Technical Report

Distilling Subject Concepts from OpenCyc

Volume 1

Overview and Methodology

TR 08-07-16-B1

July 2008

Acknowledgements

The UMBEL project would like to thank Zitgist LLC for its generous donation of time and resources in programming and writing the documentation for this Technical Report.

UMBEL would also like to thank Cycorp for its support and preparation of a more current OWL version of the OpenCyc knowledge base. We would especially like to thank Larry Lefkowitz for his internal advocacy and answering many questions.

The Cyc Foundation, notably Mark Baltzegar, has been instrumental in helping to guide us through OpenCyc and to share with us many Foundation resources and projects currently in progress. The effort to date would not have been possible without this assistance.

– *Michael K. Bergman*, editor
– *Frédérick Giasson*, editor

UMBEL (Upper Mapping and Binding Exchange Layer) *is a lightweight ontology structure for relating Web content and data to a standard set of subject concepts. Its purpose is to provide a fixed set of reference points in a global knowledge space. These subject concepts have defined relationships between them, and can act as binding or attachment points for any Web content or data.*

Connecting to the UMBEL structure thus provides ***context*** *to Web data. In this manner, Web data can be linked, made interoperable, and more easily navigated and discovered. The project Web site is at http://www.umbel.org.*

UMBEL defines "subject concepts" as a distinct subset of the more broadly understood concept such as used in the SKOS RDFS controlled vocabulary or formal concept analysis or the very general concepts common to some upper ontologies. Subject concepts are a special kind of concept: ones that are concrete, subject-related and non-abstract. We further contrast these with named entities, which are the real things or instances in the world that are members of these subject concept classes. The UMBEL "backbone" is this set of reference subject concepts.

Table of Contents

List of Tables

List of Figures

This volume describes the purposes of UMBEL and the pivotal role to it of OpenCyc[1], the open-source version of the Cyc knowledge base.[2] It describes why OpenCyc was chosen as the basis for the UMBEL subject concepts, and the year-long distillation process and methodology followed to vet OpenCyc into a size and form suitable for UMBEL.[†]

REPORT SERIES OVERVIEW

This document is one of three volumes in the Technical Report series, *Distilling Subject Concepts from OpenCyc*. These three volumes, in combination with the ontology documentation, complete the first public distribution for UMBEL version 0.70, released July 16, 2008.

This current three-volume series describes the selection and vetting of UMBEL's 20,000 subject concepts from OpenCyc. These three volumes are:

- *Distilling Subject Concepts from OpenCyc, Vol. 1: Overview and Methodology*, **TR 08-07-16-B1**, this volume, the basic introduction and explanation of terminology and the distillation process
- *Distilling Subject Concepts from OpenCyc, Vol. 2: Files Documentation*, **TR 08-07-16-B2**, the listing and description of the various files accompanying this process, and
- *Distilling Subject Concepts from OpenCyc, Vol. 3: Appendices*, **TR 08-07-16-B3**, supporting materials and detailed backup.

In addition, key specifications for the UMBEL ontology itself are documented in two volumes:

- *UMBEL Ontology, Vol. 1: Technical Documentation*, **TR 08-07-16-A1**, that overviews the ontology schema, vocabulary and use, and
- *UMBEL Ontology, Vol. 2: Subject Concepts and Named Entities Instantiation*, **TR 08-07-16-A2**, which is an explanation of the N3 files in the ontology distribution.

INTRODUCTION TO UMBEL

UMBEL (Upper-level Mapping and Binding Exchange Layer) is a lightweight ontology for relating named entities and instances or external ontologies and their classes to UMBEL subject concepts. UMBEL subject concepts are conceptually related together using the SKOS[3] and the OWL-Full[4] ontologies. They form a structural 'backbone' comprised of subject concepts and their semantic relationships. And, by linking UMBEL to external ontologies, we *explode the domain* of the linked classes by leveraging its conceptual structure.

Connecting to the UMBEL structure thus provides ***context*** to Web data. In this manner, Web data can be linked, made interoperable, and more easily navigated and discovered. UMBEL's project Web site is at http://www.umbel.org.

UMBEL defines "subject concepts" as a distinct subset of the more broadly understood *concept* such as used in the SKOS/OWL-Full controlled vocabulary, conceptual graphs, formal concept analysis or

[†] References and explanatory notes are found under the concluding Endnotes section.

the very general concepts common to many upper ontologies. We define subject concepts as a special kind of concept: namely, ones that are concrete, subject-related and non-abstract.

UMBEL contrasts subject concepts with abstract concepts and with named entities. Abstract concepts represent abstract or ephemeral notions such as truth, beauty, evil or justice, or are thought constructs useful to organizing or categorizing things but are not readily seen in the experiential world. Named entities are the real things or instances in the world that are themselves natural and notable instances (members) of subject concepts (classes).

More detailed distinctions are provided under *Terminology and Definitions* below.

Purpose

UMBEL is like a map of an interstate highway system, a set of road signs to help find related content and a way of getting from one big place to another. Once in the right vicinity, other maps (or ontologies), more akin to detailed street maps, may then be necessary to get to specific locations or street addresses.

By definition, these more fine-grained maps are beyond UMBEL's scope. But UMBEL can help provide the ***context*** for placing such detailed maps in relation to one another and in relation to the Big Picture of what related content is about.

UMBEL provides the mapping points for the many, many (indeed, millions of) named entities that are the notable instances of its subject concepts. Examples might include the names of specific physicists, cities in a country, or a listing of financial stock exchanges. UMBEL mappings enable us to link a given named entity to the various subject classes of which it is a member.

And, because of relationships amongst subject concepts in the backbone, we can also relate that entity to other related entities and concepts. The UMBEL backbone traces the major pathways through the content graph of the Web. The UMBEL backbone graph can be visualized at large or small scale. The UMBEL subject concept graph can also be applied to any form of Linked Data, be it public, open, proprietary or enterprise.

Basic Approach

UMBEL thus sets for itself objectives that include an identification of subject concepts and their relationships; a premise of emphasizing representational concepts over unattainable precision or exactitude; and a means for relating any notable thing of the world to that structure.

This technical report (TR) focuses mainly on the process whereby 20,000 subject concepts have been distilled from OpenCyc and vetted for their relational structure. This document also touches briefly on the process of mapping 1.5 million named entities that come from Wikipedia, via Yago, to that structure.

Current Distribution

See **Volume 2, Files Documentation**, for a listing of the various zip and constituent files in the current UMBEL subject concept structure; also see http://www.umbel.org/documentation.html.

Structure and Use

The UMBEL backbone is, in essence, a content graph of subject nodes related via a set of relational or hierarchical predicates, or edges. About 800 nodes represent *Abstract Concepts*, and are included for graph integrity and consistency purposes (see below). The current graph thus has 20,896 total concept nodes, 739 of which are abstract, and 48,701 edges.

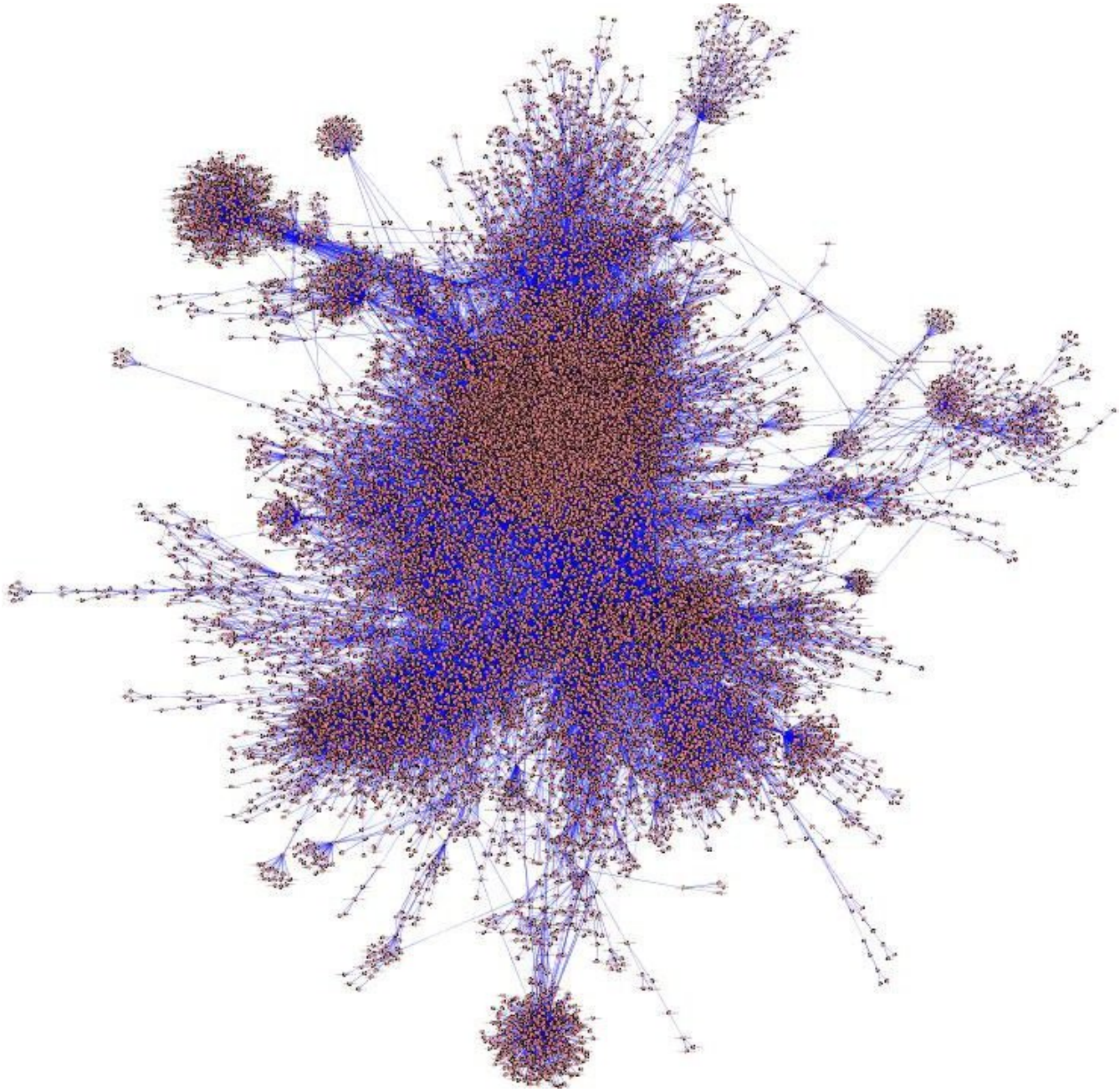

Figure 1. 'Big Graph' in Force-directed View

The 'Big Graph'

The UMBEL 'Big Graph' is shown in *Figure 1* above with all 20,896 nodes. The UMBEL backbone is an OWL-Full ontology where all data instances are described using this ontology and the SKOS Ontology. The ontology specification and vocabulary is documented in the accompanying, *UMBEL Ontology, Vol. 1: Technical Documentation*, **TR 08-07-16-A1**.

That document focuses on the UMBEL ontology and vocabulary; this document focuses on the actual subject concepts used within that structure, and their vetting and selection from OpenCyc.

This Big Graph and its relations are also available as Cytoscape [5] text files; see further http://www.umbel.org/documentation.html.

The adoption of Cytoscape was very useful to the vetting process because we can view the entire subject concept graph, as *Figure 1* shows, manipulate and filter it, report and publish it, show its overlay in relation to the progenitor OpenCyc, and analyze it for completeness and gaps. The use of Cytoscape in this manner is documented in the concluding portions of this report.

While the *Figure 1* view provides a pretty good perspective on large clusters and relationships, it is also hard to see the relationship to individual nodes and labels. Another view – what is termed a *layout* in the software – provides a different way of seeing this same structure.

This is shown in *Figure 2* on the following page. All nodes and edges (blue) are displayed. This is just about at the limit of our graphing program, Cytoscape, which we estimate is limited to about 30 K nodes.

(Use tips for Cytoscape are also provided in Appendix G of the separate volume, *Distilling Subject Concepts from OpenCyc, Vol. 3: Appendices*, **TR 08-07-16-B3**.)

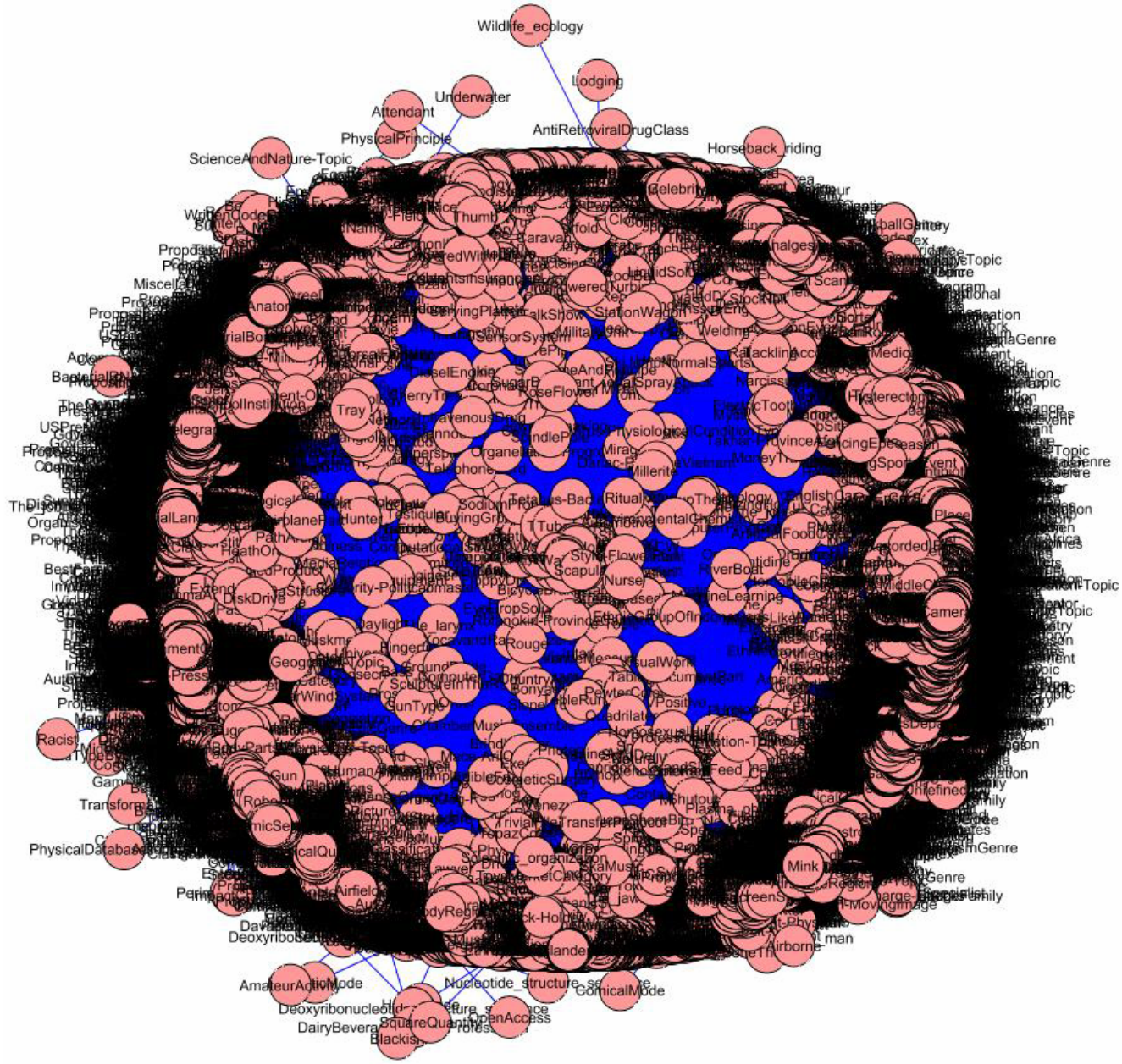

Figure 2. 'Big Graph' in Organic View

The Top 750

Through the manipulation of the topological coefficient, which is a relative measure for the extent to which a node shares neighbors with other nodes, we can zoom in on the Top 750 (actually, 759!) node gateways or hubs; this figure on the next page fairly approximates the upper structure or hierarchy within the graph:

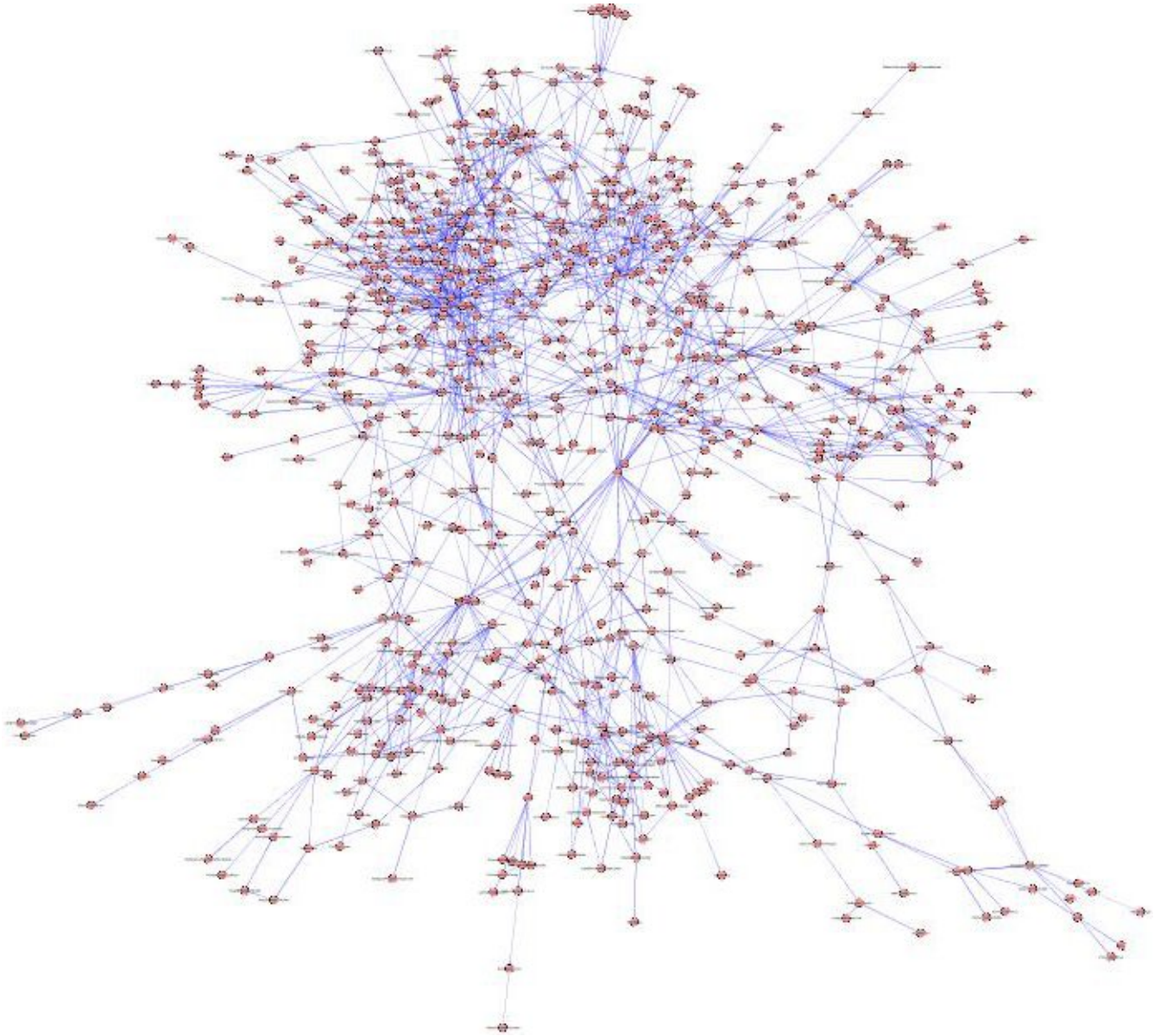

Figure 3. 'Top 750' Nodes

The Top 350

By tightening the coefficient further, we can get a view of the Top 350 (actually, the top 336). Were the system live and not a captured jpeg, we could zoom in and read the actual node labels, as shown by this next figure:

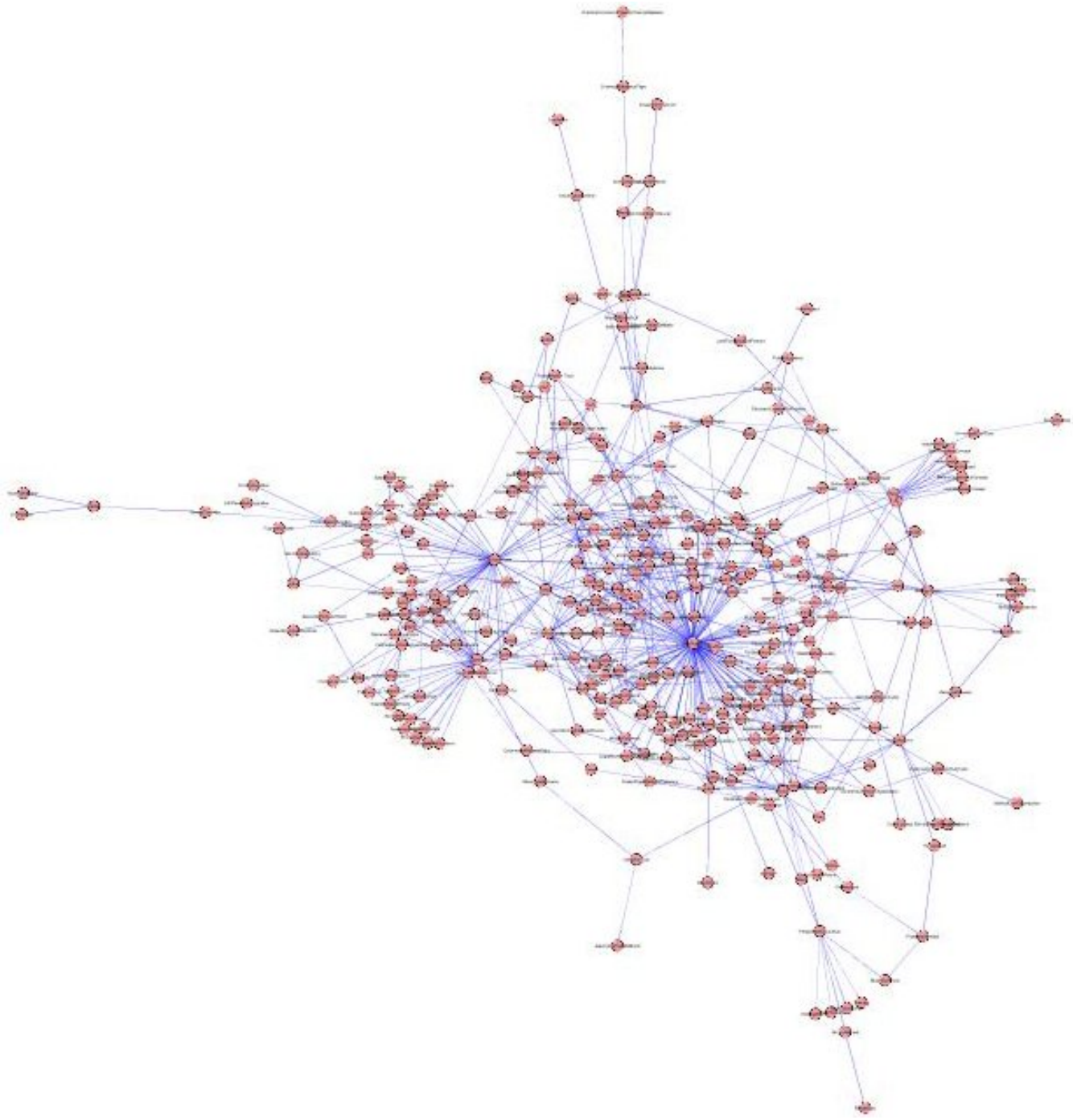

Figure 4. 'Top 350' Nodes

Two Degrees of Separation: Saab Example

The real value from a graph structure, of course, is that now selections can be made based on relationships, neighbors and distances for various reasoning, inference or relatedness purposes. This next diagram begins by inputting "saab", and then getting all nodes within two car concept links:

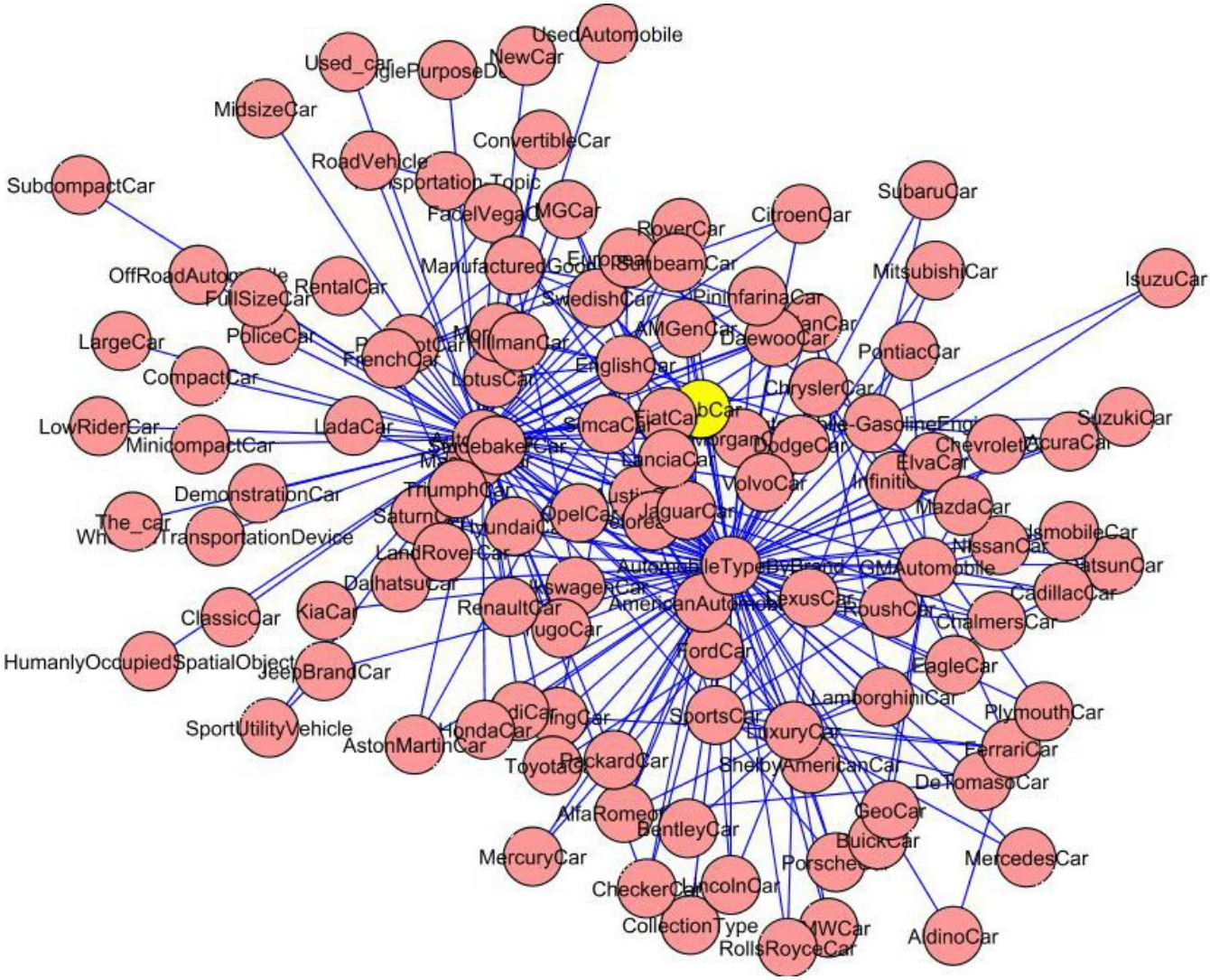

Figure 5. Saab Two Degrees of Separation Example

The Saab Neighborhood

Alternatively, for the same “saab” car concept, we can obtain all directly related links (in yellow), followed by some pruning of car types to make the subgraph more readable and interesting:

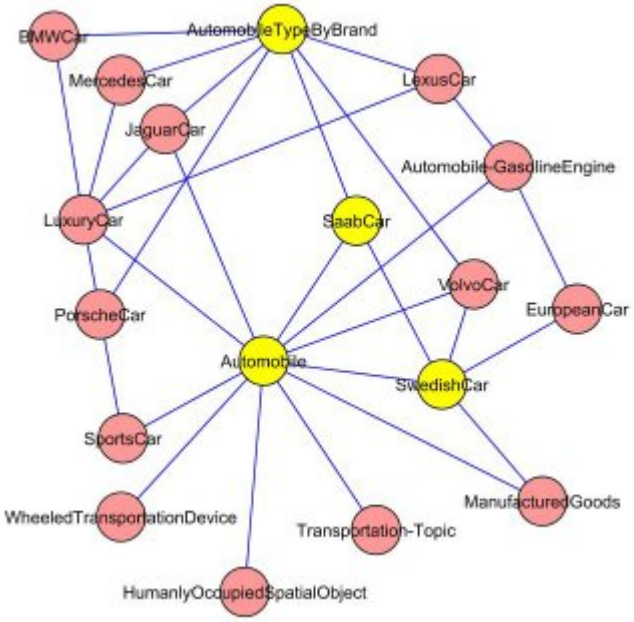

Figure 6. Saab Neighborhood Example

This ability to manipulate and navigate the large UMBEL subject backbone at will brings immense benefits to retrieve and analyze contextually relevant information based on simple rules and the provision of a starting seed node.

Because of its common sense grounding in OpenCyc, these relationships are very logical and clean.

UMBEL Web Services

There are eleven different UMBEL Web services presently offered in conjunction with the ontology and its subject concepts. Description of these services is provided in Appendix H of the separate volume, *Distilling Subject Concepts from OpenCyc, Vol. 3: Appendices*, **TR 08-07-16-B3**. UMBEL's Web services site is at: http://umbel.zitgist.com.

WHY OPENCYC?

The combination of the *representativeness* of UMBEL's subject concepts (the scope of the ontology) and their *relationships* (the structure of the backbone) is fundamental. These factors in turn express the *functional* capabilities of the system. The use of OpenCyc as the source basis for UMBEL is pivotal to these capabilities.

First Things First: The Importance of Context

A reference structure of almost any nature has value. A reference structure provides ***context***, which in turn provides fixed points in the information space for relating distributed datasets to one another. Further, a reference structure of concepts has the further benefit of providing a logical reference structure for instances as well.

While Wikipedia is perhaps the most comprehensive collection of well-known Web instances, no single source can or will be complete in scope. Thus, many public and private sources of entities will emerge as contributing sources.

How do each of these rich instance sources relate to one another? What is the subject concept or topical basis by which they overlap or complement? What is the framework and graph structure of knowledge to give this information context?

These are the benefits brought by a structure of reference concepts, independent from the specifics of the reference structure itself.

Over time, it is likely that a few Web-based reference structures will emerge and compete and get supplemented by still further structures. This evolution is expected and natural and desirable in that it provides choice and options.

Alternative Approaches

Once the importance of ***context*** is embraced, the next logical question is: *What shall be the framework to provide this context?*

Since the Web's inception, there have been various alternatives tried or in ascendance for organizing and bringing structure to Web content. Some of these may be too static and inflexible, others perhaps too arbitrary or parochial. All approaches to date have had little collective success.

Here are some of these alternate approaches:

- *Existing library systems* — Dewey Decimal Classification, Library of Congress, UDC and many other library classification schemes have been touted for the Web. None have enjoyed broad acceptance. Some reasons cited for this failure are physical books are very different from free digital bits; Web schema need to evolve quickly; and lack of stewards and curation
- *Market share* — at various times certain successful vendors have held temporary minor ascendance with content organizational frameworks, generally directory structures. Examples include About, Yahoo!, Open Directory Project (DMOZ), Northern Light, etc. Yet even at their peaks, market shares were low, external adoption was rare, scope was questioned and arbitrary, with interest in directories now nearly absent
- *WordNet* [6]— though of strong interest and use to computational linguists, and quite popular for many content analyses, WordNet has seen little consumer or commercial interest. However, the synset structure and its coverage is extremely valuable for concept disambiguation, and therefore has a role in UMBEL (as it does in many other online systems)
- *Standards efforts* — some sporadic success and some notable failures have occurred in the standards arenas. Generally, the successful initiatives tend to be in close communities where there are clear financial benefits for adherence, such as in the exchange of financial or commerce data; broader and more ambitious efforts have tended to be less successful
- *Professional organizations and associations* — areas such a finance, pharmaceuticals, biologists, physicists and many bounded communities have enjoyed sporadic and sometimes notable success in developing and using domain-specific schema; none have yet transferred beyond their beginning boundaries to the broader Web

- *Government initiatives* — there are episodic successes for government-sponsored initiatives in content organization, mostly in metadata, controlled vocabularies and ontologies, often where contractors or suppliers may be compelled to comply. NIH's National Library of Medicine (and other NIH branches) have also seen significant domain successes, due to its foresight and its receptive biology, genetics and medical communities
- *Upper ontologies* [7]— UMBEL investigated this area considerably in the early months of the project. Most of the upper ontologies have relatively sparse subject concept content, being geared to smaller, abstract-oriented upper structures. Some such as SUMO and DOLCE and now PROTON, have concerted initiatives to extend to middle- and domain-level ontologies. More comprehensive knowledge bases, such as Cyc and its OpenCyc sibling, also may be placed into this category. To date, penetration of these systems into general Web or commercial realms has been quite limited
- *Wikipedia* [8]— a clear and phenomenal success, Wikipedia and related initiatives like Wikinvest and Wikicompany and scores more have proven to be a rich fount for named entities and article-length content. However, so far, such systems have not proven as coherent and useful for the category and content organization structures in which that content is embedded. This is an area of keen academic and collective interest and it may still result in useful organizational schema as these systems mature. [9] However, they have not yet done so, and while a rich source for entities and data, UMBEL decided to pass on their use for backbone structure at this time
- *No collective structure* — tagging or folksonomies or doing nothing have perhaps the greatest market share at present.

An early decision was to minimize development time by basing UMBEL on an extant structure drawn from one of the categories above.

Rationale for OpenCyc

Cyc is the granddaddy of knowledge bases geared to human content and knowledge. Because of its more than 20-year history, Cyc brings with it considerable strengths and some weaknesses.

Amongst all alternatives, Cyc rapidly emerged as the leading candidate for UMBEL. While its strengths warranted close attention, its weaknesses also suggested a considerable effort to overcome them. This combination compelled the need for a significant investigation and due diligence.

First, here are some of OpenCyc's strengths:

- *Venerable and solid* — through an estimated 1000 person-years of engineering and effort, the Cyc structure has been tested and refined through many projects and applications. While a few years back such groundings were unparalleled in the field, we are also now seeing some Internet-wide projects tap into the law of large numbers to get significant inputs of human labor. Cyc has also tapped this venue for ongoing expansion of its KB using the online FACTory game [10]
- *Community* — there is a large community of Cyc users and supporters from academic, government, commercial and non-profit realms. Moreover, the formation of The Cyc Foundation has also served as a vehicle for tapping into volunteer effort as well
- *Upgrade Path* — OpenCyc has an upgrade path to the more capable ResearchCyc, full Cyc and the services of Cycorp

- *Comprehensive* — no existing system has the scope, breadth and coverage of human concepts to match that of Cyc (however, sources for named entities such as Wikipedia have recently passed Cyc in scope)
- *Common sense* — since its founding as a project and then backed by the standalone Cycorp, Cyc has set for itself both a more pragmatic but harder challenge than other knowledge systems. Cyc has set out to capture the common sense at the heart of human reasoning. This objective means codifying generally unstated logic and rules-of-thumb not unlike teaching a baby to walk and talk and read, which often involve often unstated assertions. However, as Cyc has gained this foundation, it has also led to a more solid basis for its reasoning and conceptual relationships
- *Power and inference* — ultimately the purpose of a knowledge base is to support reasoning and inference by computer when presented with a (often small) starting set of assertions or facts. Cyc has literally thousands of microtheories now governing its inference domains, giving it a scope and power unmatched by other systems. The importance of such reasoning is not the silly science fiction of autonomous intelligent robots, but as achievable aids to make connections, determine relationships and filter and order results
- *Robust supporting capabilities* — such knowledge base-wide capabilities can also be deeply leveraged in such areas as entity extraction, machine translation, natural language processing, risk analysis or one of the other dozens of specialty modules available in Cyc, and
- *Free and open* — last, but not least, is the fact that a mostly complete Cyc was released as a free and open source version in 2002. OpenCyc has now been downloaded more than 100,000 times and is in production use for many applications. Non-profits and academics can also obtain access to the full capabilities of the Cyc system through ResearchCyc. This open character is an absolute essential because leading Web applications and leading innovators of the Web eschew proprietary systems.

Another advantage arose after the decision to base UMBEL on OpenCyc. Namely, that advantage is the willingness and assistance of Cycorp and the Cyc Foundation to help contribute to UMBEL's aims. This support has been an essential factor in the preparation of UMBEL for public release. Their ongoing commitment also bodes well for future updates and releases.

Drawbacks to OpenCyc

In its then-current state, OpenCyc had a number of drawbacks in its application to UMBEL and its use with open-world Web content. Most, if not all of these, have now been addressed to some degree of completeness.

But, in that initial investigation, these were some of the areas of concern:

- *Obscure upper ontology* – the Cyc upper ontology, shown in the figure below, is perhaps not as clean as more recent upper ontologies (Proton,[11] for example, is a very clean system). The various sub-classifications of 'Thing' and degrees of "tangibility" seem particularly problematic. However, since these are not direct binding concepts for UMBEL and provide appropriate "glue" for the upper portions of the graph, these criticisms can be easily overlooked

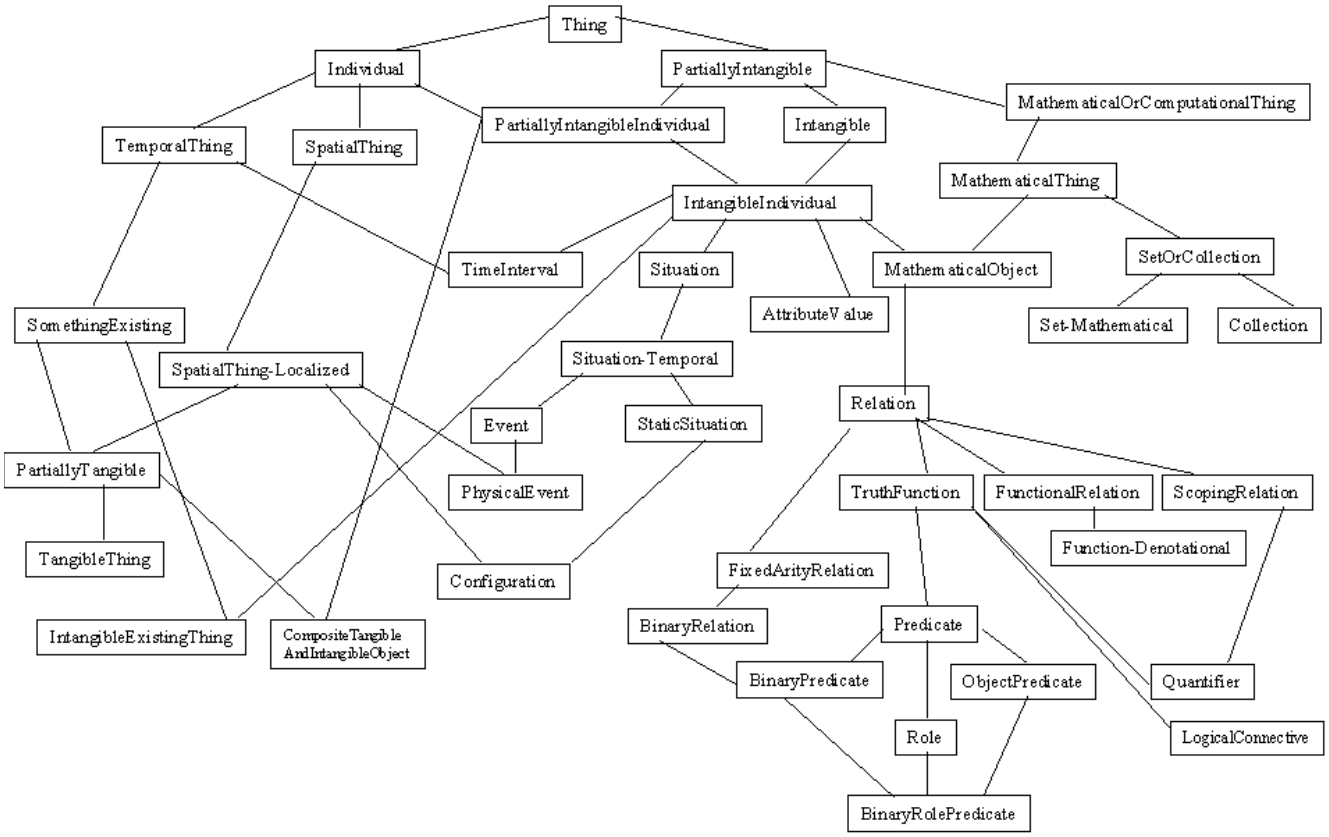

Figure 7. Cyc Upper Ontology

- *Cruft* – twenty years of projects and forays into obscure domains (many for the military or intelligence communities) had left a significant degree of cruft marbled through the knowledge base. Indeed, as latter statistics indicate, perhaps about 30% of the concepts in earlier versions of OpenCyc are holdovers from prior projects or relate to internal Cyc-only topics
- *Reasoning concepts* – another 15% or so of OpenCyc concepts are abstract or for reasoning purposes, such as reasoning over colors, beliefs, the sizes of objects, their orientations in space, and so forth. These are certainly legitimate concepts and appropriate to Cyc's purposes, but not needed or desired for UMBEL's purposes
- *Greater expressivity* – Cyc is grounded in the LISP language and has many higher-order logic constructs. Paradoxically, this greater expressiveness makes translation to the first-order logic (FOL) RDFS basis of UMBEL more difficult
- *Older conventions* – also related to these groundings in an earlier era are the reliance on functions and functional predicates for many relations, and the absence of the current triple data model underlying RDF. There are also conceptual and terminological differences with some Web standards

- *Documentation* – while complete reference materials can be found, it is difficult to do so and introductory and entry-level tutorials could stand to be augmented
- *Named entities* – for many years, but now especially with the emergence of Wikipedia, Cyc has been criticized for its relative paucity of named entity coverage and imbalances of what it does contain. While strictly correct, such criticism misses the mark of Cyc's special purpose and contributions as a solid conceptual and common sense framework. Those framework portions of the system are more immutable, and can be readily mapped to named entity sources. Indeed, perhaps Cyc will now see new vigor as the Web becomes a superior source for contemporary named entity coverage while Cyc fulfills its more natural (and needed) structural role.

As first encountered, one impression of OpenCyc was that of a very solid structure, but somewhat obscured and deserving of a fresh cleaning.

The Decision and Implementation Overview

Nearly five full months of due diligence were devoted to the question of the suitability of OpenCyc as the conceptual and relationship grounding for UMBEL.

On balance, OpenCyc's benefits significantly outweighed its weaknesses. This balance also stood considerably superior to all potential alternatives. An important factor through this deliberation was the commitment of Cycorp and The Cyc Foundation to the aims of UMBEL, and the willingness of those organizations to lend time and effort.

The decision was thus made in October 2007 to base UMBEL on OpenCyc and to undertake the (eventual) two person-years of effort to clean and vet the OpenCyc knowledge base for UMBEL's purposes.

The implementation effort covered by this document was conducted in three phases from October onward:

- **Phase 1: Basic Vetting** – clean up and restrictions from the starting OpenCyc concepts, completed by the end of January 2008
- **Phase 2: Structure Refinement** – an internal UMBEL exercise using the Cytoscape large-graph visualization tool for checking the integrity of the ontology and structure in February 2008
- **Phase 3: Review and Finalization** – external review by Cycorp, the Cyc Foundation, and other interested parties. This latter phase resulted in significant changes to the underlying OpenCyc and took nearly five months, culminating in the current release.

As this process unfolded, the project also made two pivotal decisions with respect to OpenCyc and its use:

1. **All UMBEL subject concepts are based on existing concepts in OpenCyc**. This means UMBEL inherits the proven structure and relationships extant in OpenCyc

2. **No new subject concepts will be added to UMBEL that are not included in OpenCyc**. This means that UMBEL's structure will not diverge from the structural relations already in OpenCyc. This decision preserves the use of UMBEL as a sort of contextual middleware between unstructured Web content and the inferential and tools infrastructure within OpenCyc

(and beyond into ResearchCyc and Cyc for commercial purposes) and back again to the Web. We term this "round-tripping" and the capability is available for any of the 20,000 subject concepts vetted from OpenCyc within UMBEL.[12]

Fortunately, Cycorp has been responsive and made changes to the OpenCyc concept structure and its conversion to OWL in support of needs and observations brought forth by the UMBEL project. This is discussed further under *Phase 3* below.

The Relationship with Cycorp and OpenCyc

UMBEL is thus based on a faithful but reduced subset extraction of concepts and relationships from the OpenCyc version of the Cyc knowledge base. As such, UMBEL is a lightweight reflection of these sources, but not nearly as capable nor complete.

Use and relations based on UMBEL may therefore not be an accurate representation of what might be obtained in working directly with the source Cyc or OpenCyc knowledge bases.

Our project is most grateful for the ongoing involvement and contributions of Cycorp and the Cyc Foundation in support of UMBEL. However, any errors or omissions in UMBEL are the responsibility of the project alone.

TERMINOLOGY AND DEFINITIONS

To understand what was to be vetted and distilled from OpenCyc it is necessary to understand the three basic definitions (plus one) at the core of UMBEL. After definitions, these terms are contrasted to help show the distinctions.

Basic Definitions

Here are the basic definitions used by UMBEL. To understand them and their distinctions is to understand the nature of the UMBEL subject concept backbone.

Subject Concepts

Subject concepts are a special kind of concept: namely, ones that are concrete, subject-related and non-abstract. Note in other systems or ontologies, similar constructs may alternatively be called *topics*, *subjects*, *concepts* or perhaps *interests*. UMBEL has adopted the term *subject concept* to distinguish from these uses, which have different nuances of meaning and use, as well as to highlight the subject or topic nature of UMBEL's concrete concepts.

All subject concepts are a class and while they have a preferred label (using SKOS terminology), they are *representative* or a *proxy* for that concept, and not to be confused with the thing itself. Every UMBEL subject concept can be expressed and referred to by a different preferred label in alternate languages.

Indeed, in a given language, different preferred labels may be swapped out without affecting the identity or use of the subject concept itself. This labeling entity is known as *hasSemset* and is defined below. Each subject concept is related to at least one s*emset*.

Subject concepts are the core constituents to the UMBEL framework. **All subject concepts are**

based on existing concepts in OpenCyc. About 20,000 of them have been distilled and are part of the UMBEL backbone.

Semsets

Semsets are semantically close terms or phrases synonymous or nearly so with the meanings of a subject concept. *Semsets* are akin to WordNet *synsets* or Cyc *aliases*, but can also include more contemporary jargon or slang as may be drawn from Web tagging or folksonomies. (For example, Web 2.0, Web 20, web20, web_20, web-20, etc., can be expanded variants.) The term *semset* has been chosen to distinguish this consolidated meaning.

Semsets may apply to either *subject concepts* or *named entities*. In the latter case, their use is closer to the sense of an alias (such as nicknames, or "great satan" or "uncle sam" for the "United States").

Semsets are related to subject concepts and named entities by the property *umbel:hasSemset*. Each semset has one, and only one, preferred label, which is the standard handle for referring to its governing subject concept or named entity. Alternatively, a semset can have zero or more alternative labels. Each semset has a relation to a Lingvoj[13] instance to refer to the language used to write the semset.

Abstract Concepts

Abstract concepts represent abstract or ephemeral notions such as truth, beauty, evil or justice, or are thought constructs useful to organizing or categorizing things but are not readily seen in the experiential world. They are included in the UMBEL specification because they help maintain the integrity of the UMBEL subject concept graph.

Like subject concepts, abstract concepts are based strictly on those already in OpenCyc.

Unlike subject concepts, abstract concepts may ***not*** be used as binding points to external ontology classes or named entity instances.

For various domain extraction or relatedness determinations, abstract concepts may be excluded from UMBEL's internal processing. They are kept in the baseline UMBEL graph, though may only be shown as intersection points in some visualizations.

Named Entities

Named entities are the real things or instances in the world that are themselves natural and notable class members of subject concepts. Named entities are the instances of the *subject concepts* in the standard definition of the term.

Subject Concepts v. Abstract Concepts

The following table helps draw the distinction between subject concepts and abstract concepts. Please note that Appendix B, Listing Of Abstract Concepts,[14] provides the full listing of the 740 or so abstract concepts presently within UMBEL. Looking at those can help draw the distinction.

Subject Concepts	Abstract Concepts
▪ Nouns or noun phrases ▪ These are concrete kinds of things or ideas in	▪ These are either: 1) abstract (truth, beauty, evil) concepts, or

Subject Concepts	Abstract Concepts
the real world ▪ Broad, collective, reference concepts, often hierarchically related ▪ Similar to "topics" or "subjects", these other terms are used in somewhat different ways in alternative schemas ▪ Collections or classes of like "kinds" of items ▪ Quite stable in scope, breadth and structure ▪ Grounded in the OpenCyc knowledge base, which is the source of its relationships and graph structure ▪ Named entities are members of subject concepts ▪ Are the binding points to external ontology classes or named entity instances	2) artificial thought constructs for organizing things but not encountered as standalone concepts in their own right (*e.g.*, PartiallyTangibleThing) ▪ Collections or classes of like "kinds" of items ▪ Class members may be either other abstract concepts or subject concepts ▪ Class members are never named entities ▪ Tend to reside higher in the subsumption structure ▪ Hidden from the UMBEL subject concept reference "backbone" structure ▪ Grounded in the OpenCyc knowledge base, which is the source of its relationships and graph structure ▪ Can not be used as binding points to external ontologies or named entity instances

Table 1. Subject Concepts v. Abstract Concepts

Subject Concepts v. Named Entities

The following table helps draw the distinction between subject concepts and named entities. Please refer to Appendix C, Named Entity Assignments[14] for a description of certain “gray” categories as to whether they should be treated as one or the other.

For example, most geographical places clearly belong to the named entity category. But, on somewhat arbitrary grounds, all nations, countries, states and provinces were assigned as subject concepts so that they would act as classes with other entities mapped to them. It should also be noted that entities or concepts in the gray zone may be treated both as a named entity and a subject concept.

Subject Concepts	Named Entities
▪ Broad, collective, reference concepts. In a hierarchical category structure, subject concepts represent the "root" or "branch" nodes ▪ Nouns or noun phrases ▪ Called "subject concepts" (or sometimes as a shorthand, "concepts"). Similar to "topics" or "subjects", these other terms are used in somewhat different ways in specific in alternative schemas and are therefore not	▪ Atomic, specific objects, often famous or well-known, that belong to reference "types" such as persons, places, organizations, events, products, time intervals, etc. In a hierarchical category structure, named entities represent the "leaves" ▪ Nouns or noun phrases ▪ Called "named entities" not entities alone, to prevent confusion with other general senses of the term "entity" and in keeping with named

Subject Concepts	Named Entities
used interchangeably here ▪ These are not abstract (truth, beauty, evil) concepts, but concrete about kinds of things or ideas in the real world; abstract concepts are often properly part of what are known as "upper ontologies" but they are not applicable for UMBEL's purposes ▪ Collections or classes of like "kinds" of items ▪ Quite stable in scope, breadth and structure ▪ Grounded in the OpenCyc knowledge base, which is the source of its relationships and graph structure ▪ Basis for the UMBEL subject concept reference "backbone" structure ▪ Named entities are members of subject concepts	entity recognition (NER). ▪ Very concrete, atomic entities ▪ The number and scope is fluid and growing, and potentially of huge size as specific objects are named ▪ Often expressed as a proper noun (with some capitalization), but not necessarily so. Common animal, plant, object, substance names also can be named entities ▪ Major sources are Wikipedia (YAGO), and similar such as Wikinvest, Wikicompanies, etc. ▪ Named entities are exclusive from inclusion in UMBEL ▪ Every named entity belongs to at least one subject concept.

Table 2. Subject Concepts v. Named Entities

METHODOLOGY OVERVIEW AND STATISTICS

We now shift gears to discuss the methodology and process of actually distilling OpenCyc into subject concepts. Summary statistics at key stages of the process are also presented.

Overall, this vetting process resulted in distilling out about 88% (or 8 out of 9) of all initial OpenCyc concepts, reducing from the starting 173,700 concepts to the final count of 20,157 UMBEL subject concepts (and 739 abstract concepts).

Random sampling also suggests, however, that the current vetting process may have missed another *bona fide* 2500 subject concepts still within OpenCyc. It is anticipated the number of concepts in UMBEL will fluctuate for some period to come.

Three Phases, Twelve Rounds

The distillation of OpenCyc to UMBEL can best be summarized as a process of successive refinement. Each step of the process led to new insights and understanding of the knowledge base, learning that was then applied to its subsequent step to further the distillation.

Much of a rough outline for the process noted below was already completed when the final methodology became apparent. This initial trial-and-error walkthrough pointed to better procedures, nomenclature, file naming conventions, and scoring and categorization schema. Armed with this learning, we returned to the beginning and re-ran the process anew with these improved procedures.

This learning enabled us to structure the distillation process into three phases:

- **Phase 1: Basic Vetting** – clean up and restrictions from the starting OpenCyc concepts
- **Phase 2: Structure Refinement** – an internal UMBEL exercise using the Cytoscape large-graph visualization tool for checking the integrity of the ontology and structure
- **Phase 3: Review and Finalization** – external review by Cycorp, the Cyc Foundation, and other interested parties. This latter phase culminated with the current ontology and distribution.

Through the first two phases of review, the analysis progressed through 12 distinct rounds of evaluation. Each round represented a new removal, or "whittling", of existing concepts. Each round, and supporting files if relevant, is discussed below.

The third phase began at the end of February 2008 and benefited from much active involvement by Cycorp. Once translated into the UMBEL structure, there remained some documentation and relationship gaps. These were addressed through further review, each combining Cycorp and UMBEL project efforts.

Process Overview

Because all of these phases and rounds can get a bit confusing, *Figure 8* captures the summary flow of the process of distilling the 173,000 starting OpenCyc concepts to the resulting 20,000 UMBEL subject concepts and 700 organizing abstract concepts (light yellow). This flowchart is not strictly equivalent to the actual phases and rounds followed, but is conceptually accurate and the size of the bubbles are proportional to the amount of whittling that occurred.

The distillation process resulted in nearly 90% of starting OpenCyc concepts being removed (green circles) from UMBEL consideration. Each distillation progression led to a smaller pool of candidates for next steps (bright yellow).

The beginning cut removed Fn concepts (non-atomic terms that define items as a function, Fn, of other items) and verbs and predicates. This left a pool of putative concepts, which then had "reasoning" concepts removed (concepts for describing things and placing them spatially, for example), leaving noun concepts.

Then, internal Cyc concepts and project-specific concepts were excluded, resulting in the final candidate pool. These candidates were then evaluated for segregation into true subject concepts and abstract concepts useful to UMBEL's organization; non-concept entities were also removed at this step, and certain orphans and errors were also removed in final review.

Nearly all issues identified in the early phases of reviewing OpenCyc have either been completely addressed, or are in the process of being so, by Cycorp itself in the scope of the OpenCyc knowledge base or the procedures used to create its OWL versions. Subsequent versions of OpenCyc should have very few residual concerns.

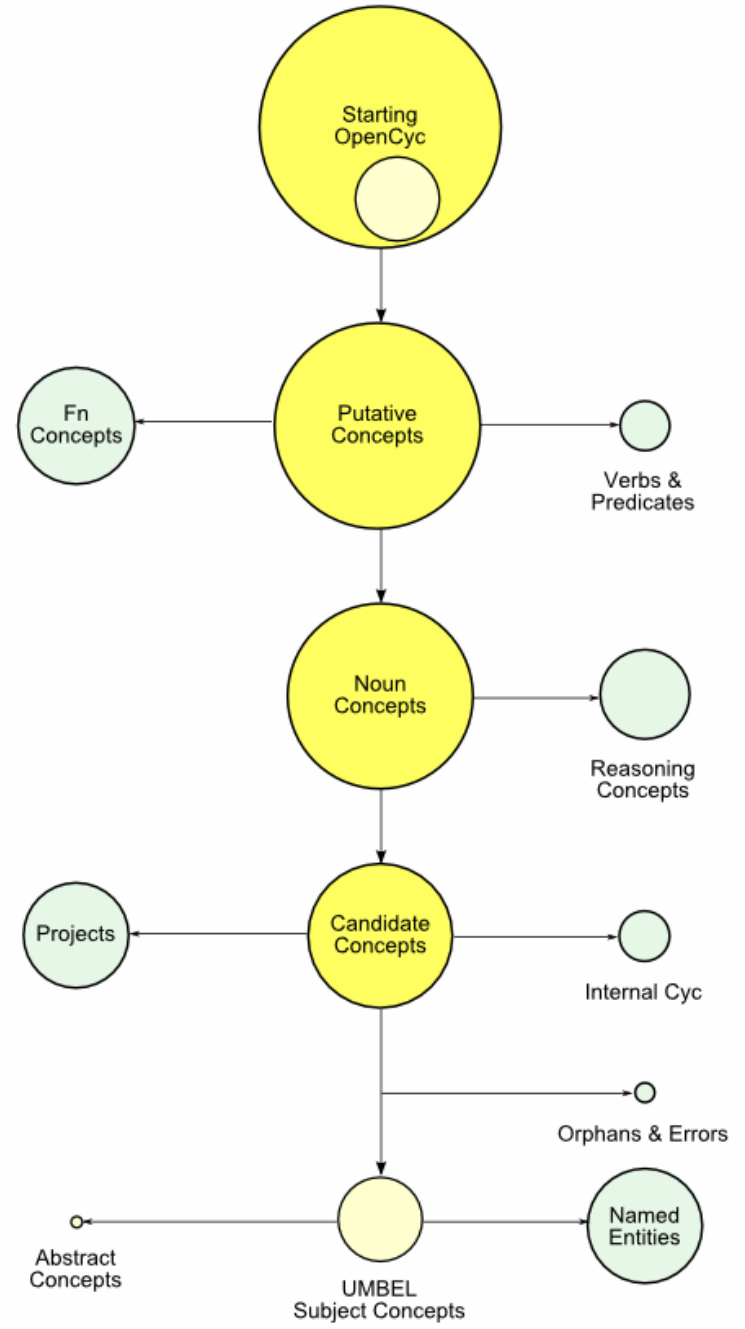

Figure 8. Overview of the OpenCyc Distillation Process

Phase 1 Results Summary

Phase 1 solely involved the listing and inspection of text files from the knowledge base, generally in CSV (comma separated values) files that are easily viewed and manipulated in a standard spreadsheet.

Here is the summary analysis of the Phase 1 basic vetting (note that Classes are the same as OpenCyc Collections):

Category	Classes	Individuals	Missing	TOTAL
Starting	57755	54744	61201	173700
nonFn	51799	52008	29471	133278
Clean (NE + SC)	26931	32939	1996	61866
Named Entities	5530	32494	1101	39125
Subject Concepts	21401	445	895	22741

Table 3. Phase 1 Summary Statistics

Note that the input concepts from OpenCyc were either from collections (classes), individuals or "missing". Missing is a special category that we defined as an OpenCyc entry that is a subject of a triple in the OWL version of the KB, but for which the entry is not defined as a Cyc Individual (`<rdf:type> <cyc:Individual>`) or defined as a class (`<rdf:type> <rdfs:Class>`). (This category was unfortunately named 'Missing' in that the concepts existed but were inferred as classes but not labeled as such in the original OWL file. This issue has subsequently been fixed.)

The process followed for each of the three categories of Classes, Individuals and 'Missing' was nearly identical, with the results of each path then combined to conclude this phase; see further the *Figure 9* flowchart below.

Early rounds in this phase focused on removal of predicates and Fn (functional concepts).

The next rounds, premised that the concepts were now subject concept candidates, became more involved, ultimately resulting in an assignment of a concept to one of about ten categories. Most of these categories reflected removals. Identification for removal could result from the concept being an internal Cyc one and not of external interest; abstract concepts not appropriate to the subject concept focus of UMBEL; or residual concepts from earlier projects. Since games and software were individually large, these were called out separately.

The assignment results from these rounds is shown in the table below:

Category	Classes	Individuals	Missing	TOTAL	TOTAL %
1 – Subject Concepts	21401	445	895	22741	17.06%
2 – Internal Cyc	2592	468	4439	7499	5.63%
3 – Not Useful Concepts	10628	1539	11617	23784	17.85%
4 – Cycorp Projects	6851	9174	6298	22323	16.75%
5 – Verbs, Predicates, Fn	4366	5	2906	7277	5.46%
6 – Software Concepts	409	5527	2213	8149	6.11%
7 – Game Concepts	1	2356	2	2359	1.77%
9 - Named Entities	5530	32494	1101	39125	29.36%
99 – Unclassified	21			21	0.02%
TOTALS	51799	52008	29471	133278	100.00%

Table 4. Phase 1 Concepts and Removals by Category

Appendix D [14] presents detailed discussion of what is included or not in these various OpenCyc Scoring Categories.

Some of the summary findings from this table and the previous *Table 3* are:

- There are some 173 K entries in the starting OpenCyc, about 61 K of which are "missing" (subsequently addressed by the basic OWL-creation procedures)
- About 3/4 of these entries are nonFn; that is non-predicates; these are the candidates for scrutiny
- Of the 133 K or so nonFn entries, about 65% are either project-specific entries or are concepts (such as internal Cyc entries, detailed software or games entries, attitudes, shapes, abstract notions, colors, various collection types, and the like) (the description of the categorization rules and categories are found under Appendix D). **These are not suitable candidates as either broadly useful named entities (NEs) or subject concepts (SCs) !!**
- Thus, only about 1/3 of all entries are suitable candidates for either NEs or SCs
- Detailed review of these candidates suggests on the order of about 23 K SCs (subject concepts) within OpenCyc (note, this estimate is refined to 21 K in Phase 2 and then 20 K in Phase 3)
- Further, external sources such as Wikipedia (via YAGO) are a much stronger resource for named entities than OpenCyc; see also *Table 8*, the OpenCyc-YAGO Named Entity Analysis.

Percent Distillations by Category

Analysis of the various lists shows the magnitude of these distillations. For example, these percentages of Cyc appear to not be directly related to its core purpose or of benefit to general users:

Category	All OpenCyc % Useful	All OpenCyc % Not Useful	nonFn OpenCyc % Useful	nonFn OpenCyc % Not Useful
Fn and Predicates		23.3%		---
Internal Cyc		4.3%		5.6%
Verb and Action Concepts		4.2%		5.5%
Software, etc. Projects		4.7%		6.1%
Games Projects		1.4%		1.8%
Other Projects (many)		12.9%		16.7%
___Subtotal Projects		18.9%		24.6%
Abstract / Reasoning Concepts		13.7%		17.8%
Named Entities	22.5%		29.4%	
Subject Concepts	**13.1%**		**17.1%**	
TOTAL	35.6%	64.4%	46.4%	53.6%

Table 5. OpenCyc Distilation to UMBEL Candidates

Some of the summary findings from this table are:

- Two-thirds of existing OpenCyc entries are not useful to UMBEL's purposes
- Even after excluding more functional (Fn) entries, still more than half of OpenCyc entries are not useful to UMBEL's purposes (the nonFn entries are the basis for remaining points)

- One-quarter of OpenCyc entries seem solely related to past projects and not of general interest or use to the core knowledge base
- About six percent of entries in the KB are related solely to internal Cyc purposes
- A further nearly six percent of entries are related to verbs and actions and not concepts *per se*
- A still further nearly 18 percent of entries, while actually conceptual in nature, are related more to abstract ideas, or modifiers or qualifiers; in other words, not concrete subject concepts
- Thus, fewer than half of the entries in the KB are noun or noun phrase candidates for subject concepts
- And, of these, only about a third of them are actual concepts in nature, as opposed to individual instances or named entities of those concepts
- In sum, only about one of every six or seven entries in the KB is a suitable candidate as a subject concept, equivalent to about 23 K entries.

Phase 2 Results Summary

In Phase 2 the emphasis shifted from categorical removals to structure-wide evaluations. The various rounds in this phase either were aimed at restoring the integrity of the Big Graph by re-establishing linkages for fragments resulting from Phase 1, or making the final refinements of the vetted concepts into subject concepts, abstract concepts or named entities.

The methodology also changed in this phase to rely on Cytoscape and analytic tools. The net result of these efforts was to either add nodes (concepts) or delete them.

This refinement phase led to 1112 subject concepts being added back in, mostly missing members of large classes. One category in particular was the addition of all country names because of the decision on that named entity split (see Appendix C [14]). However, on the order of 50-100 concepts were added to re-establish the integrity of the upper levels of the ontology (as abstract concepts useful for organization and ontology purposes but not concepts generally used in the real world).

This detailed review also led to the identification of earlier errors and orphans, resulting in a further 2272 concepts being removed.

This refinement then led to a concept count of 21581 (522 of which later deemed to be abstract concepts). The summary statistics are as follows:

	Phase 1	Phase 2
Same	20469	20469
Dropped	2272	
New		1112
TOTAL	**22741**	**21581**

Table 6. Phase 2 Changes to UMBEL Concepts

Further analysis based on random checking of specialized concepts found 19 of those checked out of 171 to be within OpenCyc, but not properly picked up in previous review rounds. Extrapolating from this limited sample suggests there may be on the order of 19 overlooked concepts for every 152 actually included in the current system, an overlook rate of 12.5%.

This extrapolation leads to an estimate of perhaps about 2500 *bona fide* subject concepts in OpenCyc that are not presently in the system.

Phase 3 Results Summary

The beta results of Phase 2 enabled the project to put in place some Web services and to do direct graph analysis and tie-ins with Cycorp and Cyc Foundation Web services. This phase marked a shift to Cycorp taking the lead on internal Cyc KB issues.

Cycorp with its various versions of the Cyc knowledge base has a general flow for packaging and delivery as follows:

Cyc → OpenCyc → OWL version of OpenCyc → SemWeb endpoints

All UMBEL material is based on the OWL version of OpenCyc, and the SemWeb endpoints (URI access) are new to the system.

In its efforts during Phase 3, many of which are not listed here and many of which were internally planned and independent of any influence by UMBEL, Cycorp did make changes at the OpenCyc and OWL layers of the above flow. They also added the new category of the SemWeb endpoints, directly useful to UMBEL and UMBEL's users.

Important refinements useful to UMBEL made during this period include:

- General clean-up of the concepts packaged in OpenCyc (attention to projects, redundancies, cruft, etc.)
- Automated ways to add definitions for about 7,000 concepts
- Mapping of more diverse predicates to fit within the SKOS properties of broaderTransitive and narrowerTransitive (such as might be applied to some of the problem relations in areas such as *Fields of Study*)
- More refined and automated methods for the OWL file creation from OpenCyc
- A new "GUID" system for providing unique identifiers for concepts (which is now language neutral, more flexible)
- URI naming of OWL properties to support dereferencing and SemWeb endpoints
- Corrected typos and consistency, and
- Many, many more.

The expression of these changes may be found in UMBEL and completely new and updated OpenCyc and OWL versions of OpenCyc. (See http://www.opencyc.org/ for details; also, updates may be posted a bit later than the release of this report.)

From UMBEL's standpoint, these changes caused a complete renewed inspection of the entire draft ontology and further scripting and packaging of the build routines in going from the OpenCyc OWL to the UMBEL ontology and then to the staged Web services.

This automation should serve the project well as rapid refinements are anticipated as a result of public release.

Also, because of that review, there also was a further refinement to the numbers and structure of the UMBEL ontology. Using a similar format to what has been reported before, here are the statistics:

	Phase 2	Phase 3
Same	20893	20893
Dropped	688	
New		3
TOTAL	**21581**	**20896**
SCs	21059	20157
ACs	522	739

Table 7. Phase 3 Changes to UMBEL Concepts

Version Numbers and Versioning

The current UMBEL ontology is version 0.70.

All efforts through Phase 1 had no version numbers. The first testable ontology in a full graph occurred at the conclusion of Phase 1 and was nominally given the version number 0.01. This corresponded to the 'alpha' version of the ontology.

During Phase 2 when most of the Cytoscape testing and analysis occurred, a further five major interim releases were produced (v. 0.1x to v. 0.5x). By the completion of that testing, an UMBEL version was made available for limited beta testing and Web services were posted online (though still without dereferencable URIs). This 'beta' release was based on version 0.54.

Upon receipt of revised OpenCyc updates from Cycorp in Phase 3, all internal version iterations were in the v. 0.6x series. Thus, with this first current public release, we have incremented to version 0.70.

We anticipate some rapid changes forthcoming in terms of the scope of the subject concepts, further clean up of concept relations, and a move from some of the new UMBEL properties from experimental and unstable.

By the intended time of UMBEL v. 1.00, we anticipate the scope of subject concepts to be relatively stable and questions regarding the ontology vocabulary to be largely settled.

PHASE 1 ROUND UP: BASIC VETTING

We now shift to discuss the specific rounds within each phase, and the files and results associated with each. The basic vetting of Phase 1 involved a succession of evaluations and removals of starting concepts from the OpenCyc knowledge base. As noted, all evaluations in this phase involved manual file inspections.

The basic vetting of Phase 1 is also supported by these appendices:[14]

- Appendix C: Named Entity Assignments

- Appendix D: Opencyc Scoring Categories
- Appendix E: Opencyc Categorization Issues, and
- Appendix F: General Processing Tips

It should be noted that the results of early review rounds have been archived and can be made available upon request (see Volume 2). They are historical and of questionable importance now since all earlier steps have been captured in the latest complete zip distribution. Nonetheless, these older files are noted within **[square brackets]** below.

Basic Vetting Flowchart

The process followed for each of the three categories of Classes (Collections), Individuals and 'Missing' was nearly identical, with the results of each path then combined as this flowchart shows:

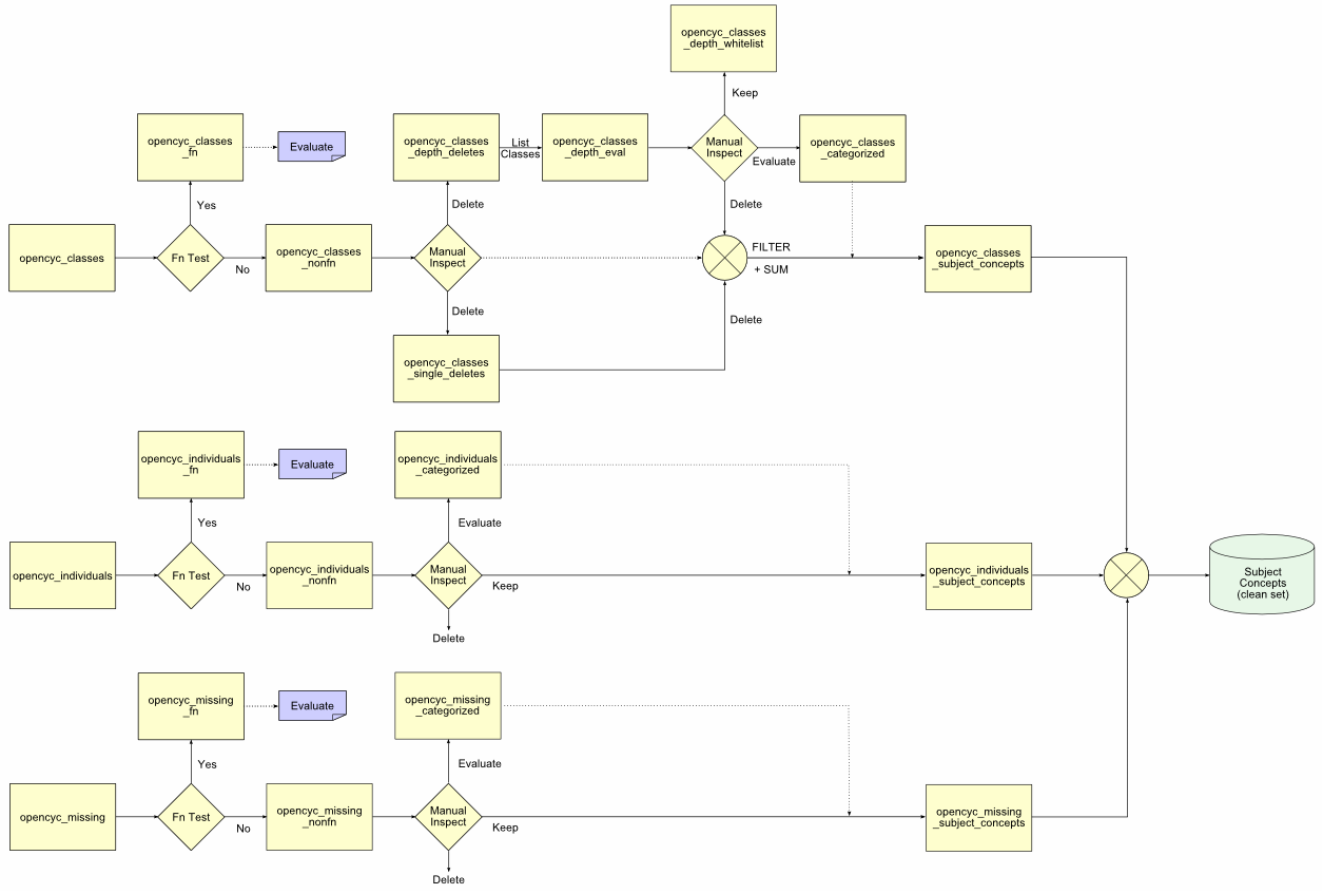

Figure 9. Basic Vetting Flowchart

From first exposure, the analysis and evaluation of the OpenCyc knowledge base (KB) in order to create a "clean", canonical set of subject concepts (SCs) have progressed through a number of stages. These stages, or "Rounds" as they are called below, reflect UMBEL's growing knowledge and experience with the KB.

Initial Vetting Rounds

Round 1: Initial Explorations

Prior to about 2007-10-31, various probes and explorations were made against the OpenCyc KB structure. Various clusters and aggregate concepts within OpenCyc including Facets, Domains, Microtheories, Topics, Types, orders, etc., were probed and tested. The objective was to understand the overall structure of the KB and perhaps structural methods for filtering it to meet UMBEL's purposes.

This period was mostly one of learning and many apparent structural constructs proved to not meet our filtering needs.

The early test files may be found in **[OpenCyc_pre071031.zip]**.

Round 2: Probing Collections (Classes)

Initial investigations had led to a focus on the OpenCyc "collections", the corresponding concept to OWL classes. Through inheritance and other aspects, various inclusion ("whitelist") and exclusion ("blacklist") lists were assembled based on manual review of the concepts for concrete noun and noun phrases within the KB. This was the first detailed review of all entries within the KB, though understanding was still fairly naïve.

This phase marked the beginning of standard file names and labeling, and the use of storing intermediate results as comma-delimited (CSV) files.

These various review files, especially the lists, may be found in **[OpenCyc_20071107.zip]**.

Round 3: Fn Removal

This inspection led to the first major filtering decision regarding the KB, namely the removal of entries that were designated with the Fn functional designator. These were found to not represent fundamental concepts, but combined concepts that matched ideas like a "fruit" with a "tree" type. As noted elsewhere, the decision to exclude Fns is problematic, and may be wrong in some instances. In general, however, most Fns were observed to have been chosen for nearly predicate reasons. Some of the earlier OpenCyc Fn concepts have been later reified and re-expressed as *bona fide* classes.

This intermediate series of lists has not been archived.

Round 4: Splits into Classes, Individuals and Missing Subsets

This round represented a major turning point in the scrutiny and understanding of the KB. First, it was realized that, in an RDF/OWL sense, that some OpenCyc Individuals more properly should be classified as a class, and that some Collections fit better as individuals or instances. Second, it was also discovered that some subjects of the OWL triples in the OpenCyc conversion were listed as neither OWL classes nor RDF types. In actuality, nearly one-third of all concepts were so treated.

These understandings caused the need, then, to review three categories of OpenCyc entities for subject concept inclusion: Classes (or, Collections in OpenCyc terminology), Individuals, and Missing (the name adopted for the mischaracterized entries).

These added complications also warranted that more strict attention be paid to workflow, naming and documentation, and the use of a wiki for tracking purposes. This phase led to the creation of a formal lists process, documentation of the subject concepts workflows, and other specific documentation.

This round was also the focus of a major manual review of the complete KB.

The list results of this major round is archived under **[OpenCyc_20071119.zip]** or **[OpenCyc_20071121.zip]**. The 20071121 series is more up to date and complete, and includes the SQL processing procedures used in many of the lists creation. The 20071119 series is kept for comparative purposes, since this was a period of very active change in the lists.

Round 5: Detailed Delete Categories

To this point, the review of entries within the KB was largely expressed in binary terms: include or exclude. Yet with growing complexity and growing numbers of exclusions, it was becoming clear that the exclusions fit into various patterns. Some entries were internal to the operation of Cyc; many others were related to past projects and not of broad interest; still others were perhaps too abstract or related to descriptive qualities that were not directly relevant to an objective of canonical subject concepts.

Some of these possible deletions, such as internal Cyc entries, would be needed moving forward but could be excluded from consideration in a specific subject concept mapping. Some of these possible deletions, such as for past projects, might be useful for permanent deletion.

Seeing these distinctions and hoping for sophisticated review by Cycorp or other Cyc users,[15] the project decided to be more precise in its reasons for identifying entries for possible exclusion.

This commitment led to a more detailed categorization scheme, documented as Appendix D, OpenCyc Scoring Categories.[14] Patterned methods for identifying these possible deletions were also developed in many cases. Finally, a complete review of the KB was undertaken to apply these more detailed assignments.

Because of these details, each of the specific list files was separately documented and posted to the internal project wiki. This version is **[OpenCyc_20080107.zip]**.

Round 6: OWL v 1 versus v 2 Difference Analysis

Certain issues and problems discovered in this review brought to light that the OWL version of OpenCyc was some years old and out of sync with recent improvements to the Cyc progenitor. With Cycorp's gracious support, an effort was undertaken by Cycorp to update and improve the OWL version. The new version created from that effort was designated by UMBEL as version 2 (v 2).

Upon release of v 2, an effort was undertaken to analyze and document the differences between the two versions. This effort was also necessary to retroactively apply the previous assessments as to deletion and retention entries. Automated methods were applied to the identical entries between versions, with the deltas between the versions manually categorized.

The results of this analysis are documented and posted under **[OpenCyc_20080111.zip]**.

Round 7: Detailed Delete Categories Applied to OWL v 2

With these basic techniques and procedures complete, all efforts were brought up to date with the new v 2 release. (The v 2 release was an updated OWL file for OpenCyc provided by Cycorp.)

The results of this effort became the new master for moving forward, documented and posted under **[OpenCyc_20080114.zip]**.

Vetting Completed: Round 8: Named Entity (NE) v Subject Concepts (SC)

Independent of these efforts was the growing conviction that a clarifying distinction in the nouns and noun phrases in the candidate entries was warranted. These entries either belonged as named entities (NEs), the atomic instances, or belonged to the various subject concepts (SCs) that were the classes for these entities. Further elaboration of these distinctions and their rationale is provided in *Table 2* above. Ambiguities and uncertainties associated with these distinctions are also discussed in Appendix E, OpenCyc Categorization Issues,[14] and the next section.

The actual review of the full KB to assign these distinctions was very lengthy and represented another complete, detailed manual review.

Of all archival lists, this one is perhaps the most deserving of review. These lists are found and posted under **[opencyc_vetting_20080226.zip]**, one of the three major file distributions for this TR.

OpenCyc-YAGO Analysis

YAGO coverage was also analyzed. The top part of this chart simply recaps the numbers from the recent assignments; the bottom part presents some key ratios, interpreted below.

Category	Classes	Individuals	Missing	TOTAL
OpenCyc-Subject Concept	21401	445	895	22741
OpenCyc-Named Entity	5530	32494	1101	39125
YAGO-Subject Concept	1764	158	35	1957
YAGO-Named Entity	2382	12694	54	15130
YAGO-SC + NE	4259	12853	89	17201
YAGO - Internal NEs *	---	---	---	1531588
YAGO OK %	55.93%	98.76%	60.67%	87.96%
YAGO Error %	41.42%	1.23%	39.33%	11.38%
YAGO Coverage %	43.07%	39.07%	4.90%	38.67%
OpenCyc Coverage %	---	---	---	2.55%

(*) means they are called **Individuals** in YAGO.

Table 8. OpenCyc-YAGO Analysis

- OpenCyc (coverage) covers only a very small percentage -- about 2.5% -- of the 1.5 M named entities in YAGO
- On the other hand, YAGO itself only covers about 39% of the named entities within OpenCyc. This is likely due to some errors in the Oliver string similarity algorithm[16] or, also likely, that there remain project-specific named entities within OpenCyc

- However, where there is overlap, Oliver applied to YAGO did correctly match 88% of the named entities in the OpenCyc dataset, with matches to OpenCyc individuals approaching 99%.

These results suggest a good correspondence of named entities between the two datasets, but also that YAGO is a much, much richer source of the entities than OpenCyc.

PHASE 2 ROUND UP: STRUCTURE REFINEMENT

In Phase 2 the emphasis shifted from categorical removals to structure-wide evaluations. The various rounds in this phase either were aimed at restoring the integrity of the Big Graph (see *Figure 1* or *Figure 2*) by re-establishing linkages for fragments resulting from Phase 1, or making the final refinements of the vetted concepts into subject concepts, abstract concepts or named entities.

Round 9: Graph and Link Analysis

The resulting set of 23 K subject concept (SC) candidate nodes from Phase 1 was analyzed according to two tests in this round.

The first test identified fragments that were separate from the major link connections (the "BigGraph"). Fragments with many internal links likely belonged in the pool, but were possibly lacking a proper connecting parent node (possible inadvertently deleted in one of the prior Rounds). Alternatively, fragments with only one node (“orphans”) were likely mis-assigned (perhaps should have been assigned as a named entity or was an oversight deserving earlier deletion).

The second test looked at incoming (parent) and outgoing (child) classes across the candidate pool. Single total parent link nodes (thus, also by definition, having no children) are leaf nodes and possibly suggest the node is an orphan, and similar to above, deserving earlier deletion. Single total child link nodes suggest a possibly unimportant parent, and therefore possibly deserving deletion. And, very large link counts may be missing intermediate sub-category children (with the existing links skipping directly to grandchildren) or may be missing a critical parent that would also help distinguish children properly.

This round was the major source of deletions and insertions in Phase 2.

Round 10: Mapping of Named Entities

Prior steps had now produced a test structure of sufficient quality and completeness to begin further tests and assignments. In this round, named entities from YAGO were linked to the draft UMBEL structure (see Appendix C, Named Entity Assignments,[14] for more information).

This round produced the following mappings:

Number of named entities in the Named Entities Dictionary:	**1 452 225**
Number of subject concepts in UMBEL:	**21 581**
Number of types used to type named entities:	**57 366**
Number of types linked to UMBEL subject concepts:	**3785**
Number of named entities linked to UMBEL subject concepts:	**239 208**

Table 9. Mapping of YAGO Named Entities to UMBEL

Round 11: Graph Visualization

In this stage, we adopted the Cytoscape graph visualization software (actually developed for bioinformatics use) in order to "see" all candidate subject concepts. Through filtering and other techniques, we are able to identify key concept "hubs" and to see lightly connected or largely disconnected portions of the network graph (or parts having unusually long connections), or identify other problem areas.

It should be noted that early versions of the Big Graph shown in *Figure 1* and *Figure 2* had many disconnected separate graphs ("fragments") and lacked the contiguous aspect now shown in the figure. The creation of this "giant component" came about through one of the last review rounds, which re-introduced many abstract concepts. This integral structure now appears at many levels of slicing and dicing the ontology, again suggesting pretty fair integrity.

Round 12: Abstract Concepts and Packaging

The resulting UMBEL structure was quite clean and compelling. Though appropriate OpenCyc nodes were missing from the structure due to earlier refinements and deletions, analysis of the graph, the extraction of relations and sub-graphs, and other tests suggested a generally clean structure.

The structure was ready for the final step of this phase to identify the "glue" abstract concepts in the structure.

Like many prior rounds, pattern searches and various attribute sorts were applied to the base graph in CSV format. Key patterns for searching for possible abstract concepts included Thing, Type, Object, By, Concept, Topic, Path and Events. Useful sorts included degree (focusing on the highest numbers), topological coefficient (lower scores being of more interest), and eccentricity.

The result of this final vetting is the Appendix B, Listing Of Abstract Concepts.[14] About 522 concepts were so identified through this process. These files have been made available as **[umbel_final_20080226.zip]** and **[umbel_cytoscape_20080226.zip]**, with the latter as input files to Cytoscape.

PHASE 3: REVIEW AND FINALIZATION

See the section on the *Phase 3 Results Summary* on p. 24 above. That section presents the relevant details, since much of the activity during this phase was conducted by Cycorp.

Also, see *Distilling Subject Concepts from OpenCyc, Vol. 2: Files Documentation*, **TR 08-07-16-B2**, for the listing and description of the various files accompanying this release.

Finally, for direct downloads of files or ontology or documentations, see http://www.umbel.org/documentation.html.

CONCLUSIONS

The benefits of UMBEL will appear in its use. Early testing and application to internal projects have shown some unique and tremendous advantages for relating to the UMBEL structure and deriving

relationships between concepts and entities from it. Its use as an organizing and reference framework for external ontologies – existing and new ones whose construction can also be aided by UMBEL – is proving itself daily.

Actual use of UMBEL – plus ongoing analytic techniques being applied to its graph completeness and integrity – should cause rather rapid improvements and updates to its roster of subject concepts. The methods used for such improvements will be documented as addenda to the current technical volumes. We also expect to see rather frequent additions to the portfolio of UMBEL Web services.

The continued attentiveness of Cycorp to ongoing improvements in OpenCyc and its inter-relationships with UMBEL is also gratifying. This commitment bodes well for the longevity and ability to adapt for UMBEL.

ENDNOTES

[1] OpenCyc is the open source distribution of the Cyc knowledge base, first released in early 2002. The OpenCyc knowledge base and APIs are available under the Apache License. There are presently about 100,000 users of this open source KB, which has the same ontology as the commercial Cyc, but is limited to about 1 million assertions. The most recent release is v. 1.02. An OWL version is also available, and is the basis for UMBEL's work . The Cyc Foundation is the non-profit organization formed to help promote further extensions and development around OpenCyc.

[2] Cyc is a common sense knowledge base developed and commercialized by Cycorp, Inc, Austin, TX; the project and company is an outgrowth of the early MCC advanced computing initiative, which spun Cycorp off as a separate company in 1994. Cyc is a renowned leader in artificial intelligence and machine learning. Cyc has been applied to many specialty projects and domains over its history, resulting in rich knowledge and assertions across all areas of human knowledge. Today, Cyc has contains more than: 16,000 predicates; 1,000 reasoning modules; 300,000 concepts; 4,000 physical devices; 400 event-participant relationships; 11,000 event types; 171,000 "names" (chemicals, persons, places, etc.); 1,100 geospatial classes and 500 goespatial predicates; and 3.2 million assertions.

[3] http://www.w3.org/TR/2008/WD-skos-reference-20080125/

[4] http://www.w3.org/TR/owl-ref/

[5] Cytoscape is a bioinformatics software platform that is well-suited for visualizing and analyzing large-scale graphs of any form. Cytoscape is partially based on GINY and Piccolo, among other open-source toolkits. It has many third-party plug-ins and is being continuous extended into general RDF and ontology use.

[6] WordNet® is a large lexical database of English provided in open source by the Cognitive Science Laboratory of Princeton University. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. See http://wordnet.princeton.edu/.

[7] Examples include the Suggested Upper Merged Ontology (SUMO), the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), PROTON, Cyc, John Sowa's Top-Level Categories and BFO (Basic Formal Ontology). Most of the content in their upper-levels is akin to broad, abstract relations or concepts (similar to the primary classes, for example, in a Roget's Thesaurus), though Cyc is a clear exception with its stated emphasis on capturing "common knowledge."

[8] See http://en.wikipedia.org, for example.

[9] See, for example, this listing of about 100 academic articles devoted to structure and linguistic uses of Wikipedia: http://www.mkbergman.com/?p=417.

[10] FACTory is a game that lets people enter knowledge into the Cyc knowledge base. Via this online game, Cyc tries to determine the truth or falsehood of a series of facts. When enough people have agreed that a fact is true or not, Cyc considers it confirmed and stops asking about it. See http://game.cyc.com/helpfiles/HowToPlay.html.

[11] There are many aspects that make PROTON one of the more attractive reference ontologies. The PROTON ontology (PROTo ONtology), developed within the scope of the SEKT project, is attractive because of its understandability, relatively small size, modular architecture and a simple subsumption hierarchy. However, despite its clean design, the lack of a mature knowledge base guiding its structure made it a less attractive candidate for UMBEL.

[12] More on this role of UMBEL is presented in M. Bergman, *The Role of UMBEL: Stuck in the Middle with You . . .*, May 11, 2008, http://www.mkbergman.com/?p=441.

[13] http://www.lingvoj.org/

[14] See *Distilling Subject Concepts from OpenCyc, Vol. 3: Appendices*, **TR 08-07-16-B3**.

[15] Admirably, this hoped-for sophisticated review indeed has taken place, vindicated the attention to detail.

[16] The PHP Oliver string similarity algorithm, 1993; see http://us2.php.net/similar_text.