AIR FORCE RESEARCH LABORATORY
ROME, NEW YORK

STATEMENT OF WORK

FOR

REAL TIME NETWORK FORENSIC ANALYSIS

PCSN G-6-2310

13 NOVEMBER 2006

CONTRACT No. FA8750-07-C-0014

ATTACHMENT No. 1

TABLE OF CONTENTS

1.0 OBJECTIVE.

1.1 The objective of this effort is to determine the possibility for inferring Semi-Structured Data Formats automatically, and to develop the architecture for a larger Network Forensics Analysis Workbench.

2.0 SCOPE.

2.1 The scope of this effort includes study and analysis of inferring Semi-Structured Data Formats automatically, and to develop the architecture and a prototype Network Forensics Analysis Workbench.

3.0 BACKGROUND.

3.1 The purpose of this effort is to determine whether one of the core technical pieces of the long-term objective is feasible. In particular, the purpose is to explore to what extent we can infer automatically the structure of the semi-structured ad hoc data that forensic analysts encounter in practice. Without automated or semi-automated inference, network analysts must painstakingly attempt to understand the structure of the data by hand and perform the time consuming, error-prone, and tedious task of manually constructing a format parser, which must then be linked by hand to other tools such as query engines and visualization components.

3.2 To tackle the task of format inference, PADS data description language will be used as the target for the inference system. This language has been specifically designed to describe the kinds of data typical for forensic analysis. The inference system will leverage the volume of data to be analyzed to separate the punctuation encoding the structure of the data from the actual payload. The idea is to then infer the structure of the data by examining the patterns formed by distinctive base types and punctuation, and use functional dependency analysis to recover bounds on array lengths and union tag information.

3.3 If this approach to inferring the structure of ad hoc data sources is successful, it should enable forensic analysts to access the content of a new data source in minutes rather than in days. This will remove a major roadblock to enabling forensic analysis to occur in real-time, thereby supporting more flexible and reactive computer network defense.

4.0 TECHNICAL REQUIREMENTS.

4.1 The contractor shall design, develop, document, implement, evaluate and deliver a format inference algorithm for ad hoc data and propose architecture for a Network Forensics Analysis Workbench that includes the inference algorithm.

4.1.1 Evaluate the effectiveness of format inference for real-time network forensics, by having a collection of data sources relevant to that domain. (CDRL A003)

4.1.1.1 Identify and study the kinds of ad hoc semi-structured data sources that are critical to real-time network forensics.

4.1.1.2 Collect and archive representative samples of such data (as privacy and security concerns allow) of sufficient size to measure the effectiveness and scalability of our inference system.

4.1.2 Design, develop, document, and implement an Inference Algorithm to support the evaluation described. (CDRL A004)

4.1.2.1 Identify an appropriate collection of base types for the domain of forensic analysis including IP addresses, MAC addresses, path names, URLs, email addresses, dates in multiple formats.

4.1.2.2 Develop algorithms that efficiently identify occurrences of distinctive base types in ad hoc data and discover the punctuation tokens used in a given data source.

4.1.2.3 Develop techniques to bootstrap the identified occurrences of distinctive base types and punctuation into a description of the higher-level structure of the data using frequency counts and iterative refinement.

4.1.2.4 Develop algorithms to find dependencies between array lengths, union tags and other values.

4.1.2.5 Prototype the resulting inference algorithm and study the behavior with respect to the quality of the inferred description and the performance characteristics.

4.1.3 Use the ideas of the LaunchPADS system to propose a Forensic Data Workbench Architecture. (CDRL A005)

4.1.4 Evaluate and document the effectiveness of the inference algorithm. (CDRL A006)

4.1.4.1 Develop a means to measure the quality of the inferred descriptions. Compare the descriptions produced with benchmarks including simple heuristics for loading CSV and related formats into Excel, techniques for extracting payload from web pages generated by merging a template and data, (if such tools are available), and hand-written descriptions produced by experts in the data source.

4.1.4.2 Use statistical tools available from the PADS system to measure the percentage of data from a data source that conforms to the inferred specifications.

4.1.4.3 Measure the speed of inference with current approaches, and measure the time requirements of the inference system empirically.

4.1.4.4 Document the circumstances under which the inference algorithm is or is not able to produce a data description.

4.1.5 Develop and conduct software demonstrations. (CDRL A007)

4.1.5.1 Develop compelling software demonstrations that will illustrate research progress and the potential for practical application of appropriate technical areas.

4.1.5.2 Conduct demonstrations to illustrate the benefits of the approach as developed in the program.

4.1.6 Deliver all computer software developed, assembled, or acquired to the Government in accordance with the Contract schedule and the following. Software developed and delivered under this effort is to be completely maintainable and modified with no reliance on any non-delivered computer programs or documentation.

4.1.6.1 Deliver all computer software developed under this effort as source and object (executable) code on electronic media. Include the commented source listings and source coded for the target computer system. (CDRL A008)

4.1.6.2 For all software purchased or licensed as a component of the software delivered, transfer licensing and maintenance agreements to the Government upon the completion of this effort. (CDRL A009)

4.1.7 Reporting and documentation.

4.1.7.1 Continually determine the status of the effort and report progress toward accomplishment of contract requirements. (CDRL A001)

4.1.7.2 Conduct oral presentations at such times and places designated in the contract schedule. Provide status of technical progress made to date in performance of the contract during presentation. (CDRL A002)

4.1.7.3 Document the details of all technical work accomplished and information gained during performance of this acquisition to permit full understanding of the techniques and procedures used in evolving technology or processes developed. Include all pertinent observations, nature of problems, positive and negative outcome, and design criteria established where applicable. Document procedures followed, processes developed, "lessons learned," and other useful information. (CDRL A010)