



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

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

M. Robel, I. Hutcheon, M. Kristo, R. Lindvall, N.
Marks

July 1, 2014

International Conference on Advances in Nuclear Forensics
Vienna, Austria

July 7, 2014 through July 10, 2014

Disclaimer

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.

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

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

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 analysis¹. However, for many types of material, such as uranium ore concentrates, no such validated models exist that capture the complexity and variability of the associated signatures. In these cases, the only tool available for forensic analysis is the comparative process, whereby 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 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. In this application, the goal is to verify that a collected sample's characteristics 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 investigations². Hence, for the purposes of this paper, we will be referring to the Uranium Sourcing Database project as a nuclear forensics database project.

Database design, administration, and management personnel considerations

The first task involved with the establishment of a nuclear forensic (NF) database, after the management 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, from the ground up, 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 the nuclear forensics work group. This is a great approach if resources are limited, but places an increased workload on staff and requires the need for one or more staff members to possess or develop specialized skills in database design and implementation. A significant advantage to developing database design without the aid of an experienced 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 and collaboration with the database administrator/designer and the analytical staff.

Designing a database for nuclear forensic data

The term database is used in many different ways, but for our purposes, a database is 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 design of a database structure is guided by a number of potentially competing priorities. Designing and implementing a database requires evaluating these priorities and developing a structure for how the data will be stored.

One example of the competing design goals which we have confronted in the development and application of the Uranium Sourcing Database is the tradeoff between performance and accessibility. This trade-off manifests in different aspects of database development. Higher performance (i.e. faster computing) tends to come with more complex designs that require a higher level of developer expertise and specialization. While a sophisticated database that has been optimized for performance may run queries much faster than a less sophisticated design, a couple of factors make this of secondary importance for a nuclear forensic database.

First, nuclear forensic databases are likely to be small compared to “big data” applications, so that computational efficiency is not a priority. The Uranium Sourcing Database, for example includes hundreds of thousands of analyses, but is still in the small to medium-sized range where performance is not an issue. In general a somewhat less efficient data structure may run substantially slower than a more efficient design, but given the relatively small size of nuclear forensic database, and since *programmer time* is typically more valuable than *CPU time*, it is often a reasonable tradeoff.

Most nuclear forensics databases are likely to fall into the “small” database category, with perhaps tens of users, fewer than 10^7 records, and less than 40GB of data. At this scale, computing efficiency is

generally less of a priority than the limited human resources available for designing and implementing an efficient data structure.

Database normalization is the process of organizing the fields and tables of a database to minimize redundancy

The degree of data normalization is an example of a case where a less efficient (slower to process queries) design is likely preferable. The more normalized the database is, the more tables will be required for certain queries that combine many different attributes³.

Queries that involve more tables will run more slowly, but at the small to medium database scale, this effect will be relatively small, if even detectable. Hence, computational efficiency should not be a consideration for the degree of normalization of a nuclear forensic database.

In summary, at this scale, human resources are more expensive than computing resources, so ease of use by a potentially non-specialist database administrator is more important than ease of use for the hardware processing the commands or compliance with industrial scale database norms that make a system more complex and cryptic.

The preceding discussion notwithstanding, there are cases where more complexity is warranted. One of the first design decisions to make is whether the database will be a flat file or relational database. The flat format, which can be implemented by simply organizing data and metadata into rows and columns in Microsoft Excel, is more accessible than the relational database format and its associated software and programming language. However, once one goes beyond a single computer, single user system, the Excel database approach becomes problematic; version control, access control, robustness, concurrence, as well as worksheet and file size limits are some of the reasons why this approach doesn't scale well. Hence, we have adopted a relational database approach for the Uranium Sourcing Database.

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 settled on emphasizes the importance of samples and measurements, since these are the starting point for a nuclear forensic investigation.

Database Structure

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 verification checks. For example, all valid results must have corresponding sample and result 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 (fig. 2).

The **Sample** table contains information about each of the samples in the database. The **Sample** table contains information about the sample material collected, including 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, 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.

Relationships between the **Sample** table and other tables in the database are shown in Figure 1.

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. The **Result** table is also linked to 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 pdfs from which the data was originally derived. We find it is useful to have an archive of the source documents that is easily accessible in the database.

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 house” specifically for the purpose of populating the database. Once an analyst finishes a set of measurements (e.g., strontium isotopic ratios), they send a document containing the data to the analytical lead for the database, who then vets, formats, and uploads the data to the database. If an external report is required for a sample analysis, the data are downloaded from the database into a reporting template.

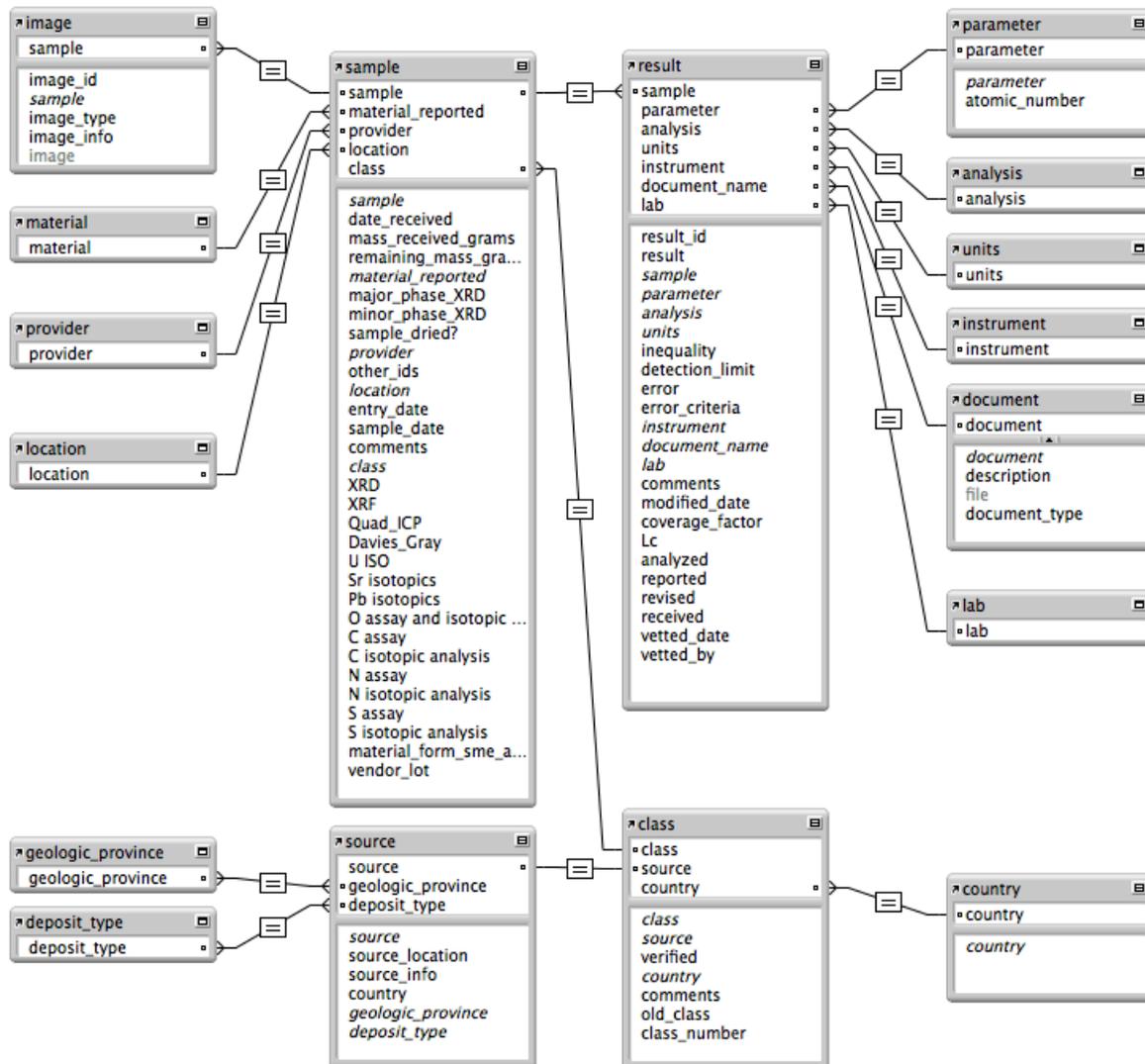


Figure 1. Uranium Sourcing Database core database diagram.

Choosing a platform for the database

We have experimented with three different platforms for the Uranium Sourcing Database: Oracle; Microsoft SQL Server; and Filemaker Pro. There are a wide variety of database management software platforms that provide the necessary functionality for a nuclear forensic database. There are also a few key considerations that can help guide the selection of a particular platform. As discussed above, we feel a relational database is appropriate for a nuclear forensic database. This rules out the spreadsheet approach. Another primary consideration is user access and version control. While desktop systems like Microsoft Access are relatively user-friendly, they are not designed for multiple users. Filemaker pro is unusual in that it is both relatively user-friendly and also designed for multiple users on a network. That said, it is non-standard; most database systems use structured query language (SQL); Filemaker doesn't. We used it for the Uranium Sourcing Database because it was already part of the standard image, making implementation over the network simple. Which leads to another important factor to consider: institutional support. If your institution already has a preferred platform, this is probably the one you should use.

One intriguing configuration is the use of MS Access as a user interface front end on a SQL Server database. This combination has the potential to capitalize on the best qualities of Access (ease of use for the end user) and SQL Server (robustness, multiple users, and network 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, it may be reused with the new platform with relatively minor changes. Another common concern is the cost of the software. The good news is that 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.

Data types

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 memory⁴. 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 are not designed to easily accommodate scientific data. This is because the decimal data types specify the number of digits allocated before and after the decimal. This is simply too rigid for scientific data. There is a provision for scientific notation, which preserves significant figures. Rather, the database administrator often 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. The 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: the database software probably won’t readily sort text numbers 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. 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 $\mu\text{g/g U}$) 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 much more useful. Furthermore, if a file repository is used, there is a record of the original data in its original form, reducing the 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 which 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. This way every 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.

Data entry

The database can be populated manually, by typing in one entry at a time, or it can be uploaded in bulk or batch operations. There are some cases where entries should be manually typed, but in most cases, a batch/bulk import operation is the best option. This is 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.

Utilizing a NF database

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 send arbitrary commands). Queries can be performed by executing SQL commands (for most systems) or through a graphical interface. Since the NF database administrator and end user(s) are not necessarily going to be SQL coding experts, it is likely that at least two graphical interfaces will be desired: one for the administrator and one for the end-users. There are two categories of graphical user interface for databases: off-the-shelf (OTS) software (e.g., SQL Server Management Studio Express) and custom application layers. The OTS solution requires less development work, but it requires a higher skill level to utilize (though not as high as the command line SQL interface). We have found that a variety of OTS applications meet the administrator interface requirements; no custom interface is needed. But these same applications tend to be overwhelming to the end-user. A custom application layer can be designed to provide the exact functionality desired for the end-user, but it requires a significant software development effort. Some platforms provide graphical 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).

For a nuclear forensic analysis, querying the database is only the beginning; a subject matter expert will need to review the data in the context of the query and draw upon outside knowledge not captured in the database to draw conclusions. For this reason, the NF database will likely only be interfaced directly by a small group of technical experts, who will use it 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/attribution⁵.

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.

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 Safeguards 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 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>.

² 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>.

³ <http://it.toolbox.com/blogs/enterprise-solutions/database-normalization-performance-storage-tradeoffs-15545>

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>.