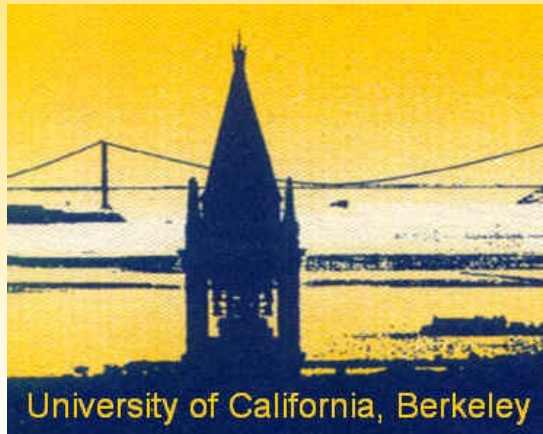
Database Heterogeneity in a Scientific Application
IASSIST 2012 Poster, Washington DC, June 6, 2012

- **Fredric C. Gey, Chloe Reynolds, Ray Larson, Electra Sutton**

  - **Institute for the Study of Societal Issues and**
  - **The Information School**
  - **University of California, Berkeley**
  - **http://metadata.berkeley.edu/nuclear-forensics**

- **Funding source: National Science Foundation Grant: #1140073**
  Recasting Nuclear Forensics as a Digital Library Search Problem (2011-2012)
  Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF)



Nuclear Safeguards

University of California, Berkeley

- Social issue: As countries abandon their nuclear ambitions, what happens to their existing nuclear facilities?
  - Ukraine
  - South Africa
  - Congo
- These may be targets by criminals or terrorists to obtain illicit nuclear materials
- Currently USA & Russia have 95% of all weapons grade nuclear material
- IAEA (International Atomic Energy Agency, Vienna) will have a role to play for decades to come
- http://www.iaea.org/NewsCenter/News/2008/forensicskit.html

Database Heterogeneity in a Scientific Application

IASSIST 2012 Poster, Washington DC, June 6, 2012

- **Nuclear Forensics** is the science of tracing the sources of smuggled nuclear material, from
  - Spent fuel from a nuclear reactor
  - Refined uranium ore (yellow cake) from mines
  - Other sources
- **Characteristics of nuclear elements** (isotopes) can be used as clues to origin in much the same way as in traditional forensics (fingerprints or DNA matching)

Database Heterogeneity in a Scientific Application

IASSIST 2012 Poster, Washington DC, June 6, 2012

- **Nuclear Forensics** is the science of tracing the sources of smuggled nuclear material, from
  - Spent fuel from a nuclear reactor after fission
  - Enriched uranium or plutonium in the nuclear fuel
  - Refined uranium ore (yellow cake) from mines
- Since 1992, more than **800 incidents of interdicted nuclear material** have been reported,
- Nuclear forensics is a key component of fighting global terrorism

Nuclear material could come from any of about 500 nuclear power plants worldwide

University of California, Berkeley

(Worldwide Nuclear Power Plants with Earthquake Zones)

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

A Case of Nuclear Murder

- On November 1, 2006, Alexander Litvinenko, former Russian Federal Security officer was poisoned by Polonium-210 isotope while having lunch with associates at a London sushi restaurant. He died of radiation poisoning three weeks later.
- According to doctors, "Litvinenko's murder represents an ominous landmark: the beginning of an era of nuclear terrorism."
- Polonium-210 (^{210}Po) is the only isotope of Polonium with a significant half-life (138 days). It decays by emitting alpha particles which can be easily shielded by even pieces of paper or the human skin
- British authorities traced the material to a specific nuclear reactor in Russia

Nuclear Forensics

- Dealing with terrorists nuclear intentions has two aspects – detection and forensics
- Large projects for improving detection (i.e. sensing radiation from outside shipping containers) are underway
- Equally large projects (>$US100M) are underway for forensics in the USA and EU
- These projects are creating digital libraries of the composition of existing nuclear material samples collected from mines or nuclear processing plants worldwide
- The search aspect against these libraries has heretofore proceeded on an ad-hoc case-by-case basis

Nuclear Reactor Database
Unifying Multiple Datasets

We wanted a comprehensive detailed database about worldwide nuclear reactors including geographic coordinates

Searches for "nuclear dataset" and similar terms

- **200+ datasets found on web**
  - **Many more web links still to explore, likely to yield even more datasets**
- **80+ datasets downloaded (arbitrary subset)**
  - **Sorted into useful (65) / not useful (15) categories**
  - **Not useful example: Nuclear capacity by country**
- **Before consolidation, the 65 combined useful datasets have 130+ columns and 50,000+ records**

Nuclear Reactor Database
Unifying Multiple Datasets

- The fun begins of aligning the data
  - The 50,000+ records / 130+ columns will likely reduce to around 2,000 records / 100 columns
  - Reduction of columns
    - Consolidation (‘Date of Shutdown’ and ‘Shutdown Date’)
    - Removal of unwanted columns (e.g. ‘Architect of Plant’ or ‘2003 Capacity’)
  - Reduction of records
    - Consolidation of name variants (‘Arkansas Nuclear One, Unit 2’ and ‘Arkansas Nuclear-2’ and ‘Arkansas Nuclear 2’ or ‘USGS’ and ‘U.S. Geological Society’ or ‘UNIV OF TEXAS’ and ‘UNIVERSITY OF TEXAS’)
  - Data Normalization
    - Handle multiple date formats – such as ‘Spring 2011' and 'Estimated December 11th' (no year) and ‘2/3/2004’

Nuclear Reactor Database
Unifying Multiple Datasets

- **Data Normalization (continued)**
  - **Parse** 'locations' column into separate columns ('Arrowhead, CA' would become city = 'Arrowhead' and state = 'CA' while '33.921563, -78.0202677' would break into latitude and longitude columns). Another example that needs to be parsed is 'Reactor and Containment Type.'
  - Investigate **conflicting data**, possibly resolve via dataset update dates
  - **Capture metadata** – update operating 'status column' where dataset title "Power Reactor Sites Undergoing Decommissioning" or "U.S. Nuclear License Renewal Filings"; or where column names imply an operational status: "Date Canceled" or "Operating License Expires." Also, add records for 'Turkey Point-1' and 'Turkey Point-2' when a 'Turkey Point-3' exists.
- **Normalization challenges**
  - Dataset **time overlaps**. Dataset A may have been last updated 2/4/04 and dataset B may have been updated 10/1/09. In most cases, I would want to uses the information from dataset B. However, I must be **careful**.
  - If the two datasets showed a mismatch of location (city and state, for example), such a conflict would indicate that the facilities are not the same because that information should be the same regardless of a dataset's update date. Other conflicts (e.g. 'operating status') could legitimately be resolved by retaining the noted in the most-recently updated dataset

Nuclear Reactor Database
Unifying Multiple Datasets

- **About dataset update dates**
  - **Although a dataset may have been 'last updated' on 2/4/04, the 2/4/04 updates may have only been changes to certain records or columns in that dataset. The contents of a particular field for a particular record might not have been updated since any older date, such as 11/1/03. Despite our utmost care, consolidating datasets is an imperfect science.**
  - **Furthermore, some data sources are only relaying information compiled and published by another agency (the original source may or may not be clearly marked). Tracing the original data providers is another challenge, in addition to finding the most current information.**

Nuclear Forensics Search
Grant home page

- http://metadata.berkeley.edu/nuclear-forensics
- Contacts
  - Fred Gey (gey at berkeley dot edu)
  - Ray Larson (ray at ischool.berkeley dot edu)
  - Electra Sutton (electra at berkeley dot edu)