

Knowledge Engineering for the Ferret Analytical Engine

James David Mason

Oak Ridge Y-12 Plant

Oak Ridge, Tennessee

Managed by Lockheed Martin Energy Systems, Inc.
for the U.S. Department of Energy
under contract DE-AC05-84R21400





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 \Rightarrow material, attribute of part \Rightarrow 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, generic roles, parts of specific systems)
- Materials (metals, alloys, plastics, other materials)



Some information is clearly 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 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”.

Ferret uses three kinds of data structures:

- Trees, with simple but recursive root-branch-leaf structure
- Tables of actions, with simple row-and-cell structure
- Rules files, which are “inverse topic maps,” to link within the trees and between the trees and the tables

The Ferret knowledge base for 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.



The “inverse topic map” is the key to Ferret operation.

- 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.



CONFIDENTIAL
(CLASSIFIED FOR TRAINING
PURPOSES ONLY)

CG-CR-1

Classification Guide for Automobile Technology (u)

U.S. DEPARTMENT OF ENERGY

Office of Declassification
Washington, DC 20585

Use of this document for declassification is only intended
for personnel who have been trained and authorized by
DOE.

RESTRICTED DATA

This document contains Restricted
Data as defined in the Atomic Energy Act
of 1954. Unauthorized disclosure subject
to administrative and criminal sanctions.

DERIVATIVE

CLASSIFIER: Henry Ford, Jr.
Project Manager
Technical Guidance Division

June 1, 1997



CONFIDENTIAL
(CLASSIFIED FOR TRAINING
PURPOSES ONLY)

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



Ferret classification actions table

<!ELEMENT actionset (rule*)>

<!ELEMENT rule (guide, item, classification, guidetext, note*)>

<!ELEMENT classification EMPTY>

<!ATTLIST classification

category (NA | RD | FRD | NSI | none-FRD | none-RD |
none-NSI | RD-FRD | NSI-RD | NSI-FRD) "RD"

level (NA | U | C | S | TS | U-TS | U-S | U-C | C-TS | C-S | S-TS) "S"

exemption CDATA #IMPLIED >

<!ELEMENT guide (#PCDATA)>

<!ELEMENT item (#PCDATA)>

<!ELEMENT guidetext (#PCDATA)>

<!ELEMENT note (#PCDATA)>



Conversion of guidance to rules is straightforward.

```
<rule><guide>CGCR1</guide><item>110.2</item>  
<classification category="RD" level="C">  
<guidetext>Basic technology associating  
carburetors with fuel supply  
systems.</guidetext></rule>
```

```
<rule><guide>CGCR1</guide><item>110.3</item>  
<classification category="RD" level="S">  
<guidetext>Identification of EFI as part of a  
specific engine or vehicle make or  
model.</guidetext></rule>
```



Ferret implication tree structure

<!ELEMENT tree (root, branches*)>

<!ELEMENT root (concept)>

<!ELEMENT concept (term, synonyms?)>

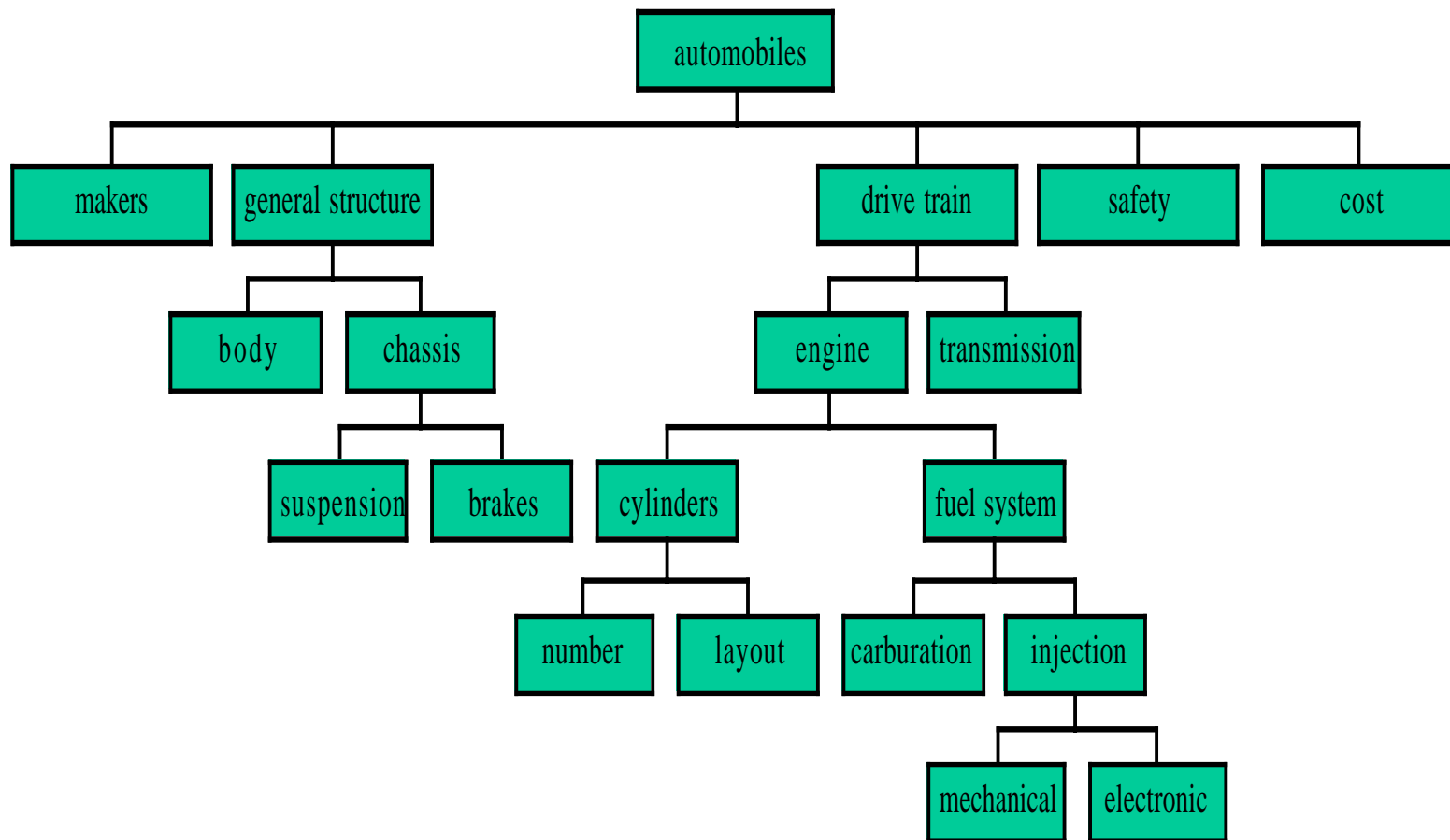
<!ELEMENT branches ((concept | tree)*)>

<!ELEMENT synonyms (term*)>

<!ELEMENT term (#PCDATA)>



Implication tree for *Car Guide*



Tree file for car guide

```
<tree><root><concept><term>automobile</term></concept>  
</root><branches>
```

```
<tree><root><concept><term>drivetrain</term>  
</concept></root><branches>
```

```
<tree><root><concept><term>engine</term>  
</concept></root><branches>
```

```
<tree><root><concept><term>fuel  
system</term></concept></root><branches>
```

```
<tree><root><concept><term>fuel  
injection</term></concept></root><branches>  
<concept><term>electronic fuel  
injection</term><synonyms<term>EFI</term>  
</synonyms>
```



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.

Components of the Ferret “inverse topic map”

- **Topics:** rules that connect between results from the implication processor and actions to be taken
- **Associations:** “conjunctive implications” to be processed within the implication processor
- **Themes:** concepts where implication trees must be linked or superimposed by the implication processor

Ferret action rules are like topics, but inverted.

<!ELEMENT rule (concept+, action+)>

<!ATTLIST rule

id ID #IMPLIED

scope_range CDATA #REQUIRED

scope_type (collection | document | paragraph | sentence |
phrase | word) "document" >

- Concepts, like conventional topic occurrences, are addresses into the knowledge base. Concepts receive the results of implication processing.
- Actions, which correspond to topic names, are also addresses, but into the classification actions base. When all the concepts in a rule are true, the action is called.



Ferret “conjunctive implications” are a hybrid of topics and associations.

<!ELEMENT conj_impl (concept, concept+, target_concept)>

- Some implications depend on finding several concepts from disjoint tree locations.
- “Conjunctive implication” rules feed back into the implication processor rather than calling external actions.
- “Conjunctive implication” rules can program the processor to execute logic operations.

Ferret “implication links” bridge trees of concepts.

<!ELEMENT impl_link (concept, concept)>

<!ATTLIST impl_link type (crosslink | overlay) “overlay” >

- “Implication links” share aspects of associations and themes.
- A cross-link decreases redundant entry of concepts into trees.
- An overlay allows implications to flow both vertically and horizontally.



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

Ferret design is independent of classification application.

- Implication trees and “topic map” rules are 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-AC05-84OR21400. Accordingly, the U.S. Government retains a paid-up, nonexclusive, irrevocable, worldwide 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.

