Superfund Site Analysis Using Web 2.0 Technologies

Russell Anderson
russellkanderson@gmail.com

Musa Jafar
mjafar@mail.wtamu.edu
Department of Information and Decision Management
West Texas A&M University
Canyon, TX 79015

Jim Rogers
jrogers@mail.wtamu.edu
Department of Life, Earth & Environmental Sciences
West Texas A&M University
Canyon, TX 79015

Abstract

The data required to plan the clean-up of an environmentally contaminated site is available, but scattered throughout research journals. In this paper, we present a process used to collect and organize this data in a single data repository. We discuss information quality issues encountered in building the repository and how these issues were resolved. We also present the design of an application which makes the data available and usable to environmental analysts via the World Wide Web. In constructing the site, numerous Web 2.0 technologies were employed. We describe their application.

Keywords: Web 2.0, data cleansing, entity resolution, interface design, Ajax, protective concentration levels

1. INTRODUCTION

In 1980 the congress of the United States passed the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA), which charged and empowered the Environmental Protection Agency with oversight of the cleanup of abandoned hazardous waste sites. "Superfund" is the name given to this program.

In evaluating site cleanup requirements, for each contaminant present, the question must be answered: "What is a safe and acceptable level of the contaminant at the site?" This level is referred to as the Protective Concentration Level (PCL). Typically PCLs are computed for species growth, reproduction, and/or mortality.

In this paper we describe the design and construction of a database and the development of a tool to derive PCLs for a given site. We focus on issues tackled with respect to data quality, entity resolution, model design, model implementation, and tool availability and usability. The overall objec-

tive was construction of a functional tool that allowed the analyst to:

1. Quickly compute a PCL for each contaminant of concern at the site.
2. Supplement and/or override toxicity data from the database with new or corrected data, professional domain knowledge, and expert judgment.
3. Generate documentation supporting the computed PCLs for submission to clean-up plan reviewers.

2. INFORMATION QUALITY ASSURANCE

In a keynote address given at the EPA's 23rd Annual National Conference on Managing Environmental Quality Systems, Dr. Richard Y. Wang (2004), Director of the MIT Information Quality Program said the following: "The lessons learned in applying solutions to solve DQ [data quality] problems in other settings could be adopted to manage the quality of environmental data, which in turn would enhance the EQP Quality Community's ability to contribute to environmental protection efforts." In this section, we review those lessons and their potential application.

Information quality assurance methodologies divide the process into two phases: data cleansing and entity resolution. The goal of data cleansing is to detect and remove all errors and inconsistencies in the data (Rahm, 2000). According to Rahm, the types of errors to be detected include:

- Invalid values – illegal nominal values, out-of-range numeric values, and unexpected values based on variance or deviation;
- Misspellings;
- Missing values;
- Identifier values not matching a related entity – a referential integrity violation.

Müller (2003) offers an alternative error taxonomy. He classifies errors as:

- Syntactic – value does not conform to domain or inconsistencies in units of measure and abbreviations;
- Semantic – is contradictory or violates one or more integrity constraints;
- Coverage – missing values.

In entity resolution, database entries are mapped to either shared or uniquely known real world entities. In earlier data quality research, this was sometimes referred to as "deduplication" (Bhattacharya, 2007).

Byung-Won On (2007) separates entity resolution (ER) problems into three categories:

- Split ER – multiple instances of the same entity appear due to variants in the instance identifier (or name).
- Mixed ER – instances of different entities are merged due to similarities in spelling or pronunciation of the names.
- Group ER – entities are grouped based on content only because they lack identifier or name values.

Müller (2003) proposed three steps in the data cleansing process:

1. Audit the data to identify the types of errors present.
2. Choose and/or create appropriate methods to automatically locate and remove errors.
3. Apply the methods to the tuples in the data collection.

Although Müller recommends automation of as much as possible, he considers data cleansing to be a semi-automatic process. Usually it requires the involvement of a domain expert to detect and correct some anomalies.

In contrast, Rahm (2000) defined five phases of data cleansing:

1. Data analysis – metadata collected and summary statistics computed;
2. Definition of the transformation workflow and mapping rules;
3. Transformation;
4. Verification of results after transformations have been applied;
5. Backflow of the cleansed data to original sources as needed.

Data for PCL Computations

As recommended by both Müller and Rahm, we began the data cleansing process with an audit of the data and a review of the types of errors found. The primary values used in computing the PCLs are toxicity data gleaned from thousands of text-based research reporting effects of a given contaminant on a given species. From a data quality perspective, inconsistencies of the data include:

1. Lack of uniformity of the measures reported. For example, many of the studies reported the concentration level in which 50% of the test subjects died (LD 50), while others reported the "No ob-

served adverse effect level" (NOAEL), yet others reported the "Lowest observable adverse effect level (LOAEL).
2. Some of the studies were conducted using adult subjects while others used juveniles.

These data were manually collected by graduate students in the Environmental Science program. From each study, they recorded the contaminant and species of the study, the body mass of the species when available, the toxicity level reported, and the toxicity type (LD 50, NOAEL, LOAEL, etc.).

The lab based toxicity values found in the literature could not be used directly as PCLs because of adjustments that needed to be made based on the species' exposure to the contaminant in its natural environment. Values used to make these adjustments included:

- Water and fat solubility of the contaminant (log k_{ow});
- Trophic level (position in food chain) of the species;
- Food, water, and soil sediment ingestion rates;
- Percent of time spent by the species within the bounds of the clean-up area;
- Percent of contaminated food in the diet.

These adjusting values were mostly found in public domain and government agency databases.

Domain experts defined the algorithm used to convert toxicity values of different types as found in the literature, to a single uniform toxicity reference value (TRV), to which the adjustments were applied to compute the final PCL for the species (Figure 1). Because there was not a consensus among experts with respect to the adjustment factor values or even the formulas to use, we found it necessary to allow the analyst using the application to override many of the values and formulas used in computing the PCLs.

The Database

The database designed for the project contained the following tables:

- Habitat – list of all site habitats supported by the tool. The habitat defines the list of species in the habitat. Identified by habitat name.
- Chemical – list of contaminants for which data has been collected. Identified by CAS number. Includes contaminant specific adjustment factors.
- AnimalClass – list of all species classes for the species of interest. Includes default ingestion rate adjustment values. Identified by class name.
- Specie – list of all species for which data is available. Identified by common species name. Includes species specific adjustment factors, body weight, and class name.
- SpecieHabitat – list of species in a given habitat. Identified by common species name and habitat name.
- TRV – toxicity reference values from literature. Identified by CAS number, common species name, concentration type (LOAEL, NOAEL, LD50), and PCL type (mortality, reproduction, or growth). Includes TRV, body weight of test subjects, and bibliographic reference information.
- SurrogateAssignment – surrogate species to be used when TRV values of species are missing. Identified by CAS number, common species name, concentration type (LOAEL, NOAEL, LD50), and PCL type (mortality, reproduction, or growth). Includes surrogate species common name.

The database also contains other tables used to manage users of the application and to save the state of an in-progress site analyses.

Data Cleansing

The data when collected by the domain experts was recorded in Microsoft Excel spreadsheets. Since most of the data was collected before we got involved in the project, much of the data was redundant. For example, the log k_{ow} of each contaminant was repeatedly recorded for each species against which it was applied. After finishing the construction of the production database, we estimated that 75% of the data in the original spreadsheets was redundant.

As a first step in preparation for the PCL application, we went through a data cleansing process followed by a transfer of the data from the spreadsheet to the production relational database management system (mySQL). Each toxic concentration value found in the literature was recorded along with identifying information of the CAS num-

ber of the contaminant and the species name. The CAS number is a numeric identifier assigned to chemicals by the Chemical Abstracts Service of the American Chemical Society. A CAS number is separated by hyphens into three parts. The first part is an assigned number up to seven digits. The second part is an assigned two digit number, and the last is a single check digit. The species name is the common name.

Because data entry was a manual process, and because data were collected from hundreds of sources, errors in the data were common. Some of these errors included: leading and trailing spaces, and double spaces between words. Short scripts were written to automatically locate and correct these errors. To locate problem CAS numbers, the format and the check digit were validated. Invalid numbers required a manual correction.

To validate numeric values, a reasonableness check was performed on all TRV concentrations and adjustment factors.

Entity Resolution

Another frequent error was in the spelling of the species name. As pointed out by Bhattacharya (2007), an effective first step in entity resolution is to sort the tuples by the attribute in question. Hence, we first sorted by species name. This brought into proximity names such as "Red-Winged Blackbird" and "Red Winged Blackbird". Once alternative spellings were identified, a single agreed upon spelling was chosen and the database was updated.

As a second level of species name validation, referential integrity was checked from the TRV and SurrogateAssignment tables against the species table. This led to the discovery of additional species name inconsistencies such as "Mallard" and "Duck, Mallard".

Missing Values

In doing the computations, we also had to deal with missing values. Many of the toxicity studies from which data were collected pertained to domestic species, not the species commonly found at the clean-up site under review. To get around this problem, toxicity values for a given contaminant from surrogate species were assigned to the species in question with an adjustment for differences in body weight. When adjustment values, such as food ingestion rate, were missing for a given species, they were estimated using a formula based on the animal class of the species. Again the formulas used for estimation were defined by domain experts.

3. IMPLEMENTATION OF THE PCL ANALYSIS SYSTEM

Architecture

The objective of the project was to make as widely available as possible, the data collected for analysis and the tools implemented to conduct the analysis. The first option considered was to create a stand-alone application packaged with a data repository. These could be distributed on CDs at conferences or downloaded from a web site. The second option was to create a web-based application with access to a centralized data repository.

The advantage of the stand-alone application is that it provides a richer development environment with ready access to features for designing and creating a more powerful, yet easy to use, user interface. For example, a stand-alone application can easily include both 2-d and 3-d vector graphics for creating powerful analysis visualizations.

The primary advantage of the web application is that it provides a single data repository and analysis application with updates that are immediately available and visible to all users. It also provides better access control over its use and allows collection of usage statistics such as frequency of use and features most and least used. Also, given the capability to save one's analysis on the server, it facilitates sharing and collaboration using the ability to authorize others to load, review and edit a project.

Mainly because we knew that the database would be changing, growing and evolving over time, we chose to go with the web-based implementation. The database was implemented using mySQL. Application server side processing (mostly data access) was implemented on a Tomcat 6.0 server using JavaServer Pages (JSP). On the client side, as much as possible, we chose technologies that were browser independent and eliminated or minimized the need for browser plug-ins. All client side scripting was

done in javascript using only features of the HTML DOM (W3C, 2007) and javascript methods, properties, and classes that were available in all of the most widely used browsers (Firefox 2 and 3 and Internet Explorer 7; Chrome had not yet been released).

Design Objectives

The primary objective of the analyst using the tool is to find a target PCL for each contaminant of concern at the site. Usually the target PCL will be the lowest PCL of each of the individual species inhabiting the site.

Before designing the interface, we listed the tasks that the user would complete in performing a site analysis. These tasks included:

1. Specification of analysis parameters to include: contaminant to be evaluated, species to be included in study, and PCL types to be computed (growth, reproduction, and/or mortality);
2. Assessment of available data applicable to analysis specification, including:
   a. TRVs of contaminant available in database for species in study;
   b. Surrogates available for species missing TRV data;
   c. Food, water, and soil/sediment ingestion rates for each species;
   d. Solubility factors for contaminant used to compute accumulation in animal;
3. Ability to supply missing values or override existing values (with source reference or justification) and have these new values automatically included in computations;
4. Ability to sort resulting PCLs in order to quickly select lowest value. Normally the lowest PCL of all of the species at the site is chosen as the target PCL, since it would be the worst case value;
5. Generation of contaminant and species support documents that can be added as supplements to the analyst's final report.

Application Design

Based on the task list, we broke the process down into two main steps: first, define the problem parameters, then second, allow the analyst to view and manipulate the available data and resulting calculations. When an analysis is first initiated, the analyst is presented with a screen to select the habitat, which automatically defines the species to be evaluated; the contaminant, and the PCL types to be computed. Habitat is chosen from a select box; contaminant may be chosen either by CAS or chemical name from select boxes; PCLs for growth, reproduction, and mortality are individually selected using check boxes. See Figure 2.

To facilitate task 2 above (assessment of available data), we decided to give the presentation of data and results a tabular or spreadsheet-like layout. Values of the chosen contaminant are presented at the top (Figure 3), while species specific values, TRVs, and PCL computations are presented in the table – one row per species and PCL type (Figure 4).

To facilitate task 3, all modifiable values are placed in text boxes. Color encoding of each cell background is used to indicate the source of the cell value. White indicates that the value is found directly in the database; pale yellow is used for values calculated from other values found in the database; khaki is used for values entered by the user; and light green is used to indicate values that are calculated from analyst entered values (Figure 5). When a user enters a value into a cell, all downstream values are immediately recomputed and the backgrounds are color encoded; thus allowing the user to see the effects of a change.

With respect to task 3, the user is also allowed to add annotations to any cell that contains a user entered value – providing reference information or rationale about the new value (Figure 6). This feature was deemed especially important for reviewers of an analysis in order to assess the validity of the change and ultimately the legitimacy of the resulting PCL.

To support the analyst in collecting support documentation for the analysis (task 5), various entries on the data page were linked to Adobe Acrobat (pdf) formatted charts and descriptions. Selecting the contaminant name opens a page describing the chemical and listing its quantitative properties. Selecting the species name opens a page describing the species, its quantitative properties and a picture of the species. Selecting the habitat opens a document containing the food web for the selected habitat. Once opened, each document may be saved on

the user's local machine and packaged with the final site analysis report.

To give the analyst flexibility in designing and presenting results, after all changes to the data page have been completed, the results may be exported. When the user selects the "Export" button, its onClick event handler collects all on-screen data into a CSV format and sends it to the server via an Ajax post [Powell, 2008]. Upon receipt, the data is temporarily stored in a CSV formatted file and notification is sent back to the browser (again using Ajax) that the file is available. The browser, upon receiving notification, redirects to the newly created file and it is opened using the browser's default handler for CSV files. For example, if the analyst is using MS Windows and has MS Excel installed, then the document is automatically opened in Excel. At this point the analyst can sort the data (task 4), eliminate any rows or columns deemed unimportant, and format the output according to needs.

4. TESTING

During development, beta versions of the site were presented to the Texas Ecological Working Group which is chaired by the Texas Commission on Environmental Quality Ecological Risk Assessment Group. It is comprised of ecological risk professionals in federal and state government agencies, Universities and private consulting firms. The site was also presented at the Texas Commission on Environmental Quality Trade Fair in 2007 and 2008 as a preliminary preview and request for comments on the usefulness and applicability of the site. Over two hundred conference participants attended the presentation. The general consensus was that the site provided a valuable and user-friendly risk assessment tool. Comments on the improvement of the site were evaluated for possible incorporation into the site.

Many of the working group members reported that they were able to use the system with minimal training and expressed satisfaction with the design. Using the application, analysts generated PCLs in less than one half hour – a task that previously took two to three weeks to complete. An additional benefit was that the application allowed risk assessors to run alternate exposure scenarios during meetings, resulting in expedited decision making and superfund site closures.

5. SUMMARY

The preceding has described a successful implementation of a web-based "superfund" site analysis tool. Challenges associated with the implementation were in two parts – database construction and user interface design. The construction of a functional database required extensive data cleansing. Novel ways to work-around missing data were defined by a committee of domain experts. Entity resolution was accomplished by first sorting on species name (the most problematic field) to group similar names. These were usually corrected by a spelling or hyphenation change. We then performed a join on species name columns found in the TRV and surrogateAssignment tables with a table of valid species names. Those species names not matching a name in the list required correction – such as changing "Duck, Mallard" to just "Mallard".

To solve the missing data problem required domain experts. A committee of experts provided formulas used to estimate missing adjustment multipliers. When a value was missing at least one formula was always available. For some values, multiple formulas were considered, with one selected for application based on the data available. To work around the problem where toxicity research was missing for a given species and contaminant, domain experts were asked to select surrogate species for which toxicity data was available. They also defined mappings based on body weight in applying the surrogate data.

With respect to the interface design of the chosen web architecture, the need for the analyst to see a "big picture" of the data was solved by creating a tabular, spreadsheet-like presentation. Like spreadsheet applications, changes in data were immediately cascaded to all downstream cells. Also, to highlight the effects on PCLs of analyst supplied data, color encoding of the background was implemented. The background colors designate values that have been supplied directly by the analyst, those values derived from other values in the database, and values derived from the analyst supplied values.

Additional features of the application include the ability to: save and retrieve work-in-progress, export to a spreadsheet compatible format, add reference annotations to user-supplied data, and generate support documentation.

6. FUTURE WORK

In a future version of the system, we plan to add a feature allowing analysts to submit new research data and/or challenge existing data. Upon submission, the data would be reviewed by a panel of domain experts who would decide if the data should be added to the master database. We consider the process similar in functionality and benefit to the GenBank (NCBI, 2009) process for submitting and collecting DNA sequences.

7. REFERENCES

Bhattacharya, I. (2007) "Collective Entity Resolution in Relational Data." *ACM Transactions on Knowledge Discovery from Data*, 1(1) March.

Fisher, C., E. Lauría, S. Chengalur-Smith, R. Wang, (2008) *Introduction to Information Quality* (4th Edition) MITIQ, Boston, MA.

Lee, T. and S. Bressan (1997) "Multimodal Integration of Disparate Information Sources with Attribution." *Proceedings of the Entity Relationship Workshop on Information Retrieval and Conceptual Modeling.*

Lee, Y. W., L. L. Pipino, J. D. Funk and R. Y. Wang (2006), *Journey to Data Quality,* MIT Press, Cambridge, MA.

Müller, H., and J. C. Freytag (2003) "Problems, Methods, and Challenges in Comprehensive Data Cleansing", Technical Report HUB-IB-164, Humboldt University, Berlin.

National Center for Biological Information (NCBI) (2009) GenBank Overview. www.ncbi.nlm.nih.gov/Genbank/.

On, B. (2007) "Data Cleaning Techniques by Means of Entity Resolution" Doctoral Thesis. UMI order # AA13284975, Pennsylvania Sate University.

Powell, Thomas (2008) *Ajax: The Complete Reference*. McGraw-Hill.

Rahm, E., and H. H. Do, (2000), "Data Cleaning Problems and Current Approaches", *IEEE Data Engineering Bulletin*, 23(4), Dec, p. 3-13.

Wang, R. Y. and D. M. Strong (1996) "Beyond Accuracy: What Data Quality Means to Data Consumers", *Journal of Management Information Systems* 12(4), p. 5-34.

Wang, R. Y. (2004) "Data Quality: Theory in Practice", Keynote Address of EPA 23rd Annual National Conference on Managing Environmental Quality Systems, Tampa, Florida.

World Wide Web Consortium (W3C) (2003) Document Object Model (DOM) Level 2 HTML Specification. www.w3.org/TR/DOM-Level-2-HTML.

Appendix

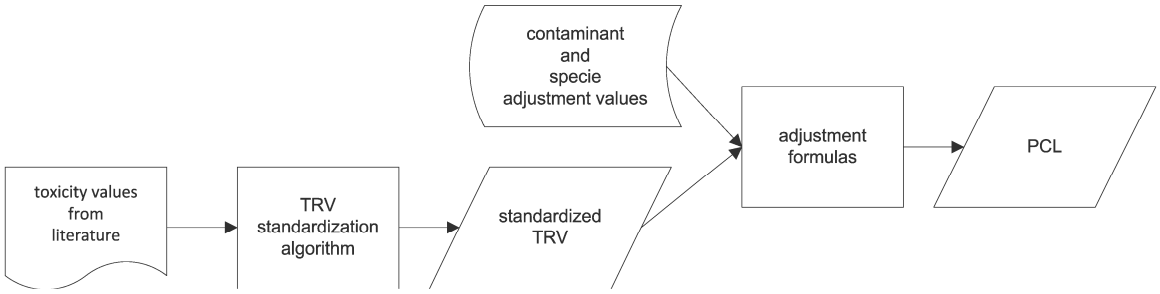

Figure 1 – PCL Computation Flow

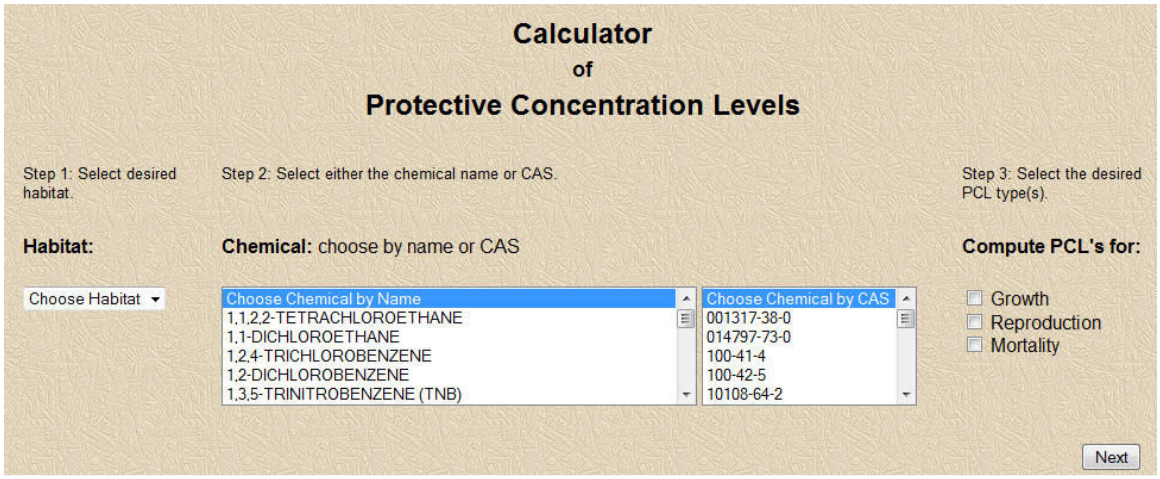

Figure 2 – Analysis Specifications

Figure 3 – Chemical Header

Specie	Body Mass	Trph Lvl	FCM	BCF	BAF	Scle Fctr a	Scle Fctr b	Food IR	Water IR	Soil Sed IR
Barred Tiger Salamander	50	4	26.669	26002	693453	0.0113	0.79	0.009431	0	0.00047
Leopard Frog	100	4	26.669	22002	586771	0.0111	0.75	0.007335	0	0.00037
Barn Swallow	16	3	14.355	26002	373261	0.638	1.2	0.2615	0	0.02432

PCL Type	Literature NOAEL	Literature LOAEL	Literature LD 50/LC 50	Surrogate Used	TRV NOAEL	TRV LOAEL	Computed PCL	Range %	Time %	Food %	Adjusted PCL
GROW		195.2		none	19.52	195.2	0.016416	100	100	100	0.016416
MORT				BULL FROG	23.1338	231.338	0.0194551				0.0194551
GROW				N/A	-	-	-	100	100	100	-
MORT				BULL FROG	20	200	0.0255578				0.0255578
GROW	2.43			none	2.43	24.3	0.000136925	60	100	100	0.000228208
MORT				BOBWHITE QUAIL	3.7249	37.249	0.00020989				0.00034982

Figure 4 – Species Data and Computations

(In browser, images above appear as a single left to right table.)

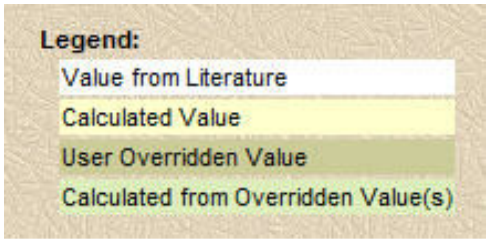

Figure 5 – Background Color Legend

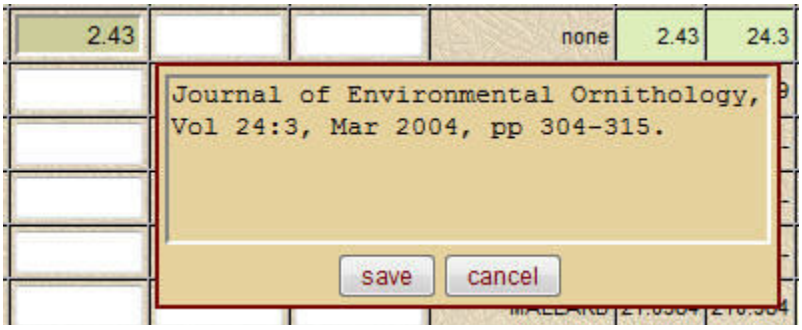

Figure 6 – Cell Annotation Editor