AFRL-IF-RS-TR-2004-255
Final Technical Report
September 2004

DISPARATE ONTOLOGY UNDERSTANDING, BROKERING, LINKING AND ELABORATION (DOUBLE)

Cycorp, Incorporated

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. L233

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2004-255 has been reviewed and is approved for publication

APPROVED: /s/

NANCY A. KOZIARZ
Project Engineer

FOR THE DIRECTOR: /s/

JAMES W. CUSACK, Chief
Information Systems Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	SEPTEMBER 2004	Final Feb 02 – Jul 04

4. TITLE AND SUBTITLE
DISPARATE ONTOLOGY UNDERSTANDING, BROKERING, LINKING AND ELABORATION (DOUBLE)

5. FUNDING NUMBERS
C - F30602-02-C-0021
PE - 62301E
PR - DAML
TA - 00
WU - 19

6. AUTHOR(S)
Stephen Reed

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Cycorp, Incorporated
3721 Executive Center Drive, Suite 100
Austin Texas 78731-1615

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFSA
525 Brooks Road
Rome New York 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2004-255

11. SUPPLEMENTARY NOTES
AFRL Project Engineer: Nancy A. Koziarz/IFSA/(315) 330-2828/ Nancy.Koziarz@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The primary work under the "Disparate Ontology Understanding, Brokering, Linking, and Elaboration (DOUBLE)" effort is OpenCyc, the publication of an open-content version of Cycorp's upper ontology together with a freely distributable knowledge base store and deductive inference engine. OpenCyc is one of the largest and most comprehensive reference ontology for the Semantic Web. The other part of the effort consisted of developing an automatic web page annotator, that takes unstructured web page text as input; and outputs OWL statements corresponding to relationships between entities parsed from the text. This tool is intended to facilitate OWL adoption by automating the initial stages of the process and flattening the otherwise steep learning curve.

14. SUBJECT TERMS
Ontology, Semantic, OWL

15. NUMBER OF PAGES
27

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

TABLE OF CONTENTS

1.0 Introduction

This final report describes research progress and results accomplished by Cycorp Inc. during the development of the Disparate Ontology Understanding, Brokering, Linking, and Elaboration (DOUBLE) system for DARPA's program. The principal investigator (PI) at Cycorp Inc. is Doug Lenat. The project manager for DAML at Cycorp is Stephen Reed. The DARPA program manager currently is Dr. Mark Greaves.

The DAML program had two previous program managers, Dr. Jim Hendler and Murray Burke. The Cycorp effort diverged from the 2001 Statement Of Work as the program managers provided feedback and guidance. The original SOW was aligned with Jim Hendler's requirement for Cycorp to provide a suite of software tools with the capability of interpreting, elaborating, and translating locally developed DAML ontologies. The primary DAML deliverable from Cycorp during the Hendler tenure was OpenCyc, the publication of an open-content version of Cyc's upper ontology together with a freely distributable knowledge base store and deductive inference engine. OpenCyc remains the largest and most comprehensive reference ontology for the Semantic Web. Given the first necessary step of creating OpenCyc, we then conducted experiments on ontology interpretation (subsequently referred to as ontology mapping). In these experiments we gathered disparate ontologies (e.g. NAICS and UNSPCS) and mapped them to Cyc. We also provided java source code at SourceForge for DAML file import to, and export from Cyc. An ontology elaboration tool was demonstrated by supplying the Horus Project (a DAML program collaborator) with an enriched vocabulary of terms, given a set of sample seed terms in the Horus vehicle ontology. Rudimentary ontology translation was demonstrated via a scripted series of Cyc "Asks" (queries) that showed Cyc translating between NAICS and UNSPSC commercial organization ontologies. Care was taken not to commingle DARPA RKF and DARPA DAML tools at the direction of Jim Hendler.

In late 2002, Murray Burke reviewed our progress at the DAML PI meeting and requested that we focus on automatic term mapping – understanding a non-Cyc ontology [*O*] and developing the means to assert *O*'s content into the Cyc. He further asked that we develop novel knowledge-based algorithms for this task. So we set aside further work on ontology translation, until such time as automatic mapping tools would be available to greatly facilitate the downstream translation of ontologies. Presumably because Murray Burke also managed the DARPA RKF program, he reversed the previous policy with regard to RKF tool inclusion, allowing Cycorp to incorporate Rapid Knowledge Formation-developed tools in support of DAML program deliverables. At this time the World Wide Web Consortium (W3C) Semantic Web Activity published the OWL (Web Ontology Language) draft that immediately superceded the previous DAML standard for Semantic Web ontology authoring. In April of 2003, Cycorp published the OpenCyc content in OWL.

Inspired in part by the progress made by Cyc ontological engineers, who greatly improved Cyc's vocabulary for scripts (i.e. tightly coupled events), it seemed likely that,

by representing novel term-mapping algorithms as scripts, we could benefit the use of RKF-style tools to enter scripts (e.g. via analogy), to clarify scripts, and to answer questions about scripts. Given the reuse of RKF-developed tools for script development, a portion of our DAML-funded effort in 2003 was spent on investigating script interpreters and, subsequently, script compilers with the goal of creating a toolkit for automatic ontology mapping. Some progress was also made on a tool for semi-automatic schema mapping from a relational database schema into Cyc.

Our final DAML tool consists of an automatic web page annotator, that takes unstructured web page text as input, and outputs OWL statements corresponding to relationships between entities parsed from the text. This tool is intended to facilitate OWL adoption by automating the initial stages of the process and flattening the otherwise steep learning curve.

2.0 Objectives

It is useful to consider Cycorp DAML program objectives in three time periods, corresponding to the tenure of the three DAML program managers: Dr. Jim Hendler, Murray Burke and Dr. Mark Greaves.

Jim Hendler, as program manager, directed Cycorp to deliver a suite of tools for ontology interpretation, elaboration and translation. The Semantic Web, as envisioned by Tim Berners-Lee and Dr. Hendler, (as subsequent events have confirmed), is populated by a variety of disparate ontologies, creating the need for ontology translation as a means to automatically create a universal Semantic Web rather than a series of Semantic Web cultures – separated by ontological (e.g. term and relationship naming) differences.

Murray Burke, as program manager of both RKF and DAML DARPA programs, encouraged the re-use of RKF tools, and directed Cycorp (via PI-meeting review and feedback) to emphasize novel automatic term-mapping tools.

Dr. Mark Greaves, as the current program manager, directed Cycorp to complete its work on OWL tools for the Semantic Web. Because the DAML program budget was reduced in FY04, Dr. Greaves chose not to fund Cycorp for FY04 or FY05. Consequently our tool-development schedule was curtailed as those in-progress tools, for example the script execution infrastructure were de-emphasized in favor of other OWL tools that could best re-use / extend RKF technology and be completed using FY03 funds. The Cycorp Web Page Annotator and OWL-Full compliant import/export tools were recently completed.

3.0 OpenCyc

In 2001, under funding provided by Dr. Jim Hendler through the DARPA RKF contract, Cycorp announced its intention to release the Cyc Upper Ontology to the public together with a freely distributable deductive inference engine and KB editing tools as OpenCyc. Our intention was to provide Cyc's reference ontology to Semantic Web developers for inclusion in their own ontologies, and to position OpenCyc as a Semantic Web server. We began the development of java software exposing the Cyc API as native java objects, so that DAML and Semantic Web developers could more easily integrate OpenCyc into their own java applications. The OpenCyc project at http://www.sourceforge.net/projects/opencyc contains the open source java API code. We launched the OpenCyc web site at http://www.opencyc.org to contain news and documentation about the OpenCyc effort.

3.1 OpenCyc Releases:

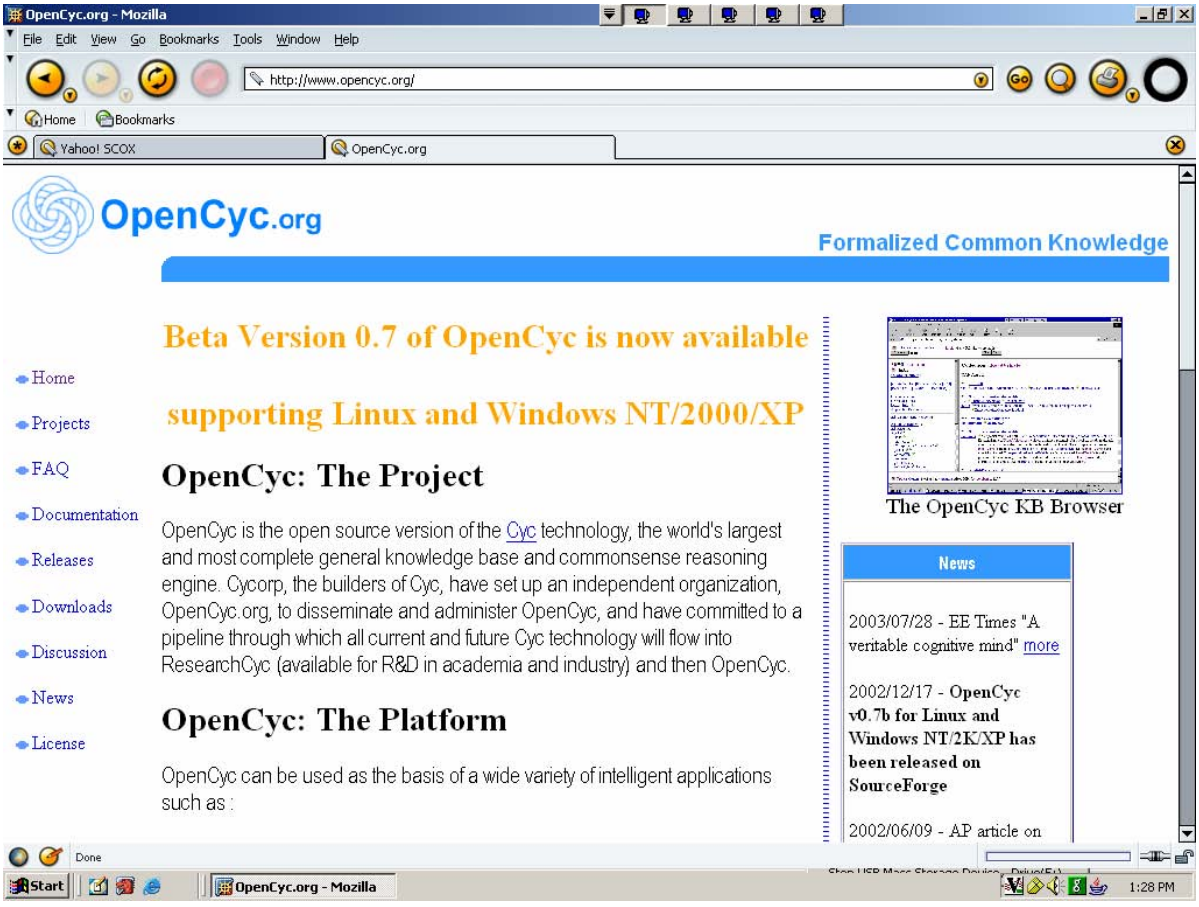
Release 0.6.0 for Linux 2002-04-02	33,103 downloads
Release 0.7.0 for Linux 2002-12-15	2,527 downloads
Release 0.7.0 for Windows/NT/2K/XP 2002-12-15	12,099 downloads

3.2 Comparing Native OpenCyc and Its OWL Version

During 2001, DAML program participants debated the feature set and characteristics of what became DAML+OIL, and subsequently OWL. OpenCyc is a knowledge base whose logical language (CycL) is primarily first order predicate calculus with second order and modal logic features. According to Cycorp's policy, commonsense knowledge representation requires the most expressive (capable) logical language. Dr. Hendler, and the majority of DAML participants (in particular those at the W3C) were highly motivated to build the Semantic Web upon the less expressive W3C RDF (Resource Definition Format) logical language. Furthermore the DAML participants decided that the most expressive elements of first order logic would be omitted in order to make Semantic Web inferences tractable. When exporting Cyc content into OWL, only binary relationships can be extracted, rules are omitted, and functional terms are omitted. But fortunately it turns out that 95% of Cyc assertions are Ground Atomic Formulae (GAFs), and that the great majority of Cyc relationships are binary. Consequently, an export of Cyc into OWL format preserves substantial information. And on the other hand, because Cyc's logical language (CycL) is more expressive than OWL, any OWL document can be completely imported into the Cyc KB.

3.3 OpenCyc.org

Below is a screen shot of the web site launched to document OpenCyc for the Semantic Web. This portal manages all the public Cyc documentation, tutorials and API docs. See Appendix 2 for a list of tutorials.



3.4 OpenCyc Ontology Improvement

With the goal of increasing the utility, consistency, correctness and organization of the OpenCyc Knowledge Base with regard to the Semantic Web, Cycorp in 2001 began the comprehensive review and cleanup of all the existing and newly proposed Upper Ontology concepts and relationships. Because DAML is the first government project for which Cyc provides a widely applicable commonsense ontology (as opposed to a military or intelligence focused domain ontology), our ontological engineers were required to sweep over the more than 5,000 terms intended for release. We had a number of new organizing methods to enforce across the board, among them the required presence of definitional assertions for each term. In 2002, we increased the resources spent on the Upper Ontology sweep and cleanup, as the most general concepts at the top of the KB pyramid were completed. Two of the most comprehensive changes to date:

- The substitution of collection-membership for objects that otherwise would have attributes. Rather than say that *an object is red*, we now say *that an object is a member of the set of red things*. These two representations are logically and linguistically equivalent, but the collection-membership version has inference implementation benefits. Furthermore it should be more straightforward for Semantic Web developers and ontology users to understand and use a single uniform representation for these relationships, rather than the two previous means to the same end.
- The identification and movement of context-free assertions to a new set of microtheories above the contextual microtheory hierarchy. Mainly the context-free assertions are those that are true in every Cyc context. Cyc inference gains from having to look in a few new microtheories for frequently required knowledge (such as the arity of a given predicate).

4.0 Published Ontologies

Dr. Jim Hendler directed DAML program participants to jump-start the Semantic Web with a series of "homework assignments" with the goal of publishing a number of DAML+OIL ontologies. Fulfilling this goal, the preliminary Upper Cyc ontology was exported into DAML in 2001 via the java API software and published at http://www.daml.org/ontologies/274. We imported the North American Industrial Classification System NAICS) codes, required for business reporting by the US Census Dept., into Cyc and exported the ontology into DAML via the java API software and published the result at http://www.daml.org/ontologies/179. We imported the DAML ontology for the Universal Standard Products and Services Classification (UNSPSC) Code, which was published by Stanford KSL at http://www.daml.org/ontologies/106. In the aftermath of 9-11 we assisted the Horus project with the elaboration of their transportation ontology. The elaboration of a user's ontology is one of the three main tasks in Cycorp's Intent Of Work for DAML FY 2001. We also contributed additional relevant portions of Cyc's reference ontology to Horus. Cycorp ontological engineers began participation in the IEEE Standard Upper Ontology working group at http://suo.ieee.org with the notion of eventually proposing OpenCyc as a standard.

4.1 Published OpenCyc OWL-Full

In June, 2004, funding from another DARPA program, IPTO ResearchCyc, permitted the identification and organization of more than 60,000 terms in the Cyc Knowledge Base. These terms were exported as OWL-Full using a tool developed last year by Cycorp within the DAML program. This greatly enlarged and enriched OWL-Full file has been submitted to SchemaWeb and is hosted at: http://www.cyc.com/2004/06/04/cyc.owl. The popular Protégé ontology editor, using its OWL plug-in, can import the entire ResearchCyc OWL skeleton ontology given sufficient memory and computer time (currently nine hours on an AMD 2100+ Linux computer).

5.0 Ontology Mapping

In 2001, as previously mentioned, we imported the DAML ontology for the Universal Standard Products and Services Classification (UNSPSC) Code, which was published by Stanford KSL at http://www.daml.org/ontologies/106. We then prepared a demo for the July 2001 Principal Investigators' meeting showing how a representative sample of terms from the NAICS and UNSPSC could be mapped to the Cyc reference ontology. The demo consisted of a series of Cyc assertions and queries that showed how Cyc could perform the essential steps of a term-mapping algorithm that was based upon a noun-phrase parse of the term names and analysis of the containing taxonomic hierarchy. We also showed how the Cyc RDF (the XML scheme upon which DAML extends) linking tool could be used to map a representative sample of terms from NAICS into Cyc concepts. In 2001, we began thinking about a Schema Mapping Tool for ontology mapping, and for mapping any structured knowledge source into Cyc's reference ontology. Our research to date (mainly the OpenDirectory linking project) demonstrates that the effort of ontology mapping is the bottleneck for ontology interpretation / translation (i.e. for semantic integration) and that specialized tools will be required to enable Semantic Web developers to map their own structured knowledge sources to Cyc. In 2002, Stephen Reed and Doug Lenat published the paper "Mapping Ontologies Into Cyc" in the AAAI Workshop on Ontologies for the Semantic Web.

5.1 Mapping OpenDirectory's Taxonomy to Cyc Terms

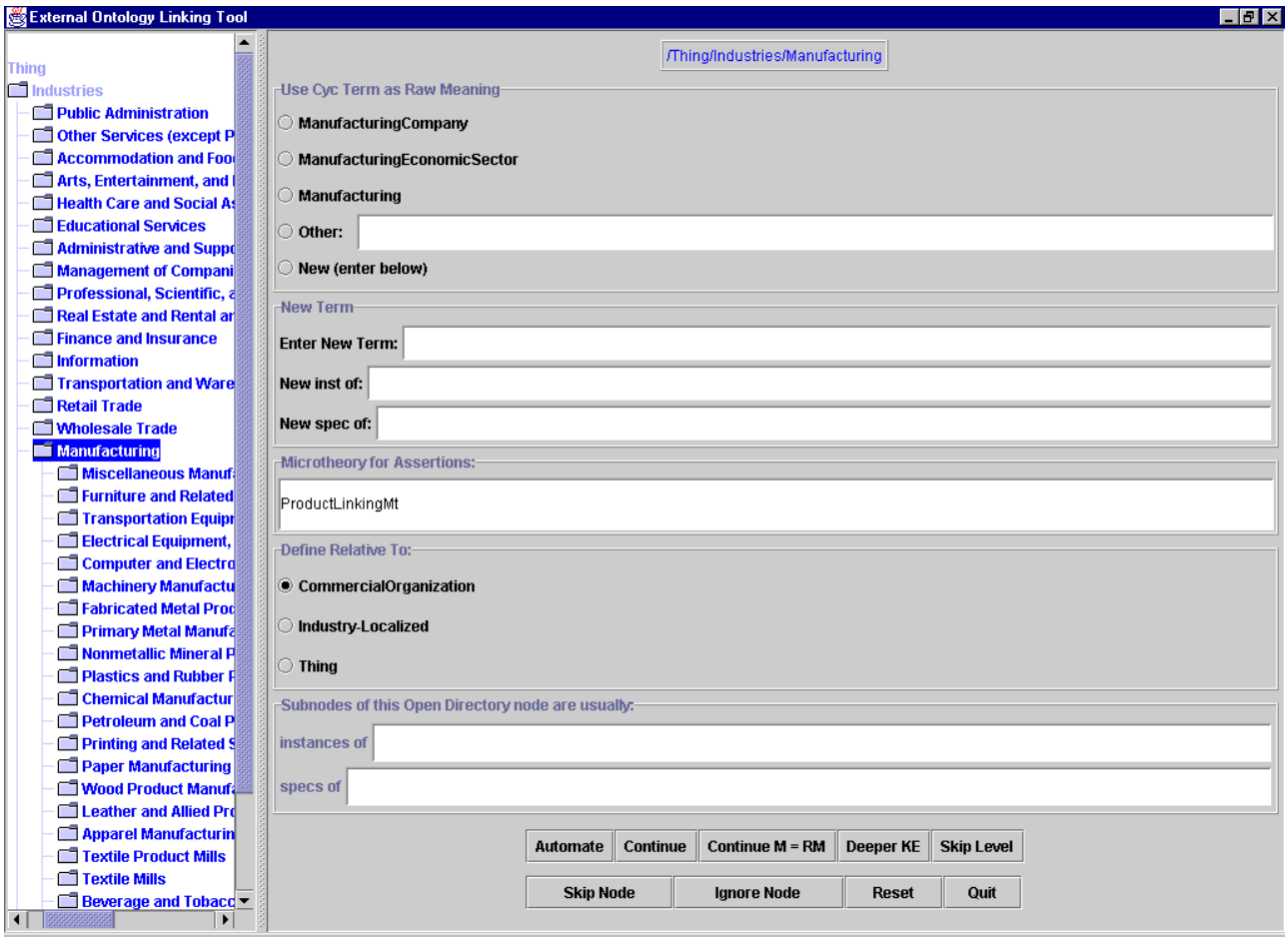
Below is a sample of the OpenDirectory RDF taxonomy that was mapped into Cyc as part of a commercially funded project. Our DAML effort investigated extensions to this work, along the lines of reducing the maintenance burden due to frequent changes in the OpenDirectory source ontology.

OpenDirectory Term and Related Terms

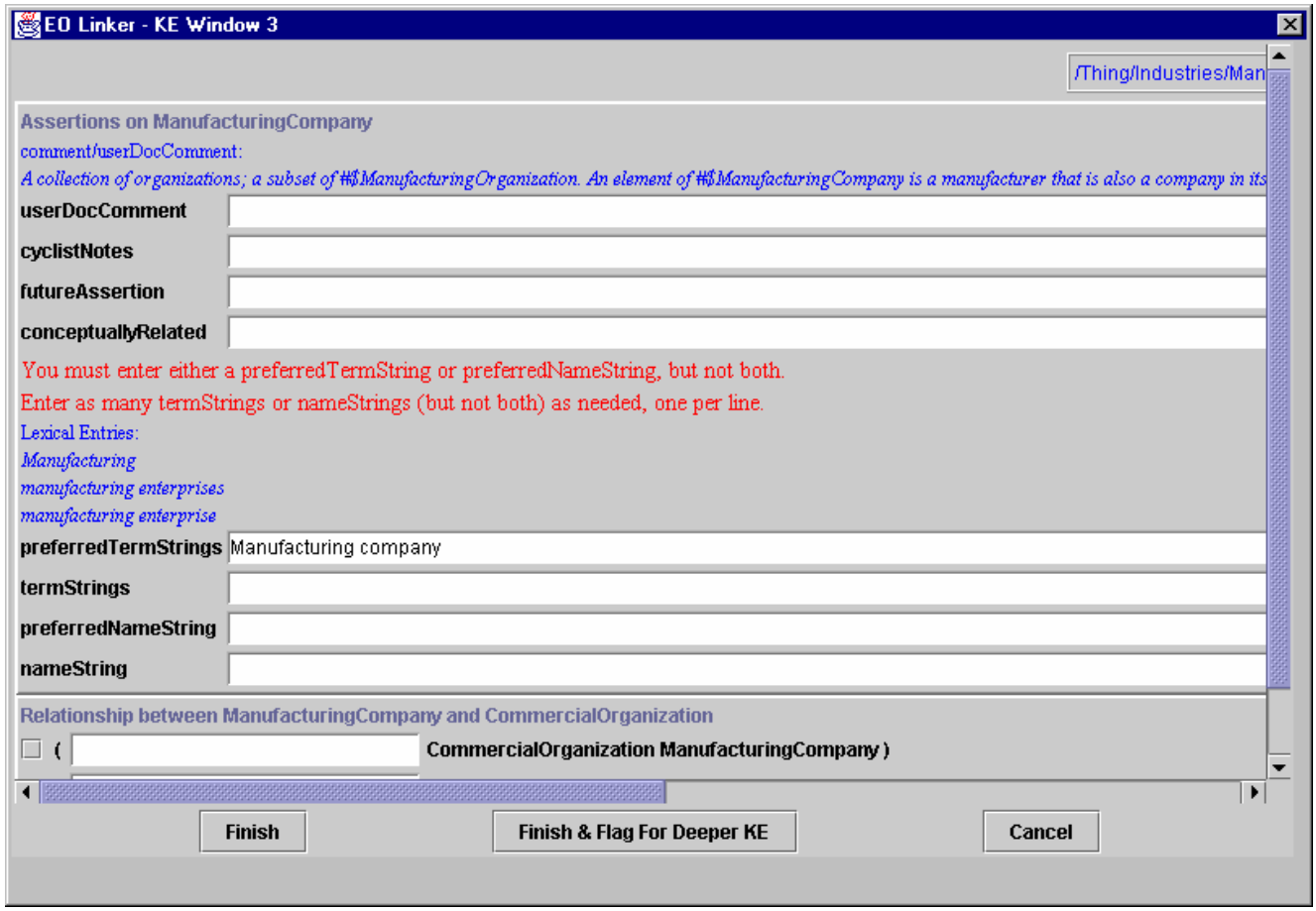
```
<Topic r:id="Top/Arts/Literature/Genres/Mystery">
  <catid>905</catid>
  <d:Description>This category is dedicated to mystery genre fiction sites.
</d:Description>
  <narrow r:resource="Top/Arts/Literature/Genres/Mystery/Reviews"/>
  <narrow r:resource="Top/Arts/Literature/Genres/Mystery/Authors"/>
  <narrow r:resource="Top/Arts/Literature/Genres/Mystery/Organizations"/>
  <narrow r:resource="Top/Arts/Literature/Genres/Mystery/Academic_Mysteries"/>
  <narrow r:resource="Top/Arts/Literature/Genres/Mystery/Whodunit_Mysteries"/>
  <narrow
r:resource="Top/Arts/Literature/Genres/Mystery/Historical_Mysteries"/>
  <narrow r:resource="Top/Arts/Literature/Genres/Mystery/Magazines_and_E-
zines"/>
  <related r:resource="Top/Arts/Writers_Resources/Fiction/Mystery"/>
  <related r:resource="Top/Society/Crime/Books_and_Authors"/>
</Topic>

OpenDirectory Top/Arts/Literature/Genres/Mystery mapped to Cyc Term via→

(#$overlappingExternalConcept
     (#$GenreFn #$LiteraryWork-CW #$Mystery-Genre) #$OpenDirectoryMt "905")
```

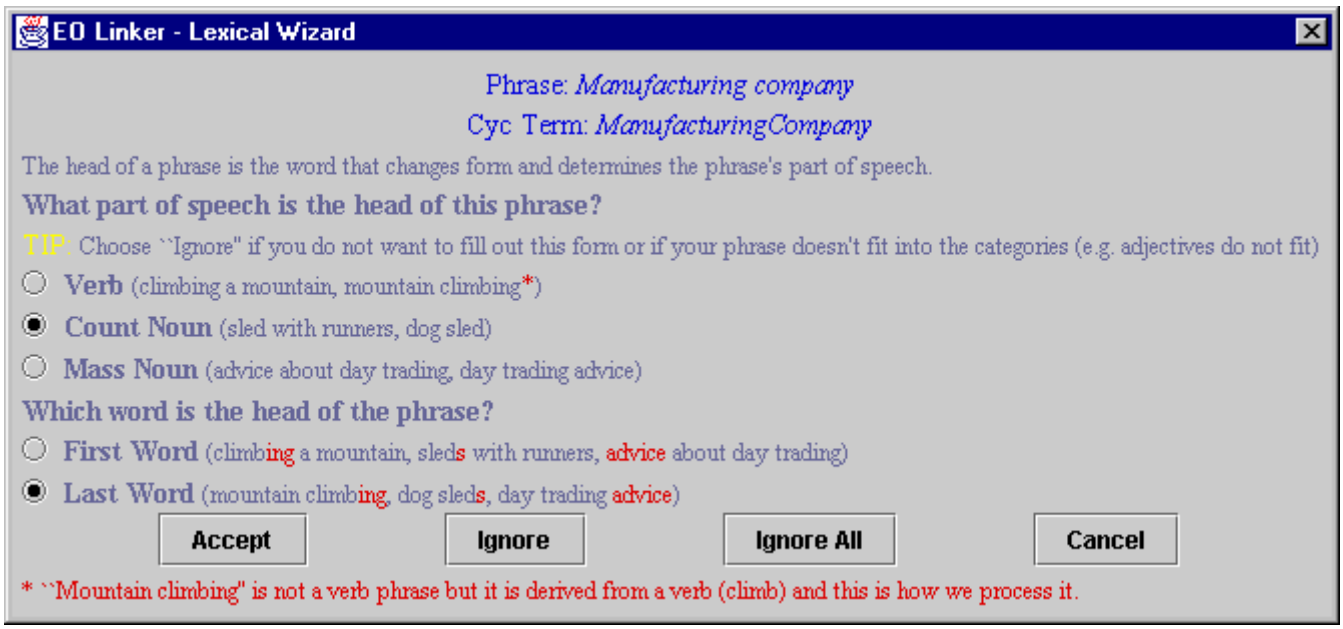
5.2 Ontology Mapping Tool

Below is a screen shot of a tool developed for the OpenDirectory Linking project (commercial sponsor) that was re-purposed for the NAICS ontology. Because this tool can import RDF information (e.g. OpenDiretory structure definition files), it was straightforward to have it import DAML+OIL.



Here are two further screen shots of the ontology-mapping tool:

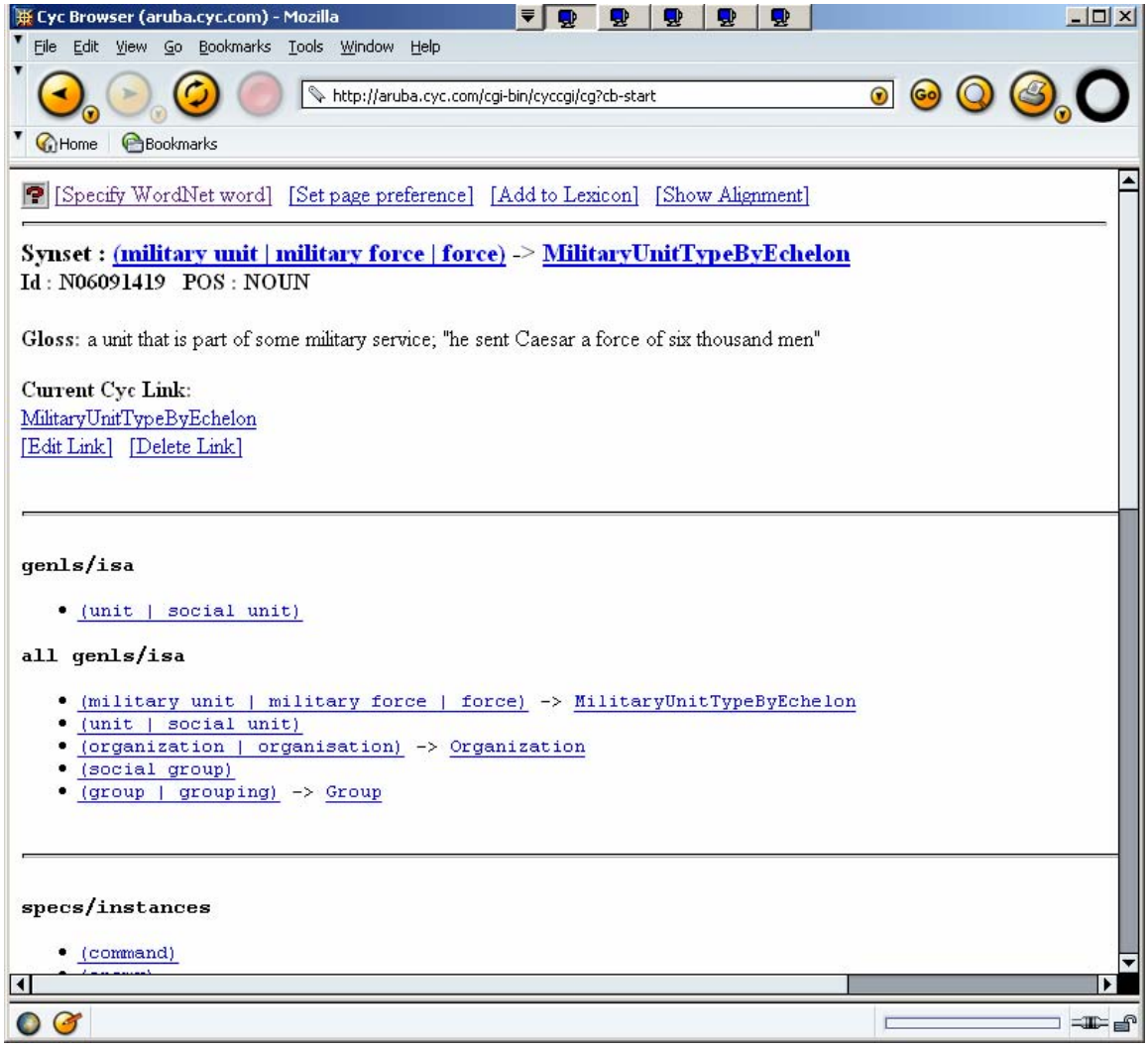




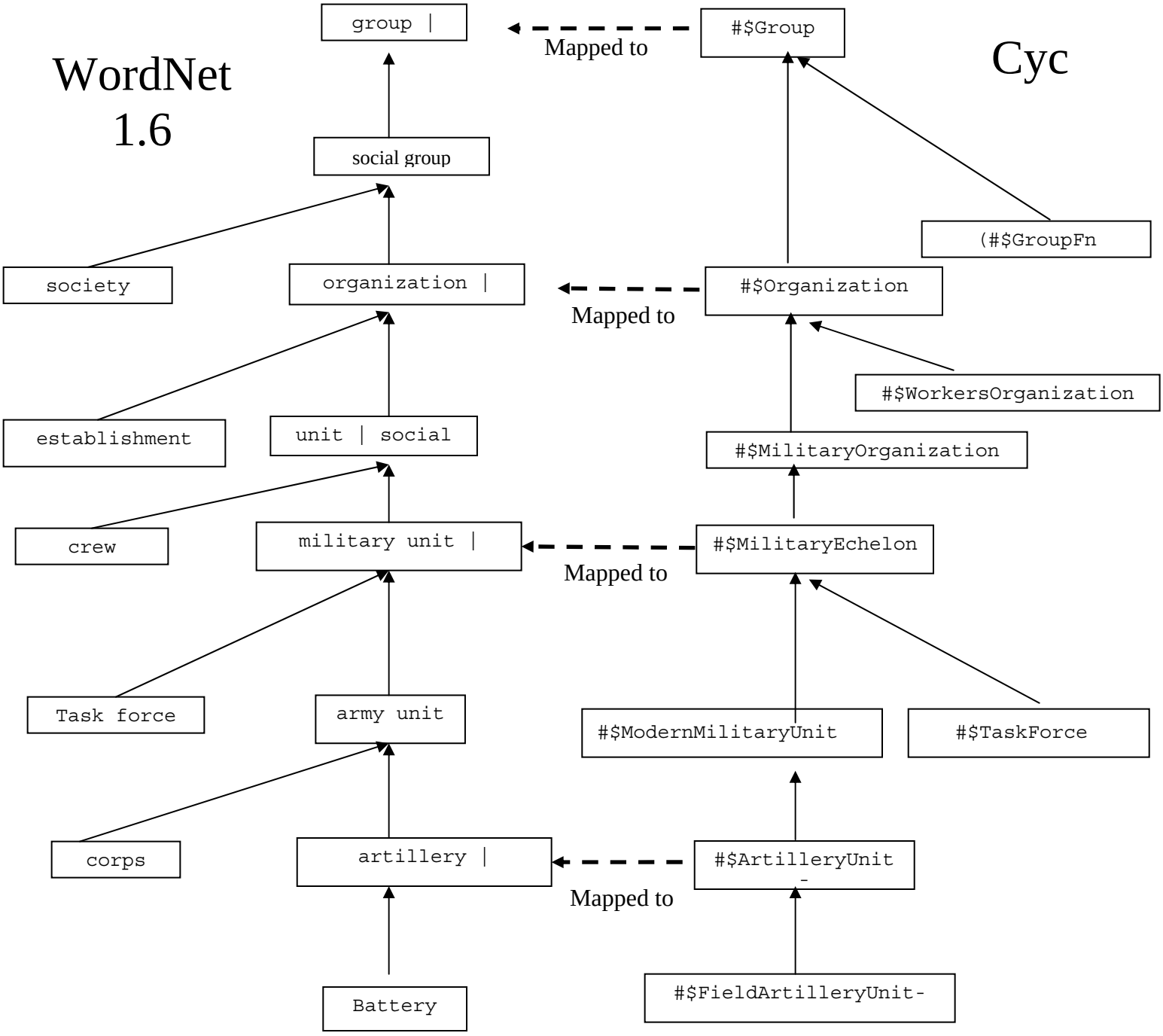
5.3 Mapping WordNet to Cyc

WordNet is the world's largest lexical knowledge base, or machine-readable dictionary. In WordNet, English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexicalized concept. Different relations link the synonym sets. Cycorp has permission from the developers of WordNet at Princeton University to include WordNet with each OpenCyc distribution. Prior work at Cycorp created a WordNet to Cyc mapping tool that was examined for possible generalization as an ontology-mapping tool.

Below is a screen shot of the WordNet ontology mapping tool illustrating the entry linking WordNet "military unit" with Cyc's #$MilitaryUnitTypeByEchelon:



WordNet/Cyc Mapping Illustration



In this illustration, WordNet 1.6 synsets (Synonym Sets) are mapped to equivalent concepts in Cyc. For example, WordNet "artillery" is mapped to Cyc #$ArtilleryUnit. Upward arrows indicate the respective concept generalization hierarchies in WordNet and Cyc.

5.4 Using Mapped Ontologies in Cyc Inference

After mapping a portion of the UNSPSC and NAICS commercial organization ontologies, Cycorp developed a use case for ontology translation. At the time, early Web Service Directories provided by IBM and Microsoft utilized both the UNSPSC and NAICS for web service lookup by organization characteristic keywords. In this scenario a hypothetical user is interested in finding web services provided by vendors of "waterproof plywood". An informal reasoning chain illustrates the kinds of Cyc queries and Cyc commonsense domain information that might be used in an automatic term mapping tool.

> **User's query:** "waterproof plywood"
> **UDDI (Directory Web Service) :** No results matching **plywood.** Failing this direct lookup by keyword, the (hypothetical) Cyc ontology translation service is automatically invoked.
> **Cyc:** Parses the query as-- the collection of **?X** for which
> **(#$and (#$isa ?X #$Plywood)**
> **(#$keepsOut ?X (#$LiquidFn #$Water))).**
> And because **#$CanopyTheShelter, #$Submarine, #$TruckTrailer** are known to be waterproof, then Cyc concludes that shelters, watercraft, and containers all include things which are waterproof.
>
> Cyc has the following knowledge from the mapped **NAICS** (National Industrial Classification System) ontology and knows that these are conceptually related to plywood (and may have knowledge of stronger relationships between these concepts):
> "321" Wood Product Manufacturing
> "3212" Veneer, Plywood and Engineered Wood Product
> "321212" Softwood Veneer and Plywood Manufacturing
> "321211" Hardwood Veneer and Plywood Manufacturing
>
> Likewise Cyc has the following knowledge from the mapped **UNSPSC** (United Nations Standard Product & Services Code) ontology:
> "7311" Wood and paper industries"
> "301515" Roofing materials
>
> Cyc then searches UDDI on the above terms that Cyc knows to be closely related to plywood, yields the following interesting result:
>
> **UDDI**: web page search results for UNSPSC/7311 ("Wood and paper industries" above) include this text:
>
> > Hardwood International - Thanks for coming to our online store. We have the finest selection of Hardwood, Hardwood Flooring, moldings, plywood, **marine plywood**, and veneer from around the world.

Hardwood International is returned to the user as a vendor of waterproof plywood by Cyc because "**marine**" is a word that Cyc associates with #$Sea, and #$Sea is composed of #$Water, and that water is conceptually related to waterproof.

5.5 Mapping A Sample Of The Army CALL Thesaurus

Another DAML program participant, DRC Corp., developed a DAML ontology for the Army Center For Lessons Learned (CALL) Thesaurus. The DRC thesaurus ontology is a meta-ontology, namely it describes the sorts of relationships that occur between terms defined in the CALL thesaurus.

call:Term – The object identifies the subject term defined in the CALL Thesaurus

call:name – The object is a name of the subject term defined in the thesaurus

call:UF – The object is a substitution phrase that subject term is Used For (i.e. a synonym).

call:NT – The object term is Narrower Term than the subject term

call:BT – The object term is Broader Term than the subject term

Using the above relationships, translated into OWL by Cycorp, we manually imported a small portion of the 18,000-term CALL thesaurus (i.e. Aircraft) into Cyc. Here is a sample:

```
<call:Term rdf:ID="GroundEffectMachines">
  <call:name>ground effect machines</call:name>
  <call:UF>ground-effect machines</call:UF>
  <call:UF>hovercraft</call:UF>
  <call:BT>Aircraft</call:BT>
  <call:BT>SurfaceEffectVehicles</call:BT>
  <call:NT>AirCushionVehicles</call:NT>
  <call:NT>CaptiveAirspaceCraft</call:NT>
</call:Term>
```

We then applied our Noun Phrase parser to automatically map CALL term names into equivalent Cyc terms. We achieved 60% accuracy (i.e. precision). In the above fragment, "hovercraft" was automatically mapped but "captive airspace craft" failed to automatically map (due to insufficient Cyc knowledge about concept "captive" in the context of "airspace craft"). Automatic term mapping using the Cyc Noun Phrase parser was accomplished using the four following methods.

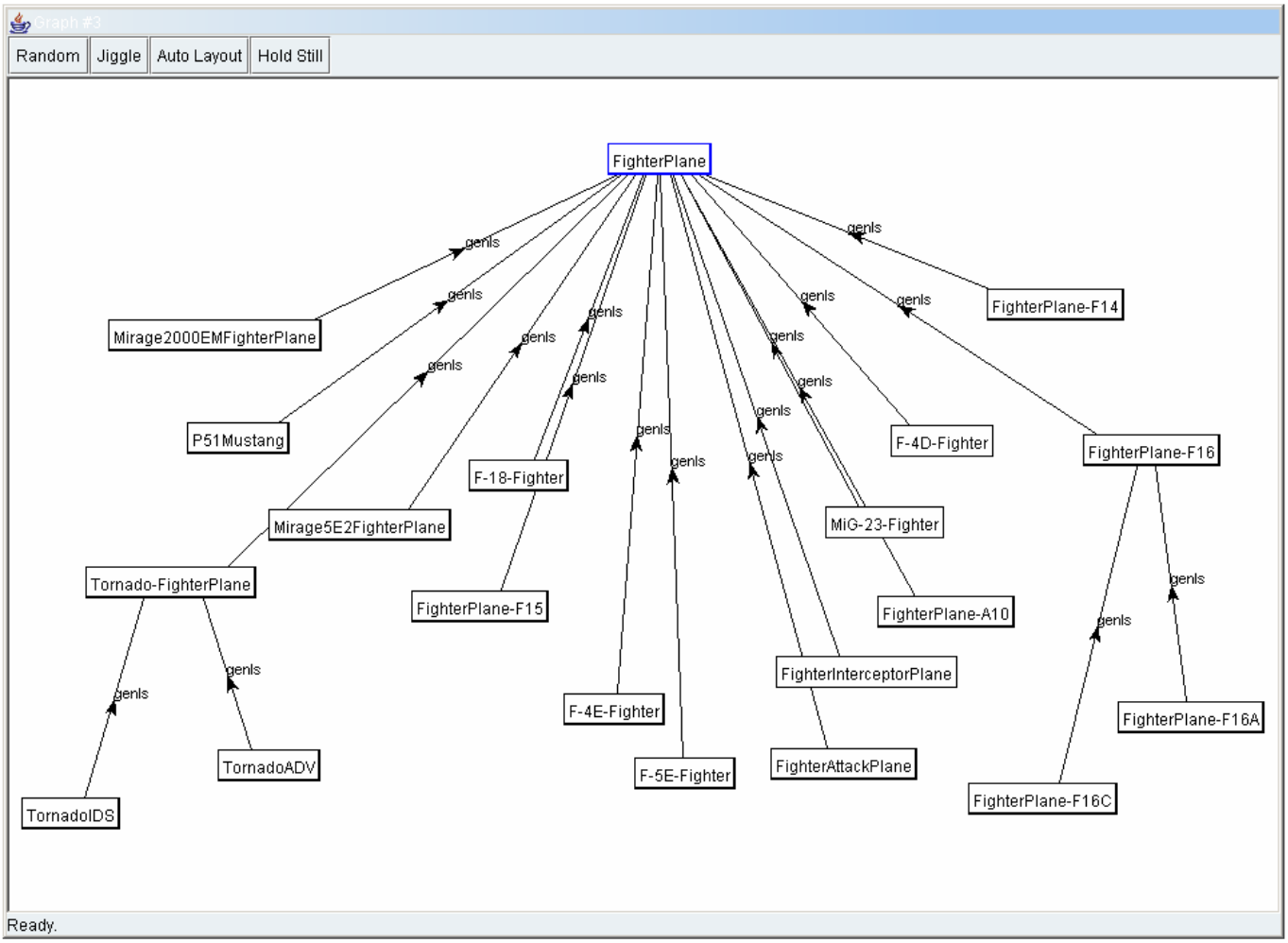
- Parsing via lexical links from single-word CALL term names to Cyc concepts. "drone" → cyc:UnmannedAircraft Parsing via functional roles. In this example, attack is a functional role of the aircraft. "attack aircraft" → cyc:((#$SubcollectionOfWithRelationToTypeFn #$AttackOnObject #$instrument-Generic#$AirTransportationDevice))
- Parsing distinctions according to word morphology. In this example, "towed" and "towing" have quite different mappings. "towing vehicles" →

cyc:((#$SubcollectionOfWithRelationToTypeFn #$TowingARoadVehicle #$instrument-Generic #$TransportationDevice-Vehicle))

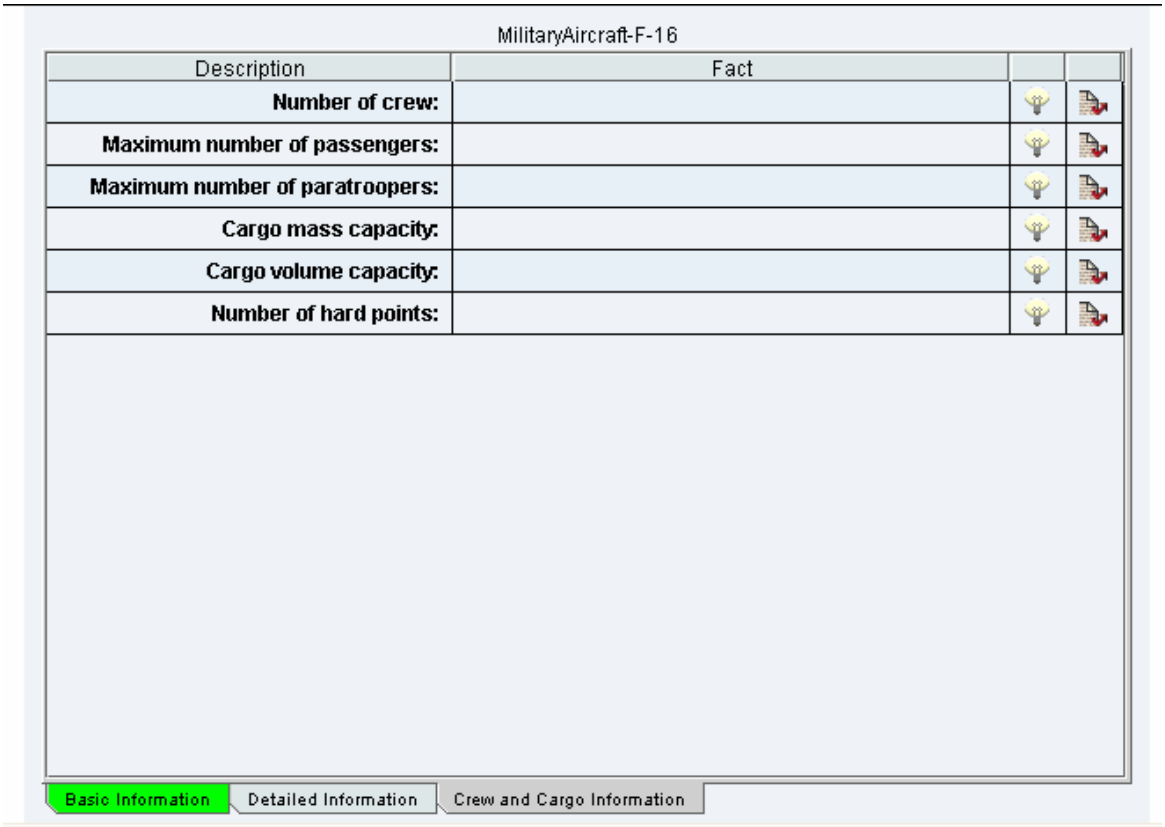
- Parsing some noun phrases into Cyc concepts that include a predicate (relationship). "rocket aircraft" → cyc:((#$SubcollectionOfWithRelationToTypeFn #$AirTransportationDevice #$poweredBy #$RocketEngine)) In this case the relevant Cyc predicate is #$poweredBy.

Here is a screen shot of the Center for Army Lessons Learned Thesaurus web page:

Below is an illustration of the Cyc ontology for #$FighterPlane into which CALL “F-16” was mapped to Cyc FighterPlane-F16 using the Cyc ontology graphing tool. The “genls” links illustrate the Cyc type subsumption hierarchy. Although this graph does not show multiple genls links for a term, Cyc’s concept hierarchy allows multiple inheritance.



The Cyc Fact Entry Tool is show below for the term MilitaryAircraft-F-16. After an ontology is mapped into Cyc, this tool can be used to elaborate relevant facts about the focal concept.



5.6 Adding Unified Modeling Language (UML) to Cyc

During late 2002 and early 2003, we added the Unified Modeling Language (UML) core concepts and relationships to Cyc. A total of 238 modeling concepts were represented in Cyc, including 61 binary predicates (OWL properties). Our research effort was to determine the utility of UML state charts for representing the procedures required for complex term mapping. We constructed a UML state machine interpreter that could execute a procedure represented in Cyc using UML vocabulary. A simple state machine was designed that counts from one to ten and then terminates. This state machine was input into Cyc as assertions using Cyc microtheories to contain reusable UML components such as procedure definitions. Below is a CycL assertion describing the body of a UML procedure that increments a given input number by one.

```
(#$umlBody #$UMLProcedure-IncrementNumber
  (#$ProgramAssignmentFn
    (#$SoftwareParameterFromSyntaxFn
         #$UMLProcedure-IncrementNumber-OutputPin-X)
      (#$PlusFn 1
        ((#$FunctionToArg 2 softwareParameterValue)
         (#$SoftwareParameterFromSyntaxFn
            #$UMLProcedure-IncrementNumber-InputPin-X)))))
```

Our experiments demonstrated that a state machine designed in UML could be asserted into Cyc and executed via a Cyc application. Because we chose to represent dynamic states also as Cyc assertions, the cycle time of the executing state machine was limited by Cyc's assertion and retraction speed of approximately 10 operations per second. This speed is at least four orders of magnitude too slow to usefully implement algorithms for complex term mapping. Consequently, later in 2003 we developed a means of compiling CycL script (procedure-describing) knowledge into efficient java programs, as described in the next section.

5.7 Compiled Behavior Language for Complex Term Mapping Algorithms

We learned from the UML state machine experiment in early 2003 that keeping dynamic procedural state in Cyc as KB assertions resulted in implementations that are too slow for complex term mapping algorithms. Cycorp ontological engineers, funded by other programs, had elaborated Cyc's scripting vocabulary to the point where it could be used to represent complex (i.e. nested, reusable, multi-role, multi-agent) procedures.

The Semantic Web consists of numerous incompatible ontologies (e.g. UNSPSC and NAIC ontologies for commercial organization description). Algorithms for automatic term mapping that achieve compatibility among disparate ontologies, are complex. For an example refer back to section 5.4.

Cyc Behavior Language is a robust, semi-declarative specification for Cyc's deliberative and reactive behavior. CBL is intended for use as a general purpose ***CycL Script Compiler***, enabling Cyc to become an actor with regard to scripted events it knows about. CBL is implemented on the Java platform, and is closely coupled to the Cyc inference engine, KB store, Natural Language, and Knowledge Entry/Query functions.
CBL shares numerous features with these published agent control languages (ACLs), hierarchical planning and robot control architectures:

NIST Real-Time Control System, James Albus
Believable Agents, Bates (CMU Oz Project)
Tapir, King, Atkin, Westbrook, Cohen (UMass)
Task Description Language, Simmons (CMU)
Procedural Reasoning System, Lee, Huber, Durfee, Kenny (UMich)
Simple Hierarchical Ordered Planner, Nau (UMBC)

CBL includes the most useful features of the surveyed ACLs:

- Real-world actuators, sensors and actuator-sensors, including Semantic Web service providers (actuators) and service consumers (sensors)
- Actuators accept commanded actions
- Sensors produce sensations, e.g. a measurement or perception
- Tasks consist of either an action (task-sequence or action) or a goal
- Multiple methods may be defined to accomplish each task
- Tasks succeed or fail, and further methods are tried in case of failure
- Tasks propagate downwards through the control node hierarchy
- Sensations/perceptions propagate upwards through the control node hierarchy
- Messages can be sent between nodes, e.g. configuring sensors, task status
- Method lexically scoped and node shared state variables may be declared
- Task, sensor, message and exception processing run in separate threads within each control node
- Task commands, sensations, messages and exceptions are all transmitted asynchronously

Reinforcement learning is provided for method choice, action ordering, action subsets, and relevant objectsBecause of its close coupling to Cyc, CBL has these novel features:
CBL can be in principle be derived indirectly via deduced scripts, from non-event, commonsense knowledge in Cyc.
CBL can be directly extracted from Cyc's Script representation.
Cyc Ask, Assert, Unassert and Create-Constant operations are CBL task built-ins.
In principle, one programs in CBL by asking Cyc to perform a task, and by subsequently teaching Cyc how to address the performance failures.

Note that the script compilation effort did not progress far enough to execute an automatic term mapping algorithm in the DAML program.

6.0 Automatic Web Page Annotation With OWL

The Semantic Web has a great deal more unstructured text than structured (e.g. database) content. DARPA customers, government intelligence agencies, and civilian Semantic Web users would benefit from a tool that automatically annotates text web pages with OWL statements. The existence of such a tool would substantially lower the barrier to entry into the community of semantic web content providers, and reduce the learning load required to produce OWL annotated pages to that required to gain editing skills, rather than “from scratch” authoring skills.

In response to Dr. Mark Greaves’ DAML program statement of direction in fall of 2003, Cycorp constructed a tool that re-uses RKF style natural language parsers to create Cyc assertions from parsed web page text, and then uses the DAML-funded OWL export tool to create OWL statements for the parsed text. The tool pipeline uses Apache.org java components to parse and traverse web pages. The Xerces-Java component uses the CyberNeko HTML parser to create a DOM (Document Object Model) of the target web page. Apache Xalan contains an Xpath implementation that will access a DOM object according to a given Xpath locator expression. For example the below Xpath expression extracts the Sunday weather forecast from the National Weather service NWS web page for Austin, Texas on Friday, January 30, 2004:

/HTML/[2]/BODY[7]/TABLE[7]/TR[1]/TD[2]/TABLE[1]/TR[1]/TD[17]/B[0]

The descriptive text located in the web page by the above Xpath is "Mostly cloudy in the morning then becoming partly cloudy. Areas of fog in the morning. Highs in the mid 60's. Southwest winds 10 to 15 mph."

We applied Cyc's natural language parsers to such web pages to extract CycL concepts and axioms that then are emitted as OWL for automatic web page markup.

On the following page there is a typical terrorism news story followed by a sample of the OWL markup extracted from the English sentences in the news story.


```
    <guid>bd5893c1-9c29-11b1-9dad-c379636f7270</guid>
  </Agent-Generic>
- <Bomb rdf:ID="SKF-0000221211">
    <rdfs:label xml:lang="en">some bomb (weapon)</rdfs:label>
    <guid>bd5893c1-9c29-11b1-9dad-c379636f7270</guid>
  </Bomb>
- <Leader rdf:ID="SKF-0074145264">
    <rdfs:label xml:lang="en">some head (person)</rdfs:label>
    <guid>bd5893c1-9c29-11b1-9dad-c379636f7270</guid>
  </Leader>
- <Organism-Whole rdf:ID="SKF-0074360693">
    <rdfs:label xml:lang="en">some organism</rdfs:label>
    <guid>bd5893c1-9c29-11b1-9dad-c379636f7270</guid>
  </Organism-Whole>
- <Event rdf:ID="SKF-0218914796">
    <rdfs:label xml:lang="en">some event</rdfs:label>
    <guid>bd5893c1-9c29-11b1-9dad-c379636f7270</guid>
  </Event>
- <Campaign rdf:ID="SKF-0221856368">
    <rdfs:label xml:lang="en">some campaign</rdfs:label>
    <guid>bd5893c1-9c29-11b1-9dad-c379636f7270</guid>
  </Campaign>
- <Bombing rdf:ID="SKF-0251705840">
    <rdfs:label xml:lang="en">some bombing</rdfs:label>
    <guid>bd5893c1-9c29-11b1-9dad-c379636f7270</guid>
    <instantiatesScript rdf:resource="#Bombing" />
    <possessiveRelation rdf:resource="#SKF-7547163201" />
    <instantiatesScript rdf:resource="#Bombing" />
    <possessiveRelation rdf:resource="#SKF-7547163201" />
  </Bombing>
- <Leader rdf:ID="SKF-0729272883">
```

6.1 Issues Regarding Web Page Annotation

- Skolem (anonymous) term names are opaque in most OWL tools – we need to incorporate other further techniques to match entities identified elsewhere in the document.CycL parsing is performed on the phrase level and does not connect phrases via discourse analysis.Adding discourse and context analysis could filter wrong parses, improving the precision of the results.

6.2 Web Page Annotation Performance

- Total number of phrases attempted: 210
- Number of phrases Cyclified: 79
- Average duration of attempt: 5 seconds

7.0 Conclusion

OWL amplifies the current XML (eXtensible Markup Language) Web Services / Semantic Web revolution to the degree that it provides a scalable way to add semantics. OWL is creating a world where agents, search engines, and other programs can read semantic markup to decipher the real meaning of a web page. OWL-literate agents are able to retrieve computer readable facts, integrate and reason about those facts, answer questions, solve problems, and generally bring a new level of intelligence to the WWW that is unimaginable with previous technology.

OWL's success is enabled in part, by solving three key problems addressed by Cycorp's research: (1) publishing a comprehensive reference ontology – **OpenCyc** and **mapping** it to existing ,valuable, mission-critical domain ontologies, including database schemas, and otherwise structured knowledge sources, (2) educating novice users about ontological techniques and providing tools to create, extend and revise semantic markup easily (launching OpenCyc.org), and (3) developing the tools, namely **web page annotation**, that harvest the semantically rich but ontologically inconsistent Semantic Web.

We found that a richly axiomatized knowledge base, the Cyc KB, can provide the foundation for software tools that intelligently coordinate information contained in heterogeneous ontologies built on OWL semantics. These tools capitalize on the KB's broad real-world knowledge and commonsense reasoning to enrich and link simple OWL ontologies.

To perform the tasks of interpreting, elaborating, and translating heterogeneous ontologies, Cycorp developed several innovations, including:

- lexical tools that employ the Cyc KB to provide intelligent parsing and interpretation of noun phrases in OWL tags created by others as described in section 5.5 - *Mapping A Sample Of The Army CALL Thesaurus*;
- KB-based tools for automatically constituting and reasoning about complex concepts that are implied by the OWL tags, but are not explicitly present, as described in section 5.3 *Mapping WordNet to Cyc*, and in section 5.4 - *Using Mapped Ontologies in Cyc Inference*;

- management of microtheory inheritance, together with more sophisticated and fine-grained "wrapping" of assertions, to specify contexts as described in section 6.0 *Automatic Web Page Annotation With OWL*.
- behavior language, consisting of Cyc knowledge about scripted actions, compiled into executable java programs and intended to implement complex term-mapping algorithms, as described in section 5.7 *Compiled Behavior Language for Complex Term Mapping Algorithms*

Cycorp's central hypothesis is that a complex knowledge base, specifically one containing both large amounts of world knowledge and contextually sensitive meta-knowledge about ontological structures, can function as a hub for interrelating thousands, and potentially millions, of local ontologies written in OWL. We tested the full power of Cyc's lexical and conceptual knowledge to map, reason about, and to some extent, translate candidate ontologies in diverse domains, such as WordNet, Horus, Army CALL Thesaurus, DAML experimental ontologies, UML, OpenDirectory, NAICS, and UNSPSC. On the basis of our completed research, we believe that a highly beneficial area for future research is highly automated ontology term mapping, including UML, database schemas, and other structured descriptions enabling the Semantic Web to function as an integrated whole as opposed to islands of automation.

Appendix 1 – Published Paper

In 2002, Stephen Reed and Doug Lenat authored Mapping Ontologies Into Cyc that was published during the AAAI Workshop on Ontologies for the Semantic Web.

Appendix 2 – OpenCyc Tutorial Links

DAML funding made possible the organization, revision, and augmentation of a substantial corpus of Cyc documentation, published at OpenCyc.org for the benefit of Semantic Web ontology authors using OpenCyc as reference ontology or starting point for their domain ontology. Each of the below links returns a richly annotated slide presentation on its topic.

- **KB Browser Interface Overview**
  - [pdf] [zip] **Interface Overview**
- **Foundations of Knowledge Representation in Cyc**
  - [pdf] [zip] **Why Use Logic?**
  - [pdf] [zip] **CycL Syntax**
  - [pdf] [zip] **Collections and Individuals**
  - [pdf] [zip] **Microtheories**
- **Predicates and Denotational Functions**
  - [pdf] [zip] **The Basics**
  - [pdf] [zip] **Arity**
  - [pdf] [zip] **Argument Types**
  - [pdf] [zip] **Second-order Predicates**
  - [pdf] [zip] **More On Functions**
- **Errors in Representing Knowledge**
  - [pdf] [zip] **Errors with Constants, Variables and Reliance on NL**
  - [pdf] [zip] **Errors with Specialization, Generalization & Rules**
  - [pdf] [zip] **Other Errors**
- **Survey of Knowledge Base Content**
  - [pdf] [zip] **Introduction**
  - [pdf] [zip] **Fundamental ExpressionTypes**
  - [pdf] [zip] **Top Level Collections**
  - [pdf] [zip] **Time and Dates**
  - [pdf] [zip] **Spatial Properties and Relations**
  - [pdf] [zip] **Event Types**
  - [pdf] [zip] **Information**
  - [pdf] [zip] **More Content Areas**
- **OE Example: Events and Roles**

[pdf] [zip] **Events in Cyc**

[pdf] [zip] **Roles and Event Predicates**

[pdf] [zip] **Actor Slots**

[pdf] [zip] **Sub-events**

Writing Efficient CycL: Some Concrete Suggestions

[pdf] [zip] **Writing Efficient CycL: Part 1**

[pdf] [zip] **Writing Efficient CycL: Part 2**

Inference in Cyc

[pdf] [zip] **Logical Aspects of Inference**

[pdf] [zip] **Incompleteness in Searching**

[pdf] [zip] **Incompleteness from Resource Bounds and Continuable Searches**

[pdf] [zip] **Efficiency Through Heuristics**

[pdf] [zip] **Inference Features in Cyc**