ARI-MA Recasting Nuclear Forensics Discovery as a Digital Library Search Problem

NSF/Domestic Nuclear Detection Office Academic Research Initiative Grant Proposal from UC Berkeley

UC Berkeley proposes a grant focused on nuclear forensics discovery. Nuclear forensics is the science of identification of source and characteristics of smuggled nuclear materials possibly seized by authorities [AAAS/APS Report 2008]. Nuclear material identification are of utmost importance to international threat reduction and this project would be a significant step in this effort. The proposers wish to recast the nuclear materials identification process as a search problem against a digital library of standard nuclear materials samples and their digital signatures. This should prove useful to supply a conceptual, algorithmic approach to nuclear material identification and origination. Among the elements of the problem and questions to be resolved are:

Can the identification of nuclear samples which represent the nodes in a nuclear decay chain be approached as a weighted, labeled directed graph matching problem? Can the standard XML representations of chemical materials and compounds be extended to represent the nuclear isotope decay chain process? From existing representations, nuclear decay theory can be used to extrapolate downstream (times after analysis) or upstream (time prior to sample analysis, up to time zero). Can a simulation approach be used to develop a pseudo-digital library derived from the differential equations of nuclear decay to test algorithmic research without having to operate under the veil of secrecy?

The UC Berkeley PI has extensive experience in search algorithms, with past funding from NSF and DARPA. The Co-PI is a digital library expert who can implement an educational component on nuclear forensics within an information search curriculum. The senior consultant has over 20 years experience in forensics, including 10 years experience with nuclear materials at Los Alamos National Laboratory. This highly experimental grant would run for three years to answer these questions, develop a digital library of nuclear signatures, and field educational outreach to encourage more search specialists to dedicate research attention to the problems. Cooperation has been obtained from projects at Los Alamos National Laboratory (nuclear materials samples library) and Lawrence Berkeley Laboratory (nuclear ontology). During the grant, the PI would expect to develop international relationships with researchers in the IAEA and the EU Joint Research Commission's Institute for Trans-Uranium Elements (ITU) at the University of Karlsruhe. The results of the project should include open-source code for nuclear forensics search to be made available to DNDO agencies, national laboratory groups and to streamline the process of nuclear materials identification.

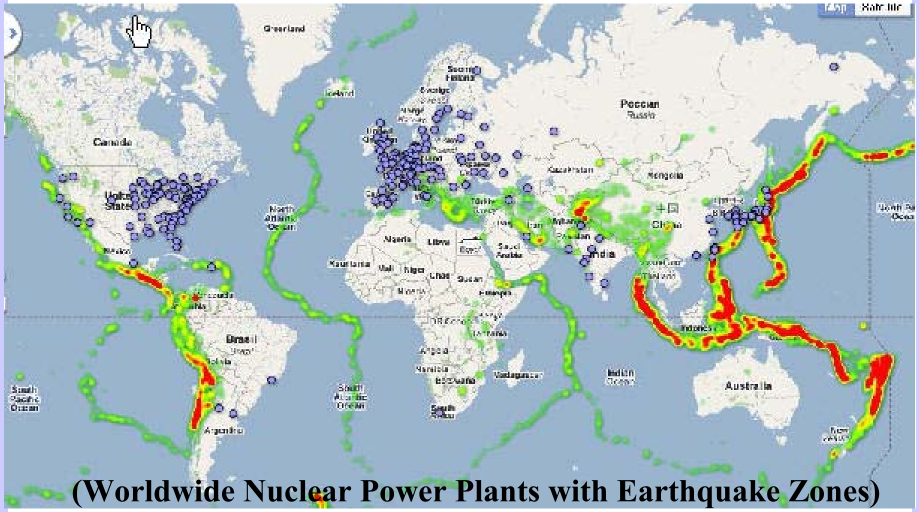
Intellectual Merit: This grant proposes the first computer science algorithmic approach to the nuclear forensics search problem as a special directed graph matching problem. The success of this grant will encourage other computer scientists to work on nuclear forensics search

Broader Impact: Success in nuclear forensics search is a critical component to fighting terrorist activity and preventing disastrous individual terrorist nuclear attacks.

Reference

APS/AAAS Joint Working Group: **Nuclear Forensics: Role, State of the Art, Program Needs,** 2008.

RECASTING NUCLEAR FORENSICS DISCOVERY AS A DIGITAL LIBRARY SEARCH PROBLEM



(Worldwide Nuclear Power Plants with Earthquake Zones)

http://maptd.com/worldwide-map-of-nuclear-power-stations-and-earthquake-zones

BROADER IMPACT

- Introduces the search community to the area of nuclear forensics
- Creates deployable software for signature building and matching
- Helps in the struggle against mass destruction terrorism, in particular
- will support the activities of the Nuclear Smuggling International Technical Working Group

INTELLECTUAL MERIT

- Applies modern search technology to the nuclear forensics matching problem
- Creates new models for matching (graph, classification, rule-based)
- Creates nuclear signature building methods
- Creates educational materials bridging between search specialty disciplines and nuclear forensics scientists

SCHEDULE AND BUDGET

- Year 1: Develop models, create international partnerships, workshop on scientific search
- Year 2: Refine and test models, prepare educational materials, pilot software
- Year 3: Field a course in nuclear forensics combining search and radio chemistry, field test software
- $1.041M (3 senior scientists, 3 graduate student research assistants for 3 years)

ARI-MA: RECASTING NUCLEAR FORENSICS DISCOVERY AS A DIGITAL LIBRARY SEARCH PROBLEM

Primary NSF Directorate: CISE
Secondary NSF Directorate: ENG

Introduction

The dirty bomb scenario and nuclear forensics

One of the most frightening scenarios of individual terrorism has been the deployment of a conventional explosive device laden with nuclear material, a so-called "dirty bomb." This device (technically referred to as a radiological dispersal device or RDD [GAO 2009]) would release major amounts of radiation poisoning onto an urban population, causing untold human suffering. According to [Mayer, Wallenius and Fanghänel 2007] "Since the beginning of the 1990s, when the first seizures of nuclear material were reported, the IAEA recorded more than 800 cases of illicit trafficking of nuclear or other radioactive materials." Security agencies worldwide continue to work to prevent this scenario from happening. The two aspects of prevention are detection and forensics. Millions of dollars are being spent on improvement of devices to detect contraband radioactive material which might be hidden in shipping containers. The flip side of detection is forensics – if a significant amount of smuggled nuclear material is seized, can it's origin be traced to both track down the would-be terrorists and to prevent further smuggling activities. To do this, a seized sample can be analyzed to ascertain its "nuclear signature" which can be compared to an archived library of digital nuclear signatures which have been abstracted by radio-chemical analysis of a large number (tens of thousands) of nuclear samples from uranium mines and nuclear manufacturing facilities worldwide. Currently a number of laboratories domestically (Lawrence Livermore, Los Alamos, Pacific Northwest National Laboratory) and internationally (Institute for Trans-Uranium Elements in Karlsruhe, Germany) are assembling such samples, subjecting them to analysis, and creating digital libraries of their nuclear signatures (see [Smith et al 2004] as an example of published work).

Finally, the nuclear power plant disaster earlier this year in Fukoshima, Japan has made it clear that, almost unnoticed, large amounts of spent uranium fuel rods are being stored at nuclear power plants worldwide. According to [West 2005], "An RDD is any device that uses conventional explosives to spread radiological material with the intent to cause damage or injury. One of the more hazardous types of material that could be used in an RDD is spent nuclear fuel." Potential availability of this spent nuclear material poses an additional threat to international security of nations and the globe itself.

A case of nuclear murder

The death by radiation poisoning of Alexander Litvinenko in 2006 ushered in a "new era of nuclear terrorism" [Patterson 2007]. Litvinenko was given a lethal dose of Polonium 210 secreted in a drink while having lunch in a London restaurant [Goldfarb and

Litvinenko 2008] He died three weeks later. Because Polonium 210 emits alpha particles in its radioactive decay, lethal amounts of ^{210}P can be shielded within a paper box – the decay produces fatal radiation only when directly exposed to the internal organs of a human being.

Nuclear Forensics Human Capital

Recent reports [GAO 2009] and [NRC Nuclear Forensics 2010] have emphasized the need for fostering and preserving human capital in the nuclear forensics area by coordinating between the various agencies involved in nuclear forensics investigation, as well as development of educational initiatives to build our nation's capacity to meet the nuclear forensics challenges. The emphasis, however, is on expanding human capital in nuclear sciences, in particular, radiochemistry and omits the contribution that could be made from computer and information science. This proposal argues for another dimension of human capital -- developing expertise for nuclear forensics matching as a search problem against a specialized digital library of nuclear sample signatures. The DNDO-ARI solicitation makes specific reference to an educational component, but its major emphasis is also upon radiochemistry, detection technologies, and rapid response and recovery. This proposal's educational component will focus upon introducing information scientists, and, in particular, specialists in search technology, to the unique challenges of nuclear forensics matching. Professionals from the search community may have significant contributions to make in developing new algorithms and database structures which facilitate rapid matching techniques which scale to searching digital libraries of up to one hundred thousand nuclear signatures representing analyzed nuclear samples.

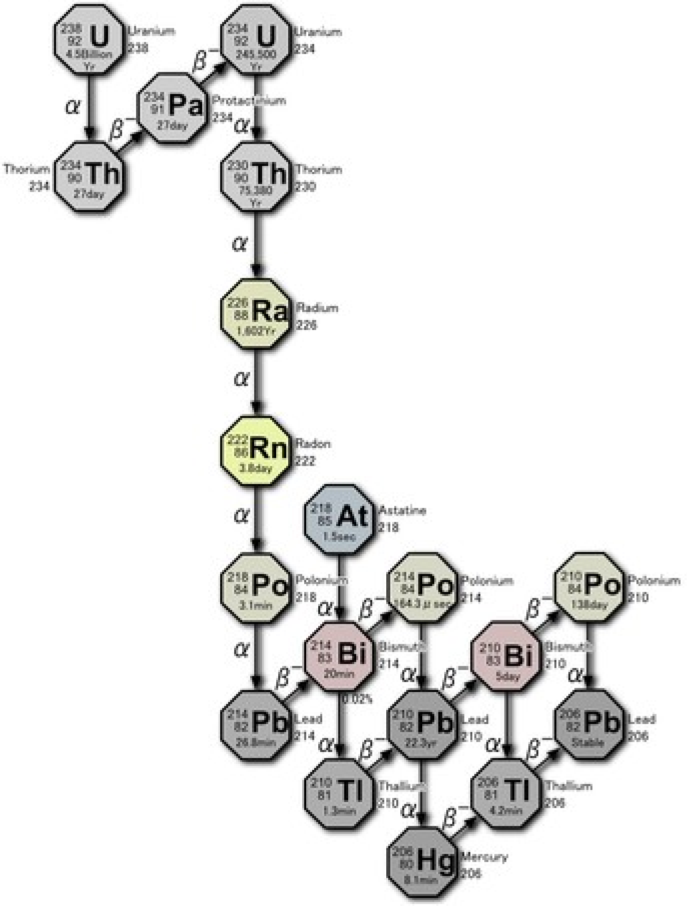
Approaching Nuclear Forensics as a Search Problem

Given a nuclear sample obtained from whatever process (interdiction, for example), the problem is to identify its source. Such identification requires clues to match against a dataset of samples for which sources and compositions have been identified. The process, abstractly, is not that different from matching fingerprints or DNA samples from a crime scene – both require a library against which the match will be made, and specialized matching technologies which execute the search. In the case of nuclear forensics, the library will consist of two broad classes of samples: Uranium Oxide Composites (UOC) from Uranium mining operations, or Radioactive Materials which have been produced by sophisticated nuclear manufacturing (as from a nuclear power plant). An example of the latter is the Polonium used to poison Liventenko, which UK authorities eventually traced to a particular nuclear facility in Russia. Polonium isolation would require special processing in order to extract a "pure" specimen, yet other trace elements from the manufacturing process, would provide clues as to the origin of the specimen. The primary products of a nuclear reaction process are Plutonium and various Uranium isotopes. Plutonium is produced from the burnup of the original Uranium fuel. According to the 2008 American Association for Advancement of Science report [AAAS/APS 2008], "even if two plutonium samples are similar in one respect ... their origin may still be identified based on an analysis of selected isotope ratios ...

(which) change significantly with the age of the material."   Table F.1 of the report further classifies reactor types used for plutonium production as "graphite-moderated", "heavy-water-moderated", or "enriched-driver-fuel", independent of the geolocation of the reactor. Thus simple search discrimination can quickly identify the type of reactor used for production of the material.   According to the report, ratios of various plutonium isotopes (238,240,241,242) with respect to Pu-239, definitively identify the reactor type. These ratios can provide a fundamental clue in discrimination between reactor types.

A Directed Graph Search Model

Both plutonium and enriched uranium samples are characterized by their nuclear decay isotope production.   Figure 1 displays the decay chain for U-238



The figure shows the parent and child isotope in each decay reaction, as well as the type of decay (alpha, beta).

If we examine this figure abstractly from a mathematical concept point-of-view, it appears to resemble a directed graph, in particular, a directed acyclic, weighted graph. The nodes of the graph are the particular isotopes of decay. The edges of the graph display the particular decay directions.   The graph is 'directed' because decay proceeds from parent isotope to child (product) isotope. It is 'acyclic' because there are no cycles in the decay chain, i.e. no child isotope will subsequently decay to a parent isotope, producing an infinitude or cycles. Finally the edges are 'weighted' in that the weights applied to the edges between parent and child consist of the decay rates (half-life). Thus, given the digital signature of an interdicted sample of nuclear material siezed by authorities, we wish to identify its origin, and the conceptual search problem becomes one of a graph matching, in particular, of matching between weighted directed acyclic graphs. For background on graph matching, see [Bengoetxea 2003].

.

Represented as a Graph G = (V,E), a nuclear sample consists of a finite number of vertices (sometimes referred to as nodes) v_1 ... v_n representing elements in a decay chain, i.e. For Uranium above, n=19, v_1 = ^{238}U v_2 = ^{234}Th and v_{19} = ^{206}Pb the terminal stable element of lead. Associated with each vertex at time $t_{m,}$ is an amount $m(t_m),$ the measured mass of the element at the time of measurement. The edges (or arcs) between elements represent the decay direction: thus $e_{7,8}$ = (^{226}Ra,^{222}Rn), representing the decay path from Radium to Radon.

We can simplify our representation by saying that a seized material sample at time $t_{m,}$ is referred to as $G_s(t_{m,})$. Let us further say that there exist a digital library of k samples each measured at different times LIB={$G_1(t_1)$ .... $G_k(t_k)$}. We wish to match the seized sample to appropriate library samples. But there are differences in times of measurement – to do the match we have to forwardly compute each of the library samples from $t_{k,}$ to time t_m (or backwardly compute the seized sample from time time t_m to time $t_{k,}$). Thus we seek a similarity function: SIM ($G_s(t_{m,})$,$G_i(t_i)$ ε LIB) = SIM($G_s(t_i)$=f_b($G_s(t_{m,})$,$G_i(t_i)$ ε LIB) for the ith sample in the library and where f_b is the backward computation function.

Of course this is the simplest model – in reality, all samples will have additional geolocation clues L (manufacturing impurities or geologically specific elements) which may or may not have a time dependency. Thus G = (V,E,L) for a more complex model. It will be the first priority of the project to develop a time-varying graph model which accurately captures the complexity of the nuclear forensic discovery (matching) problem.

The proposers have no experience in graph matching, but have found one of the classic papers in weighted directed graph matching is by Shinji Umiyama [Umeyama 1988] which uses eigen decomposition to analytically solve the matching problem. It does not deal with the complexity of temporal dependent graph attributes. Literature searchs for the subject "time dependent graph matching" yielded papers on traversing transportation networks where time dependency (e.g. rush hour) affects the shortest time traversal of the graph. It seems that our model of time-dependent graph matching may provide a new vision of graph matching which have a temporal component.

Nuclear Forensics Matching as an Automatic Classification Problem

Another, alternative approach to the forensics matching problem is that of automatic classification within machine learning. Nuclear materials samples are geocoded as to origin (the classification) and a sample under investigation is automatically classified as to geospatial location from a finite number (hundreds) of locations (either nuclear processing facilities, nuclear power plants, or uranium mines). If the training sample (the digital library of nuclear sample signatures) is in the tens of thousands, machine learning to identify location by features (presence and amount of decay byproduct elements, trace elements related to processing, or significant elements characterizing the geology of particular mines). This approach has been taken by Roble and associates [Robel, Kristo and Heller 2009] of Lawrence Livermore National Laboratory (LLNL) – their work will be discussed further below in our review of prior research. Machine learning for text categorization has been extensively studied by information retrieval specialists, applying the following methodologies "including Support Vector Machines, k-nearest neighbor, ridge regression, linear least square fit and logistic regression" [Yang, Zhang and Kisiel 2003] -- scalability of these algorithmic approaches to large numbers of categories is the focus of that paper, and will also be a focus of this research. A large part of the art of automatic classification is feature selection, i.e. what are the particular clues which characterize the class (geolocation) – for spent fuel it may be Plutonium isotope ratios or metallic alloys used in container manufacturing, while for Uranium ore it may be geologically specific elements which are found in the ore samples. The Pis have experience in applying machine learning techniques, in particular Dunning's Log Likelihod Ratio method [Dunning 1993] to automatic text classification in an interactive environment [Gey et al 2001].

Nuclear Forensics Identification by Rule Based Matching

A completely different approach to matching is to identify what human experts would do in attempting to find the origin of a nuclear sample based upon forensic signatures. In this approach, an ordered set of rules can be created which can be instantiated in software which will execute the matching process against the existing library of signatures.

Axiomatic rule based knowledge generation can be a starting point for signature formulation. The assumption deduced is that the isotope measurements reported from a laboratory analysis of nuclear material can be projected into the past and future according to radioactive decay chain binary relationship rules between parent/daughter pairs of isotopes. In adopting this starting point the proposed work can implement decay chain rules on the set of reported isotopes in a provable automated routine. Reasoning can then be imposed on the new set of data generated since it is valid as a consequence of rule based axiomatic computations.

The isotopic constituents of an actinide morph due to radioactivity therefore the laboratory measurements of a material will differ from one measurement event to the next. The signature of interdicted materials will be cataloged in laboratories at the time of measurement, however matching reports to find materials that are from the same samples but different from times entails constructing a process to map one to the other. For this task the proposed work will employ an axiomatic system where (like Euclidean geometry) it is postulated to project into another time using deduction. Since the initial measurements are facts and that the decay chain activity is scientifically provable, then the projection that is a result of this activity is a valid morph of the material. The proposed work will achieve this process by creating a signature using the initial measurements from nuclear samples, compute the projections of this material through time stages by building a graph of radioactive parent/daughter relationships which will be used to construct the postulated future and past states of the material. The construction works over measurements and scientifically factual decay chain processes therefore it can be reproduced and is provable and therefore forensic evidence. The forensic aspect of the proposed work relies on the deductive process of geometry since it asserts the existence of objects and provides a method to construct them.

We will develop and evaluate the effectiveness of all three approaches to the nuclear forensic discovery and identification problem.

Data sets for nuclear forensic search evaluation

Since this proposal is about search and search evaluation, it is crucial to have some data upon which to evaluate the effectiveness of our methodologies. We expect to include, among others to be assembled, the following data sets:

- Radioactive decay chain parent/daughter isotopes and radioactive properties
  - This data set will be used to build the graph of radioactive decay that is the proposed work's nuclear material prediction routine. Nudat, National Nuclear Data Center, Brookhaven National Laboratory http://www.nndc.bnl.gov/nudat2/ The entire set of over 165 thousand records was obtained directly from *Database Manager and Web Programming: Alejandro Sonzogni,* NNDC, Brookhaven National Laboratory, sonzogni@bnl.gov by senior investigator Ms. Sutton for a data model she designed at Los Alamos National Laboratory (LANL). The data set contains each isotope for every element, decay daughters, radiation type, half life and energy levels.
  - The data set is unclassified

- *Actinide measurement test results.*
  - This data set will be used to design the proposed project's signature formulation routines. It is produced by researchers from six national laboratories and the Atomic Weapons Establishment. Ms. Sutton already uses the data set for a statistical collaboration application that she is developing and a commitment has been agreed upon whereby the data set will be available to the proposed project [*see supplemental documents for the LANL letter of commitment*]. The data set is a report of all measurements made on actinide samples. The tests are blind and re-run on the same materials every year in order to validate practices and calibrations. The results lead to standardization of the testing process and therefore are the most reliable data set available for the design stage of the proposed work. The set contains the assay actinide and it's isotopic and trace element amounts. It is an example of the input data needed from any user of the finished product. It is a perfect application primer.
  - The data set is unclassified

- *Uranium Oxide QC data set.*

- This proprietary dataset was used by LLNL researchers for their principal components analysis to infer location from indicator trace elements which are geospecific [Robel, Kristo and Heller 2009]. The data set has 3436 sample measurements for 21 sites.
- The data set has licensing restrictions to Department of Energy. Our team has contacted Dr Robel about the availability of the data and will explore whether existing cooperative agreements between LLNL and UC Berkeley's Nuclear Engineering Department can be extended to allow access to the data.

- *Simulated Data Sets*
  - In their paper "Characterization of nuclear fuel using multivariate statistical analysis" [Robel and Kristo 2007], the authors state:
    "It is impractical to obtain such data from real nuclear fuel sample analyses. Instead, the ORIGEN-ARP code package (Bowman, 2000) was used to predict U and Pu isotopic compositions at various burnup values."
    In the AAAS/APS report on nuclear forensics [APS, AAAI 2008], simulated data was also used to characterize different types of nuclear reactors. The computer code used there was called MCODE. It is similar in function to ORIGEN-ARP but uses more computing resources because it implements a Monte Carlo method.
    For this project we expect to also generate test data sets using one or the other of the code packages mentioned above. Predictive data generated by simulation models has the advantage that it can produce large amounts of 'observations' which can be used to train Baysean classification approaches (machine learning) to the nuclear forensic matching problem. In our approach, we also will consider introducing random variations into the generated values to simulate measurement errors which might be found in actual data.

XML Metadata Structures

All metadata structures describing the nuclear materials data will be in XML conforming, wherever possible, to existing XML standards (e.g. CMLSpect [Kuhn et al 2007]). A major XML schema standard for atomic and molecular structure representation has just been released by the IAEA, XSAMS (Schema for Atomic, Molecular and Solid Data) [IAEA 2011, INDC 2010].

E. Sutton has prepared an unclassified nuclear ontology in XML format which will be extended during the project. This ontology should be a useful supplement to Lawrence Berkeley National Laboratory's Simulations Algorithms and Models (SAM) Program Office Ontology Project. In the supplementary documents we have a letter of support from LBL describing its potential contribution to their development. PI Larson has extensive experience with XML search methodologies, having participated in a number of worldwide evaluations (the INEX evaluations) of such search techniques; he uses XML as the underlying structure supported by his CHESHIRE series of indexing and search engines.

Geospatial visualization of nuclear materials locations

The following map displays worldwide nuclear plant locations overlaid with earthquake zones.[1]

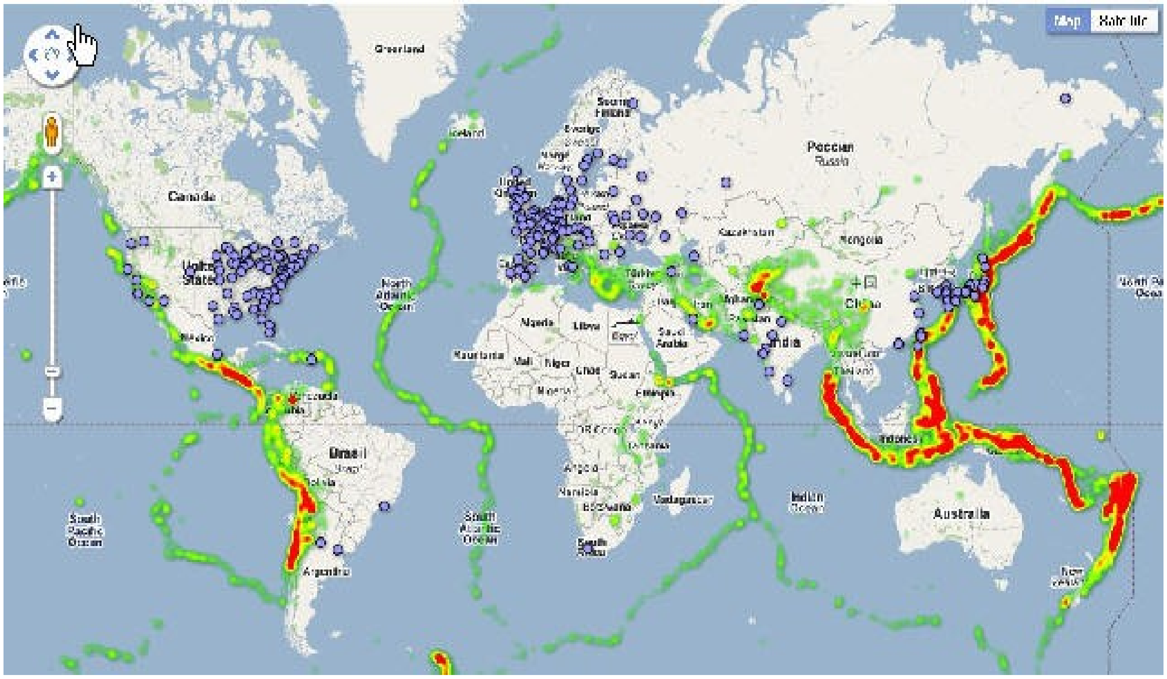


Figure 2: Map of Worldwide Nuclear Plants overlaid on Seismic Zones

The PIs have extensive experience in map displays for location-specific information from text, coupled with quantitative information (census data). They have used Google Earth to prepare dynamic maps of geotemporal information. We expect to prepare significant geospatial visualizations of both nuclear power, spent fuel storage, and uranium mines to help make our work understandable to the public.

Prior Research

Most of the prior work on nuclear forensics has concentrated on chemical analytical techniques and radio-chemistry. Much of this work is explained and summarized in the book **<u>Nuclear Forensic Analysis</u>** [Moody, Hutcheon and Grant 2005], which covers not only nuclear reactions and characteristics of reaction byproducts as clues to source, but also covers techniques for analyzing organic and inorganic materials which may provide clues to where the forensic evidence came from. For example (page 283) “deviations in the isotope ratios of elements such as O and Pb provide signatures that can be used for geolocation.” and chapter 19.7 (pp 395-397) *Geolocation and Route Attribution: Real-World Examples.* The references found in this book give a broad overview of the published work in analytical techniques.

[1] **Source:** http://maptd.com/worldwide-map-of-nuclear-power-stations-and-earthquake-zones/

A major source of published research on nuclear forensics discovery and attribution are the papers by Klaus Mayer and associates at the Institute for Trans-Uranium Elements (ITU) in Karlsruhe Germany. In [Mayer, Wallenius and Fangänel 2007], the authors date 1992 as the first instance of an ITU analysis of seized nuclear material, an analysis "sufficient to attribute the intended use of the material being fuel pellets for a Russian ...nuclear reactor…(with) the possible origin of the material .. either UMP in Kazakhstan or Elektostal in Russia." The paper also describes the evolution of analysis of geolocation methods based upon impurities, in one case "having concentrations >1000ug/g of … Al,Ca,Cr,Fe,Mg,Mo,Na,Ni and P."

The ITUs 2005 paper [Mayer, Wallenius and Ray 2005] describes the major analytical techniques applied in forensic investigation, including radiometric methods, gamma and alpha spectrometry, mass spectrometry, and microstructural techniques. Their 2006 paper [Mayer, Wallenius and Ray 2006] presents two case studies of seized samples, one uranium pellets from Lithuania and the other uranium powder from the Czech republic. The paper reports upon a 1 day, 1 week and 2 months report of analyses conducted on the samples which convincingly determined the origin of the materials. Their 2007 article [Švedkauskaite-LeGore et al 2007] analyzed worldwide uranium ore samples using hierarchical cluster analysis based upon amounts of Zirconium and Chromium in the samples. They then also analyzed various lead isotope levels, reasoning that "Fortunately, lead isotope 204Pb is a measurable lead isotope, which is not of radiogenic origin and can, therefore, be used to determine the amount of natural lead present in the sample."

A quite readable overview of the ITU's work and the elements of nuclear forensics can be found in Mayer's 2009 presentation [Mayer, Wallenius and Schubert 2009] at the IAEA nuclear security forum in Vienna.

The essence of the published works by ITU and associates can be used by this project to determine the essential clues for matching and deciding geolocation. The actual forensic matching that had been published by the group is relatively small in number (a few case studies), whereas we wish to work with much larger datasets to produce and evaluate scalable algorithms.

The published work which most closely resembles the proposed work is that of Martin Robel and colleagues at Lawrence Livermore National Laboratory. Their work has specialized in principal components analysis for reduction of the dimensionality of the problem. Their 2007 report [Robel and Kristo 2007] uses simulation models of the reaction process to produce data, upon they apply the dimensionality reduction. Their 2009 paper [Robel, Kristo and Heller 2009] is the first published large scale test of a geolocation strategy and evaluation of same. The authors claim their statistical discrimination approach outperforms standard KNN classification. They test against two levels of classification, by country and by site. The PIs have performed a multilevel relevance evaluation for patent classification [Gey and Larson 2008]. The QC dataset which LLNL used (described above) is what is known in the machine learning community as an "imbalanced dataset", because the top five sites have about 37 percent of all the samples (1277/3436), and seven sites have fewer than ten samples (including Kazakhstan site 4 with only one sample). Not shown in the paper is that the median sample size is 27 samples, the mean is 81.8 with standard deviation 111. Figure 3 at the end of the paper is a new graphical display of Table 2 from the paper, overlaid with the mean of the samples per site.

The PI has experience working with imbalanced datasets. In [Kim et al 1999] he and his students used sampling to the mean and median for classes with training observations of size greater than the mean to re-balance an automatic text classification algorithm. Since that time considerable attention was paid to the problem of imbalanced datasets by the machine learning community [Chalwa, Jakowicz and Kolcz 2004]. We would expect to apply the latest techniques for machine learning in the presence of imbalanced datasets in development of our algorithms.

The UC Research Team

The UC Berkeley research leaders consist of three well-qualified individuals with complementary skills. PI Dr Fredric Gey has developed document ranking algorithms for search based upon logistic regression probabilistic modeling techniques. He has continued to apply these techniques to cross-language information retrieval and has co-chaired three SIGIR (Special Interest Group on Information Retrieval of the Association for Computing Machinery) workshops (2002, 2006, 2009) on multilingual information access. Dr Gey also has 30 years experience with management and curation of datasets and databases (primarily in the social sciences), first with Lawrence Berkeley Laboratory (LBL) and then as Assistant Director of the UC Data Archive & Technical Assistance within UC Berkeley's Institute for the Study of Social Issues. Dr Gey's undergraduate coursework (B.S. in Mathematics) at Harvey Mudd College included 2 years of both chemistry and physics. Dr Gey is a guest scientist in the Advanced Computing for Science Department at Lawrence Berkeley Laboratory and, in that capacity, may be able to access the cloud computing resources at LBL for specific tasks which fit both this DNDO-ARI project and the LBL mission. He can also coordinate cooperative activities between this project and relevant LBL groups. Dr Gey will serve as overall coordinator of the project and overseer of execution of the data management plan for the project data.

Co-PI Professor Ray Larson has developed query expansion techniques using clustering algorithms similar in nature to latent semantic indexing. Professor Larson has also been a pioneer in the field of geographic information retrieval (GIR), as evidenced by his paper “Spatial Ranking Methods for Geographic Information Retrieval (GIR)” which was awarded best paper in the 2004 European Conference on Digital Library Research. Professor Larson teaches the data management course at the School of Information at UC Berkeley and can coordinate with PI Gey in implementation of the data management plan. Professor Larson will also be the lead in implementing an educational outreach program for the project. He will introduce the special concept of nuclear forensic search into both his data management course at the UCB Information School and the Fundamentals of Information course at the school .

Dr Gey and Professor Larson have a long history of collaboration together on a series of DARPA and IMLS grants. They currently are co-organizers of evaluation of geotemporal search in both English and Japanese as part of the Asian language search evaluation series NTCIR. For more information see: http://metadata.berkeley.edu/NTCIR-GeoTime Prior to that they co-organized GeoCLEF evaluation of geographic search in European languages. They both have a lengthy experience with evaluation methodologies, and often serve on conference review committees reviewing papers on search evaluation.

Ms. Electra Sutton has over 20 years experience in the technology of forensics, first with Federal Bureau of Investigation and currently with Los Alamos National Laboratory where she has served as chief architect for a $30million project assembling nuclear samples worldwide. If this project is funded, Ms. Sutton will take an academic leave from Los Alamos to work 35% time on this project. Ms. Sutton holds numerous classification qualifications (Top Secret, Q, SC) with the Department of Energy and other agencies. Ms. Sutton will function as liaison to National Laboratories and Federal Agencies where data resides in a classified environment. In this role, Ms. Sutton will take search and classification algorithms developed by the project and transfer the technology to the classified environment. Where feasable, she will also test the algorithms upon classified data and bring the test results (cleared for anonymity) to our published research in the open literature. Ms. Sutton will function as the subject area expert for nuclear forensics data. She will also function as the chief implementation supervisor for the project, providing overall XML data structure for nuclear isotopes and other non-radioactive elements which characterize geographic signatures of for both Uranium mines and nuclear reactor production facilities. Finally, Ms. Sutton will oversee computer code production by graduate students and undergraduate students hired by the project.

The research team will be also include three graduate student research assistants. In recruiting these graduate students, we will primarily seek graduates studying in information technology areas (information school or computer science) who have an undergraduate background in chemistry or physics, or graduate students in chemistry, physics or nuclear engineering who have an interest in informaton technology and its application to their specialties. We have not included undergraduates in the first year, because the work is highly experimental – if the objectives of the first year are realized, we may seek supplemental funds for the following years. We will also apply for Research Experience for Undergraduates funds from the REU program.

Modes of Dissemination and Education.

The PI's expect to write and present papers at professional search and digital library conferences (SIGIR – the Special Interest Group on Information Retrieval of the Association for Computing Machinery, CIKM – the Conference on Information and Knowledge Management, JCDL – Joint Conference on Digital Libraries, an ACM and IEEE cooperative conference) which describe project results and rigorous evaluation of nuclear forensic search models developed by the project. At SIGIR-2012 in Portland, the PIs will propose and organize a workshop on “Scientific Search” of which nuclear forensics search will be a component. Another venue for dissemination would include the Scientific and Statistical Database (SSDBM) conference.[2] During years 1 and 2, course materials will be developed to introduce computer and information scientists to the mechanisms and data structures for nuclear forensic search. During year 3, the PIs expect to collaborate with UC Berkeley’s Nuclear Engineering Department in fielding a course on nuclear forensics.

Results of prior NSF funding

PI Dr Fredric Gey had a small role as a Senior Person on NSF grant # 0637122 “Eco-informatics Project: Semantics Management and Semantics Services**.”**(2007-2010). For this grant, Dr Gey developed a syntax to capture mappings between classifications and a mathematical model of mappings between instance-based classification (special hierarchical ontologies) systems. See Dr. Gey's CV for specific references. Dr. Gey was PI of NSF grant “Probabilistic Document Retrieval for Full-Text Document Collections Using Logistic Regression,” (1996-2000) which developed and tested new probabilistic approaches to text retrieval based on logistic regression and showed that logistic regression retrieval algorithms work well for English and foreign language retrieval and also crosslingual retrieval using English to European languages (French, German, Italian,Spanish) Japanese to English and English to Japanese/Chinese/Korean and English to Arabic bilingual retrieval. The grant provided support for five graduate students, including full support for the PhD research of Dr. Aitao Chen, now with Yahoo research and some support for Langjie Jack Xu who later became Vice President and Director of Search Technologies for EBay Inc. This grant led to two significantly larger, follow-on research grants on translingual information management from the DARPA Information Management and TIDES programs (1997-2003).

Ray R. Larson was the Principal Investigator on the DLI-International grant entitled "Cross-Domain Resource Discovery: Integrated Discovery and Use of Textual, Numeric, and Spatial Data" (NSF Award No. IIS-9975164, $306,004 10/1/99-9/30/02). This undertook the investigation and development of methods for distributed information retrieval across a variety of different servers and data types, and forms the foundation of the current proposal.

Prof. Larson was a faculty investigator on the NSF/NASA/ARPA Digital Library Initiative Project entitled "The Environmental Electronic Library: A Prototype of a Scalable Intelligent Distributed Electronic Library" (PIs: Robert Wilensky and Michael Stonebraker, Award No. IRI-9411334, $4 million 9/94-8/98).

He was also a faculty investigator affiliated with the Digital Libraries Initiative, phase 2 grant to U.C. Berkeley (PIs: David Forsythe, Robert Wilensky, $5 million), entitled Reinventing Scholarly Information Access” but received no direct support from that project. This project developed a prototype digital library focused on the California environment accessible at http://elib.cs.berkeley.edu. The project explored many areas

[2] SSDBM 2011 in Portland: http://www.ssdbm2011.ssdbm.org/

ranging from document decoding and scanning to information categorization and retrieval. The project, inter alia, showed that effective retrieval from OCR data is possible to implement (The Cheshire II system is the primary text search and retrieval engine for the project). The project also made several breakthroughs in computer vision research and in the design of scalable information retrieval methods for image retrieval (also using Cheshire II in conjunction with selection and matching methods developed by computer vision researchers).

Summary

A literature search by the PI's has revealed that neither the database research community nor the information retrieval research community are aware of the unique search challenges posed by the nuclear forensics identification problem. Each of these communities have significant resources which could be brought to bear in solving the fundamental problems of large-scale nuclear forensic discovery. Among other goals, a primary focus of this ARI proposal is to interest these communities in the issues of nuclear forensic identification and to supply education.

Intellectual Merit

This is the first project of its kind to apply modern search technology to the nuclear forensics matching and identification problems. The methodologies explored (directed graph similarity, automatic classification, and rule-based matching) are expected to break new ground in providing rigorously evaluated approaches to nuclear signature matching. It is expected that published results from the project will encourage other computer search specialists to turn yet further attention to the nuclear forensics area. Just as in traditional forensics (fingerprint matching, DNA matching) benefited from algorithms developed by information search specialists in their area, we expect the area of nuclear forensics to be significantly improved by focussed attention by researchers in the search area.

Broader Impact

Nuclear forensics, expecially pre-detonation forensics, are a crucial component of prevention of nuclear terrorist incidents. International threat reduction, safeguards, attribution, and treaty verification initiatives rely on the ability to produce valid forensic assessments of nuclear materials. Effective, efficient, and fair administration of these initiatives requires accurate assessments of the origin, age, intended use, production method, and threat characteristics of nuclear materials. The basis for forensics assessment is accurate and precise measurements of the radioactive, chemical, and physical characteristics of suspect or verification nuclear materials. More complete collection of forensic evidence can be produced by calculating derived material characteristics from these measurements. Of value is the calculation of nuclear material signatures as they provide a unique identifier, context such as clues to material processing history, and reproducible and therefore provable forensic evidence. In particular, the projects results should help support the work of the Nuclear Smuggling International Technical Working Group.

Management Plan and Activity Timeline

While this project is larger than the traditional NSF single investigator and graduate student researcher, it is not a very large project from a management point of view – three investigators and three supporting graduate students. The PI was data administrator for a 6 year project (1992-1998) developing a database about California welfare recipients which employed over 50 graduate student researchers in assembling data from county administrative computer systems. The key to effective management is clear definition of goals and tasks and weekly project meetings between the investigators and graduate student researchers to assess progress and identify bottlenecks.

Year 1:

- **project initiation, hiring of graduate student researchers**
- **development of theoretical models**
- **development of evaluation methodologies**
- **creating international partnerships**
- **development of XML metadata structures for nuclear materials data**
- **assembling data sets**
- **design and prototyping of nuclear forensic matching software**
- **workshop on scientific search at SIGIR 2012 conference**

Year 2:

- **testing of theoretical models**
- **refinement of evaluation methodologies**
- **release of XML metadata structures and a nuclear decay ontology**
- **assembling further data sets, particularly from international sources**
- **development of pilot software for nuclear forensic matching**
- **initial technology transfer to national laboratories**
- **development of course material targeted to information search professionals**

Year 3:

- **further refinement of models and evaluation**
- **implementation and fielding of search software**
- **transfer of search technology to national laboratories and other agencies**
- **initial fielding of a course in on nuclear forensics in partnership with the UC Berkeley Nuclear Engineering Department**
- **archiving of final data sets and project materials in the UC institutional repository**

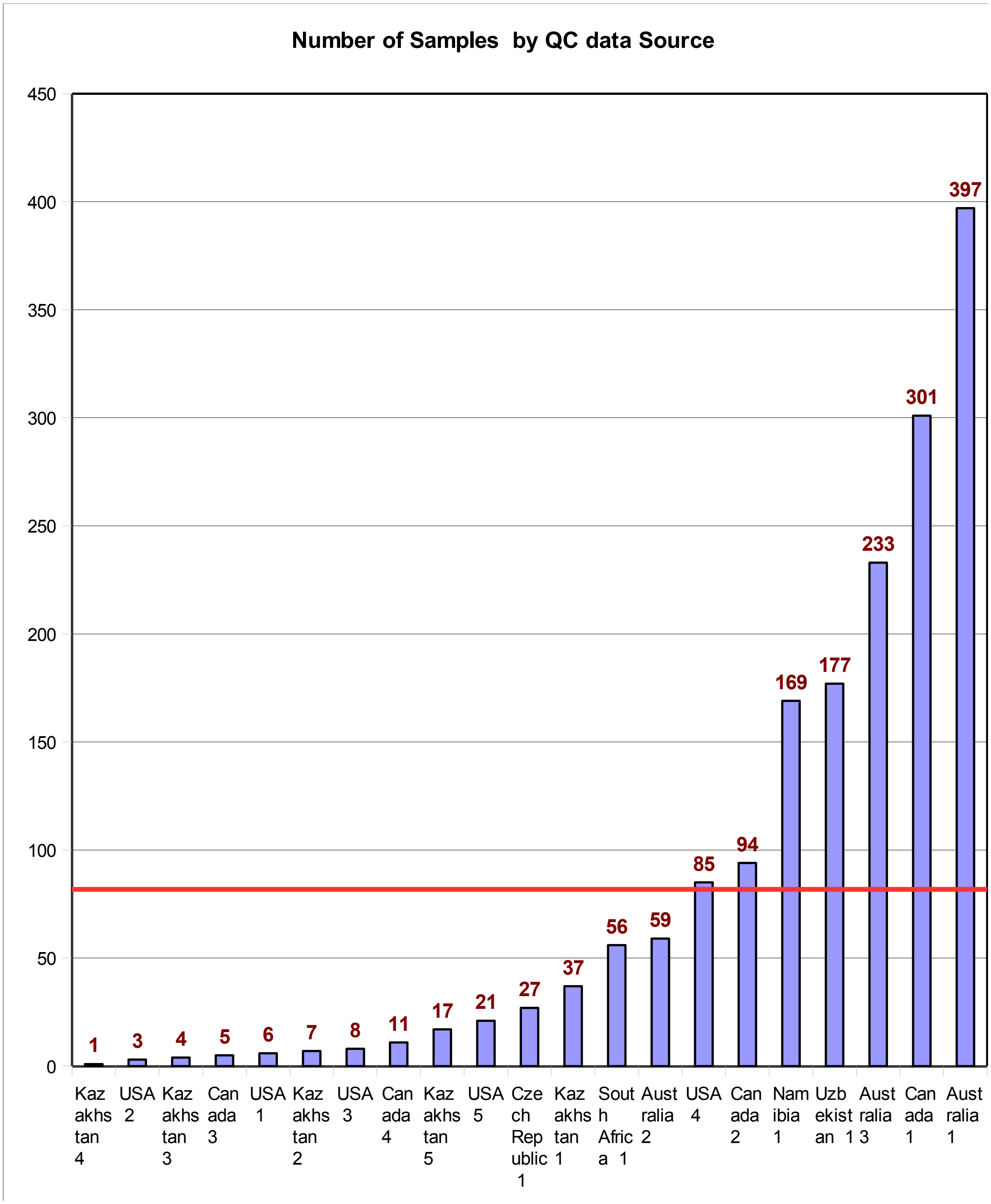


Figure 3: Bar Chart of Uranium sites from LLNL QC data with mean sample size

Recasting Nuclear Forensics Discovery as a Digital Library Search Problem: References Cited

[APS, AAAI 2008] American Physical Society (APS)/ American Association for the Advancement of Science (AAAI) Joint Working Group: **Nuclear Forensics: Role, State of the Art, Program Needs,** 2008.

[Bengoetxea 2003] Endika Bengoetxea, **Inexact Graph Matching Using Estimation of Distribution Algorithms, PhD Thesis,** (in particular, "Chapter 2, The graph matching problem,", available at http://www.sc.ehu.es/acwbecae/ikerkuntza/these/Ch2.pdf) Ecole Nationale Supérieure des Télécommunications (Paris), 2003.

[Chalwa, Jakowicz and Kolcz 2004] "Editorial: **Special Issue on Learning from Imbalanced Data Sets",** *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced data sets*, V 6, No 1, June 2004.

[Dunning 1993] Ted Dunning. 1993. "Accurate methods for the statistics of surprise and coincidence.*". Computational Linguistics* V.19, No. 1 (March 1993), 61-74.

[GAO 2009] General Accountability Office (GAO) report, **Nuclear Forensics: Comprehensive Interagency Plan Needed to Address Human Capital Issues,** *(abbreviated version),* available at http://www.gao.gov/new.items/d09527r.pdf

*[Gey and Larson 2008] F Gey and R Larson, "Relevance Levels in Patent Mining," *Proceedings of the 2nd International Workshop on Evaluating Information Access*, Tokyo, Japan, Dec 16, 2008, pp 57-59.

*[Gey et al 2001] Gey, F, M Buckland, A Chen and R Larson (2001). "Entry Vocabulary - a Technology to Enhance Digital Search." In: *Proceedings of HLT 2001, the First International Conference on Human Language Technology Research,* James Allan, Editor, San Diego, California, March 2001, pp. 91-95.

[Goldfarb and Litvinenko 2007] Alex Goldfarb with Marina Litvinenko, **Death of a Dissident: The Poisoning of Alexander Litvenenko and the Return of the KGB,** Free Press, New York, 2007.

[IAEA 2002] International Atomic Energy Agency (IAEA) Staff Report, "Tracing the Source: Nuclear Forensics and Illicit Nuclear Trafficking," October 2002.

[Keegan, Richter et al 2008] Elizabeth Keegan, Stephan Richter, Ian Kelly, Henri Wong, Patricia Gadd, Heinz Kuehn, Adolfo Alonso-Munoz, "The provenance of Australian uranium ore concentrates by elemental and isotopic analysis," Applied Geochemistry 23 (2008) 765–777.

[Kim et al 1999] Kim, Y, B Norgard, A Chen and F Gey "Using Ordinary Language to Access Metadata of Diverse Types of Information Resources: Trade Classification and

Numeric Data” *Proceedings of the 62nd Annual Meeting of the American Society for Information Science,* Washington DC, Oct 31-Nov 4, 1999, pp 172-180.

[Kuhn et al 2007] Stefan Kuhn, Tobias Helmus, Robert J. Lancashire, Peter Murray-Rust, Henry S. Rzepa, Christoph Steinbeck, and Egon L. Willighagen "CMLSpect, an XML Vocabulary for Spectral Data," J. Chem. Inf. Model. 2007, 47, 2015-2034.

[Mayer, Wallenius and Ray 2005] K Mayer, M Wallenius and I Ray “Nuclear forensics—a methodology providing clues on the origin of illicitly trafficked nuclear materials,” Analyst, 2005, 130, 433–441 (a Journal of the Royal Chemical Society).

[Mayer, Wallenius and Ray 2006] K Mayer, M Wallenius and I Ray “Nuclear forensic investigations: Two Case Studies, Forensic Science International 156 (2006) pp 55-62.

[Mayer, Wallenius and Fangänel 2007] K. Mayer, M. Wallenius, T. Fanghänel, “Nuclear forensic science — From cradle to maturity,” Journal of Alloys and Compounds 444–445 (2007) 50–56.

[Mayer, Wallenius and Schubert 2009] K. Mayer, M. Wallenius, and A. Schubert , “Data Interpretation in Nuclear Forensics”, presentation at the IAEA Nuclear Security Symposium – Vienna, 30 March – 3 April 2009, available at **http://www-pub.iaea.org/mtcd/meetings/PDFplus/2009/cn166/CN166_Presentations/Session%209/013%20Mayer.pdf**.

[Moody, Hutcheon and Grant 2005] K Moody, I Hutcheon , P Grant, **Nuclear Forensic Analysis,** CRC Press, 2005.

[NRC Nuclear Forensics 2010] National Research Council Committee on Nuclear Forensics, National Academies report, **Nuclear Forensics, A Capability at Risk,** National Academies Press, 2010.

[Patterson 2007] A Patterson, "Ushering in the era of nuclear terrorism", *Critical Care Medicine*, v. 35, p.953-954, 2007.

[Robel and Kristo 2007] Martin Robel and Michael J. Kristo, “Characterization of nuclear fuel using multivariate statistical analysis, ” Lawrence Livermore National Laboratory report

[Robel, Kristo and Heller 2009] M. Robel, M. J. Kristo, and M. A. Heller, “Nuclear Forensic Inferences Using Iterative Multidimensional Statistics,” Institute of Nuclear Materials Management 50th Annual Meeting Tucson, AZ. July, 2009, 12pp.

[Scott 2005] Mark Robert Scott, **Nuclear forensics: attributing the source of spent fuel used in an RDD event,** Masters Thesis, Department of Nuclear Engineering, Texas A & M University, 2005, available at: http://txspace.di.tamu.edu/bitstream/handle/1969.1/2368/etd-tamu-2005A-NUEN-Scott.pdf?sequence=1

[Smith et al 2004] L. E. Smith, E Ellis, A. Valsan, C Aalseth and H Miley, “A Coincidence Signature Library for Multicoincidence Radionuclide Analysis Systems” IEEE Transactions on Nuclear Science, v.51, No. 3, June 2004, pp 1044-1048.

[Švedkauskaite-LeGore et al 2007] J. Švedkauskaite-LeGore, K. Mayer, S. Millet, A. Nicholl, G. Rasmussen and D. Baltrunas (2007). Investigation of the isotopic composition of lead and of trace elements concentrations in natural uranium materials as a signature in nuclear forensics. Radiochimica Acta: Vol. 95, No. 10, pp. 601-605.

[Yang Zhang and Kisiel 2003] Y. Yang, J Zhang and B Kisiel, “A Scalability Analysis of Classifiers in Text Categorization,” *Proceedings of SIGIR 2003, the 26th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 96-103.