for appraisal. These datasets were then appraised using the published decision trees provided in Johansen and Reis (1994). In order to maintain consistency between the appraisals, the original the references cited in the Johansen and Reis (1994) protocols were obtained to supplement the guidelines. The 1993 appraisals were then obtained from the historical files in order to compare results. In 1993, two

appraisers carried out all the appraisals in the benthic catalogue. Upon inspection of the 17 datasets, it was determined that one of the two original appraisers was underrepresented in the test sample. Therefore, an additional four datasets were chosen using the initial random list but also searching by appraiser.

The appraisal process described in this work was designed to provide an effective methodology for researchers to appraise measurements. It was designed as a tool to aid experts not to test individuals. The methodology was designed specifically to avoid individual bias consequently, in the following discussion there will be no consistency of identification between appraisals. All individual appraisers and appraised references documented in this thesis will remain anonymous.

Once the first stage of the evaluation was complete, the second stage was begun. In that case, an individual (Appraiser A2) was chosen to do the appraisals. Appraiser A2 was an experienced user of decision trees who had carried out numerous appraisals of biological datasets. Appraiser A2 had never used the decision trees designed for benthic datasets. Appraiser A2 was given the test set and put to the task of training using the datasets. In order to avoid bias, an individual not involved in the actual testing supervised the training. Once the training set was complete, Appraiser A2 was given 10 datasets. These were chosen randomly from the original set of 21 datasets. A summary of the overall dataset ratings is presented in Table 3.7. The ratings are based on the 0-4 Scale used in CODIS with "n" being used to indicate that the individual appraiser did not appraise that dataset.

In order to provide additional information it is useful to examine how agreement was broken down by individual decision tree. This is displayed in Table 3.8. The values presented in the table involve a one-to-one comparison between each appraiser and all others who appraised that dataset. The Z values provided in Table 3.8 are the outcome of a sign test where agreement between appraisers was viewed as a positive and disagreements as a negative (Johnson and Bhattacharyya, 1992).

Table 3.7 QA/QC Results from Benthic Evaluation

Dataset	A 1	A 2	A 3	A 4
1	4	3	3	3
2	2	4	2	n
3	2	2	2	n
4	2	4	2	n
5	2	3	2	2
6	2	2	2	2
7	2	0	2	n
8	4	4	N	2
9	4	2	4	4
10	2	2	4	2
11	2	N	2	n
12	2	N	2	n
13	2	N	2	n
14	2	N	2	n
15	2	N	2	n
16	2	N	2	n
17	2	N	2	n
18	4	N	4	n
19	2	N	2	n
20	2	N	N	2
21	4	N	N	4

Table 3.8 Interrater Agreement by Decision Tree

Benthos	Agree	Test	% Agreement	Z Value	Sig 90%	sig 95%
Collection	41	56	73.2	3.45	Yes	No
Storage	50	55	90.9	6.07	Yes	No
Analysis	46	56	82.1	4.81	Yes	No
QA/QC	35	56	62.5	1.87	No	No
Overall	36	56	64.2	2.13	No	No

The results, while positive, only showed significance for three of the four trees and only at the 90% level of significance.

A more powerful statistical tool was available for the analysis of inter-rater reliability: the Cohen's Kappa (κ). It provides a coefficient of interjudge agreement for nominal scales (Cohen, 1960). The difficulty with κ is that it requires large sample sizes in order to provide reliable results (Cicchetti, 1976). In the evaluation of the benthic appraisals, the

only grouping that provided for a sufficiently large sample size for the κ to be reliable was a comparison between raters A1 and A3. A comparison of the two produced a 95% κ value of 0.654 ± .449 while the 99% κ was 0.654 ± .591. Cicchetti (1991) provides a scale to judge the agreement for two reviewers:

<0.4	– Poor
0.41-0.59	- Fair
0.60-0.74	- Good
0.75-1.00	- Excellent

Checchetti points out that low levels of κ can be produced not only by low levels of overall agreement, but also by large discrepancies in agreement on the various rating categories available to reviewers (Cicchetti, 1991).

In summary, good agreement was seen between appraisals A1 and A3 with significant agreement seen for all appraisals using the collection, storage and analysis decision trees. Seven of the eight occurrences of differing ratings involved a single appraisal disagreeing with the other appraisals on the suitability of documentation. Only one disagreement (dataset 1) involved a direct disagreement between appraisals as to the quality of a dataset. This disagreement involved the appropriateness of a standard used in the derivation of the dataset.

The lowest level of agreement was encountered in the QA/QC decision tree. An examination of the comments associated with the disagreements derived from this tree appears to indicate that this lower level of agreement reflects the continued debate in the benthic field as to what a reliable and reproducible measurement entails. This will be discussed in detail later in this section. The "Overall" agreement of 64% reflects the effect of the weakest link principle and the uncertainty in the QA/QC decision tree. This value is comparable to other methods of appraisal, like peer review, which will also be discussed in detail later in this chapter.

3.3.8 Evaluation of Continental Chemistry Decision Trees

Given the experience with the Continental Benthos decision trees it was determined that a similar procedure be carried out using the Continental Chemistry decision trees. The methodology used for the first evaluation was changed for this second test. In this test three appraisers were given the same 10 datasets chosen at random from the chemistry files in order to carry out an appraisal. The historical appraisals were also included. In the original appraisal process for Continental Chemistry carried out in 1992-1993, all the datasets were appraised by one appraiser using decision trees developed at that time (Appraiser T1). For the evaluation, refined decision trees developed for NCIS were used (these are discussed in Chapter 4).

The three individuals chosen included two experienced appraisers and an individual who had never carried out appraisals using decision trees. All three appraisers were initially supplied with a training set of five datasets with the inexperienced appraiser being trained by one of the experienced appraisers. At the end of the training task all three appraisers compared results in order to ensure that a common base of understanding. Following that meeting all three appraisers were given 10 randomly chosen datasets for the test. Communication between appraisers was limited in order to avoid bias. Following the test, the historical appraisals were retrieved and the four sets of appraisals were compared. The outcome of these appraisals is presented in Table 3.9. The historical appraisal is included to reflect the effect of the changing of decision tree structure on the actual outcome of the trees. Three tables follow. The first (Table 3.9) presents the appraisal outcomes, the second (Table 3.10) compares the three modern appraisals and the third (Table 3.11) compares the agreement between all four sets of appraisals.

Table 3.9 Results from Chemistry Evaluation

Dataset	Historical T1	Appraiser T2	Appraiser T3	Appraiser T4
1	3	0	1	2
2	2	2	2	2
3	3	1	2	4
4	3	1	1	4
5	3	2	2	2
6	3	2	2	2
7	3	2	2	2
8	2	2	2	2
9	2	2	2	2
10	2	2	2	2

Table 3.10 Comparison of Agreement of Modern Chemistry Appraisals

Modern CC	Agree	Test	% Agreement	Z Value	Sig 90%	Sig 95%
Collection	19	30	63.3	1.46	No	No
Storage	26	30	86.7	4.01	Yes	No
Precision	22	30	73.3	2.56	No	No
Accuracy	22	30	73.3	2.57	No	No
Analysis	28	30	93.3	4.75	Yes	No
Overall	22	30	73.3	2.56	No	No

Table 3.11 Comparison of Agreement of All Chemistry Appraisals

All CC	Agree	Test	% Agreement	Z Value	Sig 90%	Sig 95%
Collection	38	60	63.3	2.07	Yes	No
Storage	39	60	65	2.32	Yes	No
Precision	40	60	66.7	2.58	Yes	No
Accuracy	34	60	56.7	1.03	No	No
Analysis	54	60	90	6.20	No	No
Overall	34	60	56.7	1.03	No	No

The Z statistic is the same as that used in the benthos evaluation. Due to the limited sample number, a full κ could not be determined. In this case, however, it was possible to use a similar methodology developed by Williams (1976) to compare the joint agreement of several raters with another rater. This technique requires that all raters rate all datasets and was therefore not appropriate for use in the benthic evaluation. Williams' technique

produces a measure of nominal scale agreement (I_n). He described the interpretation of I_n as follows:

> Let a specimen be selected at random and rated by a reference laboratory which itself has been randomly selected from the n reference laboratories. If the specimen was also rated by the participant laboratory, this second rating would agree with the first at a rate I_n of the rate that would be obtained by a second randomly selected reference laboratory (Williams, 1976).

The values of I_n for the overall ratings were calculated for Table 3.9. For the three modern raters the vales were:

T2 vs. modern	1.07
T3 vs. modern	1.07
T4 vs. modern	0.88

When considering the entire record the values were:

T1 vs. all	0.55
T2 vs. all	1.27
T3 vs. all	1.27
T4 vs. all	1.13

In both cases, agreement between T2, T3 and T4 was high but not significant (a value of 0.70 is significant at the 75% level and 1.38 is significant at the 90% level) (Johnson and Bhattacharyya, 1992). It was notable that the three modern appraisals differed considerably from the historical appraisals based on the original decision trees.

The data from this analysis, while insufficient for statistical significance, appeared to indicate a number of trends that could be investigated in a future research activity. The outcomes from the early chemistry decision trees differed markedly from those of the newer chemistry decision trees. The benthos decision trees appeared to show an intermediate consistency while the modern chemistry trees showed a high level of agreement between appraisers. The author maintains that this hierarchy of reliability is a progression based on an understanding of process and a more intense peer review process. The final chemistry decision trees underwent detailed scrutiny through numerous tests and external evaluations. This improved scrutiny resulted in the production of improved trees that present consistent results. An additional, not unexpected, consideration emerged

from the data. It appeared that the more experienced appraisers showed the highest level of agreement. There was insufficient data to determine whether this high level of agreement was the result of increased familiarity with the trees or a more detailed knowledge of the field developed through repeated appraisals.

3.3.9 Peer Review

In order to calibrate the foregoing evaluation of the decision trees it is necessary to compare it to a readily recognized standard. Peer review provides a useful comparison although it has been criticized. Ciccetti (1991) discusses the problems of peer review in detail.

A number of papers have been produced which discuss the reliability of peer review. In their benchmark paper produced for *Science* Cole, Cole and Simon (1981) evaluated reviews of research proposals submitted to the National Science Foundation and National Academy of Sciences. They found that correlation between reviewer of proposals in chemical dynamics, economics and solid-state physics (n=50 for each) ranged from 0.60-0.66 with the highest correlation in economics and the lowest in chemical dynamics (Cole, Cole and Simon, 1981). Hodgson (1997) did a similar study, which involved determining the agreement and correlation between the peer review systems at the Heart and Stroke Foundation and the Medical Research Council of Canada. She examined grants that were simultaneously submitted in the same funding year to these two funding agencies and examined the results based on a six level ordinal scale. She found that raw agreement within a whole-digit range was 53% with a 95% κ of 0.198-0.382. Raw agreement between two funding agencies on the binary fundable/not fundable question was 73% with a 95% κ of 0.382-0.522 (Hodgson, 1997). In another major study, a review of articles submitted to the *Journal of the American Medical Association* showed that one fourth of the time re-reading of a journal article results in a reversal of the decision as to publish or not (Garfunkel et al., 1990).

The peer review study indicated that the outcomes derived by the decision trees in CODIS were entirely consistent and comparable in reliability and reproducibility with peer review. In some cases, it appeared that the decision trees produced more reliable results than peer review, although one must recognize that low levels of reliability in peer-review evaluations may reflect the low levels of consensus at the research frontier of scientific disciplines (Cole, 1991). In addition the increased efficiency of the CODIS decision trees may be affected by the training of appraisers. If peer reviewers underwent the same level of training, their repeatability might also improve.

3.4 Insights Derived from the Decision Tree and Appraisal Evaluations

The process of evaluating the decision trees highlighted a number of significant cases and issues that will be discussed in the following sections. In order to ensure anonymity in the following discussion there will be no consistency of identification between appraisals, and references to datasets will not be included.

3.4.1 Insufficient Process Knowledge and Differing Expert Opinion

Uncertainty is an issue to be addressed in all scientific endeavours and as such had to be addressed by the decision trees. As an example, in one typical case, two appraisals differed over a change in analytical technique. In the report in question, vandals disrupted a research project measuring benthic communities, this resulted in the loss of some samples. In order to preserve the experiment, the investigators chose to use a "Modified Surbur" method to collect samples. The investigators indicated that this modified collection methodology would provide useful data and avoid a complete loss of data for the survey.

Differences emerged in the appraisal of this report. Appraisal α1 noted that there were questions surrounding the use of a Surber sampler in a riverine environment and that additional detail was required in order to determine if the sampling was correctly carried out. In support of this decision, appraisal α1 cited two journal articles: Chutter (1972) and Kroger (1972). Chutter noted that the Surbur method could give highly variable results in determining population information depending on how the net was submerged. Kroger suggested that the Surbur method systematically underestimated standing crop because small invertebrates were able to crawl through the sampler's fine mesh while the backwash created by the tapered end of the net carried others out. Appraisal α2 cited a differing reference (Zelt and Clifford, 1972) which suggested that as long as pore size was reported, then the Surber was an effective tool for population studies. Since the report being appraised included a reference to the mesh size, appraisal α2 deemed that the reporting was acceptable.

This raised an issue common to every developing field: there exists the possibility that experts will disagree on how effectively different tools work. The decision trees were incapable of completely accommodating these discrepancies. Consequently, when the decision trees were initially being applied a standard approach (a convention) was created. It was decided amongst the CODIS cataloguers that the trees be interpreted in a conservative manner with the benefit of the doubt going to the researcher. If the appropriate academic community recognized a methodology, then it would be considered acceptable. This decision was conservative because of the way it affected the overall rating. When applying the weakest-link principle, a higher rating would not influence the overall rating in the same way as a lower rating. Where lower ratings existed, the lower rating would be the weakest link. A raised rating would not increase the rating of the dataset to a value higher than its weakest link, but a lower rating would reduce the overall rating for the entire dataset. Should subsequent research or debate achieve a scientific consensus then appraisals could be reconsidered. Consequently, in the case of disagreements between experts as to the validity of a methodology, the benefit of the doubt went to the researcher.

3.4.2 Uncertainty in Appraisals

The decision trees required appraisers to make decisions. Any time a decision is made, the potential exists for disagreements. In the analysis of the benthic datasets, a number of cases emerged. In one study, benthic invertebrates were being collected for life-history analysis. Samples were collected in a well-designed manner and then frozen on site before being brought to a lab for analysis. Appraisal β1 indicated that inadequate information was provided on storage while appraisals β2 and β3 did not. In this case, appraisal β3 included the following comment:

> Sample frozen in the field with dry ice. Note-assume that appropriate container type was used.

This assumption, which was noted in appraisal β2, was that in the case of simple storage for the purposes of transport, any storage container that was sealed would be appropriate because issues of contamination were not relevant.

In another dataset, the stomach contents of caddisfly were examined to identify diet. The prey items were identified on the basis of sclerotized/hard structures with reference to slide-mounted species of benthic taxa. Appraisal γ1 indicated that the lack of specific keys for the determination of stomach contents limited the usability of the data and gave the Analysis section a rating of 2. Appraisal γ2 noted that the mounted slides that had been previously identified by experts and were being used by qualified individuals would serve the purpose effectively, in lieu of specific keys, and so awarded a rating of 4 for the Analysis section.

A recurring disagreement arose between appraisals regarding the number of tows or samples taken. In the case to be considered benthic population studies were being carried out through the use of crab traps. The experiment called for a specific number of animals for the tests and trapping was carried out until the appropriate number of animals were caught. Appraisal φ1 indicated that the number of samples collected was ambiguous and

awarded a rating of 2 for collection. Appraisal φ2 suggested that the number of sampling episodes was not relevant as long as the number of animals was known.

The final case to be considered dealt with the identification of experts in order to confirm the accuracy of results. In this case, appraisal δ2 disagreed with the other appraisals regarding Question 4 of the QA/QC decision tree, which states:

> Where applicable, were taxonomic samples identified with a known key and/or by qualified individuals?

Appraisal δ3 presented the view that:

> Again, no mention is made of taxonomic keys/refs and/or comparison with archived reference collections although researchers are highly respected in their fields.

Appraisal δ2 noted that:

> Acknowledgments thanked a distinguished researcher for help in the identification of the organisms in the study. Cited individual is a qualified individual and therefore the use of the or in the question is applicable i.e. were taxonomic samples identified by qualified individuals? Yes [sic].

All these cases re-iterated the reality that experts often differ. Clearer guidelines could improve several of these cases, and better documentation by authors might address the rest, however, it would be unrealistic to assume that experts will not make unreliable decisions. This issue can only be addressed through the complete documentation of appraisals and an effective program that allows appraisers to get feed-back from authors or other experts. This issue was addressed in the NCIS system by including authors in the appraisal process.

3.4.3 Potential Sources of Additional Outside Information

An improperly worded question in a decision-tree or additional knowledge outside the assessment could result in decreased repeatability of dataset appraisals. One example was observed several times. In the Benthic QA/QC decision tree Question 3 stated: "Where

applicable, were portions of the samples re-examined to determine sorting efficiency?" The guideline that accompanied this question was equally vague. This resulted in inconsistencies both between appraisals and between datasets by the same expert appraiser. From the commentary included with the appraisals, it appeared that some appraisers used this question to lower the ratings of papers in which they had low overall confidence. In one case a self-published report done by a consulting company was awarded a rating of 2 due to a lack of documentation on sorting efficiency. A similar report, published by a highly respected researcher in a major journal, was awarded a rating of 4 with no data on sorting efficiencies.

Another example arose where additional information was available outside the assessment. In that case, the guidelines did not incorporate the fact that the full knowledge of time-of-year in a study of stomach contents would result in a qualitative difference in the results of a study. Most expert appraisers recognized this problem. One appraisal awarded a rating of 4 while the other appraisals all gave a rating of 2.

This discussion only reiterates that the decision trees were an attempt to create an effective expert system that would allow the appraisal of datasets. The appraisal process itself, however, could not rely on automatons. The appraisers, in many cases, were experts in their own right, and as such, would be expected to hold opinions. The aim of the guidelines was to direct those opinions and ensure sufficient documentation to provide for the needs of future users.

3.4.4 Currency and Adaptability of Guidelines

The aim of the guideline development process was to develop guidelines that reflected the best of current research and technique. These guidelines were then used to appraise work that had been done in the past when many of the techniques recommended were not available. Of particular concern would be the case where an outdated methodology was

demonstrated to produce faulty values. In that instance, an appropriate appraisal must indicate that fact. A subtler scenario existed where methodologies had changed and standards of the past were no longer standard, but neither were they generally considered wrong. Consider the use of an artificial substrate for benthic colonization that was used in the 1970's. At that time, it was considered standard to allow a month for colonization of these artificial substrates, which were then evaluated for community structure. In appraising two reports one appraisal included the following comments:

> "Only 4 weeks colonization period; 6-8 weeks is recommended, Johansen and Reis (1994). Note -If 4 weeks is acceptable, a collection rating of 4 is awarded"
>
> " Approximate (sic) 4 weeks colonization period with tray samplers vs. the recommended 6-8 weeks (Johansen and Reis, 1994). Note- if 4 weeks is sufficient then a collection rating of 4 is awarded.

Different published reports held different views on colonization period. The B.C. Ministry of Environment recommendation was 6-8 weeks for colonization (MELP, 1994). The APHA Standard methods merely stated that 6 weeks was recommended. Neither indicated that a month would produce incorrect results, only that better results would be seen by waiting 6-8 weeks. Recent work clarified the debate. Clements (1991) indicated that the length of time required to obtain equilibrium communities in trays varied among streams. Clements (1991) showed that community composition in a second order stream was highly variable among days but benthic communities collected from fourth and sixth order streams after 18-20 days colonization were similar to those collected on day 30. These results suggested that longer colonization periods may be necessary to characterize the benthic communities of small streams but that shorter times were entirely appropriate for larger streams. Given this added information, and knowledge of the stream size, it was decided to accept the original methodology as reliable and reproducible. In similar cases the lack of documentation on stream size resulted in the assigning of a rating of 2 except where additional or general knowledge was available regarding the stream size in which case that information could be considered.

3.4.5 Additional Issues in Assigning Ratings

The flow of questions in the decision trees resulted in the possibility that one dataset could be assigned differing ratings due to ambiguities in documentation. In particular, it was possible to grant the rating of 2 in lieu of a rating of 0 or 1. An example consider the previous case involving the Surber sampler where a long-term study was being carried out but a number of additional issues came into play. The vandalization of gear not only forced the use of questionable gear, but also almost certainly ensured that comparison of the values collected for that year with other years in the study would be inappropriate to due to differences in study design. In addition, this was a long-term study being carried out by a private firm. They underwent a major staff changeover during the course of the study and had to change taxonomists. Appraisals $\eta 1$ and $\eta 2$ used this detail to award a rating of 2 based on lack of information demonstrating internal consistency of the data. However, appraisal $\eta 3$ saw the work differently. As appraisal $\eta 3$ put it in awarding the dataset a rating of 1:

> From a comparison point of view the data is highly unreliable. First 3 years had six replicates next 3 years had 3 replicates. As well the taxonomists capabilities and/or abilities changed over the duration of the study.

It would seem unjust to assign a lower rating for difficulties not directly under the control of the researcher; this appeared to be what appraisals $\eta 1$, and $\eta 2$ represented. In this case, however, appraisal $\eta 3$ followed the logic of the appraisal completely by requiring documentation of internal consistency within the project. While it would be easier to have awarded a data rating of 2 the appraisers had to maintain their consistency throughout their appraisals. If this meant awarding a questionable dataset a rating of 0 or 1, then that was what should have been done.

In some cases, however, the decision trees only gave the option of a zero when a 1 was an appropriate rating (i.e. level of certainty in error). In the previously described discussion regarding length of colonization of an artificial substrate, an issue with the decision trees was identified. There were occasions where only a rating of zero was available but a

rating of 1 might be more appropriate. A rating zero is considered wrong, while a rating of 1 indicates documented uncertainty. If one were to assume that colonization period varied depending on stream size as suggested in Clements (1991), then a colonization period of 4 weeks would certainly not be wrong, but it might not be optimum as defined by APHA. A zero would, therefore, not be an appropriate appraisal rating and an appraisal might hesitate to give the zero. A rating of 1, however, would be appropriate and would be much more likely to be assigned.

A more obvious flaw in the decision tree structure involved the benthos storage decision tree. In that decision tree, there was no way to award a rating of zero. Thus, an obvious flaw could not be indicated by an appropriate rating. This flaw in the tree will be eliminated in the next round of peer review.

The second part of this discussion makes it appear that it would be appropriate to deal with these difficulties by providing an alternative access to either a 0 or a 1 rating in a single decision tree question. The first discussion, however, readily indicates that when given the choice appraisers tend to award the higher rating. This difficulty has yet to be effectively addressed in the benthic decision trees as they stand. Since the benthic decision trees are not currently being used to appraise datasets, this issue has not been addressed. If, however, a new cataloguing task arose that required the use of benthic decision trees then the only solution that would be consistent with the design criteria of the CODIS approach would be to scrap the original decision trees and redesign them to meet the requirements of the system.

3.4.6 Anonymity of Reviewers

Anonymity is a pillar of the peer review system, with reviewers identities being known only to editors and not to the author being reviewed. The preservation of anonymity has been argued to protect reviewers from retribution (Zentall, 1991) and to avoid the review

process from becoming too "personalized" (Greene, 1991). On the other side of the argument are those who suggest that anonymity in peer review protects careless and obstructive reviewers (Campanario, 1996). Many arguments against anonymity have been put forward including Adams (1991) who suggests that forcing reviewers to include their name with their evaluations would result in more constructive criticism (Adams, 1991). The issue may be moot, since as Rourke (1991) suggests, the advent of freedom of information movement means that journals will eventually be forced to adopt a policy of signed reviews.

The appraisal processes described in this case study shared many problems with peer review and the arguments continue regarding the anonymity of the appraisal process. In the case of retrospective analyses, where hundreds of datasets needed to be appraised, the decision tree methodology provided individual researchers with the tools to effectively appraise the work of numerous other researchers. This task could be tedious, but was critical to the effectiveness of the process. As Roos et al. (1989) put it: "assessing data quality is not glamorous, but it is highly desirable". The possibility existed in a very small minority of cases that appraisals risked offending individuals who had the ability to harm or interfere with their career. Anonymity provided a protection against this unusual occurrence.

Anonymity could often be critical to ensure appraiser cooperation for a more general case. The decision tree methodology developed for CODIS interjected error checking into what was fundamentally a trust relationship between peers. This error checking had the potential to affect continuing working relationships. Consequently, individuals appraising the work of their peers would often seek to remain anonymous in order to preserve working relationships.

The arguments against anonymity were equally persuasive. It was pointed out that public documentation can provide the best protection against reprisals. This protection would not available to anonymous reviewers. Another consideration when dealing with

anonymity addressed concerns presented in this case study. It was demonstrated that while the decision trees could be designed to reduce bias they could not be designed to be entirely objective since every appraiser brought his or her own body of knowledge with them into the appraisal. Numerous researchers address this issue in the literature. One school of thought professes that objectivity does not exist and that all research is value-laden (Moss, 1994). Staeheli and Lawson (1994) suggested that rather than objectivity, researchers should aim for critically examined partiality. While not doing so expressly, this is how the NCIS system handled the difficulty of anonymity of appraisals (see Chapter 4). In NCIS, scientists were expected to appraise their own data. Thus, there was a recognized bias that could be limited through an effective documentation of the appraisal and an effective QA/QC process to scrutinize the self-appraisals. This approach will be dealt with in more detail in chapter 4. In the CODIS case, anonymity of the appraisals was critical and the issues of appraiser bias were dealt with through documentation and QA/QC.

In the general case, the issue of anonymity of appraisals could be addressed in one of two ways. The first would be to follow the example used in research journals where reviewers remained anonymous to the data provider but known to the data manager. The alternative to anonymous appraisals involved incorporating the investigator into the appraisal of her/his work. As will be discussed later, in this research study the latter was found to be most effective. Incorporating researchers in the analysis of their work provided an educational element and resulted in improved documentation of subsequent research.

3.4.7 Training

Training is an issue in any area where an evaluation takes place. Crandall (1991) and Delcomyn (1991) both pointed out that the ability to write a useful review in the peer review process improves with experience. Adams (1991) pointed out, however, that the learning process for most reviewers was "haphazard" and "uncertain". While training in

the use of decision trees was not possible at the start of this research, more recent appraisers underwent a variety of training regimes. Over the course of this project, approximately 30 individuals were trained to carry out appraisals. While this number is insufficient to provide statistical verification, it did provide a great deal of anecdotal evidence on appropriate training methodology and on training speeds. With regards to application of the protocols, experts, as would be expected, had the least difficulty understanding the methodology but tended to be the most resistant to a strict adherence to rating guidelines. The general feeling they had towards the dataset providers was one of collegiality, where datasets that lacked documentation would not be downgraded since "everyone does that".

One interesting aspect of the training was that as individuals became more familiar with the decision trees they tended to rely less and less on the decision trees, trusting their own judgment instead. Less experienced appraisers were found to rely more heavily on the guidelines provided. The strongest demonstration of this came in the case of one of our most experienced appraisers. In a number of cases, this appraiser included specific comments in the QA/QC section of appraisal forms that were clearly derived from questions in the analysis tree. It was only seven months later when that appraiser was finalizing the forms for input that these errors were noted and corrected. Fortunately for this research, the appraiser documented the changes, which brought this to the attention of the group. Subsequently, two other appraisers pointed out that they too had often appraised datasets without reference to the decision trees. One candidly admitted that while he/she used the decision trees for her/his early appraisals after gaining experience he/she ceased using them for day-to-day appraisals. She/he only referred to the trees when specific concerns were raised, or in order to consult a specific guideline.

The lessening reliance by appraisers on their decision trees could lead to a serious issue, which could challenge QA/QC evaluators: that of interpretation drift. An evaluation of the appraisals done by the benthic cataloguers noted a significant change in notation associated with the Analysis and QA/QC decision trees. Specifically this dealt with how the appraisers handled the use of taxonomic keys for the identification of samples.

Taxonomic keys were specifically mentioned in question 4 of the QA/QC tree "Where, applicable, were taxonomic samples identified with a known key and/or by qualified individuals?" Numerous appraisals dated from the beginning of their appraisal process bore the following comment under QA/QC rating "No mention of keys used for I.D." In the development of the trees, however, notes from telephone conversations with UVic indicated that a decision was made to move the use of keys to the Analysis decision tree. It was decided that taxonomy was a measurement, and as such keys were a tool to carry out that measurement. While the decision was made and documented in telephone logs, the actual trees were never changed nor the guidelines updated.

It was clear that shortly after the decision was made the appraisers chose to interpret the analysis tree as if it included a requirement for taxonomic keys while still considering it in the QA/QC tree as well. This was supported by the appraisal documents that showed positive comments on taxonomic keys appearing in both the QA/QC and analysis notes, but that negative comments towards keys, and resulting lower ratings were only observed in the analysis trees.

"Interpretation creep", as exemplified in this case, is not always serious, since all the datasets were being treated in the same manner. Where interpretation creep becomes an issue is in a dispersed environment like DFO. If this type of creep were to occur at the regional level then the possibility would exist that similar datasets might receive very different treatment in different constituencies. This would defeat the aim of a common system and limit the effectiveness of any such system

In summary, a number of important issues have been raised and addressed. The rating outputs of the decision trees were of comparable reliability to peer review. The reliability of the decision trees improved in an environment of consensus, continued training and effective review, however, the decision trees, as written, were weighted towards disagreement and appraiser agreement had an upper limit (i.e. appraisals could only get so good due to inherent uncertainties).

3.5 Case Study Outcomes

The creation of databases presupposes the notion that data are valuable and therefore worthy of preservation. The value in data lies in the fact that once collected they can be used for more than one purpose. The CODIS case study took threads from various sources including reliability ratings, data rating charts, database structuring theories and search criteria and evolved each to become compatible with an overall system. This overall system involved the creation of a formalized process to take a collection of measurements and using structuring and appraisal tools to reproducibly create datasets with associated reliability indicators. These datasets and their associated ratings were subsequently input into an exact search system developed for this work which presented improved data structures. The simultaneous development of the CODIS program and the cataloguing task greatly clarified the critical necessity for rigorous protocols and definitions and indicated areas of consideration in continued research on the information requirements of environmental databases. The activity of inputting data into the CODIS prototype demonstrated that the approach used in CODIS to structuring data and evaluating data quality was both effective and comparable to similar systems in use at the time.

The ultimate utility of the software system called CODIS continues to be investigated. CODIS provides the most comprehensive assembly of datasets available for the West Coast of Canada and the Canadian Arctic. The comprehensive nature of the data stored in CODIS provides users with unparalleled detail of the historical record. The software shell and contents also have some applicability outside the boundaries of the research area comprised by their data. The CODIS source code and documentation has been made available on the internet. This should provide an opportunity to determine if the model will flourish or be ignored. The software shell and cataloguing tools are under consideration for use in new cataloguing efforts in the British Columbia interior and the Arctic (Smiley, B., pers. comm.) and the evaluation process is ongoing and will continue for the foreseeable future.

The development and application of the CODIS prototype also exposed a number of issues that drove the next stage of the research program. Consider a plausible scenario using the CODIS database. A researcher accesses CODIS to assemble data to test a hypothesis regarding the effects of chlorophenolic compounds on benthic communities. The researcher is trying to recover data involving chlorophenolics in benthic organisms as well as chlorophenolics in the sediments and benthic communities in those sediments. A search of CODIS uncovers two datasets involving the concentrations of chlorophenols in benthic organisms and a number of studies involving either chlorophenols in the sediment or benthic communities in sediments. CODIS has therefore successfully carried out its task. The researcher has a bibliographic list of datasets that have been appraised for reliability. The obvious next task for the researcher is to determine which of the datasets has appropriate data for use in the study. This involves obtaining the individual bibliographic references and carefully examining each to identify which of these datasets are useful for hypothesis testing. The CODIS appraisal process has aided the researcher in identifying reliable datasets but an additional activity remains. This additional task, the determination of the appropriate use of data stored in various media, is the basis of the next case study.

Chapter 4 NCIS Case Study

4.0 Introduction

The role of environmental databases is to assemble and store data. These data can consist of measurements of environmental variables collected at geographical locations at specific times, or controlled measurements collected as part of a research activity in a laboratory. What is often missing from these databases is the additional supporting information that clearly indicates what types of data are being stored. It is this annotative documentation which allows individuals who are unfamiliar with a collection of data to determine the data's applicability for a specific research objective (Stafford 1993). This annotative data supplies secondary users with the necessary scientific and experimental "context" to make use of the raw data. The contextual information provides an indication of the reliability and applicability of the data in databases for secondary users. Komarkova and Bell (1986) indicate that descriptive and annotative data are an essential part of a scientific database. A process that is able to reproducibly associate contextual information with the included sets of data increases the likelihood that the data in that database will be used appropriately.

Context is dependent on the information user. As an example, consider pulp-mill stack-gas emission data. It might be used by a regulator to confirm compliance to a regulatory requirement, but might also be used by a process engineer to indicate process characteristics of the stack. Emission data may be of no use to the process engineer if all the values were non-detects, but would be perfectly acceptable to the regulator who only wanted to ensure that the emissions were below regulatory requirements. The differences between these two uses involve a difference in context between a simple monitoring use and a more complex process-control use. The two types of tasks have very different requirements of the data. The CODIS data structuring and appraisal process focussed on the appraisal of individual datasets, without reference to their context. CODIS assumed that through documentation, multidisciplinary data could be made available to all users. This documentation procedure did not acknowledge that multi-disciplinary users might

lack the expertise to appraise the usefulness of data outside their area of expertise. Following the successful implementation of CODIS, this research project went on to address the issue of developing tools to associate contextual information with data for the Department of Fisheries and Oceans, Canada (DFO) National Contaminants Information System (NCIS).

The aim of this portion of the research project was to develop a methodology to formally associate contextual information with data and provide the tools to allow investigators to store contextual information with their archived data for NCIS. The methodology developed was governed by a structured set of appraisal "protocols". A protocol was defined as a set of procedures that created specific restrained outputs from diverse inputs (Fyles, 1996). The protocols developed for NCIS consisted of decision trees and associated guidelines and conventions. The protocols governed the activity of creating contextual information and associating it with datasets. This case study examined the development and implementation of the protocols developed to carry out this task in NCIS. This involved examining the process developed to create contextual information for NCIS and identifying key features. This case study included a discussion of the methodology developed to appraise the reliability of the contextual information and an attempt to understand its strengths and weaknesses. The CODIS case study included a detailed evaluation of the decision tree methodology. That level of detailed evaluation was not carried out in this case study because the methodology had not been fully implemented. This case study relied on an anecdotal evaluation of the protocols to create contextual information and a retrospective analysis to determine whether the context evaluation actually worked. The outcome of this case study was a practical understanding of how such protocols should be designed and implemented. This knowledge will be incorporated into the overall model presented in Chapter 6.

4.1 Introduction and Overview of NCIS

DFO, since its inception, has collected a vast amount of information regarding chemical contaminants and their various effects on environmental compartments. Much of this information still needs to be archived (Keeley, 1998). When the Green Plan Toxic Chemicals Program was instituted, it became apparent that an organized system to archive the resulting data was required. NCIS was designed as a structure to describe the data collections and hold the data from all these projects (Keeley, 1998). NCIS had three major components, which formed a logical structure for the data and information. The directory level held information about the projects under which data collections were made. The inventory held information about regional data holdings and was designed to allow users to identify data of interest. The archives were the repositories of the measurements and each region controlled its own archive's structure (Keely, 1998).

The inventory level of NCIS was of particular interest to this project. It was designed as a relational database built on a distributed architecture with the Marine Environmental Data Service (MEDS) in Ottawa and DFO regions taking part in the design, building and operation. Keeley (1998) stated that the inventory level was designed to hold metadata about contaminants involved in the archives; details of collection, storage and analysis procedures; where and when observations were made; and the organism studied.

Each participant region operated a server for the database. Queries using client software ran on one or more computers in each region and could be made across the DFO communications network to examine holdings in the local region or other regions. Upon identification of data of interest, a request could be made for the data. Depending on access privileges, data (and any additional information pertaining to the data) would be extracted from the appropriate archives. Provision of data was not a component internal to the NCIS, rather, data managers were required to make contact with the requester to transfer the data (Keeley, 1998).

The three NCIS levels provided for a system that both archived raw data and provided detailed inventory metadata to provide users with effective tools to search the database. As a data repository, NCIS could function without any indicators of the reliability or indications of the context of the information contained in it. However, that strategy would not take the full advantage of the quality assurance programs already in place in the DFO laboratories that generated results for NCIS. The research activity described in this work had two tasks: to develop a method to appraise the data being incorporated into the archive and to create and incorporate a new metadata element to describe the context of the data in the inventory level of NCIS.

4.1.1 Experimental and Survey Events

In order to insulate users from the large masses of contaminant measurements held in the archives, NCIS collected smaller assemblages of "like-data" about contaminants into "events". The definition of an event was related to the definition of a dataset:

> Each event is made up of individual data records that share a number of common characteristics. For example, all data in an event must measure the same contaminant, in the same organism or abiotic medium, that was collected in the same way, at about the same time, and at about the same place (Fyles et al., 1996).

Events in NCIS were split into two types: "survey events", and "experimental events". Strain and Sowden (1995) discussed the distinction between the two in detail, some basic concepts are repeated here:

> A survey study is a study into field conditions that have not been manipulated by the researcher.
>
> An experimental study is a study in which an organism or ecosystem has been manipulated or a biological response has been measured in a field study.
>
> A survey study must always have associated location information while location information may or may not be associated an experimental study.
>
> A contaminant survey event can arise only from contaminant survey studies and represents similar measurements associated with one or more locations. A contaminant survey event cannot record biological observations associated with

> the samples collected in survey studies. It can only record chemical contaminant data.
>
> An experimental event can arise either from experimental studies, or from survey studies involving biological responses. It represents ALL the chemical and biological responses that occur as a consequence of a given experimental manipulation. An experimental event generated from experimental work conducted in a natural environment can be associated with one or more locations (Strain and Sowden, 1995).

Contaminant survey events consisted only of appraised measurements. These measurements were appraised using CODIS-like decision trees and were not evaluated for context. For the purposes of the system, being a survey was their context. Experimental events included appraised measurements and additional metadata elements to describe context. Each individual measurement in an experiment event was called a "response" (Strain and Sowden, 1995). The use of this term in NCIS was intended to avoid confusion by allowing users to recognize whether a value or measurement was derived from a survey or an experimental event. This provided users with the ability to search NCIS for particular types of experimental events and to refine data-driven searches for applicability.

NCIS was designed to incorporate retrospective and continuing data. Consequently, data contributors were involved in the creation of the metadata, including the determination of events. So while event creation was controlled though a protocol, the data-contributor had the ability to influence the scope of individual events. The most obvious control involved specifying the spatial and temporal coverage. The data-contributor may have felt that it made sense for a clump of data to be presented to the data-user as a single package. For example, a researcher might consider it important to keep measurements carried out to evaluate the effects of new pollution control equipment together, insisting that measurements before and after implementation be considered as a group. The reasons for this could include a concern that a user might inadvertently miss contextually important data because of the restricted time period or spatial extent of the user's query.

4.2 Discipline-specific Appraisal Protocols

As will become clear later, the appraisal of an experimental event required the appraisal of the individual measurements that made up the event. This process was done using discipline-specific appraisal protocols and decision trees. Each protocol and decision tree was developed following the methodology presented in the CODIS case study. As in the CODIS case study, a common parallel structure united the discipline-specific appraisal protocols but each had its own peculiarities. The entire process for appraising measurements was documented in Fyles et al. (1996) which presented detailed protocols for the appraisal of chemical, biological and toxicological responses along with their associated decision trees. At over 200 pages, the Fyles et al. (1996) document was too comprehensive to be reproduced in detail here, although critical protocols and decision trees are presented. The overview for the appraisal of chemical responses was presented in the CODIS case study and will not be discussed here. A summary of the methodologies for biological responses and bioassay experiments follows. The process for appraising biological responses was created by Mark Pawluk and Dr. Fyles. It is being included here for completeness. The majority of the text for these methodologies was taken directly from Fyles et al. (1996).

4.2.1 Appraisal methodology for biological responses

A biological response was defined as a measurement designed to reveal the effect of contaminants on an organism. The organisms studied in an experimental event may have one or multiple responses to a contaminant. For example, in an experiment that involved exposing rainbow trout to 2-phenoxyethanol, one could measure blood gas and acid-base balance as well as corticosteroid/catecholamine levels in conjunction with a behaviour response. Multiple responses for the same exposure resulted in an inherent overlap in response classifications. For instance, in studies that involved independent responses such as behavioral and biochemical measurements, the behavioral component might require a relatively unstressed living organism, and therefore, must surely precede withdrawal of

terminal samples for blood chemistry measurements from any individual. On the other hand, tertiary measurements such as reproductive fitness were definitely dependent upon precursory biochemical responses (e.g. sex hormone levels). In this case, preliminary and ongoing endocrinological analyses would further support the observed behavioral response. The appraisal protocol followed this experimental logic - measurements requiring remote observations were usually considered before measurements involving greater degrees of intervention.

A flow chart for the appraisal of biological responses is given in Figure 4.1. The appraisal began with a list of controlled terms for the biological response measurements and organisms. The first step of the appraisal involved determining the measurements of interest (e.g. cortisol) and the associated response parameter (e.g. biochemical effects - endocrinological). The second step of the appraisal involved the identification of the organisms involved using the "Appraisal of Identification" decision tree (Fyles et al., 1996). For some measurements, the identity of the organism was not critical. In the case of cultured specimens, the identification process was usually straightforward. For field caught specimens, the identification process was more involved, and was appraised more closely. The outcome of this first step was an identification rating based on the five-tier scale presented earlier. In the third step, the appraisal was directed to one of six sections depending on the type of measurement considered. Remote observations were considered first, followed by measurements on whole/live organisms, measurements on sub-samples of organisms that were either alive or were alive immediately prior to sampling, and finally measurements which did not necessarily require live specimens. Each of the sections appraised relevant aspects of the measurement process: sampling/collection, storage, QA/QC, and the actual analysis.

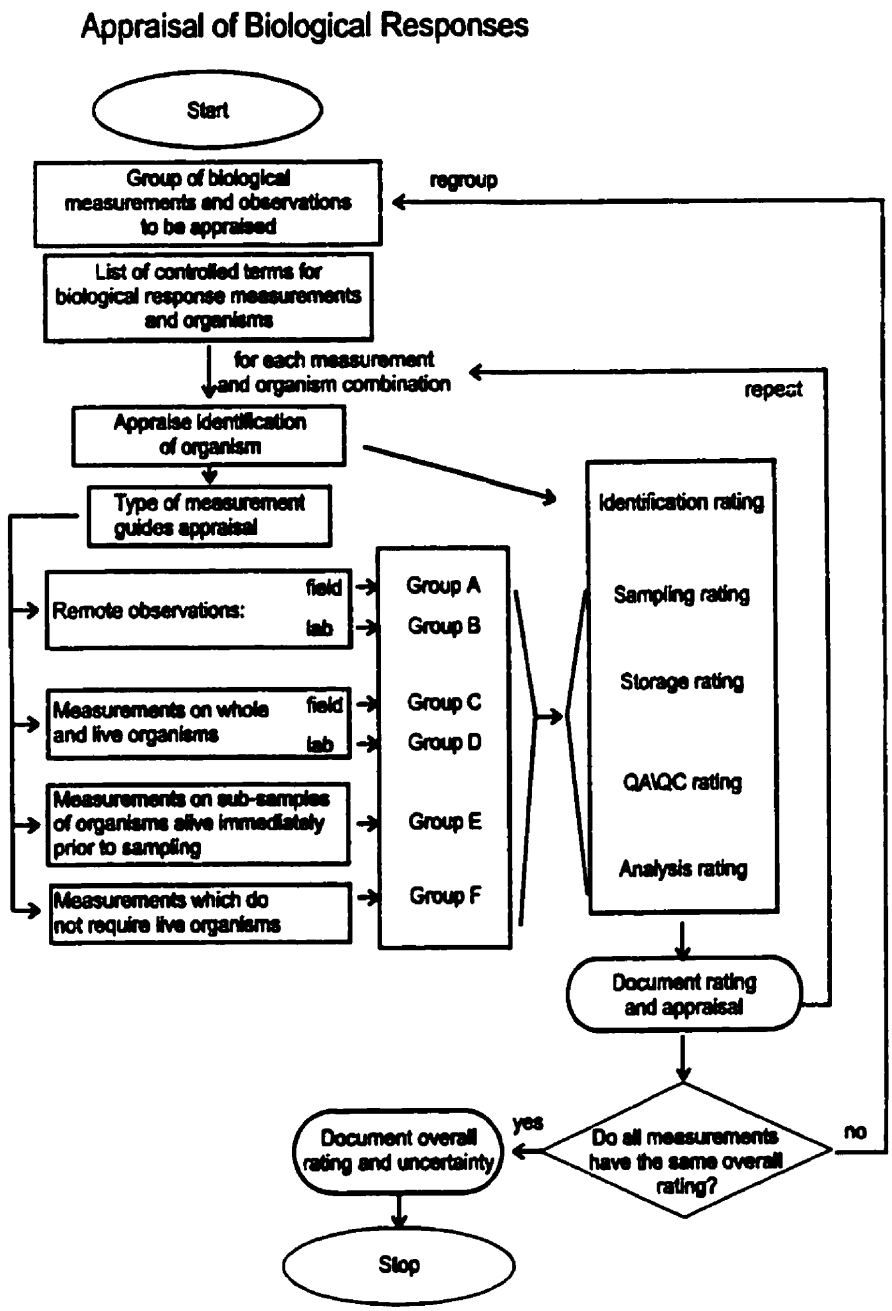


Figure 4.1 Appraisal Process for Biological Responses (Fyles et al., 1996)

Groups A and B (remote observations), and C and D (measurements on whole and live organisms) were divided by the designations "field" and "lab". This distinction was based on the extent of environmental manipulation around the organism during the measurement. Field meant a natural or unmanipulated environment. Conversely, lab meant a manipulated or controlled environment. The extent of the manipulation could vary from minor alteration of the flow within a stream or contaminant plume, to a completely isolated experimental apparatus. No matter what the physical location of the measurement, or the degree of the environmental manipulation, all such measurements were appraised as lab measurements.

The field/lab distinction extended to the separation of cultured stock versus captured wild stock. The appraisal of collection for field caught stock was more rigorous than for cultured stock, but the appraisal of storage for both types of organisms was similar. The critical issue was the extent to which experimental manipulation of the organism interfered with the response measurement.

The process was repeated for each measurement and organism until all measurements and organisms had been considered. Due to substantial duplication in the appraisal processes streamlining questions were added to groups C to F to simplify the amount of work. In essence, if a measurement with restricted handling and storage considerations had already been appraised, it was not necessary to redo the same type of appraisal for a measurement that had less stringent handling requirements. From the example above - if the appraisal of the behaviour measurements was satisfactory, then sample handling of the organism prior to removal of the blood sample had certainly been satisfactory.

Once all measurements and organisms had been considered, the final step was to establish that all measurements grouped for appraisal achieved the same overall rating. If they had, then the appraisal could proceed through the remaining steps of the experimental events appraisal. If they had not, then the group needed to be redefined to place measurements

of the same reliability in a single event and those of different reliability in additional event(s).

4.2.2 Appraisal Methodology for Bioassay Experiments

Bioassay experiments dealt with toxicological experiments involving organisms and populations in a laboratory setting. Toxicology is the study of the harmful action of chemical stressors on biota, including humans. It involves both an understanding of chemical distribution and effects as well as an understanding of biological mechanisms in order to identify hazards (Loomis, 1974). The measure of toxicology is toxicity, which is defined as the potential or capacity of a test material to cause adverse effects on living organisms (APHA et al., 1992). The first objective of toxicology is to determine the dose-response where dose is a measure of weight of test substance per unit weight of test animal (Environment Canada, 1993) and response is the measured biological effect of the variable tested (APHA et al., 1992). Dose-response can be modified by variables such as temperature, chemical form, availability and exposure as revealed by bioassays and measurements of ambient concentrations; this leads to an assessment of risk. In the bioassay appraisal process toxicology was defined to include both acute and chronic toxicity testing. Acute toxicity involved relatively short-term lethal or other effects, usually defined as occurring within four days for fish. Chronic toxicity involved a stimulus that lingered or continued for a relatively long period of time. Historically, a toxicity test was "chronic" if it exceeded a fixed number of days (*i.e.* 70). Now, chronic is defined with respect to total organism life span (usually as one-tenth of the life span). A chronic toxic effect could be measured in terms of reduced growth or reduced reproduction in addition to lethality (APHA et al., 1992). Some examples of measures of toxicology include LC_{50}, LD_{50}, EC_{50}, inhibiting concentration, subchronic NOEL and chronic NOEL.

The bioassay appraisal methodology was an amalgamation of the biological and chemical responses protocols and a flow chart for the full appraisal of both components is given in

Figure 4.2. The appraisal process began with a list of controlled terms for the bioassay experiment. The first step of this appraisal involved appraising the collection and storage of the chemical contaminant under consideration. These appraisals dealt with the collection of the chemical sample used in the experiment not with the biological indicator in the test. Appraising collection was imperative in the case of mixtures and effluents that had not been thoroughly characterized or where synergistic and antagonistic effects were undetermined. Appraising collection would be considered less relevant in the case of studies on known and pure chemicals (i.e., toxicity of laboratory grade solvents). The output from the contaminant collection and sampling decision trees were rating values and comments.

The next phase of the analysis involved all activities surrounding the biological indicators as well as the actual bioassay procedures and quality assurance/quality control (QA/QC). The first two decision trees in this section evaluated the sampling/collection/rearing of the biological indicator and any subsequent handling of that indicator. The analysis decision tree identified if appropriate procedures were used in the bioassay and whether the reliability of the measurement was established. Once the analysis had been appraised, it was necessary to examine the documentation of QA/QC to determine if it met standard requirements. The final step in the appraisal of bioassays involved full documentation of all tests carried out including the statistical procedures that were used. The statistical significance tree provided the means for a systematic evaluation of the statistical tests applied in the analysis, as well as documenting the tests that were carried out. The protocols and decision trees presented followed the guidelines set out by the Steering Committee on Identification of Toxic and Potentially Toxic Chemicals for Consideration by the National Toxicology Program, of the U.S. National Research Council (1984).

The discipline-specific appraisal protocols developed for NCIS were first published in 1996. After that they underwent review in the workshop described previously and were edited and revised. The protocols shared many of the strengths and weaknesses of those developed for CODIS and discussed in the CODIS case study.

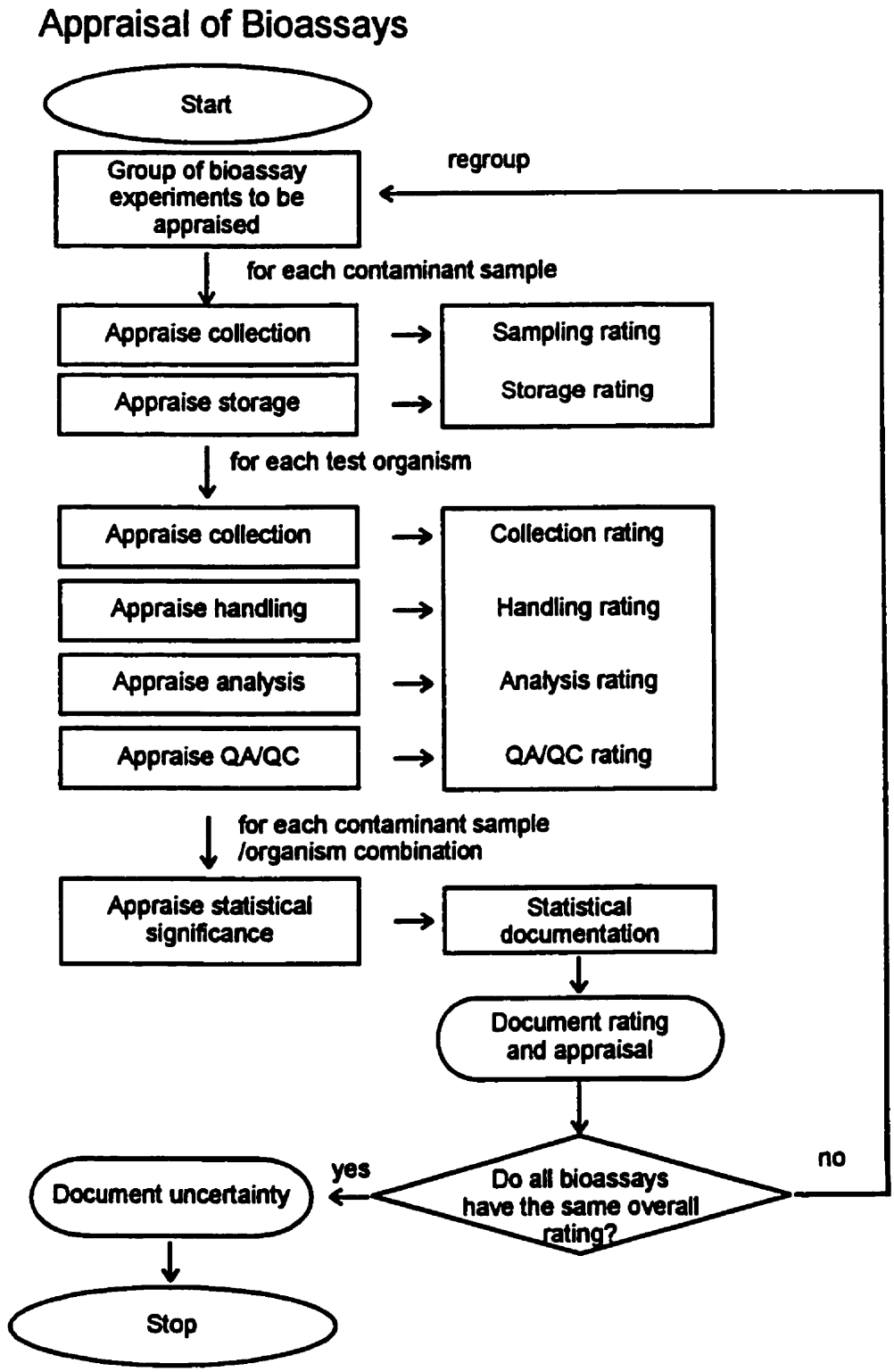


Figure 4.2 Appraisal Process for Bioassay Experiments

They continue to be used to appraise contaminant survey events being entered into NCIS. The number of users of the protocols remains relatively small, however, and insufficient data is available at this point to carry out any evaluation of the protocols' effectiveness.

4.3 Overview of Experimental Event Appraisal

4.3.1 Experimental Pedigree

The experimental event appraisal process was designed to examine and document the context of the experimental event. It considered the intention of the investigator, the design and execution of the experiment and summarized that information in a pedigree, which reported the context of the experiment to the user. A pedigree was defined as the most confident statement about an event that could be supported by the outcome of the study as it actually occurred. It was a statement that defined the scope and limitations of the study and reflected the uncertainty of the results. The concept of a pedigree was first suggested by Costanza et al. (1992) for use with the NUSAP system to communicate data quality in policy-relevant research. A broader, more inclusive application that included reliability, uncertainty and achievement of the goals of the research design was sought for the NCIS application. In this context, the pedigree provided an indication of the uncertainty of an experiment or study and the reliability of the results rather than merely the quality of the measurement process. This use of pedigree was an extension of the concept of "lineage" that was incorporated into the Federal Geographic Data Committee metadata standards (Federal Geographic Data Committee, 1994).

In NCIS, a pedigree described the reliability and applicability of the data in the event. It consisted of two types of information, 1) exact terms, which included the experiment type and formal statement and 2) descriptive fields which provided summary details of the experiment. The process of developing the pedigree is displayed in Figure 4.3 and will be addressed in more detail later in this work. The process involved four major stages based on decision trees.

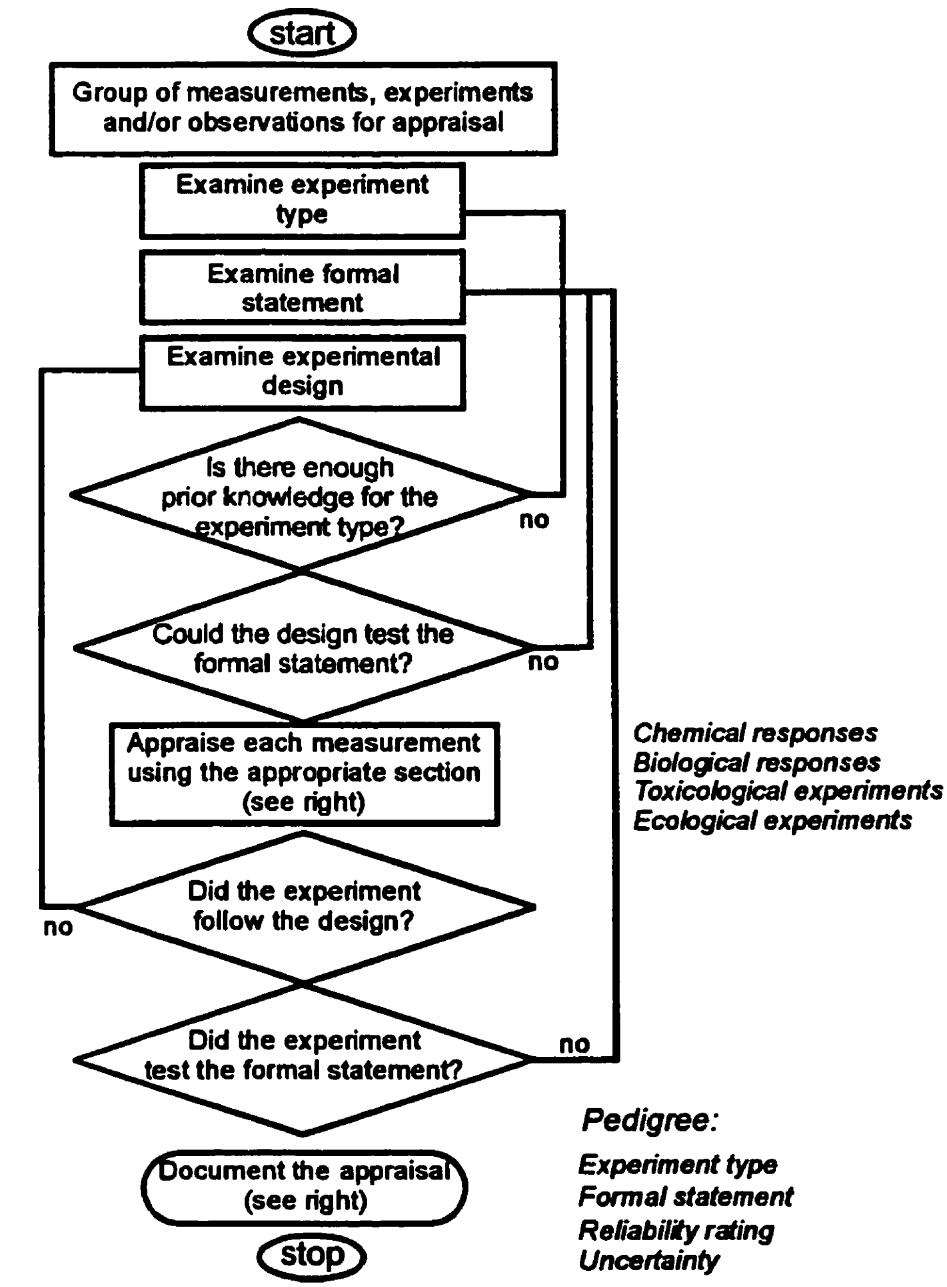


Figure 4.3 General Overview of Appraisal Process

In the first stage, a statement of the aims of the experimental event (the experiment type and formal statement) would be generated using a controlled list of possibilities. This generated the exact terms portion of the pedigree. This was followed by an analysis of the experimental design required by the formal statement chosen. If the design was

compatible with the formal statement, the measurement processes that generated the experimental data would be appraised using appropriate reliability appraisal protocols for the individual measurements involved. If the formal statement was not compatible, a new formal statement would be chosen, and the design re-appraised until compatibility was achieved. Once the measurements that made up the dataset were appraised, the experiment outcome would be examined to ensure that the design goals of the experiment were actually achieved. Finally, summary text would be created to describe the event.

4.3.2 Worked Example of the Overall Process

The following worked example of the process was presented for use with the protocols that were developed for the NCIS system (Fyles et al., 1996). Consider a report of the effects of chlorine on arctic char (*Salvelinus alpinus*) (Jones and Hara, 1988). This experiment involved the measurement of behavioural effects of variable levels of chlorine in water. The experiment was initially intended to test the effects of acidification on the behaviour of the sample fish, but the goals of the study were adjusted when the holding water supply became contaminated with a large dose of chlorine from a municipal water supply.

The first step in the appraisal process involved determining the measurements being considered. In this case the measurements were "behavioural", "locomotion" and "orientation" all of which were controlled terms in the NCIS system. The next step defined an experiment type and formal statement based on the intent of the experiment. In this case, the experiment type would be considered "research" as it was used to develop process knowledge and its formal statement would be of the form "to develop process knowledge". The testable null hypothesis for the experiment was "addition of chlorine will not affect the behaviour of char". The next portion of the analysis involved comparing the experimental design with the formal statement. In this case, the design was based on an

earlier design for the effects of acidification on char but was adapted to fit the circumstances of the experimental situation.

The behavioural nature of the observations and the inability of the experimentor to replicate all the conditions of the test resulted in the design relying on internal replication rather than external standards. This was entirely consistent with other work in this field. Following the experimental design appraisal, the execution of the experiment was evaluated. Given that this was a biological experiment that occurred in a lab environment with live fish; the protocols for "Biological Response Appraisal" (Group D: Measurements on whole and live organisms in the lab), were used (Fyles et al., 1996). This evaluation was carried out in its entirety and the result, a "measurement rating" was produced. In this case, the outcome was a reliability rating of "4" for sampling, storage and analysis and a reliability rating of "3" for QA/QC. The "3" was based on the lack of external standard. The overall rating, based on the weakest-link principle, was "3".

Following the experimental execution step, the experiment outcome was appraised. This involved evaluating the experiment to determine if sufficient QA/QC had been carried out throughout the experiment to ensure that the results were statistically significant. This appraisal produced a number of statistical results (power, accuracy, precision, spread etc.) which were incorporated in the summary pedigree. This analysis also asked the questions "Did the outcome actually test the formal statement?" "Was the uncertainty suitably addressed?" and most importantly, "Do all experiments in the list have the same pedigree?" The summary pedigree read:

Hypothesis Tested:	Addition of chlorine will not affect behaviour of char
Measurement Rating:	Biological response rating "3", results precise but accuracy impossible to determine
Measurement Uncertainty:	Internally consistent experiment, α known, no ß given.

The summary pedigree provided the text fields in the pedigree. The experiment type and formal statement were not part of these text fields but were the exact terms of this experiment. As a result, a user searching using the exact terms "research", "behaviour"

and "chlorine" would find this experimental event and the summary pedigree would supply the added contextual information to determine whether the experiment would be appropriate for secondary use.

4.4 Experiment Type and Formal Statements

In the research community, there are three kinds of scientific activity: monitoring, research and surveys. In NCIS, the term "survey" had a specified definition. In order to carry out a general discussion, it was necessary to use an alternative term. Consequently, the term "observation" was used instead of "survey".

There are no agreed upon, natural definitions for observation, monitoring and research, but as early as 1975, MacKay and MacDonald identified the key features that distinguish the three terms. They pointed out that monitoring was an activity carried out on well-understood systems (MacKay and Macdonald, 1975). Research was identified as an activity that preceded monitoring and was designed to develop process knowledge; as they put it:

> in the large, interdependent systems that we deal with in the environment, we often have a poorly developed understanding. We are not sure of the key parameters that control the system. Hence, it is logical that research be carried on so that monitoring programs can be properly designed and adequately implemented.... often we can discern a progression from a research activity which studies a particular area at one time, or perfects a new technique or checks for the presence of some substance, to a monitoring program in which data is routinely collected to be available for management decisions....as a general rule, baseline studies, such as description of the present state of an area where planned industrial activity may have an environmental impact, have been considered as research projects. Those cases where there is a high probability of continuing measurements of the type made in baseline studies represent examples of research projects which could become monitoring projects (MacKay and MacDonell, 1975).

4.4.1 Induction, Deduction and Theoretical Knowledge

The distinctions between monitoring, research and observation lead directly to a fundamental question in positivist science, that is how does one carry out any form of

generalization? One of the major distinctions between monitoring, research and observation can be linked directly to the two primary methods of deriving generalizations and laws: induction and deduction. In very simple terms induction can be described as arguing from the particular to the general, while deduction argues from the general to the particular (Medawar, 1969). Medawar defined induction as:

> A scheme or formulary of reasoning which somehow empowers us to pass from statements expressing particular "facts" to general statements which comprehend them. These general statements (or laws or principles) must do more than merely summarize the information contained in the simple and particular statements out of which they were compounded: they must add something, say more than that which has been said already...Inductive reasoning is *amplitative* in nature. It expands our knowledge, or at all events our pretensions to knowledge [quotations and italics are his] (Medawar, 1969).

The two primary assumptions of simple induction are that science starts with observation and that observation yields a secure basis from which knowledge can be derived (Chalmers, 1982).

Deduction, on the other hand, presupposes the theory-dependence of observation. What an observer sees depends in part on his/her past experience, his/her knowledge and his/her expectations (Chalmers, 1982). This past experience brings in the concept of theoretical knowledge, or process knowledge. Process knowledge describes the dynamics and interrelationships within natural, biophysical, and social systems (Cornford and Blanton, 1993).

Induction and deduction are the two principles that underpin "monitoring", "research" and "observation". The debate continues, however, as to the values and drawbacks of induction and deduction as scientific methods (Chalmers, 1982). This thesis accepts that both inductive data collection (observation and research) and deductive data collection (monitoring and research) have their place in environmental science. The important factor to remember is that each has its limitations and drawbacks. Theory-driven observation (deduction) is often only as good as the theory driving the observation; while theory-free

observation (induction) can often result in incomplete datasets unable to answer fundamental questions.

From a working scientist's perspective, experiments are carried out in order to get results that can then be used to answer questions. Little care is paid as to whether this is done through inductive or deductive means. The difficulty in defining terms for use in NCIS arose from the fact that while many research projects begin with inductively derived starting points; they often proceed to use traditional deductive techniques in analyzing the observations that have been made. In terms of the three activities of the previous section (monitoring, research and observation), NCIS-defined experiment types viewed monitoring as predominantly a deductive activity and observation as predominantly inductive. The experiment type research encompassed both inductive and deductive methods. As developed below, formal statements served to clarify the distinctions.

4.4.2 Monitoring

Monitoring continues to be one of the primary tasks of DFO and takes on many formats. Cornford and Blanton (1993) stressed that monitoring, like prediction, required an understanding of parameters and processes, as well as their detailed relationships in time and space. They distinguished three types of monitoring:

> **Research (Trend) Monitoring** is the continuous measurement, in either time or both space and time (i.e., time-series), of parameters that may have only partially known relationships (or several plausible hypotheses), for the purpose of detecting trends in the periodicity or magnitude of possible cycles or repetitive events.
>
> **Scientific Monitoring** involves repetitive time-series measurements (over a particular area) and is only possible when processes and their interrelationships are known, and hence there is sufficient knowledge to assess the effects of some external influence on known trends and/or cycles. Initiation of this type of monitoring also requires an adequate understanding of both process and data interrelationships to establish a proper temporal and spatial sampling strategy.
>
> **Surveillance (Compliance or Enforcement) Monitoring** is normally a government regulatory requirement and is designed primarily to determine conformity with some

> predetermined standards. This term could be termed **Comparative Monitoring** and may be used as part of a follow-up strategy for a development to determine the extent to which project implementation conforms to original predictions or conditions (Cornford and Blanton, 1993).

For the purpose of NCIS, monitoring included all null-hypothesis-bounded experimental or survey events. The null hypothesis could be explicit or implicit and must be based on pre-existing knowledge of process. The two requirements for monitoring were that some preexisting process knowledge was incorporated into the experimental event design, and that some conditional null hypothesis, whether explicitly described or implicitly assumed was being tested.

One additional type of monitoring considered was modelling. In order to create and test models it is necessary to have some process knowledge, and the resulting modelling serves to test a null hypothesis. Therefore, evaluating models or iteratively refined correlations were considered monitoring. Any collection of data that did not meet these two criteria and was not a modelling study would, therefore, either fall into the research or observation categories.

4.4.3 Research

Under the system described by MacKay and Macdonell (1975), research preceeds monitoring. Research is used to develop an understanding of natural processes so that monitoring programs can be developed. These monitoring projects are then used to assure scientists that the results obtained from their research and the resulting theories are correct. "Research" developes the process knowledge required to design effective monitoring programs. Cornford and Blanton (1993) defined a hierarchy of research based on process knowledge:

> **Descriptive (Baseline) Research** generally refers to the acquisition of initial data to assess time and/or space patterns, and permit a more comprehensive data collection

> strategy to be designed for the purpose of delineating trends and the operating processes within the system under investigation.
>
> **Process Research** concentrates on determination of the relationships between and among parameters to provide an understanding of system dynamics.
>
> **Time Series Research** may be broadly defined as an extension of trend monitoring for events that have extremely long cycles or are of sufficiently large magnitude that traditional research strategies must be altered to examine system dynamics (Cornford and Blanton, 1993).

Dane (1990) pointed out that the relationship between research and theory is an extremely strong one; research results are always placed in the context of existing theory, and existing theory provides a framework for new ideas about what to research. Stonehouse and Mumford (1994) suggested that when causal mechanisms are understood with some confidence, scientific information could be used with greater precision and flexibility. Thomas (1992) effectively summed up the entire relationship between research and monitoring:

> Two essential types of information are required for the satisfactory resolution of cause-effect relationships. First, the cause-effect relationship being monitored must be understood to the point that appropriate parameters can be chosen to produce meaningful test criteria. Second, each factor and its relative contribution to the natural variability of the index parameters must be known to the degree required by the pre-set level of confidence expected in the monitoring result. When the above information is unavailable, monitoring cannot proceed. In its place, basic research programs must be developed to provide the essential information (Thomas, 1992).

Research can also be carried out on well-understood systems and processes, but in all cases, the intent of research is to improve the level of process knowledge available. In research, process knowledge is generally incomplete or incompletely understood; consequently, other factors are needed so that the process is formalised and scientific. In normal science, this is carried out through careful control over external conditions. The data produced through research is carefully collected and any numerical data resulting from the process is manipulated by identified statistical methods. Research studies fundamentally require that some form of null hypothesis be proposed and that testing the hypothesis serve as the basis of the activity.

The critical consideration of the NCIS experiment type research was that the study tested a null hypothesis. In addition, it was necessary to control or limit external and internal factors and state experiment and numerical acceptance values. This effort to control or shape conditions would usually be part of the documentation of the event. The specific requirements for the experiment type research were: at least one, explicitly described, conditional null hypothesis be tested; a formalized method to interpret the data produced in the experiment be stated; and a demonstrated effort be made to control external effectors in the experiment. Any collection of data that did not meet these three criteria would either fall into the monitoring or observation categories.

4.4.4 Observation

The experiment type observation included predominantly inductive studies. For NCIS, observations were any experimental event in which one of the following was true:

a) there was no testing of a null hypothesis (either stated or understood);
b) process knowledge was not included in the experimental design;
c) external conditions were either not controlled or not incorporated into the design so as to limit results to the causative agent;
d) numerical or statistical goals were not determined before the experiment; or
e) there was no initial intent to collect the data.

Since observations in NCIS were associated with experimental events and the majority of events include a null hypothesis relatively few events were expected to be assigned this experiment type.

Excerpt 4.1 presents the decision tree created to provide a simple method of identifying experiment type. The initial four questions were designed to tease out the difference between "observation" and "monitoring". The subsequent questions distinguished between "monitoring" and "research". The concluding question clarified the research/observation distinction.

Determination of Type of Experimental Event

1) Was there a stated, specific purpose for the data collection?
- No → Observation
- Yes → Continue

2) Was pre-existing theoretical knowledge incorporated into the data collection design?
- No → Continue
- Yes → Go to Question 4

3) Was the stated purpose of the data collection to develop process or theoretical knowledge?
- No → Observation
- Yes → Continue

4) Was the data collected for the purpose of testing a stated or understood null hypothesis?
- No → Observation
- Yes → Continue

5) Was the data collected to model theoretical relationships or to monitor changes in stated theoretical relationships?
- No → Continue
- Yes → Monitoring

6) Were controls in place to limit the possible outcomes of the experiment?
- No → Monitoring
- Yes → Continue

7) Were statistical requirements and acceptance criteria for the results clearly stated?
- No → Observation
- Yes → Research

Excerpt 4.1 Experiment Type Decision Tree

4.4.5 Formal Statements

Given the large number of experimental events anticipated to be stored in NCIS, three experiment types were found to provide too coarse a screen for the development of appraisal protocols or for searching purposes. In order to create a finer screen it was essential to create a secondary tier of classification that would separate similar experimental reports into well-defined categories that shared a sufficient number of characteristics. Experimental appraisal processes could then be specialized based on the experimental requirements of

each characteristic group. These additional categories were referred to as "formal statements".

Formal statements described the intellectual capital invested in the experimental event and the original intentions of the experimental process. Formal statements in NCIS were derived through an iterative process designed to identify the most appropriate formal statement supported by the data. Formal statements were initially derived from the expectations of the scientists carrying out the data collection as indicated by their stated intentions. The process of determining a formal statement began by identifying what the author intended to study and assigning a temporary formal statement that reflected that intention. The temporary formal statement was finalized at the end of the process, once it had been tested to see if all measurements in the event supported it. If an experimental event was found to have more than one different pedigree, then the entire event would have to be regrouped and the process, including the determination of formal statements would have to be reconsidered.

Following the Cornford and Blanton (1993) model, one of the crucial elements for these formal statements was the existence of a null hypothesis. The weakness of the Cornford and Blanton (1993) model was that it assumed that hypotheses could be proven. This research took the more conservative approach described in Stonehouse and Mumford (1994), namely that a hypothesis could only be disproven. In essence, this work combined the Cornford and Blanton (1993); Thomas (1992); MacKay and MacDonell (1975); and Stonehouse and Mumford (1994) models to generate an initial working set of levels of formal statements ranging from studies conducted without hypotheses, through to studies working from a strong base of prior theory and hypothesis testing. This initial set of levels could then be divided into a scale of working categories. This list of working categories was not hierarchical and the this list was expected to form a continuum. Formal statements were derived for each experiment type and are displayed in Table 4.1. The internal categories for formal statements within an experimental type are listed in no particular order, and no hierarchy is implied.

Table 4.1 Formal Statements for Experimental Events

Experiment Type	Formal Statements	Goal of Experiment	Nature of the Experiment
Monitoring	Scientific Monitoring	To improve understanding of theory, based on well established process knowledge	Repetitive time-series measurements (over a particular area) on well understood processes. This includes modelling studies
Monitoring	Compliance Monitoring	To identify variance from accepted norms, null hypothesis not stated but understood	Similar to scientific monitoring but monitoring limited in time or space. Monitoring is mandated and only methodology is under the scientist's control.
Monitoring	Research Monitoring	To test a null hypothesis based on demonstrated processes	Working theory derived from several iterations of proposing and testing hypotheses with experimental work (uses field tested process knowledge).
Monitoring	Trend Monitoring	To test a null hypothesis based on inferred processes	Testing working theories based on several iterations of proposing and testing hypotheses regarding process knowledge (generally in the lab) but little real world data (Field testing process knowledge).
Research	Baseline Research	Involving testing to develop process knowledge	Testing hypotheses regarding effects of contaminants or conditions and includes toxicity testing
Research	Process Research	Based on untested working hypotheses used to derive relationships between and among parameters to improve system understanding	Process knowledge research, testing process hypotheses.
Research	Foundational Research	Based on untested working hypotheses derived from general observations not process knowledge	Testing of the hypothesis
Research	Deductive Research	Involving one or more untested working hypotheses using well established techniques	Research using "accepted" theory or technique to derive information or data.
Observation	General Observations	Made as part of a survey	Data collection for data collection sake as a part of a general data survey
Observation	Fortuitous Observations	Made as part of another research study	Data collected for another purpose or as a part of another research study
Observation	Incidental Observations	Incidental data collected without specific intent	Data obtained through haphazard collection.

4.5 Experimental Design Appraisal

The goal of experimental design appraisal process was to determine the extent to which an experimental design reflected the formal statement assigned to the experiment. It asked the question: Is the experiment able to meet the objectives required by the type of formal statement? It also acknowledged that non-scientific considerations could impact on experimental design. Thus, the goal of the appraisal was to document what the design could achieve if executed as planned. The analysis of experimental design has been extensively studied (Green, 1979; Scheiner and Gurevitch, 1993 and Montgomery, 1984) and there is a rich literature on this subject. The protocols developed were intended to reflect that literature. The protocols guided the user through established frameworks in experimental design and emphasized practical methods to appraise whether the experiment being considered had the ability to meet the goals of the formal statement. This appraisal process suggested areas for consideration in experimental design as well as presenting suggested methods to achieve various ends.

There were a number of fundamental difficulties that had to be addressed in the experimental design appraisal, as Eberhardt and Thomas (1991) suggested:

> Field experiments in ecological and environmental research usually do not meet the criteria for modern experimental design. Subsampling is often mistakenly substituted for true replication, and sample sizes are too small for adequate power in tests of significance. In many cases, field study objectives may be better served by various kinds of sampling procedures, even though the resulting inferences will be weaker than those observable through controlled experiments (Eberhardt and Thomas, 1991).

Given the difficulties presented above regarding sampling design, the potential complexity of any system to appraise experimental design was evident. In order to simplify the process for a practical design appraisal, some very basic considerations for every experimental design had to be addressed before any experimental design was concluded. Macdonald et al. (1991) listed a number of these issues:

- How many samples are likely to be needed to characterize a parameter with a specified degree of uncertainty?

- How many samples are likely to be needed to determine if there is a difference between locations, or a change over time?
- Where and when should samples be taken?
- Which parameters should be measured?
- How will the precision and accuracy of the data be assured? (Macdonald et al., 1991).

Every experiment had potential sources of confusion or error. Hurlburt (1984) suggested a number of methods to deal with general sampling design problems:

- Control treatments could eliminate errors due to temporal change and procedural effects
- Experimental bias could be dealt with through the use of randomized assignment of experimental units to treatments, randomization in conduct of other procedures and "Blind" procedures in measurements with a large subjective element;
- Experimenter-generated variability (random error) could be dealt with through replication.
- Initial or inherent variability amongst experimental units required replication of treatments, interspersion of treatments and concomitant observations
- The impingement of chance could be dealt with through the replication of treatments and interspersion of treatments (Hurlburt, 1984).

The problems of design of monitoring programs also had to be considered. Bernard et al. (1993) suggested that the four things to consider in designing a sampling program for a monitoring task were time, space, stressors and indicators. Thomas (1992) suggested that the design of a monitoring program would involve making decisions on what to sample; sample replication (sample size); where to sample; when to sample; and how to sample. He suggested that guidance for making decisions on these decisions would come from baseline data on biological, chemical and physical parameters; experience in previous studies in the area; and scientific judgement (Thomas, 1992).

All the experts presented above suggested variations on a general theme with respect to the evaluation of experimental design. Each approach emphasized a number of fundamental characteristics common to all well-executed designs, which included proper planning, effective QA/QC, and control over variables. Planning would provide a groundwork to identify the appropriate study methodology and the number and locations of samples to be collected. Planning would ensure that the statistical requirements for the

tests were identified and the methodology accommodated those requirements. QA/QC would ensure that the sample collected was not contaminated or altered and control of external variables would ensure that the sample was representative of the matrix.

4.5.1 Practical Issues

A number of practical considerations had to be addressed before any appraisal of experimental design could be concluded. Three of these could seriously impede the progress of an experiment and are considered in detail below, they are sample size, error types, and power analysis. A number of additional issues were addressed in Eberhardt and Thomas (1991). They included: the use of multivariate analyses, checking for internal replication, identifying control strategies, confirming randomization of treatments, and avoiding pseudoreplication.

With regard to sample size, Spellerberg (1993) pointed out that there is no wholly satisfactory way of determining the number of sample replicates required in environmental sampling but did suggest that some general guidelines. In discussing this it must be understood that Spellerberg (1993) did not immediately include the factors of sub-sampling, but rather, he dwelt on the more fundamental question of what was the minimum number of samples needed to give a defined degree of certainty. The reason a minimum number was considered involves the fact that having more data than needed has generally not been considered a major disadvantage in experimental design, but having too few samples can lead to the financial consideration suggested by Hayne (1978).

> Where such preliminary examination indicates doubt that the available resources can support an experiment powerful enough to reliably detect a reasonable level of response, then probably the experiments should not be run (Hayne, 1978).

Determining a minimum sample size varies on conditions and type of experiment. Manly (1992) suggested that the equation:

$$n = \frac{N\sigma^2}{(\frac{N\delta^2}{4+\sigma^2})}$$

Where N is the overall population size
σ is the population standard deviation
δ is the magnitude of effect (the true difference between the two population means being tested or the accepted margin of error)

gave the sample size that was required in order to achieve the margin of error as a function of population size. Thomas (1992) pointed out that σ^2 was determined by baseline studies, measured in pilot programmes, or in some cases guessed based on previous experience.

Manly (1992) continued that if population size was unknown or is very large then the equation became:

$$n \approx 4\sigma^2/\delta^2$$

Manly (1992) suggested that this equation was especially appropriate if the population size was unknown since it gave an upper limit to the sample size that was required for all population sizes. Both of these equations required the population standard deviation (σ) to be known. When an approximation was needed, then the range of values divided by 4 would give an approximate value for σ. The theory was that for many distributions, the effective range would be the mean plus or minus about two standard deviations (Manly, 1992). Both Manly and Thomas stressed that the process by which the effective sample size was determined was iterative. In order to determine an appropriate sample size to monitor an effect, one required an idea of the magnitude of the standard deviation of that effect. The action of carrying out those iterations was research. Without adequate knowledge of the magnitude of the effect it would be impossible to determine an appropriate sample size.

Two types of errors could arise over the decision to reject a null hypothesis being tested, types II or I. A type I error (α) is the declaration of the hypothesis to be false when it is

actually true; a type II error (ß) is the failure to falsify the hypothesis when it is actually false (Scheiner and Gurevitch, 1993). Scientists are most familiar with the pitfalls of type I errors, and have been, for some time, obligated to publish α's for all statistical tests done (Toft and Shea, 1983). Accordingly, there is a critical level of α of 0.05, which is a widespread, almost inviolate, convention (Toft and Shea, 1983). In contrast, type II error is rarely estimated in research, and there is no single critical level for its widespread use (Toft and Shea, 1983). Toft and Shea (1983) pointed out that the reason for this is that scientists are by nature cautious, and to jump prematurely to a false conclusion is considered more of an impediment than failing to detect a result (which may be discovered in the next experiment).

Table 4.2. Four Possible Outcomes for a Statistical Test of a Null Hypothesis. The probability for each outcome is given in parentheses. (From: Peterman, 1990 and Toft and Shea, 1983)

	Decision	
State of Nature	Do Not Reject Null Hypothesis	Reject Null Hypothesis
Null Hypothesis True	1) Correct (1-α)	2) Type I error (α)
Null Hypothesis False	3) Type II Error (ß)	4) Correct (1-ß)

Power analysis is a test of the likelihood that a type II error has been made and is defined as 1-ß in Table 4.2. Power reflects the probability of correctly rejecting H_o (Peterman, 1990). In the environmental field, the cost of a type II error often exceeds the cost of type I errors (Peterman, 1990). As an example, in fisheries research if a fish stock is being studied and is rapidly declining in abundance, but is being managed as if it were relatively constant because of low-power data, a type II error could lead to collapse of the stock and loss of all future revenue. If the stock were incorrectly believed to be declining when it was stable due to a type I error the cost would only be a reduced catch and resulting revenue. The cost of reduction in fishing in the latter case would be smaller than that caused by a complete loss of the fishery (Peterman, 1990; Clark, 1976).

Thomas (1992) indicated that the power of a statistical test is determined by five study design parameters:

1) Significance level (α) of the test;
2) Number of sampling locations;
3) Number of replicates;
4) Minimum detectable difference specified for the monitoring variable; and
5) Residual error variance (i.e. natural variability within a system)

The greatest increase in power would be produced by a reduction in the residual term. Although an increase in sample size could improve the power of a test, a far greater increase in power could be achieved by identifying additional sources of variation and removing them from the residual by design (Thomas, 1992). Like the case of sample size, statistical power could be increased through an increased understanding of natural variation and effect size. Environment Canada and Fisheries and Oceans Canada (EC/DFO, 1993) and Green (1979) in their discussions of power and spatial considerations both pointed out that the determination of site numbers to increase power must be derived in an iterative manner and that the solution converged to a stable value of sites after a few iterations (EC/DFO, 1993).

Given these considerations an experimental design decision tree was created for NCIS. It had two major tasks: to determine whether an appropriate analysis method was used, and to ensure that the actual analysis was designed to match the formal statement. Both tasks were incorporated into a single decision tree, which appears as Excerpt 4.2. The tree, as presented, does not include the detailed guidelines which can be found in Fyles et al. (1996).

1)Is there sufficient information available to allow for an analysis of the experimental design?
- no → the experiment type is an "observation"

2) Is the experiment type "observation"?
- no → continue
- yes → proceed to the Experimental Execution/Outcome Appraisal

3) Does a conditional null hypothesis either stated or understood exist?
- no → restate formal statement and experiment type and restart decision trees
- yes → document hypothesis and continue

4) Is the experiment type "research"?
- no → this implies that the experiment is "monitoring"; Proceed to question 10
- yes → continue

5) Is there a formal method to interpret the data produced by the experiment?
- no → restate formal statement and experiment type and restart decision trees
- yes → document method to interpret data and continue

6) Are there sufficient demonstrated controls to correct for the effect of external effectors in the experiment?
- no → restate formal statement and experiment type and restart decision trees
- yes → document controls and continue

7) Is the experiment a bioassay?
- no → proceed to question 9
- yes → continue

8) Does the bioassay use an appropriate experimental design/analysis method?
- no → restate formal statement and experiment type and restart decision trees
- yes → document a summary of the method and proceed to Experimental Execution/Outcome Appraisal

9) Does the research experiment use appropriate methodologies?
- no → restate formal statement and experiment type and restart decision trees
- yes → document a summary of the method and proceed to Experimental Execution/Outcome Appraisal

10) Are n (estimated population size), σ (the population standard deviation), δ (the magnitude of the effect), α (the significance level of the test), and β (the power of the test) known and incorporated into the monitoring design?
- no → continue
- yes → proceed to Experimental Execution/Outcome Appraisal

11) Are there extenuating circumstances that would allow the monitoring to continue without sufficient process knowledge?
- no → restate formal statement and experiment type and restart decision trees
- yes → document understood process knowledge and extenuating circumstances and proceed to Experimental Execution/Outcome Appraisal

Excerpt 4.2 Experimental Design Decision Tree from Fyles et al. (1996)

4.6 Experimental Execution and Outcome Appraisal

It is generally accepted that the appraisal of experimental execution is a fundamental step in every experiment in the natural sciences; it is practised either informally or formally every time results are published. As Hurlbert (1984) put it:

> Experimental design and experimental execution bear equal responsibility for the validity and sensitivity of an experiment. Yet in a practical sense, execution is a more critical aspect of experimentation than is design. Errors in experimental execution can and usually do intrude at more points in an experiment, come in greater numbers of forms, and are often subtler than design errors. Consequently, execution errors generally are more difficult to detect than design errors, both for the experimentor himself and for readers of his reports. It is the insidious effects of such undetected or undetectable errors that makes experimental execution so critical (Hurlbert, 1984).

In its simplest form, the experimental execution appraisal asked the question: Was the work done correctly? More exactly, it asked, was the experiment carried out as described in the experimental design? If the answer was "yes" then the execution was documented, if "no" the appraisal asked: did changes from design to execution significantly affect the ability of the experiment to test the formal statement?

The experimental execution appraisal process was designed to analyse whether the experiment actually did what it was designed to do. The experimental execution and outcome step had to complete two tasks: appraise the measurements that make up the experimental event; and determine whether the experiment achieved the desired outcome based on the appraised experimental design. The appraisal step evaluated the individual measurements that made up the experiment using the discipline-specific decision trees presented previously. Once the individual ratings had been derived for each measurement in the experiment then an appraiser could proceed to the second step, which determined whether or not the experimental execution/outcome achieved the desired outcome of the design using the decision tree displayed in Excerpt 4.3.

1) Was the experimental execution documented in a manner that would allow for an appraisal to be carried out?

- no → assign experiment type as "observation", assign an overall reliability rating of "2" and complete documentation using the Pedigree forms
- yes → continue

2) Was the experiment type "observation"?

- no → proceed to question 4
- yes → continue

3) Did all observations in the group receive the same reliability rating?

- no → regroup measurements into events based on equal reliability ratings and re-examine experiment type and formal statement for each event
- yes → document the "observation" using the Pedigree forms

4) Was the event carried out according to design so as to meet documented requirements stated in experimental design?

- no → proceed to question 8
- yes → continue

5) Did all measurements in the event have the same reliability rating?

- no → regroup measurements into events based on equal reliability ratings and re-examine experiment type and formal statement for each event
- yes → continue

6) Did the experiment test the documented null hypothesis?

- no → re-examine formal statement and null hypothesis
- yes → continue

7) Was a reliability measurement rating less than "3" given?

- no → assign overall reliability based on weakest-link principle and complete documentation using the Pedigree forms
- yes → document experimental outcome using the Pedigree forms

8) Were unexpected processes encountered?

- no → continue
- yes → restate experiment type, formal statement and null hypothesis

9) Were insufficient data (replicates/controls/blanks) collected to meet the stated design?

- no → continue
- yes → re-examine formal statement and null hypothesis

10) Arriving here assumes that an abrupt change to some controlled or understood condition or process influenced some variable in the outcome of the experimental event,

Re-examine the experiment type using the information in guideline 10 to establish how to accommodate this condition

Excerpt 4.3 Experimental Execution/Outcome Appraisal Decision Tree from Fyles et al. (1996)

4.7 Pedigree Creation

The pedigree was a statement about the goals of the experiment, the reliability of the measurements made as part of the experiment, and an expression of the confidence level of the results. For some "monitoring" or "research" experiments the goals, measurements, and confidence levels could be expressed concisely using statistical concepts. For other experiments, particularly involving "observations", the measurement reliability could be easily indicated using the 0-4 scale, but the confidence level would be given as an expert opinion from the appraiser. The generation of a pedigree was the logical extension of the QA/QC process applied to the entire experimental activity. The essence of QA/QC applied to a measurement was documented adherence to detailed procedures. By extension, a component of the pedigree was simply the documentation of adherence to the appraisal protocols.

The full pedigree in NCIS included controlled terms and descriptions. The first decision involved determination of experiment type from the three choices provided. This was followed by the determination of the formal statement type. The first descriptive field provided a location where the hypotheses tested could be explicitly stated. For well-structured monitoring programs and experiments testing null-hypotheses, defined acceptance values could be known. These could be documented as design requirements. The appraisal of the measurements using the appropriate discipline-specific decision trees followed. Each of these processes returned a reliability rating and documentation to this part of the appraisal. Once all measurement appraisals were complete, the experimental outcome was documented. This might be a statistical statement concerning the hypothesis tested, or might simply be a qualitative statement about the strengths and limitations of the results in the event. The opportunity existed for appraisers to include some appraiser's comments so that specific comments that the appraiser felt were relevant to understanding the appraisal could be recorded.

4.3 Application of Appraisal

4.8.1 Tutorials

In order to train appraisers, tutorials were developed. The tutorials demonstrated how to appraise a variety of different types of data collection efforts using templates and examples that appraisers could apply to their own work. A full tutorial document was prepared that included five sample appraisals made up of one for contaminant surveys, two for biological responses, one for bioassays and one complete experimental event appraisal. In order to demonstrate the practicality of the appraisal process each tutorial used actual articles from the academic press. The example presented in section 4.3.2 was one such article. The papers were chosen based on a number of criteria including relevance to the discipline being appraised and unusual properties that could be used to emphasize particular points in the appraisal process. The Jones and Hara (1988) paper discussed in section 4.3.2, as an example, not only presented an interesting behavioural paper but also demonstrated how to deal with an experiment where external factors resulted in significant changes in the methodology in mid-experiment.

It was recognized that creating tutorials that appraised real papers had the potential to chagrin authors and infringe copyrights. Consequently, before the tutorials were completed, copyright holders were contacted and their leave to use the material obtained. In addition, individual authors were reached and presented with our intentions and the completed tutorials. In all cases the authors agreed to the use of their work and in some cases additional insight was gained, which improved the document.

The development of tutorials served as a useful measure of the usability of the appraisal protocols. It exposed a number of areas of difficulty, which were addressed using FAQs (frequently asked questions). The FAQs covered a number of areas of concern including the difficulty in determining experiment type; the use of overly broad or open questions; and the use of controls in research. Once the tutorials were complete it was necessary to

train individual scientists to use the appraisal protocols in order to evaluate their own research. This was carried out at a workshop.

4.8.2 Workshop

The initial version of the appraisal protocols developed for NCIS was completed in August 1996 and was distributed widely throughout DFO. Following this distribution, a workshop was held at the University of Victoria. It was attended by 15 senior scientists and data managers from DFO and the EPA and served as the first test of the system by reviewers. These reviewers had the task of learning to use the protocols, deciding whether they should be changed, and if so, how? In order to prepare them for the workshop, all attendees were supplied with the tutorials and the protocols document in advance of their arrival.

The workshop was broken into a number of sections with the experimental events protocols being analyzed near the end of the process. This gave the workshop participants an opportunity to become comfortable with the discipline-specific decision trees and the appraisal process. As was the case with many of our training exercises, several of the attendees expressed concern over the practicality of the system prior to beginning the process. These concerns when expressed verbally usually were of the form: "this system can never be applied effectively". As was the case in our previous experience, this concern was quickly allayed when it became apparent that the system was designed not to supplant experts but to provide them with an added tool in their work.

The decision to start with the basic decision trees also appeared successful. By the time the participants had reached the section on appraising experimental events, they appeared very comfortable with the task of appraising individual measurements. The task of appraising experimental events, however, caused a good deal of controversy. The major problem dealt with the concept of formal statements. The participants generally agreed

that that our attempt to make the system user-friendly had resulted in a break from the strict use of controlled language. The participants interpreted the process of assigning formal statements to be a task that did not meet the requirements of standardization and controlled language seen in the other areas of the appraisal process. In particular, they felt that the assignment of formal statements was too open to individual interpretation. It became apparent that while the system was intended to include two tiers of controlled terms and a tier of comments, in practice it was treated as a single tier of controlled terms and a tier of detailed comments.

After a good deal of discussion, it was agreed that non-specialist users would benefit from the two tiers of controlled terms. Whether this added benefit was outweighed by the effort needed to create the formal statements was the issue to be addressed. In the end, the group decided that the abstract increase in usefulness for an unknown user did not warrant the increased effort by the known scientists. It was decided to combine the two tiers of controlled terms from the process used in NCIS by expanding the number of experiment types from three to seven based on a combination of formal statement with experiment type. The list of experiment types was expanded from the three proposed previously: research, monitoring and observation, to seven: observation, research – exploratory, process and confirmatory; and monitoring – scientific, compliance and trend. This list lost the idea of deductive research as well as the role of model development as both a monitoring and research task. This was acceptable to the review group for their particular task of creating a functional NCIS system for DFO. It did not address important aspects of the general model presented in this work and will be addressed later for that purpose. Following the discussion, a new decision tree was created (Excerpt 4.4) to combine the concept of formal statement with experiment type and new definitions for experiment types were presented (Table 4.3).

1) Was there a stated, specific purpose for the data collection?
- No → Experiment type = Observation.
- Yes → Continue

2) Was pre-existing theoretical knowledge incorporated into the data collection design?
- No → Continue
- Yes → Go to Question 4

3) Was the stated purpose of the data collection to develop process or theoretical knowledge?
- No → Experiment Type = Observation
- Yes → Continue

4) Was the data collected for the purpose of testing a stated or understood null hypothesis?
- No → ExperimentType = Observation
- Yes → Continue

5) Was the data collected to model theoretical relationships or to monitor changes predicted by stated theoretical relationships?
- No → Continue
- Yes → The Experiment type is some type of Monitoring. Go to question 11

6) Were controls in place to limit the possible outcomes of the experiment?
- No → The Experiment type is some type of Monitoring. Go to question 11
- Yes → Continue

7) Were acceptance criteria for the results clearly stated?
- No → Experiment type = Observation
- Yes →Arriving here indicates some type of research Research is involved continue

8) Was the experiment intended to establish a quantitative effect level?
- No → continue
- Yes → Experiment type = Research - corroboration

9) Was the experiment intended to refine a previously tested hypothesis based on statistically significant experimental results?
- No → continue
- Yes → Experiment type = Research - corroboration

10) Was the experiment intended to establish the statistical basis of a previously tested hypothesis?
- No → Experiment type = Research – exploration
- Yes → Experiment type = Research – process

11) Was the monitoring designed to establish compliance to an established norm?
- No → continue
- Yes → Experiment type = Monitoring – compliance

12) Was the monitoring designed to establish a trend in research or response?
- No → Experiment type = Monitoring – research
- Yes → Experiment type = Monitoring – trend

Excerpt 4.4 Determination of Experiment Type

Table 4.3 Experiment Types

Experiment Type	Goal of Experiment	Nature of the Experiment
Monitoring – Compliance	Monitoring to identify variance from accepted norms. Null hypothesis not stated but understood.	Monitoring is mandated and only methodology is under the scientist's control.
Monitoring – Research	Monitoring to test null-hypothesis based on demonstrated process knowledge.	Based on working theory derived from several iterations of proposing and testing process knowledge hypotheses with experimental work (Testing using field tested process knowledge understood to a first order of magnitude). This is second level research.
Monitoring – Trend	Monitoring to test null hypotheses based on inferred process knowledge.	This involves testing working theories based on several iterations of proposing and testing hypotheses regarding process.
Research – Corroboration	Testing to develop or refine previously tested process knowledge.	This includes toxicity testing and involves testing hypotheses regarding effects of contaminants or conditions.
Research – Process	One or more untested working hypotheses used to determine the relationships between and among parameters to provide an understanding of system dynamics.	This is process knowledge research, testing process hypotheses.
Research – Exploration	Involving one or more untested working hypotheses derived from general observations not process knowledge.	This is pure baseline research designed to develop process knowledge.
Observation	Data collected as part of a general survey, as part of another research study, or incidental to other activities.	Data collected as part of a general survey, as part of another research study, or incidental to other activities.

4.8.3 Practical Application of the Protocols

The first version of the NCIS protocols was completed in 1996 with the workshop producing changes that were distributed in October 1996. Since that time the protocols have been available for use when entering data into NCIS. Changes in focus at DFO, however, have limited our ability to investigate the effectiveness of the appraisal protocols for experimental events. (Sowden, T., pers. comm., 1999). As mentioned previously, the full protocols are used only to appraise experimental events, while survey events are appraised using only the appropriate discipline-specific protocols. In light of decreased

funding at DFO, it was felt that NCIS had to appear to have a great deal of data very quickly; this could best be accomplished by entering survey rather than experimental events (Smiley, B., pers. comm., 1998). In March 1999, it was estimated that NCIS contained about 6000 events in the inventory referencing over 2 million observations but almost all these were survey events (Keeley, B., pers. comm., 1998). Specific DFO districts have addressed the need to input data in different ways and as a result some regional variation can be seen.

In the Maritimes it was decided to prioritize data input to survey events rather than experimental events in order to maximize the amount of data that could be input into the system (Vromans, A, pers. comm., 1998). A relatively large amount of data of this type was entered into the NCIS without being appraised (Vromans, A., pers. comm., 1998). The QA/QC information associated with this data was relatively complete and was stored in the archive with the data. Plans exist to perform an appraisal en mass for these data since all criteria have been met to rate them as level 4 data although the rating details may not be entered for these data (Vromans, A., pers. comm., 1998). In this region it was agreed that all subsequent data would be appraised using the appraisal document procedure. The principal investigator along with the data manager would perform the appraisal. To date, two complete datasets comprising approximately 16 events have been successfully appraised in this manner (Vromans, A., pers. comm., 1998).

In Newfoundland, the effort stressed computerizing information from scientists nearing departure due to retirement. Two senior scientists were trained to use the appraisal protocols with the regional data manager (Linda Fancy) serving as their instructor (Fancy, L., pers. comm., 1998). The data manager worked directly with the investigators. All the data input to date has been survey events with the majority being chemical in nature. The investigators reported little difficulty using the discipline-specific protocols, although the data manager reported that user-friendliness could be improved in order to reduce the need for "hand-holding" (Fancy, L., pers. comm., 1998). The major complaints by the

investigators involved technical disagreements with specific guidelines (Fancy, L., pers. comm., 1998).

In the Laurentian district 689 survey events and 16 experimental events had been input into the system by March 26, 1999 (Guay, C., pers. comm., 1998). The Laurentian district did not, however, use the protocols directly to do the appraisal. The protocols, instead, served as a reference that accompanied their in-house appraisal system. This was due to issues of language and complexity. The Laurentian district operates primarily in French and the decision trees and guidelines have only been produced in English. Administrators felt that the English protocols were unduly complex for the French technicians and since the district was being pushed to quickly add data to NCIS a decision was made to develop a less complex method (Trembley, G., pers. comm., 1998). The district managers, therefore, developed a "walkthrough" for the evaluation of survey events. It consisted of a simple translation of the chemistry collection and storage decision trees and combined the precision and accuracy decision trees into a single QA/QC tree. The guidelines in the walkthrough were much less detailed than those from the full, discipline-specific appraisal protocols. As an example, they consider an acceptable collection, storage or analysis method to be:

> a method that is in agreement with the most current accepted and published data in any particular given field (Trembley, G., pers. comm., 1998).

At time of publication, the Great Lakes district had not implemented the appraisal protocols to their data. They were developing a software tool to input data into the system and since the software was not complete had not entered or appraised any data (Carswell. A., pers., comm., 1998).

The Pacific branch also emphasized the input of survey data into the system. They hired an experienced data appraiser and entered large amounts of biological and chemical data into the system (Sowden, T., pers. comm., 1999). Co-op students from the University of Victoria were hired and successfully trained to appraise survey events using the protocols. Some experimental events were entered into the system but in all cases, the appraisal was

carried out by an experienced appraiser (Pawluk, M. pers. comm., 1998). By using a trained appraiser instead of the researchers, this activity failed to provide an opportunity to examine how investigators might approach appraisals. Instead, investigators were only involved when questions arose with their data collection efforts (Pawluk, M. pers. comm., 1998).

In summary, it would appear that DFO has not emphasized inputting experimental events into NCIS. The system managers could not determine precisely how many experimental events had been appraised; however a low-level of commitment to the appraisal process for context is evident.

4.9 Evaluation of the Decision Trees to Appraise Experimental Events

The 1996 workshop provided a great deal of useful feedback on the development and implementation of the process for assigning contextual information. As is evidenced by the discussion surrounding assigning experiment types and formal statements, the workshop attendees were both interested and informed reviewers. That said, the outcome of the workshop was not entirely positive. A great deal more criticism was anticipated over the "nuts and bolts" of the decision trees themselves. It has been proposed that the reason that the decision trees were not more soundly criticised at the workshop had to do with the fact that the reviewers did not intend to use them directly. In the CODIS case study it was noted that experienced appraisers often carried out appraisals without directly referring to the decision trees. If this were the case, then the fact that the reviewers did not criticise the decision trees might indicate that they felt that their expertise would essentially eliminate the need to use the decision trees. The appraisers, being experienced scientists, no doubt felt that they could identify errors in experimental designs and document those occurrences. This would explain their allowing the decision trees to "get by" untouched. This supposition cannot be effectively tested.

In order to provide a more complete evaluation of the appraisal process, a retrospective analysis of the decision trees was carried out. This analysis included a number of tests involving real datasets as well as virtual experiments similar to those used in the CODIS case study. As with the CODIS case study, the major activity of the analysis involved appraising journal articles in order to provide an anecdotal evaluation of the protocols to evaluate context. Seven issues of the *Canadian Journal of Fisheries and Aquatic Sciences* (1998-1999) were obtained and of the 122 articles, 30 were chosen for context evaluation. The choice of journal was intended to reflect the expertise of the author since the intent was to compare the output of the protocols with the expert opinion.

In each case, the article was read and in the absence of the appraisal protocols a decision was made regarding the type of activity. Following that determination, the trees were applied to the article and the outcome determined. Of the 30 articles examined the expert determination and the contextual protocols produced comparable outcomes for 29. The single disagreement involved the secondary use of pre-existing data and will be discussed in detail below. In addition to the one direct disagreement, the analysis also uncovered one serious omission, a number of minor problems and some logic errors in the various decision trees. The minor issues involved missing information in the guidelines, which could be easily solved through simple additions of information. The logic errors will be discussed below.

The omission involved the use of guidelines. A key feature of the decision tree approach developed for this work was the inclusion of guidelines to assist the appraiser. The experiment type decision tree in the experimental events appraisal process (Excerpt 4.4) did not include guidelines. This lack of guidelines limited the ability of the decision tree to deal with unusual cases.

The one disagreement between the expert appraiser and the experimental type decision tree involved the secondary use of pre-existing data. This example exposed a minor problem involving documentation of the intention of a research activity, as well as an

underlying inconsistency in the way the decision tree distinguishes between monitoring and research. In order to effectively consider this inconsistency it is necessary to deal with the lesser problem first.

Consider a case where data collected for one purpose is subsequently used for another. A typical example would be where carefully controlled regulatory data was subsequently used for research purposes. The experiment type decision tree (Excerpt 4.4) could incorrectly assign any subsequent work using this data as "observation". The first four questions in the decision tree only consider the first time the data was collected, in this example that would be for regulatory purposes. They do not acknowledge the second case where intent existed to assemble pre-existing data. It would be possible to design a research program to take the limitations of the pre-existing data into account. Through the appropriate break-up of the original data an entirely acceptable dataset could be derived for research purposes. It must be made clear that the first four questions could also address the intent to assemble pre-exiting data for the new use. This error was minor and could be solved through an appropriately worded guideline.

Of more significance was the issue of monitoring and research as determined by the experiment type decision tree. This related to the above discussion involving the secondary use of pre-existing data. In that case it was determined that the exercise was either some type of research or monitoring. The issue that had to be addressed was whether the decision tree was able to effectively distinguish between the two. Consider a real example where research is being conducted on the behaviour of Scotian Shelf Silver Hake (*Merluccius bilinearis*) using data derived from regulatory observers (Gillis, 1999). In this example the researcher made use of the detailed catch and cruise data collected by regulatory observers on 12 Russian and 12 Cuban trawlers. The records from the observers included information about catch composition, gear configurations, the initial and final ship positions, and the initial and final times of each trawl (Gillis, 1999). The research described in Gillis (1999) involved testing a hypothesis on the variability in catch rates as a result of trawl length and vessel interactions. The theoretical base of the

research dealt with the recognition that while catch-per-unit-effort has been used to evaluate populations of fish; this relationship might break down due to interferences and the reaction of the fish to the fishing effort. To an outside observer this experimental activity would clearly be "research-exploration". It involved the development of process knowledge and had a strong design thus excluding "observation"; and the background knowledge was clearly insufficient for it to be any type of monitoring.

Using the decision tree in Excerpt 4.4 the following answers were derived:

1- Yes, there was a stated purpose for the collection;
2- No, pre-existing process knowledge was incorporated into the design;
3- Yes, the data was collected to develop process knowledge
4- Yes, the data was collected to test a stated null hypothesis
5- Questionable, without guidelines either yes or no might be applicable
6- No, controls were not placed on the data.
7- Yes, acceptance criteria were clearly stated.

Based on these answers a number of difficulties were identified and will be discussed below. The difficulty in deriving an answer for question 5 reflected poor wording and the lack of appropriate guidelines. It was evident that the first part of the question addressed activities specifically involved in monitoring, this was not well expressed in the question. The second part of the question was entirely too broad. "Was the data collected to monitor changes in stated relationships?" Without a detailed guideline, the user would be most likely to say "yes" thus incorrectly classifying the activity as "monitoring".

Question six asked about the use of controls. While controls were necessary for research, the question did not acknowledge that the data might be structured in such a manner as to be controlled from the outset, as was the case in this example. The researcher did not incorporate controls thereafter because they were unnecessary. The negative response also resulted in the activity being classified as monitoring which subsequently was refined to "monitoring-research". If an appraiser chose to disregard the "controls" issue, the result would be to have sent the user to questions 9 and 10, which were insufficiently well

worded to distinguish between research-corroboration and research-exploration. A clear set of guidelines might have been able to address many of these issues.

The experimental design and outcome decision trees included guidelines and consequently dealt effectively with the majority of cases. They did, however, contain areas of concern. In both trees, "no" responses often called for experiment type to be reconsidered. In the example discussed previously should the appraiser have decided that the lack of controls resulted in a negative response to question 6, the outcome would be to reclassify the data using the experiment type decision tree. That tree would re-iterate that the work was research (or monitoring) and would create a loop. Since no process existed to exit the loop the result would be that the appraiser would have to disregard the question to carry out the appraisal. In addition, the experimental execution/outcome tree (Excerpt 4.3) had a logic error in question seven; both the yes and no answers had the same outcome. If question seven was legitimate then a negative response should have a differing outcome from a positive response.

Following the examination of the decision trees it was apparent that the decision tree methodology developed to appraise disciple-specific results could also be used to appraise experimental events. For the purposes of this research the trees served their tasks and while errors existed, these errors were of a practical nature and did not impugn the processes developed. It was equally apparent that substantial work will be required to refine the decision trees for continued use in DFO. This new research will require the addition of detailed guidelines and improvements on the wording of questions to reduce user uncertainty.

4.10 Evaluation of the Overall NCIS Protocols

The experimental event appraisal protocols detailed in Fyles et al. (1996) provided a tool to evaluate and document contextual information to be included with scientific data in information systems. The experience in the NCIS case study demonstrated that it is

possible to develop and implement such a process in a real system. What the NCIS case study failed to demonstrate was whether an appraisal system could be applied effectively in NCIS or any other similar system. The process of developing the protocols for NCIS provided valuable insight into the issues that must be addressed for designers of subsequent systems or appraisal methodologies. These insights were incorporated into the general model presented in Chapter 6. By evaluating the factors that limited the effectiveness of the approach described here, future researchers can both anticipate and bypass these roadblocks in the next generation of information systems.

In the definitional phase of the NCIS project, it was stated that survey events would not be appraised. By setting up a process though which users could input data into NCIS without appraisal, a substantial loop-hole was created. When the appraisal protocols were developed, it was anticipated that a small minority of datasets would enter NCIS as survey events since most of the data collection activities performed by DFO are monitoring in one of its types. As is evidenced by the anecdotal data presented above, this assumption was in error. Allowing data entry without appraisal was clearly a mistake. As a result, NCIS now contains some 6000 datasets and 2 million measurements that are not associated with contextual and reliability metadata. Compare this to the CODIS system, which contains some 4600 DS_IDs and approximately 1 million appraised measurements. The compilation of the CODIS data required over 14 person-years of funding. This suggests the level of effort that will be required to rectify the error in NCIS.

Given the potential benefits of appraised data for secondary users, the question must be asked: why don't data-originators include appraisals with their data? In discussions with data managers, a number of arguments were made including work load and time demands on appraisers. One of the most telling was of a lack of demand. As one data manager put it:

> We all know the arguments about the merits of doing appraisals. But of all the inquiries I've had no one has asked for or cared about appraised data. I'm afraid its come down to a matter of letting the need dictate the effort. So far no need, so no effort. It'd be nice to be proactive and work to some uncertain future demand, but this just isn't realistic. If someone wants to appraise data they get from NCIS, we

> have the QA stuff, the descriptive information, and the protocols such as they are. It'll be a case of do-it-yourself. That's the best we can do for now (Sowden, T., pers. comm., 1999).

This argument demonstrates how a lack of acceptance by researchers and education of users can limit the development of a process like the protocols to appraise experimental events. The difficulty lies in the form of this argument. Most users do not know about the existence of appraised data, therefore, they cannot know to request it. Since users are not requesting appraised data, the organization does not see the need to appraise data. Only through an education program about the existence and importance of appraised data can awareness be raised to the point where users will request that data. The reason that user and scientist are treated as being linked is that NCIS was designed primarily as an internal system. Thus, the scientists who are importing their data into NCIS are also the primary consumers of the data derived from NCIS. Without an increased education of the DFO scientists as to the value of appraised data, it is unlikely that a demand will appear. Some will continue to argue that it is not needed in any case since the scientists at DFO deal with such a specialized field that virtually all their scientists are able to appraise each other's data. The creation of an appraisal system might, therefore, be seen as bureaucrats trying to impose tighter controls on individual researchers.

Other practical issues must also be considered. Through discussions with data managers it became apparent that one of the major reasons for not implementing the appraisal system was a lack of time and resources. Learning to use the appraisal protocols takes an investment in time as does applying them. While it may be argued that this investment will be recompensed in increased productivity later on, it must also be acknowledge that the specialist in his or her own field benefits the least from appraised data. The major benefits of appraised data are derived by those individuals who can use the appraisals to inform themselves about the reliability and applicability of potentially useful work outside their field of expertise.

In the case of NCIS, the institution (DFO) receives the greatest benefit of appraisal. This is because the availability of appraised experiments results in better access to pre-existing datasets, which may allow DFO scientists to avoid costly duplication of experiments and monitoring activities. Unfortunately, no comprehensive institutional procedures exist within DFO to compensate the individual appraiser for work that increases institutional efficiency. Indeed, several attendees at the workshop argued that appraisers might see a reduction in their output as a result of taking the time to learn to and then appraise their datasets. This decrease might be interpreted to be a decrease in productivity, which could threaten advancement. It is only through the development of institutional procedures that provide appraisers with an opportunity to develop and apply their appraisal skills that such skilled practitioners will be trained.

An additional issue that emerged from discussions with DFO researchers was the concern that external scrutiny of one's data had the potential to damage one's reputation within DFO. A conscientious appraiser, upon recognizing that errors occurred in the experimental process, would be required to give a lower rating to their own data. The researchers have indicated that a system like NCIS has the strong potential to be used to evaluate individual performance. Managers could use NCIS to determine which researchers were performing the fewest experiments or were producing the lowest rated measurements. In an institutional setting where resources are tight, this could result in the loss of prestige and resources. This use of the appraisal metadata could result in a "chilly climate" for appraisers and has the potential to instil bias into any appraisal activities.

4.11 Case Study Outcomes

The NCIS case study served to clarify a number of critical issues and provided a number of advances for developers of environmental databases. An overall framework was developed to appraise measurements derived from both experimental activities and the sampling of environmental variables. This framework included a methodology to assign controlled terms to complex experimental and sampling activities as well as protocols to

appraise these activities. The creation of the framework identified key features of an appraisal system for experimental events and required the development of teaching tools to aid in training appraisers. Through the testing of the system in workshops and in in-house analyses, weaknesses in the initial design were identified and alternatives to improve the individual decision trees and the overall framework were proposed. The lack of adequate data limits the ability of this work to determine whether or not the system can be applied effectively in an institutional setting, but it did provide critical insight on the pitfalls associated with such an effort and proposed methods to avoid these pitfalls. These advances will be incorporated into the general model in Chapter 6.

NCIS is still in development and as a result it is impossible to demonstrate that the system itself is an improvement over archival systems. It is clear from the discussion in the case study that the NCIS model presents a number of significant advances over comparable systems. The inclusion of inventory and directory levels along with the archives provides a means for the single system, NCIS, to serve a great variety of uses. The inclusion of a method to relate the context of the measurements in the archive to potential users is a fundamental advance in the field. A serious consideration in this system, however, is the separation of experimental and survey events. Segregating survey events and not requiring context appraisal undermine the advantages provided by the system.

Chapter 5 EMS and Truncation Case Study

5.0 Introduction

The Environmental Protection Program (EPP) of the British Columbia Ministry of Environment, Lands and Parks (MELP) is the branch that has responsibilities for protecting air, land and water resources in the province (LGS, 1995). EPP has made use of numerous computerized archives: the Environmental Quality Information System (EQUIS), the System for Environmental Assessment and Management (SEAM) and most recently their Environmental Monitoring System (called EMS hereafter). EPP systems have been of continuing interest to this project. They have served as an environment from which tools and ideas have been derived, techniques tested, and models explored. Over 25% of the DS_IDs in the Continental Chemistry discipline in CODIS were derived from EQUIS and SEAM. A gap-analysis of SEAM provided many of the early ideas that were incorporated into CODIS and provided significant insight into general issues surrounding the long-term storage of environmental information. MELP EMS is of particular interest to this project as it was developed in parallel to the model developed in this work. The designers of CODIS and MELP EMS had regular contacts during the development of MELP EMS and a member of the CODIS development team sat on the MELP QA/QC Working Group. Consequently, EMS was used to examine the requirements of archival data to facilitate secondary usage.

EMS is MELP's most recent monitoring data repository. It is an Oracle/Unix application using client server technology which captures physical/chemical data as well as quality assurance, biological and toxicological data (MELP, 1997). MELP EMS has been designed to store high quality monitoring and compliance data, which it receives electronically from analytical laboratories. This ability to electronically download monitoring and compliance data provides a cost-effective and efficient means to reduce the number of data gaps in monitoring information available to users of the system (MELP, 1997). The EMS design includes the ability to store both raw measurement values as well as all the important accompanying quality assurance (QA) information

(Clark, M., pers. comm., 1998). EMS also incorporates an automated process to assign data reliability indicators (their Data Grader or QA Index). The QA Index function, which will be discussed in detail later, can automatically assign a "rating" to the data in the archive based on the accompanying QA information (Clark, M., pers., comm., 1998). This QA index can provide users with an indication of the quality of the data stored in EMS.

The appraisal process developed for CODIS was designed to indicate the reliability of datasets. It assumed that the actual values that comprised the datasets were exact and did not incorporate issues of uncertainty for individual values. The appraisal process developed for NCIS addressed the reliability of experimental events. It incorporated issues of uncertainty and reliability and it too assumed that the actual values that made up the measurements and data were exact. Neither process considered the requirements of the actual values that make up the data in archives. The reason for developing effective metadata systems is to encourage the secondary use of pre-existing data. This activity presupposes that the data being sought is worth the effort expended in its identification. Any activity that reduces the quality or reliability of the data in archives reduces the overall value of the any system associated with that archive.

An important issue to the designers of EMS was the inclusion of truncated data in their system. Data truncation is widely practiced by analytical environmental laboratories. It is based on a rich scientific tradition of reporting significant figures in reports, publications, and presentations throughout the scientific community. Traditionally, only values that are found to exceed the uncertainty imposed by the measurement process by a defined amount, are reported, together with an estimate of uncertainty. The reporting of significant figures is part of the process of communicating the precision of the data. Appraisal of uncertainty, calculation of precision and accuracy, and the truncation of values to report data of standard significance, is a critical final step in preparing data for communication.

Methods for the analysis of large amounts of information are growing rapidly in sophistication and complexity, together with data-management systems to accommodate ever larger amounts of data. Numerous individual scientists, analysts, and decision-makers are involved at various points in the process, each one examining data, appraising uncertainty, and truncating data to reflect the uncertainty to that point in the process. Increasing complexity is hence affecting the way significant figure truncation has been traditionally applied, and has the potential to alter the data in ways that no longer reflect the natural spread of values in the environment under study. This raises the risk that decisions will be made based on data, which have become unintentionally biased as they were being analyzed.

"Truncation" involves two distinct practices. The first applies to individual values, which are shortened by removal of digits. Digit truncation is typically done according to a rounding algorithm. Calculations, such as determination of a mean of a digit-truncated data set, are affected to a minor extent by the rounding process, but are generally of minor concern for decision-makers. The second sense is distribution truncation or censoring of data sets in which numerical values are replaced with text indicating the value lies below a defined lower level, typically the "reporting limit" which is related to the "detection limit". Censoring here is used in its strict statistical sense and is not intended to imply that laboratories are intentionally withholding valuable information or have anything other than data users' best interest in mind. Distribution truncation is a chronic issue in trace environmental analysis. It is not uncommon for a set of analytical results to contain a majority of reported values which are "less than" or "not detected".

The EMS data structure has been designed to accommodate either truncated data or data with uncertainty. Since data truncation is considered standard in the environmental analytical field the client laboratories, which were directly downloading data into EMS, insisted on providing truncated data. Consequently, research was carried out to investigate the issue of truncated data including how truncation is handled both in the laboratory and by data users. This research had two independent goals, those for EMS and those for this thesis research. The EMS goals involved an investigation of truncation

as it applied to MELP's data management and analysis needs. The aim was to address the following questions:

- What are the consequences of data truncation?
- What types of analyses and decision-making are most critically affected by data truncation?
- How can the impact of data truncation be limited?
- How do other agencies handle the data truncation issue? and
- How should MELP deal with data truncation?

The thesis goals for this research included the EMS goals but included one additional question:

- How does data truncation affect the design and implementation of metadata systems?

This question goes to the heart of the role of metadata systems, the identification of useful data for secondary analyses.

This case study reports the outcome of this truncation project. It begins with an analysis of EMS and then considers the issues surrounding data truncation. Once data truncation has been considered, the implications of truncation will be examined in light of the needs of data-driven environmental management and decision-making.

5.1 MELP EMS

In 1971, the Pollution Control Branch and the Environmental Laboratory of British Columbia's Ministry of Environment undertook the creation of a major computer data storage and retrieval system (Ellis and Clark, 1977). EQUIS was the outcome of this project. EQUIS was designed to handle inventories of effluent, air emission and refuse discharges in the province as well as store and process air and water quality monitoring information (Ellis and Clark, 1977). In conjunction with the creation of EQUIS, a laboratory management system (LABMAN) was also developed (Ellis and Clark, 1977). LABMAN included the capability to interface with the main EQUIS data files (Ellis and Clark, 1977). The EQUIS system was designed to be run on early IBM systems and

initially ran on an IBM 360 Model 40 under OS-MFT using magnetic tape and disk packs with random and sequential access (ISAM) (Clark and Ellis, 1976). The data system structure was based around two "major" files which detailed site information and test result information and several "minor" files which served as dictionaries to translate the larger files (Clark and Ellis, 1976).

By the early 1980's, technological advances allowed for much more powerful systems than EQUIS and in 1985, the System for Environmental Assessment and Management (SEAM) was developed. SEAM was a VAX RMS based application written in VAX Datatrieve and VAX Basic which supported terminal data entry of field and laboratory results from a dedicated link to the Ministry Laboratories as well as Lotus formatted data files and data formatted in the federal-provincial pulp and paper format (LGS, 1995). The construction of the SEAM data component carried over from EQUIS with "major" files being used for searches and "minor" files being used to translate the coding in the major files. The data files, while large, were "flat" [SEAM was not a relational database] but using the complicated coding system allowed a tremendous amount of information to be known about a dataset. An important addition in the SEAM system was the inclusion of Specific Parameter Analytical Route Codes (SPARCODEs). A SPARCODE categorized the combination of parameter and work route (LGS, 1995) which meant that samples were linked to their method of analysis and associated detection limits.

By the 1990's, with the advent of powerful GIS systems, it was recognized that SEAM no longer met EPP requirements (LGS, 1995). A new system was needed to store physical/chemical data; deal with critical biological data; and handle quality assurance information (BCMELP, 1997b). The outcome of the development process was the MELP EMS program in use today. MELP EMS is an Oracle/Unix based application using client/server technology (BCMELP, 1997b). MELP EMS contains physical/chemical data like its predecessors but in addition, it also captures quality assurance and biological data (BCMELP, 1997b). MELP EMS contains information on monitoring locations, samples and results and gets most of its information directly from the laboratories that analyze

monitoring samples and permit holders who submit monitoring data to the ministry (User's Manual, 1997). The original version of MELP EMS went on-line on November 4, 1996 and subsequent upgrades have been incorporated into the current system (BCMELP, 1997a).

MELP EMS organizes its information into three layers where each layer relates to a single type of data (User's Manual, 1997):

> *Monitoring locations*
> geographical sites where samples are collected for the purposes of monitoring the environment. Each monitoring location is a discrete geographical point defined by its latitude, longitude and elevation
>
> *Samples*
> physical or biological samples or specimens collected at a monitoring location
>
> *Results*
> discreet measurements (a discrete result), summaries of continuous measurements over a time period (a continuous result), or a descriptive comments (User's Manual, 1997).

MELP EMS allows a user to move effectively between all three types of data. The data structure allows for a variety searches between the different data types including searches based on geographic considerations (User's Manual 1997).

All data in EMS is geo-referenced to a collection site, and each site is identified based on its initial purpose (Peppin, N., pers. comm., 1996). A site used initially for compliance monitoring will be listed as a compliance-monitoring site. This purpose information can be used to provide an indication of context for the data. This contextual information, however, is not absolute. Data collected at a pre-existing site for reasons other than those for which the station was established will still be associated with the initial purpose (Peppin, N., pers. comm., 1996). Consequently, if a contaminant spill occurred near a compliance site any post-spill measurements taken at that site would be listed as compliance data by MELP EMS. Thus, a secondary user would face the potential that data might be incorrectly associated with data collection activities that were not compatible with the actual purpose for which the data was collected.

5.1.1 Electronic Data Interchange and Quality Assurance Index

Of particular interest to this work was the implementation of the electronic data interchange (EDI) in MELP EMS. EDI provides laboratories and permit holders with the ability to electronically transfer their data into MELP EMS (LGS, 1995). The system allows for data to be downloaded to the system through a web page, by FTP server or even through e-mail (LGS, 1995). The ease of entry of data into the system is intended to ensure that the system quickly becomes data rich. The scope of the data input limits the possibilities when considering appraising the data. In today's financial climate, it was considered unrealistic to expect EPP to supply sufficient numbers of human appraisers in order to carry out appraisals of all their data. As a result, MELP EMS incorporated an automated system called the quality assurance index (QA index). The QA index is based on an algorithm that examines the quality assurance information that accompanies the data (EMS FAQ, 1997b). Consequently, the QA index indicates a basic level of scientific confidence associated with a specific dataset. The algorithm used in the QA index includes:

- *Basic Result Validation* which converts non-continuous results to a floating-point representation;
- *Unit-Based QA Validations* which compare some reported results with pre-defined or expected limits that are known for specific units (i.e pH can only run from 0 to 14);
- *Method-Based QA Validations* which compare reported results with expected characteristics that are defined in the parameter dictionary;
- *Calculation of inferable parameters* which calculates a result for relationships that are understood (i.e. calculates Total Hardness of water from Ca and Mg values); and
- *Inter-Parameter QA Validations* which in some circumstances can check the validity of a reported value based of relationships between parameter qualities (Peppin, N., pers. comm., 1996).

The outcome of these algorithms is a QA indexing value between A (top value) and F (lowest value).

The QA index is assigned automatically but EMS incorporates a manual override:

> the override would allow the Ministry to manually downgrade or upgrade a set of results for a user specified selection criteria based on local knowledge. This facility would be used, for example, if a lab fails an EDQA audit for a constituent or if an analyzing agency has repeatedly used an incorrect methodology (LGS, 1995).

In order to ensure that the override is only used in appropriate cases another function was included into the system.

> If a new index is manually entered, a comment justifying the reasons for the override is required. If a manual override exists, then the manual index will be displayed with the associated reason (LGS, 1995).

5.2 Data Truncation and Censoring

5.2.1 Digit truncation

Digit truncation is simply the removal of a digit, usually so as to reduce precision in a numeral. It can be done to indicate significant figures (defined later) or to simplify the look of a spreadsheet. Traditionally in digit truncation, the uncertainty of a measurement process is assessed, a limit of quantification (LOQ) is established, and only values in excess of LOQ are reported. In digit truncation, if a value is determined to be 123.456 at an intermediate stage of an analysis and subsequently the limit of quantification is determined to be one decimal place, then the measurement becomes 123.4 in the final report.

Rounding off is a special case of digit truncation, which is widely practised by analytical environmental laboratories. Significant figures are made up by combining certain digits (i.e. individual digits which due to their measurement methodology are known to be correct) with the first uncertain digit in a measurement (i.e. if the number 1.725 had 4 significant digits then the numbers 1.72 would be certain digits and 5 would be the uncertain digit) (Zumdahl, 1986). Significant figures protocols exist for most fields that involve numerical measurements. In the case of rounding-off, truncation is carried out using standardized procedures that ensure that only the significant figures are listed. As an example the rules for rounding-off in chemistry are:

1. In a series of calculations, carry the extra digits to the final result, then round off
2. If the digit to be removed
    a. is less than 5, the preceding digit stays the same. For example 1.33 rounds to 1.3
    b. is greater than 5, the preceding digit is increased by 1. For example, 1.36 is rounded to 1.4.
    c. is equal to 5, the preceding digit is not changed if it is even and is increased by 1 if it is odd. For example, 1.35 rounds to 1.4, but 1.25 rounds to 1.2 (Zumdahl, 1986).

The term "digit truncation" has essentially been supplanted by the term rounding off. The reason for this may be due to the fact that analysts do not wish to confuse the special case of rounding off with non-sensitive digit truncation.

5.2.2 Distribution truncation

When the term truncation is used in the literature, it generally relates to the second type of truncation, called distribution truncation, which applies to a population or distribution of data elements. Such truncation is a common occurrence in environmental analysis where a set of analytical results contains some data that lies below a defined limit of quantification (LOQ). The LOQ, like the limit of detection (LOD), is related to the uncertainty in the measurement expressed as a measurement standard deviation (σ). Typically LOQ is set at 10σ and LOD at 3σ, which corresponds to an uncertainty in the measurement of $\pm$ 30% at the 99% confidence level (this will be discussed in more detail below). As with digit truncation, at some point during the analytical process a value of 123.456 is determined for a measurement and is found to lie below the LOQ. The value for that measurement is then reported as "less than the LOQ". The complete data set has thus been truncated by one measurement, since no numerical value can be assigned. The same is true of measurements which are determined to lie below the LOD and which are reported as "not detected". Distribution truncation is a chronic issue in trace environmental analysis. It is not uncommon for a set of analytical results to contain a majority of reported values "less than the LOQ" or "not detected".

Distribution truncation is a type of data censoring. A set of results is said to be censored when values are reported as unknown (or deliberately ignored) although their existence is known (Kendall, 1982). When a data set contains non-detected observations it is referred to as left-censored. The reference to "left" refers to the left-hand tail of the data distribution where the left is the low end of the distribution and the right is the high end. For the remainder of this work, the specific term censoring will be used in lieu of the general term truncation or the special case of left censoring.

There are two common types of censoring. In type I censoring a measurement is made and the observations which have values below a predetermined limit are censored (Schneider, 1986). In general the predetermined limit is either the LOQ or the LOD. Type II censoring occurs when the number of observations is fixed and only a proportion of observations are reported (El-Shaarawi and Esterby, 1992). This occurs when fixed time samples are taken of a distribution, which might vary with time.

5.3 Analysis of Data Censoring

In order to understand how censored data should be handled it is first important to understand the terms to be used in this discussion: these being LOD and LOQ. While there are numerous definitions for LOD and LOQ, most share similar properties. The definitions that will be accepted for this work were presented by the American Chemical Society (ACS) Committee on Environmental Improvement (ACSCEI) (Keith et al., 1983) and subsequently adopted by the ACS.

5.3.1 Definitions

The ACSCEI (Keith et al., 1983) defined the limit of detection as:

> the lowest concentration level that can be determined to be statistically different from a blank. Let S_t represents the total value measured for a data set and S_b the value for the blank, and σ the standard deviation for these measurements. The analyte signal is then the difference $S_t - S_b$. It can be shown that for normal

> distributions $S_t - S_b > 0$ at the 99% confidence level when that difference $(S_t - S_b) > 3\sigma$. The ACS recommended value of LOD is 3σ (Keith et al., 1983).

The ACSCEI defined the limit of quantification as:

> the level above which quantitative results may be obtained with a specified degree of confidence. Given the conditions described in the LOD definition the value for LOQ $=10\sigma$ is recommended, corresponding to an uncertainty of $\pm$ 30% in the measured value ($10\sigma \pm 3\sigma$) at the 99% confidence level (Keith et al., 1983)

The limit of detection of an analytical procedure is regarded as being the lowest concentration of an analyte that can be distinguished with reasonable confidence from a field blank (here defined as a hypothetical measurement containing zero concentration of an analyte) (Royal Society Analytical Methods Committee, 1987). The question of detection of a given analyte is often one of the most important decisions in low-level analysis (Keith et al., 1983). The question that must be answered is whether the measured value is significantly different from that found for the sample blank (Keith et al., 1983). The Royal Society of Chemistry, Analytical Methods Committee (RSCAMS) has recommended that the limit of detection of an analytical system be defined as the concentration or amount corresponding to a measurement level 3σ units above the value for the zero analyte. The quantity σ is the standard deviation of responses of the field blanks (RSCAMS, 1987). This agrees with the previous ACSCIE decision that signals below 3σ should be reported as "not detected" and the limit of detection should be given in parentheses: ND (LOD=value) (Keith et al., 1983). The ACSCIE also has determined that signals in the "region of less-certain quantitation" (3σ to 10σ) should be reported as numerical values with the limit of detection given in parentheses (Keith et al., 1983).

Data measured at or near the limit of detection have two problems. The uncertainty can approach and even equal the reported value, and confirmation of the species reported is virtually impossible (Keith et al., 1983). The reason for this is that traditional methods for determining detection limits usually rely on the assumptions that observed responses are normally distributed and that the variance of these measurements does not depend on concentration level, at least over a narrow range of "low" concentration of interest

(Clayton, Hines and Elkins, 1987). Whether low concentrations can be detected depends on the medium containing the analytes, the methodologies used for preparing the field samples for measurement, and the analytical techniques for measuring the analytes (Lambert, Peterson and Terpenning, 1991).

A data analyst faced with environmental data containing values below the LOD (nondetects) might assume that all nondetects are zeros, all nondetects are smaller than the smallest numerical measurement ("detect") or, if a detection limit is reported, that all nondetects are below the LOD (Lambert, Peterson and Terpenning, 1991). These assumptions can be incorrect. The confidence interval for a measurement that overlaps zero is a valid statistical outcome in the determination of the LOD (Porter, Ward and Bell, 1988). Zero or negative values are often considered to be outliers, but when working near the LOD, a certain number of analyses by chance alone are expected to be zero or negative (Keith et al., 1983).

LOD is estimated in the response (or signal) domain, but is usually reported in terms of concentration or amount (mass) (RSCAMS, 1987). The relationship between the response and the concentration domains is the calibration (RSCAMS, 1987). Calibration is the checking of physical measurements against acceptable standards, including measurements of time, temperature, mass, volume, electrical units and others (ACSCEI, 1980). It is imperative that no data should be reported beyond the range of the calibration of the methodology (Keith et al., 1983). It must be emphasized, therefore, that the LOD and LOQ are not intrinsic constants of a measurement methodology but depend upon the precision attained by the laboratory while using it on a day-to-day basis.

5.3.2 Literature Recommendations on Handling Data near LOD and LOQ:

Because environmental samples are typically heterogeneous, a large number of measurements ordinarily must be collected and examined to obtain meaningful

compositional data (Keith et al., 1983). Additionally, the number of measurements in environmental data sets is usually small, which makes them difficult to analyze (Gleit, 1985). The number of individual measurements that need to be collected and examined will depend on the data requirements of the plan or model (ACSCEI, 1980). Unfortunately, environmental testing is often done where the expected levels and the standard deviation of the population are not known in advance and where the measurement error cannot be predicated accurately, nor can it be assumed to be negligible (Keith et al., 1983).

The ACSCEI has made a number of recommendations on the intended use of data and results. These recommendations are derived from Keith et al. (1983) and are summarized below. The ACSCEI recommends that the intended use of data should be addressed explicitly in the planning process. Intended results are those that answer a question or provide a basis on which a decision can be made. The most important factor to be considered when determining the level of quality control is the consequence of being wrong. If an analytical result is to be used in a screening program or to adjust a process parameter, an unvalidated analytical method may be sufficient and appropriate. On the other hand, if regulatory compliance is the reason for an analysis, a validated analytical method is usually required. Conclusions as to whether a signal is detected, whether a positive signal is confirmed to be an analyte, how much uncertainty is contributed by the sampling, and the risk of systematic error are best made by those involved in the study and should be included in any report. Reports should contain sufficient data and information so that users of the conclusions can understand the interpretations without having to make their own interpretations from raw data. Analytical chemists must always emphasize to the public that the single most important characteristic obtained from one or more analytical measurements is an adequate statement of the uncertainty interval (Keith et al., 1983).

The results of the ACSCEI study lend strong support to the argument generally supported in ASTM standards that data should not be routinely censored by laboratories (Gilliom,

Hirsch and Gilroy, 1984). These standards suggest that uncensored data should always be retained in permanent records available to data users even if policy makers of a laboratory decide that some censoring or other form of qualification is necessary before public release of data (Gilliom, Hirsch and Gilroy, 1984). Measurement data should not be discarded unless the lack of statistical control in the measurement process is clearly demonstrated (Gilliom, Hirsch and Gilroy, 1984). Notably, Clark and Whitfield (1994) have also suggested that while environmental laboratories and scientists should conform to their national or international standard methodologies with regard to the use of significant figures and left-censoring, that results reported to electronic data bases should come in pairs of numbers. The first value should be the 'official result' identical to that which would appear on paper, and the second value would be the raw, unmodified result (Clark and Whitfield, 1994). In conclusion, the literature contains a number of strongly worded recommendations with regard to censored data.

Rao and Ku (1991):

- Censored and/or truncated data sets tend to complicate statistical analysis.

Porter, Bell and Ward (1988):

- Reporting a measurement and an estimate of observation error provides more information than does reporting a "non-detect" or simply reporting the measurement value. A significant improvement in the information content of near-detection-limit data would occur if one simply reported results of all analyses plus an estimation of observation error.

Royal Society Analytical Methods Committee (1987):

- Any censoring of measurements falling below the detection limit (or even below zero) may result in incorrect estimates of both precision and bias at low analyte levels. Therefore, the actual concentration measurements observed should be recorded and used, even when they fall below the detection limit or zero

Gilliom, Hirsh and Gilroy (1984):

- For all classes of data evaluated, trends were most effectively detected in uncensored data as compared to censored data even when the data censored were highly unreliable. Censoring data at any concentration level may eliminate valuable information. The more reliable the data censored, the greater the information lost and the more detrimental the effects of censoring.

5.3.3 Statistical Tools to Work with Censored Data:

An important consideration in any discussion of data censoring is how to arrive at summary statistics for a collection of measurements, which includes values below the detection and quantification limits. When attempts are being made to estimate simple statistical information like mean, standard deviation, median, or moment, a number of different methodologies can be used. These methods can be broken down into four different types, which have drastically different effects on the outcome of the summary statistics. These four approaches are: a) to ignore the censored values, b) to substitute the censored values with set values, c) to use the characteristics of an assumed distribution to estimate summary statistics (called distributional methods), and d) to replace the censored values with values based on a statistical distribution and calculate the summary statistics (called robust methods). The following will investigate how each of these approaches affects the summary statistics.

The most commonly used method of dealing with censored information is to discard the censored observations prior to calculating the summary statistics (Gilliom and Helsel, 1986). Thus, if ten measurements were made in an experiment and three had values below the LOD then the summary statistics would be based only on the seven measurements that had numerical values. This methodology while benefiting from simplicity, has little statistical or scientific support. "Non-detects" or "less-thans" are data and cannot be ignored in any serious scientific endeavor. Discarding censored observations will always result in both higher bias and higher root mean square estimate than any method that incorporates the detection limit (Gilliom and Helsel, 1986).

Of the methodologies that do not discard "non-detects" simple substitution methods, such as replacing all "less thans" with zero or the detection limit, are most commonly used (Helsel and Cohn, 1988). The three most common values used to replace "less thans" are a) zero, b) half the LOD, and c) LOD. Of the three the third is the most conservative (the one which returns the highest value) estimator for the mean while the first is the least

conservative (Gleit, 1985). Though replacing the "non-detects" with the LOD will bias a data set, it has often been suggested by the USEPA as their accepted procedure merely because of its conservatism (Gleit, 1985). Much like methodologies that ignore censored values, methodologies that simply substitute them with a single value are found to produce biased and highly variable estimates (Gilliom and Helsel, 1986).

An additional flaw in the EPA suggested method of using the LOD is that statistical outcomes will vary with an increase or decrease in methodology detection limits (MDL). This has serious repercussions while considering the use of historical information for current and future decision-making because mean values can be made to appear to decrease based entirely on improvements in MDL.

Expressions for the expected mean and standard deviation in simple substitution methods show (a) the impossibility of obtaining unbiased estimates when a single value is used to replace the censored observations, and (b) that the direction and magnitude of bias depends upon the expected proportion of censored values and the distributional characteristics of the data (El-Shaarawi and Esterby, 1992). Further, the problem of bias is not solved by taking a larger number of measurements since the bias is independent of number of measurements, that is, it remains constant no matter how large a sample size is collected (El-Shaarawi and Esterby, 1992).

While methods are available that appropriately incorporate data below the limit used for the purpose of reporting estimation, hypothesis testing, and regression (Helsel, 1990), most methods to replace censored observations with values are based on distributional assumptions (El-Shaarawi and Esterby, 1992). Distributional methods use the characteristics of an assumed distribution to estimate summary statistics (Helsel, 1990). Rather than replace the censored values with data in order to calculate the summary statistics, the censored values are assumed to follow a distribution such as the lognormal and based on that assumption, distribution estimates of summary statistics are computed (Helsel, 1990). In the lognormal distributional method, for example, it is assumed that

measured environmental data represents repeated samples from a lognormal probability distribution where only values above the LOD are known (Travis and Land, 1990). These values are often enough to define the right hand tail of the lognormal distribution, from which it is then theoretically possible to reconstruct the entire distribution and thus obtain knowledge of the mean and standard deviation (Travis and Land, 1990).

Environmental data sets are usually small which makes them difficult to deal with statistically (Gleit, 1985). In routinely collected sets of measurements, when either the number of measurements is small or the proportion censored is large, it may not be possible to adequately check distributional assumptions (El-Shaarawi and Esterby, 1992). When the data do not match the observed distribution, this method may produce biased and imprecise estimates (Helsel, 1990). In other cases, the point of the data collection effort might be to determine the distribution of the data. In that case the data might be found subsequently to contain censored observations. In those cases, distributional methods will not be appropriate (Helsel and Gilliom, 1986). Instead it may be necessary to collect a larger set of measurements designed for the purpose of establishing appropriate distributional assumptions, and then using these assumptions in the analysis of the smaller sets (El-Shaarawi and Esterby, 1992).

Distributional methods also have biases as a result of attempts to transform the data to meet the statistical requires of the distributions. There is transformation bias inherent in computing estimates of the mean and standard deviation for any transformation (Helsel, 1990). So even when the data fits the distribution, estimates of the mean and standard deviation computed in transformed units will be biased when they are retransformed (Helsel, 1990).

Robust methods combine observed data above the LOQ with extrapolated below-limit values, assuming a distributional shape, in order to compute estimates of summary statistics (Helsel, 1990). In this case a distributional fit to the data above the reporting limit is used only to extrapolate a collection of values below the LOQ. These extrapolated

values are not considered estimates for specific measurements but are used collectively only to estimate summary statistics (Helsel, 1990). Robust methods have produced consistently small errors for estimating the mean, standard deviation, median and interquartile range in simulation studies (Gilliom and Helsel, 1986). These methods have substantial positive bias when estimating sample moments for small or moderate sized data sets even from distributions that truly match their parent distributions (Helsel and Cohn, 1988). When the parent distribution is not known, the desirable, theoretical sampling properties of likelihood-based procedures do not necessarily apply (Helsel and Cohn, 1988).

In environmental samples where the distribution of the data is uncertain and the majority of observations lie at or near the LOD, data censoring will result in bias with regard to summary statistic estimates. As Helsel (1990) puts it: "The deletion of censored data or fabrication of values for less-thans leads to undesirable and unnecessary errors". While this conclusion is firm, some qualifications do apply. In any population where the majority of values are significantly above the LOD, censored data do not present a serious interpretation problem (Gilliom and Helsel, 1986). In addition, some methodologies exist that when used correctly allow censored values to contain nearly as much information for estimating population moments and quantriles as would the same observations had the detection limit been below them (Helsel and Cohn, 1988). Finally, robust methods have produced consistently small errors for estimating the mean, standard deviation, median and interquartile range in simulation studies (Gilliom and Helsel, 1986). It must be understood, however, that in order for these methodologies to provide reliable estimates, several considerations must be met. These include:

- large number of measurements,
- large proportion of values above LOQ, and
- some hypothesis which makes assumptions about the distribution of the measurements.

These conditions are seldom satisfied in contemporary environmental data sets.

5.4 Examination of the Practical Aspects of Censoring

While the literature appears to strongly discourage data censoring, real-world practice is somewhat different. In examining data censoring for MELP it became apparent that little documentation existed on actual laboratory practice with regards to data censoring. As a result, we undertook a survey of environmental laboratories in order to review standards and practices in the area of data handling. Other members of our research group carried out the survey of laboratory practice in parallel to the statistical work. The full details of the survey will not, therefore, be presented here. Instead, an overview of the work and the critical conclusions will appear here.

5.4.1 Survey design

A survey list of accredited laboratories was generated from directories compiled by the Canadian Association for Environmental Analytical Laboratories, Washington State Department of Ecology, and Oregon State Department of Human Resources - Health Division and the International Council for the Exploration of the Sea. The total number of laboratories on these lists is >1000. We reasoned that larger and more diverse facilities would be more likely to have considered the questions surrounding censoring and selected approximately 30 of the largest of them in each group to receive our survey. The number of laboratories initially surveyed in Canada, USA, and internationally was 33, 29, and 32, respectively. Surveys were forwarded either by fax or in the case of many of the international sites by airmail. Within two weeks of sending off these surveys we had received a response rate of almost 40%. The excellent return rate resulted in an expansion of the study. Consequently, we decided to select another roughly 60 laboratories drawn randomly. International laboratories were excluded from this batch since the time available precluded the use of the postal system. A "module" in MS Excel was constructed to generate a random list of laboratories from our directory lists. A random laboratory list was generated for both Canada and the USA, essentially consisting of 36

labs in Canada and 30 labs in the USA. Inevitably, a proportion of the random labs was already surveyed and only previously unsurveyed labs were contacted in this second batch. A grand total of 150 laboratories were surveyed with 57 (38%) responses being received. The summary results appear below with more detailed results available in Appendix 1.

5.4.2 Survey Outcome

The majority of laboratories surveyed reported supplying data that is either truncated and/or rounded. The censoring is largely based either on significant figures or associated with the analytical methodology. It appears that many of the labs that may not truncate their data do at some point round it. Rounding itself is computed chiefly by using one form of algorithm or another. Almost one in four responses that stated that they report rounded values apparently actually truncate their data. Based on conflicting responses to examples posed on the survey it would appear that the terms "truncation" and "rounding" are often confused.

Although a majority of labs are not asked to deliver "raw" data plus uncertainty, approximately one quarter have experienced this request. Clients making this request have attributed this need to statistical requirements or simply because it was the routine method of reporting data. Unfortunately, a moderate percentage of the laboratories that responded positively did not state why the data was needed in this form. In most cases the analytical laboratory obliged the client and supplied them with "raw" data plus uncertainty, although one third said they would not. For those that would not, the main issue was the dissemination of data uncertainty.

Reporting of non-detected values was broken down into three groups, those that use "<", values, or text. Almost 60% use some version of "less than" (e.g., <LOD, <LOQ etc.), whilst 25% utilize actual values and the remainder use some form of text (e.g., "not detected", "nil", etc.).

Essentially, the same number of laboratories reported values that were "detected but below the limit of quantification" as did not. Those that did report these values were broken down as described above. In the case of reporting values that were detected but below the limit of quantification, 60% utilize actual values, whereas almost 25% use "<", and the remainder utilize text. The main issue for those that did not report these values was one of high data uncertainty and overall data reliability. In addition, several of the labs were concerned about the increase in workload associated with reporting values below the quantification limit.

The final question in the study asked:

> From an environmental research position, some statistical tools would be better served if untruncated data were used. This would involve your lab reporting "raw" data and data below the quantification limit. What do you think of this as a practice for an environmental laboratory?

We divided the textual responses into three sections: responses that "generally support" the practice of reporting "raw" data and data below the quantification limit, "neutral" opinion or responses which weighed both pro and con positions, and "generally opposed" to the practice of reporting raw data and data below the quantification limit. The overall results were as follows:

Total number of responses to the question: 57	
generally supportive	14/57 (25%)
neutral	20/57 (35%)
generally opposed	21/57 (37%)
response omitted	2/57 (3%)

Reporting "raw" data and data below the quantification limit, possibly to support statistical analysis, as a practice for environmental laboratories was a contentious issue. The most outspoken were those generally opposed to this practice, even though their numbers were similar to what we called the "neutral" fraction. This latter group includes those labs that presented both pro and con positions. A slightly smaller group generally supported the concept. The key in weighing out the responses is the large number of "neutral" laboratories. A general concern from this group is how clients will interpret this data, for example, they fear that clients might assume significance where it does not exist.

It appears that the majority of the "neutral" group would supply this data if requested, although some may have stipulations tied to with the release of this data.

5.5 Censoring and Secondary Uses of Data

Of particular interest for this research are the types of secondary data uses that are affected by censoring. Once censoring has been carried out it is impossible to recover the original data. Statistical techniques can be used to try and model censored data, but no model can completely replace censored data once it has been discarded. The question that remains to be answered, however, is what secondary uses are most affected by the censoring of data. The answer to this question is not a simple one because different uses have different requirements. This section will examine the data requirements of different activities to determine if the resulting data will be affected by censoring.

5.5.1 Observation

The easiest groups of activities to consider are observations. As discussed previously, observations are carried out without a specific hypothesis being tested. In general, censoring has little effect on data from surveys. The reason for this is that while observation data are not intended to be used out of the context of the observation, the uncertainties imposed by a lack of structure greatly outweigh any additional uncertainty placed by censoring the data.

5.5.2 Monitoring

Of greater interest to this report is how censoring of data affects the four types of monitoring discussed previously: scientific, compliance, research, and trend.

Scientific monitoring is carried out when both processes and their interrelationships are known or are at least partially understood. The outcome of scientific monitoring is data that have well defined spatial and temporal characteristics within a system that is understood sufficiently well to test for environmental variances. Scientific monitoring is designed to produce data with a broad range of interpretability. As a result, scientific monitoring data are extremely sensitive to censoring. In most cases, the system is being examined for small effects, which can be masked through improper data manipulation. As Gilliom, Hirsch and Gilroy (1984) put it

> the detrimental effects of censoring increase with the increasing reliability of data and [the] adverse effects of censoring increase with increasing data reliability because more reliable data contain more information than less reliable data and, thus, more information is lost when data are censored.

The majority of the data in EMS are compliance monitoring data. The direct compliance uses of compliance or regulatory monitoring can use censored data. This is not necessarily the case for secondary uses of the data. Compliance monitoring is generally carried out in the context of a well-understood process and, due to legal requirements, under relatively strict quality assurance/quality control regimes. Consequently, the resulting data are generally of high quality and have strong potential for additional uses such as baseline research. Censoring the data limits the additional uses to which the data might be used, and thus, reduces the value of the data. When considering data in a value-added sense, censoring has a major effect by limiting the subsequent uses for which the data might be appropriate.

Both research and trend monitoring are extremely sensitive to censoring. They both involve systems where systematic or environmental variability is not completely understood and the experimental design must accommodate this imprecision. The aim of this type of monitoring program is to develop process knowledge so that more precise monitoring and research programs can be carried out. The aim of this type of monitoring is often to attempt to distinguish a signal from the constant background of environmental and sample variability. Since the systems being monitored are, by definition, insufficiently understood, any unnecessary manipulation of the data must be avoided. Censoring data of

this sort short circuits the entire purpose of this type of monitoring. It makes trends harder to distinguish from natural variability and, as a result, will diminish the power and outcomes of this type of monitoring.

5.5.3 Research

There is little doubt that censoring has a negative impact on the effectiveness of data collection programs for research purposes. Research, by its nature, is carried out to develop process knowledge. It involves the collection of data about poorly or insufficiently understood systems, within a tightly defined data collection system. As such it requires data that are understood as completely as possible. Any manipulations associated with censoring will lessen the ability of research to carry out its fundamental purpose.

5.6 Implications of Data Truncation for EMS

EMS, as designed, can effectively incorporate uncensored data. The system architecture incorporates the fields needed to supply raw data and their associated uncertainty information. EMS employs a number of validation and edit routines to ensure that data entered into EMS are reliable (MELP, 1997). Amongst these routines is an automatic data grader which is based on an algorithm developed to provide an indication of the basic level of confidence associated with a particular data set (MELP, 1997). Since this rating is based on documentation and QA information that accompanies a data set, it can give a strong indication of the quality of the data

Since EMS was designed primarily for compliance monitoring data, it does not explicitly store context information. The EMS data structure has been designed to accommodate either censored data or data with uncertainty. Unfortunately, it does not distinguish censored from uncensored data. EMS is designed to include both data and uncertainties;

it also can store all the important quality assurance (QA) information that might accompany a data set. However, these QA fields are not mandatory and therefore are not well populated (Clark, M., pers. comm.). In addition, users can download the data without the accompanying QA information. While this should not directly affect primary users, (scientists using their own data) it could limit what a secondary user will be able to say about data sets derived from the system. Since full context information is not available, secondary users are often left to infer whether or not data are appropriate or reliable for their subsequent uses. Secondary data uses are thus not fully supported.

In conclusion, as now constituted EMS could support secondary use of its stored data but due to the lack of critical information regarding the type of data being accessed users would have to be wary. The need exists for tighter control on the data being input into EMS to ensure that the appropriate fields are fully populated. When truncated data are input into the system that fact should be "flagged" for secondary users.

5.7 Case Study Outcomes

This case study demonstrated that the implications of censoring are not limited to the single system EMS, but must be considered within the model being considered in this thesis. No matter what tool is used to recover data, the quality of information stored and the decisions derived from that information are only as reliable as the data that are input into the system. The weakest-link principal states that the strength of a chain is only as strong as the weakest link. Environmental data have uncertainties and decisions based on those data must respect these inherent uncertainties. There is, however, a qualitative difference between inherent uncertainty due to the measurement process and avoidable systematic uncertainties created as a result of data storage and manipulation. The effective secondary use of data requires that the data be complete and unbiased. Data censoring and data usefulness are mathematical and statistical considerations, but they also must be recognized as issues that affect decision-making and secondary uses.

Chapter 6 A New Approach to the Management of Environmental Information

6.0 Introduction

The three previous chapters examined a sequence of interconnected case studies. Each examined a different critical aspect of the management of environmental data. This chapter will use the insight gained from these case studies in order to formulate an improved model for the storage of environmental information. This chapter will begin with a gap-analysis and assessment of idealized systems based on the models developed in the case studies. The gap analysis will examine how effectively these models facilitate primary and secondary access to environmental information. This examination will expose the strengths and weaknesses of each individual model in order to form the basis of a new archetype. Derived from the gap analysis will be a set of recommendations for environmental information system developers and responsibilities for the various individuals involved in the process of using scientific data for environmental decision-making. Based on these recommendations and responsibilities, this chapter will conclude with a conceptual model that can serve as a template for future systems. The model will incorporate two new organizational features called infosets and metasets that capture the need for emphasis on contextual and reliability information to ensure the intercomparability of environmental datasets.

Much of the emphasis in the case studies has been on developing tools and techniques to improve computerized access to environmental information. It must be made clear however, that the model presented in this work is independent of computer platforms or even computers themselves. The development of increasingly powerful computer systems has simply highlighted the limitations of pre-existing models for the management of environmental data.

The conceptual model will define the relationship between differing types of data and information and the needs, limitations and data requirements of complex multi-disciplinary problems. A critical advance in this new model will be its recognition of the

fundamental split in the environmental information management field between primary and secondary users. These two groups have different requirements for environmental information systems and only through satisfying the needs and desires of both groups can an effective system be established.

Primary users are data producers. They engage in the various activities from which the archival records originate. "Activity" is used in this chapter to indicate any effort be it research, surveys, monitoring and/or modeling intended to improve understanding about environmental variables. Primary users can vary from individuals to institutions. In some cases, many different groups can be involved in the production of a dataset and the absolute primary user becomes difficult to discern. Consider mill effluent being tested for regulatory purposes. The mill personnel collect the samples and send them to a contract laboratory. The laboratory analyzes the samples and sends the results to a compliance database. The compliance officer evaluates the data to determine if they meet regulatory requirements and publishes the results. All three groups are primary users of the data. Each group has added to the creation of the dataset and each has an interest in ensuring that the data are not misused.

Secondary users are data consumers. They make use of the data records in the information systems for a variety of purposes including decision-making. Secondary users seldom have complete knowledge of the information they have accessed and consequently rely on the information systems to provide them with reliable and accurate data. It is not uncommon for an individual to be both a primary and secondary user. A researcher might one day enter new data into the archive and the next use data from the archive to plan or carry out new activities.

Traditional environmental information systems have been designed to facilitate access to individual measurements by secondary users, while depending on primary users to provide those measurements. An important insight derived from this research is the recognition that providing secondary users direct, unrestricted access to individual measurements and data records does a disservice to all users. Datasets are generally

created after detailed planning. The measurements that make up these datasets are carried out under specified, controlled conditions, usually to serve a particular purpose. Allowing users to access individual measurements without this added information will fail both the primary and the secondary users. The individual measurements that make up these datasets, even accompanied by their complete contextual pedigrees, lose their relevance when viewed in isolation. Consequently, a critical role of any new system must be to ensure that users are provided with complete datasets and all their critical contextual information.

6.1 Analysis of Pre-existing systems

Each case study presented in this work included an evaluation of the individual system. The following section will consider the models used to organize environmental information in each case study. The aim of this task will be to identify critical advances and remaining requirements. The effectiveness of each in serving the needs of primary and secondary users will be assessed.

6.1.1 Archival systems

Consider an ideal archival database based on the model presented in EMS (Chapter 5). Such a system would store both raw measurements and all accompanying QA information derived from all types of activities. Software tools would ensure that all mandatory data fields were entered. The data in the system would be uncensored. Detailed information about the uncertainty of the measurements would be stored. The archive would incorporate a process to assign data reliability indicators (a QA Index) based on the contents of the QA fields. All measurements in the archive would be geo-referenced allowing searching based on location. Figure 6.1 presents the relationship between data and such an archive. Could such a system adequately serve as the basis for a new information model for environmental information systems?

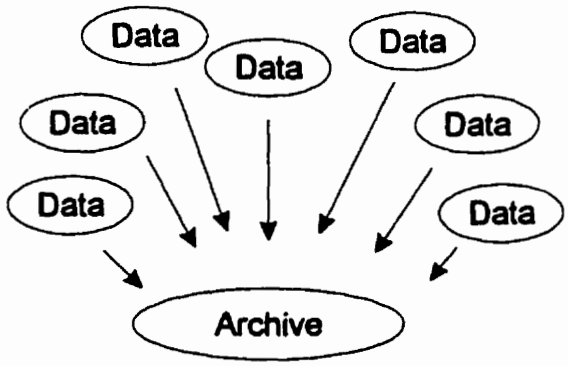


Figure 6.1 Archives and Data

No. An ideal archive, as described above, presents a number of significant advances over comparable systems currently in use, but would have fundamental flaws for both primary and secondary users. Further changes are needed in the area contextual information and the grouping of measurements. Archival systems, based on the model used in EMS, store individual measurements and cannot incorporate methodologies to evaluate the context of data. Primary users can enter their results but cannot indicate why the measurements were carried out or how individual measurements were related. Secondary users, meanwhile, would have free access to individual measurements and no opportunity to examine critical contextual information.

6.1.2 Metadata Systems

Consider an ideal metadata system based on the model presented in the CODIS case study (Chapter 3). Such a system would present a number of significant advances over current systems. The metadata system would store datasets assembled by primary users instead of raw data. The use of datasets would ensure that primary measurements were linked, which would discourage individual measurements from being used in isolation. A parallel data structure between disciplines would provide for discipline-specific data input and measurement appraisal processes. Input of critical fields would be through controlled lists that could not be manipulated by individual cataloguers. All datasets

would be appraised for reliability using guideline-driven, dichotomous decision trees. The system would function independently and would not be directly linked to the archives. It would include a reporting system to ensure that users could identify the permanent storage location of the original measurements that made up the datasets. Search tools would provide for both text and geographical searching of the database. The structure of such a system is presented in Figure 6.2. Could such a system adequately serve as the basis for a new information model for environmental information systems?

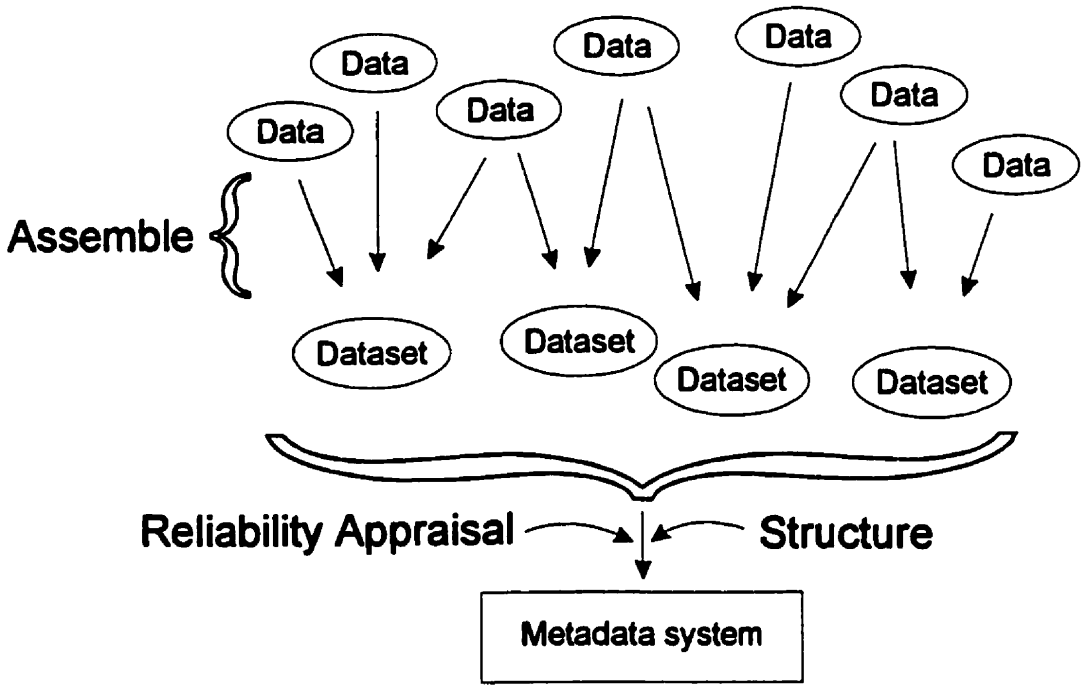


Figure 6.2 The Structure of Metadata Systems

No. An ideal metadata system would not meet the needs of both the primary and secondary users in the area of data availability and context. CODIS-style metadata systems are not linked to archival data. Consequently, users cannot directly access data. While users are made aware of potentially useful data, they are not greatly assisted in obtaining the results. In addition, such systems do not contain a function to relay the context of a dataset to secondary users. Primary users can link their datasets together through DS_ID numbers, bibliographic references, projects and/or platforms but cannot

document their reasons for assembling individual datasets or indicate whether their expectations were achieved. In addition, the requirement that datasets have the same measurement ratings may require a primary user to separate measurements collected as part of an individual activity into different datasets. The lack of links would ensure that secondary users could not easily identify datasets linked by a single activity.

6.1.3 NCIS style Combined Archival, Inventory and Directory Systems

The NCIS model of an interconnected archive, inventory and directory system (Chapter 4) presents a model that addresses many of the major needs noted above. Archives would store measurements and records. The inventories, consisting of detailed metadata, would provide users with effective search tools. The directory would provide an additional hierarchy of search terms to search regional inventories and archives. Most importantly, such systems would incorporate a methodology to evaluate and report the context of experimental events. The structure of the NCIS system is presented in Figure 6.3. Could such a system serve as the basis for a new general model for environmental information systems?

No. The NCIS structure requires modification in two critical areas, the division of surveys from experiments and the creation of a one-to-one relationship between individual measurements and experiments. With respect to contextual assessment, surveys are not fundamentally different from experiments. Secondary users might need to know the context of the survey and its pedigree in order to evaluate whether the projected secondary data use would be appropriate. Yet the NCIS model assumes the contrary and will deny full information to the secondary users. The other drawback in the NCIS model is that it limits the association of data to one experiment. This many-to-one relationship implies that even the primary user is restricted in the use of his/her own data. The primary user could not use a single measurement in two different experimental events and a secondary user making use of a dataset for subsequent work could not associate this new activity with the original data

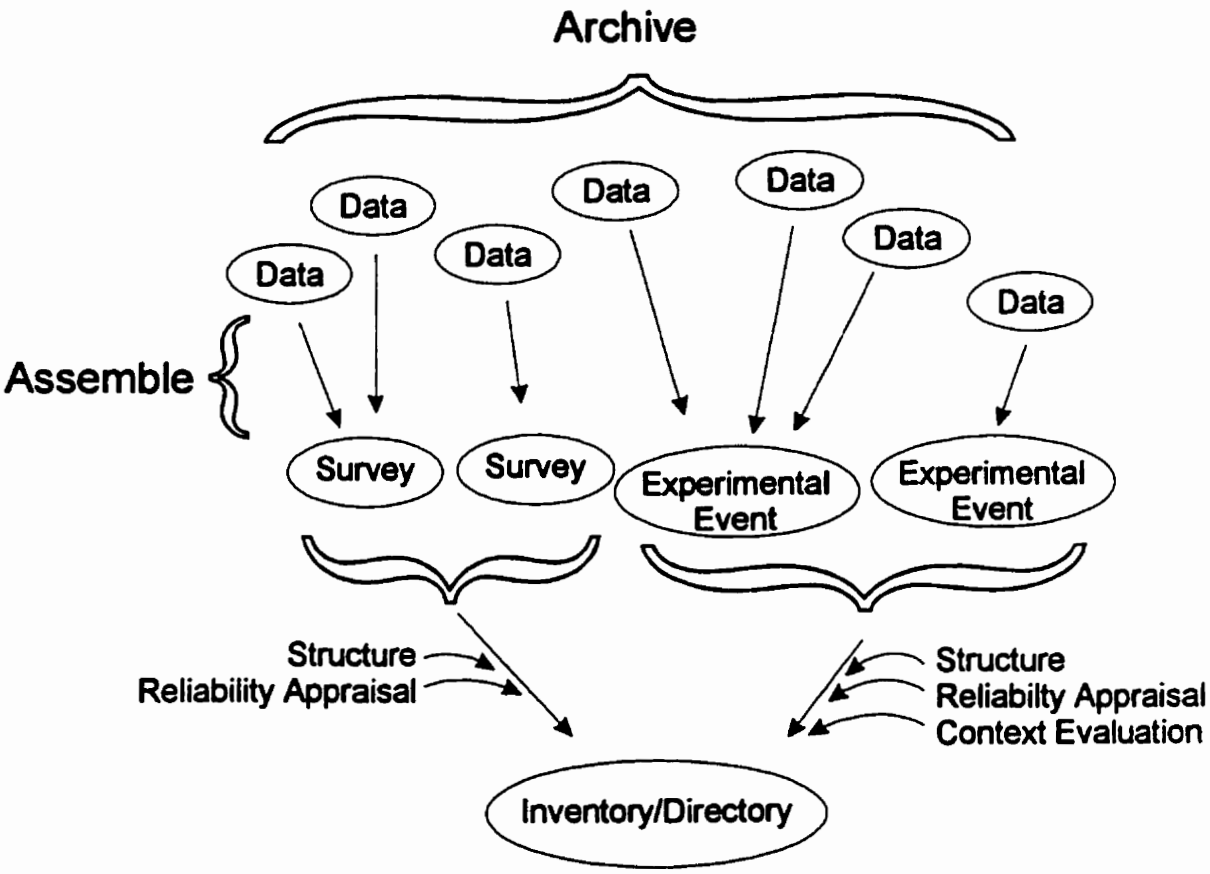


Figure 6.3 NCIS style System Architecture

Each of the models discussed above has limitations. It may be possible however, to combine and enhance them in order to develop improved models. Consider the following adaptations:

6.1.4 Metadata Systems linked to Archives

Linking metadata systems to archives (as in Figure 6.4) would eliminate some of the limitations of each individual model. Software modifications could ensure that measurements were linked through the use of datasets and reports would include the measurements and all their critical QA information. Linked metadata and archival systems would provide a valuable tool to facilitate data access.

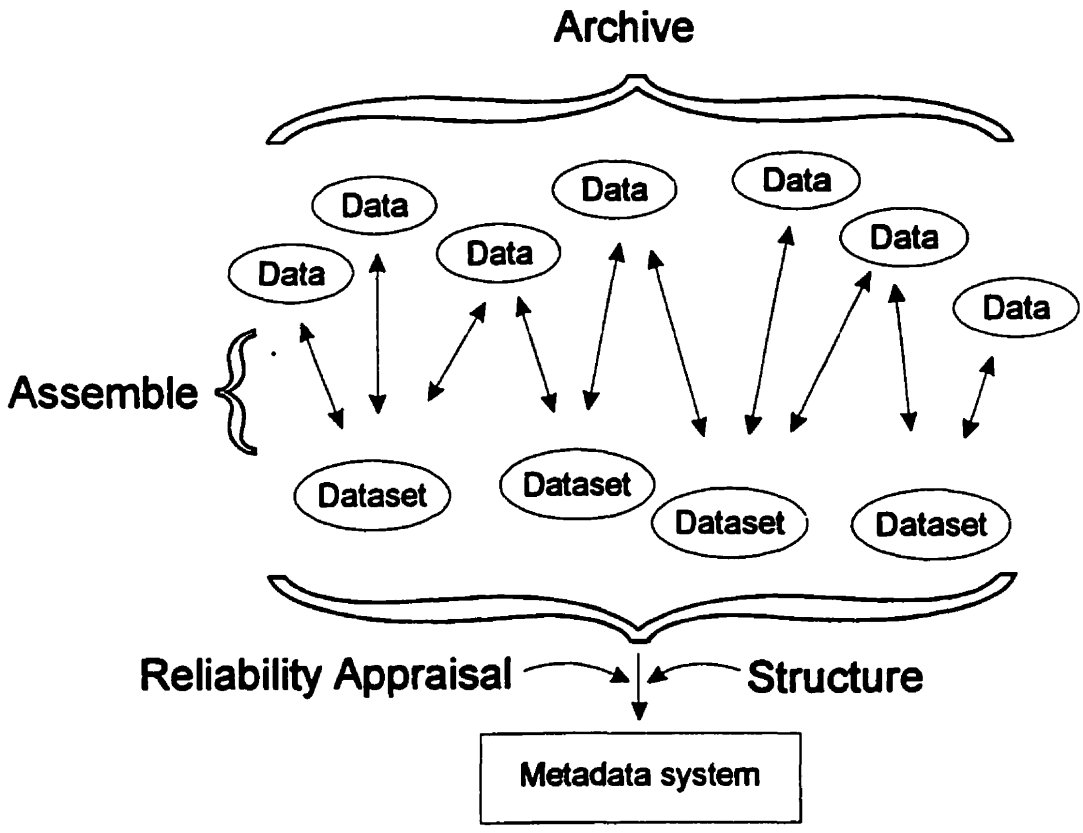


Figure 6.4 Linked Metadata and Archival Systems

Two features, however would still be lacking: 1) neither sub-system would incorporate a tool to relate the context of the measurements or metadata from primary to the secondary users; and 2) secondary users would still be provided access to individual measurements from the archives without their associated metadata.

6.1.5 Improving on the NCIS Model

An enhanced NCIS system presents another intriguing possibility. If one were to redefine "surveys" to become a new "experiment type" then all surveys would undergo contextual appraisal. This new model would meet the majority of the needs presented in this work. Primary users would be provided with a tool to effectively relate the context of their results to subsequent users.

Secondary users, however, would still potentially have access to individual measurements. In addition, this new model would emphasize the weakness of the directory/inventory split in NCIS-type systems. In the NCIS model, the inventory level holds information about regional holdings, while the directory level holds information about the projects under which data collections were made. The information stored in the directory is directly relevant to the inventory user. The division of these two systems has the potential to stratify useful information about the data holdings.

In the next two sections the recommendations for a general system and responsibilities of the developers and users are documented. We then return to propose a general model incorporating the insights developed in the analysis above and these recommendations and responsibilities that follow.

6.2 Recommendations

The following section will summarize the work completed to date through the presentation of specific recommendations. The recommendations identify the critical insights derived from the case studies and provide the foundation for effective environmental information systems. The recommendations are broken into groups of comparable impact with the earlier recommendations being substantially more important. If these early recommendations are ignored, then subsequent issues are irrelevant. Each group of recommendations will be followed by a discussion that clarifies implementation issues.

Recommendation 1: Data in archives must be stored in as complete a manner as possible and must not be censored.

Recommendation 2: All data stored in environmental information systems must be appraised for reliability using well-documented, well-supported protocols

The overwhelming reason for creating environmental information systems is to enhance the availability and usability of the measurements and experimental results stored therein. This most critical insight derived from this research is the recognition that useful environmental information systems must store complete, uncensored data that have been appraised for reliability. Including censored data in archives directly limits the utility for some secondary users. If secondary users cannot rely on the data in the archives to be complete and correct then they can not and will not make use of that data. Environmental data have uncertainties and decisions based on those data must respect those inherent uncertainties. There is, however, a qualitative difference between inherent uncertainty due to the measurement process and avoidable systematic uncertainties created as a result of data storage and manipulation techniques. In cases where the inclusion of censored data is unavoidable, a methodology must be in place to ensure that secondary users are made aware of the reduced reliability of the data.

The practical consequence of sidestepping the appraisal process as allowed in NCIS, compromises the future utility of the data. Incomplete documentation, whether of the data or of the appraisal, directly hampers future use.

These two recommendations are not difficult to implement. Through the implementation of a system comparable to the EMS EDI and QA index, it becomes relatively easy to ensure that only complete data that has been appraised for reliability are included in the system.

Appraisal provides critical information about the strengths and limitations of the included data, which provides critical insight for secondary users. Any approach to rating reliability of datasets must be both functional and repeatable. The approach used in the CODIS case study was found to be both. It relied on decision trees consisting of a series of structured questions designed to yield a unique reliability rating for each aspect of the measurement or experimental process. It included binary nodes with yes/no logic where a no resulted in the generation of a value. The dichotomous nature of the CODIS

decision trees increased the simplicity of the system. The critical advances of these decision trees were:

1. The inclusion of a strictly dichotomous or binary structure,
2. The inclusion of guidelines which gave descriptive guidance to appraisers while providing for changes in methodologies with improvements in technologies or techniques
3. Peer-review and workshop testing of the decision trees to ensure a "consensus". This was expected to limit the subjectivity inherent in the ADCAP/WESCAP charts.

The creation of the decision trees raised and addressed a number of issues. As they stand, the decision trees are weighted towards disagreement and appraiser agreement has an upper limit (i.e. appraisals can only get so good due to inherent uncertainties). The decision trees themselves are comparable to peer review and produce reliable outcomes that improve in an environment of consensus, continued training and effective supervision. Thus, using decision trees similar to those developed for CODIS, it should be possible to design a system to allow appraisers, under the supervision of an effective supervisor, to reliably and efficiently appraise scientific data.

Recommendation 3: All datasets should be appraised for context using flexible appraisal processes that consider all aspects of the activity not simply the measurement process.

Recommendation 4: Environmental Information systems must strive to maintain the link between data and their associated reliability, contextual and QA/QC information.

Recommendation 5: Datasets should not be reported without their associated contextual information.

Recommendation 6: Reliability ratings, guidelines and protocols should be accompanied by explicit date information.

The protocols developed for NCIS provide a potential tool to evaluate and document contextual information to be included with scientific data in information systems. The experience in the NCIS case study demonstrated that it is possible to develop and implement such a process for use in a real system. As discussed in the NCIS case study, defining some activities as non-experiments degrades the inventory system by providing an uneven level of contextual information within the system. Hence, even events that might be considered as simple surveys should be subjected to contextual evaluation.

The system to identify experiments must be standardized and repeatable. The initial format designed for the appraisal of experimental events in the NCIS case study provided unnecessarily broad appraiser judgement decisions. In order to avoid this problem a more direct approach is recommended. Some DFO regions have dealt with this by incorporating the principle investigator into the appraisal process. In the Newfoundland district this was found to decrease conflicts and increase confidence of the investigators.

In environmental information systems, data architecture must be respected and "mandatory" fields must be filled when data are entered. The need exists for a methodology to ensure that critical contextual information be entered and stored alongside the data. An automated system like the EDI and QA index rating presented in the EMS case study, if fully applied, would be able to carry out this task with a minimum of added effort. Any system that allows data to become dissociated with critical contextual information severely shortchanges secondary users and reduces overall usefulness of the data. This association should be reflected in reporting formats. Dataset reports should always include contextual information.

Since appraisal tools like the decision trees and guidelines are intended to reflect best current practice, it is necessary to include date information with all appraisals, decision tees and guidelines. Through this added documentation users are provided with critical details when accessing historical information.

Recommendation 7: The structuring task for inventory systems should incorporate formalized processes to assign controlled language with comprehensive conventions and definitions.

Recommendation 8: A formalized process to identify errors and determine error rates should accompany the structuring and data input task for inventories.

Recommendation 9: Investigators should be encouraged to become involved in all aspects of the dataset creation process.

The design of an effective inventory-level system is critical to the design of environmental information systems as a whole. Designing such a system has been the goal of the majority of the research described in this thesis. The correct functioning of the inventory requires the creation of datasets and associated metadata. Both the CODIS and NCIS case studies examined that process in detail.

Assembling and defining why particular data can be handled together is the process of creating datasets. Structuring tools assist this process. The aim of these tools is to translate the critical elements of the datasets using a controlled vocabulary. Some controlled vocabulary qualifiers include source, location (in time and space) and data specific elements like species, collection method etc. Every practicing scientist at one time or another has assembled data into a dataset, and every database designer has used controlled language to simplify input. What has not been considered previously is the process of assigning metadata to datasets and the minimum required metadata assigned to the dataset.

The importance of using formalized processes and conventions to structure data has been repeatedly demonstrated and the CODIS case study established the need for a structured approach. The use of look-up lists and the provision of specific conventions, where uncertainty exists, ensure consistency in the creation of datasets. The CODIS approach to structuring data entails placing the emphasis of effort at the start of the cataloging

process in order to reap the rewards of decreased input time and increased quality of data. The creation and use of look-up lists requires that the creation of an overall data structure be the initial task of any cataloguing activity. This emphasis on planning is imperative in the design of effective inventory systems. Through the effective planning and definitional phase, it is possible to limit errors and eliminate the use of impossible to define terms like "other" from inventory systems.

It was clear from the CODIS case study that the use of unstructured fields requires close control by the system manager to define what types of information is to be stored in each one (Fyles and King, 1994). In addition, it was clear that the tasks of system manager and cataloguer need to be separated in order to ensure that overall system control be maintained.

The effective implementation of data structuring can result in a number of significant improvements on traditional systems including: improved data input speed and reliability of input; improved searches through "exact search capabilities"; and improved control of data variability through the use of look-up lists controlled by a data manager.

Inventory systems for environmental research should be multidisciplinary in scope. A critical feature of CODIS-like systems is that all disciplines included must share a parallel structure. Each discipline has its own disciplinary-specific data fields while sharing common non-disciplinary system-level files with all other disciplines in the system. The parallel structure serves as a template for data input and look-up list creation that is both flexible and robust. The sharing of system level files serves to decrease unnecessary duplication and improves handling of multidisciplinary datasets.

Data should be collected in intuitive groupings of similar data-density. The CODIS structuring task demonstrated that coarseness of the groupings is an issue for any process that assembles data. If the data groupings become too large then individual groups become overfilled and searches become impractical. Too fine a sieve results in underpopulated fields. Experience creating CODIS demonstrated that of the two choices

the smaller sieve is preferable. It was found to be significantly easier to lump together several smaller groups than it was to break a larger group into smaller substituents. In the case of the CODIS Continental Chemistry media list, it was demonstrated to be relatively simple to assemble the 14 fish tissue sub-groups into three or four larger groupings. The act of breaking down the constituents list, however, was both time consuming and difficult.

The use of multi-tiered informational groupings can decrease coarseness problems and hence increase dataset usefulness. With respect to locations and areas, the use of the multi-tiered location descriptions allows data to be associated to locations within an appropriate level of reliability.

Errors in metadata systems have a profound effect on system performance. Metadata errors nullify the "exact search" capability of the system. This limits the usefulness of the inventory to primary users who have previous knowledge of the existence of a dataset. Therefore, a formalized process to determine the error rates must be incorporated into the development of any inventory system. The error rate, determined by the analysis, should be well documented in order to provide users with an understanding of the reliability of the inventory.

Principal investigators should be involved in all aspects of the creation of datasets, from the structuring of the data through the appraisals of individual measurements to the determination of the pedigree. Investigators bring unique insight and input into the dataset creation process and have a personal stake in ensuring that their data are not misused.

6.3 Critical Responsibilities

The previous section presented recommendations for the designers of environmental information systems. This section will continue that work by considering the responsibilities of individuals at each point in the process of using scientific data in

environmental decision-making. It must be emphasized that the key to all these responsibilities is documentation. Without adequate documentation, even the most effective methodology to address the data needs of environmental decision-making is of little use. Documentation aids individuals not directly associated with the measurement of a variable in understanding the strengths and limitations of that measurement. It has been emphasized many times in this thesis but bears repeating, sampling of environmental variables produce snapshots in time that cannot be retaken. Documentation provides an audit trail so subsequent workers can reconstruct the work and understand its limitations and strengths. The issue of documentation becomes an even more important consideration as standards change and users are forced to reconstruct past work.

6.3.1 Responsibilities before Data Collection

The creation of the appraisal process for experimental events, as outlined in the NCIS case study, demonstrated the importance of proper planning before measurements are carried out. Proper planning can affect both the present and future usefulness of the data derived from those measurements. The first responsibility of design is to establish a context for the data collection process. This includes indicating the strength of the process knowledge that has gone into the design and the anticipated uses of the data produced by the effort. The reasons for which the data are being collected should be made clear and anticipated secondary uses should be considered.

The appraisal process demonstrates that once the context of the data collection effort has been established the actual design of the work must be considered. The choice of an appropriate sampling design and methodology has a major influence on potential secondary uses. A small addition of effort early in the process can reap large rewards in data that is valuable for decision-making at the end of the road. A limitation that must be accepted, however, is that while sample collection design decisions must incorporate the

aim of the project for which the data is being collected, they must also optimize environmental and economic factors.

Once a data collection effort has been designed and documented, the final responsibility of the data collection stage is to ensure that the dataset is generated as planned. In most cases, where the researcher who develops the sampling program is involved in the execution of the project, the requirement for overseeing the work is a mere formality. This is not always the case, however, and oversight procedures should be carried out to ensure that data are generated according to the data collection plan.

6.3.2 Responsibility during Sampling and Storage

The role of the sampling and storage process is to collect samples that are representative of the original matrix and to deliver them to the analytical laboratory for analysis with as little an effect on the sample characteristics as possible. Individuals who collect data must document the procedures that were used to make the measurement and should make use of the best practices possible in the collection and preservation of the samples. This will ensure that the samples are representative of their original state when received by the analysts. It must be recognized that environmental variability is real and cannot be eliminated in a data collection process. The aim of the collection and storage process is to minimize the sampling variability so that variability apparent later in the process will predominantly reflect environmental conditions.

The reliability and quality of location information is an increasingly important consideration in geographical information systems and it is at the time of collection that these questions should be answered. Measurements of location should be made as precise as is both practical and required by the collection methodology. Minor investments in time at the moment of collection can go a long way to reducing uncertainty in environmental information systems.

6.3.3 Responsibilities during Analysis

The role of the analytical step is to analyze the environmental sample as presented and to report the characteristics of the sample as faithfully as possible. The obligation of the analyst is to inform not censor. Analysts often do not collect the samples they analyze and so should not be expected to be able to control samples before they arrive at the laboratory. Once a sample reaches the laboratory, however, everything that happens to the sample must be fully documented. All procedures, analyses and tests should accompany the sample so that the output from the analytical lab is not merely a number, but a complete record of how the sample was treated. Analysts are expected to carry out regular QA/QC procedures on the sample and that information must also be documented.

As mentioned previously, analysts cannot control environmental variability in the samples they receive. They must accept that variability and incorporate it in their methodology. The fact that environmental variability exists should not justify reducing the efforts required to maintain low analytical variability. Analytical variability should be the major concern of the analytical step. This variability should be documented and included with the data. This means that data should not be reported without an associated uncertainty, especially with samples at or near the detection and quantification limits. The uncertainty of a sample may make up a large percentage of the actual sample value, and at times may even exceed the value. As an example, a negative value with an associated uncertainty that brackets zero is a perfectly acceptable outcome for a sample blank and samples that are outliers should not be eliminated unless the experimental error is clearly established.

It is the responsibility of analysts not to censor data, but instead to report data with its associated uncertainty. Reporting that a sample showed no measurable concentrations of a compound is a legitimate outcome only when a range of uncertainty is placed around that report. While it is recognized that this uncertainty may be large, it does give more information than merely stating that nothing was detected. The uncertainty becomes even more critical as values approach the level of detection. Users of data must be made

to recognize that data near the detection limit are inherently uncertain and must be treated appropriately. This responsibility of the analyst, therefore, is to report fully any uncertainty.

6.3.4 Responsibilities of the Data Storage/Management System

Of particular interest to this research is the role of the data storage/management system. The role of the data manager is to ensure the effective storage and maintenance of the data produced, while causing as few modifications to fundamental data characteristics as possible. While preserving data characteristics, data managers must enforce rules regarding documentation and should not allow the structure of their information systems to be corrupted. The structure elements play a crucial role in guaranteeing that data and information in the system are accessible for use. Any corruption of the structure makes data and information less accessible. This means that either data that does not fit the data structure should not be included in the system until the data, or the system architecture, have been modified. The proliferation of unused mandatory data fields results in unwieldy systems.

The data manager must ensure that relationships in reported values are preserved. If a dataset includes measurements that are related to locations, then that relationship must appear in the information that is associated with that dataset. If the relationships between data elements are lost in a system, the likelihood that they can be reconstructed subsequent to that loss is low and the resources required to retrofit that data will be high. An essential aspect of the maintenance of data element relationships is the preservation of the link between data and their associated metadata. This is especially important when data of lower reliability is included in a system that also contains data of higher reliability. Decisions are only as reliable as the information and data upon which they are based. Therefore, users must be made aware of the strengths, and limitations of data stored in the system.

Of particular importance in any information management system is security. This is particularly the case in environmental systems that contain economically valuable information. If an analytical laboratory is expected to release valuable proprietary information, then it is the responsibility of the data manager to ensure that competitors not be able to access that information. Decision-makers, however, must be able to access that information. This security consideration is of great importance in developing an effective and trusted information management system. Inventory level systems provide a safe format through which organizations can provide information about their data holdings. Data security is ensured in inventories because archival data is not stored in the system. An individual who breaches the inventory security would not be able to access any significant proprietary data in the archive. Organizations, therefore, would be able to make known the existence of confidential data without risk that the data itself may be compromised. Users who required access to the original data could use the inventory to obtain contact information in order to contact the data owners in order to get the raw data themselves.

Data managers and researchers must also be responsible for ensuring the timeliness of the data in their systems. Inventory systems can be likened to journal abstracting services; they must be continuously updated in order to be useful. The act of creating a comprehensive inventory system is both expensive and time consuming and given that time progresses, ultimately futile without continuing funds to maintain the system. Consequently, inventory systems are most practical in environments of continued funding and mandated updates. As an example, the ongoing mandate to maintain NCIS should ensure its continuing usefulness and the timeliness of its data.

As the contents of inventory systems mature, an additional consideration confronts the data manager. With advances in techniques comes new understanding. This understanding often brings changes in standard techniques to carry out measurements or analyses. These advances have the potential to affect the rating tools and older assigned ratings. It is critical that the decision trees and their guidelines remain as living documents. The decision trees must be updated regularly to reflect changes in standard

scientific techniques. Such changes must be reflected in the guidelines. The guidelines themselves should have update logs associated with them. These logs should reflect when the guidelines were last re-evaluated and what changes were made. Significant changes in methodologies also have the potential to result in changes in the reliability ratings assigned to older datasets. Analytical techniques, now considered accurate, might later be shown to instill bias in the samples. If this is the case then older ratings may have to be revisited and revised. An alternative approach might be to add a date to the ratings to provide users with greater insight as to the likely reliability of the data.

6.3.5 Responsibilities of the Data User

In traditional information systems, secondary users have had few responsibilities. They access the system, obtain data and then make use of that data as they see fit. It is imperative that data users be especially aware of their responsibilities. The data user's work is the culmination of the efforts of all the other partners in the data-to-decision process. Data users must, therefore, acknowledge and understand the responsibilities of all the individuals who preceded them in the data-to-decision process. In addition, they must also bear specific responsibilities. Data users must take the time to understand not only the data, but also their context. Users should not just accept numbers and values, but must ensure that they understand where the values came from and how they might be appropriately used. Of particular concern to data users should be an understanding of the factors that make up the data variability. Data users have the responsibility not to use individual values outside of their context and to keep data in appropriate blocks. Data should not be disconnected from its associated uncertainty information and should not be forced for inappropriate uses.

6.4 A New Conceptual Model for Environmental Information Systems

Based on the gap-analysis of the pre-existing systems, and the recommendations and general responsibilities presented above, it is possible to suggest a conceptual model that meets all these requirements. This model is presented in Figure 6.5.

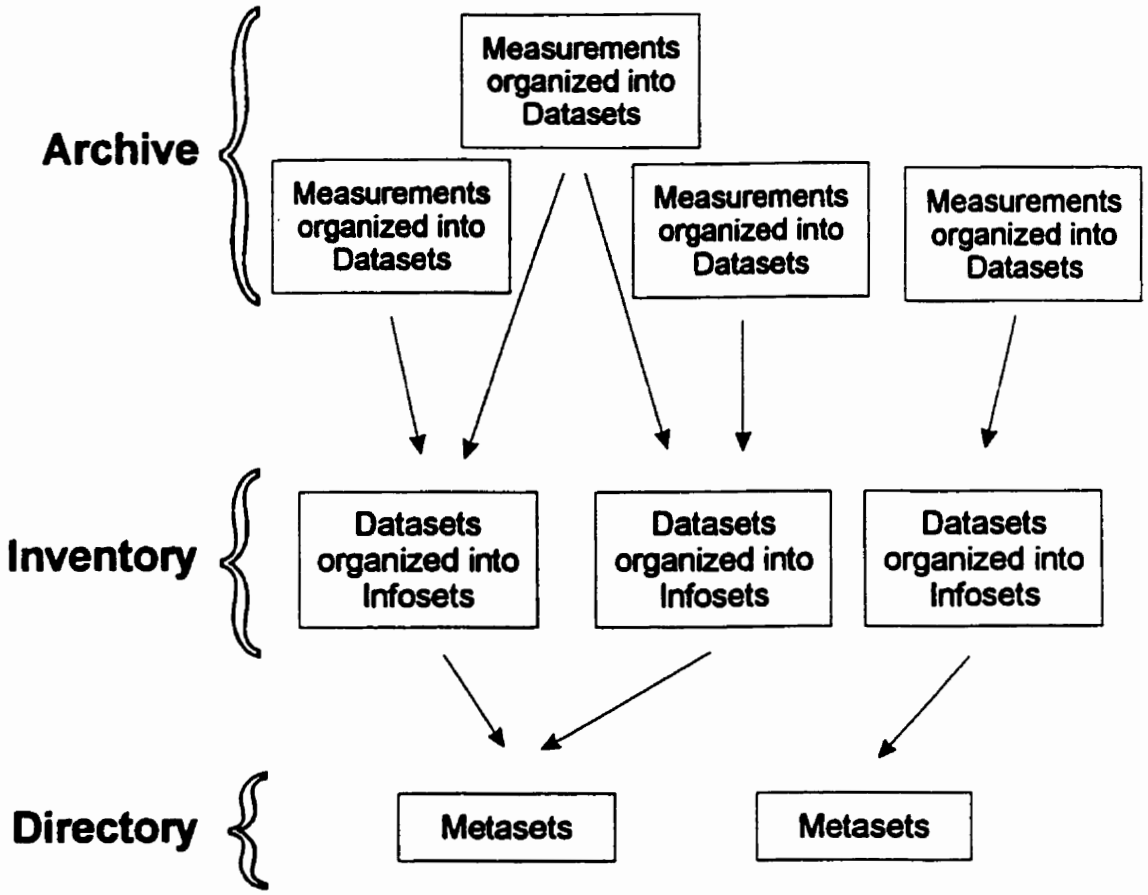


Figure 6.5 The Three-tiered Conceptual Model

The conceptual model makes use of the multi-tiered approach incorporated in NCIS. Like NCIS, it includes three levels: archives, inventories and directories. Each level represents from one-to-many data/information systems that are linked to the others as part of the larger overall system but can function independently. The multiple tiers provide a spectrum of options and tools for all types of users.

The foundation of the model is the archival level made up of one or more archives. The archives provide locations to store the output from research, monitoring and observation activities. Individual archives can vary in design based on the individual requirements of their developing agencies but all archives in the system share a number of fundamental characteristics. Data in the archive are untruncated and collected into datasets. The conventions for creating datasets incorporate an expanded definition that includes the

intentions of the primary user. These datasets are the smallest information element that can be removed from the archive. This prevents users from accessing individual measurements in the archives without their associated metadata. These new datasets are linked to critical QA information, which must be included with any report. Like datasets in CODIS, individual measurements can be related to more than one dataset. For secondary users, access to information about the datasets stored in the archives is through inventories.

The design for inventories is based on the metadata model developed for CODIS but incorporates a new information element: the infoset, which links datasets with their contextual information. The contents of individual inventories can vary, and will depend on the agencies that created them. One inventory system might have a geographical base (i.e. the Canadian West Coast and Arctic) while another might be based on a specific mandate (i.e. Green Plan Toxics data). Inventories can refer to data archives or other inventories. Since the inventories include metadata, they can provide for both exact and text-based searches. The inventories incorporate advances in GIS and mapping technology to provide the tools needed to ensure that users are able to identify data of interest using geographical characteristics as well. Indicators of reliability and context in the inventory level assist users in identifying useful data, and can serve as an additional set of search criteria.

The measurements that make up datasets are appraised for reliability using the methodology developed for CODIS. Individual datasets are associated with infosets, that serve the task of relating the context of one or more datasets based on a specified activity. These infosets are evaluated for context and have pedigrees documenting those evaluations. Since a dataset can be associated with more than one infoset, the limitations placed by the NCIS structure (one measurement related to only one experiment) are bypassed.

The highest level in this new information hierarchy is the directory. Directories provide a general overview of activities carried out in a particular area of interest and incorporate

the tools needed to identify appropriate inventories and archives. The design combines the NCIS directory with a similar model proposed by Cornford and Blanton (1993). They theorized the creation of a "meta-database" for use with Fraser River Estuary information (called FREDI). In the Cornford and Blanton model, these directories are created by central organizations and serve as general tools to aid users in identifying who is carrying out useful work. The directories do this by providing general information about projects, individuals and research activities. A comparable system is being developed by the Canadian Council of Ministers of the Environment (CCME). They have created a meta-database called "Databases for Environmental Analysis" it is a collection of descriptions documenting the contents of over 1,200 databases held by the Government of Canada and by the 12 provincial and territorial governments (CCME, 1999).

The conceptual model above presents an ideal approach to the organization of data and information resources in the environmental field. It suggests tools that, when completed, have the potential to improve significantly the data and information climate in the environmental field. Figure 6.6 provides a breakdown of the process by which the various information sub-units are created.

The process organizes the tools developed in the case studies in a manner that can be readily adapted to address the current needs of the environmental information field. It applies the lessons learned through the case studies by providing a methodology to efficiently and repeatedly produce datasets for use in an inventory system.

The process is broken down into three steps, the assembly of measurements into datasets, the process of assigning contextual information and creating infosets, and the subsequent creation of metasets from the infoset/dataset information. The first two steps in this process are deterministic and use the tools and skills developed in the case studies. The process of creating metasets has not been completely described. The following sections will define the types of information stored at each level of the conceptual model.

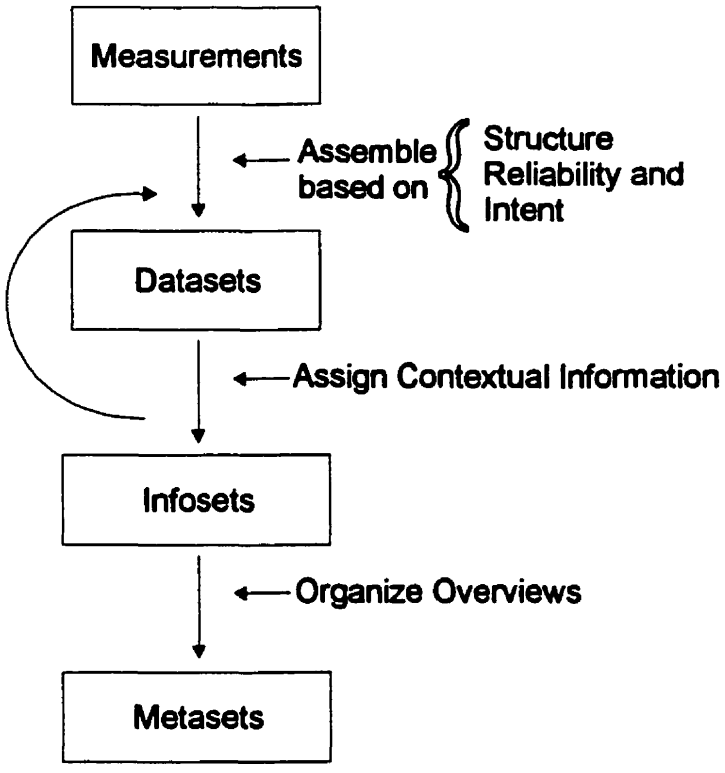


Figure 6.6 Steps in the Creation of Metasets

6.4.1 Model Details: Archives

Archives in this model share the general characteristics of ideal archives described in section 6.1. The critical change is the addition of a new data element, the dataset. While users would be able to search for and view individual measurements, datasets would serve as the smallest informational unit that could be reported. Datasets preserve the internal relationships between individual measurements as delineated by the primary user. The CODIS case study described a process for creating datasets. This new model expands on that process by including a new information element and changing one of the CODIS conventions for creating datasets.

The new definition of a dataset expands on the CODIS definition by incorporating the intentions of the primary user and including the requirement that a dataset must remain as a unit. This thesis proposes a new definition of a dataset for the conceptual model:

> A dataset is a collection of measurements, assembled for a specified purpose, and unified by one or more of the following characteristics: chemical species, biological species, physical matrix, geographical locations, or sampling methodology. The measurements must be treated uniformly, ideally by a single agent or agency and should be internally consistent with respect to sampling methodology. The measurements within the dataset need not always be of the same type but must always be treated as a unit.

The CODIS case study included a number of conventions governing the creation of datasets (Section 3.2.2). The first of these conventions was "when subdividing a large report into simpler datasets one should strive to maximize the size of the datasets." In this new model, that convention would no longer apply. Instead, the primary user must strive to group measurements into datasets based on considerations of internal consistency and inter-relatedness of the measurements. In addition, datasets and data in archives are in a many-to-many relationship. One dataset might be related to many measurements or data records and one measurement or data record might be associated with many datasets. This would allow a primary user to create several differing datasets from the same group of measurements.

6.4.2 Model Details: Inventory

The dataset serves as the basic information sub-unit in the inventory level of the new model. Consequently, understanding how datasets relate to other information elements is critical to the success of the system. Figure 6.7 delineates those relationships and could be used as the basis for an entity-relationship diagram for system designers.

Datasets in the inventory are associated with three classes of metadata: 1) global information that is shared by all datasets 2) dataset-specific information that informs users about the relationships between individual datasets, and 3) discipline-specific information that delineates the specifics of each individual dataset.

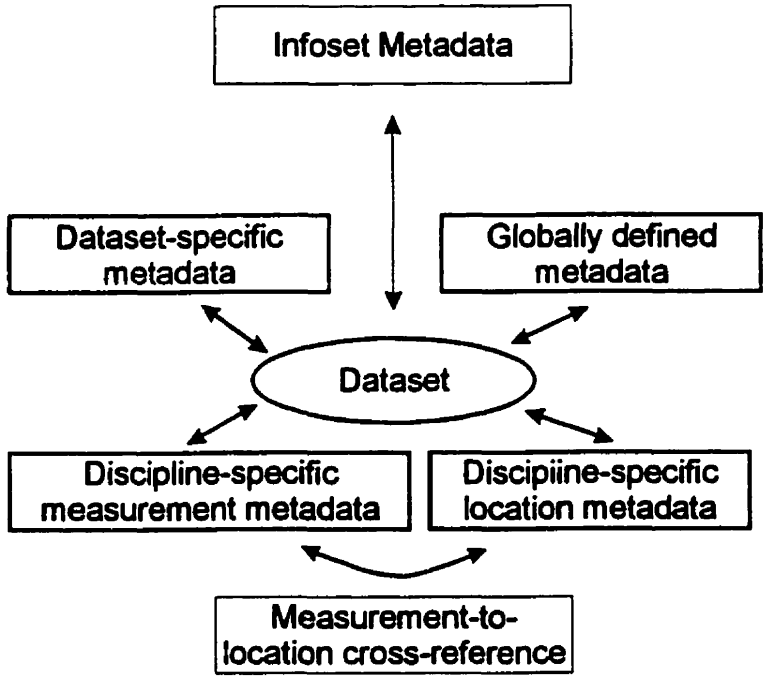


Figure 6.7 Dataset Properties

The global fields are presented in Figure 6.7 as the "globally-defined metadata". These fields are comparable to the global fields used in CODIS. They provide the information needed to carry out general searches of the datasets and report the bibliographic results.

The dataset-specific information consists of two groups of metadata: the "dataset specific metadata" and the "infoset metadata". The dataset-specific metadata provides identification information equivalent to the DS_ID file in CODIS. It indicates which disciplines are covered by the dataset, the overall dataset start and stop dates and includes identification information.

The infoset metadata is the new addition in this model. As discussed previously, the measurements that make up a dataset are assembled based on the original intentions of the primary user. The infoset metadata provides a location to store information regarding that purpose. This can be done using an "infoset-to-dataset cross-reference" table or a

number of alternative means. The table would collect datasets together much in the same way as the DS_ID table in CODIS collects measurements together.

Infosets would be created initially by the primary user. Each unique infoset would be associated to one-or-more datasets. The infoset information components would be created using the processes developed in the NCIS case study. An activity type would be defined using a formalized process to sort the activity into one of a number of potential categories such as research, monitoring, surveys, modeling or model testing. Associated with the activity type would be the summary pedigree, which would be produced using the methodology developed to appraise experimental events in NCIS. Since individual infosets may have timelines that differed from those of the included datasets, the need would exist to include infoset start and stop dates. Infosets and datasets would be in a many-to-many relationship; that is one dataset may be related to one or more infoset and one infoset may be related to one or more datasets. Individual infosets would be a searchable item in an inventory system.

The most detailed metadata for searching purposes is the discipline-specific information, which consists of "discipline-specific measurement metadata" and "discipline-specific location metadata". The discipline-specific, measurement metadata provides the critical details of the measurement or activity. It includes structured information about the specific parameters and measurements carried out, the media and/or taxon involved and the method used to carry out the measurement. In addition, the discipline-specific metadata includes the measurement ratings for the dataset.

The location metadata offers a variety of opportunities to locate the dataset in space. This might include latitude/longitude information and specific named areas. In addition, it includes critical measurement-at-location metadata. This last feature provides specific details about unique features of the measurement at specific locations. This differs from the general measurement-to-location cross-reference table, which specifies which measurements are related to which locations or sites.

Inventory systems based on metadata are able to search for the existence of data without the need to access the archive containing the original data. This has the potential to increase the availability of data at lower cost. Since inventory systems need not contain the raw data, the requirement for expensive security routines to preserve data privacy is bypassed. The existence of data (including confidential data values) can be made known without exposing that data to accidental release.

6.4.3 Model Details: Directories

The final tier in the new model is comprised of directories. Directories contain general overviews of the activities being carried out and data available in a particular area of interest and provide the tools needed to identify appropriate inventories and archives. The emphasis of this research has been on understanding the requirements of inventory systems. Based on the understanding developed in this thesis it is possible to suggest a proposed model for the properties of the information units of a directory system (Figure 6.8). The basic information unit of a directory is the metaset. Following the model used in the dataset, the metaset includes two distinct types of information: 1) higher-level organizational/identification information, and 2) metadata derived from datasets and infosets.

The "organizational" and "identification" information in the metaset will aid users in the identification of useful inventories, archives and datasets for use in their own work. The organizational information provides details on the availability, status and format of any informational resources, while the identification information would provide information regarding data and database "ownership" including the format, availability and status of any data and descriptive bibliographic details.

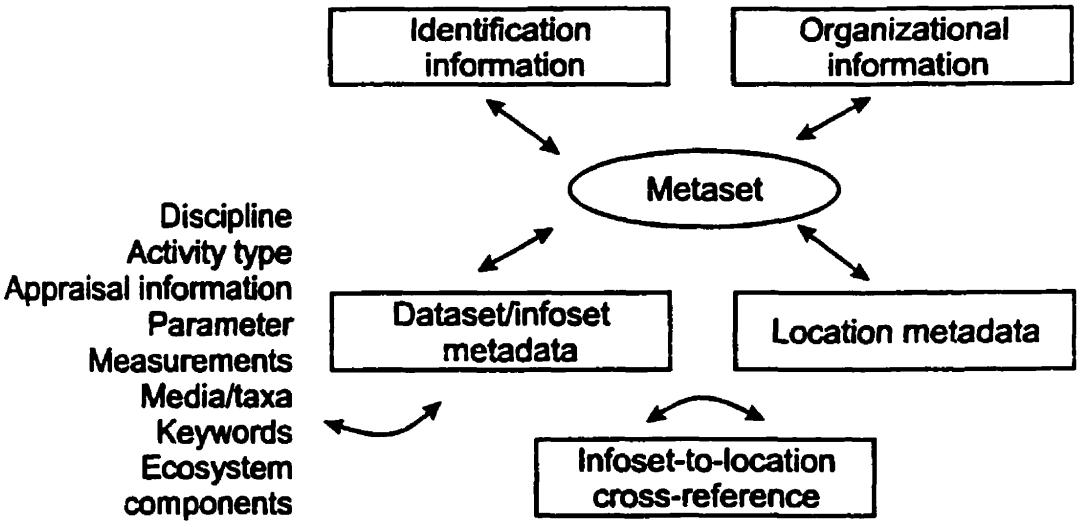


Figure 6.8 Metaset Properties

The dataset/infoset information involves areas of likely interest to searchers and is limited in detail. Thus, a searcher might be able to search the taxa field for a particular species of commercial value (be it fish, crustacean etc.,) but would not have that same level of detail for the majority of benthic invertebrates. Measurement and parameter lists will require similar condensing. The activity of creating the lists is a human task and will require a moderate amount of finesse to provide an adequate level of coarseness without any single list becoming unmanageable.

In addition to the dataset-specific information, the dataset/infoset metadata requires some new features; these include some form of keyword lists and the inclusion of a new structuring unit based on ecosystem components. Cornford and Blanton (1993) proposed the use of the basic environmental impact assessment concept of "valued ecosystem components" (VEC) for use in their prototype system FREDI. They broke down the environmental study of the Fraser Basin into aggregate classes consisting of five VECs (fish, wildlife, vegetation, plankton and benthos) and six environmental components (metals, organics, other chemicals, hydrology and geology) (Cornford and Blanton, 1993). In the model presented here, these 11 components would be combined with a still-to-be-determined further list to create a set of high-level categories with which to

sort various activities. It is anticipated that the effort involved in creating such a list would be comparable to that required to create the parameter-group list used in CODIS. This might require an extensive evaluation of potential data-density issues followed by consultation with interested researchers and a peer-review process.

6.4 Future Considerations and Further Work

This work has presented a conceptual model of a three-tiered environmental data/information system. The conceptual model is supported by practical tools and pre-existing systems. Through the combination of these pre-existing tools and systems a viable new system could be produced to ensure the accessibility of appropriate data and information for a range on environmental decision-making purposes. The practical creation of such a system entails a number of opportunities and potential problems. Archives currently exist that could be directly linked to the inventory. New archives being designed for inclusion into the system would benefit from the inclusion of some form of EDI and an automated QA Index similar to those implemented in EMS. The EDI ensures the automated inclusion of data into the archives, while the automated QA Index ensures that a minimum of QA/QC data is included with the archived data.

This thesis presents a model that could serve as the template for the inventory system. The freeware approach used to distribute programs such as CODIS would facilitate the accessibility of the inventory. In this approach an organization with interests in the field could create the inventory and make the framework and software available to interested organizations, which would then be responsible for filling the inventory.

The directory-level system proposed in the conceptual model remains primarily a theoretical construct. Cornford and Blanton (1993) through FREDI have created an initial template for such a system, but a great deal of work would have to be carried out to deal with the intricacies of designing a system of this scope. It would be likely that the scale of such a task would require either the resources of the federal or the provincial government. In order for such a system to succeed, far greater co-operation would be

required. An effective directory system should include information from governments, academic institutions and private firms.

In addition to the practical task of creating functional computer systems a number of critical education activities should be carried out. As was demonstrated in the NCIS case study, the need exists to educate researchers and decision-makers as to the value of appraising experiments and assigning reliability indicators to datasets. The current climate in many of our institutions (as typified by DFO) is one in which researchers see their research and monitoring data as having limited value beyond the task for which they were initially collected. The benefits of making data available for secondary use must be made clear to researchers and any concerns allayed. In particular, further dialogue with researchers needs to occur on the classification of their work by experiment type as recommended in Chapter 4. Through the application of context to their datasets, researchers can rest assured that their data will not be unintentionally misused. While no system exists to prevent malicious misuse of data, the association of clear contextual information with datasets can go a long way to ensuring that individuals who misuse datasets will be quickly identified and their work given the respect it deserves.

6.5 Conclusion

This research has highlighted the need for data in environmental information systems to be accompanied with indications of reliability and context and has presented a methodology to ensure that this is carried out. The fact that both NCIS and EMS incorporate many of the features developed in this work and presented here provides proof that this research has advanced the cause of improved reporting of environmental variables in information systems. It is hoped that those individuals and researchers who have been involved in this research will continue to insist that contextual information be supplied with their data and will continue to work towards ensuring that their data are stored in a manner that is both secure, yet accessible. It is also hoped that those researchers who have attended our presentations and workshops will reject the alternative of placing data into archives that do not include associated contextual information.

The final desire of this researcher is that this work will eventually influence decision-makers and the public. Decision-makers must be educated to request and expect contextual information when supplied with data for decision-making purposes and must be made aware of the dangers associated with placing undue faith in scientific results. It must be made clear that while uncertainty is a reality in any environmental activity, uncertainty can be decreased through an increased understanding of the strengths, limitations and scope of data supplied to carry out environmental decision-making.

Chapter 7 Bibliography

ACS Committee On Environmental Improvement; Subcommittee on Environmental Analytical Chemistry. 1980. Guidelines for Data acquisition and data quality evaluation in environmental chemistry. *Analytical Chemistry* 52: 2242-2249.

Adams, K.M. 1991. Peer review: An unflattering picture. *Behavioral and Brain Sciences* 14: 135-136.

Albright, L.J., T.G. Northcote, P.C. Oloffs and S.Y. Szeto. 1975. Chlorinated hydrocarbon residues in fish, crabs and shellfish of the lower Fraser River, its estuary and selected locations in Georgia Strait, British Columbia - 1972-73. *Pesticides Monitoring Journal* 9(3): 134-140.

Al-Zobaidie, A. and J.B. Grimson. 1988. *Information and Software Technology* 30(8): 484-496.

American Public Health Association (APHA), American Water Works Association and Water Environment Federation. 1992. Standard Methods for the Examination of Water and Wastewater. 18th edition. 1268 pp.

AXYS. 1994. CIS Systems Design Document. AXYS Environmental Consultants, Sydney, B.C.

Bates, D. <dbates@pangea.ca> "EMS QA/QC Minutes" e-mail to Malcolm Clark. December 17, 1996.

Bernard, D.P., D.R. Marmorek, T.M. Berry, M. Paine, C.J. Perrin, and J. Richardson, (1993): Fraser River Basin Assessment Program: Conceptual Monitoring Design. Environment Canada, Environmental Services Branch, Vancouver. 162p.

Birch, J.R., D.B. Fissel, D.D. Lemon, A.B. Cornford, R.A. Lake, B.D. Smiley, R.W. Macdonald, and R.H. Herlinveaux. 1983. Northwest Passage: physical oceanography – temperature, salinity, currents and water levels. *Canadian Data Report of Hydrography and Ocean Sciences* 5: Vol.3. 262p.

Blyth, C.A., D.J. Thomas, and S. Gormican. 1993. Protocol Development Workshop for the Contaminants Information System (CIS) May 18-20, 1993: Workshop Proceeding Report. Prepared for: Fisheries and Oceans Canada, By: AXYS Environmental Consulting. Sidney, B.C. 48p.

Bolin, B. 1994. Science and policy making. *Ambio* 23(1): 25-29.

Bowser, C. J. 1986. Historic data sets: Lessons from the past, lessons for the future. In: Michener, W.K. (ed.) Research Data Management in the Ecological Sciences. Columbia, S.C., Published for the Belle W. Baruch Institute for Marine Biology and Coastal Research by the University of South Carolina Press: 155-179.

British Columbia Ministry of Environment, Lands and Parks. 1997. B.C. Environmental Monitoring System: Frequently Asked Questions. January 1, 1997. British Columbia Ministry of Environment, Lands and Parks, Victoria, B.C. 10p.

British Columbia Ministry of Environment, Lands and Parks. 1997. B.C. Environmental Monitoring System: Using EMS. Release 1.3, April 1, 1997. British Columbia Ministry of Environment, Lands and Parks, Victoria, B.C. 164p.

Brown, J.H. 1994. Grand challenges in scaling up environmental research. In Michener, W.K., J.W. Brunt, and S. Stafford (eds.) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis Ltd, London. 555p.

Bundy, M. 1970. The management of information and knowledge. In: The Management of Information and Knowledge: A compilation of Papers Prepared for the Eleventh Meeting of the Panel on Science and Technology. Committee of Science and Astronautics, U.S House of Representatives. 130p.

Caldwell, L. K. 1990. Between Two Worlds: Science, the Environmental Movement, and Policy Choice. Cambridge University Press. New York. 224p.

Caldwell, L. K. 1970. Environment: A Challenge for Modern Society. Garden City, N.Y. Published for the American Museum of Natural History by: the Natural History Press. 292p.

Campanario, J.M. 1996. Have referees rejected some of the most-cited articles of all times? *Journal of the American Society for Information Science* 47(4): 302-310.

Canadian Global Change Program, Data and Information Systems Panel, and Royal Society of Canada. 1996. Data Policy and Barriers to Data Access in Canada : Issues for Global Change Research ; A Discussion Paper. Ottawa, Royal Society of Canada. 55p.

Carswell, A, <Carswell@DFO-Mpo.GC.CA> "RE: 1996 Protocols Workshop". Private e-mail to Blair King. April 7 1998.

Chalmers, A.F. 1982. What is this Thing Called Science? Second Edition. University of Queensland Press. 179p.

Chechile, R.A. 1991. Introduction to environmental decision making. In: Chechile, R.A., and Carlisle, S. (eds.) Environmental Decision Making: A Multidisciplinary Perspective. Van Nostrand Reinhold, New York. 296p.

Chrisman, N.R. 1994. Metadata required to determine the fitness of spatial data for use in environmental analysis. In: Michener, W.K., J.W. Brunt, and S. Stafford (eds) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis Ltd, London. 555p.

Christensen, N.L., A.M. Bartuska, S. Carpenter, C. D'Antonio, R. Francis, J.F. Franklin, J.A. MacMahon, R.F. Noss, D.J. Parsons, C.H. Peterson, M.G. Turner, and R.G. Woodmansee. 1996. The report of the Ecological Society of America committee on the scientific basis for ecosystem management. *Ecological Applications* 6(3): 665-691.

Chutter, F.M. 1972. A reappraisal of Needham and Usinger's data on the variability of a stream fauna when sampled with a Surber sampler. *Limnology of the Ocean* 17: 139-141.

Cicchetti, D.V. 1991. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences* 14: 119-134.

Cicchetti, D.V. 1976. Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry* 129: 452-456.

Clark, C.W. 1976. Mathematical Bioeconomics: The Optimal Management of Renewable Resources. John Wiley & Sons, New York, New York. 352p.

Clark, M. J. R. and J.D. Ellis. 1976. B.C. Water Resources Services and Environmental Data Base (EQUIS). Pacific Biological Station, Nanaimo, Fisheries and Marine Service: 27p.

Clark M.J.R., and P.H. Whitfield. 1993. A practical model integrating quality assurance into environmental monitoring. *Water Resources Bulletin* 29: 119-30

Clark M.J.R., and P.H. Whitfield. 1994. Conflicting perspectives about detection limits and about the censoring of environmental data. *Water Resources Bulletin* 30: 1063-79.

Clay, D. 1997. Ecosystem Monitoring, Data Management, and QA/QC in Park Science. Parks Canada, Ecosystem Science Review Reports, Fundy National Park, New Brunswick. 23p.

Clayton, C.A., J. W. Hines and P.D. Elkins. 1987. Detection limits with specified assurance probabilities. *Analytical Chemistry* 59: 2506-2514.

Clements, W.H. 1991. Characterization of stream benthic communities using substrate-filled trays: Colonization, variability, and sampling activity. *Journal of Freshwater Ecology* 6(2): 209-222.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 10(1): 37-46

Cole, S., J. Cole, and G.S. Simon. 1981. Chance and consensus in peer review. *Science* 214: 881-886.

Cornford, A., and C. Blanton. 1993. An overview of planning and management in environmental decision making with an emphasis on environmental collection and monitoring. In: Farrell, A.P. (ed.) The Aquatic Resources Research Project: Towards Environmental Risk Assessment and Management of the Fraser River Basin. November 1993.pp. 379-451. 451pp.

Cornford, A.B., D.D. Lemon, D.B. Fissel, H. Melling, B.D. Smiley, R.H. Herlinveaux and R.W. Macdonald. 1982. Beaufort Sea: Physical oceanography -- temperature, salinity, currents and water levels. *Canadian Data Report on Hydrography and Ocean Sciences*. 5(1) 279p.

Constanza, R.B., S.O. Funtowicz, and J.R. Ravetz. 1992. Assessing and communicating data quality in policy-relevant research. *Environmental Management* 16: 121-131.

Crandall, R. 1991 What should be done to improve peer reviewing? *Behavioral and Brain Sciences*.14: 143.

Dane F.C. 1990. Research Methods. Brooks/Cole Publishing, Belmont. 348p.

Data Quality Organization Newsletter. 1996. Nondetects: Planning and Statistical Analyses to Mitigate Their Effects. U.S. Department of Energy Office of Environmental Management (EM-76) Pacific Northwest National Laboratory.

Delcomyn, F. 1991. Peer review: Explicit criteria and training can help. *Behavioral and Brain Sciences*14: 144.

Duke, T.W., J.I. Lowe, and A.J. Wilson, Jr., 1970. A polychlorinated biphenyl (Arochlor 1254) in the water, sediment and biota of Escambia Bay, Florida. *Bulletin of Environmental Contamination and Toxicology* 5: 171-175.

Eberhardt, L.L. and J.M. Thomas. 1991. Designing environmental field experiments. *Ecological Monographs* 61(1): 53-73

El-Shaarawi, A,H., and S.R. Esterby. 1992. Replacement of a censored observations by a constant: An evaluation. *Water Resources Research* 26(6): 835-844.

Ellis, J. D. and M.J.R. Clark. 1977. A Computer System for Biological Information. Victoria, Province of British Columbia, Ministry of the Environment, Pollution Control Branch. 22p.

Environment Canada. 1993. Preparing the Second Priority Substances List Under the Canadian Environmental Protection Act (CEPA). Environment Canada and Health and Welfare Canada, Ottawa, Ontario. 14p.

Environment Canada and Department of Fisheries and Oceans. 1993. Technical Guidance Document for Aquatic Environmental Effects Monitoring Related to Federal *Fisheries Act* Requirements, Version 1.0.

Environmental Sciences Limited. 1988. QUIKMap System Specifications. ESL Environmental Services Limited, Sidney, B.C.

Fancy, L. 1998. Personal Communication May 26, 1998. Data Manager, Newfoundland District, Department of Fisheries and Oceans, Canada.

Farrell, A.P. (ed.) 1993. Towards Environmental Risk Assessment and Management of the Fraser River Basin. A Technical Report for the B.C. Ministry of the Environment Centre for Excellence in Environmental Research. Dept. of Biological Sciences, Simon Fraser University 451p.

Federal Geographic Data Committee. 1994. Contents standards for digital geospatial metadata (June 8). Federal Geographical Data Committee. Washington, D.C.

Fletcher, J.J., and T.T. Phipps. 1991. Data needs to assess environmental quality issues related to agriculture and rural areas. *American Journal of Agricultural Economics* August 91: 926-932.

Foehrenbach, J. 1971. Chlorinated hydrocarbon residues in shellfish (*Pelecypoda*) from estuaries of Long Island, New York. *Pesticide Monitoring Journal* 5: 242-247.

Fyles, T.M., B. King, and P.R. West. 1993a. Continental and Oceanographic Data Information System. CODIS 1.0: Protocols, Software, Compilation, and Appraisal of Meta-Data of Organic Contaminants in the Fraser River Basin. Conservation and Protection Service, Environment Canada, North Vancouver, B.C. DOE - FRAP 1993-24, 130pp.

Fyles, T.M., B. King, and P.R. West. 1993b. A protocol for the evaluation of the quality of trace organic data. Conservation and Protection Service, Environment Canada, North Vancouver, B.C. DOE -FRAP 1993-25, 33pp.

Fyles, T.M., and B.A. King. 1994. Fraser River Basin Benthic Invertebrate Catalogue, CODIS 2.0 Implementation: A QA/QC Evaluation of the Catalogue. In: Johansen, J.A. and Reis, K.E.M. Fraser River Basin Benthic Invertebrate Catalogue. Prepared for Conservation and Protection Service, Environment Canada, North Vancouver, B.C. Published as DOE FRAP 1994-1

Fyles, T.M., P.R. West and B.A. King. 1995. A meta-data approach to information management in the Georgia Basin. In: Robichaud, E (ed.) Puget Sound Research '95 Proceeding: Meydenbauer Center, Bellevue, Washington, January 12-14, 1995.

Fyles, T.M, B.A. King, M.P. Pawluk, and P.R. West. 1996. Protocols for the Appraisal of Contaminant Survey and Experimental Events in CIS. Parts I and II Prepared for Fisheries and Oceans Canada (DFO) Green Plan Toxic Chemicals Program, Institute of Ocean Sciences, Sidney, B.C.

Fyles, T.M., B.A. King, M.P. Pawluk, G.D. Robertson, B.D. Smiley, P.R. West, and C.S. Wong. 1997 CODIS 2.0: User Guide (110 pages) and CD-ROM (120 Mbyte decompressed). Department of Fisheries and Oceans.

Fyles, T.M., B.A. King, M. Pawluk, and P.R. West. 1997. Implications of Data Truncation in Environmental Decision-Making. Report prepared for B.C. Ministry of Environment, Lands and Parks. Nov. 1997. 73p.

Friedl, M.A., and C.E. Brodley. 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of the Environment* 61: 399-409.

Funtowicz, S. O. and J.R. Ravetz. 1993. Science for the post-normal age. *Futures* September: 739-755.

Garfunkel., J.M., R.H. Ulshen, H.J. Hamrick, and E.E. Lawson. 1990. Problems identified by secondary reviewers of accepted manuscripts. *Journal of the American Medical Association* 263: 1369-1371.

Gilliom, R.J., and D.R. Helsel. 1986. Estimation of distributional parameters for censored trace level water quality data 1. Estimation Techniques. *Water Resources Research* 22(2): 135-146.

Gilliom, R.J., R.M. Hirsch, and E.J. Gilroy. 1984. Effect of censoring trace-level water-quality data on trend-detection capability. *Environmental Science & Technology* 18(7): 530-535.

Gleit, A (1985): Estimation for small normal data sets with detection limits. *Environmental Science & Technology* 19(12): 1201-1206.

Gosz, J.R. 1994. Sustainable Biosphere Initiative: Data management challenges. In: Michener, W.K., J.W. Brunt, and S. Stafford (eds) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis Ltd, London. 555p.

Green, R.H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. John Wiley & Sons, Inc., Toronto. 257pp.

Greene, R. 1991. Is there an alternative to peer review? *Behavioral and Brain Sciences* 14: 149-150.

Guay, C. <Guay@dfo-mpo.gc.ca>. "Re:Your Resquest". Private e-mail communications to Blair King March 26-May21, 1998.

Gurevitch, J., L.L. Morrow, A. Wallace, and J.S. Walsh. 1992. A meta-analysis of competition in field experiments. *American Naturalist* 140: 539-572.

Harrison, S. 1999. Local and regional diversity in a patchy landscape: Native, alien, and endemic herbs on serpentine. *Ecology* 80(1): 70-80.

Harmancioglu, N.B., O. Fistikoglu, S.D. Ozkul, and M.N. Alpaslan. 1998. Decision making for environmental management. In: Harmancioglu, N.B., V.P. Singh, and M.N. Alpaslan. (eds.) Environmental Data Management. Kluwer Academic Publishers, Dordrecht pp.243-288.

Hayne, D.W. 1978. Experimental designs and statistical analyses. In: D.P. Snyder (ed.) Populations of Small Mammals Under Natural Conditions. Volume 5. Special Publication Series. Pymatuning Laboratory of Ecology. University of Pittsburgh, Pittsburgh, Pennsylvania, U.S.A pp3-13.

Helsel, D.R. 1990. Less than obvious: Statistical treatment of data below the detection limit. *Environmental Science & Technology* 24(12): 1766-1774.

Helsel, D.R., and T.A. Cohn. 1988. Estimation of descriptive statistics. *Water Resources Research* 24(12): 1997-2004.

Hindrichs, A.E. 1998. Environmental data management: Storage, handling and retrieval. In: Harmancioglu, N.B, Singh, V.P., and Alpaslan, M.N. (eds.) Environmental Data Management. Kluwer Academic Publishers, Dordrecht pp.243-288.

Hodgson, C. 1997. How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology* 50(11): 1189-1195.

Holcomb Research Institute. 1976. Environmental Modeling and Decision Making: The United States Experience. Praeger Publisher, New York. 152p.

Hurlbert, S.H. 1984. Pseudorelication and the design of ecological field experiments. *Ecological Monographs* 54(2): 187-211.

Hsu, C., M. Bouziane, L. Rattner, and L. Yee. 1991. *IEEE Trans. Software Eng.* 17(6): 604-625.

Johansen, J.A., and K.E.M. Reis, Fraser River Basin Benthic Invertebrate Catalogue. Conservation and Protection Service, Environment Canada, North Vancouver, B.C. DOE FRAP 1994-17.

Johnson, R.A., and G.K. Bhattacharyya. 1992. Statistics: Principles and Methods. John Willey & Sons, Inc. Toronto, Ontario. 686p.

Jones, K.A., and T.J. Hara. 1988. Behavioral alterations in Arctic Char (*Salvelinus alpinus*) briefly exposed to sublethal chlorine levels. *Canadian Journal of Fisheries and Aquatic Sciences* 45: 749-753.

Keeley, B. <keeley@meds-sdmm.dfo-mpo.gc.ca>. "Re: NCIS documentation". Private e-mail communications to Blair King March 25-December 10, 1998.

Keeley, B. 1998. NCIS Documentation. Unpublished Documentation for the National Contaminants Information System. Version: 15 May, 1997.

Keith. H.K., W. Crummet, J. Deegan, Jr. R.A. Libby, J.K. Taylor, and G. Wentler. ACS Committee on Environmental Improvement 1983. Principles of environmental analysis. *Analytical Chemistry* 55(14): 2210-2218.

Kendall, M,G. 1982. A Dictionary of Statistical Terms. Longman Inc., New York.

Komarkova, V. and J.L. Bell. 1986. Characteristics of scientific database management systems. In: Michener, W.K. (ed.) Research Data Management in the Ecological Sciences. Belle. W. Baruch Library in Marine Sciences Number 16. University of South Carolina Press. 419p.

Kroger, R.L.1972. Underestimation of standing crop by the Surber sampler. *Limnology of the Ocean* 17: 475-478

Lambert, D., B. Peterson, and I. Terpenning. 1991. Nondetects, detection limits, and the probability of detection. *Journal of the American Statistical Association* 86(414): 266-277.

LGS Group Inc. 1995. Environmental Monitoring System Requirements Summary, Prepared for: Ministry of Environments Lands and Parks, Victoria, B.C. June 30, 1995, 40p.

Loomis, T.A. 1974. Essentials of Toxicology, Second Edition. Lea & Febiger Inc. Philadelphia. 223p.

MacDonald, L.H., A.W. Smart and R.C. Wissmar. 1991. Monitoring Guidelines to Evaluate Effects of Forestry Activities on Streams in the Pacific Northwest. U.S. EPA Region 10. EPA 910/9-91-001.

MacKay D.R., and D.G. MacDonell. 1975. Environmental Monitoring; A Compendium of Data Gathering Activities of Environment Canada. Planning and Finance Service Report No.4. Environment Canada, Ottawa. 44p.

Manly, B.F.J. 1992. The Design and Analysis of Research Studies. Cambridge University Press. Cambridge. 353p.

Mann C. 1990. Meta-analysis in the breech. *Science* 249: 476-80.

Manning, E. W. 1992. Scientific barriers: A commentary. *Environmental Monitoring and Assessment* 20: 125-126.

McCune, B. and E.S. Menges. 1986. Quality of historical data on Midwestern old-growth forests. *American Midland Naturalist* 116: 163-172.

Medawar, P. 1969. Induction and Intuition in Scientific Thought. Methuen, London.

Medychyj-Scott, D., I. Newman, C. Ruggles, and D. Walker. (eds) 1991. Metadata in the Geosciences. Group D Publications Ltd. Loughborough, UK. 233p.

Mesley, R.J., W.D. Pocklington, and R.F. Walker. 1991. Analytical quality assurance - A review. *Analyst* 116: 975-990.

Michener, W.K., J.W. Brunt, J.J. Helly, T.B. Kirchner, and S.G. Stafford. 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7(1): 330-342.

Ministry of Environment, Lands and Parks (1997): BC Environmental Monitoring System: Using EMS. Release 1.3, April 1, 1997.

Ministry of Environment, Lands and Parks (1996): British Columbia Field Sampling Manual, 1996 Edition (Permittee). Queen's Printer, Victoria.

Ministry of Environment, Lands and Parks (MELP). 1995. Environmental Monitoring System Requirements Summary. LGS Group Inc. Victoria, B.C. June 30, 1995.

Montgomery, D.C. 1984. Design and Analysis of Experiments, Second Edition. John Wiley & Sons. Toronto. 538p.

Moore, P. W. 1975. Public Decision-Making and Resource Management: A Review. University of Toronto, Toronto. 72p.

Moss, P. 1994. Research design in feminist geographic analysis. *Great Lakes Geographer* 1(1): 31-45.

National Research Council, 1984. Toxicity Testing: Strategies to Determine Needs and Priorities. National Academy Press. Washington, D.C. 382p.

National Research Council (USA). 1995. Preserving Scientific Data on our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources. National Academy Press. Washington, D.C. 67p.

Pawluk, M. 1998. Personal Communication. Environmental Information Research Group, University of Victoria. B.C.

Peppin, N. "EMS QA/QC Phase III of EMS" e-mail message to Malcolm Clark, November 20, 1996 .

Peterman, R.M. 1990. Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences* 47: 1-15.

Porter, P.S., R.C. Ward, and H.F. Bell. 1988. The detection limit. ACS, *Environmental Science & Technology* 22(8): 856-861.

Puget Sound Estuary Program. 1991. Recommended Guidelines for Conducting Laboratory Bioassays on Puget Sound Sediments. Prepared for U.S. Environmental Protection Agency, Region 10, Office of Puget Sound. Seattle, Washington. 82p.

Rao, S.T., J.K. Ku, and K.S. Rao. 1991. Analysis of toxic air contaminant data containing concentrations below the limit of detection. *Journal of the Air and Waste Management Association* 41: 442-448.

Ratynski, R.A., and L. de March. 1988. Northwest Passage and Queen Elizabeth Islands: Biological oceanography -- Fish. 1819 to 1985. *Canadian Data Report on Hydrography and Ocean Sciences* (5)17: 423p.

Ratynski, R.A., L. de March and B.D. Smiley. 1988. Beaufort Sea: biological oceanography -- Fish. 1896 to 1985. *Canadian Data Report on Hydrography and Ocean Sciences* 5(15): Part 1, 301p; Part 2, 290p.

Reckhow, K. H. 1994. Importance of scientific uncertainty in decision making. *Environmental Management* 18(2): 161-166.

Roos, L.L., S.M. Sharp, and A. Wajda. 1989. Assessing data quality: A computerized approach. *Social Sciences Medicine* 28: 175-182.

Roots, E. F. 1992. Environmental information - A step to knowledge and understanding. *Environmental Monitoring and Assessment* 20: 87-94.

Rourke, B. 1991. Towards openness and fairness in the review process. *Behavioral and Brain Sciences*. 14: 161.

Royal Society Analytical Methods Committee. 1987. Recommendations for the definition, estimation and use of the detection limit. *Analyst* 112: 199-204.

Samli, A. C. 1996. Information-Driven Marketing Decisions: Development of Strategic Information Systems. Quorum Books. Westport. 196p.

Scheiner, S.M. and J. Gurevitch (eds.) 1993. Design and Analysis of Ecological Experiments. Chapman & Hall, New York. 445pp.

Schneider, H. 1986. Truncated and Censored Samples from Normal Populations. Marcel Dekker, Inc. New York, 273p.

Smiley, B. 1988. Personal Communication. Marine Advisor, Data Assessment. Institute of Ocean Sciences. Sidney, B.C.

Sowden, T. <SowdenT@pac.dfo-mpo.gc.ca> "RE: NCIS" Private e-mail correspondence with Blair King. March 24, 1999.

Spellerberg, I.F. 1993. Monitoring Ecological Change. Cambridge University Press. Cambridge. 334p.

Stafford, S.G. 1993. Editorial. *Environmental Monitoring and assessment* 26: 85-89

Stafford, S. G. 1994. Data, data everywhere but not a byte to read: Managing monitoring information. *Environmental Monitoring and Assessment* 26: 125-141.

Stafford, S.G., J.W. Brunt and W.K. Michener. 1994. Integration of scientific information management and environmental research. In: Michener,W.K., J.W. Brunt, and S. Stafford (eds) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis Ltd, London. 555p.

Stafford, S.G., P.B. Alabeck, K.L. Waddell, and R.L. Slagle. 1986. Data management procedures in ecological research. In: Michener, W.K. (ed.) Research Data Management in the Ecological Sciences. Belle. W. Baruch Library in Marine Sciences Number 16. University of South Carolina Press. 419p.

Staeheli, L.A. and V.A. Lawson. 1994. A discussion of "Women in the Field": The politics of feminist fieldwork. *Professional Geographer* 46(1): 96-102.

Stonehouse J.M., and J.D. Mumford. 1994. Science, Risk Analysis and Environmental Policy Decisions, UNEP Environment & Trade Monograph Series. Booklet no. 5. United Nations Environment Programme, Geneva. 79p.

Strain P., and T. Sowden. 1995. Guidelines for Creating C.I.S. "Events": How to Organize Data for Entry Into the CIS Inventory. Department of Fisheries and Oceans, Sidney, B.C. 17p.

Swain, L.G. and D.G. Walton. 1989. Fraser River Estuary Monitoring, Report on the 1988 Fish Monitoring Program. Fraser River Harbour Commission, B.C. Ministry of the Environment. 147pp.

Swain, L.G. and D.G. Walton. 1990. Report on the 1989 Boundary Bay Monitoring Program, Fraser River Harbour Commission, B.C. Ministry of the Environment. 172p

Tagatz, M.E., J.M. Ivey, N.R. Gregory, and J.L. Ogelsby, 1982. Effects of pentachlorophenol on field- and laboratory-developed estuarine benthic communities. *Bulletin of Environmental Contamination and Toxicology* 26: 137-142.

Thomas, D.J. 1992. Considerations in the Design of Effects Monitoring Strategies for the Beaufort Sea. Beaufort Sea Case Study. Environmental Studies Research Funds Report No. 118. Calgary. 54pp.

Toft C.A., and P.J. Shea. 1983. Detecting community-wide patterns: Estimating power strengthens statistical inference. *American Naturalist.* 122: 618-25.

Travis, C.C., and M.L. Land. 1990. Estimating the mean of data sets with nondetectable values. *Environmental Science & Technology* 24(7): 961-962.

Trembley, G. <TremblayGH@dfo-mpo.gc.ca> "RE: 1996 Workshop Protocols for the Appraisal of Contaminant Survey and Experimental Events" Private e-mail correspondence with Blair King. May 7, 1998.

van Beresteyn, E.C.H., M.A. van't Hof, H de Waard, P.R. Dekker, R. Neeter, H.J. Winkeldermaat, R.M. Visser, G. Schaafsa, M. van Schiak and S.A. Duursma. 1986. Design and data quality of a mixed longitudinal study to elucidate the role of dietary calcium and phosphorus on bone mineralization in pre-, peri-, and postmenopausal women. *American Journal of Clinical Nutrition* 43: 538-548.

Vromans, A. <Vromans@mar.dfo-mpo.gc.ca> "appraisal document" Private e-mail correspondence with Blair King. April 2, 1998.

Wainwright, P. 1992 PC ODIS Data Documentation. Report Prepared for Brian Smiley, Data Assessment Division, Department of Fisheries and Oceans. Sidney, B.C. 127p.

Walker, D. 1991. Introduction In: Medychyj-Scott, D., I. Newman, C. Ruggles, and D. Walker (eds) Metadata in the Geosciences. Group D Publications Ltd. Loughborough, UK. 233p.

Ward, M.O., W.L. Power, and P. Ketelaar. 1996. A computational environment for the management, processing, and analysis of geological data. *Computers and Geosciences* 22(10): 1123-1131.

West, P.R, T.M. Fyles and B. King. 1993. Environmental data management using meta-data: A meta-data survey of organic contaminants in the Fraser Basin. In: Farrell, A.P. (ed.) The Aquatic Resources Research Project: Towards Environmental Risk Assessment and Management of the Fraser River Basin. November 1993.pp. 379-451. 451pp.

Williams, G. 1976. Comparing the joint agreement of several raters with another rater. *Biometrics* 32: 619-627.

Wright, R.D. 1970. Truth and its keeper. *New Scientist* 45: 402-404.

Zelt, K.A., and H.F. Clifford. 1972. Assessment of two mesh sizes for interpreting life cycles, standing crop, and percentage composition of stream insects. *Freshwater Biology* 2: 259-269

Zentall, T.R. 1991. What to do about peer review: Is the cure worse than the disease? *Behavioral and Brain Sciences* 14: 166-167.

Zhao, S. 1991. Metatheory, metamethod, meta-data-analysis: What, why, and how? *Sociological Perspectives* 34: 377-390.

Zumdahl, S.S. 1986. Chemistry. D.C. Heath Company, Toronto, 1023p.

Appendix: Detailed Results from Data Truncation Study

Survey text

IMPLICATIONS OF DATA TRUNCATION IN DATA MANAGEMENT

The Environmental Information Research Group at the University of Victoria is currently involved in a research project that addresses data truncation and its consequences in environmental analysis. Your laboratory has come to our attention via one of the following: U.S. Environmental Protection Agency, Washington State Department of Ecology, Oregon State Department of Human Resources - Health Division, Canadian Association for Environmental Analytical Laboratories, National Association of Testing Authorities (AUS), Testing Laboratory Registration Council of New Zealand, New Zealand Biotechnology Association, International Laboratory Accreditation Co-operation, and International Council for the Exploration of the Sea. We would greatly appreciate a brief moment of your time to answer a few questions.

Data truncation ("rounding off" of values) is widely practiced by analytical environmental laboratories. It is based on a rich scientific tradition of reporting "significant figures" in reports, publications, and presentations throughout the scientific community. Traditionally, only those values which are found to exceed the uncertainty that the measurement process imposes are reported, together with their "error bars". In contrast, some powerful statistical tools like *principle components analysis* work best if untruncated data are used. Frequently much of the data used to detect trends in environmental parameters is very close to the detection limits of the analyses. We are looking at all aspects of data truncation to explore the impacts data reporting standards have on the types of environmental management decisions which can be made based on the data. Part of this project involves a survey of current practices at the laboratory level.

We would basically like to know what happens to the "raw data" in your lab as it makes its way to your client. The following questions should only take a few moments to answer and will help us resolve that issue. In order to analyze the results of the survey within a reasonable time frame, we would appreciate a response within one week. You can either Fax your response back to us (250-721-7147), please just circle the

appropriate answer (e.g., Yes / No), or E-mail your feedback (mpawluk@uvic.ca). Please feel free to attach additional comments as you wish. While the results of this survey will be used in a final report, individuals and laboratories will not be directly identified. Thank you in advance for your time and consideration.

1) First of all, if you could tell us something about yourself: Your position/title:

2) Is your data usually reported in truncated form to a number of significant digits? Yes/No (For example, truncating the value 123.456 to 123.4) What is the basis of your truncation?

3) Is your data usually reported as rounded values? Yes/No
(For example, rounding the value 123.456 to 123.5). If so, how is the rounding done?

4) Do your clients ever request that the data be given to them as "raw data plus uncertainty"? (i.e., as untruncated, unrounded data) Yes/No
If yes, do they state why they need the data in this form?
Also if yes, do you supply them with the "raw data plus uncertainty"? Yes/No

5) How do you report "non detects"? (e.g. zero values, ND, value < detection limits, etc.)

6) Does your lab report values that are "detected but below the limit of quantification"? Yes/No How are these represented? (e.g. number, <QL, etc.)

7) From an environmental research position, some statistical tools would be better served if untruncated data were used? This would involve your lab reporting "raw" data and data below the quantification limit. What do you think of this as a practice for an environmental laboratory?

Survey Numerical Results

2. Is your data usually reported in truncated form to a number of significant digits? Yes/No (For example, truncating the value 123.456 to 123.4)

Total number of responses to the question: 57
Yes 33/57 (58%)
No 21/57 (37%)
response omitted 3/57 (5%)

What is the basis of your truncation?

- 59% (or 23/39) of the responses to this question truncate based on significant figures.
- 28% (or 11/39) of the responses to this question truncate based on protocols associated with the methodology and/or analysis.
- 5% (or 2/39) of the responses to this question truncate based on report format and neatness.
- 8% (or 3/39) of the responses to this question did not state the basis of truncation. Note - the number of responses differ from that indicated in the first part of the question because several respondents gave more than one answer, depending upon the intent of the data, and also because not everyone that answered "What is the basis of your truncation?" responded with a "Yes" in the previous part.

3. Is your data usually reported as rounded values. Yes / No (For example, rounding the value 123.456 to 123.5)

Total number of responses to the question: 57
Yes 51/57 (89%)
No 3/57 (5%)
response omitted 3/57 (5%)

If so, how is the rounding done?

- 77% (or 41/53) of the responses to this question utilize some form of algorithm.
- 23% (or 12/53) of the responses to this question actually describe truncation rather than rounding. Note - not everyone that answered "how is the rounding done?" responded with a "Yes" in the previous part.

4. Do your clients ever request that the data be given to them as "raw data plus uncertainty"? (i.e., as untruncated, unrounded data) Yes / No

Total number of responses to the question: 57
Yes 13/57 (23%)
No 41/57 (72%)
response omitted 3/57 (5%)

If yes, do they state why they need the data in this form?

Total number of responses to the question: 13

- 46% (or 6/13) of the responses to this question state that they need the data in this form for statistical reasons (e.g., for statistical software, trend analysis, information purposes, etc.).
- 31% (or 4/13) of the responses to this question did not state why they needed the data in this form.
- 23% (or 3/13) of the responses to this question stated that data in this form was the routine reporting method (e.g., oil industry, radiological, etc.).

Also if yes, do you supply them with the "raw data plus uncertainty"? Yes / No

Total number of responses to the question: 18
Yes 11/18 (61%)
No 6/18 (33%)
response omitted 1/18 (6%)

Note - not everyone that answered this question responded with a "Yes" in the initial part of question 4.

5. How do you report "non-detects"? (e.g., zero values, ND, value < detection limits, etc.)

- 58% (or 42/73) of the responses to this question utilize "<", (e.g., <LOD, <LOQ, <MDL, <reporting limit, <W, <T, etc.)
- 25% (or 18/73) of the responses to this question utilize values, (e.g., zero, 0±dl, dl, number, flagged value, etc.)
- 18% (or 13/73) of the responses to this question utilize text, (e.g., ND or n.d., nil, null, etc.)
- Note - the number of responses to this question is higher than the total number of survey responses since several of the respondents gave more than one answer, depending upon the type of analysis.

6. Does your lab report values that are "detected but below the limit of quantification".Yes / No

Total number of responses to the question: 57
Yes 29/57 (51%)
No 27/57 (47%)
response omitted 1/57 (2%)

How are these represented? (e.g., number, <QL, etc.)

- 60% (or 21/35) of the responses to this question utilize values, (e.g., number, 0±dl, value with LOQ or DL, value plus uncertainty, flagged values or values in brackets, values are footnoted, etc.)

- 23% (or 8/35) of the responses to this question utilize "<", (e.g., <LOQ, <DL, <MDL, <T, <TE, etc.)
- 17% (or 6/35) of the responses to this question utilize text, (e.g., ND, TR, T, Presence/Absence, ND/MDL, etc.)

Note - the number of responses are higher than the initial *Yes*'s since several of the respondents gave more than one answer, depending upon the type of analysis (e.g., nutrients, metals, organics, etc.)

7. From an environmental research position, some statistical tools would be better served if untruncated data were used. This would involve your lab reporting "raw" data and data below the quantification limit. What do you think of this as a practice for an environmental laboratory?

We divided the textual responses into three sections: responses that "generally support" the practice of reporting raw data and data below the quantification limit, "neutral" opinion or responses which weighed both pro and con positions, and "generally opposed" to the practice of reporting raw data and data below the quantification limit. The overall results are as follows:

Total number of responses to the question:	57
generally supportive	14/57 (25%)
neutral	20/57 (35%)
generally opposed	21/57 (37%)
response omitted	2/57 (3%)

Truncation Survey Detailed Textual Responses

A/ Further survey comments (ques. 4) regarding supplying "raw" data plus uncertainty.

"Very difficult to do as "uncertainty" varies depending upon the level of analyte. Near detection limit uncertainties will be generally much greater than at higher levels."

"We supply raw data and replicates. Client can calculate uncertainty."

"In the form of results for QC samples."

"Supply with uncertainty if available and if requested, some clients only want raw data without uncertainty. Original/raw data is not altered and available within the provincial Laboratory Information Management System (LIMS)."

"Data we report is not necessarily given as raw data and there is no indication of the uncertainty in each value except by comparison to the MDL or duplicate results. We do save data in our LIMS system in an "unadjusted" form such as soils and sediments in an "as received" basis, so that the results are not biased by calculations using rounded data."

"Clients must agree with investigator that the data only be used for research and not for distribution."

"Never been asked to do so, most clients barely understand significant figures."

"Clients to date have not requested data in a particular form. Generally we specify. Almost ALL data that goes out of our lab contains uncertainty values with it to emphasize the significance of the data."

"We do not really have any clients outside the institute, and the uncertainty is known by the personnel. We would supply them with "raw data plus uncertainty" if someone asked."

"Our clients usually do not request that the data are given as "raw data plus uncertainty". We consider the data supplied by us as raw data, since the rounding procedure in our laboratory does not intend to remove the uncertainty of the method, but rather to maintain the "uncertainty level" introduced by weighting procedure. We always supply our clients with a document showing the uncertainty, which is based on relative deviation in accumulated results of the homogenised seal blubber sample (in "house control sample") that is analyzed with every series of environmental samples."

"The accuracy of the analytical method is described in the report (i.e. the performance concerning Certified Reference Material). The estimated geometrical mean values, values for last year, slopes, power, etc. are given with a 95% confidence interval."

"Clients are usually only interested in a single number as a result. We have never been asked for uncertainty data. Some clients ask for limits of detection (LoD) data."

B/ Further comments (ques. 6) regarding reporting values that are "detected but below the limit of quantification".

"Very controversial area. What's the DL? Our DL's are perhaps a bit conservative, i.e. higher than the true DL."

"No special ID of these values are made, by definition our LOQ is 3x LOD. Client can estimate number below LOQ by knowing LOD."

"<T indicates "target substance identified, quantity of substance in this sample is approximate, use caution in interpretation unless more sample data supports this result". <TE (multiple of T) indicates "target substance identified, quantity of substance in this sample is approximate following non-routine dilution of the sample to allow analysis for the target substance, use caution in interpretation unless more sample data supports this result"."

"Such data are provided only upon special request and we require the client sign a waiver acknowledging they understand the normal quality management criteria (lab is formally accredited) do not apply and that the measurement uncertainty is not known. Clients are cautioned against entering such data into an unqualified database."

"Only if specifically requested by the client. Such values are footnoted as "estimated, below the MRL but above the MDL". Usually requested for PCB's or pesticides where pattern recognition or second column confirms."

"We do report <QL for certain specialized applications. Generally these points are estmated levels based on the lowest quant point and then reported as 0 +/-DL. Often this may look like 0.12 +/- 0.23. This way the recipient is shown that the lower level of the

uncertainty would be negative, so the recipient understands the significant uncertainty associated with that data point."

"The laboratories can report values below the limit of detection/quantification. They do so by reporting the observed value or the detecting/quantification limit, and flagging the value, thereby indicating that this is a "less than" value."

"We report the detection limits as well as the uncertainty of the variables, but give the actual values (rounded)."

"To our clients we have stated that the detection limit is defined as three times the standard deviation of a zero sample, or a close to zero sample. They have also been informed that the relative uncertainty is strongly increasing down to the DL."

"Trace of a compound is generally noted in the analysis form. In practice however, they are treated as "non-detects", i.e. we call both groups "below detections limit"."

"This is only ever done in interlaboratory comparison exercises where results which are quantifiable but below the limit of detection are sometimes required, to check on bias or blank contamination."

Truncation Survey Detailed Opinion Responses

Most of the other survey questions generated Yes/No type responses with minimal text. The combined responses for those questions are presented in the body of the report. However, the following question was fundamental to the overall survey, and we have chosen representative unabridged opinions that appear to embody the average response to the statement in the positive, neutral, and negative fractions.

7. From an environmental research position, some statistical tools would be better served if untruncated data were used. This would involve your lab reporting "raw" data and data

below the quantification limit. What do you think of this as a practice for an environmental laboratory?

The following responses have been sorted sectionally according to:

A) Supports the practice of reporting raw data and data below the quantification limit

B) Neutral opinion

C) Opposed to the practice of reporting raw data and data below the quantification limit

Section "A"

"This is ok as long as there is a guarantee that the result along with all it's inter-relational data remains intact, i.e. result, units, detection limit, uncertainty, etc."

"I think it is a good idea if we have "educated" clients who understand the concept of uncertainty."

"We would be happy to do it for those clients who requested it and who have the sophistication to use it. Most of our clients do not understand statistical "niceties", although the government agencies who ultimately review the data may."

"I think it is the practice I would prefer, partly because it would get rid of the extra work done to ensure numbers are not reported below detection limits, and that is done to ensure the significant figures are correct. However, this practice may lead to confusion as the integrity of our lab has been challenged by people unfamiliar with this practice when we have used it. Some people think it is unethical to report numbers below the detection limit or to report more significant figures than they think is justified by the accuracy of the test."

"Modern statistical techniques have been shown to be capable of usefully handling untruncated data and we have been involved with programs that have successfully done this. The laboratory needs to be involved with the planning and interpretation aspects of such projects and users of such data are cautioned about entering untruncated data into a database where the values can possibly be presumed to be "absolutely" correct."

"Laboratory methodology can be adjusted to "Bracket" expected presence values."

"If requested, I would have no problem reporting such data."

"I would be cautious about reporting any value less than the MDL since the data user may not always realize the high level of uncertainty in this data. Reported values less than the MDL must be clearly flagged."

"This practice for an environmental laboratory would be acceptable if it enhances the data to be processed."

"Cumbersome but I can see the value of using untruncated values if you intend to average them with a number of data points."

"I think the practice of reporting untruncated data is ok. (Within reason, you have to realize that my mass spectrometer will report data to 15 decimal places--nothing beyond the 4th place has ANY logical value). I do feel that it is essential that people to whom these data are given are also given an explanation of the statistical validity of the data. For this reason, almost ALL data that goes out of our lab contains uncertainty values with it to emphasize the significance of the data."

"Good idea, but it must be underlined which data are below detection limits."

"I would recommend such a practice, as we never use the "limit of quantification". To our clients we have stated that the detection limit is defined as three times the standard

deviation of a zero sample, or a close to zero sample. They have also been informed that the relative uncertainty is strongly increasing down to the DL."

"I think this is a good idea. I have requested to get the raw data in this form since 1988. Before that year data is sometimes truncated causing some problems in the statistical evaluation."

Section "B"

"If can be showed that some statistical tools would be better served, I agree. Operationally would be difficult to implement as not all data would be gathered and reported in this way. Communicating this during analysis would be hard. There is also danger in reporting lots of figures as uninformed clients would likely assume significance to such values that does not exist."

"No comment, depends on customers requirements."

"This could be done if customers request it. So far, there is not much call for it. We have experienced problems in the past where researchers who are unfamiliar with statistical principles have used analytical data to calculate and report numbers such as 273.4961 where only one or two significant figures belonged. Some went so far as to find differences between this and 273.4922! In such cases, we encourage poor science by reporting too many figures in the first place."

"Pro's and con's, depends on how your data is used and interpreted. Data below detection limit are just numbers generated mathematically or electronically. Large errors and uncertainties below detection limit, depend mostly on instrumentation capability and methodology. Nowadays computers can generate lots of numbers, but are they realistic below instrument detection limit? Sometimes the number differences are great but in reality they are not. That depends on how you interpret your data. At a certain level it is

significant but not the same all the time. Every lab seems to take pride to report the lowest detection limit....." (last line of fax not readable)

"No consolidated opinion available. Depends upon instrument feed data (uncertainty, raw data) and frontend real time QC to data."

"I am well aware of the implications of left-censoring data, and the problem it causes with mathematical processing, e.g. when trying to determine unbiased means. However, I am not comfortable with releasing uncensored data, particularly negative numbers, although an unbiased measurement of a true blank should contain as many numbers as positive numbers. Perhaps the best compromise is to report censored data as a normal matter of routine, but to provide uncensored data on request, with the stipulation that only the results of the statistical analysis, and not the raw data, be revealed without prior authorization from the laboratory."

"Indifferent except where there might be software or other constraints and as long as the client understands the unreliability of the added digits."

"It would depend on what the data would be used for and if the user was aware of the instrument detection limits."

"For my data and purposes, reporting "raw" data would have little impact."

"My laboratory knowledge and experience is limited as described about. I do not have an opinion."

"Depends on time involved and if the client is willing to pay more if need be."

"Either way would be fine; would always prefer accurate quantification, but qualification would seem to be pertinent in some situations."

"For research, fine. For paying clients, it is unnecessary and potentially confusing."

"Again, it depends on the purpose of the investigation. Maybe it would be better to have untruncated data in a database, but flag those lower than the detection limit."

"I think that the observed values should never be truncated, but that only the significant digits should be shown. The benefit of extra digits in principal component analysis and other statistical models is purely artificial. The uncertainty of the values should be taken into account! Normally it is greater than the difference between the "raw" value and rounded value. Truncation may cause false results, because it always reduces the actual value, but rounding is ok."

"We always work with raw data. When data entry in the database, for each parameter, a number of significant digits is recommended. But there is no control if people enter more. All this is an old problem."

"I can not imagine, that rounding will falsify the data in an important way (at least in our work!). First, we have such a high natural variability in the observed ecosystem (i.e. Wadden Sea), that the difference between untruncated data and rounded data is neglectable. Second, the analysis itself is in consequence of a lot of possible sources of mistakes (from sampling, to preparing, to the analysis of samples) more inexact than the difference between truncated and rounded data."

"Parameter dependant. Statistical tools should cut the rounded values to the number of significant digits. Main result is that some scientists report as many digits as displayed by their tool, others report the number of significant digits and others only report the values with significant digits. In our database we archive together with the value also the number of significant digits. Another problem are values < limit of detection, either detected as values or traces; how to use these numbers always is a problem; if we couldn't avoid calculating with these values, I already took 1/5 of the detection limit value, which

appears much better than just to neglect that value leading to overestimating means; non-parametrics also sometimes helped."

"The only problem I see with this is that it may have some implication on existing data bases and reporting routines."

"I think that we have a sensible compromise in our practice. We are not trying to remove the variance or uncertainty of the method, and we are not trying to accumulate as many digits as possible. One aspect of uncertainty is recovery. We do not correct our data for recovery %, but this information is given to the client. Informing the client about practice in these issues is necessary."

"The answer is not as simple as saying only whether data should be truncated or not. Data should always be used at the level of accuracy which is justifiable, according to the analytical method used and the purpose of analysing the data. Following on from the examples I have given above, salinity would always be used to at least 2 decimal places if this has been generated using a high accuracy salinometer, and never truncated, since this would result in the loss of important information. If however, it was generated using a portable field instrument with lower accuracy, then reporting to one decimal place only may be appropriate. Similarly with zinc in sediments or biota, levels of 235 milligrams per kilogram should be compared with data of 268 milligram per kilogram. More precise data would not be justified; less precise data would result in the loss of information.
The key factor is the analytical method used to generate the data. Some methods are more accurate, and more precise than others. Data must be used only to the levels of accuracy and precision which can be justified.

Sample size is another important criterion. For example, how much water has been filtered to obtain a particular value for, say, chlorophyll a in water? How many mussels have been batched together as a homogenised mixture of tissue for metals or organochlorines analysis? How much sediment has been digested prior to trace metals

determination? Each of these factors affects the confidence level or uncertainty with which a result can be quoted."

Section "C"

"Do not at all agree with first sentences and with concept. It is completely incorrect to draw conclusions on data that has a high uncertainty. Our acceptance criteria for duplicate analysis at levels up to 5x the DL is | ±DL |. If the DL is 1, then duplicates of (1 and 2), (2 and 3), etc. meet our criteria. Therefore, if a result of 0.7638 were obtained, its uncertainty would be perhaps ±1, so reporting 0.7, 0.8, 0.76 is real silly."

"Reporting raw data is not a major problem provided that the environmental laboratory provides to the client documentation which indicates the accuracy of the measurements made. Then it is up to the clients to truncate the data as they see fit. There are many uncertainties in reporting data below the quantification limit. When dealing with complex instrumentation like HRGC/HRMS and complex sample matrices, the limit of quantitation can vary substantially between samples and from day to day due to the many variables involved. Therefore, reporting data below the quantification limit can be inaccurate and imprecise."

"Dangerous because of variation at the detection limit can actually yield negative results. Negative results are open to misrepresentation as are values below the detection limit. When does it stop? How many decimal points? When the variance happens at the second significant figure, what use is another value?"

"Data below the quantification limit is highly unreliable and we would not (and are not) confident in this data. Our clients are advised of this when we report trace data. "Raw data" reported with no regard to significant figures may imply greater analytical precision that is really there."

"Most commercial environmental lab clients don't know the difference between limit of quantification and detection. Most clients would be mislead/confused by reporting untruncated data, they wouldn't know how to evaluate the data."

"This would add more data manipulation and data entry into our reports. Our clients may be tempted to "read in" significance into data reported below quantification limit. Lot of toxicity limits have been based on calculations rather than by actual studies. Some of the values in the current guidelines (limits) are unrealistically low."

"Impractical for an environmental lab (non-research). The number of figures recorded for a measured quantity must reflect the precision with which the measurement was made. I'm sure the above practice would be suitable for a research lab."

"Not good. All data are rounded at some point. I think it is naive to think that just because the final result is not rounded that it is more useful than the rounded result. e.g. balances read to say 3 decimal places (4th place is rounded by balance), instruments (like AA/ICP) will round results as well, based on calibration curve. So what is "raw data" anyway?!"

"Absolutely dead against the concept. From years of real world experience, we have realized that while "pure statistics" if used and interpreted properly would benefit, the *average* client is not properly educated and data gets misused and abused when presented in this form. One issue which is missed by the survey is that often data is rounded or truncated by the instrument itself and by the time the analyst sees a 17 decimal "raw" value, it really isn't. e.g. $1.0 \div 7 = 0.142857142857142857....$"

"Raw data requires some "processing" of those numbers. If not, information is lost over time and hence becomes a data management issue."

"All data would have to be truncated at some point. It would be a reporting nightmare unless the data was transferred electronically (by floppydisk, or internet)."

"This is occasionally done for method detection limit studies. However, I see no reason to report environmental data to more than one figure beyond the significant figures."

"Opens up a "can of worms"."

"One of the reasons we decided to limit our significant figures was that doing so reduced the number of trivial exceedences of our WQ standards. These incidents were identified by statistical tests but when investigated, proved to be inconsequential."

"As a standard practice it would not provide our clients with useful data. It would only be appropriate to provide such "raw" data to someone knowledgeable in its interpretation. Providing such data would require going outside automated data systems and would therefore be very time consuming, and of limited use as clients are going more towards requesting electronic data deliverables with standardized formats."

"Dangerous since, i) error involved in testing, sampling, etc. is too great to place any meaning on untruncated or unrounded numbers, ii) reporting below detection limits can be legally dangerous - third party lawsuits, etc.."

"The uncertainty of such values (<QL) becomes too great (±100%)."

"This can lead to misinterpretation by administrators. Under these conditions a salient point might be missed, that better analytical sensitivity, precision, accuracy and quality control are needed."

"Not a good idea because: a) Modern analytical instruments can produce data to 4 decimal places. In many cases the 3rd and 4th decimal places are meaningless and beyond the limits of accuracy for the analytical procedure. b) Raw data from analytical instruments, below the quantification limit, can sometimes appear as a low negative value e.g. -0.0001."

"Raw data reporting would be fine. I would not be satisfied with supplying values below quantification limit as this would increase work load significantly."

"I do not think reporting untruncated results to satisfy a statistical tool is a good idea. The last digit in each numerical value is normally deemed significant. If results were untruncated, each result would have to be accompanied by a statement of uncertainty. I would object to reporting data below the quantification limit as the associated errors and uncertainty are too high."

"Not happy! What is "raw" data? Example: Absorbance on a spectro 0.003, blank 0.001, calibration curve 0 - 0.001, 5 - 0.005, 10 - 0.010. Sample calibration program on computer gives a result of 2.974..... What to report? We would report "3". Any variation due to minor changes in calibration from day to day, errors in standard prep., etc., will influence the result."