

Semantic Knowledge Source Integration: A Progress Report

Dr. James Masters and Dr. Zelal Güngördü,
Cycorp, Inc. 3721 Executive Center Dr. Austin, TX

Abstract

Semantic Knowledge Source Integration is an ongoing research project at Cycorp which is developing technologies that enable users to integrate structured knowledge sources in a variety of formats with the Cyc Knowledge Base. Once integrated, the contents of the external sources are available to Cyc for question answering as if the source were “part of” the Knowledge Base. In this paper, we give a brief overview of the Semantic Knowledge Source Integration project at Cycorp, discuss its recent, current, and future research directions, and demonstrate the added value of integrating a source with Cyc by showing how Cyc’s inference engine extrapolates knowledge from the source that could not be obtained by querying the source directly.

1 Introduction

Analyzing the vast amount of knowledge available from relational databases, queryable web pages, web services, and other forms of structured information exchange available over the Internet presents a challenge to analysts, engineers, and other decision makers in the government, military, and industry. Manually integrating the results of queries from multiple sources that do not interoperate is both time-consuming and susceptible to error and oversight.

Many research and development programs sponsored through the government have long recognized the need to integrate data sources in such a way that their contents can be shared across the enterprise. Several researchers and software developers have constructed and promoted “mediator” technologies to link multiple heterogeneous relational databases with different schemas together to allow their data to be viewed and queried from a central location.[3, 4, 6, 1] More recently, web ontology languages such as DAML+OIL, OWL, and others have gained currency with the advent of the *semantic*

web and at the academic and corporate levels at least, these technologies are revolutionizing the manner by which users, and software agents interact with information over the internet.[2] However, government agencies and entities have yet to adopt such technologies *en masse*. Instead, most of the voluminous information stored electronically is still embedded in legacy databases and file storage formats that were not designed to interact easily with other software.

While there have been many approaches and competing programs for achieving interoperability between knowledge sources, most or all have recognized that achieving a truly content based integration of information requires that the semantics of the sources are aligned in a way that is independent of their syntax. That is to say, their *ontologies* must be aligned with one another or with a third ontology so that the meaning of each source’s content is expressible using shared semantics. To this end, Cycorp¹ is developing *Semantic Knowledge Source Integration* (SKSI), a technology to integrate a variety of knowledge sources via the Cyc Knowledge Base (Cyc KB, or simply KB).[5] SKSI is an ongoing research and development project at Cycorp established to extend Cyc’s knowledge beyond the contents of its own knowledge base by integrating it with the vast amount of data available from structured knowledge sources distributed across the global information network. Each source’s content is treated as part of the Cyc knowledge base, although it remains on the source’s native server. One result is that the content of these sources may be used in conjunction with the knowledge stored in the KB to answer queries posed to Cyc.

Our approach is to map each source’s semantics into the Cyc ontology. We believe that since the Cyc ontology is the largest and most thoroughly developed ontology available, it is best suited for the task of modelling the semantics of arbitrary knowledge sources. Given the vast range of concepts described in the Cyc KB, there is a high likelihood that the semantics of any source can be mapped into existing semantics in Cyc. When a concept is discovered in a knowledge source that does not have a corresponding concept in Cyc, the CycL language is expressive enough to permit the user to define new concepts in Cyc on an ad hoc basis.²

¹<http://www.cyc.com>

²CycL is Cyc’s language for expressing common sense knowledge. It is a formal language whose syntax derives from first-order predicate calculus. For further information, the reader is re-

In addition to developing the capability to answer queries using the content of existing knowledge sources, we are also working on the capability to create, modify, and “publish” knowledge sources through Cyc in a variety of formats and accessible via popular protocols. With this capability, subsets of knowledge in Cyc, either actual assertions in the KB, or the virtual content from external sources, may be accessed through Cyc as a web service, as an object–relational database, or may be customized to implement a client specific api. The power of SKSI is that the syntax of the output is easily modifiable and is completely independent of its content.

So far, we have discussed the role of SKSI as enabling Cyc to communicate with knowledge sources from a very Cyc centric point of view. That is, it is easy to see how SKSI can benefit Cyc, but it is not obvious to see how a developer or user of a single knowledge source may benefit by integrating it with Cyc. Perhaps the most powerful and compelling argument is that integrated sources can be queried through Cyc’s inference engine.

The Cyc Inference Engine is a general theorem prover enhanced by several hundred heuristic level modules which efficiently handle special cases of reasoning that occur frequently. The general theorem prover applies *modus ponens* and *modus tollens* to ground sentences and rules stored in the Knowledge Base to try to prove user’s queries. Whenever the inference engine determines that a special purpose module would more efficiently answer part, or all of a query, it applies the heuristic instead of relying on the theorem prover. The inference engine and knowledge base provide Cyc with question answering capabilities far beyond those of databases and other structured knowledge source managers. They enable Cyc to draw logical conclusions from the data in a knowledge source that could not be obtained by using the source’s native query answering mechanisms alone. In a following section, we will provide explicit examples of this inference augmented question answering capability using two classes of heuristic level reasoning applied to knowledge in a geographical database.

In the next section, we will give a brief overview of the history, current state of development, and future plans for SKSI. Then in the following section we will discuss one of several special purpose heuristic level modules of Cyc’s inference engine and show how its application to query answering extrapolates from the prima facie information in structured knowledge sources to answer queries that may not even be possible to phrase in the source’s native query language (such as SQL). In the last section, we summarize our conclusions and discuss topics for future communications about SKSI research.

2 Overview of SKSI

The SKSI project began in February, 2002 as a referred to <http://www.cyc.com/doc/handbook/oe/02-the-syntax-of-cycl.html>.

result of funding provided by a Small Business Innovation Research grant from the Air Force’s Information Directorate. Since then, Cycorp has used the SKSI technology to integrate the Cyc KB with various structured sources that offer knowledge in different domains, and has developed the capability to integrate object–relational databases as well as queryable web pages and to communicate with software agents that implement CORBA interfaces. Examples of currently available knowledge sources include:

- (i) the Geographic Names Server (GNS) maintained by the National Imagery and Mapping Agency (NIMA)³
- (ii) the Geographic Names Information System (GNIS) maintained by the United States Geological Survey (USGS)⁴
- (iii) the National Weather Service website of the National Oceanic and Atmospheric Administration (NOAA)⁵
- (iv) the Internet Movie Database website⁶
- (v) the Office of Foreign Assets Control (OFAC) database of specially designated nationals and blocked persons with whom US citizens are not permitted to do business.⁷
- (vi) the Cycorp internal WebCalendar program⁸
- (vii) the Cycorp internal Bugzilla program⁹
- (viii) a database of historical terrorist events compiled by the International Policy Institute for Counterterrorism¹⁰
- (ix) a database of historical terrorist events compiled by the Memorial Institute for the Prevention of Terrorism¹¹
- (x) a database of historical terrorist events, threats, and hoaxes related to NBC (Nuclear, Biological, and Chemical) compiled by the Center for Non-proliferation Studies
- (xi) a Common Data Exchange (CDE) database schema developed by Cycorp in collaboration with BBN and Saffron designed to store hypotheses extracted from text by information extraction software

³<http://www.nima.mil/gns/html>

⁴<http://geonames.usgs.gov/>

⁵<http://www.nws.noaa.gov/>

⁶<http://www.imdb.com/>

⁷<http://www.treas.gov/offices/enforcement/ofac/sdn/>

⁸<http://webcalendar.sourceforge.net/>

⁹<http://www.bugzilla.org/>

¹⁰<http://www.ict.org.il/>

¹¹<http://db.mipt.org>

The integration of external knowledge sources has made it possible to increase the number of facts available to Cyc by a factor of several hundred without substantially increasing the size of its knowledge base. The Cyc inference engine can now answer complex queries which draw from knowledge stored in one or more external knowledge sources as well as the Cyc KB. In addition, Cyc users can browse the external knowledge sources in the same way as they browse the KB. They can also translate and store the contents of external knowledge sources into the Cyc KB or a portion of the KB into an external source. Our current work focuses on the following areas:

- (i) inference: developing core inference extensions motivated by SKSI and improving the heuristic level inference support available for knowledge stored in external knowledge sources;
- (ii) content coverage : mapping new knowledge sources into Cyc and extending the SKSI implementation to support new types of knowledge sources;
- (iii) usability : developing user interfaces and knowledge management tools for modelling a knowledge source's schema and querying its contents;
- (iv) versatility: creating new knowledge sources through Cyc, populating them with content, publishing virtual and actual KB content as web services, relational database tables, etc.
- (v) extensibility: developing new interfaces to treat new classes of software as structured knowledge sources, such as UNIX shell scripts (ls, ps, top) and external software agents which conform to an object-broker interface like CORBA;

3 Transitivity reasoning with SKSI

As we mentioned in Section , the Cyc inference engine consists of a general-purpose theorem prover and a number of special-purpose modules that provide heuristic-level support for various classes of inference. One set of modules handles reasoning concerning collection membership (isa) and subset relations (genls, subsetOf). Another set implements fast reasoning for genlPreds, a reflexive and transitive Cyc meta-predicate which states a "subset-like" relationship between two predicates of the same arity. The logical sentence

```
(genlPreds SPEC-PRED GENL-PRED)
```

means that GENL-PRED is a generalization of SPEC-PRED. That is,

```
(GENL-PRED ARG1...ARGN)
```

holds whenever

```
(SPEC-PRED ARG1...ARGN)
```

holds.

In addition to predicate specific heuristics for such transitive predicates, Cyc also implements simple transitivity reasoning and transitivity with respect to a certain argument position for entire classes of predicates. These kinds of heuristic-level support are available for both assertions in the Cyc KB and virtual assertions that reside in external knowledge sources that are connected to Cyc via SKSI.

3.1 Simple Transitivity Reasoning

Simple transitivity (ST) reasoning implements reasoning with transitive predicates that is further enhanced with genlPreds reasoning, at the heuristic-level. This kind of reasoning can formally be expressed by the following CycL rule:

```
(implies
  (and
    (isa ?PRED TransitiveBinaryPredicate)
    (genlPreds ?SPEC1 ?PRED)
    (?SPEC1 ?A ?B)
    (genlPreds ?SPEC2 ?PRED)
    (?SPEC2 ?B ?C))
  (?PRED ?A ?C))
```

ST inferences typically consist of a chain of transitivity steps, which would normally entail the application of the above-mentioned rule at each step. Cyc's special purpose ST reasoning ensures that ST inferences are computed in a more efficient manner, without any need for rule application.

Consider, for example, the following inference that is constructed by Cyc from a number of Cyc KB assertions (which are marked by 'KB') and a few virtual assertions that are made available to Cyc via the Usgs database (marked by 'SKSI'):

Query:

```
Mt : (ContentMtFn Usgs-KS)
HL Formula :
(geographicalSubRegions
  ContinentOfNorthAmerica
  (SchemaObjectFn Usgs-Gnis-LS
    (TheList "Arnold Lake"
      (ScientificNumberFn 3000333 1)
      (ScientificNumberFn -9624333 1))))
```

Answer:

Query was proven True.

Justification :

```
SKSI (geographicalSubRegions
      AustinCounty-Texas
      (SchemaObjectFn Usgs-Gnis-LS
       (TheList "Arnold Lake"
        (ScientificNumberFn 3000333 1)
        (ScientificNumberFn -9624333 1))))
in (ContentMtFn Usgs-Gnis-KS)
```

```
KB (genlPreds
    geopoliticalSubdivision
    geographicalSubRegions)
in DualistGeopoliticalMt
```

```
SKSI (geopoliticalSubdivision
      Texas-State
      AustinCounty-Texas)
in (ContentMtFn Usgs-Gnis-KS)
```

```
KB (geographicalSubRegions
    Southwest-USRegion
    Texas-State)
in UnitedStatesGeographyDualistMt
```

```
KB (geographicalSubRegions
    WesternUS-Region
    Southwest-USRegion)
in UnitedStatesGeographyMt
```

```
KB (geographicalSubRegions
    ContiguousUnitedStates
    WesternUS-Region)
in UnitedStatesGeographyMt
```

```
KB (geographicalSubRegions
    ContinentOfNorthAmerica
    ContiguousUnitedStates)
in WorldGeographyMt
```

One point that is worth further explanation here is the following inference step:

```
KB (genlPreds
    geopoliticalSubdivision
    geographicalSubRegions)
in DualistGeopoliticalMt
```

Even though the original query concerns a geographical-SubRegions relationship, since geopoliticalSubdivision is a specialization predicate of geographicalSubRegions in (ContentMtFn Usgs-KS), geopoliticalSubdivision assertions from the Usgs database are also used in constructing this inference.

3.2 Transitivity via Argument

In addition to ST, Cyc also employs another powerful form of transitivity reasoning at the heuristic level called transitivity via argument (TVA). TVA facts are encoded in the Cyc KB using the meta predicates transitiveViaArg and transitiveViaArgInverse. Here, we focus on transitiveViaArgInverse because the example inference we include in this section is constructed using this predicate.

The CycL meta predicate transitiveViaArgInverse is used to state that a given predicate behaves transitively, in a specified argument place, with respect to a given transitive binary predicate, in reverse argument ordering. This relationship can be expressed using the following pseudo-CycL rule:

```
(implies
 (and
  (transitiveViaArgInverse ?PRED ?BINPRED ?N)
  (?PRED ... ?ARGN ...))
 (genlPreds ?SPEC ?BINPRED)
 (?SPEC ?ARGN-PRIME ?ARGN))
(?PRED ... ?ARGN-PRIME ...)
```

For example the sentence:

```
(transitiveViaArgInverse
 objectFoundInLocation
 subRegions
 2)
```

entails the following rule

```
(implies
 (and
  (objectFoundInLocation ?OBJECT ?LOCATION)
  (subRegions ?REGION ?LOCATION))
 (objectFoundInLocation ?OBJECT ?REGION))
```

Cyc's special TVA reasoning combines this rule with several steps of ST and genlPreds reasoning in order to construct the following inference from a number of Cyc KB assertions and a few virtual assertions from the Usgs database:

Query:

```
Mt : (ContentMtFn Usgs-KS)
HL Formula :
(objectFoundInLocation
 (SchemaObjectFn Usgs-Gnis-LS
  (TheList "Arnold Lake"
   (ScientificNumberFn 3000333 1)
   (ScientificNumberFn -9624333 1))))
ContinentOfNorthAmerica)
```

Answer:

Query was proven True.

Justification :

```
SKSI (objectFoundInLocation
      (SchemaObjectFn Usgs-Gnis-LS
       (TheList "Arnold Lake"
        (ScientificNumberFn 3000333 1)
        (ScientificNumberFn -9624333 1)))
      AustinCounty-Texas)
in (ContentMtFn Usgs-Gnis-KS)
```

```
KB (transitiveViaArgInverse
    objectFoundInLocation
    subRegions
    2)
in UniversalVocabularyMt
```

KB (genLPreds
geographicalSubRegions
subRegions)
in GeographicalRegionGVocabularyMt

KB (genLPreds
geopoliticalSubdivision
geographicalSubRegions)
in DualistGeopoliticalMt

SKSI (geopoliticalSubdivision
Texas-State
AustinCounty-Texas)
in (ContentMtFn UsGis-Gnis-KS)

KB (geographicalSubRegions
Southwest-USRegion
Texas-State)
in UnitedStatesGeographyDualistMt

KB (geographicalSubRegions
WesternUS-Region
Southwest-USRegion)
in UnitedStatesGeographyMt

KB (geographicalSubRegions
ContiguousUnitedStates
WesternUS-Region)
in UnitedStatesGeographyMt

KB (geographicalSubRegions
ContinentOfNorthAmerica
ContiguousUnitedStates)
in WorldGeographyMt

In this section, we discussed various forms of transitivity reasoning in Cyc with example inferences that rely on the SKSI technology. It is important to note that it would not be possible to construct these inferences merely with facts from the Cyc KB or the external database in question. What makes them possible is the integration of facts from these two sources through Cyc's reasoning capabilities.

4 Conclusion

In this article, we discussed the continuing need for semantic, or content based, integration of the massive amount of information available in structured sources, such as databases and web sites, and recalled some of the current approaches to addressing this need. In particular, Cycorp's Semantic Knowledge Source Integration effort is working to address this need by enabling the rapid integration of structured knowledge sources with the Cyc knowledge base by mapping the syntax and semantics of a source into the Cyc ontology. This permits Cyc to treat the information content of each source as if it were part of the knowledge base. The main benefit of doing so is that users can query the external sources through the Cyc inference engine, thus taking advantage of its reasoning capabilities that extend far beyond those of the individual sources. Specifically, we discussed Cyc's special purpose heuristics for handling the transitive properties of

certain logical relations and demonstrated how their application to the content of a database can extrapolate facts that are not explicitly present in the data itself.

This article touched on only a few aspects of the development path, technical challenges, and potential advantages of SKSI. There are a number of other aspects to this technology that should be communicated to the research community at large. One topic that deserves special attention in subsequent articles is the development of the Schema Modelling Toolkit (SMT), a suite of knowledge management tools that will enable lightly trained users (as opposed to trained Cyc specialists) to easily integrate new knowledge sources with Cyc. The main technical challenge to integrating a new source with Cyc is the alignment of the source's semantics with the Cyc ontology. The SMT is intended to address this need by aiding the user in identifying the concepts in Cyc's ontology that correspond to the semantics of the source both automatically and through a clarification process with the user.

References

- [1] J. M. Blanco, Alfredo Goni, and Arantza Illarramendi. Mapping among knowledge bases and data repositories: Precise definition of its syntax and semantics. *Information Systems*, 24(4):275–301, 1999.
- [2] John Davies, Dieter Fensel, and Frank van Harmelen, editors. *Towards the Semantic Web: Ontology-driven Knowledge Management*. John Wiley and Sons, Ltd., 2003.
- [3] D. D. Karunaratna, W. A. Gray, and N. J. Fiddian. Organising knowledge of a federated database system to support multiple view generation. In Alexander Borgida, Vinay K. Chaudhri, and Martin Staudt, editors, *Proceedings of the 5th International Workshop on Knowledge Representation Meets Databases (KRDB '98): Innovative Application Programming and Query Interfaces, Seattle, Washington, USA, May 31, 1998*, number 10 in CEUR Workshop Proceedings, pages 12.1–12.10, 1998.
- [4] Vipul Kashyap and Amit P. Sheth. Semantic and schematic similarities between database objects: A context-based approach. *VLDB Journal: Very Large Data Bases*, 5(4):276–304, 1996.
- [5] James Masters. Structured Knowledge Source Integration and its applications to information fusion. In *Proceedings of the Fifth International Conference on Information Fusion*, pages 1340–1346, 2002.
- [6] Heiner Stuckenschmidt and Holger Wache. Context modeling and transformation for semantic interoperability. In Mokrane Bouzeghoub, Matthias Klusch, Werner Nutt, and Ulrike Sattler, editors, *Proceedings of the 7th International Workshop on Knowledge Representation meets Databases (KRDB 2000), Berlin, Germany, August 21, 2000*, number 29 in CEUR Workshop Proceedings, pages 115–126, 2000.