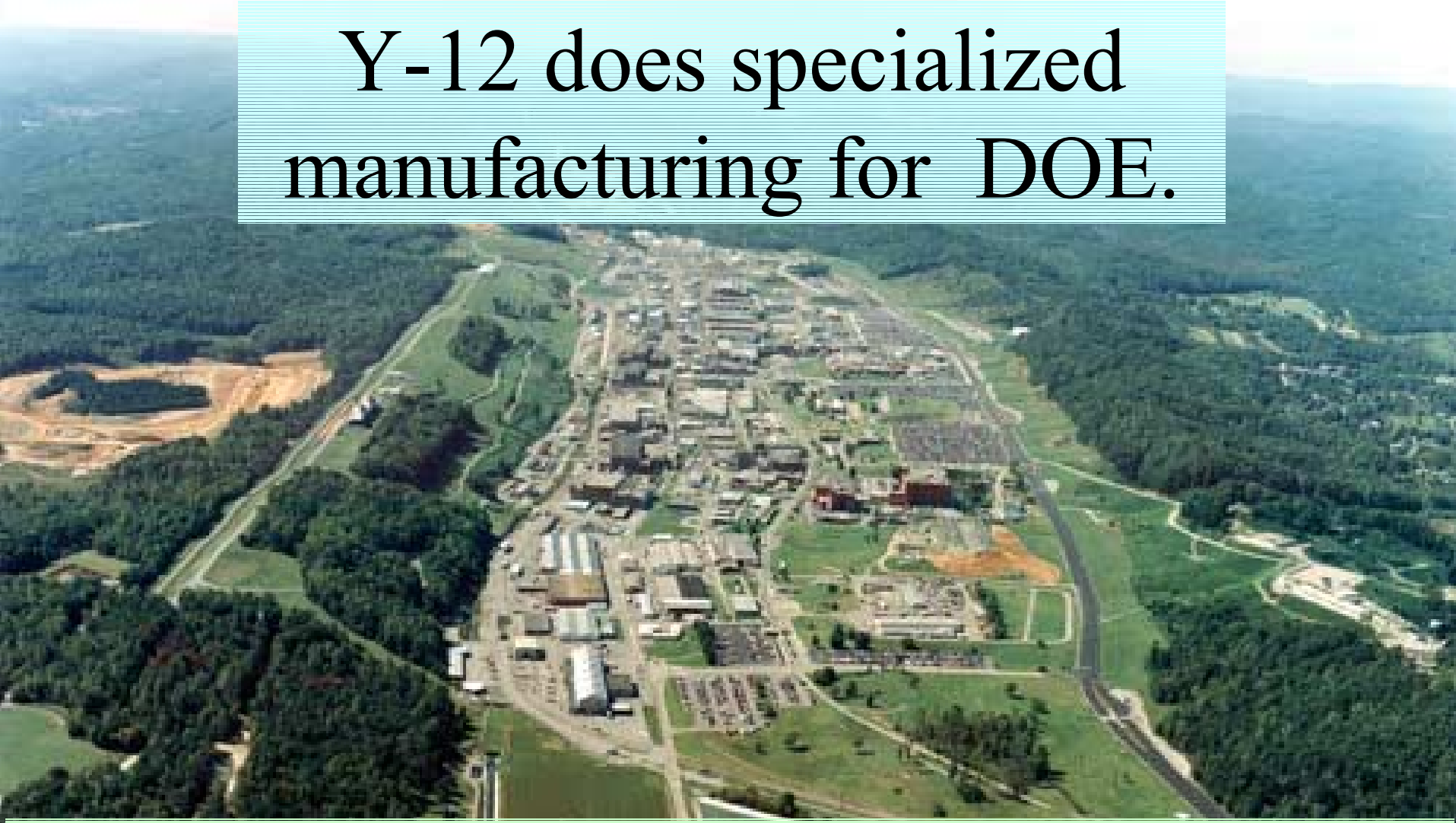
Knowledge Engineering for the Ferret Analytical Engine

James David Mason

Managed by BWXT Y-12, L.L.C.
for the U.S. Department of Energy
under contract DE-AC05-00OR22800

Y-12 does specialized manufacturing for DOE.



- Oak Ridge Centers for Manufacturing Technology
- National Prototype Center
- Major site for nuclear weapons production

Weapons manufacturing presents special challenges for information management.

- Much information is classified.
- Authorized Derivative Classifiers (ADCs) play a key role in information protection.
- The Ferret engine was developed to aid ADCs analyze documents for potentially classified information.

Classification depends on:

- Use of formal classification guidance
- Recognition of sensitive information
- Understanding of manufacturing processes
- Awareness of history and previous disclosures
- Ability to draw inferences from data

Classification process recognizes:

- Specific classified items (inventories, functional names, codewords)
- Classes of data (design of parts, specific materials)
- Implications of information (name of building ⇒ material, attribute of part ⇒ shape)
- Combinations of otherwise unclassified data (association of material with system)

The Ferret engine

- Recognizes specific concepts and combinations of concepts through simultaneous searching for many concepts,
- Follows implications of concepts,
- Associates its findings with formal classification guidance, and
- Returns marked text with appropriate classification and rules from guidance.

The Ferret knowledge base includes:

- Trees of concepts, hierarchically ordered according to their implications,
- Links among trees for complex implications,
- Rules for applying implications to classification actions, and
- Actions, including further associations or classifications to be applied, with appropriate texts from formal guidance to be presented to the user.

Trees contain the major items to be recognized:

- DOE facilities (in detail for Y-12)
- “Elaboration” on sensitive concepts
- General attributes of materials and parts
- Weapons and components (names, codenames, generic roles, parts of specific systems)
- Materials (metals, alloys, plastics, other materials)

The original Ferret knowledge base used three kinds of data structures:

- Trees, with simple but recursive root-branch-leaf structure
- Rules files, in topic map form, for links within the trees and between the trees and the tables
- Reporting structures (tables of actions, with simple row-and-cell structure)

The Ferret knowledge base for Y-12 classification includes:

- 5,400 known “words” (including stop words)
- 1,600 concepts in trees,
- 2,100 primary implications derived from trees, and
- 800 classification rules drawn from guidance.

Implications are generated from trees; then all search strings, implications, and rules are applied to candidate documents.

Ferret makes an unusual use of topic maps.

- In a traditional application, topic maps are used as a means of making individual queries about a primary fixed information base.
- In the Ferret application, the fixed information base of concepts and implications is used to query variable input data, and the “topic map” guides the processing of results from hundreds of queries.

The Ferret knowledge base has been converted to XTM.

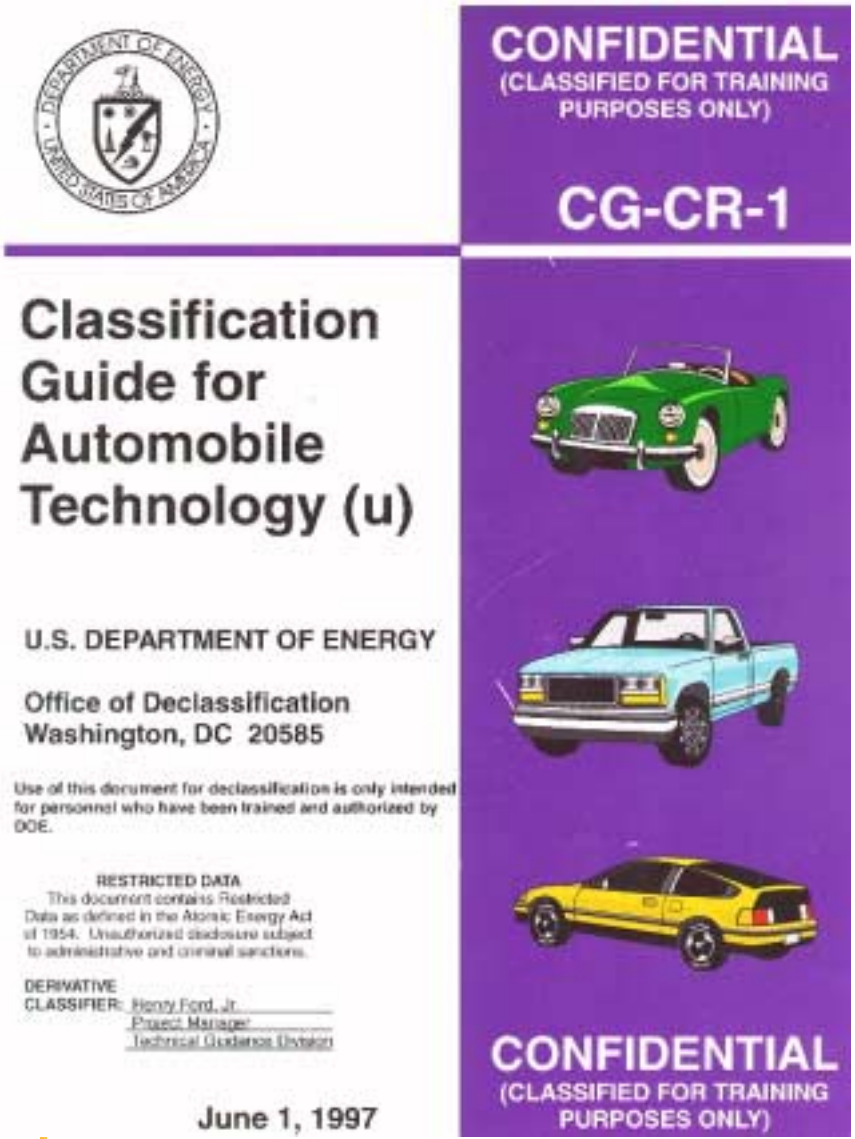
- Incorporate all knowledge structures into a single document
- Enable use of generic topic-map tools for browsing and editing
- Prepare knowledge base for other applications

The XTM version of the knowledge base includes:

- Topics for the implications and rules
- Functional topics to identify roles
- Associations
  - Implications
  - Conjunctive implications
  - Rules

Topics identify knowledge items and functional relationships.

- Implication terms (including roots of trees)
- Search terms
- Layers (for implication overlays)
- Functional roles
  - Implication
  - Conjunction
  - Antecedent (of implication)
  - Consequent (of implication)
  - Target (of rule for reporting)
  - Assignment (from rule for categorization)
  - Feedback and feedforward (to rule for internal logic)



Sample demonstration classification guide

- Contains typical rules structure
- Needs typical knowledge base
- UNCLASSIFIED

Simulated classification guidance

110	**Fuel Systems**	
110.1	Basic technology associated with fuel supply.	U
110.2	Basic technology associating carburetors with fuel supply systems.	CRD
110.3	Fact of Electronic Fuel Injection (EFI), no elaboration.	U
110.4	Information revealing theory or technology of EFI.	CRD
110.5	Identification of EFI as part of a specific engine or vehicle make or model.	SRD
110.6	Fact that a specific engine or vehicle requires high octane fuel.	SRD
110.7	Capacity of fuel tank.	U

Implication trees contain more than just words from the guides.

- Because the actual words of the guide (“theory or technology of EFI”) may not appear in candidate documents for analysis, details must be added to the implication network: “injector,” “injection pump,” “engine-control computer,” “fuel line.”
- Some concepts require lists: “vehicle make.”
- Other concepts require discriminators: “capacity of fuel tank” requires recognizing numbers and units of measure.

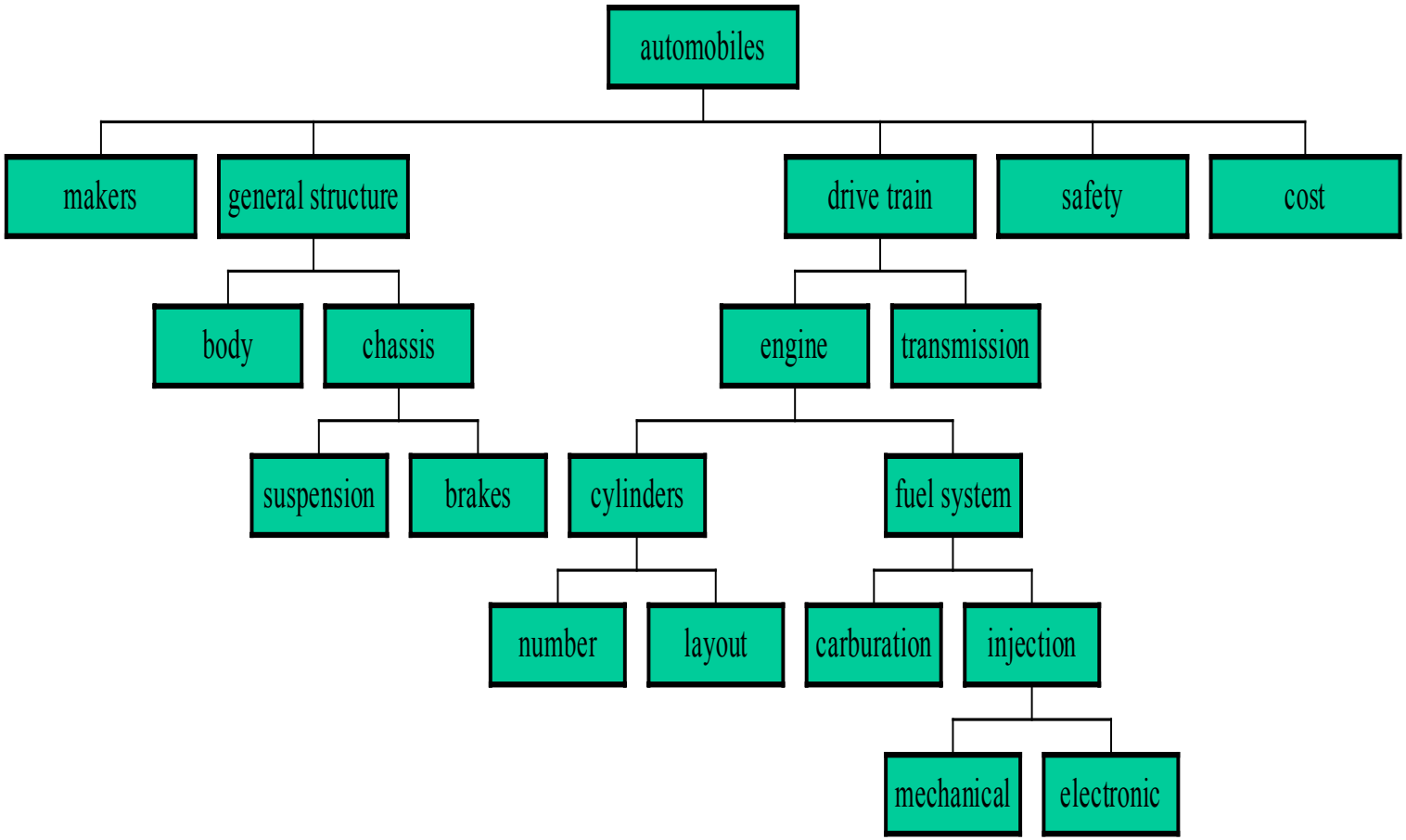
Implication flow in many domains is hierarchical.

- "The the Audi engine plant is at Gyor in Hungary."
- "Waste gate" implies a turbocharger, which in turn implies a high-performance engine.
- "Clean room" implies environmentally sensitive parts, which imply chip or disk-drive manufacture.

Some implications flow along multiple paths through overlays of implication trees.

- "Superheated steam" would imply most of the same things that "steam" would imply.
- But producing "superheated steam" depends on additional components (i.e., a superheater) and requires lubricants other than animal tallow.
- Thus a mention of petroleum lubricants for a steam engine would suggest both the implications contained in the tree "steam" and the presence of a superheater and other implications contained in the tree "superheated steam".

Implication tree for *Car Guide*



Schematic example

110.4	Information revealing theory or technology of EFI.	CRD
110.5	Identification of EFI as part of a specific engine or vehicle make or model.	SRD

- Statement: "I can get 225 horsepower from my Audi A4 by reprogramming the turbo boost and injector timing."
- Implications: injector $\Rightarrow$ fuel injection
  reprogram $\Rightarrow$ engine computer
  engine computer + fuel injection $\Rightarrow$ EFI
  Audi A4 $\subset$ "specific vehicle"
- Rules: reprogram + EFI $\Rightarrow$ 110.4
  EFI + "specific vehicle" $\Rightarrow$ 110.5

Elements from Ferret topic map

- A topic (one of many)
- A fragment of an implication tree
- A conjunctive implication
- A classification rule
- A sample categorization

A typical topic: EFI
(with synonyms)

```
<topic id="t-EFI">
 <instanceOf><topicRef xlink:href="#implication-term"/> </instanceOf>
   <baseName>
    <baseNameString>EFI</baseNameString>
   <variant>
    <parameters><topicRef xlink:href="#function-search-term"/>
                      </parameters>
    <variantName>
       <resourceData id="t-search-efi1">EFI</resourceData>
                      </variantName></variant>
   <variant>
    <parameters><topicRef xlink:href="#function-search-term"/>
                      </parameters>
    <variantName>
       <resourceData id="t-search-efi2">Electric Fuel Injection
                        </resourceData>
                      </variantName></variant>
   </baseName>
</topic>
```

An implication

waste gate $\vee$ intercooler $\Rightarrow$ turbocharging

```
<association id="impl-turbo">
  <instanceOf>
    <topicRef xlink:href="#function-implies"/></instanceOf>
      <scope><topicRef xlink:href="#layer-turbo"/></scope>
    <member><roleSpec><topicRef xlink:href="#role-consequent"/>
              </roleSpec>
     <topicRef xlink:href="#t-turbo"/> </member>
    <member><roleSpec><topicRef xlink:href="#role-antecedent"/>
              </roleSpec>
     <topicRef xlink:href="#t-waste-gate"/></member>
    <member><roleSpec><topicRef xlink:href="#role-antecedent"/>
               </roleSpec>
     <topicRef xlink:href="#t-intercooler"/></member>
</association>
```

Conjunctive implication

fuel injection ∧ control computer ⇒ EFI

```
<association id="impl-turbo">
  <instanceOf><topicRef xlink:href="#function-conjunction"/></instanceOf>
  <member>
   <roleSpec><topicRef xlink:href="#role-consequent"/></roleSpec>
     <topicRef xlink:href="#t-EFI"/></member>
  <member>
   <roleSpec><topicRef xlink:href="#role-antecedent"/></roleSpec>
     <topicRef xlink:href="#t-fuel-injection"/></member>
  <member>
   <roleSpec><topicRef xlink:href="#role-antecedent"/></roleSpec>
     <topicRef xlink:href="#t-control-computer"/></member>
</association>
```

A classification rule

EFI + “specific vehicle” ⇒ 110.5

```
<association id="report-110.5">
  <instanceOf><topicRef xlink:href="#function-rule"/></instanceOf>
  <scope><topicRef xlink:href="#scope-paragraph"/></scope>
  <member><roleSpec><topicRef xlink:href="#role-target"/></roleSpec>
    <topicRef xlink:href="report#110.5"/>
  </member>
  <member><roleSpec><topicRef xlink:href="#role-antecedent"/></roleSpec>
    <topicRef xlink:href="#t-EFI"/>
  </member>
  <member><roleSpec><topicRef xlink:href="#role-antecedent"/></roleSpec>
   <topicRef xlink:href="#t-specific-model"/>
  </member>
</association>
```

Categorization

```
<association id="category-2.1.4">
  <instanceOf><topicRef xlink:href="#function-assign"/></instanceOf>
  <member>
    <roleSpec><topicRef xlink:href="#role-feedback"/></roleSpec>
    <topicRef xlink:href="report#2.1"/>
  </member>
  <member>
    <roleSpec><topicRef xlink:href="#role-target"/></roleSpec>
    <topicRef xlink:href="report#2.1.4"/>
  </member>
  <member>
    <roleSpec><topicRef xlink:href="#role-antecedent"/></roleSpec>
    <topicRef xlink:href="#t-2.1"/>
  </member>
  <member>
    <roleSpec><topicRef xlink:href="#role-antecedent"/></roleSpec>
    <topicRef xlink:href="#t-criterion-for-2.1.4"/>
  </member>
</association>
```

Ferret design is independent of classification application.

- XTM knowledge base is generic.
- Output actions are separate from implication and rules processor.
- Anything that can be addressed can be called for output action.
- Ferret can be treated as a logic processor within its knowledge base.

Ferret is a fast processor.

- Knowledge base loads in about 10 seconds.
- Candidate documents process at 2,000 words per second.
- Software patent is pending.
- Ferret is available for government and commercial licensing.

Ferret applications

- Classification, risk assessment
- Categorization of abstracts
- Sorting of clinical diagnoses
- Pharmaceutical characterization
- Genome mapping
- Helpline diagnostics
- Cataloging, indexing, and topic map generation
- Scanning newsfeeds, e-mail, intranets
- Query expansion

Further information

- Government applications:
  Lockheed Martin Energy Systems,
  Michael Bell (mxb@y12.doe.gov),
  Robert McGaffey (rwm@y12.doe.gov),
  James Mason (mxm@y12.doe.gov)
  Peter Kortman (pjk@y12.doe.gov)
- Commercial applications: AreteQ,
  Charles Wilson (cwilson@usit.net)

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Copyright Notice

This document has been authored by a contractor of the U.S. Government under contract DE-AC05-00OR22800. Accordingly, the U.S. Government retains a paid-up, nonexclusive, irrevocable, world-wide license to publish or reproduce the published form of this document, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, or to allow others to do so, for U.S. Government purposes.