

REQUEST FOR APPROVAL

To: Ted Rauh
Deputy Director, Enforcement & Compliance Division

From: John Halligan
Branch Chief, Recycling Enforcement Branch

Request Date: March 3, 2011

Decision Subject: Approval of Scope of Work for Data Analytics for Fraud Detection and Prevention (Data Mining) Contract Proposal - (Beverage Container Recycling Fund FY 2010/2011)

Action By: March 16, 2011

Summary of Request:

CalRecycle recently implemented the DORIIS system which consolidated a number of separate data systems into a single system. As a result, there is now improved availability, accessibility, integrity, and timeliness of key CBCRF data.

CalRecycle's Recycling Enforcement Branch is currently operating with software tools that are only capable of provide rudimentary rules-based and anomaly detection of transaction data. These processes have limitations and are only capable of providing known patterns of fraud. These processes also produce high percentages of false negatives that can result in inefficient allocation of staff resources.

The former Division of Recycling engaged in a contract with Gartner consulting where it was determined that CBCRF disbursement data, when modeled, produces statistically valid indications of potential fraudulent transactions and/or fraudulent program participants.

With these two bodies of work behind us and the associated knowledge, CalRecycle is a position to implement a modern, efficient and repeatable data mining process using the available program data to address the unmet need of detecting and/or deterring program related fraud with the use of advanced data analytics (data mining).

In order to develop and implement the data mining process described above, CalRecycle Recycling Enforcement Branch staff created a scope of work for a professional services contract. The recommended scope of work is attached for your review and approval.

Recommendation: Staff recommends that you approve the attached Contract Allocation Proposal for this contract.

Deputy Director Action:

On the basis of the information and analysis in this Request for Approval, I hereby approve the scope of work for the Data Analytics for Fraud Detection and Prevention (Data Mining) contract proposal.

Dated: [MARCH 8, 2011]



Ted Rauh
Deputy Director, Enforcement and Compliance Division

Attachments:

Contract Allocation Proposal

CONTRACT ALLOCATION PROPOSAL

Project Title: Data Analytics for Fraud Detection and Prevention	
Program/Office: CBCRF Enforcement	Concept No.:
Requestor/Primary Contact: John Halligan	Fund (IWMA, Oil, RMDZ, etc.): CBCRF
Estimated Contract Amount: \$ 150,000.00	

I. PROPOSAL OVERVIEW

- The unmet need is potentially \$20 million of fraud annually that are identifiable using advanced data analytics (data mining). Currently this fraud goes undetected, and subsequently undeterred, with the tools and staff skills available to CalRecycle. This out of a potential CRV Fraud exposure for the CBCRF of \$64 million per year (8% x 800 million of CRV disbursements).
 - The original dollar values below are being revised upward because in the intervening years the amount of CRV collected and distributed has doubled in dollar value for the corresponding quantity of material (the rate paid per pound of CRV material was effectively doubled).
 - DORIIS Feasibility Study Report (FSR):
 - Program Objectives for the DORIIS Project included; “Increase Fraud detection by approximately \$10 million per year.”
 - **Fraud** – The DOR estimates that Program fraud is equal to 8% of disbursements, or approximately \$48.3 million/year.
 - Gartner Consulting contract “Fraud Detection and Data Analysis”
 - “DOJ analysis estimates \$50 million of fraud within DOC”
 - “DOC/Gartner report estimates \$50 million of fraud within DOC”
- CalRecycle recently implemented the DORIIS system which consolidated a number of separate data systems into a single system. As a result, there is now improved availability, accessibility, integrity, and timeliness of key CBCRF data.
- CalRecycle’s Recycling Enforcement Branch is currently operating with software tools that are only capable of provide rudimentary rules-based and anomaly detection of transaction data. These processes have limitations and are only capable of providing known patterns of fraud. These processes also produce high percentages of false negatives that can result in inefficient allocation of staff resources.
- The former Division of Recycling engaged in a contract with Gartner consulting where it was determined that CBCRF disbursement data, when modeled, produces statistically valid indications of potential fraudulent transactions and/or fraudulent program participants.
- With these two bodies of work behind us and the associated knowledge, CalRecycle is a position to implement a modern, efficient and repeatable data mining process using the available program data to address the unmet need of detecting and/or deterring \$20 million of fraud annually due to the use of advanced data analytics (data mining).

Brief description of project:

- See primary task at the end of this document for more detail relating to how and what will be performed in the project.
- CalRecycle will engage a contractor with extensive experience in implementing advanced data analytics to detect and deter fraud in the private and public sector. The contractor will derive a solution based upon industry best practices and base the solution upon successful implementations of data mining for fraud in other compliance/enforcement organizations..
- The project would advance CalRecycle's efforts at implementing Business Intelligence with a much higher degree of sophistication and effect. Currently CalRecycle has rudimentary query tools to perform fraud detection. The project would advance the sophistication of tools used and provide for increased efficiency and effectiveness in the direction/allocation of Investigation Section staff resources.
- The ultimate goal would be the implementation of Predictive Analytics.
- Business Intelligence Tools are software that enables business users to see and use large amounts of complex data. CalRecycle intends that these tools be installed on a desktop computer system and the analysis conducted using that system. The following three types of tools are referred to as Business Intelligence Tools:
 - Query Tools - Software that allows the user to ask questions about patterns or details in the data.
 - Multidimensional Analysis Software - Also Known As OLAP (Online Analytical Processing) - software that gives the user the opportunity to look at the data from a variety of different dimensions.
 - Data Mining Tools - Software that automatically searches for significant patterns or correlations in the data (statistical models). Data mining is driven by the data itself and is inductive in nature. OLAP software and data mining tool technologies complement each other in Business Analytics
- Predictive Analytics spin through millions of pieces of information looking for statistical patterns and simply tells you what it finds, free of the human assumptions and bias that can skew results. . Predictive Analytics provides the ability to go beyond your assumptions and discover things you wouldn't otherwise know. The key is the ability to recognize patterns or associations between seemingly disparate things.
- The contractor would work with CalRecycle to develop statistical models for identify potentially fraudulent transactions and application data for program participants.
- The Statistical models would then be integrated into an automated routine on the desktop computer used to perform the analysis. This routine will analyze CBCRP disbursement data identifying potentially fraudulent transactions and/or application data.
- The automated and repeatable data mining process would generate a report for each execution of the process. This report would be triaged by the Recycling Enforcement Branch, Risk Assessment and Data Analysis Section, who would forward transactions and/or program participants to the Investigations section for follow-up.

- Based upon the success of the model for predicting potentially fraudulent transactions, CalRecycle will consider options for broader internal deployment of the predictive model.

Outcomes of Project:

- Improved Fraud / Compliance detection process. The solution should result in significant improvement in the overall compliance detection process. These controls will have a direct impact on how much CRV is distributed.
- Data Analytics. The solution will provide capabilities for the users to use statistical functions and calculations that will facilitate detection of hidden trends and patterns with our datasets.
- Shift focus of data analytics from investigation “after the event”, which has limited success in recouping lost CBCRF monies, too focusing on fraud prevention by identifying potentially fraudulent transactions based on this information prior to CBCRF monies being disbursed. The preventative approach would also save on CBCRF funds due to a reduction in investigation and prosecution cost associated with responding after the fact (recovery of monies paid out).
- Significantly reduce the quantity of false-positive indications of fraud generated using the current tools and skills available to CalRecycle.
- Increase Investigator efficiency and associated return on investment. The data mining results indicating potential fraud will have a low percentage of false-positives. The majority of transactions and program participants referred to Investigations staff due to the data mining results should identify situations requiring enforcement actions. Investigators should have a 70% plus likely hood that a referral based on the data mining results will result in identify and/or preventing fraud.
- The efficient and effective focus on prevention will also act as a deterrent to would be fraudsters. This is premised on the concept that the majority of individuals committing fraud do so because they have observed other fraudulent activity and witness it being profitable without negative consequences.
- Get a clearer picture of new types of fraud before they emerge. This is the promise of predictive analytics; the statistical models will identify correlation that a human sifting through the same data would not identify.
- Potentially identify organized fraud networks.
- Data Profiling. By using the solution, the CalRecycle team will be able to profile applicant and transaction data. Data Profiling will enable the team to get deep insight into frequency distribution, pattern frequency distribution, percentiles and outliers within the data sets through an automated process.
- Application of advanced audit techniques to aid in administrative enforcement. The solution will provide the CalRecycle team with analytical techniques to detect hidden trends and patterns with datasets, i.e. geo coding the data, using fuzzy matching algorithms etc.
- Reduce manual work via an automated software solution.
- Handle large volumes of data with limited resources (software, hardware, staff, and skills).

- Automated process to assess what is normal, abnormal, and what is problematic.
- MS office integration. To the extent possible, the solution should provide seamless integration with MS office products i.e. Word, Excel and PowerPoint. Users will be able to access data from MS Office interface. By doing so users will be able to perform work using a familiar interface they are experienced and trained to use.
- OLAP Analysis. The solution will provide capabilities for users to develop OLAP (Online Analytical Processing) Cubes. These are highly summarized data sets that can be used to detect trends and patterns in large volumes of data. CalRecycle team can use this feature to analyze trends and patterns in large volumes of data. CalRecycle can use this feature to analyze trends and patterns across all the program participants or to analyze historical data for particular geography.

Linkage to prior projects:

- In 2006 The Department of conservation, Division of Recycling contracted with Gartner consulting for a project titled "Fraud Detection and Data Analysis for the State of California's Department of Conservation." The objectives of the contracted project are listed below.
 - Identify the underlying and interrelated structure of the data available for use in fraud detection.
 - Determine the data elements and changes in the data elements that presage fraud.
 - Identify the data elements that best discriminate between honest and fraudulent recyclers.
- The results of the project were a determination and demonstration that a model can be developed providing statistically valid results indicating potential fraud using the data available to DOR.
- Testing and implementation/deployment of the model developed was not in the scope of this contract and a follow-up contract to implement/deploy the model for DOR's uses was not developed or funded.
- The contract/project being proposed is the logical extension of the prior project/contract "Fraud Detection and Data Analysis" to implement/deploy advanced data analytics for detecting and deterring fraud.

How will the project's success/outcomes be measured?:

- A metric used to evaluate the success/outcomes will be the count (as a percentage) of false negatives produced by the repeatable data mining process (project solution). A false negative is when the data mining process identifies a possible fraudulent transaction or program participant and they are not fraudulent. The percentage of acceptability will have to be identified in the project based upon the degree of accuracy that the model is built to.
- A metric used to evaluate the success/outcomes will be count of program participant certifications revoked as a result of compliance activity being initiated by outputs from the project solution.

- A metric used to evaluate the success/outcomes will be total dollars of findings generated as a result of compliance activity being initiated by outputs from the project solution.
- A metric used to evaluate the success/outcomes will be a decrease of positive results produced by the repeatable data mining process indicating potentially fraudulent transactions and/or program participants.

Explain impact if funding for the project is not approved this FY:

- CalRecycle with its current recourse allocations (staffing) and available software tools will not be able to respond to, identify and deter, potentially 20 million dollars of fraud annually committed against the CBCRF.

II. PRIMARY TASKS, DELIVERABLES AND MILESTONES

PRIMARY TASKS:

- Conduct project planning meeting:
 - Familiarize the team with the approach and identify any problems/issues/concerns with approach.
 - Identify key documentation for review.
 - Identify subject matter experts for participation.
- Develop project work plan and schedule
- Conduct project Initiation meeting
- Identifying business objectives and requirements: This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.
 - Define the project requirements.
- Develop data understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or detect interesting subsets to form hypotheses for hidden information.
 - Review relevant documentation.
 - Interview subject matter experts.
- Prepare data for modeling: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed, multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.
 - Identify available data and filter it for data that will be relevant to the project.
 - Determine the source and access to the data used in the project.

- Coding of data, translating the data from its native state to values that can be processed by the software used to model the data.
 - Validate the quality (how clean) the available data is and determine if it will have a material impact on the models to be developed.
 - Prepare the final sets of data to use for modeling.
- Modeling the data: In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
- Identify the underlying and interrelated data structure relating to fraud.
 - Determine the changes in data elements that presage fraud.
 - Identify the elements that best discriminate between honest and fraudulent transactions.
 - Conduct Statistical Analysis:
 - Correlation analysis to determine if there is a statically significant relationship between Fraudulent and non-fraudulent data as an indicator.
 - Discriminant analysis to identify the optimal dimensions that best illustrate group differences and to identify the relative contributions of each variable.
 - Means analysis to identify the direction of the relationship between significant attributes.
 - Logistic regression to confirm the results of the remainder of the analysis and determine how “good” the model is.
- Evaluation of model(s): At this stage in the project a model(s) are built that appear to have high quality, from a data analysis perspective. Before proceeding to final development of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
- Calculate the likely predictive success of the fraud model.
 - Create a sampling plan to create a model and sampling plan to test the model developed in the prior step.
 - Update the model with any findings identified during testing.
- Deployment of solution: Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what action will need to be carried out in order to actually make use of the created model.
- Determine the most effective solution for hosting the model and providing access by CalRecycle staff.

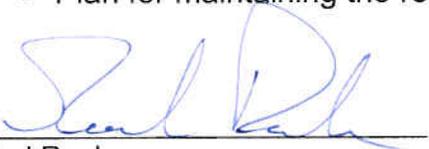
- Deploy the working model and associated software for CalRecycle staff's use.
 - Train CalRecycle staff on how to use the solution developed by the contractor.
 - Test the deployed solution.
- Close out the project

ESTIMATED PROJECT MILESTONE DATES:

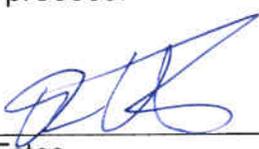
- Secure contract dollars for the project, March 2011.
- Conduct project planning meeting, April 2011.
- Draft RFP and post for bids, June 2011.
- Obtain signed contract, September 2011.
- Conduct project initiation meeting, October 2011.
- Conduct project, October 2011.
- Implement/deploy solution, April 2012.
- Close project out, May 2012.

DELIVERABLES:

- Implementation of a repeatable data mining process for CBCRF disbursement data that provides statistically valid results for use by Recycling Enforcement Branch Staff to initiate investigations of program participants resulting in the detection and deterrence of fraud against the CBCRF.
 - Software that will perform repeatable data mining process.
 - Hardware to host the software that performs the repeatable data mining process.
 - Documentation for the models developed, from a statistical perspective.
 - Documentation for the method used to deploy the solution implementation.
 - Documentation for the outputs from the repeatable data mining process.
 - Documentation for how to interpret the outputs of the repeatable data mining process.
 - Documentation for evidentiary quality of the outputs from the repeatable data mining process.
 - Training for CalRecycle staff identified as users of the repeatable data mining process outputs.
 - Plan for maintaining the repeatable data mining process.



Ted Rauh
Deputy Director
Compliance and Enforcement Division



Tom Estes
Deputy Director
Administration, Finance & Information
Technology Services Division

