Data Cleansing and Transformation of Observational Scientific Data: A Case Study

L. Singh, G. Nelson, J. Mann, A. Coakes, E. Krzyszczyk, and E. Herman. "Data cleansing and transformation of observational scientific data." ACM SIGMOD Workshop on Information Quality in Information Systems. Chicago, Illinois: ACM, 2006.

Data Cleansing & Transformation of Observational Scientific Data: A Case Study

Lisa Singh, Gregory Nelson
Georgetown University
Department of Computer Science

singh@cs.georgetown.edu

Janet Mann, Amanda Coakes,
Ewa Krzyszczyk, Elia Herman
Georgetown University
Department of Biology

mannj2@georgetown.edu

ABSTRACT

This paper investigates information quality as it pertains to observational scientific data. Specifically, we focus on presenting a case study for initial efforts related to data cleansing and data transformation of 20 years of behavioral, reproductive, demographic and ecological data on wild bottlenose dolphins located in Shark Bay, Australia. The Shark Bay dolphin population has been monitored annually by researchers since 1984 with over 13,400 surveys of dolphin groups, several thousand hours of focal follow data on individuals, and large stores of film data on both groups and individuals. It is the most comprehensive dolphin data set in research today. However, the data is inconsistent because of changing standards, variations in researcher methodology, missing data and data entry errors. To add to the difficulty, the data is scattered across multiple applications and data repositories. One of the goals of the researchers involved is to integrate the data into a single repository so it can be used for sophisticated data analysis and manual data merging can be eliminated from the data analysis procedure. After presenting our data modeling, cleansing and integration process in the context of the Shark Bay data set, we introduce a set of quality metrics specific to observational science data and used them to assess the information quality of the wild bottlenose dolphin data before and after the data cleaning and validation procedure. [1]

1. INTRODUCTION

Many large businesses and government agencies maintain enterprise wide database systems that incorporate integrity constraints, well-defined data models and data manipulation procedures. While numerous quality issues arise in these data stores [4][9], scientific data sets also present challenges, including inconsistent data recording procedures, measurement subjectivity, data entry errors, scattered data repositories, and variation in researcher data collection methodology. Some of these issues are the same ones that existed for large businesses and government agencies a decade ago. With the increased volume of research data across all disciplines, these issues need to be revisited in the context of scientific data analysis. This paper investigates information quality as it pertains to observational scientific data. We begin by presenting a case study describing our experience of modeling, cleaning, and transforming observational data on wild bottlenose dolphins. We then introduce some quality metrics specific to the domain and use them to evaluate the information quality of the wild bottlenose dolphin data before and after data cleaning and validation.

2. SHARK BAY DATABASE CREATION

2.1 Data Set Background

The Shark Bay, Australia dolphin population has been monitored annually by researchers since 1984 with over 13,400 surveys of dolphin groups, several thousand hours of focal follow data on individuals, and large stores of film data on both groups and individuals. Shark Bay has approximately 13,000km^2 of shallow clear water, few vessels, and a large dolphin population in the thousands [6]. It is the most comprehensive dolphin data set in research today with over 20 years of behavioral, reproductive, demographic and ecological data on wild bottlenose dolphins.

An international team of twelve researchers and numerous research assistants monitor the dolphins and are members of the Shark Bay Dolphin Research Project (SBDRP). The researchers developed Shark Bay Dolphin Project Guidelines and a Data Protocol Handbook that lay out expectations and procedures for data collection, contribution, authorship, and responsibilities. Researchers submit their data within three months after fieldwork is completed.

There are three types of data collected: survey, focal follow and GIS spatial data. First, extensive surveys of all animals in the study area (~300km^2) are conducted. Data gathered includes location, animal behaviors, associates, habitat, photographic information, and physical data (e.g., scars, condition, speckles). Most surveys are conducted simply by looking for specific dolphins for focal observations (or biopsy darting) and surveying groups along the way. Brief surveys, lasting 5 to 10 minutes, present a "snapshot" of associations and behaviors among dolphins. The second type of data collection involves "focal follows" on individual dolphins. A focal follow is a detailed study of a dolphin lasting approximately 2 hours. In one example, Shark Bay bottlenose dolphin mothers and calves have been studied continuously since 1988. During boat-based focal follows of specific mother-calf pairs, detailed behavioral information is gathered. The mother calf data set has approximately 200 attributes and 125,000 records. Both the survey and the focal

[1] Additional Co-author Email Addresses:

Gregory Nelson – gln@georgetown.edu
Amada Coakes – akcoakes@gmail.com
Ewa Krzyszczyk – ewakrzyszczyk@googlemail.com
Elia Herman – eliaherman@gmail.com

follow data are collected using paper forms and later entered electronically. The final data type is GIS spatial data on habitat use and ranging. Several of the researchers are also collaborating in the development of a bathymetric map of the study area.

Currently, the focal data and the survey data are in two different data formats. The survey data resides in Microsoft Access in large, independent relations that are converted to spreadsheets for analysis since much of the focal and other historic data are stored in Microsoft Excel spreadsheets. The complete data set is heterogeneous, has hundreds of attributes, contains missing values, is redundant in places and is scattered across 17 data repositories. Currently, data collected during a focal follow needs to be manually 'merged' with survey data for comprehensive analysis. Ad hoc querying of the data is also not possible since the data is manipulated in text files and spreadsheets.

2.2 Data Processing Procedure

The first challenge of this project is to develop the integrated data repository. Then in order to maintain a consistent, non-redundant, accurate data repository, we need to use intelligent applications for insertion and manipulation of the data that help enforce integrity constraints and pre-defined project standards. To get from the initial situation regarding the Shark Bay data to our goal, we employed the procedure presented in Figure 1. Many of these steps including data modeling, cleansing and integrated have been studied extensively in literature. For overviews of these topics, we refer you to [7][11]. Various tools have been developed to support these tasks. Due to space limitations, we cannot list them all here. Instead we will note the ones considered at different points in the process throughout this section. We should mention that given this project is a research endeavor, cost was a large consideration and impacted the decision to develop a considerable amount of software ourselves. We will now go through each of the steps identified and describe our experience, highlighting accomplishments and set backs throughout the process.

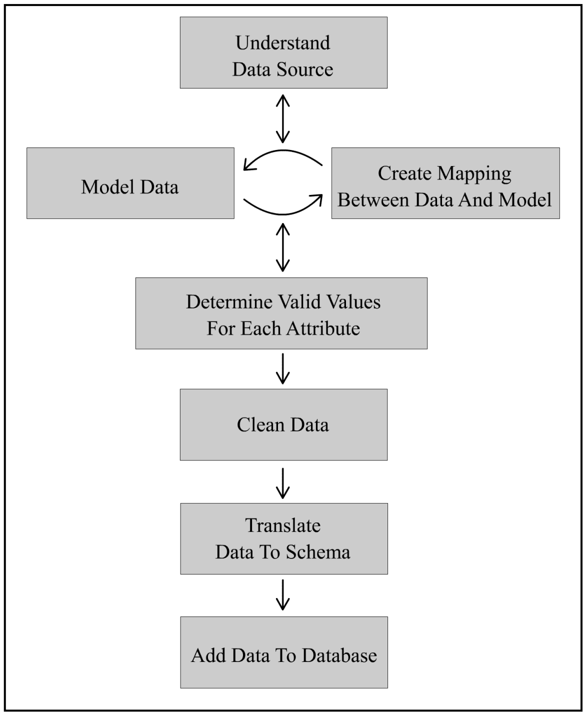

Figure 1: Data processing procedure

The first step of this process was to attempt to understand the data that existed in the SBDRP. Because there are 17 different sources of data for the integrated data repository, it is necessary to identify each source, understand the format of data stored by each source, determine the size of the source, and identify the attributes contained in each data source. The researchers on the project compiled a list of this information to help start the discussions.

We immediately noticed a few issues. First, there was a fair number of redundant attributes. Many non-key attributes existed across multiple data files. Some data files were specialized partitions of the data set used for specific analysis. Sometimes those copies of the data were more up-to-date than the original data sources. Also, because different researchers submitted surveys in different formats, one laboratory member was tasked with manually entering and standardizing the data. The sheer volume of data and the lack of manpower prevented the researchers on the project from keeping all the data up-to-date and consistent. Next, there was no data model for the data set. The data was process specific. This is not unusual for scientific data sets, particularly biological and medical data sets. Measurements were captured and stored in formats most suitable for analysis. While initial attempts at standardization had been taken, more work still needed to be done.

We should also mention that the computer scientists on the project were unfamiliar with wild bottlenose dolphins and the biologists working on the project were unfamiliar with concepts related to data modeling and database design. A communication gap was apparent from the outset. The computer scientists had to explain that changing the structure of the data did not equate to losing data values. The biologists had to explain the meaning of each value and the relationship of it to other values in the data set. Fortunately, everyone involved in the data modeling, cleaning and transformation process was passionate about the cause. We met weekly to discuss one of the data sources and identify potential important attributes. The computer scientists would then take this information and work on modeling the data elements and data set features introduced during the meetings, making adjustments to the original model if necessary. It was not unusual for the computer scientists to design a model based on one interpretation of the data seen up to that point and then change the entire model based on new information gathered at the next meeting. Every few weeks a new model would be presented to the biologists for validation. The model would typically be accompanied by numerous questions. To model all the survey data and a few additional smaller data components, took over three months, meeting once a week.

The survey component of the data is well suited for a normalized relational data model. However, the focal data is naturally multi-dimensional, so a star schema with some snowflaking for more traditional components of the data will result is faster query performance and straightforward adaptation for existing data mining applications. Because the survey data is used by more researchers than the focal follow data, we decided to begin by developing a relational model to capture the survey data. We also considered the focal data during the process to facilitate straightforward integration of it when the time comes to fully model related data sources. We focused on creating a data model that was application independent since so many applications were used by the researchers and we anticipated new ones being added in the future. We attempted to eliminate redundant information and derived attributes. When possible, we mapped to the most accurate source for a redundant attribute since data consistency across data files was an issue.

Because the data model is an integrated view of a fragmented data set, it was particularly important to make sure we captured all the necessary attributes and understood the data sources. After we initially modeled approximately half the data sources, we decided to stop modeling and begin implementing the actual database. For our initial analysis of 8 data sources, we developed a relational model containing over 200 attributes across approximately 40 tables using the ERCreator software [1]. It is an inexpensive commercial product that creates entity relationship (E/R) diagrams and generates the physical relational model from the E/R diagrams. The pictorial representation was a beneficial tool for the biologists during development.[2] A more robust tool similar to Rational Rose would have been useful for this stage, but the product we chose was reasonable and cost effective [8].

The data repository was developed using the open-source Postgres database management system developed at University of California, Berkeley on a Linux platform. Before populating data, we went through each data source associated with the model and mapped each feature to an attribute in the model. This ended up being a difficult process because many fields we were unaware of during modeling were in the data files, particularly in the case of the survey data. The initial list received did not account for historical inconsistencies or include all of the 'less used' measurements. This situation resulted because the computer scientists were unfamiliar with animal studies. We did not think to ask certain questions that a biologist would have noticed. Similarly, because the biologists had not modeled data before, they did not realize the importance of providing certain information. Viewing and mapping the raw data also allowed us to make a quick assessment about the data quality in terms of cleanliness, redundancy, and missing values.

Because the database is being created after 20 years of data has been collected and numerous protocol changes have occurred, it was necessary to integrate data validation and cleaning components into the process. By doing this, we hoped to avoid adding 'known' bad data into the database. Therefore, we designed as set of libraries that checked the data, cleaned the data when possible, transformed the data values as needed to a standard format, and inserted the transformed data into the database. [12] reports that most organizations prefer to develop in-house ETL and data cleansing tools because of the learning curve and complexity of the commercial products. In our case, the inexpensive ones we tested were reasonable, but needed much customization. Table 1 shows the different components of our cleaning and transformation library. Each component has a brief description and a list of input parameters. All the programs were written in JAVA. While the specifications of these libraries appear to be simple, the programs themselves were complicated to write. The difficulty was a direct result of the complexity of the incoming data and the different format of the final data model. Each data source stored data fields differently. Many binary fields were being converted to categorical attributes; some files contained inconsistent complex list structures; and many data types were inconsistent within single attributes, e.g. all numeric data and a dot '.'. The libraries needed to be flexible and robust enough to handle varying formats of data and general enough to validate and translate complex data types. The library took one part time student 4 months to design and implement. Our plan is to expand the libraries and integrate them with existing ETL libraries to handle updates to the database and incorporate more database operations. Two tools that we are incorporating into our process to support our developed libraries are Potter's Wheel [5] for automated data cleaning using known regular expressions and Kettle [2] for data integration of cleaned, well parsed data files.

Table 1. Data quality library components

Library Name	Description	Input Parameter Files
Data formatter	Convert a raw data file to one that can be translated into the database schema	Data file, parameter file specifying final format
Valid value checker	Verify that the data values in a data file are valid attribute values and the correct data type.	Parameter file specifying valid values for each attribute in data file
Data linker	Scan database schema to identify primary key and foreign key link requirements	Database name and password
Record translator	Takes a cleaned, validated data file, recodes the data if necessary and inserts the data into the correct tables in the database.	Output of valid value checker, database name and password

3. QUALITY METRICS FOR OBSERVATIONAL SCIENCE DATA

As we evaluate the process and results described in Section 2, we recognized a need to quantify quality for observational scientific data. Specifically, we need to define measures that assess it at each step from data collection to data entry to data use. This section introduces data quality issues specific to the domain and defines usable quality metrics that provide a means to quantify data quality during different parts of the process. It should be noted, that it is in hindsight that we realize the importance of establishing a measurable set of quality metrics that incorporate the data collection process of observation scientists.

3.1 Data Collection

Within observational sciences, protocols for collection of data vary across projects. A scientist monitors a subject for an interval of time. Example subjects include dolphins, humans, and planets. Each monitoring period can be viewed as an event consisting of a number of observations. Events include tracking a dolphin for a 30 minute period, conducting a 30 minute psychological evaluation of a person, and taking a five minute snapshot of the interaction between a planet and its moons. Each is typically recorded using one of the following methods: handwritten free-form notes, handwritten surveys, tape recordings, electronic

[2] Due to space considerations, we are unable to include the E/R diagram in the paper.

forms, digital photographs or video recordings. It is not uncommon to see thousands of handwritten observations or tape recordings about research subjects, e.g. traditional psychological evaluations of patients. In recent years, more observational data has been stored electronically, potentially in databases. However, in some cases, the data is initially captured using handwritten notes or tape recordings and then transferred to an electronic format. We now present the following quality measures associated with data collection: observation certainty, observation detail consistency, researcher vocabulary confidence and data stability.

We define *observation certainty* as the degree of confidence in the measurement itself. Did the researcher actually observe the behavior or was the behavior inferred? For example, when observing animals, some behaviors are seen first-hand – a baby being nursed by a mother. In contrast, an inferred observation is one that is not actually seen, but can be determined with a high probability of certainty based on other observations – strong circumstantial evidence exists to support the inference. An example of an inferred observation is a bite mark being recorded during an observation when an animal with a wound is seen. The actual act of biting is not observed, but the wound is such that it appears to be a bite mark. The difference between an actual observation and an inferred one is important for quality control. If a researcher is making inferences and is not identifying them as such, data analysis that relies on inferred observations may be misleading. Marking this distinction during data collection improves the understanding and degree of certainty of the data collected, thereby increasing the quality of the information.

Another important quality measure is *observation detail consistency*. Is each researcher capturing the same level of detail? Two different scientists may note the following observations of the same event:

> *Scientist 1:* A red bird is sitting on a tree eating.
>
> *Scientist 2:* A young, red robin with a damaged right wing is sitting on the lowest branch of an oak tree eating a two inch earthworm.

The more detailed the recording of the subject and the event, the closer the observation is to reality. Accurately measuring the discrepancy between the observation and reality is a difficult problem. One way to decrease the discrepancy is to develop surveys that specify the minimum amount of information needed from the observation. Because the survey is created by research participants, it is considered a reasonable approximation of reality and represents a meaningful set of data as perceived by a group of scientists. While it cannot be considered a complete reflection of reality, it does identify important features and helps standardize the level of detail across researchers.

Another quality concern involves identification of a common frame of reference or a consistent language interpretation across researchers in a group. For example, one researcher may look at a rabbit and suggest that it is large. Another researcher looking at the same rabbit may classify it as medium. We will refer to consistent interpretation among data collectors as *researcher vocabulary consistency (rvc).* If researchers have developed a survey, one approach to measuring the amount of consistency is to have each researcher use the survey to capture measurements about the same event. These surveys can be compared to calculate a quality metric, Q_{rvc}, for researcher vocabulary consistency:

$$Q_{rvc} = \frac{\sum_{i=1}^{q} nbr\_equivalent\_responses_i}{nbr\_researcher \times q}$$

Q_{rvc} is the ratio between the number of questions on the survey that the researchers have equivalent responses to and q, the total number of questions on the survey. If the ratio is particularly low, developing a document that defines a common vocabulary can improve the ratio. For example, 'large' can be specified to be 'at least twelve inches'.

Another quality issue, *data stability*, involves changes to standard data collection protocols. If changes occur frequently, data is considered unstable and may be inconsistent over time. A simple example would be if scientists' measurements were initially in inches/feet and later measurements were required to be in metric units. If the old data is not converted or the date of change is not specifically marked, then data analysis will be inaccurate. Protocol changes may result because researcher interests change, the make up of the research group changes, necessary information is missing or data is recorded inconsistently.

Table 2 presents an example stability matrix that keeps track of data changes resulting from changes in collection protocol. The columns represent the different protocol changes over time and the rows contain the features or attributes in the data set at different times. This matrix can be expanded to keep track of all changes to data over time, but it may get too large to be effective. For each new protocol or standard, changed attributes (C), new attributes (N) and removed attributes (R) are marked with corresponding letters and unchanged attributes are noted with a dash. This matrix is valuable background information prior to statistical analysis because it can help identify features that have been removed and can only be used for historical analysis, features that have been added and can only be used for future data analysis and features that have been changed and need special care during analysis.

The attributes in the matrix can be quantified to calculate the impact of a new standard on the data. If all the measures are of equal importance, then we can place a "1" in any cell containing a value and a "0" in any cell containing a dash. Then the data stability metric, Q_{ds}, can be calculated for protocol p as follows:

$$Q_{ds} = \frac{\sum_{i=1}^{n} a_{pi}}{n_p}$$

In this equation, a_{pi} represents the cell value (0 or 1) of an attribute for a specific protocol p and n_p is the total number of attributes in the database when standard p is in use. Using this measure, as the value of Q_{ds} increases, stability decreases.

Observation certainty, observation detail consistency, researcher vocabulary consistency, and data stability are all factors impacting the quality of the data. It is not unusual for computer scientist to get involved in observational scientific projects well after years of data have been accumulated and data collection procedures are well established. If this is the case, historic data collection cannot be impacted, but evaluating data using these metrics can be instrumental in improving the quality of future data.

3.2 Data Entry and Validation

Data accuracy is likely to be considered the most important quality measure. There are many definitions and measures of accuracy that have been proposed in literature. [3] defines accuracy of biological data as a combination of stability and age of data. [10] proposes using a distance function between the stored value and the actual value.[3] A simple macroscopic measure of accuracy, Q_a, is ratio between the number of correct records and the number of actual records. Sometimes this measure does not provide enough insight into the type of quality problem. For example, do the data errors exist in only a few attributes or are there sporadic errors across all the attributes? This detailed information can be determined by calculating Q_{ai} for each attribute *i* instead of for the record as a whole. Also, if the data set has a large number of attributes, it can be useful to determine the average number of errors in a record.

Many accuracy errors will be introduced during data entry after the observation is complete. We can divide data entry errors into four classes:

- Errors that result in an invalid data value. The error can be detected because it is not a valid value for a particular attribute.
- Error that result in the selection of a different valid value. The error can be detected as one based on other values for a record.
- Errors that result in the selection of another valid value that can only be detected manually by the researcher.
- Errors that result in the selection of another valid value that is plausible and cannot be detected.

Because some of the errors are detectable, validation during the data entry and data update processes is important. When a user enters a detectable invalid data value, it should not be added to the database. Instead all detectable data errors should be cleaned prior to insertion into the database. As mentioned in Section 2, numerous ETL tools focus on this aspect of data quality, particularly on the first two classes. The latter two are not easy to incorporate into an automated tool and are still open issues.

3.3 Data Usability

Data usability is extremely context dependent. Within the research domain, using observation data to learn about subjects is the ultimate goal. Learning is a somewhat vague term. Reading a large number of detailed descriptive notes provides one form of insight and data mining or statistical analysis of numeric and categorical data provides another. For scientists, both forms of observations are vital for the learning process. If a research group is more interested in computational learning verses descriptive learning, then the composition of the data set should be such that more discrete data values exist. If descriptive learning is more important, the researchers should have more free form text and raw image data. Data that is not usable should be removed.

Table 2. Data stability matrix

Attribute	Protocol 1	Protocol 2	Protocol 3
a_1	N	-	-
a_2	N	-	R
a_3	-	N	-
a_4	-	N	C

We can calculate usability, Q_u, as a weighted attribute type distribution, where the weights, w_1 and w_2, are assigned based on the needs of the researchers for discrete attributes and free-form attributes:

$$Q_u = \frac{1}{2 \times n}\left(\frac{n_{discrete}}{w_1} + \frac{n_{free-form}}{w_2}\right)$$

In the equation, n is the total number of attributes, $n_{discrete}$ is the number of discrete attributes, both binary and categorical, and $n_{free\text{-}form}$ is the number of textual, unformatted attributes. A number approaching 0 indicates data that is not meeting functional needs of the researchers.

Another usability related quality measure we present captures redundancy in the data set. Redundancy occurs when multiple fields represent the same attribute. This is particularly problematic when values of redundant attributes become inconsistent over time, i.e. one attribute is updated but the other redundant one is not. We quantify redundancy, Q_r as the ratio of attributes that are duplicates of other values or are derived from other values. Unless redundant attributes are necessary for validation, they should also be removed from the data.

3.4 SBDRP Data Quality Assessment

Many of the data quality measures described apply to the SBDRP group. Because data collection occurred prior to the computer scientists joining the project, some of the measures cannot be calculated. Currently, observation certainty is not being measured. We hope this measure will be incorporated in the future. Researcher vocabulary consistency is addressed by using a standardized survey and protocol manual. The survey does have a lot of room for inconsistency among researchers. We are developing a new survey that is more standardized and we will distribute a questionnaire with the new survey to evaluate interpretation of questions and consistency of vocabulary across all team members. Also, to help minimize vocabulary issues the project team currently has one person manually go through the data and a verify consistency. Finally, in terms of data stability, protocols have changed considerably over the life of the project. The size of the matrix prevents us from incorporating it into the paper. However, over the course of the project, valid values for almost every attribute have been modified.

In terms of quality measures related to data entry and validation, Table 3 summarizes these for two of the data files processed, demographic data and survey data. The original demographic data had 26 columns of data, 10 of which were redundant, derived or removed. Many of the columns contained multiple attribute values and needed to be split up or they contained comments that

[3] Within the scope of observation sciences, the actual value may be unknown. For this discussion, we assume that actual values are recognizable and inferred ones are noted as such.

needed to be changed to discrete attribute values. The file eventually had 42 columns, 11 of which were redundant, derived or removed. The number of redundant columns is a little high, but we were able to identify them before adding the redundancy into the database. When validating the data, 90% of the data given for processing was accurate. This was very encouraging. We investigated the inaccuracies in more detail and found that all the errors were contained in 6 columns. The majority of errors were invalid values (46%) followed by badly formatted data (39%) and values less than a specified minimum threshold (15%). The invalid values were typically values that were actually acceptable values. They were just not recorded as such.

Table 3: Quality evaluation for SBDRP data

Quality Metric	Demographic Data		Survey Data	
Q_a - accuracy	1039/1154		60/13454	
Q_a - redundancy	10/26	11/42	14/157	7/250

The survey data seemed to have less redundancy than the demographic data. The original data file had 14 redundant, derived or removed columns from a total of 157 columns. A large percentage of the 157 columns had data for multiple attributes in it. After splitting out attributes, the number of columns increases to 250. Of those, only 7 were removed because they were redundant or derived. Unfortunately, during data validation, only 60 of the 13,454 records could be processed. That is less than 1% of the data. When analyzing the errors, we found that approximately 10 different types of errors occurred in over 40% of the columns. The biologists were aware of this problem even before the accuracy calculation. Unfortunately, there are a number of inconsistencies in the survey data because of changing protocols, valid value adjustments, data entry errors and observation detail consistency. While our cleaning programs can identify the attributes with errors, most of the actually cleaning must be done manually.

After the poor results, a researcher in the lab manually cleaned approximately 246 surveys, checking 250 column values. Prior to the manual cleaning, only 2 of the records were processed. After the manual cleaning 184 records did not contain errors and were processed. That is approximately 75%, a vast improvement. As with the demographic data, the majority of errors resulted from badly formatted data and invalid values. As more cleaning is taking place, rules and transformations are being identified. For example, badly formatted data correction can be automated using regular expression transformations. We plan to use this information for incorporation of automated tools such as Potter's Wheel, in the cleaning process.

At this stage, we have now completely processed the demographic data and continue to process the survey data collection. While the survey data is being semi-manually cleaned by domain experts, we will begin developing html-based forms for future survey data. These forms will incorporate basic error and integrity checks. Once that is complete, processing the remaining survey data and the other small data repositories should be straightforward since the approach has been standardized and many of the kinks in the cleaning and transformation procedure have been ironed out. Finally, by the winter we plan to complete the model of the remaining data sources and being cleaning and transforming it.

4. CONCLUSIONS

In this paper we describe a data transformation and data cleansing procedure for a specific scientific data set. We discuss some of the quality issues specific to data collection of observational scientific data, including observation certainty, researcher observation consistency, and changing collection protocol standards. While we have worked on this project for over 6 months, there is still much to do. Because data cleansing and data validation are time consuming iterative processes, it will be months before all the historic data is in the database. During this time we will begin adding new data through html forms that incorporate robust error checking and validation procedures.

Information quality begins with information standardization. Observational sciences present even larger challenges because much of the data itself is subjective. Subjectivity across researchers needs to be reduces to maintain high quality data that is consistent and usable. If the electronic data repositories are well designed and data entry is standardized, then sophisticated data analysis can be conducted on clean, accurate data, ultimately, leading to meaningful knowledge discovery.

5. ACKNOWLEDGMENTS

Our thanks to the entire SBDRP team for staying enthusiastic throughout this integration process.

6. REFERENCES

[1] ERCreator software at http://www.modelcreator.com/.

[2] Kettle software at http://www.kettle.be/index.htm.

[3] Martinez, A. and Hammer, J. Making quality count in biological data sources. In *Proceedings of the International Workshop on Information Quality in Information Systems (IQIS 2005).*

[4] Pipino, L., Yang, W. Lee, Y., Wang, R. Data quality assessment. *Communications of ACM*. 45, 4 (April 2002).

[5] Potter's Wheel software at http://control.cs.berkeley.edu/abc.

[6] Preen, A., Marsh, H., Lawler I.R., Prince, R.I.T., Shepherd R. 1997 Distribution and abundance of dugongs, turtles, dolphins and other Megafauna in Shark Bay, Ningaloo Reef and Exmouth Gulf, *Australia Wildlife Research*. 24(1997).

[7] Raman, V., Do, H. Data cleaning: problems and current approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), 2000.

[8] Rational software at http://www.ibm.com/software/rational.

[9] Redman, T. The impact of poor data quality on the typical enterprise. *Communications of ACM*. 41, 2 (February 1998).

[10] Scannapieco, M., Virgillito, A., Marchetti, M., Mecella, M., Baldoni, R. The DaQuinCIS Architecture: a platform for exchanging and improving data quality in cooperative information systems, Information Systems, 29 2(2004).

[11] Simitsis, A., Vassiliadis, P., Sellis, T. Optimizing ETL Processes in Data Warehouses. ICDE 2005.

[12] Vassiliadis, P., Simitsis, A., Skiadopoulos, S. Conceptual modeling for ETL processes. DOLAP 2002.