SAND2016-6486D

Data Integration for SPR Cavern Analysis Tool

iSAFE
Interns for Security, Arms control and Force protection Engineering

Manager:
Shawn Taylor, 6613

Sandia National Laboratories

Katherine A. Seale
University of South Alabama
MS in Information Systems/Information Assurance, May 2017

Project Mentors:
David Hart, 6915
Giorgia Bettin, 6912

A tremendous amount of data, information, and reports are generated with operations at the four DOE Strategic Petroleum Reserve sites. Sandia National Laboratories stores and maintains a wealth of valuable historic information that is used daily. This project focuses on the development of Python code that identifies relevant reports from a shared drive, extracts important information, and integrates the data into a database. This project will utilize the database and indexing tools to efficiently store and share the vast amount of data associated with the SPR.

Requirements for this project include:

- Define the parameters of interest
- Write a Python script to search folders for relevant reports and logs
- Extract text from PDFs
- Develop an algorithm to identify and extract relevant measurements
- Import results into a database

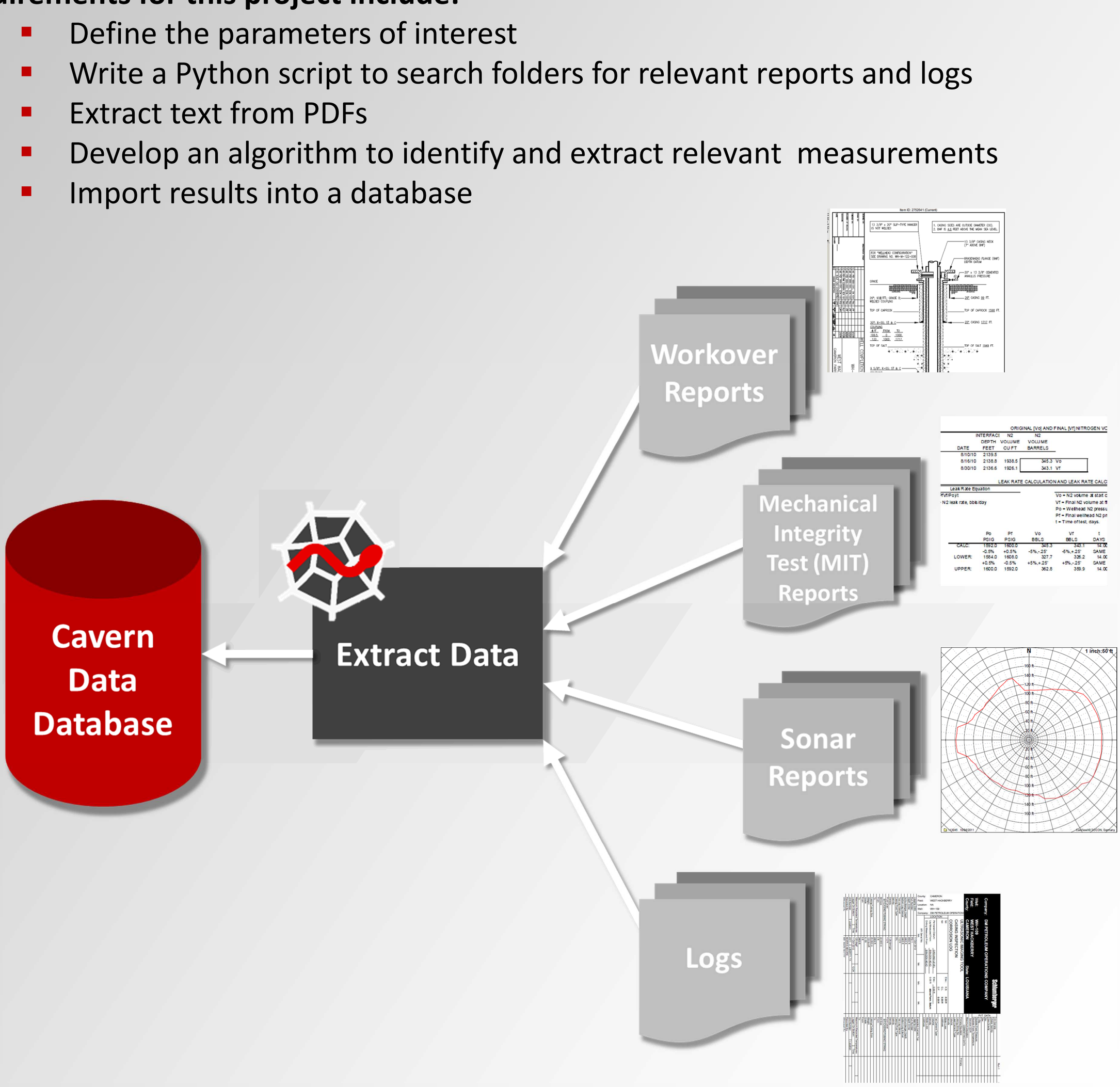

Figure 1. Flow Diagram of SPR Data Extracted from Workover Reports, MIT Reports, Sonar Reports, and Logs into the Cavern Data Database

U.S. DEPARTMENT OF ENERGY

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP

Data Integration for SPR Cavern Analysis Tool

iSAFE
Interns for Security, Arms control and Force protection Engineering

Manager:
Shawn Taylor, 6613

Sandia National Laboratories

Katherine A. Seale
University of South Alabama
MS in Information Systems/Information Assurance, May 2017

Project Mentors:
David Hart, 6915
Giorgia Bettin, 6912

Methodology

1. **Collect Sample Files for Testing**
   Use SPR Reports to Test Script
2. **Search Directories for Files**
   List Path and Filename,
   Include Subdirectories in Search
3. **Sort Files by File Extension**
   Focus on PDFs
4. **Extract Data**
   Research Methods to Parse data, Verify if PDF Contains Text,
   Sort Files by Content / Type of Report, Search for Keywords
5. **Store Results in Database**
   Practice using Temporary Database,
   Apply Methods to Preferred Database
6. **Apply Tool to Shared Folder**
   Verify that Methods Work with Big Data
7. **Present Results**
   Demo and Presentation to Team
8. **Future Work**
   Apply OCR to PDFs with Scanned Images,
   Search Other File Types, Add More Search Functions

Figure 2. Project Requirement Matrix

U.S. DEPARTMENT OF ENERGY

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP