International Conference on Advances in Nuclear Forensics
Vienna, Austria, July 7-10, 2014

The Uranium Sourcing Database Project: Practical Insights into the Establishment and Application of a Nuclear Forensics Library

Martin Robel, Naomi Marks, Ian Hutcheon, Rachel Lindvall, and Mike Kristo

Lawrence Livermore National Laboratory

ABSTRACT: The Uranium Sourcing Database is a working nuclear forensics database containing data on thousands of samples of uranium ore concentrate (UOC) and related products. The database is part of a broader effort to characterize and document distinguishing properties of UOC for use in assessing the probable source of sample of material absent any packaging or identifying marks. While this project has focused on UOC, the lessons learned are equally relevant to a wide range of nuclear and radiological materials. We will present a number of practical insights, including nuclear forensics database development and population, user interface requirements, analytical laboratory to database interface, and database utilization.

Introduction

The Uranium Sourcing Database is a working nuclear forensics database containing data on thousands of samples of uranium ore concentrate (UOC) and related products. The database is part of a broader effort to characterize and document distinguishing properties of UOC for use in assessing the probable source of sample of material absent any packaging or identifying marks[1]. While this project has focused on UOC, the lessons learned are equally relevant to a wide range of nuclear and radiological materials. We will present a number of practical insights, including nuclear forensics database development and population, user interface requirements, analytical laboratory to database interface, and database utilization.

Nuclear forensic methods can be broadly divided into two categories: predictive and comparative. Predictive forensics requires detailed, accurate, and validated models of physical processes governing the production and alteration of nuclear materials. For some types of materials, such as spent reactor fuel, such models do exist at a level of refinement that makes them useful for nuclear forensic analysis[2]. However, for many types of material, including uranium ore concentrates, no validated models exist that capture the complexity and variability of the associated signatures. For this reason, the best tools available for forensic analysis are comparative. Conclusions are drawn after considering the similarities and differences between the unknown and a reference set of known materials. The principal purpose of a nuclear forensics (NF) database is to serve as a data repository from which to draw these reference sets for comparative nuclear forensics. However, the NF database has utility beyond simply storing data for use in a comparative forensic investigation. A collection of data and metadata from a number of samples representing a variety of sources can also serve as an empirical foundation upon which to begin the development of predictive insights and models to complement comparative models in the forensic process.

In this paper, we describe our insights acquired from years of practical experience with a database of uranium ore concentrate sourced from around the world. The Uranium Sourcing Database was established as a tool to research the application of comparative signatures to the problem of safeguards verification. The goal of this work is to verify that the characteristics a collected sample are consistent with the declared source of the material. For nuclear forensics, the process is very similar, and hence the database requirements are also very similar. In fact, we have utilized the Uranium Sourcing Database for nuclear forensic investigations[3]. Hence, for the purposes of this paper, we will be referring to the Uranium Sourcing Database as a nuclear forensics database.

Database design, administration, and personnel considerations

The first task involved with the establishment of a nuclear forensic (NF) database, following the decision to implement such an effort, is to identify an individual or team to design, administer, and manage the database.

One approach to designing a nuclear forensics database, is to work with an experienced database developer to design and implement the new database. This approach has many advantages, including efficiency of implementation, potential cost savings over training internal staff, and avoiding overtaxing staff with additional duties. However, the lack of familiarity of the database developer with the particular needs of a nuclear forensics database may be a liability. Additionally, the involvement of a dedicated database administrator will usually be required beyond the initial implementation period, and it will likely be essential to have a database administrator available on an ongoing basis to maintain the database and make periodic improvements to the system.

It is also possible to develop the necessary database development skills for database design and implementation within an already established nuclear forensics work group. This is a good approach if resources are limited, but places an increased workload on staff. This approach requires at least one staff member to possess or develop specialized skills in database design and implementation. One significant advantage to database design without the aid of an external database developer is that the in-house developer is likely to possess greater familiarity with nuclear forensic data. Cultivating database skills internally to the nuclear forensics working group will also likely facilitate greater interaction with, and collaboration between, the database administrator/designer and the analytical staff.

Designing a database for nuclear forensic data

The term database is used in many different ways; in this work we define a database as a collection of data stored in some organized fashion. Data is stored within a database in one or more tables; a table is a structured list of data of a specific type. The way the tables are designed and relate to each other is referred to as the database structure. The simplest structure is a flat database, in which all of the data is stored in the rows and columns of a single table. An example of such a flat structure is an Excel worksheet, or a single table in Microsoft Access. Flat databases are quite straightforward to set up and implement, and require a minimum of specialized database knowledge. For a very small database that will be accessed by only one person at a time, it is possible that an Excel spreadsheet might be sufficient. For anything beyond the most basic and limited database, however, a more robust model that allows for multiple users across many computers is far preferable.

A data model for nuclear forensic data

The data model describes the underlying entities and relationships that a database is designed to represent and capture. The entity relationship diagram is used to develop and document the data model. There are many ways to organize a given dataset, which includes both data (e.g., measured values) and metadata (e.g., the type of instrument used to make the measurements). The data model for the Uranium Sourcing Database was developed through an iterative design process. The primary design goals for the Uranium Sourcing Database are ease of use for nuclear forensic queries; ease of use by subject matter experts; and maximizing utility for end users. The structure that we developed emphasizes the importance of samples and measurements, since these are the starting point for a nuclear forensic investigation.

Database Structure

In the Uranium Sourcing Database, data are grouped into two primary logical units (*tables*), 15 secondary derivative tables, and relationships are defined to link the tables. This structure provides efficient storage of information, and provides for built-in data validation. For example, all entries (*records*) in the result table must have corresponding sample and parameter information. The presence of lookup tables supports consistency in the data sets by limiting valid values to, for example, correct spellings and consistent abbreviations. The relational database structure is useful for efficient retrieval of subsets of data to meet user requirements.
The two principal tables in the database are the **Sample** and **Result** tables (Figure 1).

The **Sample** table contains information about each of the samples in the database. This includes sample composition, provider, and date received by the lab, among other things. Each analyzed sample has a unique Sample ID, and also sometimes additional ID numbers that were provided by the sample provider. *Sample* is the key field that links the sample to its chemical, physical, and image data found in the **Result** and **Image** tables. *Sample* also links the sample to data found in the **Class**, **Source**, and **Location** derivative tables. The date of sample receipt and mass of the sample are noted in the **Sample** table as well. The Sample table is linked to the **Class** table, which contains information about the "*class*" of the sample (usually the location of collection or production), the *source* of the sample, and the *country* of origin of the sample. The **Source** table contains information about the geologic provenance of the sample (*geologic_province*) and the type of deposit from which the sample was derived (*deposit_type*). The **Sample** table is linked to a number of lookup tables including **Material** (i.e. specific type of UOC compound), **Provider**, and **Location**. Image files associated with specific samples are included in the **Image** table. It is important to note that a given sample may have more than one associated image, i.e. several photographs, SEM images, or other graphic data. The **Image** table is therefore linked to the sample table with a many-to-one relationship.

The **Result** table contains quantitative laboratory measurements, expressed as numeric values, and qualitative results (i.e. XRD interpretations) expressed as text. The **Result** table is linked to the **Sample** table by the *Sample* ID field. The **Result** table is also linked to the **Parameter**, **Analysis**, **Units**, **Instrument**, and **Lab** tables. Another important link is between the **Result** table and the **Document** table, which links directly to the original source data for a given measurement. Typically documents in the **Document** table are either Excel files, or pdf files from which the data was extracted for uploading to the database. It has proven useful to have an easily accessible archive (i.e., a data repository) of the source documents that have been used to populate the database. Relationships between the **Sample** table, **Result** table and other tables in the database are shown in Figure 1.

Analytical laboratory to database interface

The Uranium Sourcing Database effort includes a substantial sample characterization component. In addition to data from outside sources, much data is generated in our labs specifically for the purpose of populating the database. Once an analyst finishes a set of measurements (e.g., strontium isotopic ratios), they send a file containing the data to the analytical lead for the database, who then vets, formats, and uploads the data to the database. When external reports are required for a sample analysis, the data can be easily downloaded from the database into a reporting template.

Incorporating some LIMS functionality into a NF database

Rather than stand up a separate laboratory information management system (LIMS), we have opted to utilize the Uranium Sourcing Database to track the status of sample analyses. This can be seen in the sample table in Figure 1; most of the fields are for tracking.

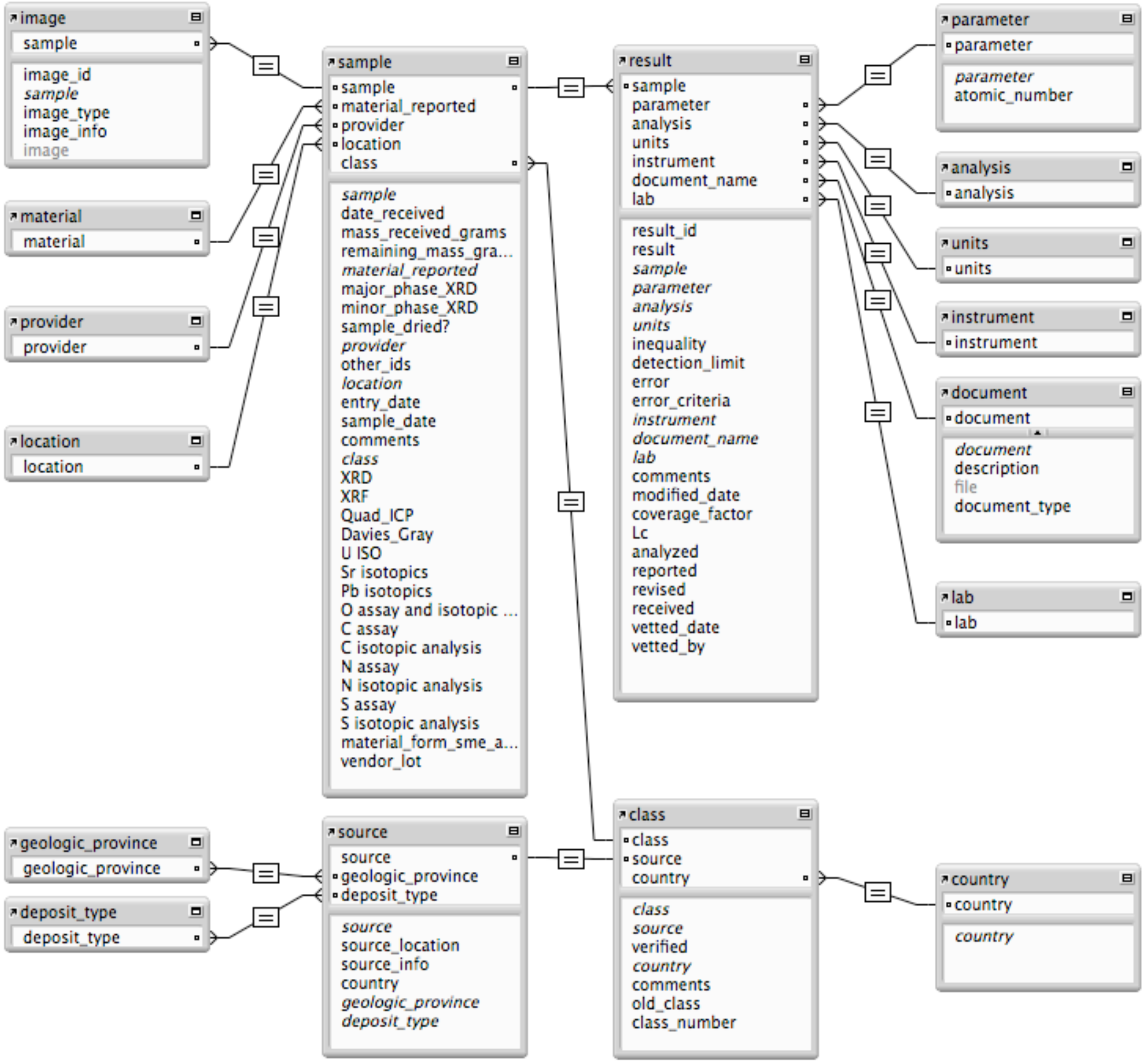

Figure 1. Uranium Sourcing Database core database diagram.

Choosing a platform for the database

There are a wide variety of database management software platforms that provide the necessary functionality for a nuclear forensic database. We have experimented with a few different platforms for the Uranium Sourcing Database; each has its own advantages and drawbacks.

Most nuclear forensic databases will likely be of a size and scale that will require a relational structure and a robust, multi-user interface. A general-purpose database management system is a software package designed to allow the creation, querying, updating, and administration of databases. A number of widely used platforms meet these requirements, including Oracle, MySQL, Microsoft SQL Server, Microsoft Access, and FileMaker Pro, to name a few. Some primary considerations in selecting a particular platform are the ability to allow and control user access, institutional support, and familiarity with the software.

While desktop systems like Microsoft Access are relatively user-friendly, they are not designed for multiple simultaneous users. FileMaker Pro is unusual in that it is both relatively user-friendly and also designed for multiple users on a network. One limitation to the FileMaker platform is that it does not use structured query language (SQL) and is therefore a non-standard database platform. Despite this limitation, we have used FileMaker Pro for the Uranium Sourcing Database because it was already part of the standard software package at LLNL. The institutional support for FileMaker Pro made implementation over the LLNL network straightforward. Our preferred approach is to employ a database with institutional support, at least for the initial database implementation.

As goals or priorities change, it may be necessary or desirable to migrate the NF database to a new platform. Choosing a relatively simple structure and avoiding the use of business logic (e.g., in the form of *stored procedures*) in the database makes migration easier. If business logic is kept at the application layer level (i.e., not part of the database), it may be reused with the new platform with relatively minor changes.

The cost of database software varies substantially, depending on the scale and level of support. For databases on the scale of nuclear forensics data, there are many no-cost options, including both open source, unlimited platforms like MySQL as well as free, size-limited versions of proprietary platforms like Oracle and SQL Server. No-cost versions typically lack customer support, but online user forums are a rich source of information as well as voluntary "crowd-source" support.

Data types

Data can be stored in database tables in a number of numeric and non-numeric formats. Selecting the appropriate format for data storage within the database has been an important component of designing and maintaining the Uranium Sourcing Database. In this context 'data type' refers to a storage format that constrains the type of information stored by a computer in a variable. For example, the 'tiny int' data type used by Transact-SQL only allows storage of integers from 0 – 255 in a variable that uses 1 byte of memory[4]. There are many different data types used by database programs and many of them have cryptic names like 'varchar(50).' This situation is further complicated by the fact that there are many deprecated data types. Since data types are not standardized across all databases, we will not go into further detail on these specifics. There are a few broad categories of data type that are needed for a NF database. These include text, integer, decimal, and time.

One important lesson learned from the Uranium Sourcing Database effort is that database data types don't necessarily easily accommodate the range of data types represented in geochemical datasets. For example, the decimal data types specify the number of digits allocated before and after the decimal. This poses a problem when attempting to store data with variability in the number of significant digits (e.g. storing isotopic data good to six significant figures cannot be stored in the same format as trace element data good only to two significant figures, unless they are stored as text).

Another common problem occurs when the database administrator receives measurement data on spreadsheets with the display setting adjusted to show the correct significant digits. Uploading these data into a numeric data type field will result in numbers with far more digits reported than intended or appropriate. It is therefore preferable to have analysts submit data to the database administrator in text format, such that the correct significant figures are preserved. It is also essential to perform quality assurance/quality control on the data, ensuring that analytical results are truncated to the appropriate number of significant figures prior to uploading data to the database. One imperfect and counter-intuitive solution to *preserving* significant figures in the database is to use a text data type for the measurement data. Unfortunately, this creates other problems: database software typically cannot sort numbers-stored-as-text properly. This will require another workaround. One could, for example, have a duplicate column with the same data stored as numbers (with the incorrect significant digits), simply to use as a field for sorting or other mathematical operations. There are probably many other solutions that will work as well. Our aim is not to declare a universal solution to this problem, but rather to call attention to it.

Populating a nuclear forensic database

Units and conventions

Data for a NF database is likely to come from multiple sources, with differing requirements and standards of reporting. Some data may be generated from laboratory analysis of samples of interest specifically for NF purposes. But there are also numerous potential sources of external data, which was originally collected for other purposes. Data collected for quality control, for example, might be reported in different units with different conventions for dealing with detection limits. There are two ways of dealing with inconsistencies in the data designated for the NF database. One option is to import the data to the database as received, and perform the necessary operations (e.g., converting reported units from ppm to μg/g U) after exporting the data for a specific query. The advantages of this approach are 1) reducing the potential for data corruption through errors in conversion, and 2) reducing the up-front work load by saving these operations until such time as they are needed. The second option is to perform all conversions prior to upload, so that the data in the database is as consistent as possible. This increases the up-front work load, but it makes the database far more useful. Furthermore, if a file repository is used, there is a record of the original data in its original form. This will minimize potential problems from corrupt conversions of data prior to uploading to the database. Regardless of which approach is used, all of the data should be vetted by a technical expert familiar with the measurements that produced the data prior to using it for forensic investigations.

The file repository

Typically, the database administrator receives data in files that have been vetted by subject matter experts. In addition to uploading these data to the database, it is highly recommended that a link to the original file be facilitated by the database structure. In this way each measurement for every sample in the database can be traced back to the source document. In the Uranium Sourcing Database, this is achieved by the use of a document field in the result table, which links to a document table, which links to a file, as illustrated in Figure 1. Some off-the-shelf analytical database solutions do not allow this linking of measurements with documents. We feel this is a critical requirement for a nuclear forensics database.

Data entry

Databases can either be populated manually, by typing in one entry at a time, or in bulk or batch operations in which data is uploaded as a group of entries. In most cases, a batch/bulk import operation is the most efficient and consistent approach. Batch and bulk imports are accomplished either with a SQL script or through a graphical user interface, or a choice of either method, depending on the database software. In many cases, the format of the data provided to the administrator for input to the database is not in the format required for bulk importing. For repetitive data formatting operations, some kind of automation is highly recommended. This automation can be programmed using a variety of languages; for the Uranium Sourcing Database, Microsoft Visual Basic for Applications (VBA) in Excel is used.

Queries

The process of interrogating the database for information is referred to as querying. The word ‘query’ can mean a request for information, but it can also refer to a block of SQL code that is not limited to requesting information (it can be used to perform other operations, including deleting data). Queries can be performed by executing SQL commands (for most systems) or though a graphical interface. Since the NF database administrator and end user(s) will not necessarily be SQL coding experts, development of at least two graphical interfaces are suggested: one for the administrator and one for other technical users, such as scientific staff.

Utilizing a NF database

In general, there are three categories of users of the NF database, and each has different interface requirements. These are the database administrator/developer, the scientist/investigator, and the 'customer.' The administrator/developer needs full control of the database, including both the structure and the content. We have found that a variety of off-the-shelf applications meet the administrator interface requirements. But these same applications tend to be overwhelming to the scientist/investigator who is typically not a database expert. A custom application layer can be designed to facilitate the functionality desired for this user group. However this requires a significant software development effort. Some platforms provide graphical user interface development environments to aid in this effort. Alternatively, a web programmer can be employed to develop a web browser interface.

The third user category, the customer, probably should not be interacting with the database through any direct interface. Instead, customer queries should probably be directed through a database point of contact, who will likely be a manager or a database administrator. The customer may be asking only for database statistics or they may be requesting utilization of the data for a nuclear forensics investigation.

There are two categories of graphical user interface for databases: off-the-shelf software (e.g., SQL Server Management Studio Express) and custom applications. The off the shelf solution requires less development work, but it requires a higher skill level to utilize (though not as high as the command line SQL interface).

In general, there are three categories of users of the NF database, and each has different interface requirements. These are the database administrator/developer, the scientist/investigator, and the 'customer.' The administrator/developer needs full control of the database, including both the structure and the content. We have found that a variety of off the shelf applications meet the administrator interface requirements. But these same applications tend to be overwhelming to the scientist/investigator. A custom application layer can be designed to facilitate the exact functionality desired for this user group. However this requires significant software development effort. Some platforms provide graphical user interface development environments to aid in this effort. Alternatively, a web programmer can be employed to develop a web browser interface. The Uranium Sourcing Database currently uses the former approach (Filemaker Pro 'layouts'), but is in transition to the latter (web interface using PHP).

The third user category, the customer, probably should not be interacting with the database through any direct interface. Instead, customer queries should probably be directed through a database point of contact, who will likely be a manager or a database administrator. The customer may be asking only for database statistics or they may be requesting utilization of the data in for a nuclear forensics investigation.

For utilization in a nuclear forensic investigation, querying the database is only the beginning; a subject matter expert will need to review the data in the context of the query and call upon outside knowledge not necessarily captured in the database. For this reason, customer queries to the NF database for drawing nuclear forensic conclusions should probably only be handled by a small group of technical experts, who will use the data and metadata to develop reports to the originator of the request for comparative NF analysis.

For complex signatures, additional data processing may be called for. In these cases, the data from the query are typically exported to Excel and/or an analysis environment like MATLAB for further processing and analysis. The Uranium Sourcing Database is populated primarily with uranium ore concentrate (UOC) data. Since samples of UOC of forensic interest don't have physical dimensions (like a fuel pellet) or serial numbers (like a sealed source), we must rely on other measurable properties for the process of comparative nuclear forensics. The relatively high abundance of elemental impurities in UOC, comprises a multivariate signature. These, along with isotope ratios are exported from the database and utilized as inputs to a multivariate analysis, such as principal components analysis (PCA) for characterization or partial least squares discriminant analysis (PLS-DA) for discrimination/classification/attribution[5].

Database summary reports involve a special kind of query, and include two types of information: that which can be derived by a direct query of the data and that which requires synthesis and interpretation and/or some kind of calculation. An example of the first type is a report documenting the number of samples in the database from a particular location. An example of the second type is a report documenting how many new sources were added to the database in the past year. Both examples are typical of the kind of information that management requires for metrics. The first example should be easily fulfilled by the most rudimentary database. The latter example requires a date-added field in the sample table, something that may not occur to the developer when deciding what kinds of information needs to be captured. It is recommended that these types of requests be given particular attention when developing the database fields to ensure that all likely requests can be addressed by a database query.

Conclusions

The Uranium Sourcing Database effort has yielded many practical insights into nuclear forensic database development and utilization. Lessons learned from this database of uranium ore concentrates should be broadly applicable to a wide variety of nuclear material types. There are a number of factors to evaluate when establishing a nuclear forensic database, from initial design to utilization in a nuclear forensic investigation. In addition to the obvious (database design, population, and management), special consideration should be given to important issues such as analytical laboratory interface; handling significant figures in the database; the linking of data to documents and source files; the unique requirements of each type of database user; and utilization of data in multivariate analysis.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 with the support of the National Nuclear Security Administration Offices of Nuclear Controls (NA-242) and Nuclear Verification (NA-243).

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

References

[1] Operating Efficiently / Engaging Globally: FY 2013 Annual Report NNSA Office of Nonproliferation and International Security, http://nnsa.energy.gov/sites/default/files/nnsa/04-14-inlinefiles/2014-04-23%20NIS_fy_2013_annual_report.pdf

[2] Chambers, A. S. (2010). A Comparison of Nuclide Production and Depletion using MCNPX and ORIGEN-ARP Reactor Models and a Sensitivity Study of Reactor Design Parameters Using MCNPX for Nuclear Forensics Purposes. (Doctoral dissertation). https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2010-05-853.

[3] Keegan E. et al, Nuclear forensic analysis of an unknown uranium ore concentrate sample seized in a criminal investigation in Australia, Forensic Science International, Volume 240, July 2014, Pages 111-121, ISSN 0379-0738, http://dx.doi.org/10.1016/j.forsciint.2014.04.004.

[4] int, bigint, smallint, and tinyint (Transact-SQL). http://msdn.microsoft.com/en-us/library/ms187745.aspx

[5] Robel, M., Kristo, M. J., and Heller, M. A. (2009). Nuclear forensic inferences using iterative multidimensional statistics. Institute of Nuclear Materials Management 50th annual meeting. LLNL-CONF-414001. https://e-reports-ext.llnl.gov/pdf/374432.pdf.