Technical Report

Distilling Subject Concepts from OpenCyc

Volume 2

Files Documentation

TR 08-07-16-B2

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

OVERVIEW

This document covers the set of files for documents and supporting data involved in the creation of the UMBEL subject concepts.

The material is presented in reverse chronological order, with current stuff first, oldest archived stuff last.

Because of the nature of the initial vetting process, the bulk of files are actually some of the oldest. Most users will find these older files of little interest.

Version Numbering

Please note that the first complete version of UMBEL was the alpha version 0.01 distributed in February 2008; it represented a months-long vetting process. The first version to receive extensive review and exposure via Web services was the beta version 0.54, first released in April 2008. The first publicly released version of the UMBEL ontology was version 0.70, released July 16, 2008.

Access Directories

All current documentation files for UMBEL may be accessed under this root directory:

- http://umbel.org/doc/, with an overview of the documentation at
- http://umbel.org/documentation.html.

All archived documentation or files may be accessed under this form of root directory:

- *http://umbel.org/doc/vnnn/*.

Where v*nnn* represents the version number of the archived files (such as “v070”).

CURRENT (v 070) TECHNICAL DOCUMENTATION

The 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; see http://www.umbel.org/doc/UMBELOntology_20080716vA1.pdf, 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; see http://www.umbel.org/doc/UMBELOntology_20080716vA2.pdf.

A three-volume series describes the selection and vetting of UMBEL’s 20,000 subject concepts from OpenCyc:[†] [1]

- *Distilling Subject Concepts from OpenCyc, Vol. 1: Overview and Methodology*, **TR 08-07-16-B1**, the basic introduction and explanation of terminology and the distillation process. This

[†] All numbered references are shown under the concluding Endnotes section.

volume, referred to below as "Volume 1" provides the methodology explanation for how most of the files listed in this volume were created. See http://www.umbel.org/doc/SubjectConcepts_20080716vB1.pdf

- *Distilling Subject Concepts from OpenCyc, Vol. 2: Files Documentation*, **TR 08-07-16-B2**, this volume, the listing and description of the various files accompanying this process; see http://www.umbel.org/doc/SubjectConcepts_20080716vB2.pdf, and
- *Distilling Subject Concepts from OpenCyc, Vol. 3: Appendices*, **TR 08-07-16-B3**, supporting materials and detailed backup; see http://www.umbel.org/doc/SubjectConcepts_20080716vB3.pdf.

CURRENT (v 070) ONTOLOGY AND DATA FILES

The UMBEL ontology itself and its instantiation files are separately described in the *UMBEL Ontology, Vol. 2: Subject Concepts and Named Entities Instantiation*, **TR 08-07-16-A2**. See http://www.umbel.org/doc/UMBELOntology_20080716vA2.pdf for a listing of these specific files and access procedures.

Input files of the current UMBEL to Cytoscape [2] are also described in that volume.

ARCHIVE: BETA (v 054) FILES

The beta files were used mostly as input drivers to Web services and for viewing and analysis within Cytoscape.

Concepts Files

The concepts files may be found under the root of http://umbel.org/doc/v054/ as:

- umbel_subjectConcepts.csv
- umbel_abstractConcepts.csv
- umbel_concepts.csv.

Cytoscape Files

The Cytoscape files may be found under the same root, and are either available as text input files to the program differentiated by layout or as graph statistics files:

- umbel_cytoscape-biolayout.cys
- umbel_cytoscape-edge-spring.cys
- umbel_cytoscape-force.cys
- umbel_cytoscape.cys
- umbel_cytoscape_organic.cys
- umbel_directed.netstats
- umbel-undirected.netstats.

ARCHIVE: INTERIM PRE-BETA FILES

During the run-up to the beta release, there were quite a few iterations using Cytoscape. All files in this series are in the basic CSV (comma separated values) format used by Cytoscape in the basic subject-predicate-object input layout. To view them properly, they must first be inputted into the program.

These files, with most recent last, appear under either of the roots of http://umbel.org/doc/v05x/ or http://umbel.org/doc/v01x-04x/:

- umbel_cytoscape_0.csv
- umbel_cytoscape_1.csv
- umbel_cytoscape_2.csv
- umbel_cytoscape_3.csv
- umbel_cytoscape_4.csv
- umbel_cytoscape_5.csv
- umbel_cytoscape_5-1.csv
- umbel_cytoscape_5-2.csv
- umbel_cytoscape_5-3.csv

In essence, this data merely tracks the addition and deletion of various concepts based on interim graph reviews. Other internal project files documented the transition decisions, but are too complicated to post without further explanation. If interested, please contact the project.

ARCHIVE: ALPHA (v 001) FILES

Significant review and vetting of the input OpenCyc files occurred before a full UMBEL graph could even be produced. The results of this process documented in Volume 1^3 was the first alpha release of UMBEL at the conclusion of the Structure Refinement Phase 2.

The lead in prior to that was the Phase 1 Basic Vetting phase. While the individual files in these two phases are described in some detail below, a zip file was also prepared with the file results from each phase.

These zip files, as well as the individual files, may be downloaded from the root directory of http://umbel.org/doc/v001/: The final round files are the ones in the current distribution, and are organized according to the methodology described in Vol. 1.

Three file packages are available for download to work with the current UMBEL set:

- opencyc_vetting_20080226.zip – the files that are the result of the Basic Vetting Phase 1, as described below. This file can be downloaded as a full zip (7,756 KB) or as individual files, as noted below
- umbel_final_20080226.zip – the final files resulting from the Structure Refinement Phase 2, available as a zip (389 KB) or as individual files, and

- umbel_cytoscape_20080226.zip – the majority of this file is zipped in its native *.cys format (4,246 KB), but it does include some calculated statistics that warrant a zip distribution. This is the direct input file including some analytical results useful for viewing and manipulation within the open source Cytoscape large-graph visualization software. It reflects the complete alpha distribution, and therefore is identical to version 0.01.

Phase 2 Structure Refinement Files

Structure refinement was the final step before the initial 'alpha' release that concluded Phase 2. These come in two sets of final files.

Final Concepts

File: umbel_concepts.csv

List Name: umbel_concepts.csv
Description: this is the text file for representing the input triples for both subject and abstract concepts in UMBEL it is the basic import file to Cytoscape
Format: three-columns: the first column is the subject using the canonical name from OpenCyc; the second column is the relationships type (subClassOf or type), and the third column is the object using the canonical name from OpenCyc **[48771 rows]**
Review File: From this point forward, review has taken place via Cytoscape
Aliases: umbel_cytoscape_xxx.csv
Notes: None

File: umbel_abstractConcepts.csv

List Name: umbel_abstractConcepts.csv
Description: These are the subset of input concepts that represent abstract concepts
Format: three-columns: the first column is the subject using the canonical name from OpenCyc; the second column is the relationships type (subClassOf or type), and the third column is the object using the canonical name from OpenCyc **[522 rows]**
Review File: From this point forward, review has taken place via Cytoscape
Aliases: None
Notes: Provided mostly for review purposes

File

Cytoscape Files

File: umbel_cytoscape.cys

List Name: umbel_cytoscape.cys
Description: This is the pivotal input to Cytoscape; this file is based on the force-directed layout and has pre-run network analysis stats (see two files below)
Format: binary

Review File: New saves create new analysis baselines
Aliases: multiples
Notes: None

File: umbel_undirected.netstats

List Name: umbel_undirected.netstats
Description: These are supplementary statistics based on NetworkAnalyzer for the Cytoscape tool; the statistics are based on undirected networks
Format: Text file specific to NetworkAnalyzer and Cytoscape inputs
Review File: None
Aliases: As named per instance
Notes: None

File: umbel_directed.netstats

List Name: umbel_directed.netstats
Description: These are supplementary statistics based on NetworkAnalyzer for the Cytoscape tool; the statistics are based on directed networks
Format: None
Review File: Text file specific to NetworkAnalyzer and Cytoscape inputs
Aliases: As named per instance
Notes: None

Phase 1 Basic Vetting

The basic vetting process for extracting concepts from OpenCyc is described in Volume 1[3]. However, because it is a useful reference point, the diagram below repeats the general vetting flow chart. Note that three separate tracks – classes (or, "collections" in OpenCyc terminology), individuals and missing – follow essentially the same methodology. These three tracks are then combined together to create the listing of basic vetted concepts for UMBEL, which are then the candidates for the concluding subject refinement rounds:

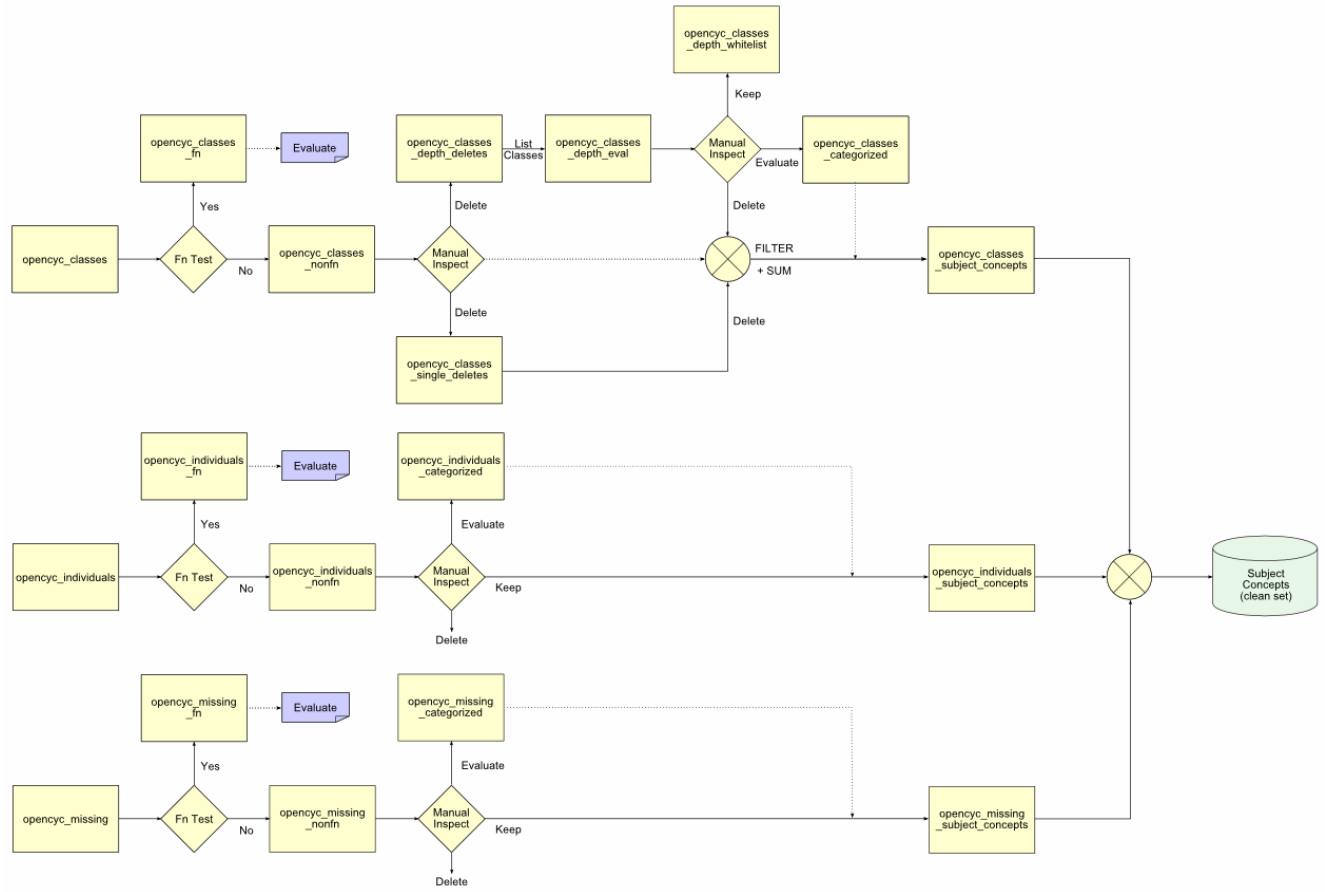

Figure 1. Basic Vetting Process for Subject Concept Candidates from OpenCyc

Classes (Collections)

Input Lists

File: opencyc_classes.csv

List Name: opencyc_classes.csv
Description: the first, baseline listing of Collections (classes) within the OpenCyc KB
Format: two-columns: pretty string, URL listing **[57755 rows]**
Review File: None; starting basis
Aliases: None
Notes: None

File: opencyc_classes_fn.csv

List Name: opencyc_classes_fn.csv
Description: a sublisting of opencyc_classes.csv that segregates out all Collections (classes) with the Fn (function) designation. At this stage, all of these are removed, but they will also

be reviewed and some may be included at a later time with further processing based on a white list. A key change from version 1 is the conversion of many earlier Fn to nonFn entities.

Format: two-columns: label, URL listing **[5954 rows]**
Review File: None
Aliases: None
Notes: None

Lists Resulting from Review

File: opencyc_classes_nonfn.csv

List Name: opencyc_classes_nonfn.csv
Description: this is the result of the first class filtering step, and the basis for all further inspections and filterings; it contains all non-Fn Collections (classes) from opencyc_classes.csv. A key change from version 1 is the conversion of many earlier Fn to nonFn entities.
Format: two-columns: label, URL listing **[51799 rows]**
Review File: MULTIPLES; see subsequent steps
Aliases: None
Notes: None

File: opencyc_classes_depth_deletes.csv

List Name: opencyc_classes_depth_deletes.csv
Description: this is first filter step applied to opencyc_classes_nonfn.csv; the "depth" designator means that all classes on this list, PLUS their children and offspring, should be removed from the KB if so designated
Format: two-columns: URL listing, label **[273 for removal + children]**
Review File: opencyc_classes_depth_deletes_eval.csv
Aliases: formerly called opencyc_class_all_deletionlist.csv
Notes: A whitelist' is necessary to reinstate inadvertent child removals; that is based on a review of opencyc_classes_depth_eval.csv that is documented in opencyc_classes_depth_whitelist.csv

---

File: opencyc_classes_depth_whitelist.csv

List Name: opencyc_classes_depth_whitelist.csv
Description: this is the hand identification of items in opencyc_classes_depth_eval_v2.csv that should be restored back in the system based on an "depth" delete, along with all of their children in one of the first filter steps; whitelist items are marked with a '1' to KEEP (whitelist) in the system
Format: single column: URL listing **[109 for inclusion]**
Review File: No subfiles anticipated
Aliases: opencyc_class_all_whitelist.csv

Notes: None

File: opencyc_classes_yago_analysis.csv

List Name: opencyc_classes_yago_analysis.csv
Description: this is automatic matching file using the Oliver string matching algorithm between entries and aliases in YAGO and entries and aliases in OpenCyc
Format: five-columns: opencyc_concept, opencyc_label, opencyc_alias, yago_ne and oliver score (between 90 and 100)
Review File: The major review file of opencyc_classes_NE-SC_analysis.csv is one result
Aliases: opencyc_clean_classes_ne_extraction_analysis.csv
Notes: None

File: opencyc_classes_NE-SC_analysis.csv

List Name: opencyc_classes_NE-SC_analysis.csv
Description: this is the intermediate results file that is the combination of earlier "clean" files (marks as "1" as accepted) and the YAGO named entity analysis. In addition, a final review for other categories is made in this file, which is then used to update prior evaluations and to create the opencyc_classes_categorized.csv file
Format: four-columns: mark type, YAGO match for NE, label, URL listing
Review File: Direct precursor to the opencyc_classes_categorized.csv file
Aliases: None
Notes: The 'mark type' is according to a multi-valued; see the separate Appendix C; multiple steps must be made to update prior review files (which then become the "clean" output)

Clean Output Lists

File: opencyc_classes_categorized.csv

List Name: opencyc_classes_categorized.csv
Description: this is a the **fully vetted and categorized list** applied to all nonFn classes; the marked items have also been split into various categories according to the **Notes** below
Format: four-columns: mark type, match with YAGO named entity, label, URL listing **[51799 entries]**
Review File: None, though subject_concepts file is an extraction
Aliases: a combination of what was formerly opencyc_classes_deletes.csv and opencyc_classes_clean.csv
Notes: The 'mark type' is according to a multi-valued; see the separate Appendix C; with the '1' categories ("Subject Concepts") moved to the opencyc_classes_subject_concepts.csv file

---

File: opencyc_classes_subject_concepts.csv

List Name: opencyc_classes_subject_concepts.csv
Description: this is a the **fully vetted and categorized list** applied to all nonFn classes; the marked items have also been split into various categories according to the **Notes** below
Format: two-columns: label, URL listing **[21401 subject concepts]**
Review File: None
Aliases: an update and a replacement with a subject concept (SC) emphasis for what was formerly opencyc_classes_clean.csv
Notes: Final output

Individuals

Input Lists

File: opencyc_individuals.csv

List Name: opencyc_individuals.csv
Description: the first, baseline listing of type:Individuals from the OpenCyc KB; the complete listing used for many subsequent comparisons and analysis
Format: two-columns: pretty string, URL listing. **[54744 rows]**
Review File: MULTIPLES
Aliases: opencyc_varied_class_individuals_Individual.csv (four columns, but older first column included parent class, Individuals in all cases)
Notes: Basis for moving forward is _nonfn instead

File: opencyc_individuals_fn.csv

List Name: opencyc_individuals_fn.csv
Description: a sublisting of opencyc_individuals.csv that segregates out all Individuals with the Fn (function) designation. Most of these will be removed, but will reviewed and some may be included on a white list for inclusion
Format: two-column: label, URL listing **[2736 rows]**
Review File: MULTIPLES
Aliases: None
Notes: None

Lists Resulting from Review

File: opencyc_individuals_nonfn.csv

List Name: opencyc_individuals_nonfn.csv
Description: a sublisting of opencyc_individuals.csv that segregates out all Individuals **without** the Fn (function) designation. This is the major listing for subsequent Individuals review

Format: single column: URL listing **[52008 rows]**
Review File: MULTIPLES
Aliases: None
Notes: None

File: opencyc_individuals_yago_analysis.csv

List Name: opencyc_individuals_yago_analysis.csv
Description: this is automatic matching file using the Oliver string matching algorithm between entries and aliases in YAGO and entries and aliases in OpenCyc
Format: five-columns: opencyc_concept, opencyc_label, opencyc_alias, yago_ne and oliver score (between 90 and 100)
Review File: The major review file of opencyc_individuals_NE-SC_analysis.csv is one result
Aliases: opencyc_clean_individuals_ne_extraction_analysis.csv
Notes: None

File: opencyc_individuals_NE-SC_analysis.csv

List Name: opencyc_individuals_NE-SC_analysis.csv
Description: this is the intermediate results file that is the combination of earlier "clean" files (marks as "1" as accepted) and the YAGO named entity analysis. In addition, a final review for other categories is made in this file, which is then used to update prior evaluations and to create the opencyc_individuals_categorized.csv file
Format: four-columns: mark type, YAGO match for NE, label, URL listing
Review File: Direct precursor to the opencyc_individuals_categorized.csv file
Aliases: None
Notes: The 'mark type' is according to a multi-valued; see the separate Appendix C; multiple steps must be made to update prior review files (which then become the "clean" output)

Clean Output Lists

File: opencyc_individuals_categorized.csv

List Name: opencyc_individuals_categorized.csv
Description: this is a the **fully vetted and categorized list** applied to all nonFn individuals; the marked items have also been split into various categories according to the **Notes** below
Format: four-columns: mark type, match with YAGO named entity, label, URL listing **[52008 entries]**
Review File: None, though subject_concepts file is an extraction
Aliases: a combination of what was formerly opencyc_individuals_deletes.csv and opencyc_individuals_clean.csv
Notes: The 'mark type' is according to a multi-value; see the separate Appendix C; with the '1'

categories ("Subject Concepts") moved to the opencyc_individuals_subject_concepts.csv file

---

File: opencyc_individuals_subject_concepts.csv

List Name: opencyc_individuals_subject_concepts.csv
Description: this is a the **fully vetted and categorized list** applied to all nonFn individuals; the marked items have also been split into various categories according to the **Notes** below
Format: two-columns: label, URL listing **[445 subject concepts]**
Review File: None
Aliases: an update and a replacement with a subject concept (SC) emphasis for what was formerly opencyc_individuals_clean.csv
Notes: Final output

Missing

Note: 'Missing' files are where the OpenCyc OWL file shows a subject in a triple, but had not been listed as an instance of *cyc:Individual* or *rdfs:Class*.[4]

Input Lists

File: opencyc_missing.csv

List Name: opencyc_missing.csv
Description: these are legitimate KB entries, but which do not show up as instances of the cyc:Individual class nor as instances of the rdfs:Class class in the standard OpenCyc OWL-Full instances data file;
Format: single column: URL listing **[61201 rows]**
Review File: MULTIPLES
Aliases: opencyc_missing_individuals.csv
Notes: None

File: opencyc_missing_fn.csv

List Name: opencyc_missing_fn.csv
Description: a sublisting of opencyc_missing.csv that segregates out all "missing" types with the Fn (function) designation. These are all removed at this point, but some are set aside for later evaluation and possible re-inclusion once better understanding is gained across the entire OpenCyc KB
Format: single column: URL listing **[31730 rows]**
Review File: None
Aliases: None
Notes: None

Lists Resulting from Review

File: opencyc_missing_nonfn.csv

List Name: opencyc_missing_nonfn.csv
Description: a sublisting of opencyc_missing.csv that segregates out all "missing" entries **without** the Fn (function) designation. This is the major listing for subsequent "missing" review
Format: one columns: URL listing **[29471 entries for review!]**
Review File: None
Aliases: None
Notes: None; the basis for subsequent review

File: opencyc_missing_yago_analysis.csv

List Name: opencyc_missing_yago_analysis.csv
Description: this is automatic matching file using the Oliver string matching algorithm between entries and aliases in YAGO and entries and aliases in OpenCyc
Format: five-columns: opencyc_concept, opencyc_label, opencyc_alias, yago_ne and oliver score (between 90 and 100)
Review File: The major review file of opencyc_missing_NE-SC_analysis.csv is one result
Aliases: opencyc_clean_missing_ne_extraction_analysis.csv
Notes: None

File: opencyc_missing_NE-SC_analysis.csv

List Name: opencyc_missing_NE-SC_analysis.csv
Description: this is the intermediate results file that is the combination of earlier "clean" files (marks as "1" as accepted) and the YAGO named entity analysis. In addition, a final review for other categories is made in this file, which is then used to update prior evaluations and to create the opencyc_missing_categorized.csv file
Format: three-columns: mark type, YAGO match for NE, URL listing
Review File: Direct precursor to the opencyc_missing_categorized.csv file
Aliases: None
Notes: The 'mark type' is according to a multi-value; see the separate Appendix C; multiple steps must be made to update prior review files (which then become the "clean" output)

Clean Output Lists

File: opencyc_missing_categorized.csv

List Name: opencyc_missing_categorized.csv
Description: this is a the **fully vetted and categorized list** applied to all nonFn "missing" entries; the marked items have also been split into various categories according to the **Notes**

	in Volume 3
Format:	three-columns: mark type, match with YAGO named entity, URL listing **[29471 entries]**
Review File:	None, though subject_concepts file is an extraction
Aliases:	a combination of what was formerly opencyc_missing_deletes.csv and opencyc_missing_clean.csv
Notes:	The 'mark type' is according to a multi-value; see the separate Appendix C; with the '1' categories ("Subject Concepts") moved to the opencyc_missing_subject_concepts.csv file

---

File:	opencyc_missing_subject_concepts.csv
List Name:	opencyc_missing_subject_concepts.csv
Description:	this is a the **fully vetted and categorized list** applied to all nonFn "missing" entries; the marked items have also been split into various categories according to the **Notes** in Volume 3
Format:	one column: URL listing **[895 subject concepts]**
Review File:	None
Aliases:	an update and a replacement with a subject concept (SC) emphasis for what was formerly opencyc_missing_clean.csv
Notes:	Final output

ARCHIVE: PRE-ALPHA FILES

Earlier review rounds produced full sets of files (see Volume 1[3] for a discussion of these early rounds). File packages are available for:

- OpenCyc_pre071031.zip
- OpenCyc_20071107.zip
- OpenCyc_20071119.zip
- OpenCyc_20071121.zip
- OpenCyc_20080107.zip
- OpenCyc_20080111.zip, and
- OpenCyc_20080114.zip.

These files are ***not*** directly available online. Please contact the project if you would like to use these files for review or analysis purposes.

ENDNOTES

[1] http://opencyc.org.

[2] Cytoscape is a large-scale graph visualization program developed in the biology community. It is available as open source and is used for large-scale visualization of UMBEL; see http://www.cytoscape.org. Installing and usage tips for Cytoscape are described in Appendix G of *Distilling Subject Concepts from OpenCyc, Vol. 3: Appendices*, **TR 08-07-16-B3.**

[3] *Distilling Subject Concepts from OpenCyc, Vol. 1: Overview and Methodology*, **TR 08-07-16-B1**; see http://www.umbel.org/doc/SubjectConcepts_20080716vB1.pdf.

[4] As Volume 1 above indicates, most of these prior issues with the older OpenCyc starting basis have now been resolved.