THE URANIUM SOURCING DATABASE PROJECT: PRACTICAL INSIGHTS INTO THE ESTABLISHMENT AND APPLICATION OF A NUCLEAR FORENSICS LIBRARY

International Conference on Advances in Nuclear Forensics

Martin Robel, Naomi Marks, Ian Hutcheon, Rachel Lindvall, and Mike Kristo
Lawrence Livermore National Laboratory

Vienna, Austria, July 7-10, 2014

Lawrence Livermore National Laboratory

NNSA/NA-24 sponsored the development of a UOC sourcing capability

7593-08

- The U-sourcing database provides the empirical foundation for developing a comparative signature approach

- The Uranium Sourcing Database contains >300 physical UOC samples and "material fingerprints" for over 4000 samples from >30 countries

- ~50 discrete signature variables

  - Major and trace element abundances
  - Isotopic composition (U, Pb, Sr, Nd, C, N, O)
  - Molecular species (U_3O_8, UO_4, ADU, AUC, …)
  - ~190,000 individual entries (data)

- Discriminant Analysis Verification Engine *(DAVE)* uses Partial Least Squares Discriminant Analysis (PLSDA) to query the database

Czech-1
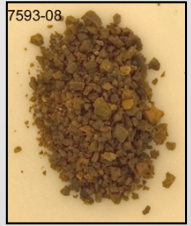
Database design, administration, and personnel considerations: internal vs. external database developer

- **External -advantages**
  - Experience, specialization
  - Efficiency
  - Potential cost savings (less training)
  - Reduced burden on internal staff

- **External –disadvantages**
  - Unfamiliar with NF db needs
  - Still R&D; not production; internal staff will still need to administer and modify db

- **Internal –advantages**
  - Greater familiarity with NF db data and requirements
  - Increased interaction between developer and analytical staff likely to result in more successful design

# Designing a database for nuclear forensics

- Flat vs relational database format
  - Flat: easy, familiar, single user, limited scale, limited queries
  - Relational: complex, unfamiliar to most, multiple user, unlimited scale, sophisticated queries, industrial strength

- The benefits of simplicity in design
  - Not a production environment; frequent changes likely
  - More transparent = more efficient for humans
  - Production environment –commuting to work
    — Get a Tesla. Sophisticated, powerful, efficient, impossible to work on
  - R&D environment –driving across the tundra
    — **Get a Jeep**. Slower, more robust, fewer features (which you don't really need), easy to diagnose, fix, and alter

A data model for nuclear forensics data

- Many ways to organize data; good structure mirrors inherent relationships in data

- Iterative design process led to current structure of Uranium Sourcing Database
  - Primary tables: Sample and Result
  - Numerous satellite tables
    — Primarily "lookup tables" –standardized lists to enforce consistency
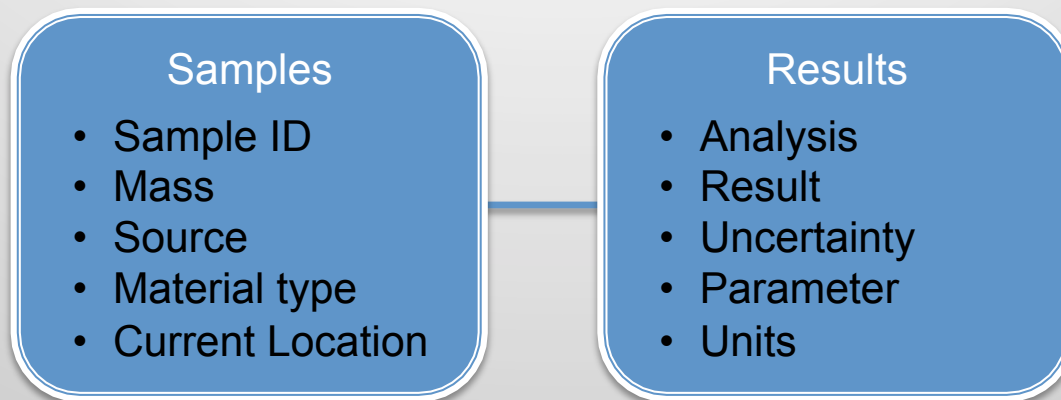
Uranium Sourcing Database -structure
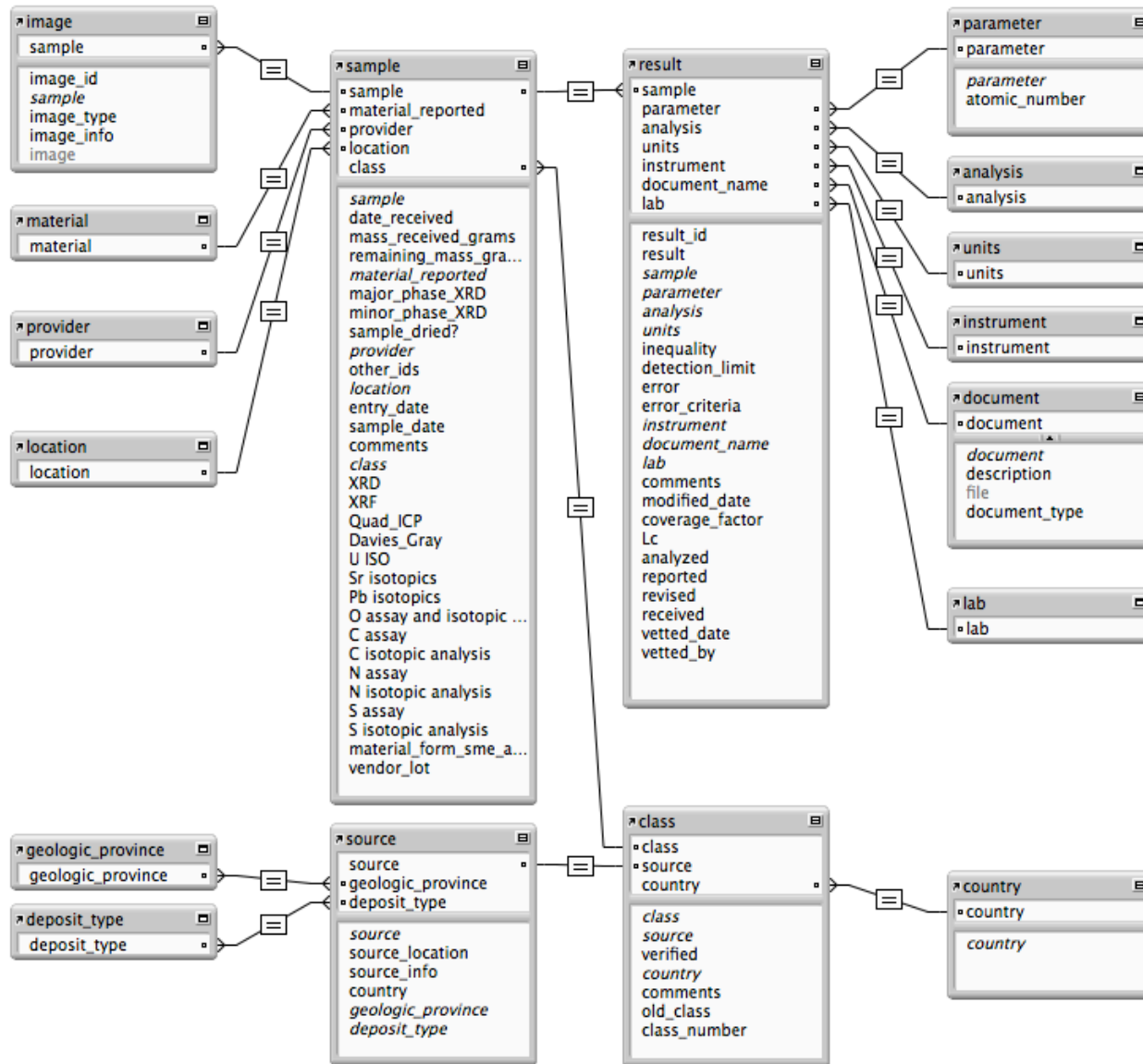
Analytical Laboratory to Database Interface

- U Sourcing Database effort includes lab analysis

- Analytical database lead

  - Receives

  - Vets

  - Formats

  - Uploads

- External analysis reports generated by database queries

- Database also used for tracking status of analyses

The U-Sourcing Database Organizes Samples According to Chemical, Physical and Isotopic Properties
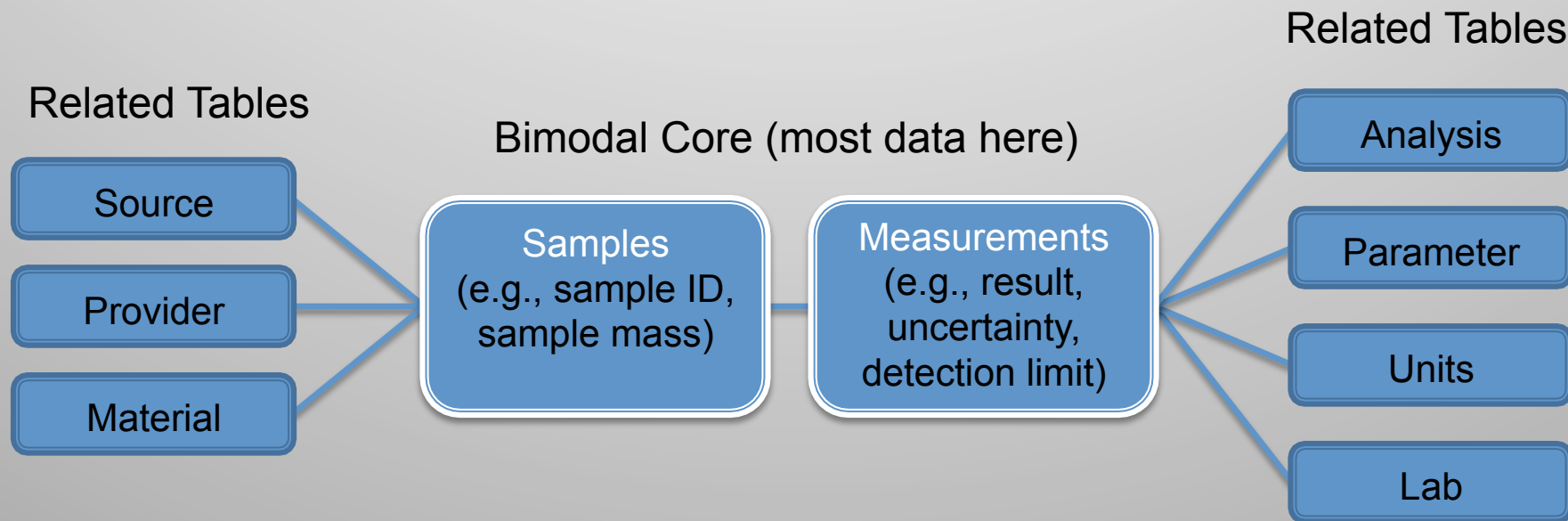
- **Bimodal core data model: Samples and Results**

Samples
- Sample ID
- Mass
- Source
- Material type
- Current Location

Results
- Analysis
- Result
- Uncertainty
- Parameter
- Units

- **Each sample may have ~70 results (measurements)**

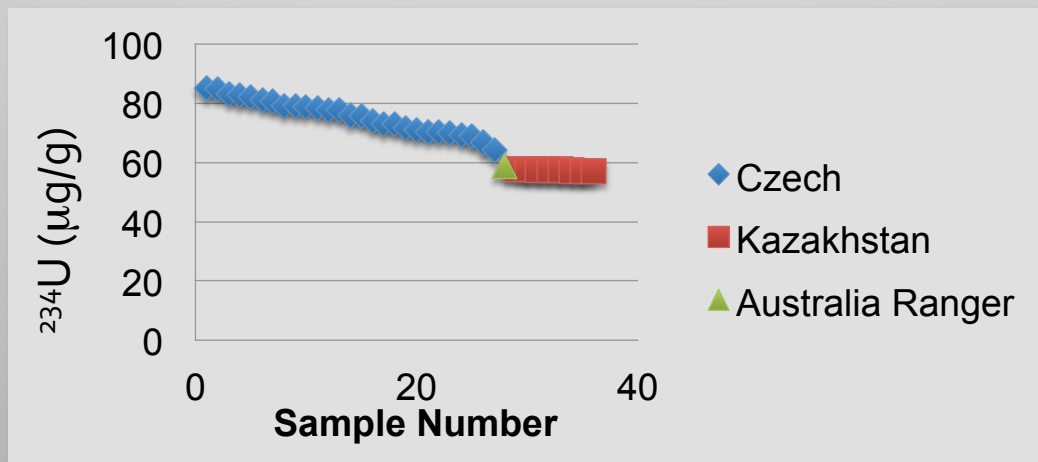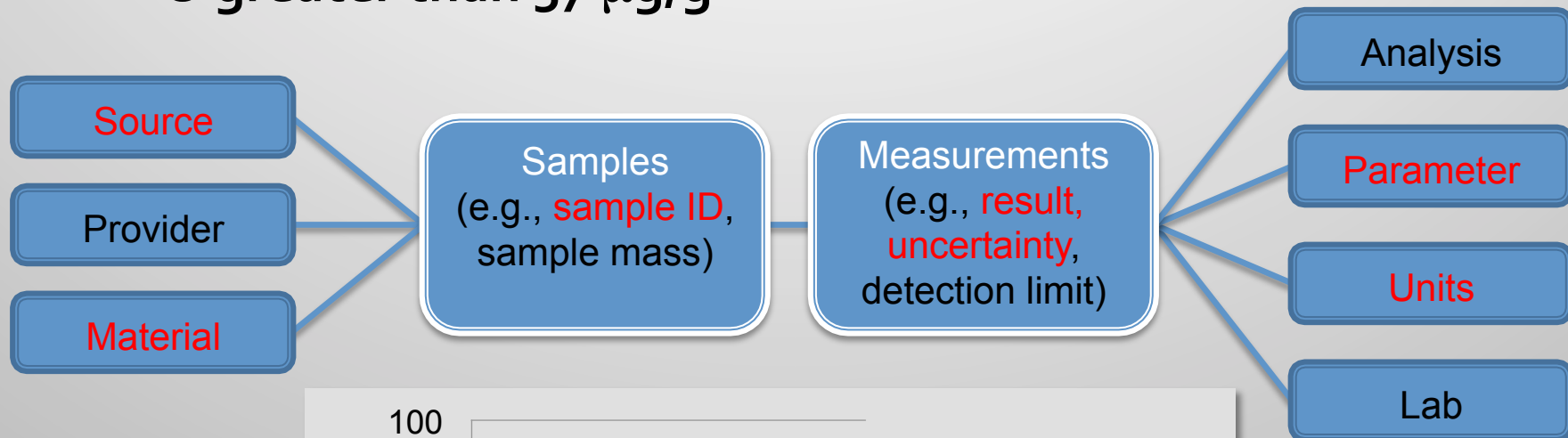| Sample ID	Result	Parameter	Unit
201	0.00733	$^{234}U/^{238}U$	Atom ratio
201	82,000	P	μg/g
201	70.5	U	wt. %

Database structure allows rapid querying of samples

- **Sample properties (data) are contained in a bimodal core structure holding variable measurements/sample**

- **Related tables contain meta-data and supporting information**

Related Tables

Related Tables

Bimodal Core (most data here)

Source		Samples (e.g., sample ID, sample mass)	Measurements (e.g., result, uncertainty, detection limit)	Analysis
Provider				Parameter
Material				Units
				Lab

Database Architecture Allows Both Simple and Multi-faceted Queries

- **Simple query: find source & sample ID for all U_3O_8 with [234]U greater than 57 µg/g**

Source

Provider

Material

Samples
(e.g., sample ID, sample mass)

Measurements
(e.g., result, uncertainty, detection limit)

Analysis

Parameter

Units

Lab

Selecting a database platform

- Many choices of relational database platform
  - Microsoft SQL Server
  - MySQL
  - Microsoft Access
  - Etc…

- Desirable features
  - User friendly
  - Multi-user (excludes MS Access)
  - Conventional (for ready supply of developers; excludes FileMaker Pro)
  - Existing institutional support

- Free versions of most platforms available
  - No customer support
  - But generally good "crowd sourced" support (forums)

Database data types

- Broad categories of data type
  - Text (text, varchar, char, etc…)
  - Numeric (int, bigint, tinyint, float, double, long, etc…)
  - Time/date (date, time, timestamp, etc…)

- The problem of significant figures
  - There is no numeric data type that meets the specific requirements of nuclear forensic analytical data
    — Numeric data types won't preserve significant figures properly
    — Rounding issues
    — Text will preserve sig figs, but at a cost (can't sort or do calcs)
    — Best solution we've found is to use text and implement a work-around for associated issues

Populating a NF database

- **Units and conventions**
  - Two ways to deal with units/reporting
    — Just import as is and then convert units as necessary after making a query
      – Requires less up-front work
      – Reduces potential for conversion errors creeping into database
      – Makes it much harder to perform searches on data
    — Standardize before importing to database
      – Requires more up-front work
      – Makes database much more useful
      – Can use file repository to preserve original data

The file repository

- All data is likely to be received in some document

- Those documents used to populate the database should be permanently stored in a file repository

- Database fields can point to source documents

- Uranium Sourcing Database does this by use of a document field in the Result table
  - Links to the document table
  - Links to the document file in the repository

Data Entry

- Manual vs batch/bulk upload
  - Some fields best updated manually
  - Most better with batch upload
    — Use a SQL (Structured Query Language) script -command line interface
    — Use upload utility in graphical user interface

- Preparation for upload
  - Data not likely received in database format
  - Manual copy and paste re-formatting tedious and error prone
  - Automation of standardized reporting template from analysts is ideal (e.g., MS Visual Basic macro in Excel)

Database user interface(s)

- Two categories of user interface
  - Off the shelf (OTS)
    - No development work necessary
    - Powerful
    - General purpose
    - Relatively difficult for non-specialist to use
  - Custom
    - Best for end-user with repetitive query needs
    - Requires significant development effort
  - Best solution is probably off-the-shelf admin interface and custom user interface(s) for non database specialists

Example OTS Interface: phpMyAdmin and MySQL

Database utilization

- Queries

  - Query is only the beginning; subject matter expert review/ interpretation is essential

  - Technical experts should directly query the database
    — For input to technical reports to external request originator

    — For research and development

  - For complex (e.g., multivariate) signatures, post-processing will be necessary

    — Export to Excel or analysis environment (e.g., MATLAB) for analysis

    — UOC signatures well suited to multivariate analysis

      – PCA

      – PLS-DA

Database summary reports

- Particular type of query with special requirements

- Two types of summary information
  - That which can be derived by a direct query of the data
    — E.g., number of samples in the database from a specific location
    — Should be easily accomplished with well designed database
  - That which requires synthesis and interpretation of database contents
    — E.g., number of sources added to the database in the past year
    — Requires a date added field in the appropriate table(s)
    — May require interpretation of whether a source is new
  - Recommend trying to anticipate such requests during development, since may require less than obvious fields

From database to analysis application

- Two ways to link analysis application to database
  - Direct queries of database
    - Requires rigorous cleansing and standardizing of database
    - Most flexible
  - Queries of "cached" data in form of datasets
    - Easier to ensure quality control of data utilized by application
    - Easier to track/document exactly which data were used as training set for a particular analysis/conclusion
    - More robust; no direct interface with database
    - Requires periodic re-construction of datasets (not automatic)

The Uranium Sourcing Database and iDAVE

- iDAVE is the pattern classification application that utilizes the data stored in the Uranium Sourcing Database

- Example of advanced post-processing of a database query

The Uranium Sourcing Database and iDAVE

Conclusions

- All but the most elementary nuclear forensic database should use a relational database system

- Internal vs. external database development entails trade-offs

- Simpler db structure is easier to maintain/modify

- Iterative design process recommended

- Preserving significant figures in database is tricky

- File repository is highly recommended

- At least two user interfaces recommended

- NF data may be best utilized with multivariate analysis

Acknowledgements

The LLNL nuclear forensics team:

Richard K. Bibby
Lars Borg
Amy M. Gaffney
Victoria G. Genetti
Julie M. Gostic
Richard Gostic
Patrick M. Grant
Roger A. Henderson
Ian D. Hutcheon
Gregory L. Klunder
Kimberly B. Knight
Carolyn Koester
Michael J. Kristo
Laurence Lewis
Rachel E. Lindvall

Naomi Marks
Audrey N. Martin
Kenton J. Moody
Christina E. Ramon
Erick C. Ramon
Martin Robel
Frederick J. Ryerson
Kerri Schorzman
Michael A. Sharp
Michael J. Singleton
Paul E. Spackman
Leonard T. Summers
Scott Tumey
Ross W. Williams
Paul T. Wooddy

Cutting-edge science validates forensic analysis capabilities
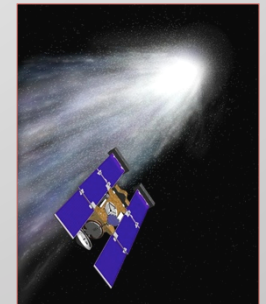
Isotopic Compositions of Cometary Matter Returned by Stardust
Science **314,** 15 Dec. 2006