Research Challenges on the BioEncyclopedia Project

Terence Critchlow

July 2005

UCRL-PRES-214157

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

The Biodefense Knowledge Center

- **This is a multi-institutional, multi-disciplinary center established June 2004**

- **The BKC Mission is to "Enable collaboration and data sharing among policy makers, responders, analysts, law enforcement personnel, and scientists in the Homeland Security community to assure that timely and authoritative biodefense information is available to those with a need to know."** ***– Bill Colston, BKC Director, August 2004***

We have developed a comprehensive program for understanding the threat from bioterrorism

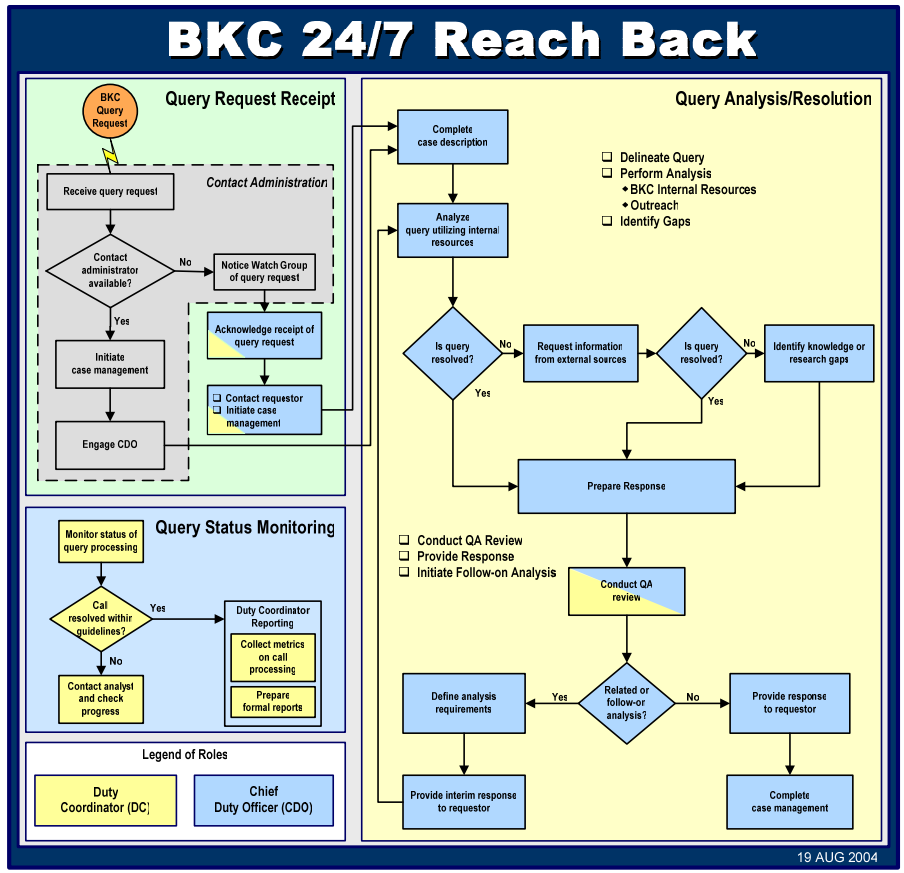
Science based

Intelligence informed

Technology enabled

In September 2004 we established a 24/7 S&T support line to DHS

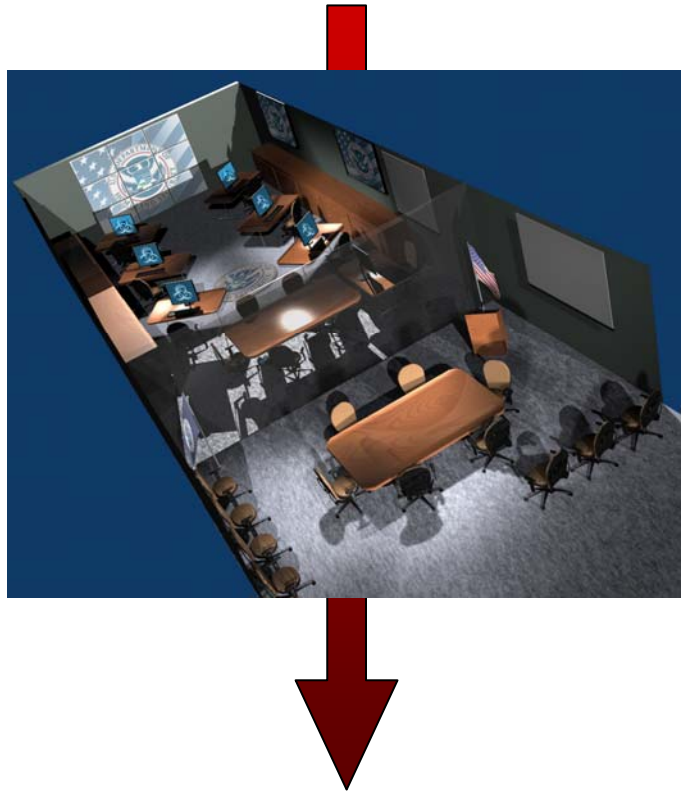
- **Single point of contact for authoritative biodefense information**
  - Large community of internal scientists/ Intel analysts
  - Access to external SMEs with peer review
- **Secure access levels**
  - Security level determined by inquiry/user
- **Unique analysis toolset**
  - Bio-medical data and research reports linked to intel data
- **Institutional memory**
  - Archive of data sources, contributors, modeling parameters, etc.



We offer various assessment 'products' generated by BKC analysts

- **Reachback reports**
  - Immediate (hours)
    - Minimum research required--chiefly based on analyst experience or info available in knowledge base from previous inquiry
  - Near-term (days)
    - Research required, but minimal information synthesis
- **Long-term studies**
  - Multidisciplinary studies
- **Information bulletins**
  - Preemptive analysis of a timely topic or a means of disseminating topical information from reachback requests

TACTICAL



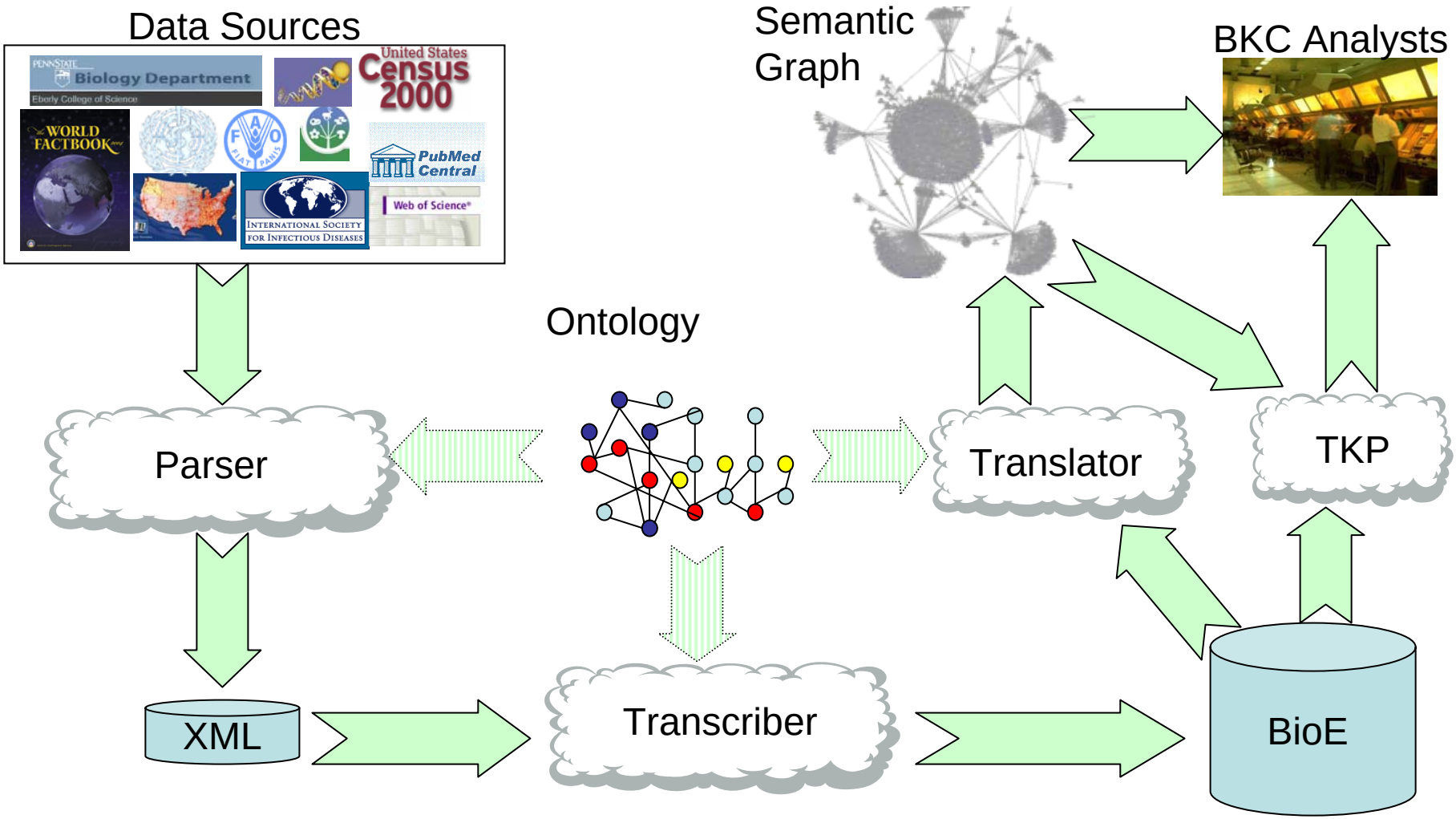
STRATEGIC

The BioE is one of three thrust areas within the BKC technology section

- **Collaborative Visualization**
  - Extend existing analysis tools to work within the unified architecture
  - Develop an integrated analyst environment to provide a consistent interface to the range of data and tools used by the analysts
- **Knowledge management**
  - Apply a semantic graph architecture to the BKC to support complex queries and interaction with other domains of interest
- **BioEncyclopedia (BioE)**
  - Integrate all biological information relevant to the BKC mission
  - Develop tailored knowledge products that provide customized interfaces to complex queries and the resulting data

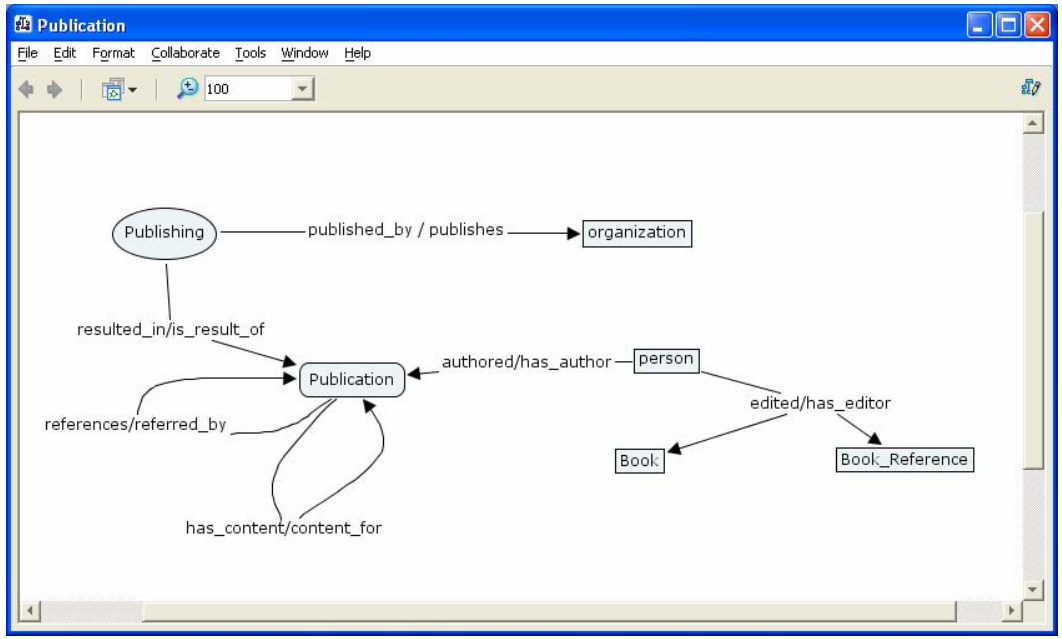
We are designing, developing, and deploying general solutions to research level problems being encountered within the BKC environment.

The BioE approach: The big picture



The BioE approach: Ontology creation

- **Globally consistent view of the world**
- **Current ontology has several hundred concepts and edges**
  - Multiple namespaces
- **Created using a graphical interface program**
  - Developed by the Institute for Human and Machine Cognition



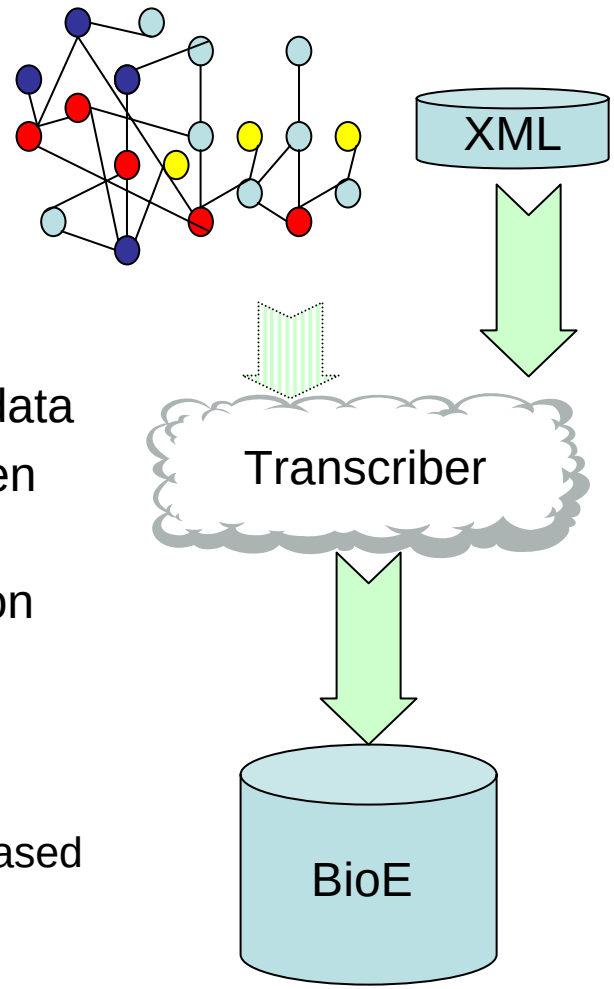
- **Challenges:**
  - Develop consistent perspective of complex information across a broad domain

The BioE approach: Parser writing

- **Data comes from autonomous sources**
- **Parsers are manually written to extract data from each source format and convert it to an XML representation**
  - Syntactic and semantic transformations
  - Converting units (inches to cm)
  - Adding implicit information
- **Currently parsing 15 structured and unstructured data sources**
  - Many more sources on the queue
  - RDBMS, HTML, XML, excel, CSV
  - Many sources contain multiple types of information
  - Much of the data contains free text

The BioE approach: BioE loading

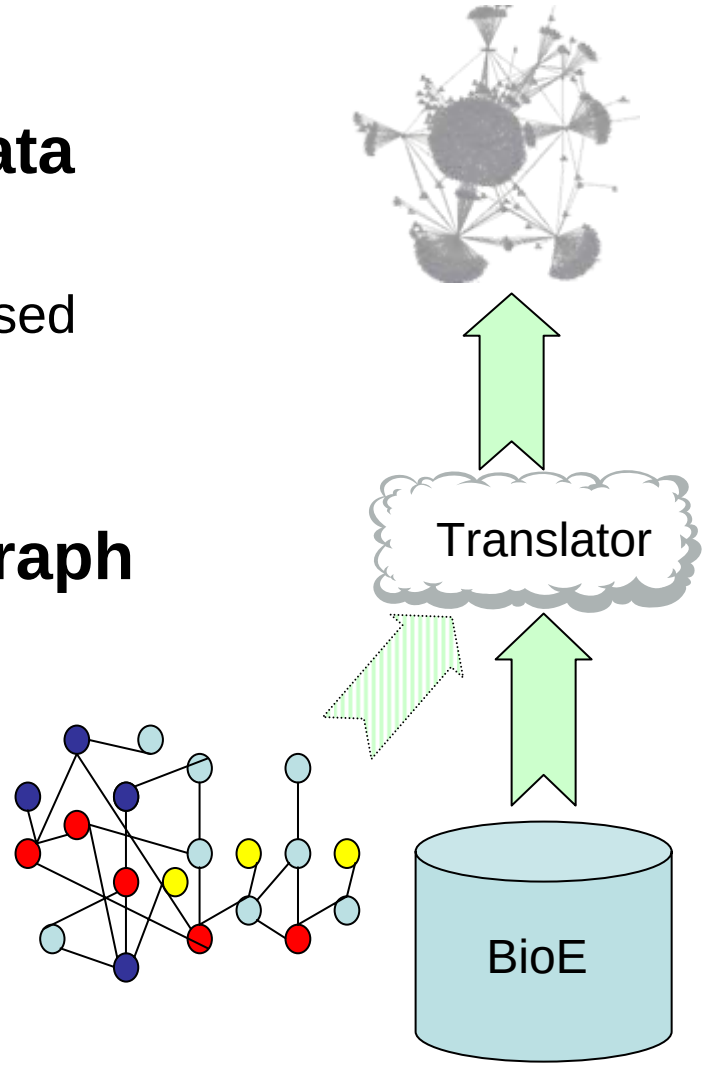
- **The transcriber reads the parser generated XML and uses the ontology to enter it into the BioE appropriately**
  - XML includes provenance information which is also transferred into the BioE
  - Performs concept based canonicalization of data
  - Identifies instances of concepts that have been previously entered
  - Instance locking used to support parallelization
- **Challenges:**
  - Represent a complex relational schema within the ontology format
  - Ensure consistency of a large, complex ontology based on a set of constraints



The BioE approach: Graph loading

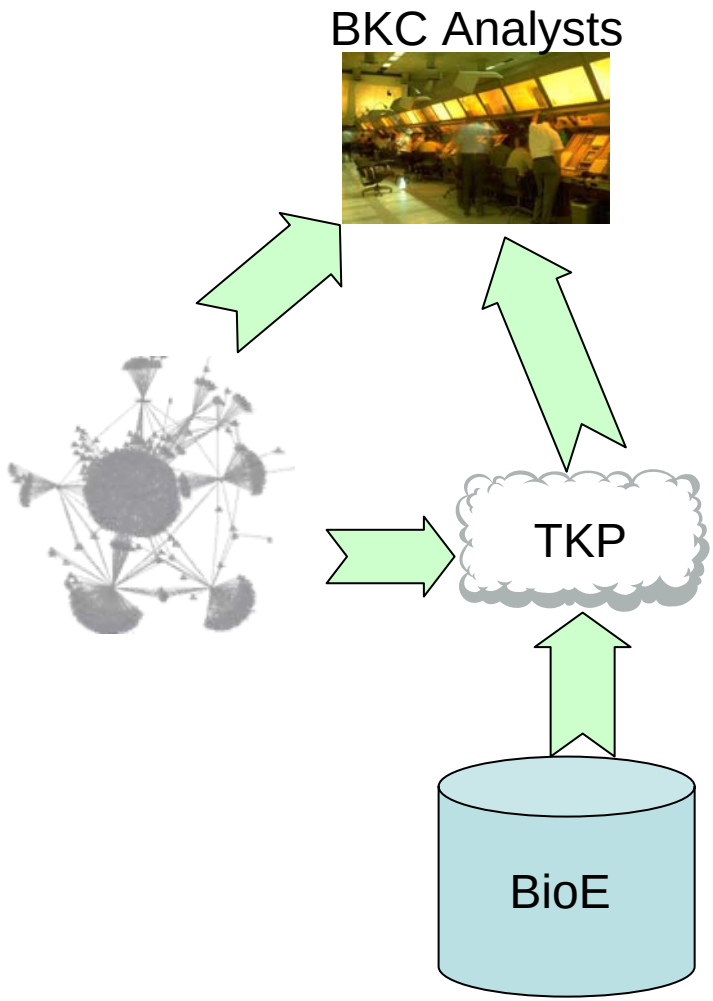
- **The translator sends both the ontology and the associated data into the semantic graph**
  - Security information must also be passed

- **Currently we are passing the graph on the order of 10M nodes and edges**
  - "Small" data, expected to increase dramatically over time



The BioE approach: Information analysis

- **Analysts can perform queries directly against the graph through viewer or can perform "*canned queries*" using a pre-defined tailored knowledge prduct (TKP)**
  - TKPs use a combination of graph queries, detailed information that was not passed to the graph, and data post-processing to provide a straight-forward way to ask a complex question
- **Direct keyword search of documents is also supported**



So what is left to do?

Looking towards the future, there are a large number of problems whose solutions are not imminent

Ontology representation standards

- **OWL is a recognized ontology representation standard**
  - Reasonable amount of software supporting this standard
- **OWL does not include sufficient information to map to an efficient relational database schema**
  - Primary keys
  - Alternate keys
  - Foreign keys
  - Inheritance
  - Complex indices
  - Data duplication
- **Need to handle dynamic ontologies**

```
<owl:Class rdf:ID="Book">
   <rdfs:subClassOf>
    <owl:Class
    rdf:ID="Publication"/>
   </rdfs:subClassOf>
</owl:Class>
<owl:ObjectProperty
    rdf:ID="Authored">
   <rdfs:domain
    rdf:resource="#Person"/>
   <rdfs:range
    rdf:resource="#Publication"/>
</owl:ObjectProperty>
<owl:DatatypeProperty
    rdf:ID="Title">
   <rdfs:domain
    rdf:resource="#Publication"/>
   <rdfs:range
    rdf:resource="http://www.w3.org
    /2001/XMLSchema#string"/>
</owl:DatatypeProperty>
```

Ontology Transformation

- **Given a source ontology and a target ontology, both in OWL, define a mapping that defines the transformation of source data into the target format**
  - Source instance maps to multiple target instances
  - Multiple source instances map to single target instances
  - Must be able to define arbitrarily complex data transformations

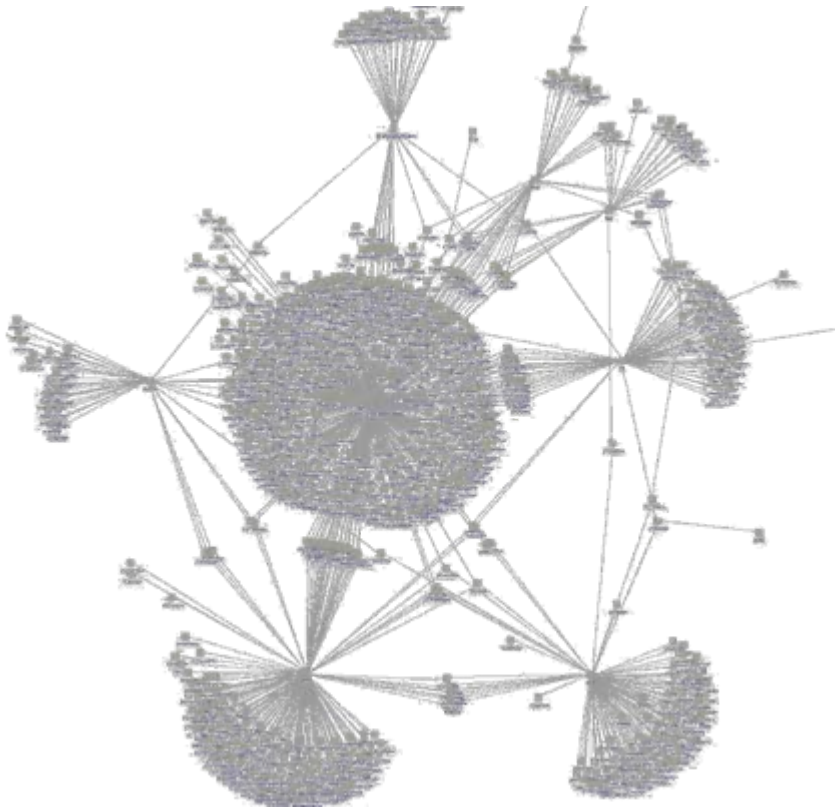
```
<concept_mapping id="pub-to-
    conference-proceedings">
    <target>
        conference_proceedings </>
    <source> pub </>
    <type> attribute_selection
        <condition>
            publication_location
            contains
            "In Proceedings of"
        </condition>
    </type>
    <attribute_mapping>
        <target> title </>
        <source> title </>
        <type> identity </>
    </attribute_mapping>
</concept_mapping>
```

Effective use of roles to enforce consistency

- **Some concepts can be thought of as a role being played by another concept**
  - Person -> Author
  - Person -> Editor
  - Person -> Disease Host
  - Person -> Disease Vector
- **Inheritance has problems**
  - Overlapping subsets
  - Multiple inheritance required
  - Constraints hard to enforce (e.g. a person is only a vector if they are also a host)
- **Relationships have problems**
  - Constraints hard to enforce
  - Unable to work well with enumerated sets of concepts
- **Need to be able to define a concept that incorporates a subset of the instances contained in another concept**
  - Defined classes in OWL
  - Hard to define when class membership is not an obvious feature of other attributes
  - Any person can be a host for any disease that infects humans
  - Disease vector status is transitory

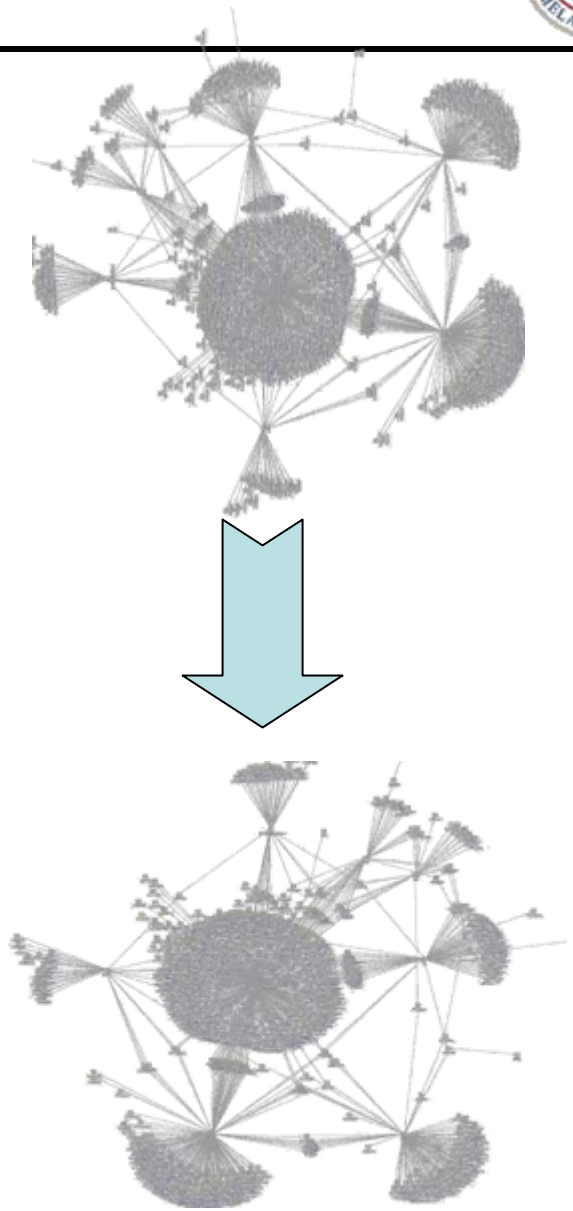
Scalable ontology definition and editing

- **Visualization is a useful way to define complex ontologies**
  - Current systems are not scalable for large-scale ontologies
  - Not able to define hierarchical views over the data
  - Do not handle namespaces effectively / correctly
- **Validation tools currently work only on OWL DL**
  - Need to be able to perform ontology validation against OWL Full definitions (or at least better than current)



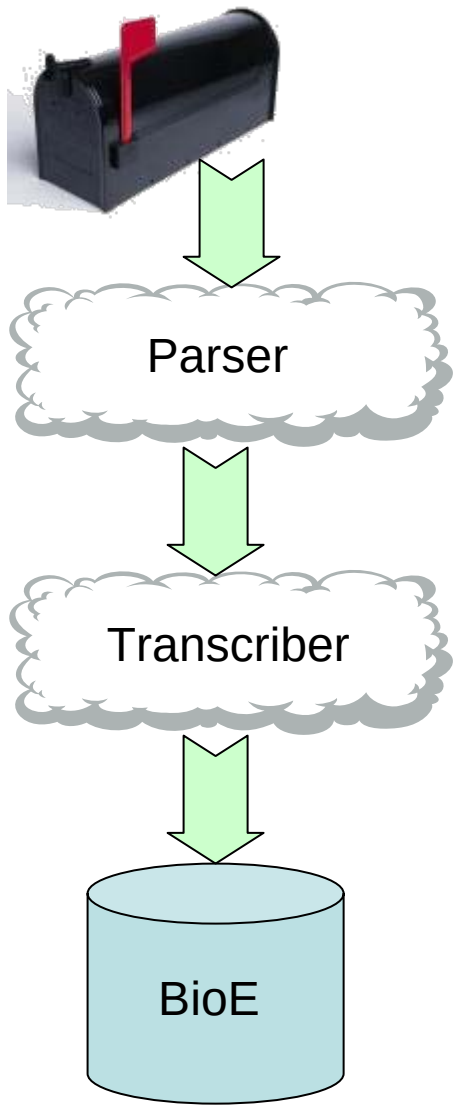
Ontology driven tools

- **How can ontologies reduce the effort required to write parsers?**
- **Currently, parser writer needs to understand both the global ontology and the data source ontology**
  - Ontology mappings break that requirement
  - What can we do with mapping?
- **Tools for defining ontology mappings**
  - What level of reasoning can you perform over that information
  - Is the mapping consistent?
  - Must work on large ontologies



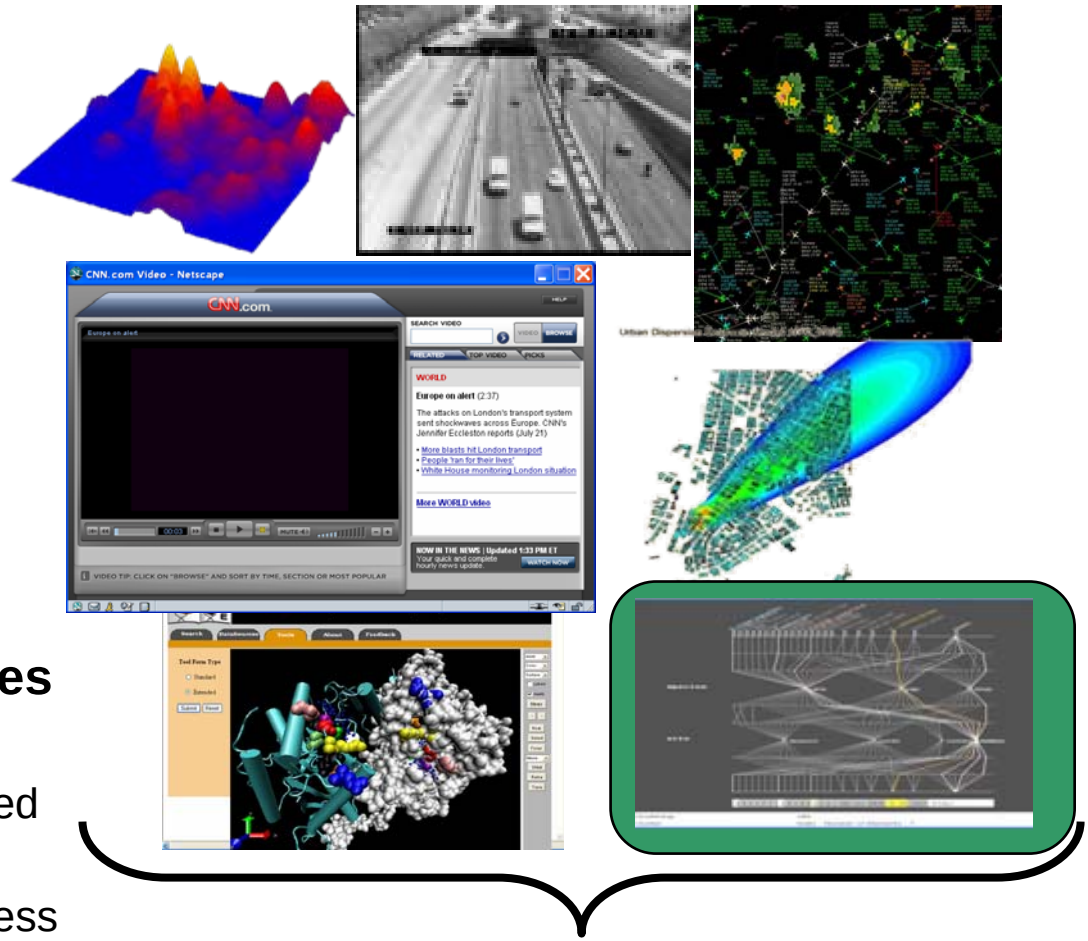
Change detection

- **New information is identified, ingested, parsed, and loaded automatically**
  - Minimal cost to data source
  - Multiple processing steps need to be defined
- **Need to identify when new information becomes available from a data source of interest**
  - File system monitoring
  - Email notifications
  - Web sites
  - ***Databases***



Effective use of multi-modal data

- **In addition to textual data, we need to incorporate information from**
  - Images
  - Video
  - Sensor networks
  - Data streams
    - News feeds
  - Simulations
- **Also need to be able to effectively query these types of information**
  - Complex data analysis required
  - Incorporate complex analysis into ingest and workflow process
- **Handle multi-step analysis of data prior to ingest**



Provenance tracking

- **Would like to be able to identify all source document references for all instances**
  - Each instance may be references in multiple sources
  - ***Is this possible?***
- **Need to track where information comes from and what processing it has undergone**
  - Data source
  - Data transfer method
  - Date obtained
  - Query that generated data (opt)
  - Parser(s) used to convert data
  - Analysis program(s) applied
  - Date loaded
  - Load method

```
<ingest> </>
<src_file>File1</>
<src_file>File2</>
<process_step>
    <input>File1</input>
    <output>File3</>
    <output>File4</>
    <desc>parsedXML-source</desc>
    <tool id=”dso parser”>
        <params>File1 configfile</>
        <host>bioe</>
        <date_ran>May 11, 2005</>
    </tool>
</process_step>
```

Natural language processing

- **Automatically extract all relevant information from a large collection of free text documents, in such a way as to support post-processing of the events represented in the corpora**
  - Identify relevant entities
  - Identify relationships between entities
  - Resolve intra-document co-references
  - Determine appropriate scenario templates
  - Extract template information
  - Resolve inter-document co-references
  - Extract multi-document template information

  - Completely automated pipeline a requirement (millions of documents)
  - Precise extraction
  - Entire corpus not known in advance
  - Technology needs to work on a variety of corpora

Event extraction from free text

Typical approach is to fill in a template that describes an event

<disease-outbreak>
- **<disease>**
- **<location>**
- **<host>**
  - **<organism>**
  - **<number>**
- **<vector>**
  - **<organism>**
  - **<transmission method>**
- **<containment measures>**

Event extraction from free text

From: ProMED-mail

Indonesia: Farm Worker is First Confirmed Human Case of Avian Influenza

JAKARTA: A farm worker in eastern Indonesia has tested positive for avian influenza virus, marking him the country's first human case of the virus. The worker from southern Sulawesi island is healthy but two tests at a Hong Kong laboratory confirmed that he had been infected by avian influenza virus. Since 2003, the highly lethal disease has struck chickens, quail and other birds in 18 Indonesian provinces on seven islands. Researchers have also determined that avian influenza virus has infected Indonesian pigs, an ominous development because swine can catch both avian and [mammalian] influenza viruses.

Typical approach is to fill in a template that describes an event

<disease-outbreak>
- **<disease>**
- **<location>**
- **<host>**
  - **<organism>**
  - **<number>**
- **<vector>**
  - **<organism>**
  - **<transmission method>**
- **<containment measures>**

Event extraction from free text

From: ProMED-mail

Indonesia: Farm Worker is First Confirmed Human Case of Avian Influenza

JAKARTA: A farm worker in eastern Indonesia has tested positive for avian influenza virus, marking him the country's first human case of the virus. The worker from southern Sulawesi island is healthy but two tests at a Hong Kong laboratory confirmed that he had been infected by avian influenza virus. Since 2003, the highly lethal disease has struck chickens, quail and other birds in 18 Indonesian provinces on seven islands. Researchers have also determined that avian influenza virus has infected Indonesian pigs, an ominous development because swine can catch both avian and [mammalian] influenza viruses.

Typical approach is to fill in a template that describes an event

<disease-outbreak>
- **<disease>** Avian Influenza
- **<location>** Indonesia
- **<host>**
  - **<organism>** Human
  - **<number>**
- **<vector>**
  - **<organism>**
  - **<transmission method>**
- **<containment measures>**

Event extraction from free text

From: ProMED-mail

Indonesia: Farm Worker is First Confirmed Human Case of Avian Influenza

JAKARTA: A farm worker in eastern Indonesia has tested positive for avian influenza virus, marking him the country's first human case of the virus. The worker from southern Sulawesi island is healthy but two tests at a Hong Kong laboratory confirmed that he had been infected by avian influenza virus. Since 2003, the highly lethal disease has struck chickens, quail and other birds in 18 Indonesian provinces on seven islands. Researchers have also determined that avian influenza virus has infected Indonesian pigs, an ominous development because swine can catch both avian and [mammalian] influenza viruses.

- **All of the information we are interested in extracting is highlighted in Red**
  - The fact this is the first human case is important
  - There is only 1 person infected
  - The diagnosis was confirmed by lab tests
  - Avian influenza virus causes avian influenza
  - The location is refined
  - Other organisms were previously infected in multiple related locations
- **Note that the article contains both information about a specific outbreak instance and historical information that sets the context**
  - Both are interesting

Event extraction from free text

From: ProMED-mail
[Indonesia]: [Farm Worker] is [First] [Confirmed] [Human Case] of [Avian Influenza]

JAKARTA: A [farm worker] in [eastern Indonesia] has [tested positive] for [avian influenza virus], marking him the country's [first] [human case] of the virus. The worker from [southern Sulawesi] island is [healthy] but [two] [tests] at a [Hong Kong] [laboratory] [confirmed] that he [had been infected] by [avian influenza virus]. [Since 2003], the highly lethal disease has struck [chickens], [quail] and [other birds] in [18 Indonesian provinces]

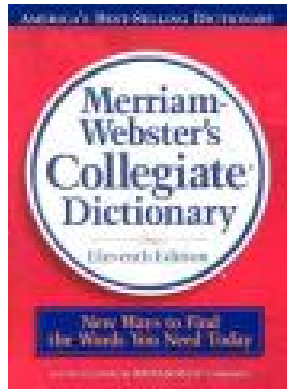
What about the information in blue

Still many unanswered questions

- Do we care where the lab work was done?
- How much of Indonesian island geography do we track?
- Are Indonesian pigs a separate breed or just a national reference?
- Do we care about the health status of the index case?
- Do we care about his occupation

This does not map cleanly to an event template. How do we represent all of this information in a computer usable format.

Dictionary definitions



- **Some concepts can be completely enumerated**
  - Diseases
  - Organisms
  - Vaccines
- **Some concepts are finite but non-enumeratable**
  - People's names
- **Using these enumerations can improve precision of free text processing**
  - Help differentiate between diseases and viruses
  - Lists must be correct, thus curation by an expert is required
- **How do you effectively deal with multiple uses of a word**
- **To what extent can entity recognition techniques be developed to reduce the effort involved in identifying concept instances**

Instance disambiguation

- **Within a document, which nouns reference the same instance of a concept?**

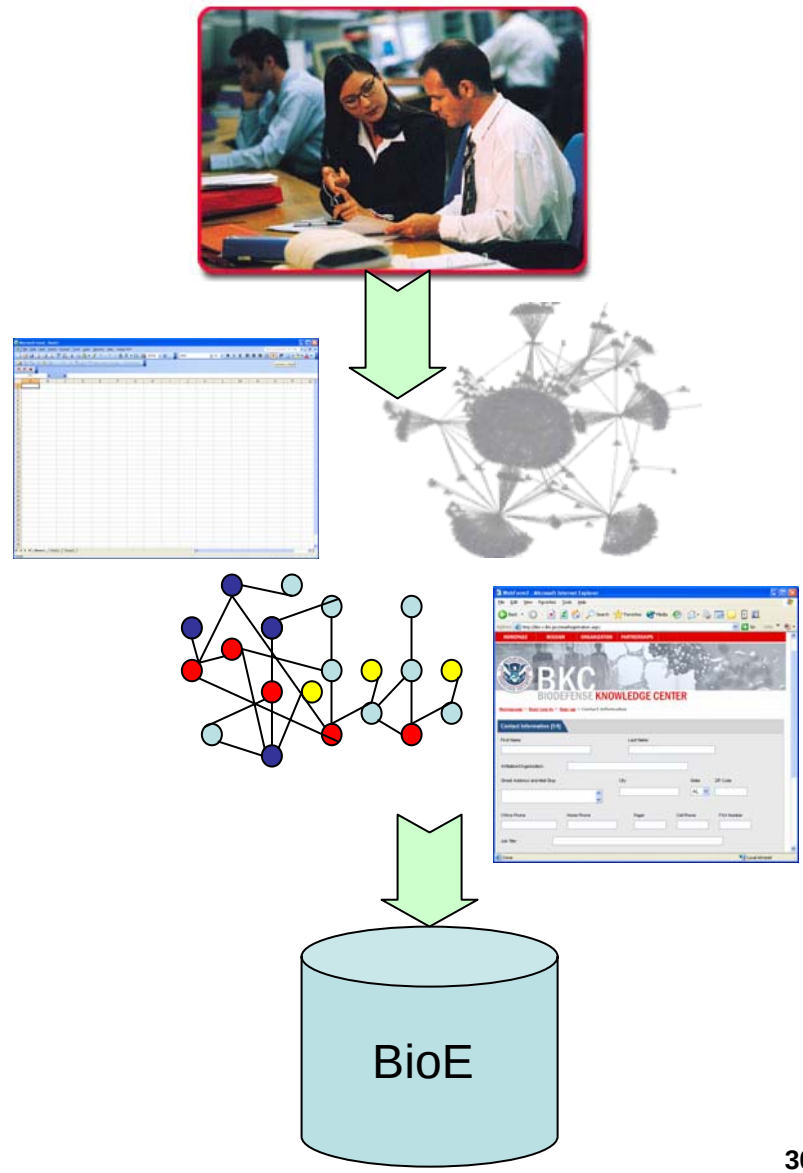
- **For a given instance obtained from a specific document, how do you know if it is the same instance as referenced in a different document?**
  - Different spellings of names
  - Different keys in different sources
  - How much real-world knowledge do you need?
  - What forms a unique key for the concept
  - What happens if this information is not provided by a source?

- **What does it mean to “be the same”**

!!?

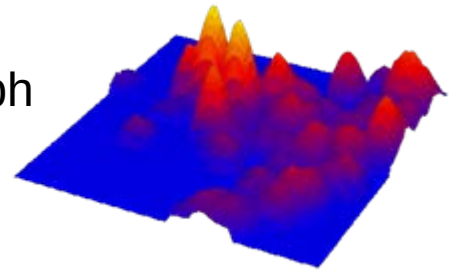
Development and incorporation of curation tools

- **Need to be able to utilize information currently contained only within analysts heads**
  - Map knowledge to facts and relationships within ontology
  - Annotate existing data
  - Extend ontology
  - Analysts are busy and are not expert computer users
- **Need to maintain provenance information**
  - Who entered information
  - When it was entered
  - How was it entered
  - Confidence in data



Extendable interfaces to TKPs

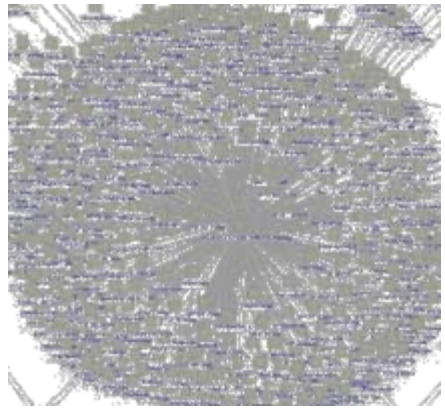
- **Ability to explore data from a variety of angles through a single query definition**
  - Can the philosophy of data cubes be applied to graph data
  - Use of workflows and modular programming to simplify creation of new queries

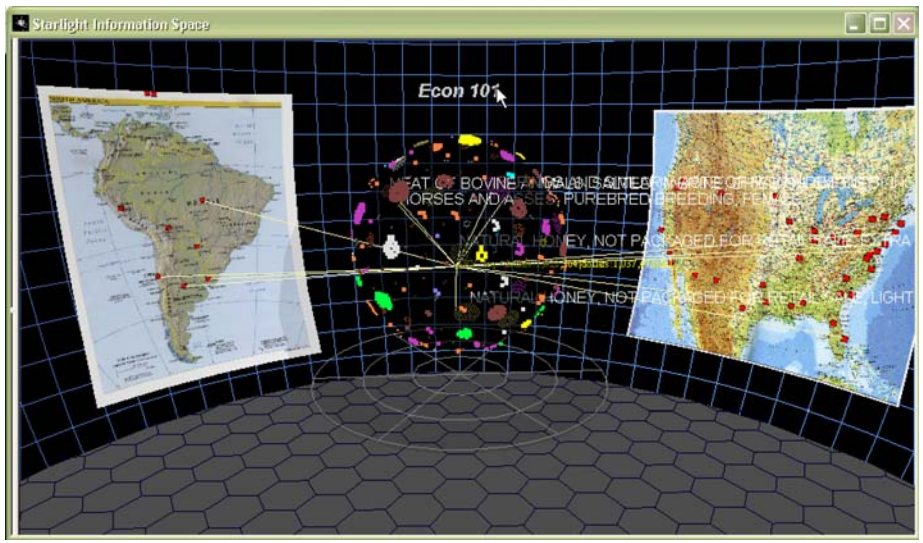


- **Need to incorporate external analysis programs into canned query interface**
  - May require additional user interactions between steps in a query execution

Intuitive, scalable interfaces

- **Need to seamlessly integrate applications through an intuitive interface**
  - Different display mechanisms for different data types
- **Need to be able to visualize large results sets**





Summary

- **The BioE has deployed an initial infrastructure to support a real-time operational facility providing BioDefense information to DHS**

- **Addressing the plethora of remaining issues will require a variety of techniques. Improved infrastructure and engineering approaches will address some of the problems, but will only provide a band aid solution for others. An aggressive long-term research effort is required to completely solve all of these problems.**

Research and Development on the BioEncyclopedia

Terence Critchlow

July 2005

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.